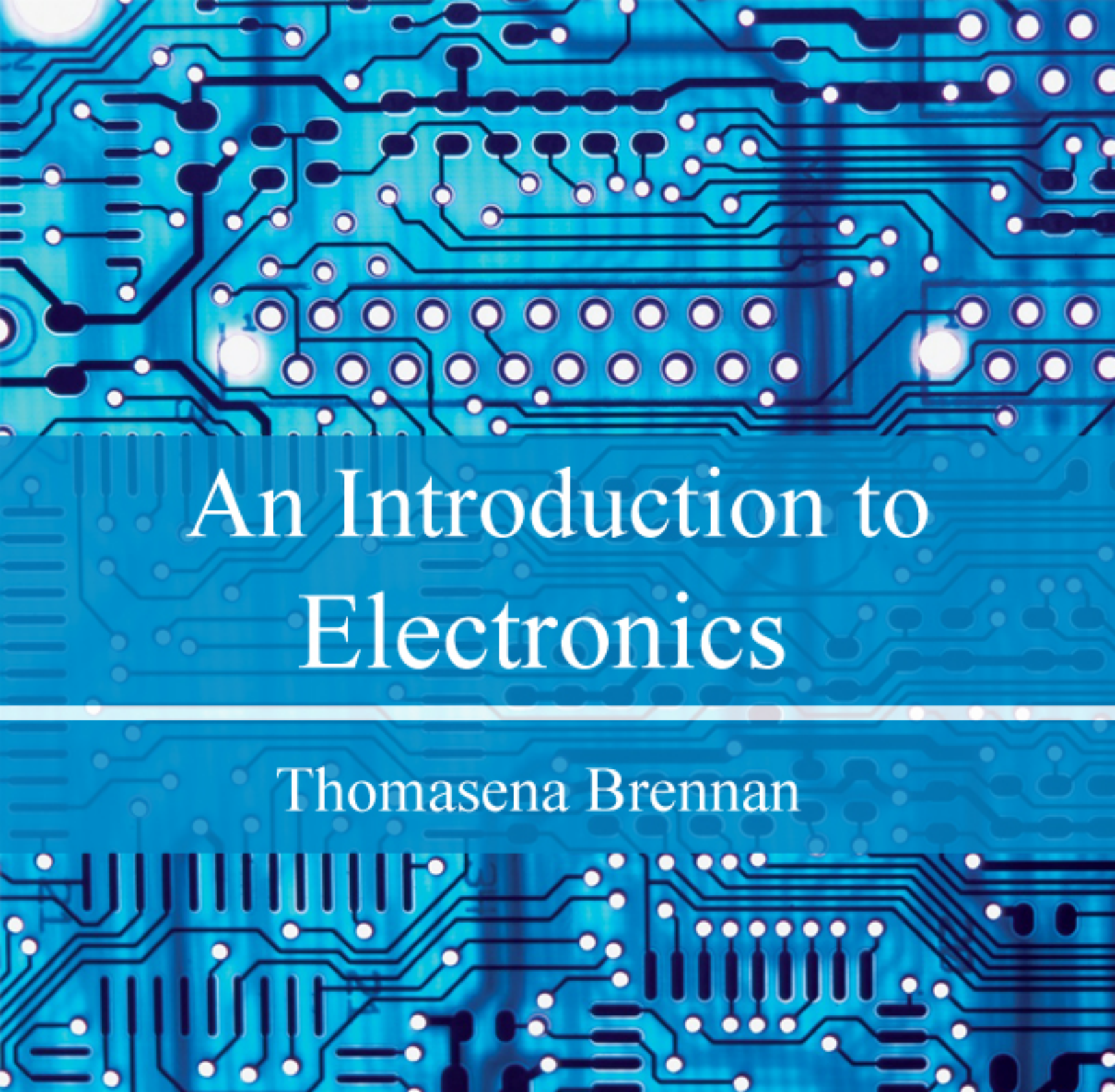


An Introduction to Electronics

Thomasena Brennan

First Edition, 2012

ISBN 978-81-323-4070-6

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

Table of Contents

Chapter 1

Electric Current

Electric current is a flow of electric charge through a medium. This flowing electric charge is typically carried by moving electrons in a conductor such as wire. It can also be carried by ions in an electrolyte, or, it can be carried by both ions and electrons in a plasma.

The SI unit for measuring the rate of flow of electric charge is the ampere, which is charge flowing through some surface at the rate of one coulomb per second. Electric current is measured using an ammeter.

Symbol

The conventional symbol for current is *I*, which may seem puzzling. It originates from the French phrase *intensité de courant*, or in English *current intensity*. This phrase is frequently used when discussing the value of an electric current, especially in older texts; modern practice often shortens this to simply *current* but *current intensity* is still used in many recent textbooks. The *I* symbol was used by André-Marie Ampère himself, after whom the unit of electric current is named, in formulating the eponymous Ampère's force law which he discovered in 1820. The notation travelled from France to England where it became standard, although at least one journal did not change from using *C* to *I* until 1896.

Conduction mechanisms in various media

In metallic solids, electricity flows by means of electrons, from lower to higher electrical potential. In other media, any stream of charged objects may constitute an electric current. To provide a definition of current that is independent of the type of charge carriers flowing, *conventional current* is defined to flow in the same direction as positive charges. So in metals where the charge carriers (electrons) are negative, conventional current flows in the opposite direction as the electrons. In conductors where the charge carriers are positive, conventional current flows in the same direction as the charge carriers.

In a vacuum, a beam of ions or electrons may be formed. In other conductive materials, the electric current is due to the flow of both positively and negatively charged particles at the same time. In still others, the current is entirely due to positive charge flow. For example, the electric currents in electrolytes are flows of electrically charged atoms (ions), which exist in both positive and negative varieties. In a common lead-acid electrochemical cell, electric currents are composed of positive hydrogen ions (protons) flowing in one direction, and negative sulfate ions flowing in the other. Electric currents in sparks or plasma are flows of electrons as well as positive and negative ions. In ice and in certain solid electrolytes, the electric current is entirely composed of flowing ions. In a semiconductor it is sometimes useful to think of the current as due to the flow of positive "holes" (the mobile positive charge carriers that are places where the semiconductor crystal is missing a valence electron). This is the case in a p-type semiconductor.

Metals

A solid conductive metal contains mobile, or free electrons, originating in the conduction electrons. These electrons are bound to the metal lattice but no longer to any individual atom. Even with no external electric field applied, these electrons move about randomly due to thermal energy but, on average, there is zero net current within the metal. Given a surface through which a metal wire passes, the number of electrons moving from one side to the other in any period of time is on average equal to the number passing in the opposite direction. As George Gamow put in his science-popularizing book, *One, Two, Three...Infinity* (1947), "The metallic substances differ from all other materials by the fact that the outer shells of their atoms are bound rather loosely, and often let one of their electrons go free. Thus the interior of a metal is filled up with a large number of unattached electrons that travel aimlessly around like a crowd of displaced persons. When a metal wire is subjected to electric force applied on its opposite ends, these free electrons rush in the direction of the force, thus forming what we call an electric current."

When a metal wire is connected across the two terminals of a DC voltage source such as a battery, the source places an electric field across the conductor. The moment contact is made, the free electrons of the conductor are forced to drift toward the positive terminal under the influence of this field. The free electrons are therefore the charge carrier in a typical solid conductor. For an electric current of 1 ampere, 1 coulomb of electric charge (which consists of about 6.242×10^{18} elementary charges) drifts every second through any plane through which the conductor passes.

For a steady flow of charge through a surface, the current I in amperes can be calculated with the following equation:

$$I = \frac{Q}{t},$$

where Q is the electric charge transferred through the surface over some time t. If Q and t are measured in coulombs and seconds respectively, I is in amperes.

More generally, electric current can be represented as the rate at which charge flows through a given surface as:

$$I = \frac{\mathrm{d}Q}{\mathrm{d}t}\,.$$

Electrolytes

Electric currents in electrolytes are flows of electrically charged particles (ions). For example, if an electric field is placed across a solution of Na^+ and Cl^- (and conditions are right) the sodium ions move towards the negative electrode (cathode), while the chloride ions move towards the positive electrode (anode). Reactions take place at both electrode surfaces, absorbing each ion.

Water-ice and certain solid electrolytes called proton conductors contain positive hydrogen ions or "protons" which are mobile. In these materials, electric currents are composed of moving protons, as opposed to the moving electrons found in metals.

In certain electrolyte mixtures, brightly-colored ions form the moving electric charges. The slow migration of these ions means that the current is visible.

Gases and plasmas

In air and other ordinary gases below the breakdown field, the dominant source of electrical conduction is via a relatively small number of mobile ions produced by radioactive gases, ultraviolet light, or cosmic rays. Since the electrical conductivity is low, gases are dielectrics or insulators. However, once the applied electric field approaches the breakdown value, free electrons become sufficiently accelerated by the electric field to create additional free electrons by colliding, and ionizing, neutral gas atoms or molecules in a process called avalanche breakdown. The breakdown process forms a plasma that contains a significant number of mobile electrons and positive ions, causing it to behave as an electrical conductor. In the process, it forms a light emitting conductive path, such as a spark, arc or lightning.

Plasma is the state of matter where some of the electrons in a gas are stripped or "ionized" from their molecules or atoms. A plasma can be formed by high temperature, or by application of a high electric or alternating magnetic field as noted above. Due to their lower mass, the electrons in a plasma accelerate more quickly in response to an electric field than the heavier positive ions, and hence carry the bulk of the current.

Vacuum

Since a "perfect vacuum" contains no charged particles, it normally behaves as perfect insulator. However, metal electrode surfaces can cause a region of the vacuum to become conductive by injecting free electrons or ions through either field electron emission or thermionic emission. Thermionic emission occurs when the thermal energy exceeds the

metal's work function, while field electron emission occurs when the electric field at the surface of the metal is high enough to cause tunneling, which results in the ejection of free electrons from the metal into the vacuum. Externally heated electrodes are often used to generate an electron cloud as in the filament or indirectly heated cathode of vacuum tubes. Cold electrodes can also spontaneously produce electron clouds via thermionic emission when small incandescent regions (called **cathode spots** or **anode spots**) are formed. These are incandescent regions of the electrode surface that are created by a localized high current flow. These regions may be initiated by field electron emission, but are then sustained by localized thermionic emission once a vacuum arc forms. These small electron-emitting regions can form quite rapidly, even explosively, on a metal surface subjected to a high electrical field. Vacuum tubes and sprytrons are some of the electronic switching and amplifying devices based on vacuum conductivity.

Current density and Ohm's law

Current density is a measure of the density of an electric current. It is defined as a vector whose magnitude is the electric current per cross-sectional area. In SI units, the current density is measured in amperes per square meter.

$$I = \overrightarrow{J} \cdot \overrightarrow{A}$$

where *I* is current in the conductor, **J** is the current density, and **A** is the cross-sectional area. The dot product of the two vector quantities (**A** and **J**) is a scalar that represents the electric current.

Current density (current per unit area) *J* in a material is proportional to the conductivity σ and electric field *E* in the medium:

$$J = \sigma E$$

Instead of conductivity, reciprocal quantity called resistivity ρ, can be used:

$$J = \frac{E}{\rho}$$

Conduction in semiconductor devices may occur by a combination of electric field (drift) and diffusion, which is proportional to diffusion constant *D* and charge density α_q. The current density is then:

$$J = \sigma E + Dq\nabla n,$$

with *q* being the elementary charge and *n* the electron density. The carriers move in the direction of decreasing concentration, so for electrons a positive current results for a positive density gradient. If the carriers are holes, replace electron density *n* by the negative of the hole density *p*.

In linear anisotropic materials, σ, ρ and D are tensors.

In linear materials such as metals, and under low frequencies, the current density across the conductor surface is uniform. In such conditions, Ohm's law states that the current is directly proportional to the potential difference between two ends (across) of that metal (ideal) resistor (or other ohmic device):

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

where I is the current, measured in amperes; V is the potential difference, measured in volts; and R is the resistance, measured in ohms. The letter I stands for the German word, "Intensität" meaning "Intensity". For alternating currents, especially at higher frequencies, skin effect causes the current to spread unevenly across the conductor cross-section, with higher density near the surface, thus increasing the apparent resistance.

Drift speed

The mobile charged particles within a conductor move constantly in random directions, like the particles of a gas. In order for there to be a net flow of charge, the particles must also move together with an average drift rate. Electrons are the charge carriers in metals and they follow an erratic path, bouncing from atom to atom, but generally drifting in the opposite direction of the electric field. The speed at which they drift can be calculated from the equation:

$$I = nAvQ,$$

where

I is the electric current
n is number of charged particles per unit volume (or charge carrier density)
A is the cross-sectional area of the conductor
v is the drift velocity, and
Q is the charge on each particle.

Electric currents in solids typically flow very slowly. For example, in a copper wire of cross-section 0.5 mm^2, carrying a current of 5 A, the *drift velocity* of the electrons is on the order of a millimetre per second. To take a different example, in the near-vacuum inside a cathode ray tube, the electrons travel in near-straight lines at about a tenth of the speed of light.

Any accelerating electric charge, and therefore any changing electric current, gives rise to an electromagnetic wave that propagates at very high speed outside the surface of the conductor. This speed is usually a significant fraction of the speed of light, as can be deduced from Maxwell's Equations, and is therefore many times faster than the drift velocity of the electrons. For example, in AC power lines, the waves of electromagnetic

energy propagate through the space between the wires, moving from a source to a distant load, even though the electrons in the wires only move back and forth over a tiny distance.

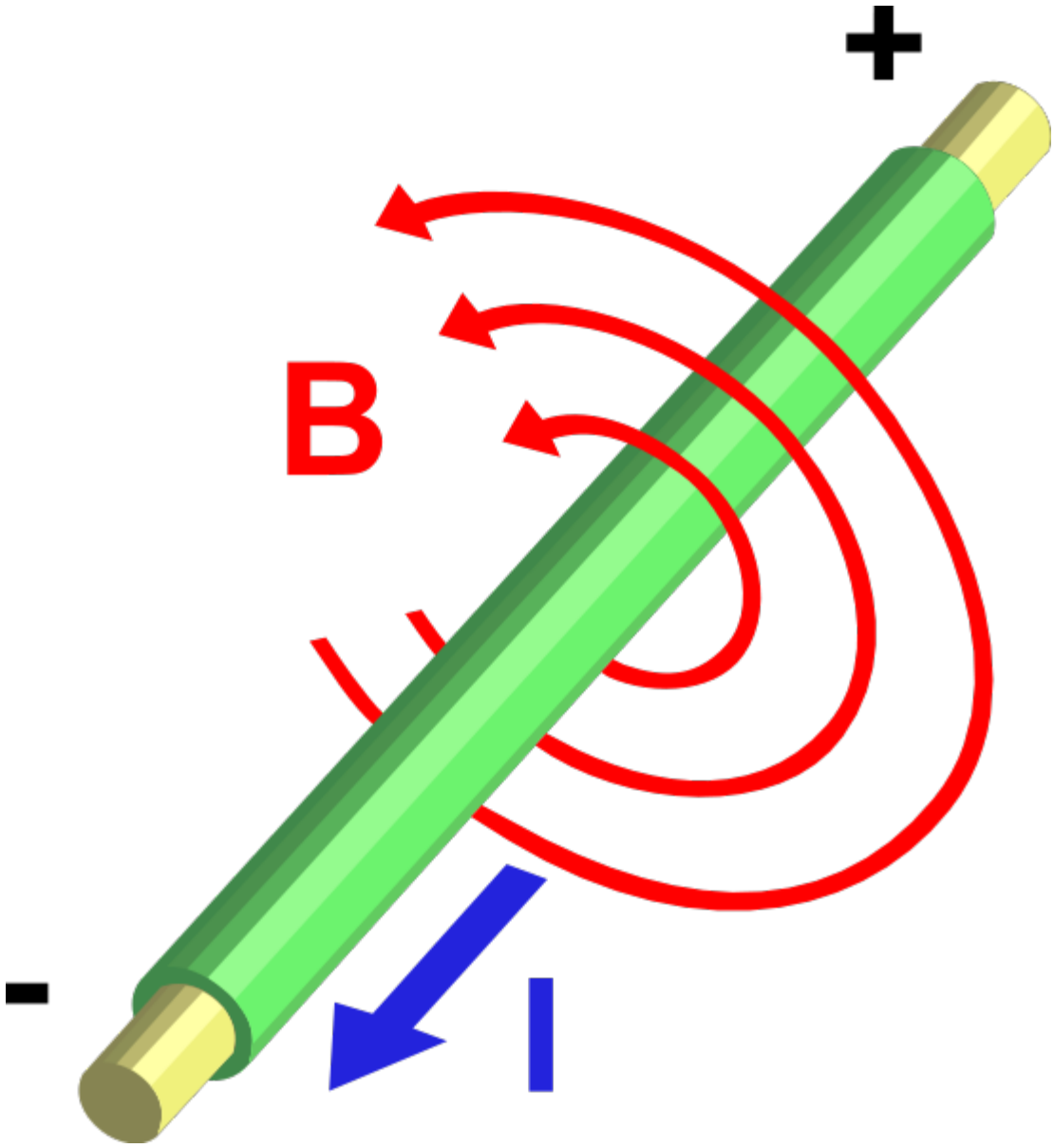
The ratio of the speed of the electromagnetic wave to the speed of light in free space is called the velocity factor, and depends on the electromagnetic properties of the conductor and the insulating materials surrounding it, and on their shape and size.

The magnitudes (but, not the natures) of these three velocities can be illustrated by an analogy with the three similar velocities associated with gases.

- The low drift velocity of charge carriers is analogous to air motion; in other words, winds.
- The high speed of electromagnetic waves is roughly analogous to the speed of sound in a gas (these waves move through the medium much faster than any individual particles do)
- The random motion of charges is analogous to heat - the thermal velocity of randomly vibrating gas particles.

This analogy is extremely simplistic and incomplete: The rapid propagation of a sound wave doesn't impart any change in the air molecules' drift velocity, whereas EM waves do carry the energy to propagate the actual current at a rate which is much, much higher than the electrons' drift velocity. To illustrate the difference: The sound and the change in the air's drift velocity (the force of the wind gust) cross distance at rates equaling the speeds of sound and of mechanical transmission of force (**not higher** than rate of drift velocity); while a change in an EM field and the **change** in current (electrons' drift velocity) both propagate across distance at rates **much higher** than the actual drift velocity. You can hear wind much earlier than the force of the gust reaches you, but you don't observe a change in an EM field earlier than you can observe the change of current.

Electromagnetism



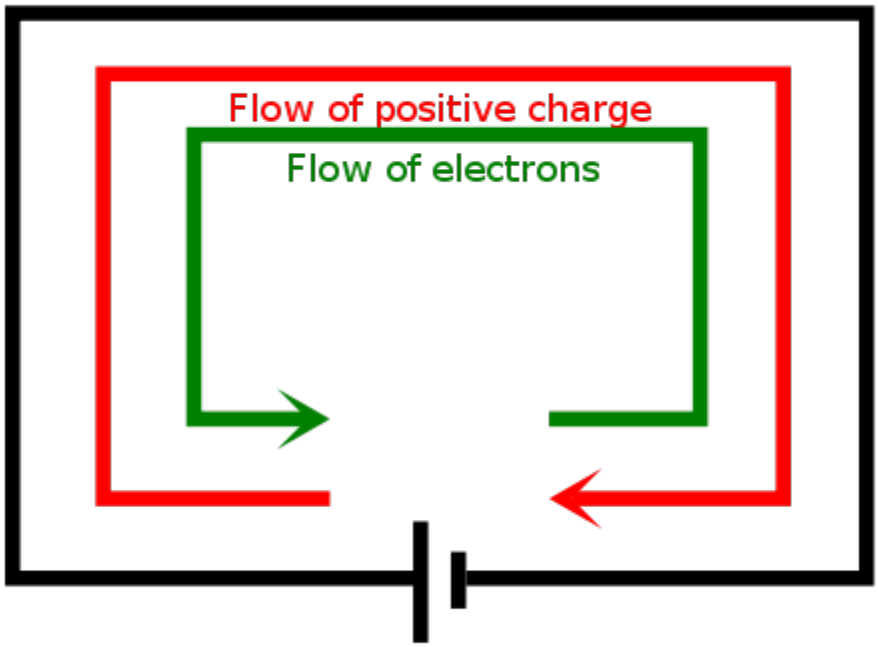
According to Ampère's law, an electric current produces a magnetic field.

Electric current produces a magnetic field. The magnetic field can be visualized as a pattern of circular field lines surrounding the wire.

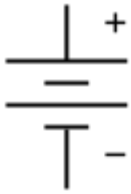
Electric current can be directly measured with a galvanometer, but this method involves breaking the electrical circuit, which is sometimes inconvenient. Current can also be measured without breaking the circuit by detecting the magnetic field associated with the current. Devices used for this include Hall effect sensors, current clamps, current transformers, and Rogowski coils.

The theory of Special Relativity allows one to transform the magnetic field into a static electric field for an observer moving at the same speed as the charge in the diagram. The amount of current is particular to a reference frame.

Conventions



The electrons, the charge carriers in an electrical circuit, flow in the opposite direction of the *conventional* electric current.



The symbol for a battery in a circuit diagram.

A flow of positive charges gives the same *electric* current as a flow of negative charges in the opposite direction. Since current can be the flow of either positive or negative charges, or both, a convention for the direction of current which is independent of the type of charge carriers is needed. Therefore the direction of *conventional current* is defined to be the direction of the flow of positive charges.

In metals, which make up the wires and other conductors in most electrical circuits, the positive charges are immobile, and only the negatively charged electrons flow. Because

the electron carries negative charge, the *electron* motion in a metal conductor is in the direction opposite to that of conventional (or *electric*) current.

Reference direction

When analyzing electrical circuits, the actual direction of current through a specific circuit element is usually unknown. Consequently, each circuit element is assigned a current variable with an arbitrarily chosen *reference direction*. When the circuit is solved, the circuit element currents may have positive or negative values. A negative value means that the actual direction of current through that circuit element is opposite that of the chosen reference direction. In electronic circuits the reference current directions are usually chosen so that all currents flow toward ground. This often matches conventional current direction, because in many circuits the power supply voltage is positive with respect to ground.

Occurrences

Natural examples include lightning and the solar wind, the source of the polar auroras (the aurora borealis and aurora australis). The artificial form of electric current is the flow of conduction electrons in metal wires, such as the overhead power lines that deliver electrical energy across long distances and the smaller wires within electrical and electronic equipment. In electronics, other forms of electric current include the flow of electrons through resistors or through the vacuum in a vacuum tube, the flow of ions inside a battery or a neuron, and the flow of holes within a semiconductor.

Current Measurement

Current can be measured using an ammeter.

At the circuit level there are various techniques that can be used to measure current:

- Shunt resistor
- Hall effect current sensor transducers
- Transformer(however dc cannot be measured)
- Magnetoresistive Field Sensors

Chapter 2

Passivity (Engineering) and Voltage Source

Passivity (engineering)

Passivity is a property of engineering systems, used in a variety of engineering disciplines, but most commonly found in analog electronics and control systems. A **passive component**, depending on field, may be either a component that consumes (but does not produce) energy (thermodynamic passivity), or a component that is incapable of power gain (incremental passivity).

A component that is not passive is called an **active component**. An electronic circuit consisting entirely of passive components is called a passive circuit (and has the same properties as a passive component). Used without qualifier, the term **passive** is ambiguous. Typically, analog designers use this term to refer to **incrementally passive** components and systems, while control systems engineers will use this to refer to **thermodynamically passive** ones.

Thermodynamic passivity

In control systems and circuit network theory, a passive component or circuit is one that consumes energy, but does not produce energy. Under this methodology, voltage and current sources are considered active, while resistors, transistors, tunnel diodes, glow tubes, capacitors, metamaterials and other dissipative and energy-neutral components are considered passive. Circuit designers will sometimes refer to this class of components as dissipative, or thermodynamically passive.

While many books give definitions for passivity, many of these contain subtle errors in how initial conditions are treated (and, occasionally, the definitions do not generalize to all types of nonlinear time-varying systems with memory). Below is a correct, formal definition, taken from Wyatt et al. (which also explains the problems with many other definitions). Given an n-port R with a state representation S, and initial state x, define available energy E_A as:

$$E_A(x) = \sup_{x \to T \geq 0} \int_0^T -\langle v(t), i(t) \rangle \, \mathrm{d}t$$

where the notation $\sup_{x \to T \geq 0}$ indicates that the supremum is taken over all $T \geq 0$ and all admissible pairs $\{v(\cdot), i(\cdot)\}$ with the fixed initial state x (e.g., all voltage–current trajectories for a given initial condition of the system). A system is considered passive if E_A is finite for all initial states x. Otherwise, the system is considered active. Roughly speaking, the inner product $\langle v(t), i(t) \rangle$ is the instantaneous power (e.g., the product of voltage and current), and E_A is the upper bound on the integral of the instantaneous power (i.e., energy). This upper bound (taken over all $T \geq 0$) is the *available energy* in the system for the particular initial condition x. If, for all possible initial states of the system, the energy available is finite, then the system is called *passive.*

Incremental passivity

In circuit design, informally, passive components refer to ones that are not capable of power gain. Under this definition, passive components include capacitors, inductors, resistors, diodes, transformers, voltage sources, and current sources. They exclude devices like transistors, vacuum tubes, relays, tunnel diodes, and glow tubes. Formally, for a memoryless two-terminal element, this means that the current–voltage characteristic is monotonically increasing. For this reason, control systems and circuit network theorists refer to these devices as locally passive, incrementally passive, increasing, monotone increasing, or monotonic. It is not clear how this definition would be formalized to multiport devices with memory – as a practical matter, circuit designers use this term informally, so it may not be necessary to formalize it.

Systems for which the small signal model is not passive are sometimes called locally active (e.g. transistors and tunnel diodes). Systems that can generate power about a time-variant unperturbed state are often called parametrically active (e.g. certain types of nonlinear capacitors) .

Other definitions of passivity

In some very informal settings, passivity may refer to the simplicity of the device, although this definition is now almost universally considered incorrect. Here, devices like diodes would be considered active, and only very simple devices like capacitors, inductors, and resistors are considered passive. In some cases, the term "linear element" may be a more appropriate term than "passive device." In other cases, "solid state device" may be a more appropriate term than "active device."

Stability

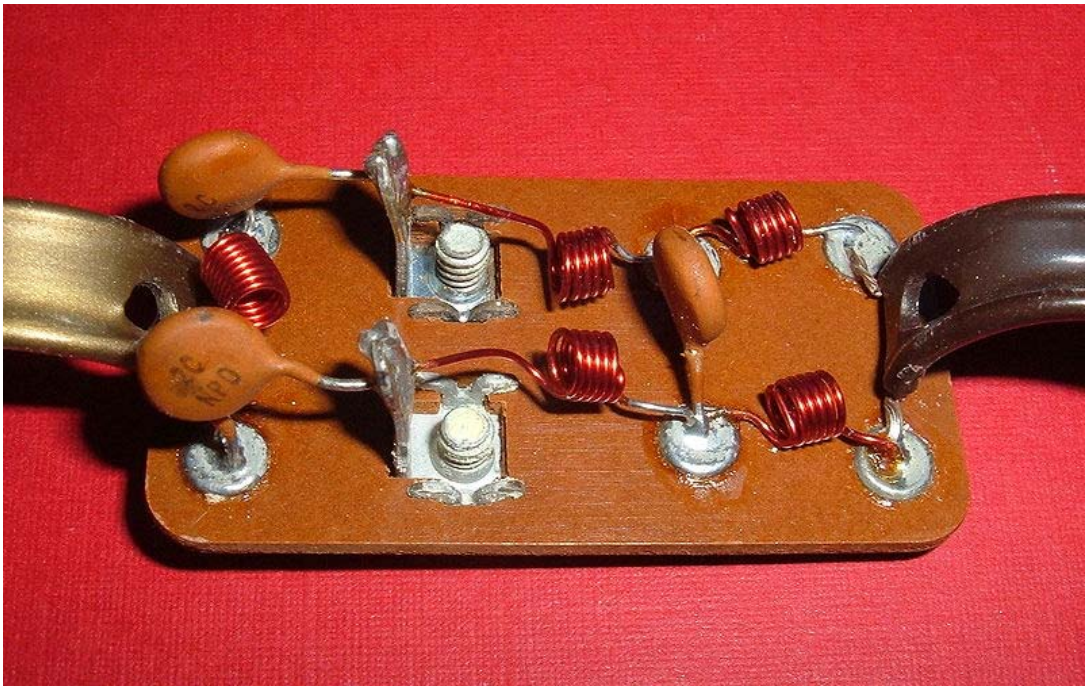
Passivity, in most cases, can be used to demonstrate that passive circuits will be stable under specific criteria. Note that this only works if only one of the above definitions of passivity is used – if components from the two are mixed, the systems may be unstable

under any criteria. In addition, passive circuits will not necessarily be stable under all stability criteria. For instance, a resonant series LC circuit will have unbounded voltage output for a bounded voltage input, but will be stable in the sense of Lyapunov, and given bounded energy input will have bounded energy output.

Passivity is frequently used in control systems to design stable control systems or to show stability in control systems. This is especially important in the design of large, complex control systems (e.g. stability of airplanes). Passivity is also used in some areas of circuit design, especially filter design.

Passive filter

A passive filter is a kind of electronic filter that is made only from passive elements – in contrast to an active filter, it does not require an external power source (beyond the signal). Since most filters are linear, in most cases, passive filters are composed of just the four basic linear elements – resistors, capacitors, inductors, and transformers. More complex passive filters may involve nonlinear elements, or more complex linear elements, such as transmission lines.



Television signal splitter consisting of a passive high-pass filter (left) and a passive low-pass filter (right). The antenna is connected to the screw terminals to the left of center.

A passive filter has several advantages over an active filter:

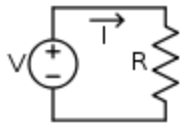
- Guaranteed stability
- Passive filters scale better to large signals (tens of amperes, hundreds of volts), where active devices are often impractical

- No power consumption.
- May be less expensive in discrete designs (unless large coils are required)
- For linear filters, may be, more linear than filters including active (and therefore non-linear) elements, depending on components required.

They are commonly used in speaker crossover design (due to the moderately large voltages and currents, and the lack of easy access to power), filters in power distribution networks (due to the large voltages and currents), power supply bypassing (due to low cost, and in some cases, power requirements), as well as a variety of discrete and home brew circuits (for low-cost and simplicity). Passive filters are uncommon in monolithic integrated circuit design, where active devices are inexpensive compared to resistors and capacitors, and inductors are prohibitively expensive. Passive filters are still found, however, in hybrid integrated circuits. Indeed, it may be the desire to incorporate a passive filter that leads the designer to use the hybrid format.

Voltage source

In electric circuit theory, an **ideal voltage source** is a circuit element where the voltage across it is independent of the current through it. A voltage source is the dual of a current source. In analysis, a voltage source supplies a constant DC or AC potential between its terminals for any current flow through it. Real-world sources of electrical energy, such as batteries, generators, or power systems, can be modeled for analysis purposes as a combination of an ideal voltage source and additional combinations of impedance elements.



A schematic diagram of an ideal voltage source, *V*, driving a resistor, *R*, and creating a current *I*

Ideal voltage sources

An **ideal voltage source** is a mathematical abstraction that simplifies the analysis of electric circuits. If the voltage across an ideal voltage source can be specified independently of any other variable in a circuit, it is called an **independent** voltage source. Conversely, if the voltage across an ideal voltage source is determined by some other voltage or current in a circuit, it is called a **dependent** or **controlled voltage source**. A mathematical model of an amplifier will include dependent voltage sources whose magnitude is governed by some fixed relation to an input signal, for example. In the analysis of faults on electrical power systems, the whole network of interconnected

sources and transmission lines can be usefully replaced by an ideal (AC) voltage source and a single equivalent impedance.

Voltage Source

Current Source

Controlled Voltage Source

Controlled Current Source

Battery of cells

Single cell

Symbols used for voltage sources

The internal resistance of an ideal voltage source is zero; it is able to supply or absorb any amount of current. The current through an ideal voltage source is completely determined by the external circuit. When connected to an open circuit, there is zero current and thus zero power. When connected to a load resistance, the current through the source approaches infinity as the load resistance approaches zero (a short circuit). Thus, an ideal voltage source can supply unlimited power.

No real voltage source is ideal; all have a non-zero effective internal resistance, and none can supply unlimited current. However, the internal resistance of a real voltage source is effectively modeled in linear circuit analysis by combining a non-zero resistance in series with an ideal voltage source (a Thévenin equivalent circuit).

Comparison between voltage and current sources

Most sources of electrical energy (the mains, a battery) are modeled as voltage sources. An *ideal* voltage source provides no energy when it is loaded by an open circuit (i.e. an infinite impedance), but approaches infinite energy and current when the load resistance approaches zero (a short circuit). Such a theoretical device would have a zero ohm output impedance in series with the source. A real-world voltage source has a very low, but non-zero output impedance: often much less than 1 ohm.

Conversely, a current source provides a constant current, as long as the load connected to the source terminals has sufficiently low impedance. An ideal current source would provide no energy to a short circuit and approach infinite energy and voltage as the load resistance approaches infinity (an open circuit). An *ideal* current source has an infinite output impedance in parallel with the source. A *real-world* current source has a very high, but finite output impedance. In the case of transistor current sources, impedance of a few megohms (at low frequencies) is typical.

Since no ideal sources of either variety exist (all real-world examples have finite and non-zero source impedance), any current source can be considered as a voltage source with the *same* source impedance and vice versa. Voltage sources and current sources are sometimes said to be duals of each other and any non ideal source can be converted from one to the other by applying Norton's or Thevenin's theorems.

Chapter 3

Current Source

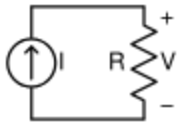


Figure 1: An ideal current source, *I*, driving a resistor, *R*, and creating a voltage *V*

A **current source** is an electrical or electronic device that delivers or absorbs electric current. A current source is the dual of a voltage source. The term constant-current **sink** is sometimes used for sources fed from a negative voltage supply. Figure 1 shows a schematic for an ideal current source driving a resistor load.

Ideal current sources

Voltage source

Current Source

Controlled Voltage Source

Controlled Current Source

Battery of cells

Single cell

Figure 2: Source symbols

In circuit theory, an **ideal current source** is a circuit element where the current through it is independent of the voltage across it. It is a mathematical model, which real devices can only approach in performance. If the current through an ideal current source can be specified independently of any other variable in a circuit, it is called an *independent* current source. Conversely, if the current through an ideal current source is determined by some other voltage or current in a circuit, it is called a **dependent** or **controlled current source**. Symbols for these sources are shown in Figure 2.

An independent current source with zero current is identical to an ideal open circuit. For this reason, the internal resistance of an ideal current source is infinite. The voltage across an ideal current source is completely determined by the circuit it is connected to. When connected to a short circuit, there is zero voltage and thus zero power delivered. When connected to a load resistance, the voltage across the source approaches infinity as the load resistance approaches infinity (an open circuit). Thus, an ideal current source, if such a thing existed in reality, could supply unlimited power and so would represent an unlimited source of energy.

No real current source is ideal (no unlimited energy sources exist) and all have a finite internal resistance (none can supply unlimited voltage). However, the internal resistance of a physical current source is effectively modeled in circuit analysis by combining a non-zero resistance in parallel with an ideal current source (the Norton equivalent circuit). The connection of an ideal open circuit to an ideal non-zero current source does not represent any physically realizable system.

Physical current sources

Resistor current source

The simplest non-ideal current source consists of a voltage source in series with a resistor. The current available from such a source is given by the ratio of the voltage across the voltage source to the resistance of the resistor. This value of current will only be delivered to a load with zero voltage drop across its terminals (a short circuit, an uncharged capacitor, a charged inductor, a virtual ground circuit, etc.) The current delivered to a load with nonzero voltage (drop) across its terminals (a linear or nonlinear resistor with a finite resistance, a charged capacitor, an uncharged inductor, a voltage source, etc.) will always be different. It is given by the ratio of the voltage drop across the resistor (the difference between the exciting voltage and the voltage across the load) to its resistance. For a nearly ideal current source, the value of the resistor should be very large but this implies that, for a specified current, the voltage source must be very large (in the limit as the resistance and the voltage go to infinity, the current source will become ideal and the current will not depend at all on the voltage across the load). Thus, efficiency is low (due to power loss in the resistor) and it is usually impractical to construct a 'good' current source this way. Nonetheless, it is often the case that such a circuit will provide adequate performance when the specified current and load resistance are small. For

example, a 5 V voltage source in series with a 4.7 kilohm resistor will provide an *approximately* constant current of 1 mA (±5%) to a load resistance in the range of 50 to 450 ohm.

A Van de Graaff generator is an example of such a high voltage current source. It behaves as an almost constant current source because of its very high output voltage coupled with its very high output resistance and so it supplies the same few microamperes at any output voltage up to hundreds of thousands of volts (or even tens of megavolts) for large laboratory versions.

Active current sources without negative feedback

Active current sources have many important applications in electronic circuits. Current sources (current-stable resistors) are often used in place of ohmic resistors in analog integrated circuits to generate a current that depends slightly on the voltage across the load.

Transistor current sources with constant input voltage

The common collector, common drain and common cathode configurations driven by a constant input voltage naturally behave as current sources (or sinks) because the output impedance of these devices is naturally high. The simple current mirror is an example of such a current source widely used in integrated circuits.

FET current sources with zero input voltage

A JFET can be made to act as a current source by tying its gate to its source. The current then flowing is the I_{DSS} of the FET. These can be purchased with this connection already made and in this case the devices are called current regulator diodes or constant current diodes or current limiting diodes (CLD). An enhancement mode N channel MOSFET can be used in the circuits listed below.

Current sources with series negative feedback

Simple transistor current source

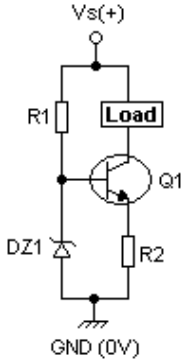


Figure 3: Typical constant current source (CCS)

Figure 3 shows a typical constant current source (CCS). DZ1 is a zener diode which, when reverse biased (as shown in the circuit) has a constant voltage drop across it irrespective of the current flowing through it. Thus, as long as the zener current (I_Z) is above a certain level (called holding current), the voltage across the zener diode (V_Z) will be constant. Resistor R1 supplies the zener current and the base current (I_B) of NPN transistor (Q1). The constant zener voltage is applied across the base of Q1 and emitter resistor R2. The operation of the circuit is as follows:

Voltage across R2 (V_{R2}) is given by V_Z - V_{BE}, where V_{BE} is the base-emitter drop of Q1. The emitter current of Q1 which is also the current through R2 is given by

$$I_{R2}(= I_E) = \frac{V_{R2}}{R2} = \frac{V_Z - V_{BE}}{R2}$$

Since V_Z is constant and V_{BE} is also (approximately) constant for a given temperature, it follows that V_{R2} is constant and hence I_E is also constant. Due to transistor action, emitter current I_E is very nearly equal to the collector current I_C of the transistor (which in turn, is the current through the load). Thus, the load current is constant (neglecting the output resistance of the transistor due to the Early effect) and the circuit operates as a constant current source. As long as the temperature remains constant (or doesn't vary much), the load current will be independent of the supply voltage, R1 and the transistor's gain. R2 allows the load current to be set at any desirable value and is calculated by

$$R2 = \frac{V_Z - V_{BE}}{I_{R2}}$$

or

$$R2 = \frac{V_Z - 0.65}{I_{R2}},$$

since V_{BE} is typically 0.65 V for a silicon device.

(I_{R2} is also the emitter current and is assumed to be the same as the collector or required load current, provided h_{FE} is sufficiently large). Resistance R_1 at resistor R1 is calculated as

$$R_1 = \frac{V_S - V_Z}{I_Z + K \cdot I_B}$$

where K = 1.2 to 2 (so that R_1 is low enough to ensure adequate I_B),

$$I_B = \frac{I_C(= I_E = I_{R2})}{h_{FE(min)}}$$

and $h_{FE(min)}$ is the lowest acceptable current gain for the particular transistor type being used.

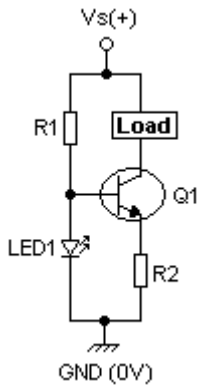


Figure 5: Typical constant current source (CCS) using LED instead of zener

The Zener diode can be replaced by any other diode, e.g. a light-emitting diode LED1 as shown in Figure 5. The LED voltage drop (V_D) is now used to derive the constant voltage and also has the additional advantage of tracking (compensating) V_{BE} changes due to temperature. R_2 is calculated as

$$R_2 = \frac{V_D - V_{BE}}{I_{R2}}$$

and R_1 as

$$R_1 = \frac{V_S - V_D}{I_D + K \cdot I_B}$$, where I_D is the LED current.

Simple transistor current source with diode compensation

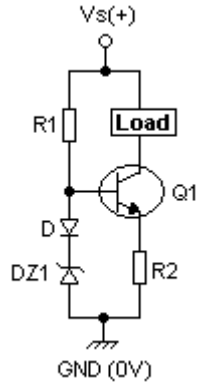


Figure 4: Typical constant current source (CCS) with diode compensation

Temperature changes will change the output current delivered by the circuit of Figure 3 because V_{BE} is sensitive to temperature. Temperature dependence can be compensated using the circuit of Figure 4 that includes a standard diode D (of the same semiconductor material as the transistor) in series with the Zener diode as shown in the image on the left. The diode drop (V_D) tracks the V_{BE} changes due to temperature and thus significantly counteracts temperature dependence of the CCS.

Resistance R_2 is now calculated as

$$R_2 = \frac{V_Z + V_D - V_{BE}}{I_{R2}}$$

Since $V_D = V_{BE} = 0.65$ V,

$$R_2 = \frac{V_Z}{I_{R2}}$$

(In practice V_D is never exactly equal to V_{BE} and hence it only suppresses the change in V_{BE} rather than nulling it out.)

R_1 is calculated as

$$R_1 = \frac{V_S - V_Z - V_D}{I_Z + K \cdot I_B}$$ (the compensating diode's forward voltage drop V_D appears in the equation and is typically 0.65 V for silicon devices.)

This method is most effective for Zener diodes rated at 5.6 V or more. For breakdown diodes of less than 5.6 V, the compensating diode is usually not required because the breakdown mechanism is not as temperature dependent as it is in breakdown diodes above this voltage.

Current mirror with emitter degeneration

Series negative feedback is also used in the two-transistor current mirror with emitter degeneration. Negative feedback is a basic feature in some current mirrors using multiple transistors, such as the Widlar current source and the Wilson current source.

Op-amp current sources...

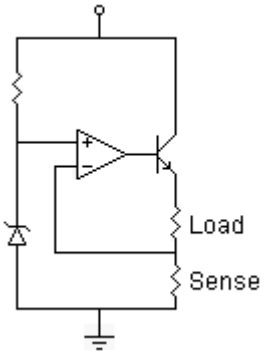


Figure 6: Typical op-amp current source. The transistor is not needed if the required current doesn't exceed the sourcing ability of the op-amp. The current will be the zener voltage divided by the sense resistor.

...with series negative feedback...

Another common method is to use a negative feedback to set the current and remove the dependence on the Vbe of the transistor. Figure 6 shows a very common approach using an op amp with the non-inverting input connected to a voltage source (such as the Zener in an above example) and the inverting input connected to the same node as the resistor and emitter of the transistor. The circuit is actually a non-inverting amplifier driven by a constant input voltage. It keeps up this constant voltage across the constant sense resistor; as a result, the current flowing through the load is constant as well.

...with parallel negative feedback

In the case of op-amp circuits (e.g., an op-amp voltage-to-current converter) sometimes it is desired to inject a precisely known current through a resistor into the inverting input (as an offset of signal input for instance). The combination of the input voltage source and the resistor will approximate an ideal current source with value V/R. The op-amp inverting input will be at virtual ground.

Current source made by a voltage regulator

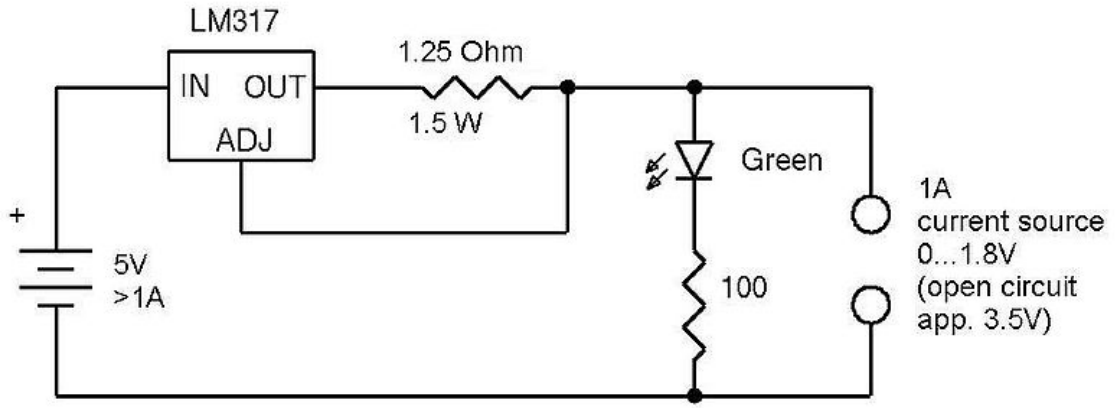


Figure 7: Constant current source using the LM317 voltage regulator

The circuit of Figure 7 using the LM317 voltage regulator is used to present a constant current source. The voltage regulator keeps up a constant voltage drop (1.25 V) across a constant resistor (1.25 Ω); so, a constant current (1 A) flows through the resistor and the load. The LED is on when the voltage across the load exceeds 1.8 V (the indicator circuit introduces some error).

Current and voltage source comparison

Most sources of electrical energy (mains electricity, a battery, ...) are best modeled as voltage sources. Such sources provide constant voltage, which means that as long as the amount of current drawn from the source is within the source's capabilities, its output voltage stays constant. An ideal voltage source provides no energy when it is loaded by an open circuit (i.e. an infinite impedance), but approaches infinite power and current when the load resistance approaches zero (a short circuit). Such a theoretical device would have a zero ohm output impedance in series with the source. A real-world voltage source has a very low, but non-zero output impedance: often much less than 1 ohm.

Conversely, a current source provides a constant current, as long as the load connected to the source terminals has sufficiently low impedance. An ideal current source would provide no energy to a short circuit and approach infinite energy and voltage as the load resistance approaches infinity (an open circuit). An *ideal* current source has an infinite output impedance in parallel with the source. A *real-world* current source has a very high, but finite output impedance. In the case of transistor current sources, impedances of a few megohms (at DC) are typical.

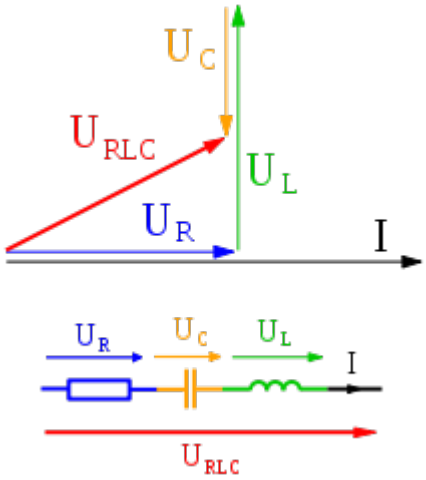
An *ideal* current source cannot be connected to an *ideal* open circuit because this would create the paradox of running a constant, non-zero current (from the current source) through an element with a defined zero current (the open circuit). Nor can an *ideal*

voltage source be connected to an *ideal* short circuit (R=0), since this would result a similar paradox of finite non zero voltage across an element with defined zero voltage (the short circuit).

Because no ideal sources of either variety exist (all real-world examples have finite and non-zero source impedance), any current source can be considered as a voltage source with the *same* source impedance and vice versa. These concepts are dealt with by Norton's and Thévenin's theorems.

Chapter 4

Phasor



An example of series RLC circuit and respective **phasor diagram**

In physics and engineering, a **phase vector**, or **phasor**, is a representation of a sine wave whose amplitude (**A**), phase (**θ**), and frequency (**ω**) are time-invariant. It is a subset of a more general concept called analytic representation. Phasors reduce the dependencies on these parameters to three independent factors, thereby simplifying certain kinds of calculations. In particular the frequency factor, which also includes the time-dependence of the sine wave, is often common to all the components of a linear combination of sine waves. Using phasors, it can be factored out, leaving just the static amplitude and phase information to be combined algebraically (rather than trigonometrically). Similarly, linear differential equations can be reduced to algebraic ones. The term *phasor* therefore often refers to just those two factors. In older texts, a **phasor** is also referred to as a **sinor**.

Definition

Euler's formula indicates that sine waves can be represented mathematically as the sum of two complex-valued functions**:**

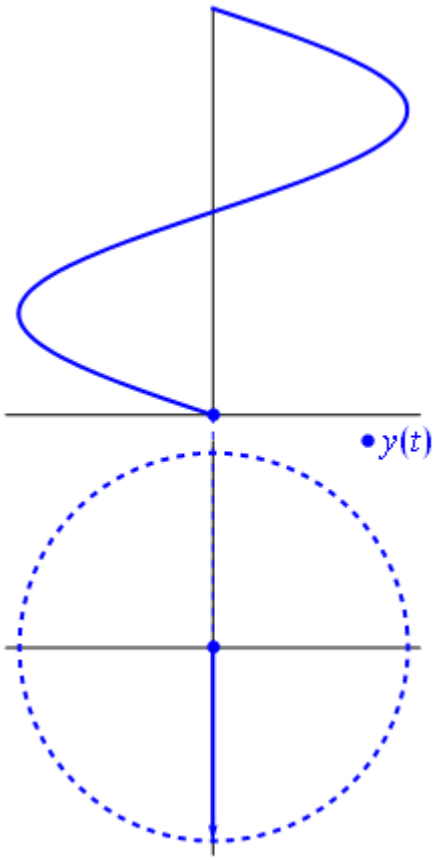
$$A \cdot \cos(\omega t + \theta) = A/2 \cdot e^{i(\omega t + \theta)} + A/2 \cdot e^{-i(\omega t + \theta)},$$

or as the real part of one of the functions:

$$\begin{aligned} A \cdot \cos(\omega t + \theta) &= \operatorname{Re}\left\{A \cdot e^{i(\omega t + \theta)}\right\} \\ &= \operatorname{Re}\left\{A e^{i\theta} \cdot e^{i\omega t}\right\}. \end{aligned}$$

As indicated above, *phasor* can refer to either $Ae^{i\theta}e^{i\omega t}$ or just the complex constant, $Ae^{i\theta}$. In the latter case, it is understood to be a shorthand notation, encoding the amplitude and phase of an underlying sinusoid.

An even more compact shorthand is angle notation: $A\angle\theta$.



A phasor can be seen as a rotating vector. The graphical representation (on paper) is at $t = 0$.

The sine wave can be understood as the projection onto the real axis of a rotating vector on the complex plane. The modulus of this vector is the amplitude of the oscillations, while its argument is the total phase $\omega t + \theta$. The phase constant θ represents the angle that the complex vector forms with the real axis at $t = 0$.

Phasor arithmetic

Multiplication by a constant (scalar)

Multiplication of the phasor $Ae^{i\theta}e^{i\omega t}$ by a complex constant, $Be^{i\phi}$ produces another phasor. That means its only effect is to change the amplitude and phase of the underlying sinusoid**:**

$$\begin{aligned}\operatorname{Re}\{(Ae^{i\theta}\cdot Be^{i\phi})\cdot e^{i\omega t}\} &= \operatorname{Re}\{(ABe^{i(\theta+\phi)})\cdot e^{i\omega t}\}\\ &= AB\cos(\omega t+(\theta+\phi))\end{aligned}$$

In electronics, $Be^{i\phi}$ would represent an impedance, which is independent of time. In particular it is *not* the shorthand notation for another phasor. Multiplying a phasor current by an impedance produces a phasor voltage. But the product of two phasors (or squaring a phasor) would represent the product of two sine waves, which is a non-linear operation that produces new frequency components. Phasor notation can only represent systems with one frequency, such as a linear system stimulated by a sinusoid.

Differentiation and integration

The time derivative or integral of a phasor produces another phasor. For example**:**

$$\begin{aligned}\operatorname{Re}\left\{\frac{d}{dt}(Ae^{i\theta}\cdot e^{i\omega t})\right\} &= \operatorname{Re}\{Ae^{i\theta}\cdot i\omega e^{i\omega t}\}\\ &= \operatorname{Re}\{Ae^{i\theta}\cdot e^{i\pi/2}\omega e^{i\omega t}\}\\ &= \operatorname{Re}\{\omega Ae^{i(\theta+\pi/2)}\cdot e^{i\omega t}\}\\ &= \omega A\cdot\cos(\omega t+\theta+\pi/2)\end{aligned}$$

Therefore, in phasor representation, the time derivative of a sinusoid becomes just multiplication by the constant, $i\omega = (e^{i\pi/2}\cdot\omega)$. Similarly, integrating a phasor corresponds to multiplication by $\frac{1}{i\omega} = \frac{e^{-i\pi/2}}{\omega}$. The time-dependent factor, $e^{i\omega t}$, is unaffected. When we solve a linear differential equation with phasor arithmetic, we are merely factoring $e^{i\omega t}$ out of all terms of the equation, and reinserting it into the answer. For example, consider the following differential equation for the voltage across the capacitor in an RC circuit**:**

$$\frac{d\ v_C(t)}{dt}+\frac{1}{RC}v_C(t)=\frac{1}{RC}v_S(t)$$

When the voltage source in this circuit is sinusoidal**:**

$$v_S(t) = V_P \cdot \cos(\omega t + \theta),$$

we may substitute:

$$v_S(t) = \mathrm{Re}\{V_s \cdot e^{i\omega t}\}$$
$$v_C(t) = \mathrm{Re}\{V_c \cdot e^{i\omega t}\},$$

where phasor $V_s = V_P e^{i\theta}$, and phasor V_c is the unknown quantity to be determined.

In the phasor shorthand notation, the differential equation reduces to:

$$i\omega V_c + \frac{1}{RC} V_c = \frac{1}{RC} V_s$$

Solving for the phasor capacitor voltage gives:

$$V_c = \frac{1}{1 + i\omega RC} \cdot (V_s) = \frac{1 - i\omega RC}{1 + (\omega RC)^2} \cdot (V_P e^{i\theta})$$

As we have seen, the factor multiplying V_s represents differences of the amplitude and phase of $v_C(t)$ relative to V_P and θ.

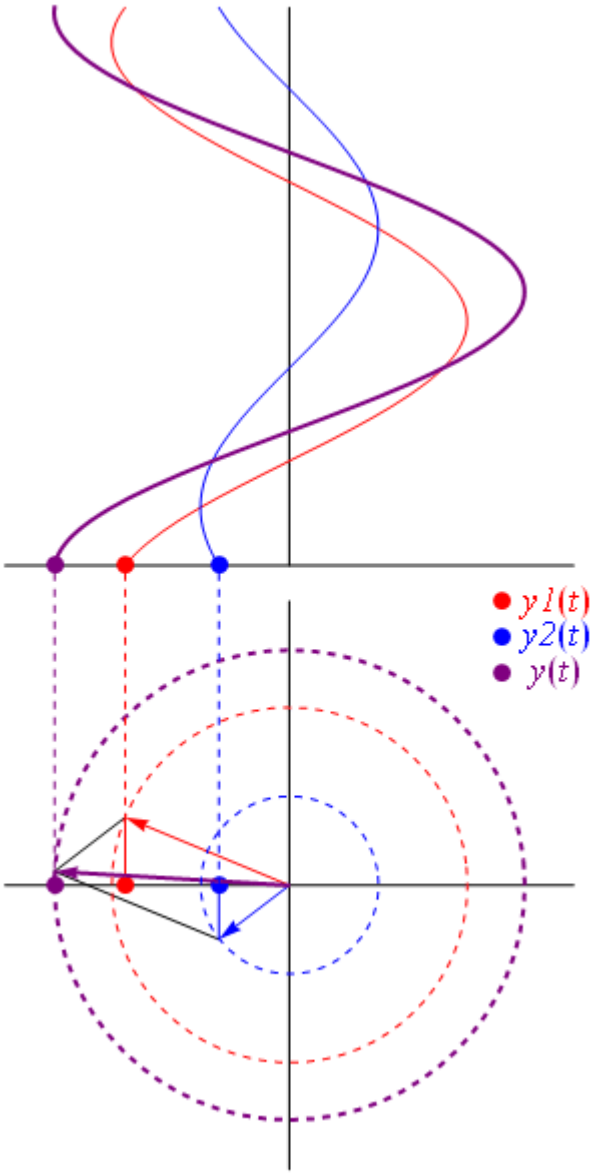
In polar coordinate form, it is:

$$\frac{1}{\sqrt{1 + (\omega RC)^2}} \cdot e^{-i\phi(\omega)}, \text{ where } \phi(\omega) = \arctan(\omega RC).$$

Therefore:

$$v_C(t) = \frac{1}{\sqrt{1 + (\omega RC)^2}} \cdot V_P \cos(\omega t + \theta - \phi(\omega))$$

Addition



The sum of phasors as addition of rotating vectors

The sum of multiple phasors produces another phasor. That is because the sum of sine waves with the same frequency is also a sine wave with that frequency:

$$\begin{aligned} A_1\cos(\omega t+\theta_1)+A_2\cos(\omega t+\theta_2) &= \operatorname{Re}\{A_1e^{i\theta_1}e^{i\omega t}\}+\operatorname{Re}\{A_2e^{i\theta_2}e^{i\omega t}\} \\ &= \operatorname{Re}\{A_1e^{i\theta_1}e^{i\omega t}+A_2e^{i\theta_2}e^{i\omega t}\} \\ &= \operatorname{Re}\{(A_1e^{i\theta_1}+A_2e^{i\theta_2})e^{i\omega t}\} \\ &= \operatorname{Re}\{(A_3e^{i\theta_3})e^{i\omega t}\} \\ &= A_3\cos(\omega t+\theta_3), \end{aligned}$$

where:

$$A_3^2 = (A_1 \cos\theta_1 + A_2 \cos\theta_2)^2 + (A_1 \sin\theta_1 + A_2 \sin\theta_2)^2,$$
$$\theta_3 = \arctan\left(\frac{A_1 \sin\theta_1 + A_2 \sin\theta_2}{A_1 \cos\theta_1 + A_2 \cos\theta_2}\right)$$

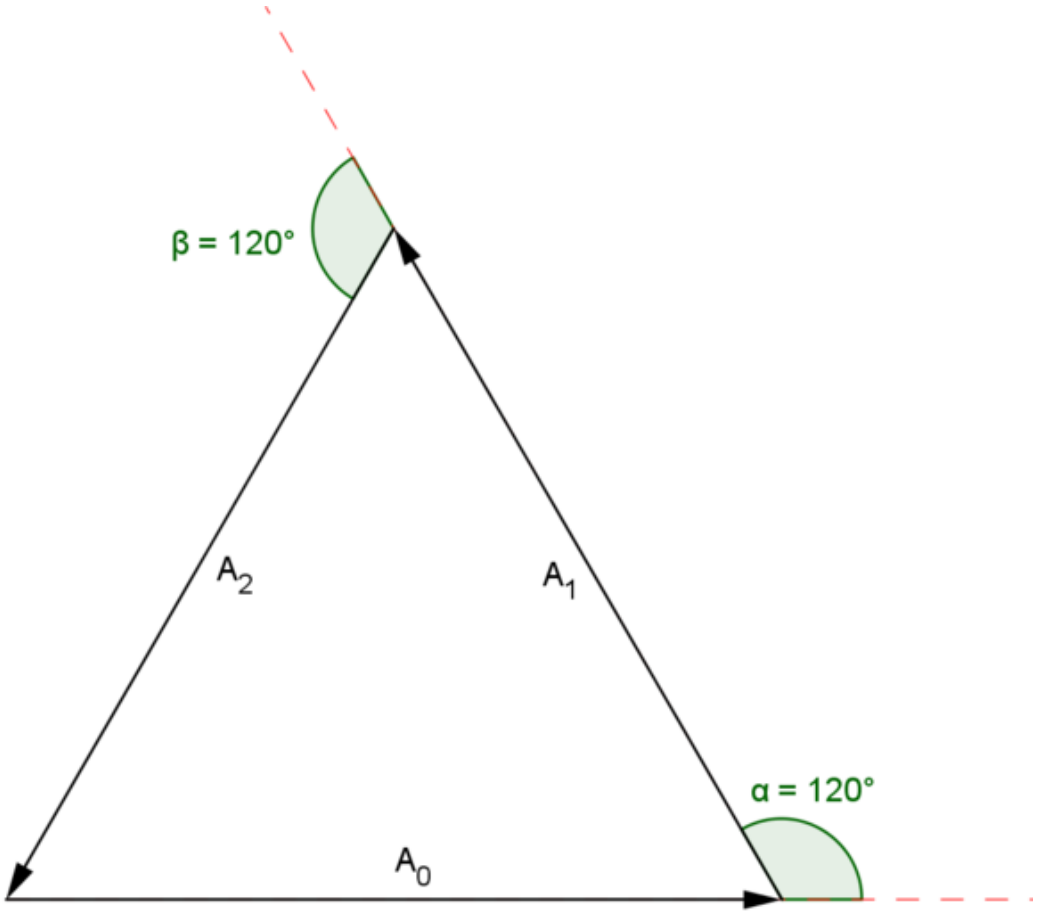
or, via the law of cosines on the complex plane (or the trigonometric identity for angle differences):

$$A_3^2 = A_1^2 + A_2^2 - 2A_1A_2\cos(180^\circ - \Delta\theta), = A_1^2 + A_2^2 + 2A_1A_2\cos(\Delta\theta),$$

where $\Delta\theta = \theta_1 - \theta_2$. A key point is that A_3 and θ_3 do not depend on ω or t, which is what makes phasor notation possible. The time and frequency dependence can be suppressed and re-inserted into the outcome as long as the only operations used in between are ones that produce another phasor. In angle notation, the operation shown above is written:

$$A_1\angle\theta_1 + A_2\angle\theta_2 = A_3\angle\theta_3.$$

Another way to view addition is that two **vectors** with coordinates $[A_1 \cos(\omega t+\theta_1), A_1 \sin(\omega t+\theta_1)]$ and $[A_2 \cos(\omega t+\theta_2), A_2 \sin(\omega t+\theta_2)]$ are added vectorially to produce a resultant vector with coordinates $[A_3 \cos(\omega t+\theta_3), A_3 \sin(\omega t+\theta_3)]$.



Phasor diagram of three waves in perfect destructive interference

In physics, this sort of addition occurs when sine waves "interfere" with each other, constructively or destructively. The static vector concept provides useful insight into questions like this: "What phase difference would be required between three identical waves for perfect cancellation?" In this case, simply imagine taking three vectors of equal length and placing them head to tail such that the last head matches up with the first tail. Clearly, the shape which satisfies these conditions is an equilateral triangle, so the angle between each phasor to the next is 120° (2π/3 radians), or one third of a wavelength $\lambda/3$. So the phase difference between each wave must also be 120°, as is the case in three-phase power

In other words, what this shows is:

$$\cos(\omega t) + \cos(\omega t + 2\pi/3) + \cos(\omega t + 4\pi/3) = 0.$$

In the example of three waves, the phase difference between the first and the last wave was 240 degrees, while for two waves destructive interference happens at 180 degrees. In the limit of many waves, the phasors must form a circle for destructive interference, so that the first phasor is nearly parallel with the last. This means that for many sources, destructive interference happens when the first and last wave differ by 360 degrees, a full wavelength λ. This is why in single slit diffraction, the minima occurs when light from the far edge travels a full wavelength further than the light from the near edge.

Phasor diagrams

Electrical engineers, electronics engineers, electronic engineering technicians and aircraft engineers all use phasor diagrams to visualize complex constants and variables (phasors). Like vectors, arrows drawn on graph paper or computer displays represent phasors. Cartesian and polar representations each have advantages.

Circuit laws

With phasors, the techniques for solving DC circuits can be applied to solve AC circuits. A list of the basic laws is given below.

- **Ohm's law for resistors:** a resistor has no time delays and therefore doesn't change the phase of a signal therefore $V=IR$ remains valid.
- **Ohm's law for resistors, inductors, and capacitors:** $V = IZ$ where Z is the complex impedance.
- In an AC circuit we have real power (P) which is a representation of the average power into the circuit and reactive power (Q) which indicates power flowing back and forward. We can also define the complex power $S = P + jQ$ and the apparent power which is the magnitude of S. The power law for an AC circuit expressed in phasors is then $S = VI^*$ (where I^* is the complex conjugate of I).
- Kirchhoff's circuit laws work with phasors in complex form

Given this we can apply the techniques of analysis of resistive circuits with phasors to analyze single frequency AC circuits containing resistors, capacitors, and inductors. Multiple frequency linear AC circuits and AC circuits with different waveforms can be analyzed to find voltages and currents by transforming all waveforms to sine wave components with magnitude and phase then analyzing each frequency separately, as allowed by the superposition theorem.

Power engineering

In analysis of three phase AC power systems, usually a set of phasors is defined as the three complex cube roots of unity, graphically represented as unit magnitudes at angles of 0, 120 and 240 degrees. By treating polyphase AC circuit quantities as phasors, balanced circuits can be simplified and unbalanced circuits can be treated as an algebraic combination of symmetrical circuits. This approach greatly simplifies the work required in electrical calculations of voltage drop, power flow, and short-circuit currents. In the

context of power systems analysis, the phase angle is often given in degrees, and the magnitude in rms value rather than the peak amplitude of the sinusoid.

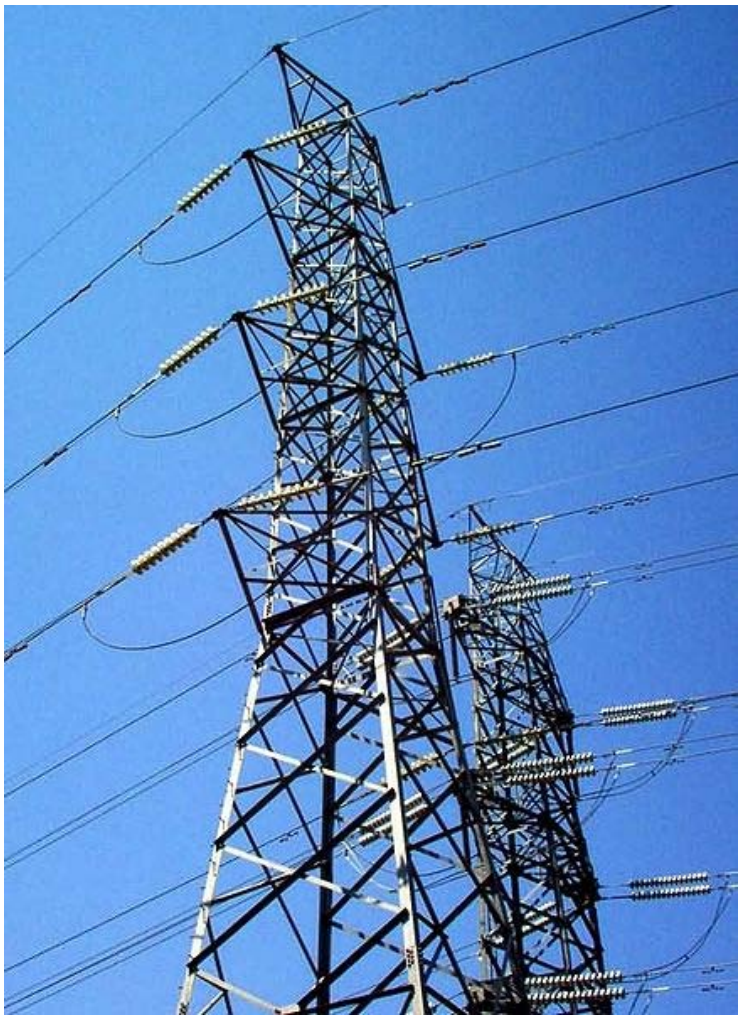
The technique of synchrophasors uses digital instruments to measure the phasors representing transmission system voltages at widespread points in a transmission network. Small changes in the phasors are sensitive indicators of power flow and system stability.

Chapter 5

Electric Power and Thévenin's Theorem

Electric power

Electric power is the rate at which electric energy is transferred by an electric circuit. The SI unit of power is the watt.



Electric energy is transmitted with **overhead lines** on pylons like these in Brisbane, Australia.

When electric current flows in a circuit, it can transfer energy to do mechanical or thermodynamic work. Devices convert electrical energy into many useful forms, such as heat (electric heaters), light (light bulbs), motion (electric motors), sound (loudspeaker), information technological processes (computers), or even chemical changes. Electricity can be produced mechanically by generation, or chemically, or by direct conversion from light in photovoltaic cells, also it can be stored chemically in batteries.

Mathematics of electric power

Circuits

Electric power, like mechanical power, is represented by the letter *P* in electrical equations. The term *wattage* is used colloquially to mean "electric power in watts."

Direct current

In direct current resistive circuits, electrical power is calculated using Joule's law:

$$P = VI$$

where *P* is the electric power, *V* the potential difference, and *I* the electric current.

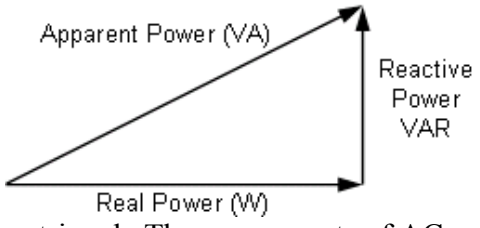
In the case of resistive (Ohmic, or linear) loads, Joule's law can be combined with Ohm's law (*I* = *V*/*R*) to produce alternative expressions for the dissipated power:

$$P = I^2R = \frac{V^2}{R},$$

where *R* is the electrical resistance.

Alternating current

In alternating current circuits, energy storage elements such as inductance and capacitance may result in periodic reversals of the direction of energy flow. The portion of power flow that, averaged over a complete cycle of the AC waveform, results in net transfer of energy in one direction is known as real power (also referred to as active power). That portion of power flow due to stored energy, that returns to the source in each cycle, is known as reactive power.



Power triangle The components of AC power

The relationship between real power, reactive power and apparent power can be expressed by representing the quantities as vectors. Real power is represented as a horizontal vector and reactive power is represented as a vertical vector. The apparent power vector is the hypotenuse of a right triangle formed by connecting the real and reactive power vectors. This representation is often called the *power triangle*. Using the Pythagorean Theorem, the relationship among real, reactive and apparent power is:

$$(\text{apparent power})^2 = (\text{real power})^2 + (\text{reactive power})^2$$

Real and reactive powers can also be calculated directly from the apparent power, when the current and voltage are both sinusoids with a known phase angle between them:

$$(\text{real power}) = (\text{apparent power})\cos(\theta)$$
$$(\text{reactive power}) = (\text{apparent power})\sin(\theta)$$

The ratio of real power to apparent power is called power factor and is a number always between 0 and 1. Where the currents and voltages have non-sinusoidal forms, power factor is generalized to include the effects of distortion.

In space

Electrical power flows wherever electric and magnetic fields exist together and fluctuate in the same place. The simplest example of this is in electrical circuits, as the preceding section showed. In the general case, however, the simple equation $P = IV$ must be replaced by a more complex calculation, the integral of the cross-product of the electrical and magnetic field vectors over a specified area, thus:

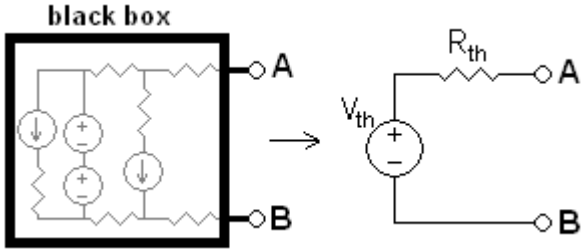
$$P = \int_S (\mathbf{E} \times \mathbf{H}) \cdot \mathbf{dA}.$$

The result is a scalar since it is the *surface integral* of the *Poynting vector*.

Thévenin's theorem

In circuit theory, **Thévenin's theorem** for linear electrical networks states that any combination of voltage sources, current sources, and resistors with two terminals is electrically equivalent to a single voltage source V and a single series resistor R. For single frequency AC systems the theorem can also be applied to general impedances, not just resistors. The theorem was first discovered by German scientist Hermann von Helmholtz in 1853, but was then rediscovered in 1883 by French telegraph engineer Léon Charles Thévenin (1857–1926).

This theorem states that a circuit of voltage sources and resistors can be converted into a **Thévenin equivalent**, which is a simplification technique used in circuit analysis. The Thévenin equivalent can be used as a good model for a power supply or battery (with the resistor representing the internal impedance and the source representing the electromotive force). The circuit consists of an ideal voltage source in series with an ideal resistor.



Any black box containing only voltage sources, current sources, and other resistors can be converted to a Thévenin equivalent circuit, comprising exactly one voltage source and one resistor.

Calculating the Thévenin equivalent

To calculate the equivalent circuit, the resistance and voltage are needed, so two equations are required. These two equations are usually obtained by using the following steps, but any conditions placed on the terminals of the circuit should also work:

1. Calculate the output voltage, V_{AB}, when in open circuit condition (no load resistor—meaning infinite resistance). This is V_{Th}.
2. Calculate the output current, I_{AB}, when the output terminals are short circuited (load resistance is 0). R_{Th} equals V_{Th} divided by this I_{AB}.

The equivalent circuit is a voltage source with voltage V_{Th} in series with a resistance R_{Th}.

Step 2 could also be thought of as:

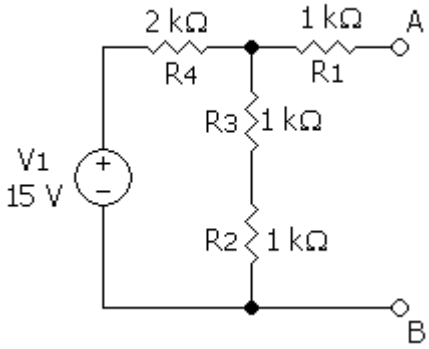
> 2a. Replace voltage sources with short circuits, and current sources with open circuits.
> 2b. Calculate the resistance between terminals A and B. This is R_{Th}.

The Thévenin-equivalent voltage is the voltage at the output terminals of the original circuit. When calculating a Thévenin-equivalent voltage, the voltage divider principle is often useful, by declaring one terminal to be V_{out} and the other terminal to be at the ground point.

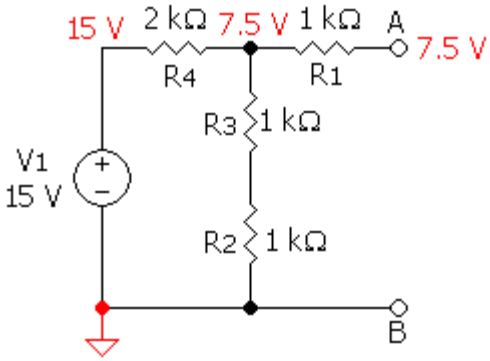
The Thévenin-equivalent resistance is the resistance measured across points A and B "looking back" into the circuit. It is important to first replace all voltage- and current-sources with their internal resistances. For an ideal voltage source, this means replace the voltage source with a short circuit. For an ideal current source, this means replace the current source with an open circuit. Resistance can then be calculated across the terminals

using the formulae for series and parallel circuits. This method is valid only for circuits with independent sources. If there are dependent sources in the circuit, another method must be used such as connecting a test source across A and B and calculating the voltage across or current through the test source.

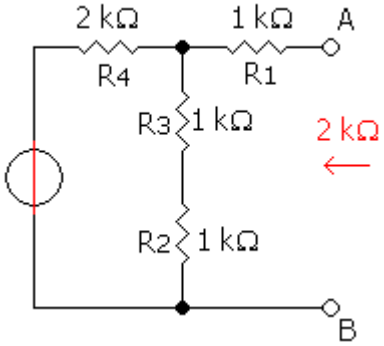
Example



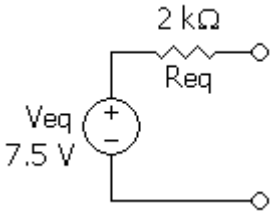
Step 0: The original circuit



Step 1: Calculating the equivalent output voltage



Step 2: Calculating the equivalent resistance



Step 3: The equivalent circuit

In the example, calculating the equivalent voltage:

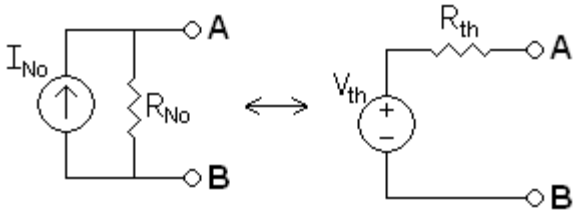
$$\begin{aligned} V_{\mathrm{Th}} &= \frac{R_2 + R_3}{(R_2 + R_3) + R_4} \cdot V_1 \\ &= \frac{1\,\mathrm{k\Omega} + 1\,\mathrm{k\Omega}}{(1\,\mathrm{k\Omega} + 1\,\mathrm{k\Omega}) + 2\,\mathrm{k\Omega}} \cdot 15\,\mathrm{V} \\ &= \frac{1}{2} \cdot 15\,\mathrm{V} = 7.5\,\mathrm{V} \end{aligned}$$

(notice that R_1 is not taken into consideration, as above calculations are done in an open circuit condition between A and B, therefore no current flows through this part, which means there is no current through R_1 and therefore no voltage drop along this part)

Calculating equivalent resistance:

$$\begin{aligned} R_{\mathrm{Th}} &= R_1 + [(R_2 + R_3) \,\|R_4)] \\ &= 1\,\mathrm{k\Omega} + [(1\,\mathrm{k\Omega} + 1\,\mathrm{k\Omega}) \,\|2\,\mathrm{k\Omega})] \\ &= 1\,\mathrm{k\Omega} + \left(\frac{1}{(1\,\mathrm{k\Omega} + 1\,\mathrm{k\Omega})} + \frac{1}{(2\,\mathrm{k\Omega})}\right)^{-1} = 2\,\mathrm{k\Omega} \end{aligned}$$

Conversion to a Norton equivalent



A Norton equivalent circuit is related to the Thévenin equivalent by the following:

$$\begin{aligned} R_{\mathrm{Th}} &= R_{\mathrm{No}} \\ V_{\mathrm{Th}} &= I_{\mathrm{No}} R_{\mathrm{No}} \\ I_{\mathrm{No}} &= V_{\mathrm{Th}} / R_{\mathrm{Th}} \end{aligned}$$

Practical limitations

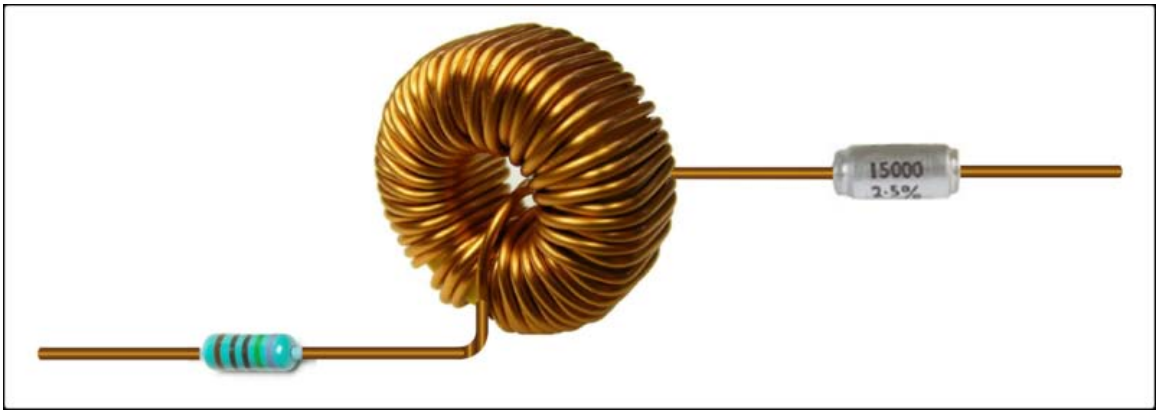
- Many, if not most circuits are only linear over a certain range of values, thus the Thévenin equivalent is valid only within this linear range and may not be valid outside the range.

- The Thévenin equivalent has an equivalent I-V characteristic only from the point of view of the load.

- The power dissipation of the Thévenin equivalent is not necessarily identical to the power dissipation of the real system. However, the power dissipated by an external resistor between the two output terminals is the same however the internal circuit is represented.

Chapter 6

RLC Circuit



A series RLC circuit: a resistor, inductor, and a capacitor

An **RLC circuit** (or **LCR circuit**) is an electrical circuit consisting of a resistor, an inductor, and a capacitor, connected in series or in parallel. The RLC part of the name is due to those letters being the usual electrical symbols for resistance, inductance and capacitance respectively. The circuit forms a harmonic oscillator for current and will resonate in just the same way as an LC circuit will. The difference that the presence of the resistor makes is that any oscillation induced in the circuit will die away over time if it not kept going by a source. This effect of the resistor is called damping. Some resistance is unavoidable in real circuits, even if a resistor is not specifically included as a component. A pure LC circuit is an ideal which really only exists in theory.

There are many applications for this circuit. They are used in many different types of oscillator circuit. Another important application is for tuning, such as in radio receivers or television sets, where they are used to select a narrow range of frequencies from the ambient radio waves. In this role the circuit is often referred to as a tuned circuit. An RLC circuit can be used as a band-pass filter or a band-stop filter. The tuning application, for instance, is an example of band-pass filtering. The RLC filter is described as a *second-order* circuit, meaning that any voltage or current in the circuit can be described by a second-order differential equation in circuit analysis.

The three circuit elements can be combined in a number of different topologies. All three elements in series or all three elements in parallel are the simplest in concept and the most straightforward to analyse. There are, however, other arrangements, some with practical importance in real circuits. One issue often encountered is the need to take into account inductor resistance. Inductors are typically constructed from coils of wire, the resistance of which is not usually desirable, but it often has a significant effect on the circuit.

Basic concepts

Resonance

An important property of this circuit is its ability to resonate at a specific frequency, the resonance frequency, f_0. Frequencies are measured in units of hertz. Here, however, angular frequency, ω_0, is used which is more mathematically convenient. This is measured in radians per second. They are related to each other by a simple proportion,

$$\omega_0 = 2\pi f_0$$

Resonance occurs because energy is stored in two different ways: in an electric field as the capacitor is charged and in a magnetic field as current flows through the inductor. Energy can be transferred from one to the other within the circuit and this can be oscillatory. A mechanical analogy is a weight suspended on a spring which will oscillate up and down when released. This is no passing metaphor, a weight on a spring is described by exactly the same second order differential equation as an RLC circuit and for all the properties of the one system there will be found an analogous property of the other. The mechanical property answering to the resistor in the circuit is friction in the spring/weight system. Friction will slowly bring any oscillation to a halt if there is no external force driving it. Likewise, the resistance in an RLC circuit will "damp" the oscillation, diminishing it with time if there is no driving AC power source in the circuit.

The resonance frequency is defined as the frequency at which the impedance of the circuit is at a minimum. Equivalently, it can be defined as the frequency at which the impedance is purely real (that is, purely resistive). This occurs because the impedance (reactance) of the inductor and capacitor at resonance are equal but of opposite sign and cancel out. Circuits where L and C are in parallel rather than series actually have a maximum impedance rather than a minimum impedance. For this reason they are often described as antiresonators, it is still usual, however, to name the frequency at which this occurs as the resonance frequency.

Natural frequency

The resonance frequency is defined in terms of the impedance presented to a driving source. It is still possible for the circuit to carry on oscillating (for a time) after the driving source has been removed or it is subjected to a step in voltage (including a step down to zero). This is similar to the way that a tuning fork will carry on ringing after it

has been struck, and the effect is often called ringing. This effect is the undriven natural resonance frequency of the circuit and in general is not exactly the same as the driven resonance frequency, although the two will usually be quite close to each other. Various terms are used by different authors to distinguish the two, but resonance frequency unqualified usually means the driven resonance frequency. The driven frequency may be called the undamped resonance frequency or undamped natural frequency and the undriven frequency may be called the damped resonance frequency or the damped natural frequency. The reason for this terminology is that the driven resonance frequency in a series or parallel resonant circuit has the value

$$\omega_0 = \frac{1}{\sqrt{LC}}$$

This is exactly the same as the resonance frequency of an LC circuit, that is, one with no resistor present, that is, it is the same as a circuit in which there is no damping, hence undamped resonance frequency. The undriven resonance frequency, on the other hand, depends on the value of the resistor and hence is described as the damped resonance frequency. A highly damped circuit will fail to resonate at all when undriven. A circuit with a value of resistor that causes it to be just on the edge of ringing is called critically damped. Either side of critically damped are described as underdamped (ringing happens) and overdamped (ringing is suppressed).

Circuits with topologies more complex than straightforward series or parallel (some examples described later) have a driven resonance frequency that deviates from $\omega_0 = \frac{1}{\sqrt{LC}}$ and for those the undamped resonance frequency, damped resonance frequency and driven resonance frequency can all be different.

Damping

Damping is caused by the resistance in the circuit. It determines whether or not the circuit will resonate naturally (that is, without a driving source). Circuits which will resonate in this way are described as underdamped and those that will not are overdamped. Damping attenuation (symbol α) is measured in nepers per second. However, the unitless damping factor (symbol ζ) is often a more useful measure, which is related to α by

$$\zeta = \frac{\alpha}{\omega_0}$$

The special case of $\zeta = 1$ is called critical damping and represents the case of a circuit that is just on the border of oscillation. It is the minimum damping that can be applied without causing oscillation.

Bandwidth

The resonance effect can be used for filtering, the rapid change in impedance near resonance can be used to pass or block signals close to the resonance frequency. Both band-pass and band-stop filters can be constructed and some filter circuits are shown later. A key parameter in filter design is bandwidth. The bandwidth is measured between the 3dB-points, that is, the frequencies at which the power passed through the circuit has fallen to half the value passed at resonance. There are two of these half-power frequencies, one above, and one below the resonance frequency

$$\Delta\omega = \omega_2 - \omega_1$$

where $\Delta\omega$is the bandwidth, $\omega_1$is the lower half-power frequency and $\omega_2$is the upper half-power frequency. The bandwidth is related to attenuation by,

$$\Delta\omega = 2\alpha$$

when the units are radians per second and nepers per second respectively. Other units may require a conversion factor. A more general measure of bandwidth is the fractional bandwidth, which expresses the bandwidth as a fraction of the resonance frequency and is given by

$$F_\mathrm{b} = \frac{\Delta\omega}{\omega_0}$$

The fractional bandwidth is also often stated as a percentage. The damping of filter circuits is adjusted to result in the required bandwidth. A narrow band filter, such as a notch filter, requires low damping. A wide band filter requires high damping.

Q factor

The Q factor is a widespread measure used to characterise resonators. It is defined as the peak energy stored in the circuit divided by the average energy dissipated in it per cycle at resonance. Low *Q* circuits are therefore damped and lossy and high *Q* circuits are underdamped. *Q* is related to bandwidth; low *Q* circuits are wide band and high *Q* circuits are narrow band. In fact, it happens that *Q* is the inverse of fractional bandwidth

$$Q = \frac{1}{F_\mathrm{b}} = \frac{\omega_0}{\Delta\omega}$$

Q factor is directly proportional to selectivity, as Q factor depends inversely on bandwidth.

Scaled parameters

The parameters ζ, F_b, and Q are all scaled to ω_0. This means that circuits which have similar parameters share similar characteristics regardless of whether or not they are operating in the same frequency band.

Series RLC circuit

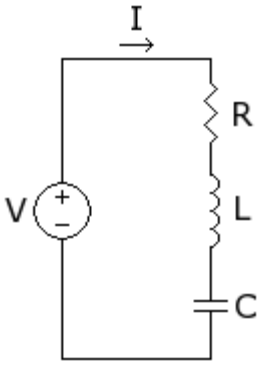


Figure 1. RLC series circuit

V - the voltage of the power source
I - the current in the circuit
R - the resistance of the resistor
L - the inductance of the inductor
C - the capacitance of the capacitor

In this circuit, the three components are all in series with the voltage source. The governing differential equation can be found by substituting into Kirchhoff's voltage law (KVL) the constitutive equation for each of the three elements. From KVL,

$$v_R + v_L + v_C = v(t)$$

where v_R, v_L, v_C are the voltages across R, L and C respectively and $v(t)$ is the time varying voltage from the source. Substituting in the constitutive equations,

$$Ri(t) + L\frac{di}{dt} + \frac{1}{C}\int_{-\infty}^{\tau=t} i(\tau)\,d\tau = v(t)$$

For the case where the source is an unchanging voltage, differentiating and dividing by L leads to the second order differential equation:

$$\frac{d^2i(t)}{dt^2} + \frac{R}{L}\frac{di(t)}{dt} + \frac{1}{LC}i(t) = 0$$

This can usefully be expressed in a more generally applicable form:

$$\frac{d^2i(t)}{dt^2} + 2\alpha\frac{di}{dt} + {\omega_0}^2 i(t) = 0$$

$\alpha$and $\omega_0$are both in units of angular frequency. $\alpha$is called the *neper frequency*, or *attenuation*, and is a measure of how fast the transient response of the circuit will die away after the stimulus has been removed. Neper occurs in the name because the units can also be considered to be nepers per second, neper being a unit of attenuation. $\omega_0$is the angular resonance frequency and is discussed later.

For the case of the series RLC circuit these two parameters are given by:

$$\alpha = \frac{R}{2L} \text{and } \omega_0 = \frac{1}{\sqrt{LC}}$$

A useful parameter is the *damping factor*, ζ which is defined as the ratio of these two,

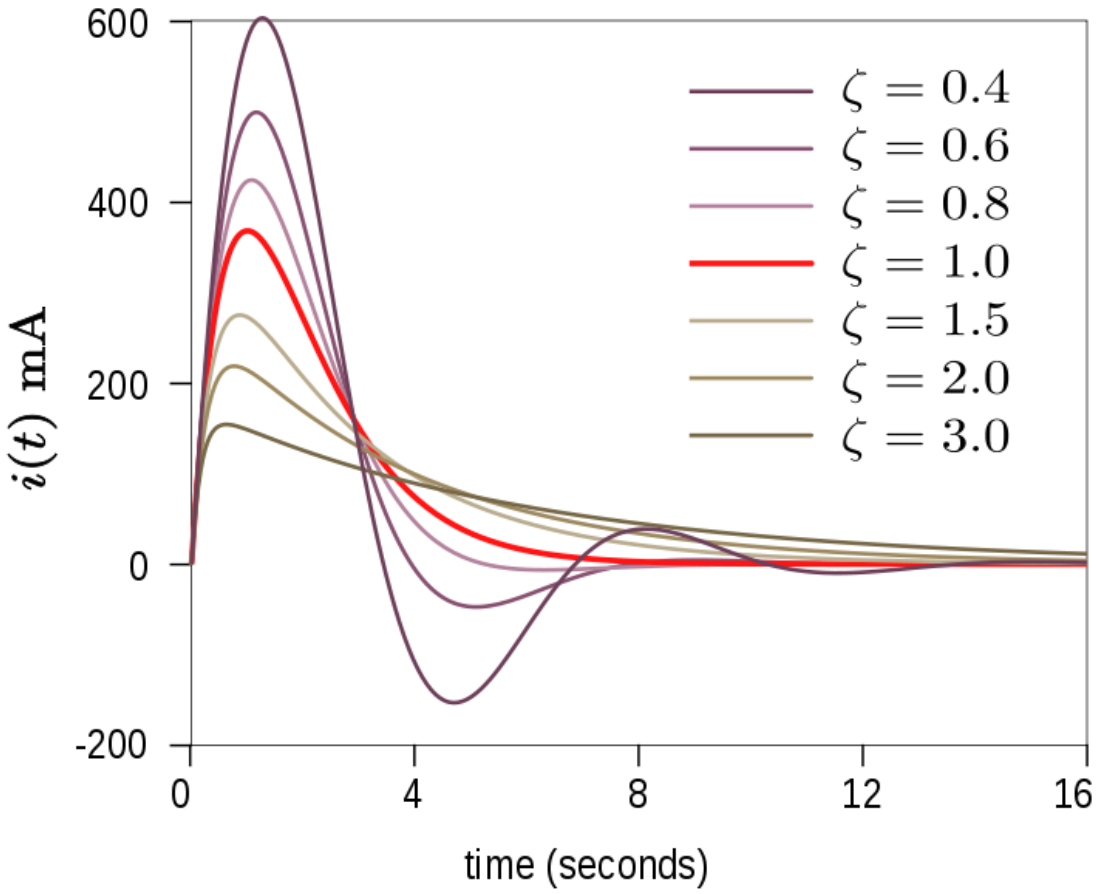
$$\zeta = \frac{\alpha}{\omega_0}$$

In the case of the series RLC circuit, the damping factor is given by,

$$\zeta = \frac{R}{2}\sqrt{\frac{C}{L}}$$

The value of the damping factor determines the type of transient that the circuit will exhibit. Some authors do not use $\zeta$and call $\alpha$the damping factor.

Transient response



Plot showing underdamped and overdamped responses of a series RLC circuit. The critical damping plot is the bold red curve. The plots are normalised for L=1, C=1 and $\omega_0=1$

The differential equation for the circuit solves in three different ways depending on the value of ζ. These are underdamped ($\zeta<1$), overdamped ($\zeta>1$) and critically damped ($\zeta=1$). The differential equation has the characteristic equation,

$$s^2 + 2\alpha s + {\omega_0}^2 = 0$$

The roots of the equation in *s* are,

$$s_1 = -\alpha + \sqrt{\alpha^2 - {\omega_0}^2}$$
$$s_2 = -\alpha - \sqrt{\alpha^2 - {\omega_0}^2}$$

The general solution of the differential equation is an exponential in either root or a linear superposition of both,

$$i(t) = A_1 e^{s_1 t} + A_2 e^{s_2 t}$$

The coefficients A_1 and A_2 are determined by the boundary conditions of the specific problem being analysed. That is, they are set by the values of the currents and voltages in the circuit at the onset of the transient and the presumed value they will settle to after infinite time.

The overdamped response ($\zeta > 1$) is,

$$i(t) = A_1 e^{-\omega_0(\zeta + \sqrt{\zeta^2 - 1})t} + A_2 e^{-\omega_0(\zeta - \sqrt{\zeta^2 - 1})t}$$

The overdamped response is a decay of the transient current without oscillation.

The underdamped response ($\zeta < 1$) is,

$$i(t) = B_1 e^{-\alpha t} \cos(\omega_d t) + B_2 e^{-\alpha t} \sin(\omega_d t)$$

By applying standard trigonometric identities the two trigonometric functions may be expressed as a single sinusoid with phase shift,

$$i(t) = B_3 e^{-\alpha t} \sin(\omega_d t + \varphi)$$

The underdamped response is a decaying oscillation at frequency ω_d. The oscillation decays at a rate determined by the attenuation α. The exponential in αdescribes the envelope of the oscillation. B_1 and B_2 (or B_3 and the phase shift φin the second form) are arbitrary constants determined by boundary conditions. The frequency $\omega_d$is given by,

$$\omega_d = \sqrt{{\omega_0}^2 - \alpha^2} = \omega_0 \sqrt{1 - \zeta^2}$$

This is called the damped resonance frequency or the damped natural frequency. It is the frequency the circuit will naturally oscillate at if not driven by an external source. The resonance frequency, ω_0, which is the frequency at which the circuit will resonate when driven by an external oscillation, may often be referred to as the undamped resonance frequency to distinguish it.

The critically damped response ($\zeta = 1$) is,

$$i(t) = D_1 t e^{-\alpha t} + D_2 e^{-\alpha t}$$

The critically damped response represents the circuit response that decays in the fastest possible time without going into oscillation. This consideration is important in control systems where it is required to reach the desired state as quickly as possible without overshooting. D_1 and D_2 are arbitrary constants determined by boundary conditions.

Laplace domain

The series RLC can be analyzed for both transient and steady AC state behavior using the Laplace transform. If the voltage source above produces a waveform with Laplace-transformed $V(s)$ (where s is the complex frequency $s = \sigma + i\omega$), KVL can be applied in the Laplace domain:

$$V(s) = I(s)\left(R + Ls + \frac{1}{Cs}\right)$$

where $I(s)$ is the Laplace-transformed current through all components. Solving for $I(s)$:

$$I(s) = \frac{1}{R + Ls + \frac{1}{Cs}}V(s)$$

And rearranging, we have that

$$I(s) = \frac{s}{L\left(s^2 + \frac{R}{L}s + \frac{1}{LC}\right)}V(s)$$

Laplace admittance

Solving for the Laplace admittance $Y(s)$:

$$Y(s) = \frac{I(s)}{V(s)} = \frac{s}{L\left(s^2 + \frac{R}{L}s + \frac{1}{LC}\right)}$$

Simplifying using parameters α and ω_o defined in the previous section, we have

$$Y(s) = \frac{I(s)}{V(s)} = \frac{s}{L\left(s^2 + 2\alpha s + {\omega_0}^2\right)}$$

Poles and zeros

The zeros of $Y(s)$ are those values of s such that $Y(s) = 0$:

$$s = 0 \quad \text{and} \quad |s| \to \infty$$

The poles of $Y(s)$ are those values of s such that $Y(s) \to \infty$. By the quadratic formula, we find

$$s = -\alpha \pm \sqrt{\alpha^2 - {\omega_0}^2}$$

The poles of $Y(s)$ are identical to the roots s_1 and s_2 of the characteristic polynomial of the differential equation in the section above.

General solution

For an arbitrary E(t), the solution obtained by inverse transform of I(s) is:

$$I(t) = \frac{1}{L}\int_0^t E(t-\tau)e^{-\alpha\tau}\left(\cos\omega_d\tau - \frac{\alpha}{\omega_d}\sin\omega_d\tau\right)d\tau$$

in the underdamped case ($\omega_0 > \alpha$)

$$I(t) = \frac{1}{L}\int_0^t E(t-\tau)e^{-\alpha\tau}(1-\alpha\tau)d\tau$$

in the critically damped case ($\omega_0 = \alpha$)

$$I(t) = \frac{1}{L}\int_0^t E(t-\tau)e^{-\alpha\tau}\left(\cosh\omega_r\tau - \frac{\alpha}{\omega_r}\sinh\omega_r\tau\right)d\tau$$

in the overdamped case ($\omega_0 < \alpha$)

where $\omega_r = \sqrt{\alpha^2 - {\omega_0}^2}$, and cosh and sinh are the usual hyperbolic functions.

Sinusoidal steady state

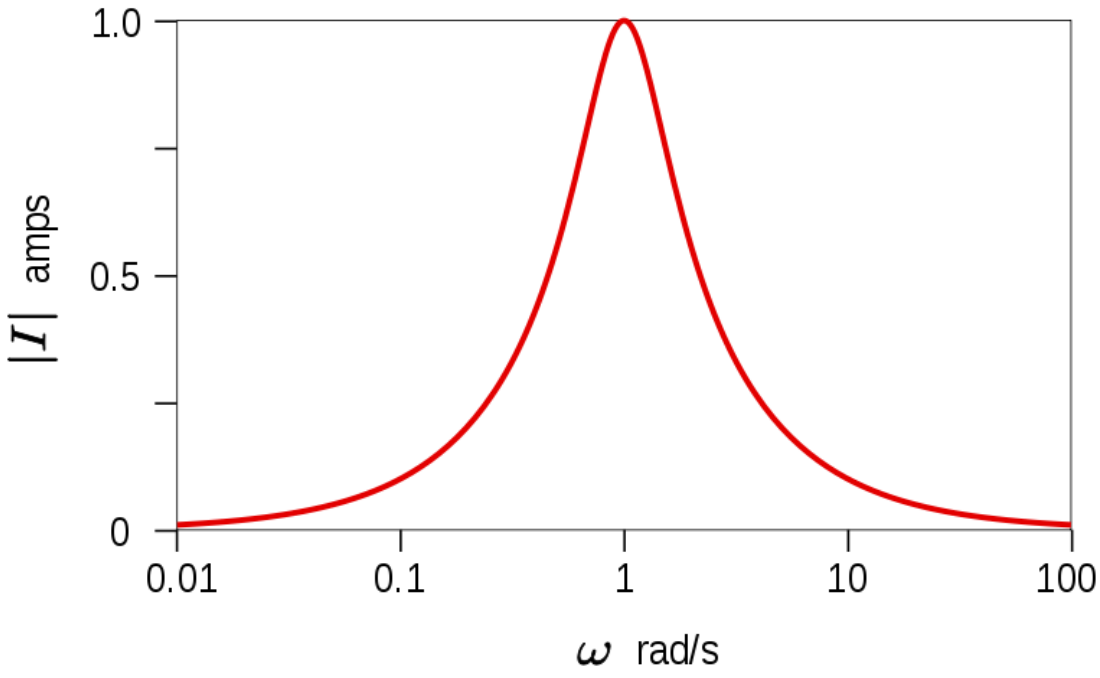


Figure 4. *Sinusoidal steady-state analysis*

normalised to R = 1 ohm, C = 1 farad, L = 1 henry, and V = 1.0 volt

Sinusoidal steady state is represented by letting $s = i\omega$

Taking the magnitude of the above equation with this substitution:

$$|Y(s = i\omega)| = \frac{1}{\sqrt{R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2}}.$$

and the current as a function of ω can be found from

$$|I(i\omega)| = |Y(i\omega)||V(i\omega)|.$$

Note that there is a peak at $i_{mag}(\omega) = 1$. This is known as the resonance frequency. Solving for this value:

$$\omega_0 = \frac{1}{\sqrt{LC}}.$$

Parallel RLC circuit

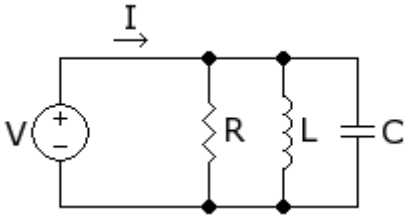


Figure 5. RLC parallel circuit

V - the voltage of the power source
I - the current in the circuit
R - the resistance of the resistor
L - the inductance of the inductor
C - the capacitance of the capacitor

The properties of the parallel RLC circuit can be obtained from the duality relationship of electrical circuits and considering that the parallel RLC is the dual impedance of a series RLC. From this consideration is immediately obtained the result that the differential equations describing this circuit will be identical to the general form of those describing a series RLC.

For the parallel circuit, the attenuation α is given by

$$\alpha = \frac{1}{2RC}$$

and the damping factor is consequently

$$\zeta = \frac{1}{2R}\sqrt{\frac{L}{C}}$$

This is the inverse of the expression for ζ in the series circuit. Likewise, the other scaled parameters, fractional bandwidth and Q are also the inverse of each other. This means that a wide band, low Q circuit in one topology will become a narrow band, high Q circuit in the other topology when constructed from components with identical values. The Q and fractional bandwidth of the parallel circuit are given by

$$Q = R\sqrt{\frac{C}{L}} \text{ and } F_{\mathrm{b}} = \frac{1}{R}\sqrt{\frac{L}{C}}$$

Frequency domain

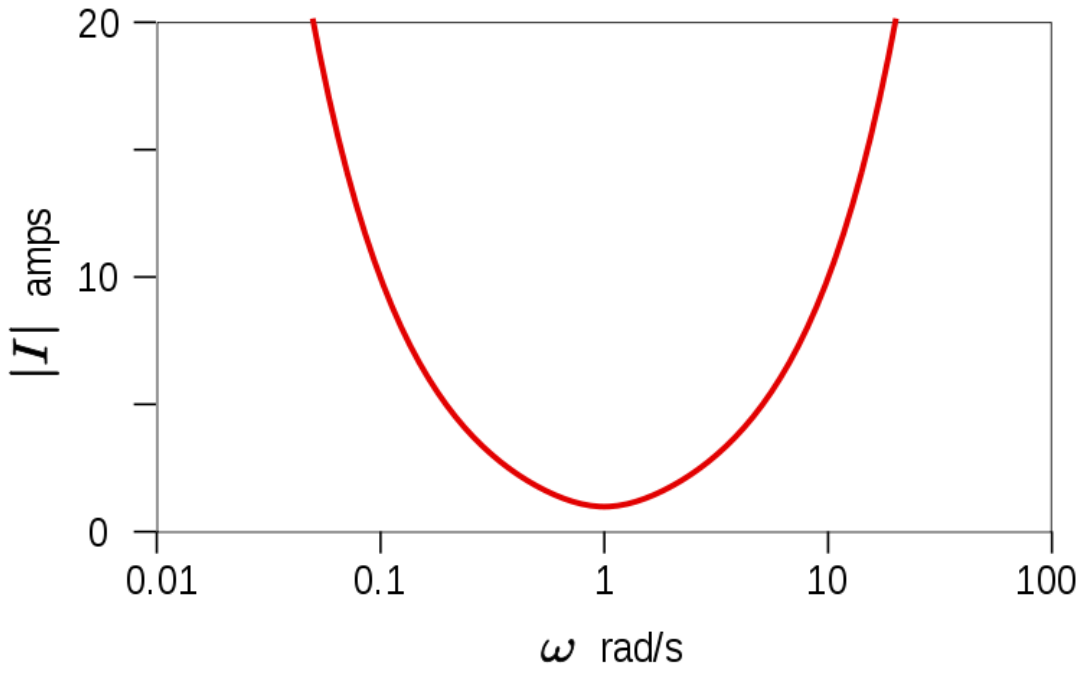


Figure 6. *Sinusoidal steady-state analysis*s

normalised to R = 1 ohm, C = 1 farad, L = 1 henry, and V = 1.0 volt

The complex admittance of this circuit is given by adding up the admittances of the components:

$$\frac{1}{Z} = \frac{1}{Z_L} + \frac{1}{Z_C} + \frac{1}{Z_R} = \frac{1}{j\omega L} + j\omega C + \frac{1}{R}$$

The change from a series arrangement to a parallel arrangement results in the circuit having a peak in impedance at resonance rather than a minimum, so the circuit is an antiresonator.

The graph opposite shows that there is a minimum in the frequency response of the current at the resonance frequency $\omega_0 = \frac{1}{\sqrt{LC}}$ when the circuit is driven by a constant voltage. On the other hand, if driven by a constant current, there would be a maximum in the voltage which would follow the same curve as the current in the series circuit.

Other configurations

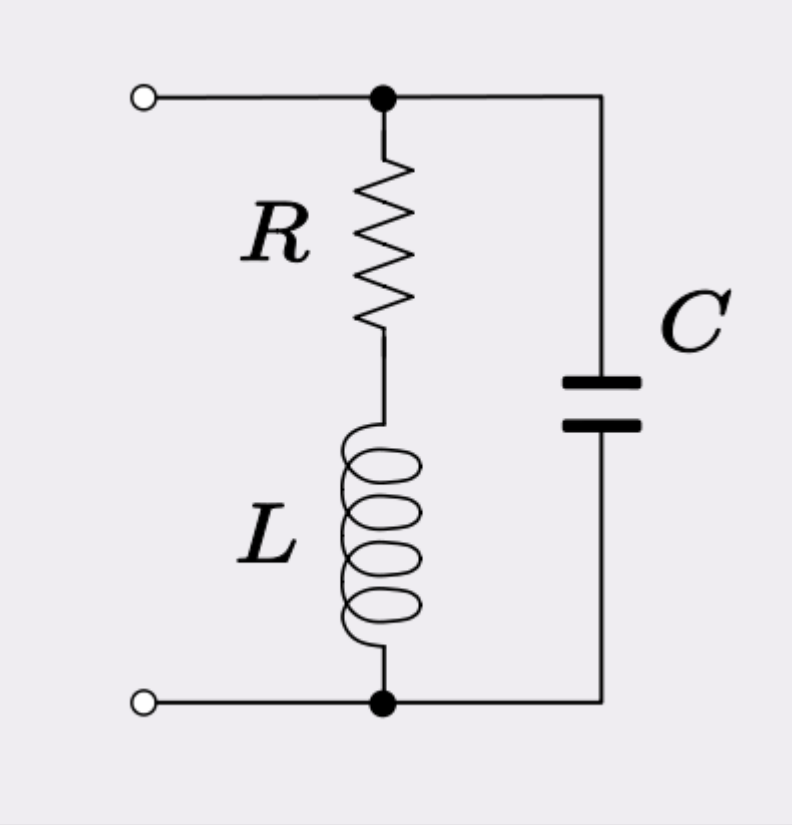


Fig. 7. RLC parallel circuit with resistance in series with the inductor

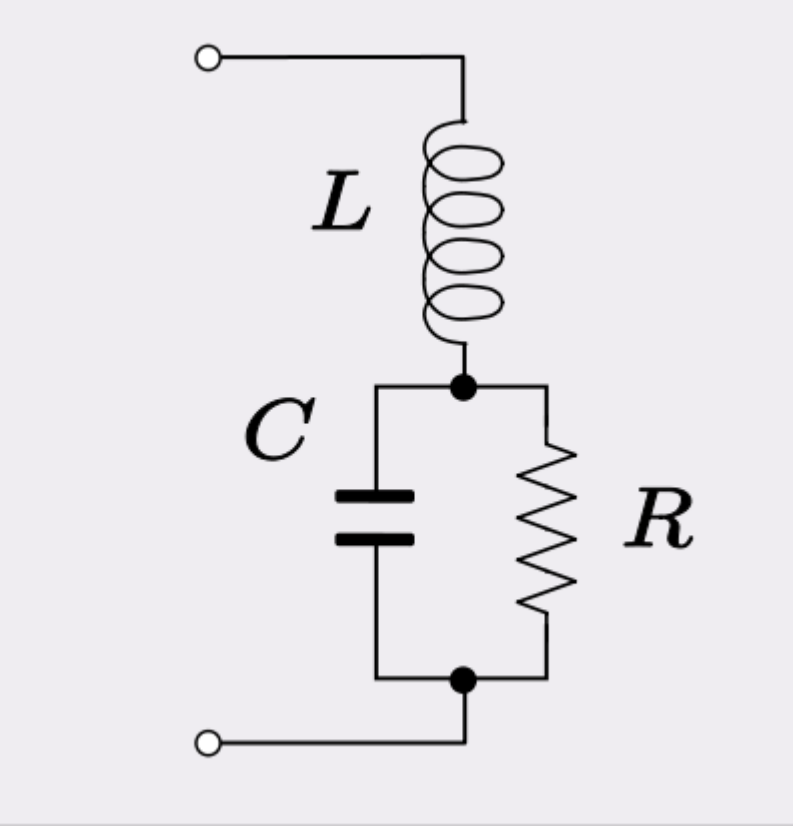


Fig. 8. RLC series circuit with resistance in parallel with the capacitor

A series resistor with the inductor in a parallel LC circuit as shown in figure 7 is a topology commonly encountered where there is a need to take into account the resistance of the coil winding. Parallel LC circuits are frequently used for bandpass filtering and the *Q* is largely governed by this resistance. The resonant frequency of this circuit is,

$$\omega_0 = \sqrt{\frac{1}{LC} - \left(\frac{R}{L}\right)^2}$$

This is the resonant frequency of the circuit defined as the frequency at which the admittance has zero imaginary part. The frequency that appears in the generalised form of the characteristic equation (which is the same for this circuit as previously)

$$s^2 + 2\alpha s + {\omega'_0}^2 = 0$$

is not the same frequency. In this case it is the undamped resonant frequency

$$\omega'_0 = \sqrt{\frac{1}{LC}}$$

In the same vein, a resistor in parallel with the capacitor in a series LC circuit can be used to represent a capacitor with a lossy dielectric. This configuration is shown in figure 8. The resonant frequency in this case is given by

$$\omega_0 = \sqrt{\frac{1}{LC} - \frac{1}{(RC)^2}}$$

Applications

Variable tuned circuits

A very frequent use of these circuits is in the tuning circuits of analogue radios. Adjustable tuning is commonly achieved with a parallel plate variable capacitor which allows the value of C to be changed and tune to stations on different frequencies. For the IF stage in the radio where the tuning is preset in the factory the more usual solution is an adjustable core in the inductor to adjust L. In this design the core (made of a high permeability material that has the effect of increasing inductance) is threaded so that it can be screwed further in, or screwed further out of the inductor winding as required.

Filters

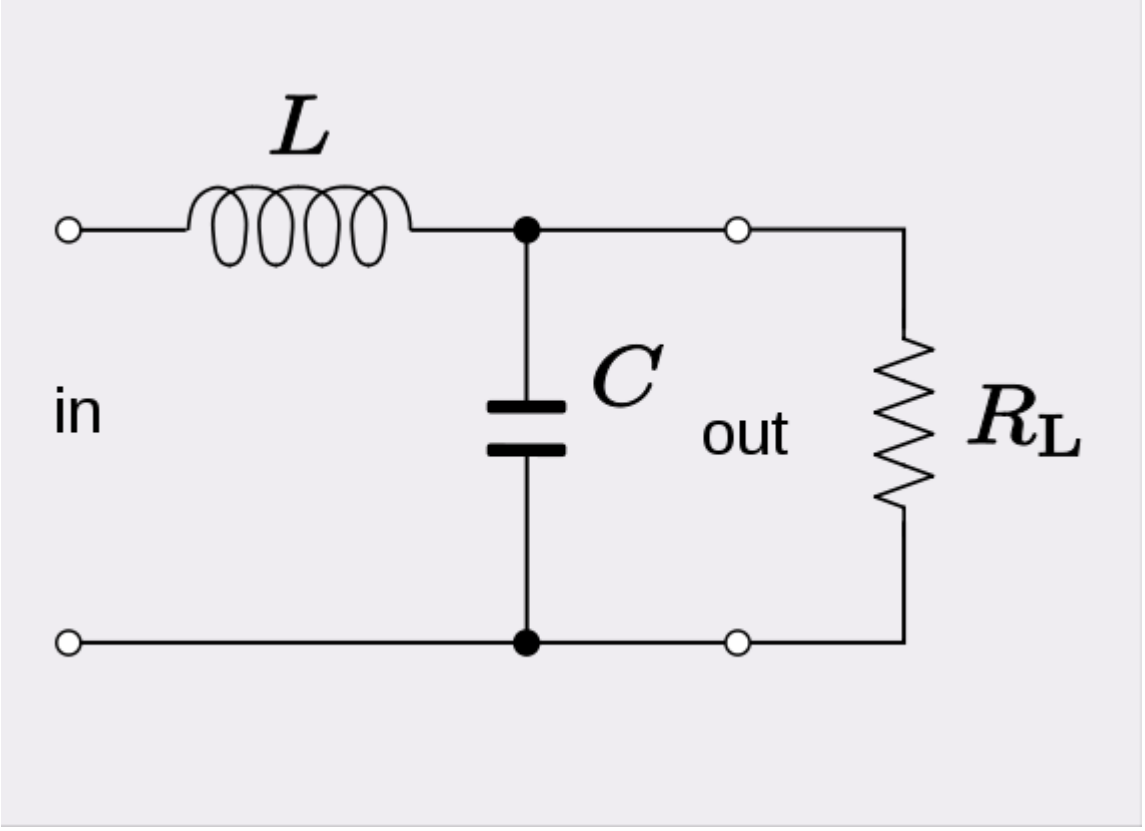


Fig. 9. RLC circuit as a low-pass filter

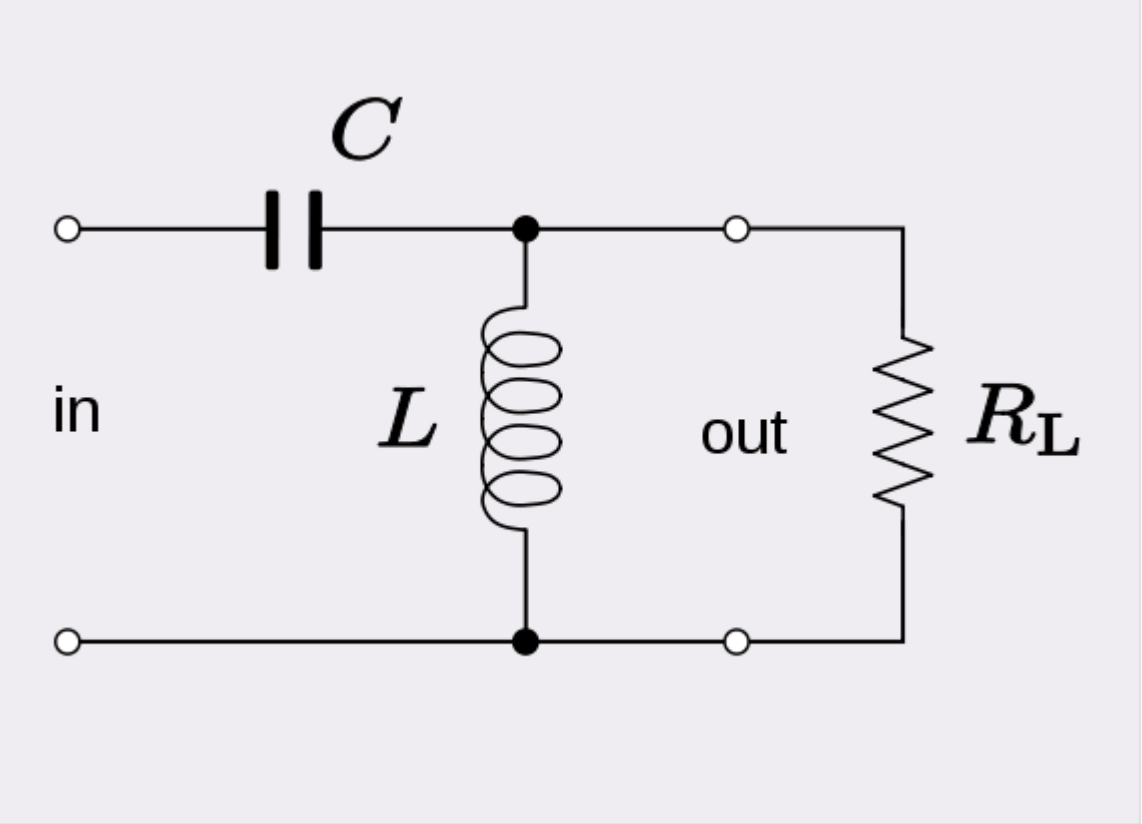


Fig. 10. RLC circuit as a high-pass filter

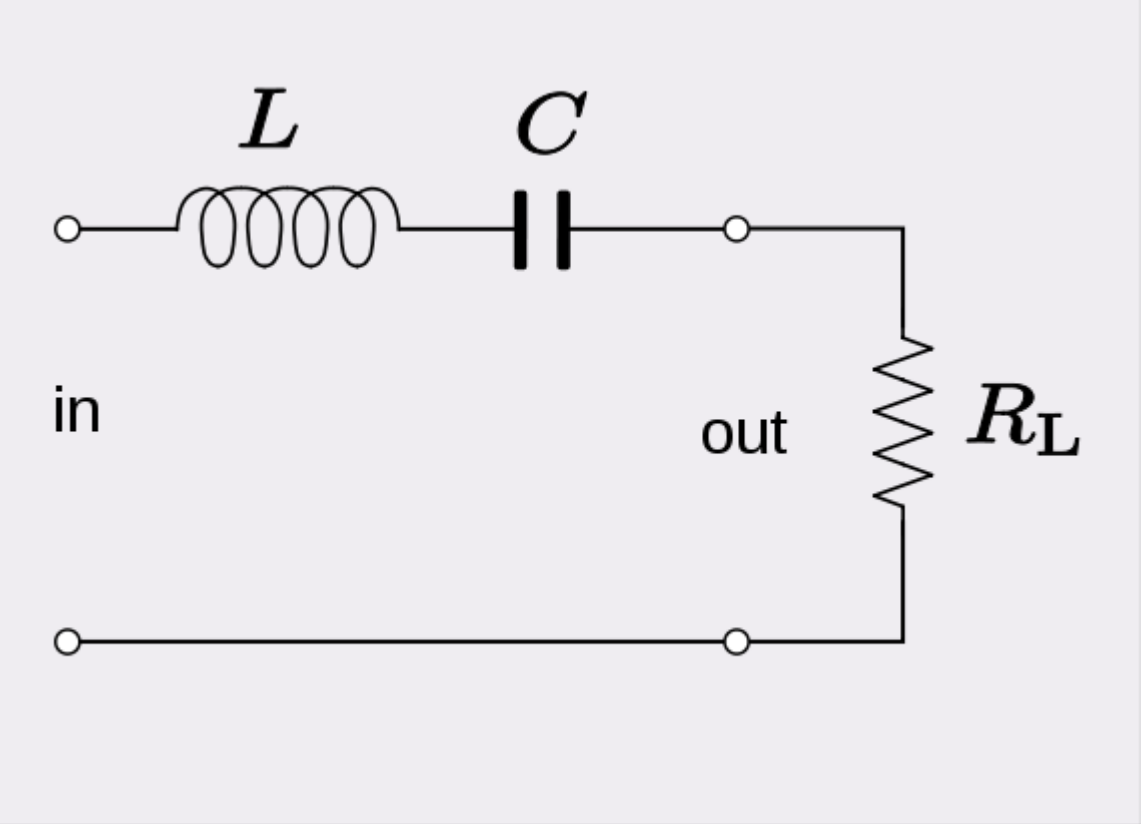


Fig. 11. RLC circuit as a series band-pass filter in series with the line

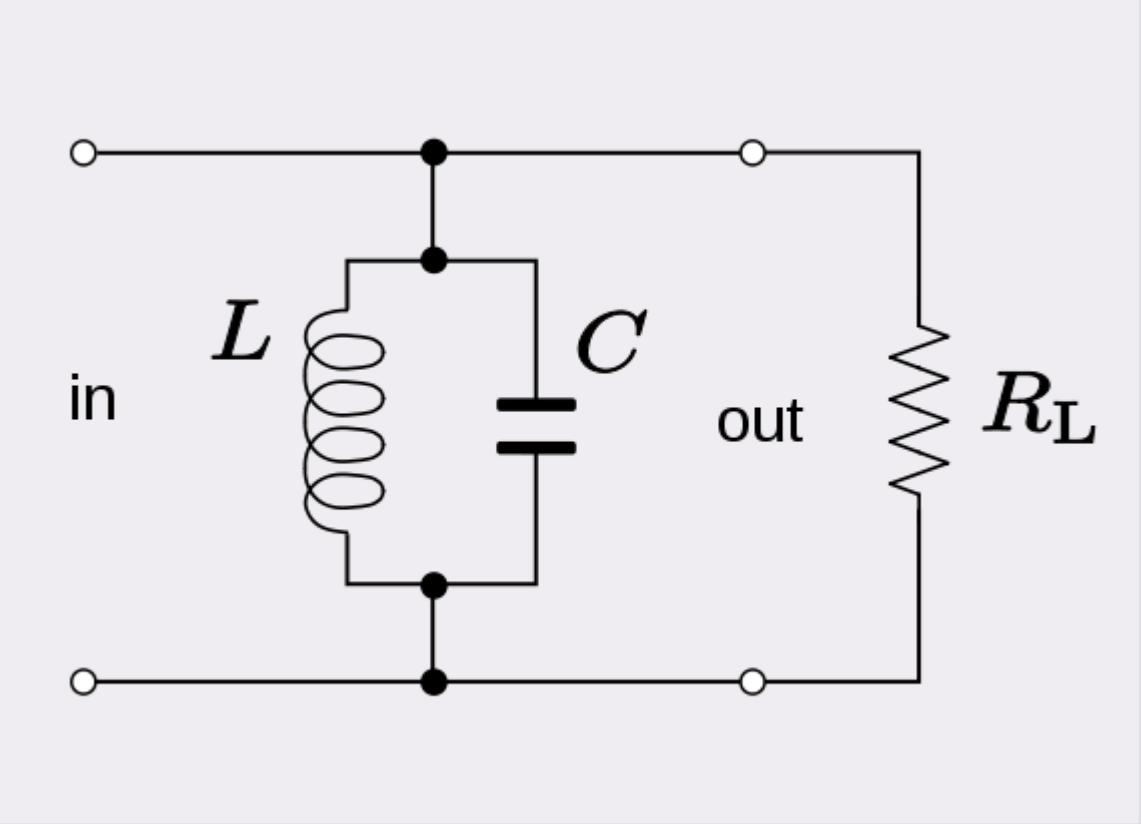


Fig. 12. RLC circuit as a parallel band-pass filter in shunt across the line

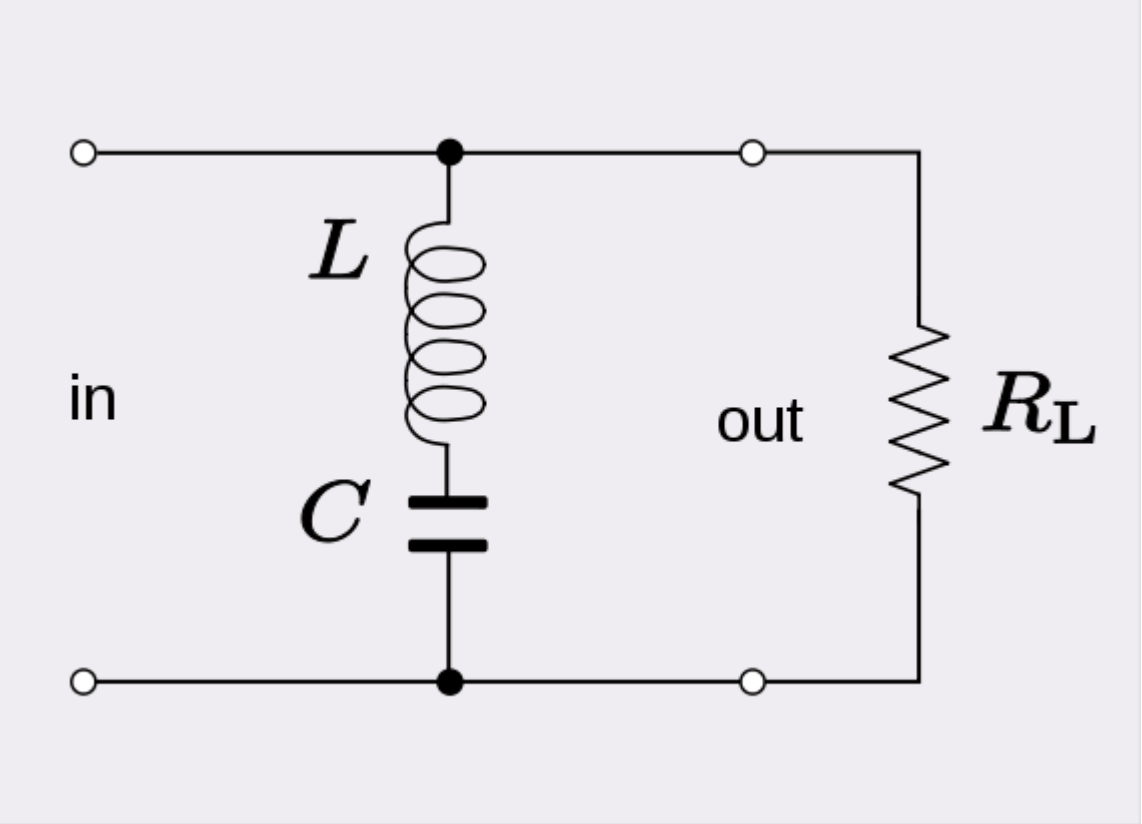


Fig. 13. RLC circuit as a series band-stop filter in shunt across the line

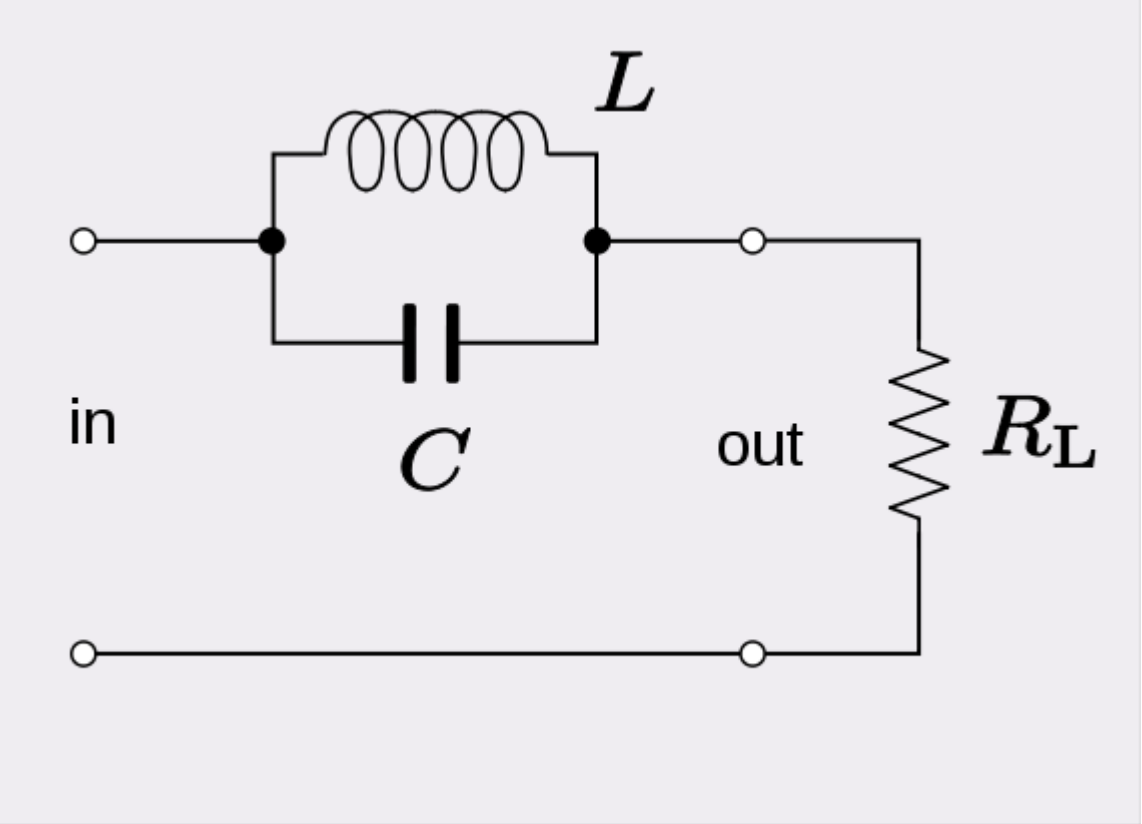


Fig. 14. RLC circuit as a parallel band-stop filter in series with the line

In the filtering application, the resistor *R* becomes the load that the filter is working into. The value of the damping factor is chosen based on the desired bandwidth of the filter. For a wider bandwidth, a larger value of the damping factor is required (and vice versa). The three components give the designer three degrees of freedom. Two of these are required to set the bandwidth and resonant frequency. The designer is still left with one which can be used to scale *R*, *L* and *C* to convenient practical values. Alternatively, *R* may be predetermined by the external circuitry which will use the last degree of freedom.

Low-pass filter

An RLC circuit can be used as a low-pass filter. The circuit configuration is shown in figure 9. The corner frequency, that is, the frequency of the 3dB point, is given by

$$\omega_c = \frac{1}{\sqrt{LC}}$$

This is also the bandwidth of the filter. The damping factor is given by

$$\zeta = \frac{1}{2R_L}\sqrt{\frac{L}{C}}$$

High-pass filter

A high-pass filter is shown in figure 10. The corner frequency is the same as the low-pass filter

$$\omega_{\mathrm{c}} = \frac{1}{\sqrt{LC}}$$

The filter has a stop-band of this width.

Band-pass filter

A band-pass filter can be formed with an RLC circuit by either placing a series LC circuit in series with the load resistor or else by placing a parallel LC circuit in parallel with the load resistor. These arrangements are shown in figures 11 and 12 respectively. The centre frequency is given by

$$\omega_{\mathrm{c}} = \frac{1}{\sqrt{LC}}$$

and the bandwidth for the series circuit is

$$\Delta\omega = \frac{R_{\mathrm{L}}}{L}$$

The shunt version of the circuit is intended to be driven by a high impedance source, that is, a constant current source. Under those conditions the bandwidth is

$$\Delta\omega = \frac{1}{CR_{\mathrm{L}}}$$

Band-stop filter

Figure 13 shows a band-stop filter formed by a series LC circuit in shunt across the load. Figure 14 is a band-stop filter formed by a parallel LC circuit in series with the load. The first case requires a high impedance source so that the current is diverted into the resonator when it becomes low impedance at resonance. The second case requires a low impedance source so that the voltage is dropped across the antiresonator when it becomes high impedance at resonance.

Oscillators

For applications in oscillator circuits, it is generally desirable to make the attenuation (or equivalently, the damping factor) as small as possible. In practice, this objective requires making the circuit's resistance *R* as small as physically possible for a series circuit, or

alternatively increasing *R* to as much as possible for a parallel circuit. In either case, the *RLC circuit* becomes a good approximation to an ideal LC circuit. However, for very low attenuation circuits (high Q-factor) circuits, issues such as dielectric losses of coils and capacitors can become important.

In an oscillator circuit

$$\alpha \ll \omega_0.$$

or equivalently

$$\zeta \ll 1.$$

As a result

$$\omega_d \approx \omega_0.$$

Voltage multiplier

In a series RLC circuit at resonance, the current is limited only by the resistance of the circuit

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

If R is small, consisting only of the inductor winding resistance say, then this current will be large. It will drop a voltage across the inductor of

$$V_\mathrm{L} = \frac{V}{R}\omega_0 L$$

An equal magnitude voltage will also be seen across the capacitor but in antiphase to the inductor. If R can be made sufficiently small, these voltages can be several times the input voltage. The voltage ratio is, in fact, the *Q* of the circuit,

$$\frac{V_\mathrm{L}}{V} = Q$$

A similar effect is observed with currents in the parallel circuit. Even though the circuit appears as high impedance to the external source, there is a large current circulating in the internal loop of the parallel inductor and capacitor.

Pulse discharge circuit

An overdamped series RLC circuit can be used as a pulse discharge circuit. Often it is useful to know the values of components that could be used to produce a waveform this is described by the form:

$$I(t) = I_0(e^{-\alpha t} - e^{-\beta t})$$

Such a circuit could consist of an energy storage capacitor, a load in the form of a resistance, some circuit inductance and a switch - all in series. The initial conditions are that the capacitor is at voltage V_0 and there is no current flowing in the inductor. If the inductance L is known, then the remaining parameters are given by the following - Capacitance:

$$C = \frac{1}{L\alpha\beta}$$

Resistance (total of circuit and load):

$$R = L(\alpha + \beta)$$

Initial terminal voltage of capacitor:

$$V_0 = -I_0 L\alpha\beta(\frac{1}{\beta} - \frac{1}{\alpha})$$

Rearranging for the case where R is known - Capacitance:

$$C = \frac{(\alpha + \beta)}{R\alpha\beta}$$

Inductance (total of circuit and load):

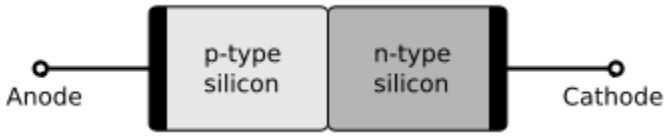
$$L = \frac{R}{(\alpha + \beta)}$$

Initial terminal voltage of capacitor:

$$V_0 = \frac{-I_0 R\alpha\beta}{(\alpha + \beta)}(\frac{1}{\beta} - \frac{1}{\alpha})$$

Chapter 7

p-n Junction



A silicon p–n junction with no applied voltage.

A **p–n junction** is formed by joining P-type and N-type semiconductors together in very close contact. The term *junction* refers to the boundary interface where the two regions of the semiconductor meet. If they were constructed of two separate pieces this would introduce a grain boundary, so p–n junctions are created in a single crystal of semiconductor by doping, for example by ion implantation, diffusion of dopants, or by epitaxy (growing a layer of crystal doped with one type of dopant on top of a layer of crystal doped with another type of dopant).

P-N junctions are elementary "building blocks" of almost all semiconductor electronic devices such as diodes, transistors, solar cells, LEDs, and integrated circuits; they are the active sites where the electronic action of the device takes place. For example, a common type of transistor, the bipolar junction transistor, consists of two p–n junctions in series, in the form n–p–n or p–n–p.

The discovery of the p–n junction is usually attributed to American physicist Russell Ohl of Bell Laboratories.

Schottky junction is a special case of a p-n junction, where metal serves the role of the n-type semiconductor.

Manufacture

Normally, p–n junctions are manufactured from a single crystal with different dopant concentrations diffused across it. Creating a semiconductor from two separate pieces of

material would introduce a grain boundary between the semiconductors which severely inhibits its utility by scattering the electrons and holes..

Note that, in the case of solar cells, polycrystalline silicon is often used to reduce expense, despite the lower efficiency caused by the grain boundaries. These boundaries are not related to the p-n junctions in the cell. If they would be the same spacially, the disturbing effects would make the solar cell useless.

Properties of a p-n junction

The p–n junction possesses some interesting properties which have useful applications in modern electronics. A p-doped semiconductor is relatively conductive. The same is true of an n-doped semiconductor, but the junction between them can become depleted of charge carriers, and hence nonconductive, depending on the relative voltages of the two semiconductor regions. By manipulating this non-conductive layer, p–n junctions are commonly used as diodes: circuit elements that allow a flow of electricity in one direction but not in the other (opposite) direction. This property is explained in terms of *forward bias* and *reverse bias*, where the term *bias* refers to an application of electric voltage to the p–n junction.

Equilibrium (zero bias)

In a p–n junction, without an external applied voltage, an equilibrium condition is reached in which a potential difference is formed across the junction. This potential difference is called built-in potential V_{bi}.

After joining p-type and n-type semiconductors, electrons near the p–n interface tend to diffuse into the p region. As electrons diffuse, they leave positively charged ions (donors) in the n region. Similarly, holes near the p–n interface begin to diffuse into the n-type region leaving fixed ions (acceptors) with negative charge. The regions nearby the p–n interfaces lose their neutrality and become charged, forming the space charge region or depletion layer (see figure A).

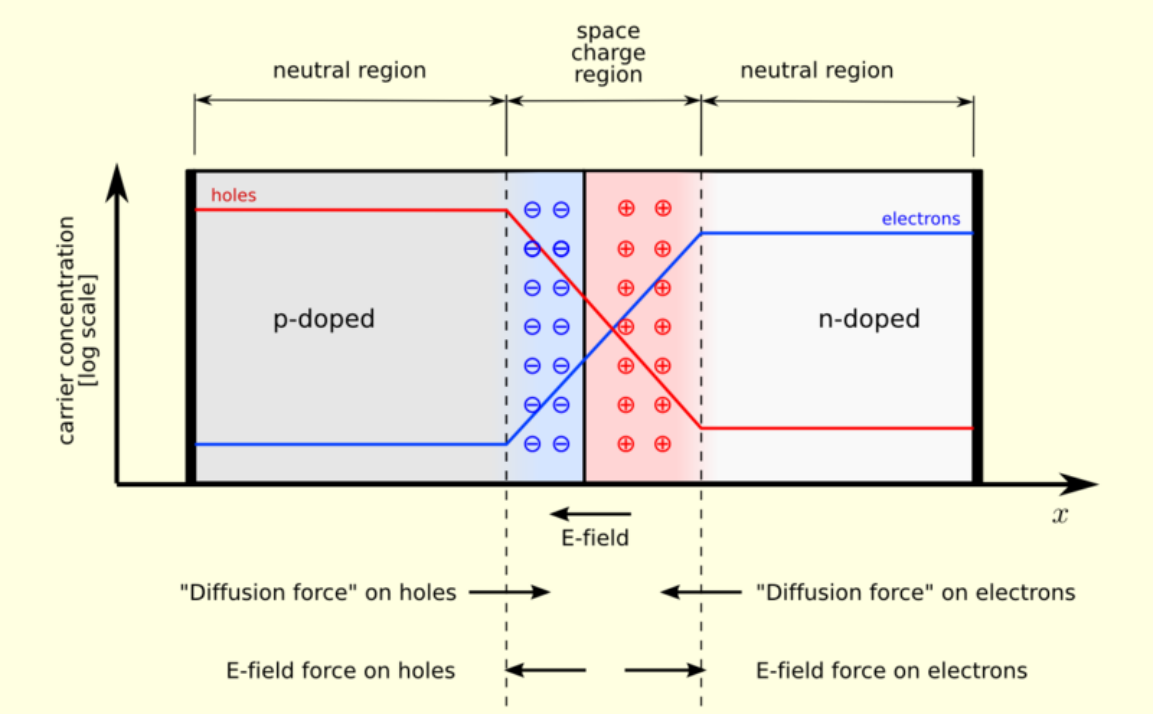


Figure A. A p–n junction in thermal equilibrium with zero bias voltage applied. Electrons and holes concentration are reported respectively with blue and red lines. Gray regions are charge neutral. Light red zone is positively charged. Light blue zone is negatively charged. The electric field is shown on the bottom, the electrostatic force on electrons and holes and the direction in which the diffusion tends to move electrons and holes.

The electric field created by the space charge region opposes the diffusion process for both electrons and holes. There are two concurrent phenomena: the diffusion process that tends to generate more space charge, and the electric field generated by the space charge that tends to counteract the diffusion. The carrier concentration profile at equilibrium is shown in figure A with blue and red lines. Also shown are the two counterbalancing phenomena that establish equilibrium.

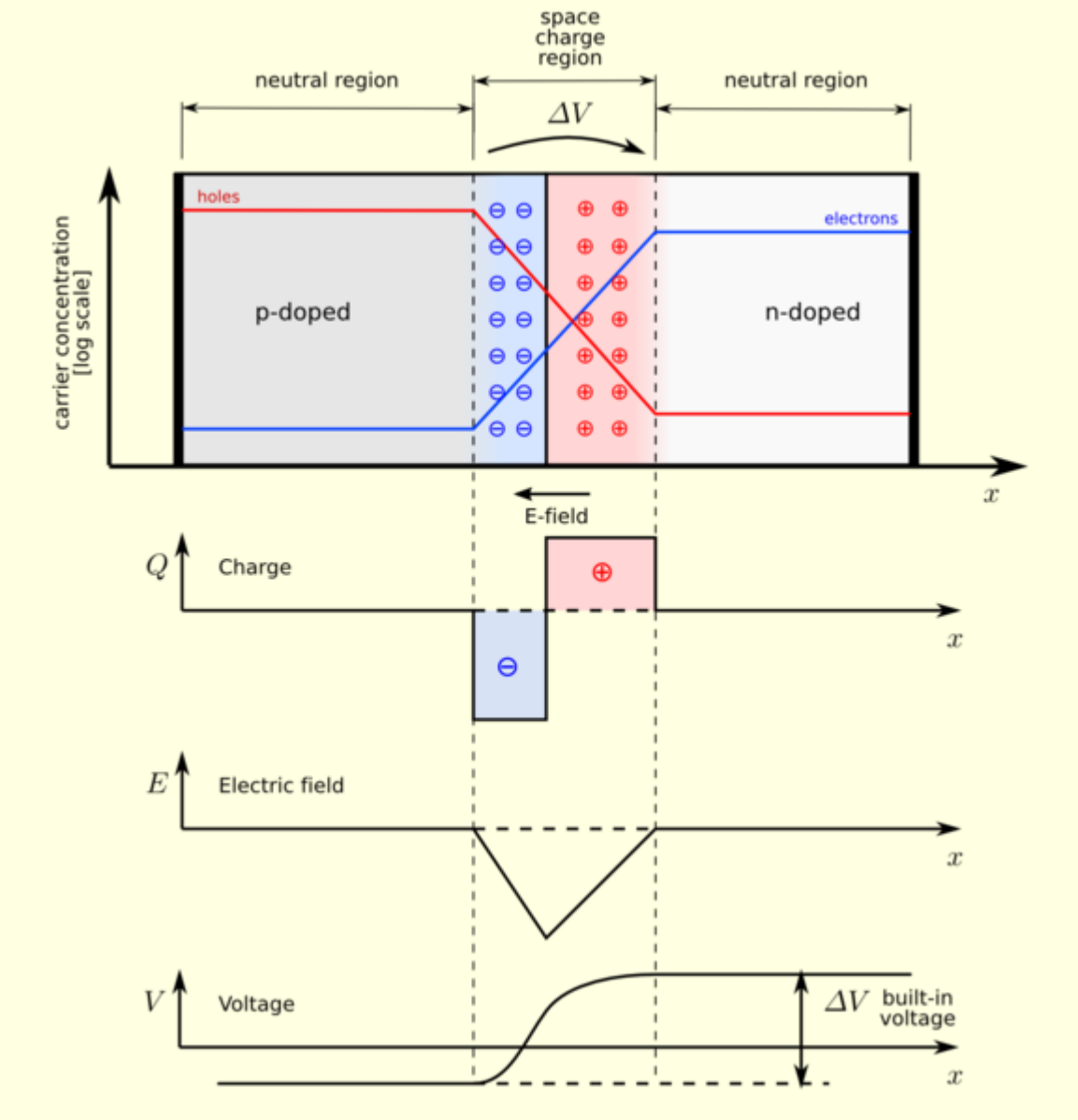
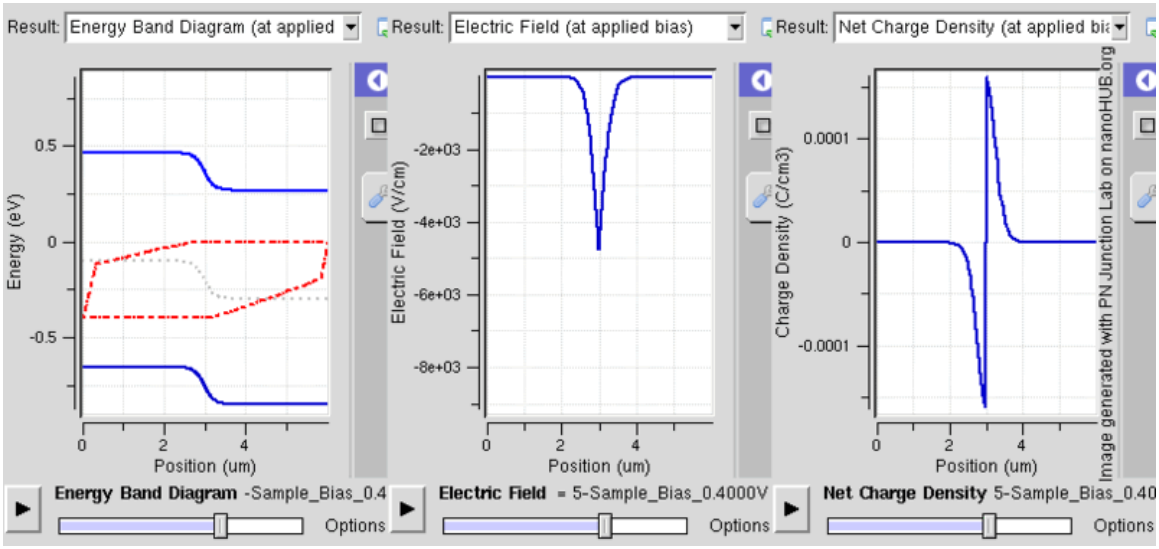


Figure B. A p–n junction in thermal equilibrium with zero bias voltage applied. Under the junction, plots for the charge density, the electric field and the voltage are reported.

The space charge region is a zone with a net charge provided by the fixed ions (donors or acceptors) that have been left *uncovered* by majority carrier diffusion. When equilibrium is reached, the charge density is approximated by the displayed step function. In fact, the region is completely depleted of majority carriers (leaving a charge density equal to the net doping level), and the edge between the space charge region and the neutral region is quite sharp (see figure B, Q(x) graph). The space charge region has the same magnitude of charge on both sides of the p–n interfaces, thus it extends farther on the less doped side (the n side in figures A and B).

Forward bias

In forward bias, the p-type is connected with the positive terminal and the n-type is connected with the negative terminal.



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.

With a battery connected this way, the holes in the P-type region and the electrons in the N-type region are pushed towards the junction. This reduces the width of the depletion zone. The positive charge applied to the P-type material repels the holes, while the negative charge applied to the N-type material repels the electrons. As electrons and holes are pushed towards the junction, the distance between them decreases. This lowers the barrier in potential. With increasing forward-bias voltage, the depletion zone eventually becomes thin enough that the zone's electric field can't counteract charge carrier motion across the p–n junction, consequently reducing electrical resistance. The electrons which cross the p–n junction into the P-type material (or holes which cross into the N-type material) will diffuse in the near-neutral region. Therefore, the amount of minority diffusion in the near-neutral zones determines the amount of current that may flow through the diode.

Only majority carriers (electrons in N-type material or holes in P-type) can flow through a semiconductor for a macroscopic length. With this in mind, consider the flow of electrons across the junction. The forward bias causes a force on the electrons pushing them from the N side toward the P side. With forward bias, the depletion region is narrow enough that electrons can cross the junction and *inject* into the P-type material. However, they do not continue to flow through the P-type material indefinitely, because it is energetically favorable for them to recombine with holes. The average length an electron

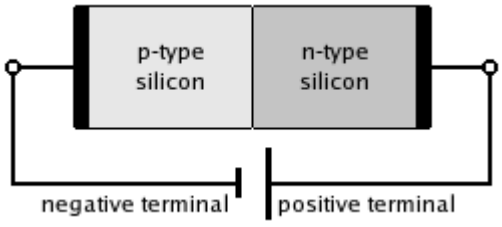
travels through the P-type material before recombining is called the *diffusion length*, and it is typically on the order of microns.

Although the electrons penetrate only a short distance into the P-type material, the electric current continues uninterrupted, because holes (the majority carriers) begin to flow in the opposite direction. The total current (the sum of the electron and hole currents) is constant in space, because any variation would cause charge buildup over time (this is Kirchhoff's current law). The flow of holes from the P-type region into the N-type region is exactly analogous to the flow of electrons from N to P (electrons and holes swap roles and the signs of all currents and voltages are reversed).

Therefore, the macroscopic picture of the current flow through the diode involves electrons flowing through the N-type region toward the junction, holes flowing through the P-type region in the opposite direction toward the junction, and the two species of carriers constantly recombining in the vicinity of the junction. The electrons and holes travel in opposite directions, but they also have opposite charges, so the overall current is in the same direction on both sides of the diode, as required.

The Shockley diode equation models the forward-bias operational characteristics of a p–n junction outside the avalanche (reverse-biased conducting) region.

Reverse bias



A silicon p–n junction in reverse bias.

Reverse biased usually refers to how a diode is used in a circuit. If a diode is reverse biased, the voltage at the cathode is higher than that at the anode. Therefore, no current will flow until the diode breaks down. Connecting the *P-type* region to the *negative* terminal of the battery and the *N-type* region to the *positive* terminal, corresponds to reverse bias. The connections are illustrated in the following diagram:

Because the p-type material is now connected to the negative terminal of the power supply, the 'holes' in the P-type material are pulled away from the junction, causing the width of the depletion zone to increase. Similarly, because the N-type region is connected to the positive terminal, the electrons will also be pulled away from the junction. Therefore the depletion region widens, and does so increasingly with increasing reverse-bias voltage. This increases the voltage barrier causing a high resistance to the flow of charge carriers thus allowing minimal electric current to cross the p–n junction. The increase in resistance of the p-n junction results in the junction behaving as an insulator.

The strength of the depletion zone electric field increases as the reverse-bias voltage increases. Once the electric field intensity increases beyond a critical level, the p–n junction depletion zone breaks-down and current begins to flow, usually by either the Zener or avalanche breakdown processes. Both of these breakdown processes are non-destructive and are reversible, so long as the amount of current flowing does not reach levels that cause the semiconductor material to overheat and cause thermal damage.

This effect is used to one's advantage in zener diode regulator circuits. Zener diodes have a certain - low - breakdown voltage. A standard value for breakdown voltage is for instance 5.6V. This means that the voltage at the cathode can never be more than 5.6V higher than the voltage at the anode, because the diode will break down - and therefore conduct - if the voltage gets any higher. This effectively regulates the voltage over the diode.

Another application where reverse biased diodes are used is in Varicap diodes. The width of the depletion zone of any diode changes with voltage applied. This varies the capacitance of the diode. For more information, refer to the Varicap article.

Electrostatics

For a p-n junction Poisson's equation becomes

$$\Delta\varphi = -\frac{\rho}{\varepsilon} = \frac{q}{\varepsilon}\left(\underbrace{n_0 - p_0}_{\substack{\text{equilibrium concentration}\\ \text{difference of free charges } (\approx 0)}} + \underbrace{N_A - N_D}_{\substack{\text{concentration difference}\\ \text{of acceptor and donor atoms}}}\right)$$

where φ is the electric potential, ρ is the charge density, ε is permittivity and q is the magnitude of the electron charge.

Since the total charge on either side of the depletion region must cancel out it is

$$\underbrace{d_p}_{\substack{\text{width of}\\ \text{electric field}\\ \text{within p-side}}} N_A = \underbrace{d_n}_{\substack{\text{width of}\\ \text{electric field}\\ \text{within n-side}}} N_D$$

From the above equations and by deploying basic calculus it can be shown that the total width of the depletion region is

$$d = d_p + d_n = \sqrt{\frac{2\varepsilon}{q}\frac{N_A + N_D}{N_A N_D}\left(\underbrace{V_{bi}}_{\text{built-in voltage}} - \underbrace{V}_{\text{external applied voltage}}\right)}$$

Furthermore, by implementing the Einstein relation and assuming the semiconductor is nondegenerate (i.e. the product $p_0 n_0$ is independent of the Fermi energy) it follows that

$$V_{bi} = \frac{kT}{q} \ln\left(\frac{N_A N_D}{p_0 n_0}\right)$$

where T is the temperature of the semiconductor and k is Boltzmann constant.

Summary

The forward-bias and the reverse-bias properties of the p–n junction imply that it can be used as a diode. A p–n junction diode allows electric charges to flow in one direction, but not in the opposite direction; negative charges (electrons) can easily flow through the junction from n to p but not from p to n and the reverse is true for holes. When the p–n junction is forward biased, electric charge flows freely due to reduced resistance of the p–n junction. When the p–n junction is reverse biased, however, the junction barrier (and therefore resistance) becomes greater and charge flow is minimal.

Non-rectifying junctions

In the above diagrams, contact between the metal wires and the semiconductor material also creates metal-semiconductor junctions called Schottky diodes. In a simplified ideal situation a semiconductor diode would never function, since it would be composed of several diodes connected back-to-front in series. But in practice, surface impurities within the part of the semiconductor which touches the metal terminals will greatly reduce the width of those depletion layers to such an extent that the metal-semiconductor junctions do not act as diodes. These "nonrectifying junctions" behave as ohmic contacts regardless of applied voltage polarity.

Chapter 8

Bipolar Junction Transistor

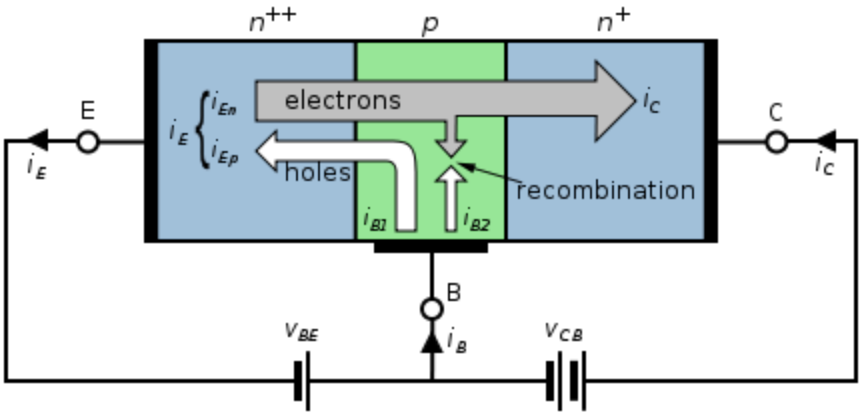
A **bipolar (junction) transistor** (**BJT**) is a three-terminal electronic device constructed of doped semiconductor material and may be used in amplifying or switching applications. *Bipolar* transistors are so named because their operation involves both electrons and holes. Charge flow in a BJT is due to bidirectional diffusion of charge carriers across a junction between two regions of different charge concentrations. This mode of operation is contrasted with *unipolar transistors,* such as field-effect transistors, in which only one carrier type is involved in charge flow due to drift. By design, most of the BJT collector current is due to the flow of charges injected from a high-concentration emitter into the base where they are minority carriers that diffuse toward the collector, and so BJTs are classified as *minority-carrier* devices.

PNP

NPN

Schematic symbols for PNP- and NPN-type BJTs.

Introduction



NPN BJT with forward-biased E–B junction and reverse-biased B–C junction

An NPN transistor can be considered as two diodes with a shared anode. In typical operation, the base-emitter junction is forward biased and the base–collector junction is reverse biased. In an NPN transistor, for example, when a positive voltage is applied to the base–emitter junction, the equilibrium between thermally generated carriers and the repelling electric field of the depletion region becomes unbalanced, allowing thermally excited electrons to inject into the base region. These electrons wander (or "diffuse") through the base from the region of high concentration near the emitter towards the region of low concentration near the collector. The electrons in the base are called *minority carriers* because the base is doped p-type which would make holes the *majority carrier* in the base.

To minimize the percentage of carriers that recombine before reaching the collector–base junction, the transistor's base region must be thin enough that carriers can diffuse across it in much less time than the semiconductor's minority carrier lifetime. In particular, the thickness of the base must be much less than the diffusion length of the electrons. The collector–base junction is reverse-biased, and so little electron injection occurs from the collector to the base, but electrons that diffuse through the base towards the collector are swept into the collector by the electric field in the depletion region of the collector–base junction. The thin *shared* base and asymmetric collector–emitter doping is what differentiates a bipolar transistor from two *separate* and oppositely biased diodes connected in series.

Voltage, current, and charge control

The collector–emitter current can be viewed as being controlled by the base–emitter current (current control), or by the base–emitter voltage (voltage control). These views are related by the current–voltage relation of the base–emitter junction, which is just the usual exponential current–voltage curve of a p-n junction (diode).

The physical explanation for collector current is the amount of minority-carrier charge in the base region. Detailed models of transistor action, such as the Gummel–Poon model, account for the distribution of this charge explicitly to explain transistor behavior more exactly. The charge-control view easily handles phototransistors, where minority carriers in the base region are created by the absorption of photons, and handles the dynamics of turn-off, or recovery time, which depends on charge in the base region recombining. However, because base charge is not a signal that is visible at the terminals, the current- and voltage-control views are generally used in circuit design and analysis.

In analog circuit design, the current-control view is sometimes used because it is approximately linear. That is, the collector current is approximately β_F times the base current. Some basic circuits can be designed by assuming that the emitter–base voltage is approximately constant, and that collector current is beta times the base current. However, to accurately and reliably design production BJT circuits, the voltage-control (for example, Ebers–Moll) model is required. The voltage-control model requires an exponential function to be taken into account, but when it is linearized such that the transistor can be modelled as a transconductance, as in the Ebers–Moll model, design for circuits such as differential amplifiers again becomes a mostly linear problem, so the voltage-control view is often preferred. For translinear circuits, in which the exponential I–V curve is key to the operation, the transistors are usually modelled as voltage controlled with transconductance proportional to collector current. In general, transistor level circuit design is performed using SPICE or a comparable analogue circuit simulator, so model complexity is usually not of much concern to the designer.

Turn-on, turn-off, and storage delay

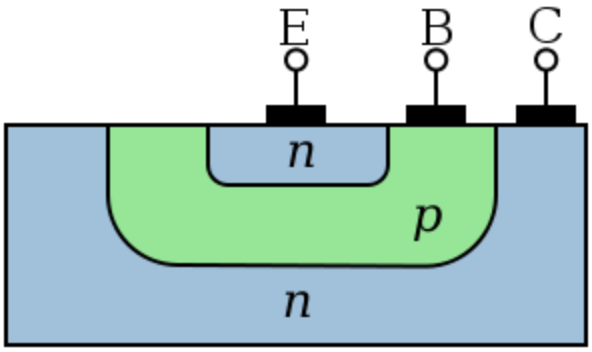
The Bipolar transistor exhibits a few delay characteristics when turning on and off. Most transistors, and especially power transistors, exhibit long base storage time that limits maximum frequency of operation in switching applications. One method for reducing this storage time is by using a Baker clamp.

Transistor 'alpha' and 'beta'

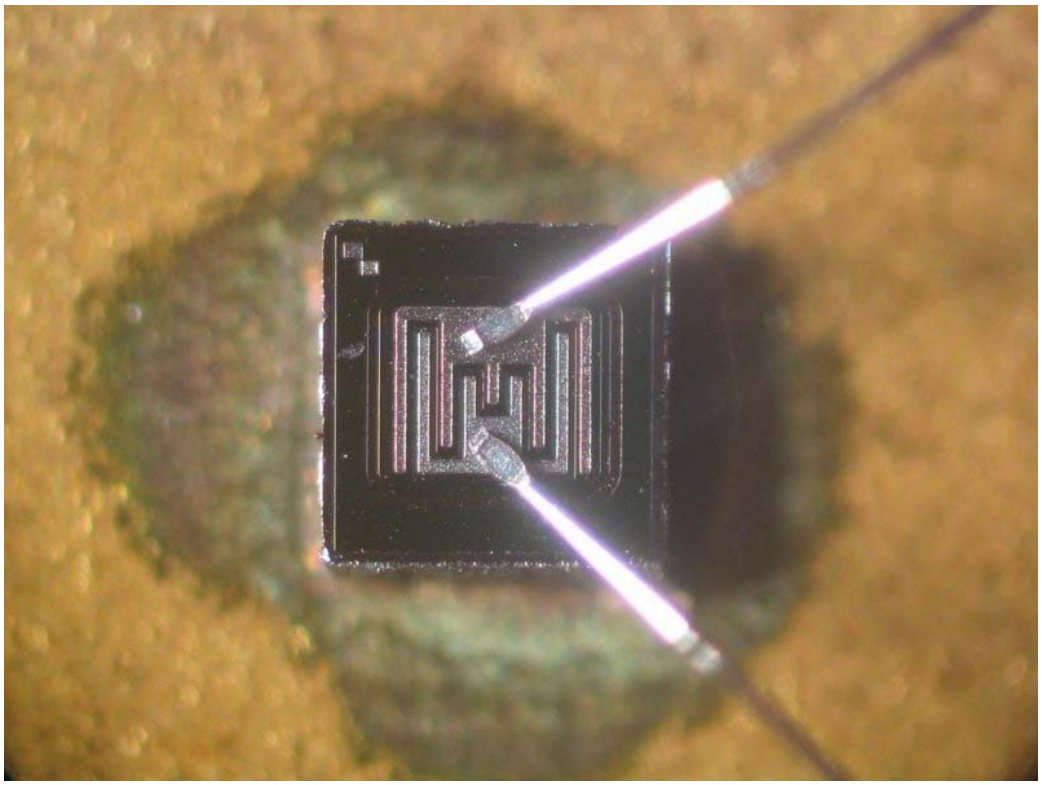
The proportion of electrons able to cross the base and reach the collector is a measure of the BJT efficiency. The heavy doping of the emitter region and light doping of the base region cause many more electrons to be injected from the emitter into the base than holes to be injected from the base into the emitter. The *common-emitter current gain* is represented by β_F or h_{fe}; it is approximately the ratio of the DC collector current to the DC base current in forward-active region. It is typically greater than 100 for small-signal transistors but can be smaller in transistors designed for high-power applications. Another important parameter is the common-base current gain, α_F. The common-base current gain is approximately the gain of current from emitter to collector in the forward-active region. This ratio usually has a value close to unity; between 0.98 and 0.998. Alpha and beta are more precisely related by the following identities (NPN transistor):

$$\alpha_F = \frac{I_\mathrm{C}}{I_\mathrm{E}}$$
$$\beta_F = \frac{I_\mathrm{C}}{I_\mathrm{B}}$$
$$\beta_F = \frac{\alpha_F}{1-\alpha_F} \iff \alpha_F = \frac{\beta_F}{\beta_F + 1}$$

Structure



Simplified cross section of a planar *NPN* bipolar junction transistor



Die of a KSY34 high-frequency NPN transistor, base and emitter connected via bonded wires

A BJT consists of three differently doped semiconductor regions, the *emitter* region, the *base* region and the *collector* region. These regions are, respectively, *p* type, *n* type and *p* type in a PNP, and *n* type, *p* type and *n* type in a NPN transistor. Each semiconductor region is connected to a terminal, appropriately labeled: *emitter* (E), *base* (B) and *collector* (C).

The *base* is physically located between the *emitter* and the *collector* and is made from lightly doped, high resistivity material. The collector surrounds the emitter region, making it almost impossible for the electrons injected into the base region to escape being collected, thus making the resulting value of α very close to unity, and so, giving the transistor a large β. A cross section view of a BJT indicates that the collector–base junction has a much larger area than the emitter–base junction.

The bipolar junction transistor, unlike other transistors, is usually not a symmetrical device. This means that interchanging the collector and the emitter makes the transistor leave the forward active mode and start to operate in reverse mode. Because the transistor's internal structure is usually optimized for forward-mode operation, interchanging the collector and the emitter makes the values of α and β in reverse operation much smaller than those in forward operation; often the α of the reverse mode is lower than 0.5. The lack of symmetry is primarily due to the doping ratios of the emitter and the collector. The emitter is heavily doped, while the collector is lightly doped, allowing a large reverse bias voltage to be applied before the collector–base junction breaks down. The collector–base junction is reverse biased in normal operation. The reason the emitter is heavily doped is to increase the emitter injection efficiency: the ratio of carriers injected by the emitter to those injected by the base. For high current gain, most of the carriers injected into the emitter–base junction must come from the emitter.

The low-performance "lateral" bipolar transistors sometimes used in CMOS processes are sometimes designed symmetrically, that is, with no difference between forward and backward operation.

Small changes in the voltage applied across the base–emitter terminals causes the current that flows between the *emitter* and the *collector* to change significantly. This effect can be used to amplify the input voltage or current. BJTs can be thought of as voltage-controlled current sources, but are more simply characterized as current-controlled current sources, or current amplifiers, due to the low impedance at the base.

Early transistors were made from germanium but most modern BJTs are made from silicon. A significant minority are also now made from gallium arsenide, especially for very high speed applications.

NPN

The symbol of an NPN bipolar junction transistor

NPN is one of the two types of bipolar transistors, consisting of a layer of P-doped semiconductor (the "base") between two N-doped layers. A small current entering the base is amplified to produce a large collector and emitter current. That is, an NPN transistor is "on" when its base is pulled **high** relative to the emitter.

Most of the NPN current is carried by electrons, moving from emitter to collector as minority carriers in the P-type base region. Most bipolar transistors used today are NPN, because electron mobility is higher than hole mobility in semiconductors, allowing greater currents and faster operation.

PNP

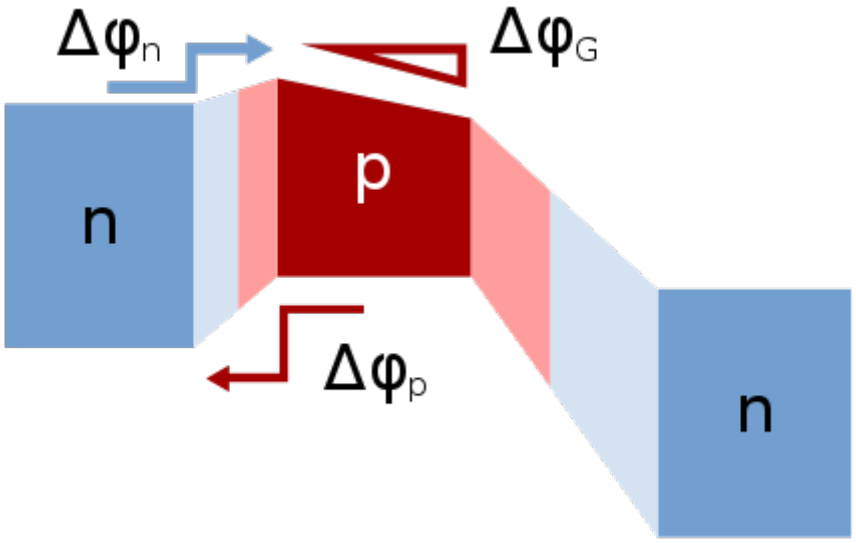
The symbol of a PNP Bipolar Junction Transistor.

The other type of BJT is the PNP, consisting of a layer of N-doped semiconductor between two layers of P-doped material. A small current leaving the base is amplified in the collector output. That is, a PNP transistor is "on" when its base is pulled **low** relative to the emitter.

The arrows in the NPN and PNP transistor symbols are on the emitter legs and point in the direction of the conventional current flow when the device is in forward active mode.

A mnemonic device for the NPN / PNP distinction, based on the arrows in their symbols and the letters in their names, is *not pointing in* for NPN and *pointing in* for PNP.

Heterojunction bipolar transistor



Bands in graded heterojunction NPN bipolar transistor. Barriers indicated for electrons to move from emitter to base, and for holes to be injected backward from base to emitter; Also, grading of bandgap in base assists electron transport in base region; Light colors indicate depleted regions

The heterojunction bipolar transistor (HBT) is an improvement of the BJT that can handle signals of very high frequencies up to several hundred GHz. It is common in modern ultrafast circuits, mostly RF systems. Heterojunction transistors have different semiconductors for the elements of the transistor. Usually the emitter is composed of a larger bandgap material than the base. The figure shows that this difference in bandgap allows the barrier for holes to inject backward into the base, denoted in figure as $\Delta\varphi_p$, to be made large, while the barrier for electrons to inject into the base $\Delta\varphi_n$ is made low. This barrier arrangement helps reduce minority carrier injection from the base when the emitter-base junction is under forward bias, and thus reduces base current and increases emitter injection efficiency.

The improved injection of carriers into the base allows the base to have a higher doping level, resulting in lower resistance to access the base electrode. In the more traditional BJT, also referred to as homojunction BJT, the efficiency of carrier injection from the emitter to the base is primarily determined by the doping ratio between the emitter and base, which means the base must be lightly doped to obtain high injection efficiency, making its resistance relatively high. In addition, higher doping in the base can improve figures of merit like the Early voltage by lessening base narrowing.

The grading of composition in the base, for example, by progressively increasing the amount of germanium in a SiGe transistor, causes a gradient in bandgap in the neutral base, denoted in the figure by $\Delta\varphi_G$, providing a "built-in" field that assists electron transport across the base. That drift component of transport aids the normal diffusive

transport, increasing the frequency response of the transistor by shortening the transit time across the base.

Two commonly used HBTs are silicon–germanium and aluminum gallium arsenide, though a wide variety of semiconductors may be used for the HBT structure. HBT structures are usually grown by epitaxy techniques like MOCVD and MBE.

Regions of operation

Applied voltages	B-E Junction Bias (NPN)	B-C Junction Bias (NPN)	Mode (NPN)
E < B < C	Forward	Reverse	Forward active
E < B > C	Forward	Forward	Saturation
E > B < C	Reverse	Reverse	Cut-off
E > B > C	Reverse	Forward	Reverse-active

Bipolar transistors have five distinct regions of operation, defined by BJT junction biases.

The modes of operation can be described in terms of the applied voltages (this description applies to NPN transistors; polarities are reversed for PNP transistors):

- Forward active: base higher than emitter, collector higher than base (in this mode the collector current is proportional to base current by β_F).
- Saturation: base higher than emitter, but collector is not higher than base.
- Cut-Off: base lower than emitter, but collector is higher than base. It means the transistor is not letting conventional current to go through collector to emitter.
- Reverse-active: base lower than emitter, collector lower than base: reverse conventional current goes through transistor.

Applied voltages	B-E Junction Bias (PNP)	B-C Junction Bias (PNP)	Mode (PNP)
E < B < C	Reverse	Forward	Reverse-active
E < B > C	Reverse	Reverse	Cut-off
E > B < C	Forward	Forward	Saturation
E > B > C	Forward	Reverse	Forward active

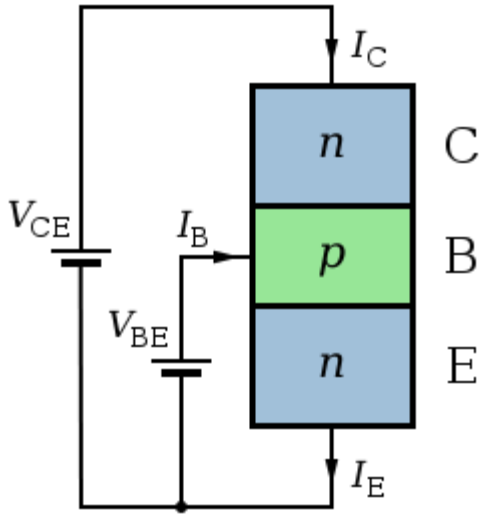
In terms of junction biasing: ('reverse biased base–collector junction' means Vbc < 0 for NPN, opposite for PNP)

- **Forward-active** (or simply, **active**): The base–emitter junction is forward biased and the base–collector junction is reverse biased. Most bipolar transistors are designed to afford the greatest common-emitter current gain, β_F, in forward-active mode. If this is the case, the collector–emitter current is approximately proportional to the base current, but many times larger, for small base current variations.

- **Reverse-active** (or **inverse-active** or **inverted**): By reversing the biasing conditions of the forward-active region, a bipolar transistor goes into reverse-active mode. In this mode, the emitter and collector regions switch roles. Because most BJTs are designed to maximize current gain in forward-active mode, the β_F in inverted mode is several (2–3 for the ordinary germanium transistor) times smaller. This transistor mode is seldom used, usually being considered only for failsafe conditions and some types of bipolar logic. The reverse bias breakdown voltage to the base may be an order of magnitude lower in this region.
- **Saturation**: With both junctions forward-biased, a BJT is in saturation mode and facilitates high current conduction from the emitter to the collector (or the other direction in the case of NPN, with negatively charged carriers flowing from emitter to collector). This mode corresponds to a logical "on", or a closed switch.
- **Cutoff**: In cutoff, biasing conditions opposite of saturation (both junctions reverse biased) are present. There is very little current, which corresponds to a logical "off", or an open switch.
- **Avalanche breakdown region**

Although these regions are well defined for sufficiently large applied voltage, they overlap somewhat for small (less than a few hundred millivolts) biases. For example, in the typical grounded-emitter configuration of an NPN BJT used as a pulldown switch in digital logic, the "off" state never involves a reverse-biased junction because the base voltage never goes below ground; nevertheless the forward bias is close enough to zero that essentially no current flows, so this end of the forward active region can be regarded as the cutoff region.

Active-mode NPN transistors in circuits



Structure and use of NPN transistor. Arrow according to schematic.

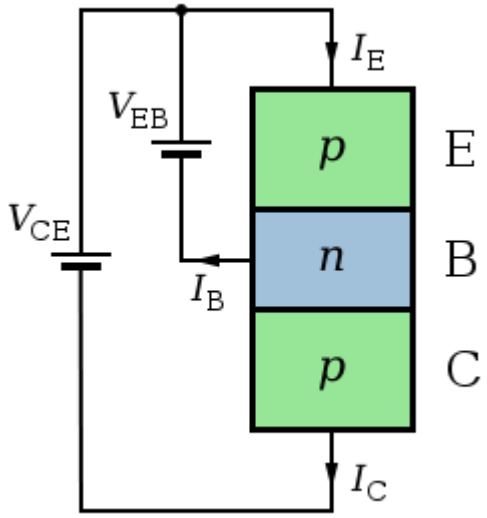
The diagram opposite is a schematic representation of an NPN transistor connected to two voltage sources. To make the transistor conduct appreciable current (on the order of

1 mA) from C to E, V_{BE} must be above a minimum value sometimes referred to as the **cut-in voltage**. The cut-in voltage is usually about 600 mV for silicon BJTs at room temperature but can be different depending on the type of transistor and its biasing. This applied voltage causes the lower P-N junction to 'turn-on' allowing a flow of electrons from the emitter into the base. In active mode, the electric field existing between base and collector (caused by V_{CE}) will cause the majority of these electrons to cross the upper P-N junction into the collector to form the collector current I_C. The remainder of the electrons recombine with holes, the majority carriers in the base, making a current through the base connection to form the base current, I_B. As shown in the diagram, the emitter current, I_E, is the total transistor current, which is the sum of the other terminal currents (i.e., $I_E = I_B + I_C$).

In the diagram, the arrows representing current point in the direction of conventional current – the flow of electrons is in the opposite direction of the arrows because electrons carry negative electric charge. In active mode, the ratio of the collector current to the base current is called the *DC current gain*. This gain is usually 100 or more, but robust circuit designs do not depend on the exact value. The value of this gain for DC signals is referred to as h_{FE}, and the value of this gain for AC signals is referred to as h_{fe}. However, when there is no particular frequency range of interest, the symbol β is used.

It should also be noted that the emitter current is related to V_{BE} exponentially. At room temperature, an increase in V_{BE} by approximately 60 mV increases the emitter current by a factor of 10. Because the base current is approximately proportional to the collector and emitter currents, they vary in the same way.

Active-mode PNP transistors in circuits



Structure and use of PNP transistor.

The diagram opposite is a schematic representation of a PNP transistor connected to two voltage sources. To make the transistor conduct appreciable current (on the order of 1

mA) from E to C, V_{EB} must be above a minimum value sometimes referred to as the **cut-in voltage**. The cut-in voltage is usually about 600 mV for silicon BJTs at room temperature but can be different depending on the type of transistor and its biasing. This applied voltage causes the upper P-N junction to 'turn-on' allowing a flow of holes from the emitter into the base. In active mode, the electric field existing between the emitter and the collector (caused by V_{CE}) causes the majority of these holes to cross the lower P-N junction into the collector to form the collector current I_C. The remainder of the holes recombine with electrons, the majority carriers in the base, making a current through the base connection to form the base current, I_B. As shown in the diagram, the emitter current, I_E, is the total transistor current, which is the sum of the other terminal currents (i.e., $I_E = I_B + I_C$).

In the diagram, the arrows representing current point in the direction of conventional current – the flow of holes is in the same direction of the arrows because holes carry positive electric charge. In active mode, the ratio of the collector current to the base current is called the *DC current gain*. This gain is usually 100 or more, but robust circuit designs do not depend on the exact value. The value of this gain for DC signals is referred to as h_{FE}, and the value of this gain for AC signals is referred to as h_{fe}. However, when there is no particular frequency range of interest, the symbol β is used.

It should also be noted that the emitter current is related to V_{EB} exponentially. At room temperature, an increase in V_{EB} by approximately 60 mV increases the emitter current by a factor of 10. Because the base current is approximately proportional to the collector and emitter currents, they vary in the same way.

History

The bipolar point-contact transistor was invented in December 1947 at the Bell Telephone Laboratories by John Bardeen and Walter Brattain under the direction of William Shockley. The junction version known as the bipolar junction transistor, invented by Shockley in 1948, enjoyed three decades as the device of choice in the design of discrete and integrated circuits. Nowadays, the use of the BJT has declined in favor of CMOS technology in the design of digital integrated circuits.

Germanium transistors

The germanium transistor was more common in the 1950s and 1960s, and while it exhibits a lower "cut off" voltage, typically around 0.2 V, making it more suitable for some applications, it also has a greater tendency to exhibit thermal runaway.

Early manufacturing techniques

Various methods of manufacturing bipolar junction transistors were developed.

- Point-contact transistor – first type to demonstrate transistor action, limited commercial use due to high cost and noise.

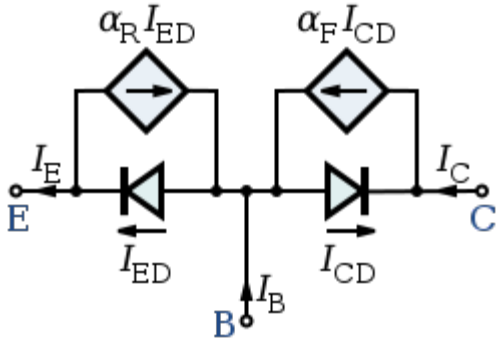
- Grown junction transistor – first type of bipolar junction transistor made. Invented by William Shockley at Bell Labs. Invented on June 23, 1948. Patent filed on June 26, 1948.
- Alloy junction transistor – emitter and collector alloy beads fused to base. Developed at General Electric and RCA in 1951.
    - Micro alloy transistor – high speed type of alloy junction transistor. Developed at Philco.
    - Micro alloy diffused transistor – high speed type of alloy junction transistor. Developed at Philco.
    - Post alloy diffused transistor – high speed type of alloy junction transistor. Developed at Philips.
- Tetrode transistor – high speed variant of grown junction transistor or alloy junction transistor with two connections to base.
- Surface barrier transistor – high speed metal barrier junction transistor. Developed at Philco in 1953.
- Drift-field transistor – high speed bipolar junction transistor. Invented by Herbert Kroemer at the Central Bureau of Telecommunications Technology of the German Postal Service, in 1953.
- Diffusion transistor – modern type bipolar junction transistor. Prototypes developed at Bell Labs in 1954.
    - Diffused base transistor – first implementation of diffusion transistor.
    - Mesa transistor – Developed at Texas Instruments in 1957.
    - Planar transistor – the bipolar junction transistor that made mass produced monolithic integrated circuits possible. Developed by Dr. Jean Hoerni at Fairchild in 1959.
- Epitaxial transistor – a bipolar junction transistor made using vapor phase deposition. Allows very precise control of doping levels and gradients.

Theory and modeling

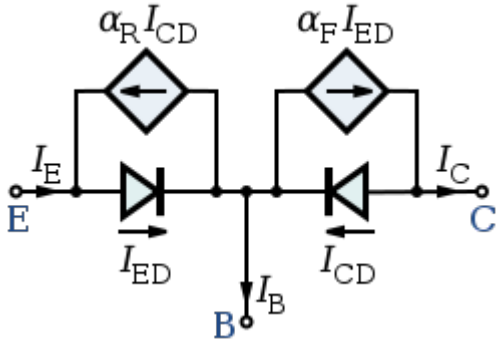
In the discussion below, focus is on the NPN bipolar transistor. In the NPN transistor in what is called **active mode** the base-emitter voltage V_{BE} and collector-base voltage V_{CB} are positive, forward biasing the emitter-base junction and reverse-biasing the collector-base junction. In active mode of operation, electrons are injected from the forward biased n-type emitter region into the p-type base where they diffuse to the reverse biased n-type collector and are swept away by the electric field in the reverse biased collector-base junction.

Large-signal models

Ebers–Moll model



Ebers–Moll Model for an NPN transistor. * I_B, I_C, I_E: base, collector and emitter currents * I_{CD}, I_{ED}: collector and emitter diode currents * α_F, α_R: forward and reverse common-base current gains



Ebers–Moll Model for a PNP transistor.

The DC emitter and collector currents in active mode are well modeled by an approximation to the Ebers–Moll model:

$$I_E = I_{ES}\left(e^{\frac{V_{BE}}{V_T}} - 1\right)$$
$$I_C = \alpha_T I_{ES}\left(e^{\frac{V_{BE}}{V_T}} - 1\right)$$

The base internal current is mainly by diffusion and

$$J_{n\,(\text{base})} = \frac{qD_n n_{bo}}{W} e^{\frac{V_{EB}}{V_T}}$$

where

- V_T is the thermal voltage kT / q (approximately 26 mV at 300 K ≈ room temperature).

- I_E is the emitter current
- I_C is the collector current
- α_T is the common base forward short circuit current gain (0.98 to 0.998)
- I_{ES} is the reverse saturation current of the base–emitter diode (on the order of 10^{-15} to 10^{-12} amperes)
- V_{BE} is the base–emitter voltage
- D_n is the diffusion constant for electrons in the p-type base
- W is the base width

The α and forward β parameters are as described previously. A reverse β is sometimes included in the model.

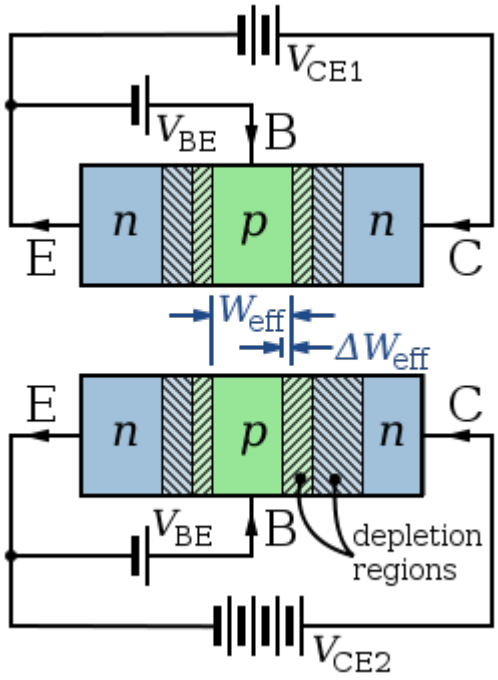
The unapproximated Ebers–Moll equations used to describe the three currents in any operating region are given below. These equations are based on the transport model for a bipolar junction transistor.

$$
\begin{aligned}
i_{\mathrm{C}} &= I_{\mathrm{S}}\left(e^{\frac{V_{\mathrm{BE}}}{V_{\mathrm{T}}}} - e^{\frac{V_{\mathrm{BC}}}{V_{\mathrm{T}}}}\right) - \frac{I_{\mathrm{S}}}{\beta_R}\left(e^{\frac{V_{\mathrm{BC}}}{V_{\mathrm{T}}}} - 1\right) \\
i_{\mathrm{B}} &= \frac{I_{\mathrm{S}}}{\beta_F}\left(e^{\frac{V_{\mathrm{BE}}}{V_{\mathrm{T}}}} - 1\right) + \frac{I_{\mathrm{S}}}{\beta_R}\left(e^{\frac{V_{\mathrm{BC}}}{V_{\mathrm{T}}}} - 1\right) \\
i_{\mathrm{E}} &= I_{\mathrm{S}}\left(e^{\frac{V_{\mathrm{BE}}}{V_{\mathrm{T}}}} - e^{\frac{V_{\mathrm{BC}}}{V_{\mathrm{T}}}}\right) + \frac{I_{\mathrm{S}}}{\beta_F}\left(e^{\frac{V_{\mathrm{BE}}}{V_{\mathrm{T}}}} - 1\right)
\end{aligned}
$$

where

- i_C is the collector current
- i_B is the base current
- i_E is the emitter current
- β_F is the forward common emitter current gain (20 to 500)
- β_R is the reverse common emitter current gain (0 to 20)
- I_S is the reverse saturation current (on the order of 10^{-15} to 10^{-12} amperes)
- V_T is the thermal voltage (approximately 26 mV at 300 K ≈ room temperature).
- V_{BE} is the base–emitter voltage
- V_{BC} is the base–collector voltage

Base-width modulation



Top: NPN base width for low collector-base reverse bias; Bottom: narrower NPN base width for large collector-base reverse bias. Hashed regions are depleted regions.

As the applied collector–base voltage ($V_{CB} = V_{CE} - V_{BE}$) varies, the collector–base depletion region varies in size. An increase in the collector–base voltage, for example, causes a greater reverse bias across the collector–base junction, increasing the collector–base depletion region width, and decreasing the width of the base. This variation in base width often is called the "Early effect" after its discoverer James M. Early.

Narrowing of the base width has two consequences:

- There is a lesser chance for recombination within the "smaller" base region.
- The charge gradient is increased across the base, and consequently, the current of minority carriers injected across the emitter junction increases.

Both factors increase the collector or "output" current of the transistor in response to an increase in the collector–base voltage.

In the forward-active region, the Early effect modifies the collector current (i_C) and the forward common emitter current gain (β_F) as given by:

$$i_C = I_S e^{\frac{v_{BE}}{V_T}} \left(1 + \frac{V_{CB}}{V_A}\right)$$

$$\beta_F = \beta_{F0}\left(1 + \frac{V_{\mathrm{CB}}}{V_{\mathrm{A}}}\right)$$
$$r_{\mathrm{o}} = \frac{V_{\mathrm{A}}}{I_{\mathrm{C}}}$$

where:

- V_{CB} is the collector–base voltage
- V_{A} is the Early voltage (15 V to 150 V)
- β_{F0} is forward common-emitter current gain when $V_{\mathrm{CB}} = 0$ V
- r_{o} is the output impedance
- I_{C} is the collector current

Current–voltage characteristics

The following assumptions are involved when deriving ideal current-voltage characteristics of the BJT

- Low level injection
- Uniform doping in each region with abrupt junctions
- One-dimensional current
- Negligible recombination-generation in space charge regions
- Negligible electric fields outside of space charge regions.

It is important to characterize the minority diffusion currents induced by injection of carriers.

With regard to pn-junction diode, a key relation is the diffusion equation.

$$\frac{d^2\Delta p_{\mathrm{B}}(x)}{dx^2} = \frac{\Delta p_{\mathrm{B}}(x)}{L_{\mathrm{B}}^2}$$

A solution of this equation is below, and two boundary conditions are used to solve and find C_1 and C_2.

$$\Delta p_{\mathrm{B}}(x) = C_1 e^{x/L_{\mathrm{B}}} + C_2 e^{-x/L_{\mathrm{B}}}$$

The following equations apply to the emitter and collector region, respectively, and the origins 0, 0', and 0'' apply to the base, collector, and emitter.

$$\Delta n_{\mathrm{B}}(x'') = A_1 e^{x''/L_{\mathrm{B}}} + A_2 e^{-x''/L_{\mathrm{B}}}$$
$$\Delta n_{\mathrm{c}}(x') = B_1 e^{x'/L_{\mathrm{B}}} + B_2 e^{-x'/L_{\mathrm{B}}}$$

A boundary condition of the emitter is below:

$$\Delta n_{\mathrm{E}}(0'') = n_{EO}(\exp(qV_{\mathrm{EB}}/kT) - 1)$$

The values of the constants A_1 and B_1 are zero due to the following conditions of the emitter and collector regions as $x'' \to 0$and $x' \to 0$.

$$\Delta n_{\mathrm{E}}(x'') \to 0$$
$$\Delta n_{\mathrm{c}}(x') \to 0$$

Because $A_1 = B_1 = 0$, the values of $\Delta n_{\mathrm{E}}(0'')$ and $\Delta n_{\mathrm{c}}(0')$ are A_2 and B_2, respectively.

$$\Delta n_{\mathrm{E}}(x'') = n_{\mathrm{E0}}(\exp(qV_{\mathrm{EB}}/kT) - 1)\exp(-x''/L_{\mathrm{E}})$$
$$\Delta n_{\mathrm{C}}(x') = n_{\mathrm{C0}}(\exp(qV_{\mathrm{CB}}/kT) - 1)\exp(-x'/L_{\mathrm{C}})$$

Expressions of $I_{\mathrm{E}n}$ and $I_{\mathrm{C}n}$ can be evaluated.

$$I_{\mathrm{E}n} = -qAD_{\mathrm{E}}\frac{d\Delta_{\mathrm{E}}(x'')}{dx}|_{x''=0''}$$
$$I_{\mathrm{C}n} = -qA\frac{D_{\mathrm{C}}}{L_{\mathrm{C}}}n_{\mathrm{C0}}(\exp(qV_{\mathrm{CB}}/kT) - 1)$$

Because insignificant recombination occurs, the second derivative of $\Delta p_{\mathrm{B}}(x)$ is zero. There is therefore a linear relationship between excess hole density and x.

$$\Delta p_{\mathrm{B}}(x) = D_1 x + D_2$$

The following are boundary conditions of Δp_{B}.

$$\Delta p_{\mathrm{B}}(0) = D_2$$
$$\Delta p_{\mathrm{B}}(W) = D_1 W + \Delta p_{\mathrm{B}}(0)$$

Substitute into the above linear relation.

$$\Delta p_{\mathrm{B}}(x) = -\left(\frac{\Delta p_{\mathrm{B}}(0) - \Delta p_{\mathrm{B}}(W)}{W}\right)x + \Delta p_{\mathrm{B}}(0)$$

.

With this result, derive value of $I_{\mathrm{E}p}$.

$$I_{\mathrm{E}p}(0) = -qAD_{\mathrm{B}}\frac{d\Delta p_{\mathrm{B}}}{dx}|_{x=0}$$
$$I_{\mathrm{E}p}(0) = \frac{qAD_{\mathrm{B}}}{W}(\Delta p_{\mathrm{B}}(0) - \Delta p_{\mathrm{B}}(W))$$

Use the expressions of I_{Ep}, I_{En}, $\Delta p_B(0)$, and $\Delta p_B(W)$ to develop an expression of the emitter current.

$$\begin{aligned}\Delta p_B(W) &= p_{B0}\exp(qV_{CB}/kT)\\ \Delta p_B(0) &= p_{B0}\exp(qV_{EB}/kT))\\ I_E = qA&\left(\left(\frac{D_E n_{E0}}{L_E}+\frac{D_B p_{B0}}{W}\right)\left(\exp\left(\frac{qV_{EB}}{kT}\right)-1\right)-\left(\frac{D_B}{W}p_{B0}\right)\left(\exp\left(\frac{qV_{CB}}{kT}\right)-1\right)\right)\end{aligned}$$

Similarly, an expression of the collector current is derived.

$$\begin{aligned}I_{Cp}(W) &= I_{Ep}(0)\\ I_C &= I_{Cp}(W)+I_{Cn}(0')\\ I_C = qA&\left(\left(\frac{D_B}{W}p_{B0}\right)(\exp(qV_{EB}/kT)-1)-\left(\frac{D_C n_{C0}}{L_C}+\frac{D_B p_{B0}}{W}\right)(\exp(qV_{CB}/kT)-1)\right)\end{aligned}$$

An expression of the base current is found with the previous results.

$$\begin{aligned}I_B &= I_E - I_C\\ I_B = qA&\left(\frac{D_E}{L_E}n_{E0}(\exp(qV_{EB}/kT)-1)+\frac{D_C}{L_C}n_{C0}(\exp(qV_{CB}/kT)-1)\right)\end{aligned}$$

Punchthrough

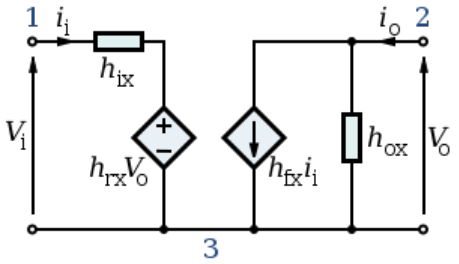
When the base–collector voltage reaches a certain (device specific) value, the base–collector depletion region boundary meets the base–emitter depletion region boundary. When in this state the transistor effectively has no base. The device thus loses all gain when in this state.

Gummel–Poon charge-control model

The Gummel–Poon model is a detailed charge-controlled model of BJT dynamics, which has been adopted and elaborated by others to explain transistor dynamics in greater detail than the terminal-based models typically do . This model also includes the dependence of transistor β-values upon the direct current levels in the transistor, which are assumed current-independent in the Ebers–Moll model.

Small-signal models

h-parameter model



Generalized h-parameter model of an NPN BJT.
*Replace **x** with **e**, **b** or **c** for CE, CB and CC topologies respectively.*

Another model commonly used to analyze BJT circuits is the "h-parameter" model, closely related to the hybrid-pi model and the y-parameter two-port, but using input current and output voltage as independent variables, rather than input and output voltages. This two-port network is particularly suited to BJTs as it lends itself easily to the analysis of circuit behaviour, and may be used to develop further accurate models. As shown, the term *"x"* in the model represents a different BJT lead depending on the topology used. For common-emitter mode the various symbols take on the specific values as:

- x = 'e' because it is a common-emitter topology
- Terminal 1 = Base
- Terminal 2 = Collector
- Terminal 3 = Emitter
- i_i = Base current (i_b)
- i_o = Collector current (i_c)
- V_{in} = Base-to-emitter voltage (V_{BE})
- V_o = Collector-to-emitter voltage (V_{CE})

and the h-parameters are given by:

- $h_{ix} = h_{ie}$ – The input impedance of the transistor (corresponding to the emitter resistance r_e).
- $h_{rx} = h_{re}$ – Represents the dependence of the transistor's I_B–V_{BE} curve on the value of V_{CE}. It is usually very small and is often neglected (assumed to be zero).
- $h_{fx} = h_{fe}$ – The current-gain of the transistor. This parameter is often specified as h_{FE} or the DC current-gain (β_{DC}) in datasheets.
- $h_{ox} = 1/h_{oe}$ – The output impedance of transistor. The parameter h_{oe} usually corresponds to the output admittance of the bipolar transistor and has to be inverted to convert it to an impedance.

As shown, the h-parameters have lower-case subscripts and hence signify AC conditions or analyses. For DC conditions they are specified in upper-case. For the CE topology, an approximate h-parameter model is commonly used which further simplifies the circuit

analysis. For this the h_{oe} and h_{re} parameters are neglected (that is, they are set to infinity and zero, respectively). It should also be noted that the h-parameter model as shown is suited to low-frequency, small-signal analysis. For high-frequency analyses the inter-electrode capacitances that are important at high frequencies must be added.

Etymology of h_{FE}

The 'h' refers to its being an h-parameter, a set of parameters named for their origin in a ***h****ybrid equivalent circuit* model. 'F' is from ***f****orward current amplification* also called the current gain. 'E' refers to the transistor operating in a *common* ***e****mitter* (CE) configuration. Capital letters used in the subscript indicate that h_{FE} refers to a direct current circuit.

Applications

The BJT remains a device that excels in some applications, such as discrete circuit design, due to the very wide selection of BJT types available, and because of its high transconductance and output resistance compared to MOSFETs. The BJT is also the choice for demanding analog circuits, especially for very-high-frequency applications, such as radio-frequency circuits for wireless systems. Bipolar transistors can be combined with MOSFETs in an integrated circuit by using a BiCMOS process of wafer fabrication to create circuits that take advantage of the application strengths of both types of transistor.

Temperature sensors

Because of the known temperature and current dependence of the forward-biased base–emitter junction voltage, the BJT can be used to measure temperature by subtracting two voltages at two different bias currents in a known ratio .

Logarithmic converters

Because base–emitter voltage varies as the log of the base–emitter and collector–emitter currents, a BJT can also be used to compute logarithms and anti-logarithms. A diode can also perform these nonlinear functions, but the transistor provides more circuit flexibility.

Vulnerabilities

Exposure of the transistor to ionizing radiation causes radiation damage. Radiation causes a buildup of 'defects' in the base region that act as recombination centers. The resulting reduction in minority carrier lifetime causes gradual loss of gain of the transistor.

Power BJTs are subject to a failure mode called secondary breakdown, in which excessive current and normal imperfections in the silicon die cause portions of the silicon inside the device to become disproportionately hotter than the others. The doped silicon has a negative temperature coefficient, meaning that it conducts more current at higher temperatures. Thus, the hottest part of the die conducts the most current, causing its

conductivity to increase, which then causes it to become progressively hotter again, until the device fails internally. The thermal runaway process associated with secondary breakdown, once triggered, occurs almost instantly and may catastrophically damage the transistor package.

If the emitter-base junction is reverse biased into avalanche (zener) mode and current flows for a short period of time, the current gain of the BJT will be permanently degraded.

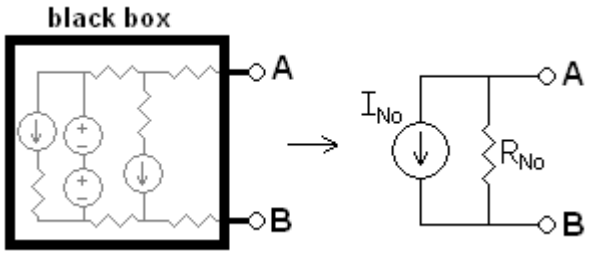
Chapter 9

Norton's Theorem and Kirchhoff's Circuit Laws

Norton's theorem

Norton's theorem for linear electrical networks, known in Europe as the **Mayer–Norton theorem**, states that any collection of voltage sources, current sources, and resistors with two terminals is electrically equivalent to an ideal current source, I, in parallel with a single resistor, R. For single-frequency AC systems the theorem can also be applied to general impedances, not just resistors. The **Norton equivalent** is used to represent any network of linear sources and impedances, at a given frequency. The circuit consists of an ideal current source in parallel with an ideal impedance (or resistor for non-reactive circuits).

Norton's theorem is an extension of Thévenin's theorem and was introduced in 1926 separately by two people: Siemens & Halske researcher Hans Ferdinand Mayer (1895–1980) and Bell Labs engineer Edward Lawry Norton (1898–1983). Only Mayer actually published on this topic, but Norton made known his finding through an internal technical report at Bell Labs.



Any black box containing only voltage sources, current sources, and resistors can be converted to a Norton equivalent circuit.

Calculation of a Norton equivalent circuit

The Norton equivalent circuit is a current source with current I_{No} in parallel with a resistance R_{No}. To find the equivalent,

1. Find the Norton current I_{No}. Calculate the output current, I_{AB}, with a short circuit as the load (meaning 0 resistance between A and B). This is I_{No}.
2. Find the Norton resistance R_{No}. When there are **no dependent sources** (i.e., all current and voltage sources are **independent**), there are two methods of determining the Norton impedance R_{No}.

    - Calculate the output voltage, V_{AB}, when in open circuit condition (i.e., no load resistor — meaning infinite load resistance). R_{No} equals this V_{AB} divided by I_{No}.

    or

    - Replace independent voltage sources with short circuits and independent current sources with open circuits. The total resistance across the output port is the Norton impedance R_{No}.

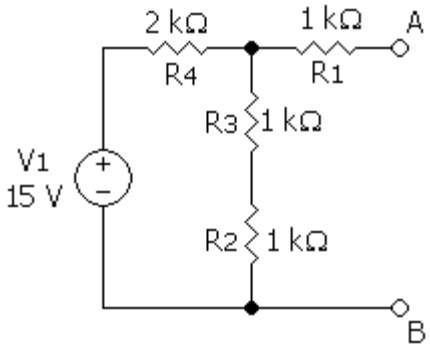
    or

    - Use a given Thevenin resistance: as the two are equal.

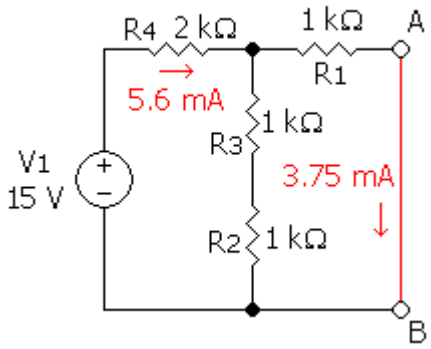
    However, when there are **dependent sources**, the more general method must be used. This method is not shown below in the diagrams.

    - Connect a constant current source at the output terminals of the circuit with a value of 1 Ampere and calculate the voltage at its terminals. The quotient of this voltage divided by the 1 A current is the Norton impedance R_{No}. This method must be used if the circuit contains dependent sources, but it can be used in all cases even when there are no dependent sources.

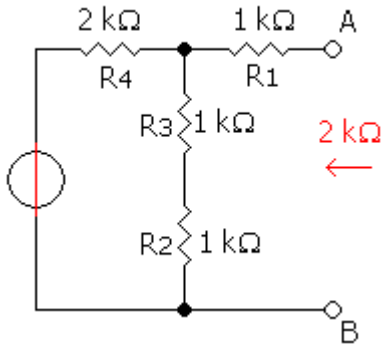
Example of a Norton equivalent circuit



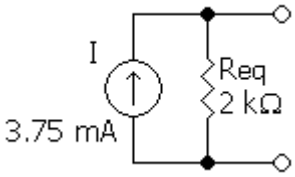
Step 0: The original circuit



Step 1: Calculating the equivalent output current



Step 2: Calculating the equivalent resistance



Step 3: The equivalent circuit

In the example, the total current I_{total} is given by:

$$I_{\text{total}} = \frac{15\text{V}}{2\,\text{k}\Omega + 1\,\text{k}\Omega \| (1\,\text{k}\Omega + 1\,\text{k}\Omega)} = 5.625\text{mA}$$

The current through the load is then, using the current divider rule:

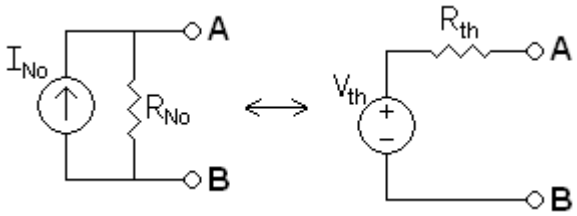
$$I = \frac{1\,\text{k}\Omega + 1\,\text{k}\Omega}{(1\,\text{k}\Omega + 1\,\text{k}\Omega + 1\,\text{k}\Omega)} \cdot I_{\text{total}}$$
$$= 2/3 \cdot 5.625\text{mA} = 3.75\text{mA}$$

And the equivalent resistance looking back into the circuit is:

$$R_{\text{eq}} = 1\,\text{k}\Omega + 2\,\text{k}\Omega \| (1\,\text{k}\Omega + 1\,\text{k}\Omega) = 2\,\text{k}\Omega$$

So the equivalent circuit is a 3.75 mA current source in parallel with a 2 kΩ resistor.

Conversion to a Thévenin equivalent



A Norton equivalent circuit is related to the Thévenin equivalent by the following equations:

$$R_{Th} = R_{No}$$
$$V_{Th} = I_{No} R_{No}$$
$$V_{Th}/R_{Th} = I_{No}$$

Queueing theory

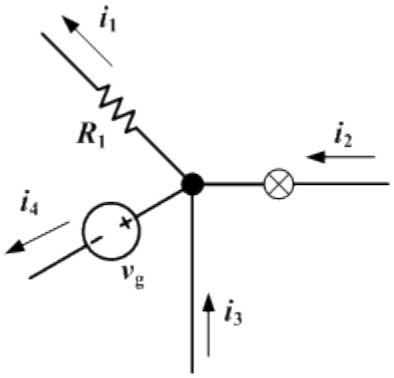
The term "Norton equivalent" is also used in queueing theory for a similar concept. In a reversible queueing system, it is often possible to replace an uninteresting subset of queues by a single (FCFS or PS) queue with an appropriately chosen service rate.

Kirchhoff's circuit laws

Kirchhoff's circuit laws are two equalities that deal with the conservation of charge and energy in electrical circuits, and were first described in 1845 by Gustav Kirchhoff. Widely used in electrical engineering, they are also called Kirchhoff's *rules* or simply Kirchhoff's *laws*.

Both circuit rules can be directly derived from Maxwell's equations, but Kirchhoff preceded Maxwell and instead generalized work by Georg Ohm.

Kirchhoff's current law (KCL)



The current entering any junction is equal to the current leaving that junction. $i_1 + i_4 = i_2 + i_3$

This law is also called **Kirchhoff's point rule**, **Kirchhoff's junction rule** (or nodal rule), and **Kirchhoff's first rule**.

The principle of conservation of electric charge implies that:

> At any node (junction) in an electrical circuit, the sum of currents flowing into that node is equal to the sum of currents flowing out of that node.
> or
> The algebraic sum of currents in a network of conductors meeting at a point is zero. (Assuming that current entering the junction is taken as positive and current leaving the junction is taken as negative).

Recalling that current is a signed (positive or negative) quantity reflecting direction towards or away from a node, this principle can be stated as:

$$\sum_{k=1}^{n} I_k = 0$$

n is the total number of branches with currents flowing towards or away from the node.

This formula is also valid for complex currents:

$$\sum_{k=1}^{n} \tilde{I}_k = 0$$

The law is based on the conservation of charge whereby the charge (measured in coulombs) is the product of the current (in amperes) and the time (which is measured in seconds).

Changing charge density

Physically speaking, the restriction regarding the "capacitor plate" means that Kirchhoff's current law is only valid if the charge density remains constant in the point that it is applied to. This is normally not a problem because of the strength of electrostatic forces: the charge buildup would cause repulsive forces to disperse the charges.

However, a charge build-up *can* occur in a capacitor, where the charge is typically spread over wide parallel plates, with a physical break in the circuit that prevents the positive and negative charge accumulations over the two plates from coming together and cancelling. In this case, the sum of the currents flowing into *one* plate of the capacitor is *not* zero, but rather is equal to the rate of charge accumulation. However, if the displacement current d**D**/dt is included, Kirchhoff's current law once again holds. (This is really only required if one wants to apply the current law to a point on a capacitor *plate*. In circuit analyses, however, the capacitor as a whole is typically treated as a unit, in which case the ordinary current law holds since exactly the current that enters the capacitor on the one side leaves it on the other side.)

More technically, Kirchhoff's current law can be found by taking the divergence of Ampère's law with Maxwell's correction and combining with Gauss's law, yielding:

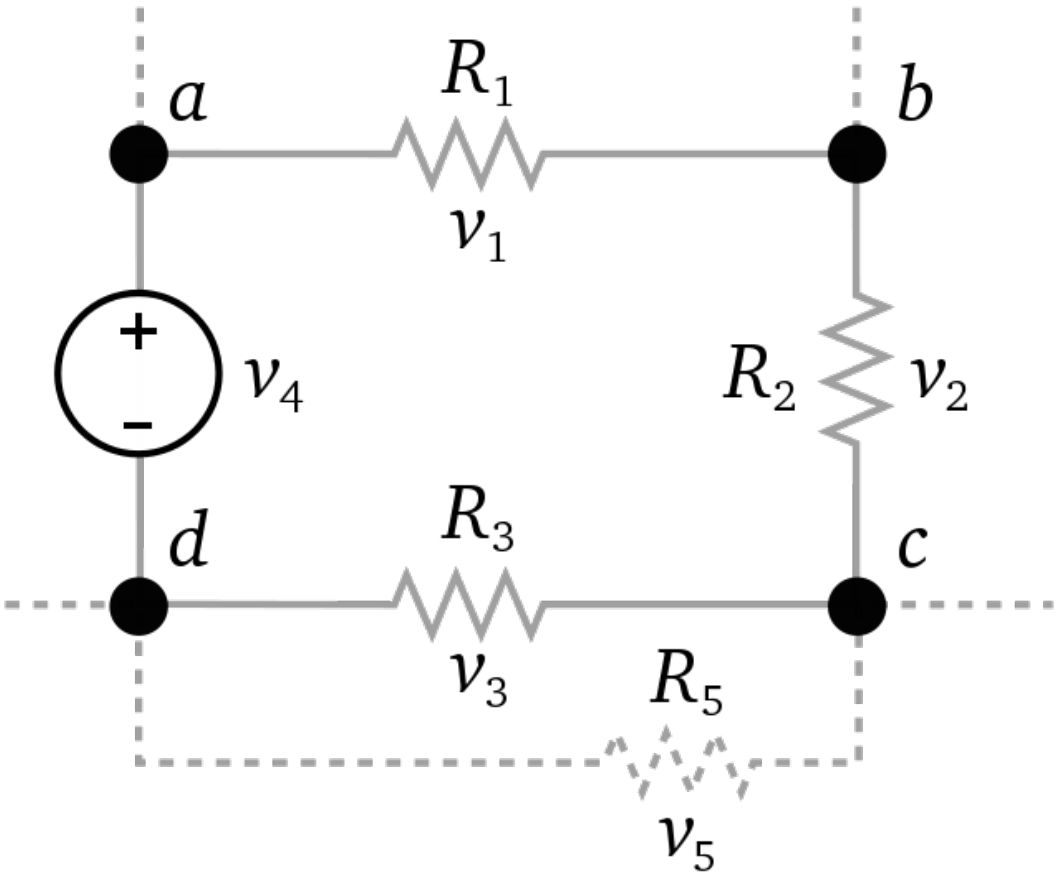
$$\nabla \cdot \mathbf{J} = -\nabla \cdot \frac{\partial \mathbf{D}}{\partial t} = -\frac{\partial \rho}{\partial t}$$

This is simply the charge conservation equation (in integral form, it says that the current flowing out of a closed surface is equal to the rate of loss of charge within the enclosed volume (Divergence theorem)). Kirchhoff's current law is equivalent to the statement that the divergence of the current is zero, true for time-invariant ρ, or always true if the displacement current is included with **J**.

Uses

A matrix version of Kirchhoff's current law is the basis of most circuit simulation software, such as SPICE.

Kirchhoff's voltage law (KVL)



The sum of all the voltages around the loop is equal to zero. $v_1 + v_2 + v_3 + v_4 = 0$

This law is also called **Kirchhoff's second law**, **Kirchhoff's loop (or mesh) rule**, and **Kirchhoff's second rule**.

The principle of conservation of energy implies that

The directed sum of the electrical potential differences (voltage) around any closed circuit is zero.
or
More simply, the sum of the emfs in any closed loop is equivalent to the sum of the potential drops in that loop.
or
The algebraic sum of the products of the resistances of the conductors and the currents in them in a closed loop is equal to the total emf available in that loop.

Similarly to KCL, it can be stated as:

$$\sum_{k=1}^{n} V_k = 0$$

Here, *n* is the total number of voltages measured. The voltages may also be complex:

$$\sum_{k=1}^{n} \tilde{V}_k = 0$$

This law is based on the conservation of "energy given/taken by potential field" (not including energy taken by dissipation). Given a voltage potential, a charge which has completed a closed loop doesn't gain or lose energy as it has gone back to initial potential level.

This law holds true even when resistance (which causes **dissipation** of energy) is present in a circuit. The validity of this law in this case can be understood if one realizes that a charge in fact doesn't go back to its starting point, due to dissipation of energy. A charge will just terminate at the negative terminal, instead of positive terminal. This means all the energy given by the potential difference has been fully consumed by resistance which in turn loses the energy as heat dissipation.

To summarize, Kirchhoff's voltage law has nothing to do with gain or loss of energy by electronic components (resistors, capacitors, etc.). It is a law referring to the potential field generated by voltage sources. In this potential field, regardless of what electronic components are present, the gain or loss in "energy given by the potential field" must be zero when a charge completes a closed loop.

Electric field and electric potential

Kirchhoff's voltage law could be viewed as a consequence of the principle of conservation of energy. Otherwise, it would be possible to build a perpetual motion machine that passed a current in a circle around the circuit.

Considering that electric potential is defined as a line integral over an electric field, Kirchhoff's voltage law can be expressed equivalently as

$$\oint_C \mathbf{E} \cdot d\mathbf{l} = 0,$$

which states that the line integral of the electric field around closed loop C is zero.

In order to return to the more special form, this integral can be "cut in pieces" in order to get the voltage at specific components.

Limitations

This is a simplification of Faraday's law of induction for the special case where there is no fluctuating magnetic field linking the closed loop. Therefore, it practically suffices for explaining circuits containing only resistors and capacitors.

In the presence of a changing magnetic field the electric field is not conservative and it cannot therefore define a pure scalar potential—the line integral of the electric field around the circuit is not zero. This is because energy is being transferred from the magnetic field to the current (or vice versa). In order to "fix" Kirchhoff's voltage law for circuits containing inductors, an effective potential drop, or electromotive force (emf), is associated with each inductance of the circuit, exactly equal to the amount by which the line integral of the electric field is not zero by Faraday's law of induction.