26. Electrical Engineering

Martin Poppe

Some of the most sophisticated systems are the result of close cooperation between mechanical and electrical engineers. This chapter aims to provide mechanical engineers with an introduction to electrical engineering. It explains its basic laws and components, and how they are used to create electrical and electronic systems.

The fundamental laws of electrodynamics are presented using terminology that is also common to mechanical engineers. In this way, the reader will, for example, understand why inductors are made by winding wires around a core and why, in different applications, completely different types of capacitors and resistors are used.

This chapter not only explains how electrical machines and generators work. It also shows how strong machines may become intelligent strong machines. For this purpose, it describes the functioning of transistors and shows how electronic networks can be analyzed. It also explains how semiconductor devices are able to switch and steer high tensions and large currents.

The chapter ends with a glance at one of the most challenging fields common to electrical and mechanical engineers: the storage of power.

26.1 Fundamental Laws

Electrical engineering is the practical application of the laws of electrodynamics for the design of machines and systems. And electrodynamics is that part of physics that analyzes a single property of matter named *charge*. In this section, the fundamental properties of charged objects are discussed. This includes the formation of electric and magnetic fields by charges as well as their influence on the motion of charges. In this context, potentials, tensions, and currents are introduced.

K.-H. Grote, H. Hefazi (Eds.), *Springer Handbook of Mechanical Engineering*, Springer Handbooks,
https://doi.org/10.1007/978-3-030-47035-7_26

26.1.1 Charge

Charge, denoted by the letter Q and measured in coulombs (C), is a signed quantity, as it exists in both positive (+) and negative (−) forms. If charges move from one end of a wire to the other, the first end will lose charge while the other end gains charge. The rate of change is called the *current*

$$I = \frac{\mathrm{d}Q}{\mathrm{d}t} \tag{26.1}$$

and is measured in amperes (A) or coulombs per second. In technical applications, the transport of charges is done by elementary particles called *electrons*, named after the Greek word for amber, as amber charges up when being rubbed against wool. Electrons come in vast numbers, each carrying the same negative amount of charge called the *elementary charge e*.

$$e = 1.6021766 \times 10^{-19}\,\mathrm{C}\,. \tag{26.2}$$

Hence, if there is a current of 1 A from a plug to a lamp, some 6.25×10^{18} electrons per second will move from the lamp to the plug (electrons always move in the direction opposite to the current due to their negative charge). In the presence of such large numbers, rather than looking at individual electrons, it is meaningful to work with the density of charge per volume V

$$\rho = \frac{Q}{V} = \frac{-eN}{V} = -en\,, \tag{26.3}$$

where N is the number of electrons and $n = N/V$ is the density of electrons. It is know from experience that the current at one end of a wire is always the same as that at the other. In other words, no charge is ever lost. This statement, usually called *charge conservation*, is assumed to be a law of Nature.

Conservation of charge leads to a fundamental equation of electrodynamics: the *continuity equation*. This equation relates the loss of charge density in a small volume to the current density through the surface of this volume. It may be deduced from the following mathematical rearrangement: For any constant and stationary volume V, one can write

$$\begin{aligned}
&\frac{\mathrm{d}Q}{\mathrm{d}t} = 0\\
\rightarrow\;& \frac{\mathrm{d}}{\mathrm{d}t}\iiint \rho \mathrm{d}V = \iiint \left[\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \boldsymbol{v})\right] \mathrm{d}V = 0\\
\rightarrow\;& \frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \boldsymbol{v}) = 0\,.
\end{aligned} \tag{26.4}$$

The integrand must be zero because the integral equation is to be zero for an arbitrary volume. Mathematically, the vector $\boldsymbol{v}$ appearing in (26.4) is the speed of the points at which the density is calculated. Physically, this speed may be identified as the speed of the carriers of charge.

The bottom line of (26.4) is the continuity equation. It states that, for any infinitesimal volume, the rate of loss of charge density $-\partial \rho / \partial t$ inside that volume equals the divergence of the charge density multiplied by the speed of the charges $\boldsymbol{v}$. In plain terms, i.e., applied to a finite volume: the loss of charge in a given volume equals the current passing through its boundaries.

The product occurring in the continuity equation is called the *current density*

$$\boldsymbol{J} = \rho \boldsymbol{v}\,. \tag{26.5}$$

Hence, the continuity equation is often written

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \boldsymbol{J} = 0\,. \tag{26.6}$$

This equation describes the effect of charge conservation for very small volumes. The current density is a quantity of great practical importance, as it connects the motion of carriers of charge to the electric current I. The current passing through an oriented area $\boldsymbol{A}$ is given by

$$I = \iint (\rho \boldsymbol{v}) \cdot \mathrm{d}\boldsymbol{A} = \iint \boldsymbol{J} \cdot \mathrm{d}\boldsymbol{A}\,. \tag{26.7}$$

Integrals of the type used in (26.7) are called *fluxes*, being represented by the Greek letter Φ. Mathematically speaking, a current I is the flux of a current density $\boldsymbol{J}$ through the oriented surface $\boldsymbol{A}$. For a mathematician, $I = \Phi_{\boldsymbol{J}}$.

While the current density $\boldsymbol{J}$ is a vector, the current I is not. Although it is said to have a direction, this simply expresses whether the change of charge balance between the two faces of a plane is meant to increase from left to right or from right to left. However, the quantity *current* does not contain any information about the angle between the electrons' movements and the surface they cross.

Application Example: The Speed of Electrons in a Power Supply Cable

If a current passes along a wire, electrons move in a direction perpendicular to its cross section A, as shown in Fig. 26.1. In this case, the equation relating the current and current density, (26.7), simplifies to $I = \rho v A = -envA$, where n is now the density of electrons. The evaluation of this product becomes simple if the following geometrical relation is understood: the number ΔN of electrons passing through the surface $\boldsymbol{A}$ per time

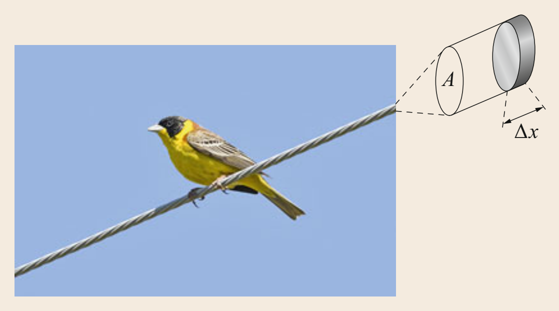

Fig. 26.1 A power cable with cross section A. The volume $V = A\Delta x$ is filled with electrons in a time $\delta t = \Delta x/v$, where v is the speed of the electrons (photo: © bennytrapp/stock.adobe.com)

interval Δt equals the number $\Delta N = n\Delta V$ of electrons in a volume $\Delta V = A\Delta x = Av\Delta t$. Therefore,

$$I = \frac{\Delta Q}{\Delta t} = -\frac{e\Delta N}{\Delta t} = -envA\,.$$

Clearly, n differs from material to material. Aluminum, for example has a chemical valence of 3 and a molar volume of $V_{\text{mol}} = 10.0\,\text{cm}^3/\text{mol}$, so its electron density is

$$\begin{aligned} n &= 3\frac{N_{\text{A}}}{V_{\text{mol}}} = 3\frac{6.02\times 10^{23}\,\text{mol}^{-1}}{10.0(0.01\,\text{m})^3\,\text{mol}^{-1}} \\ &= 1.81\times 10^{29}\,\text{m}^{-3}\,, \end{aligned}$$

with N_{A} being the Avogadro number. Assuming a current of 100 A traversing an aluminum wire with a cross section of $A = 1\,\text{cm}^2$, the velocity of electrons turns out to be surprisingly small:

$$\begin{aligned} v &= \frac{I}{enA} \\ &= \frac{100\,\text{A}}{(1.6\times 10^{-19}\,\text{C})(1.81\times 10^{29}\,\text{m}^{-3})(10^{-4}\,\text{m}^2)} \\ &\approx 3.5\,\frac{\text{mm}}{\text{s}}\,. \end{aligned}$$

The scale of this speed is thus that of snails not cheetahs. And, with the exception of semiconductor devices, this remains true for almost all products of electrical engineering.

26.1.2 Forces and Fields

Charge would be a completely unknown quality of matter if it were not connected to a set of forces to which carriers of charge (and nothing else) are susceptible. The experimental proof of the presence of a charged body is the presence of forces from other charged bodies.

The question of *how a charged particle knows or feels the presence of another charged body* leads to a key concept of electromagnetism, viz. electromagnetic fields. Forces between carriers of charges are, roughly speaking, regarded as the result of a two-step process: every charge modifies the space in its neighborhood. And it it this very modification that makes any other charge feel its presence. No carriers of charge ever notice other charges directly; they only feel the modification of the space.

The modification of space around a charged particle is called a *field of force* because the effect of a force proves its existence. Electromagnetism defines two fields. The *field of the electric force* $\boldsymbol{E}$ is usually simply called the *electric field*. This field is generated by every charged particle. If carriers of charge are moving, i.e., if there is an electric current, a second field, the *field of the magnetic force* $\boldsymbol{B}$, is generated in addition to $\boldsymbol{E}$. For historical reasons, the field of the magnetic force is often called the *flux density*. The origin of this term is an analogy which was thought to be fundamental in the 19th century.

The force due to the electric field is called the *Coulomb force*, the force due to a magnetic field is called the *Lorentz force*, and both together form the *electrodynamic force*. The term "electrodynamic force" (German: *elektrodynamische Kraft*) was introduced by *Einstein* [26.1] when he realized that a description of electromagnetism is valid in all moving frames if, and only if, both the Coulomb force and Lorentz force are taken into account. This is not to be confused with the *electromotive force*, which is not a force but a highly misleading pseudotechnical term that, for the sake of consistency, will not be used in this book. For a carrier of charge Q traversing these fields at a velocity $\boldsymbol{v}$, one has

$$\begin{aligned} &\boldsymbol{F}_{\text{C}} = Q\boldsymbol{E} && \text{Coulomb force}\,, \\ &\boldsymbol{F}_{\text{L}} = Q(\boldsymbol{v}\times\boldsymbol{B}) && \text{Lorentz force}\,, \\ &\boldsymbol{F}_{\text{e}} = Q(\boldsymbol{E}+\boldsymbol{v}\times\boldsymbol{B}) && \text{Electrodynamic force}\,. \end{aligned} \tag{26.8}$$

There is one fundamental difference between the Coulomb force and the Lorentz force: the Coulomb force can change the velocity of charge carriers and thereby their kinetic energy. In contrast, the Lorentz force always acts perpendicular to the direction of motion, leaving the speed and thus kinetic energy unchanged.

The generation of fields by charges is described by a set of four differential equations, called *Maxwell's equations* [26.2],

$$\begin{aligned} &\operatorname{div}\varepsilon_0\boldsymbol{E} = \rho\,, && \operatorname{div}\boldsymbol{B} = 0\,, \\ &\operatorname{rot}\boldsymbol{E} = -\frac{\partial}{\partial t}\boldsymbol{B}\,, && \operatorname{rot}(\mu_0^{-1}\boldsymbol{B}) = \boldsymbol{J} + \frac{\partial}{\partial t}\varepsilon_0\boldsymbol{E}\,. \end{aligned} \tag{26.9}$$

In (26.9), ρ is the charge density, $\boldsymbol{J}$ is the current density, ε_0 is the permittivity of free space, and μ_0 is the permeability of free space. The linearity of these equations guarantees the principle of linear superposition (linear superposition may be used up to photon energies below the rest energy of two electrons [26.3]) to apply to charge densities, current densities, and the fields $\boldsymbol{E}$ and $\boldsymbol{B}$.

Equations (26.8) and (26.9) form the backbone of what is known as *classical electrodynamics*. Its successor, named *quantum electrodynamics*, is the most stringently tested theory in the world [26.4, 5]

The set of equations (26.9) may also be expressed with the help of the *nabla operator* $\nabla = (\partial/\partial x, \partial/\partial y, \partial/\partial z)$ as follows

$$\begin{aligned} &\nabla \cdot \varepsilon_0 \boldsymbol{E} = \rho\,, && \nabla \cdot \boldsymbol{B} = 0\,, \\ &\nabla \times \boldsymbol{E} = -\frac{\partial}{\partial t}\boldsymbol{B}\,, && \nabla \times (\mu_0^{-1}\boldsymbol{B}) = \boldsymbol{J} + \frac{\partial}{\partial t}\varepsilon_0 \boldsymbol{E}\,. \end{aligned} \tag{26.10}$$

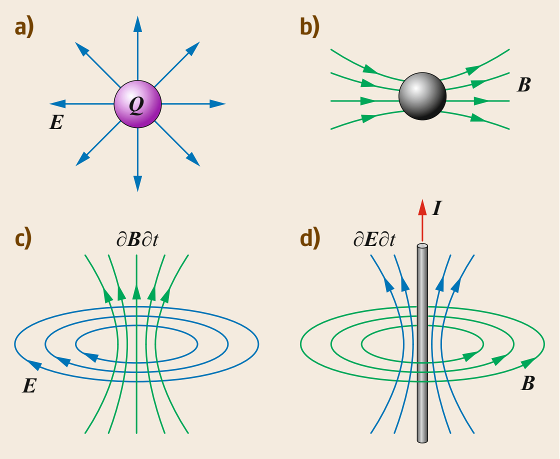

Fig. 26.2a–d Interpretation of Maxwell's equations according to [26.6]. **(a)** Charges are the sources of electric fields. $\nabla \cdot \boldsymbol{E} = \rho/\varepsilon_0$. **(b)** The magnetic field has no sources $\nabla \cdot \boldsymbol{B} = 0$; it is a purely rotational field. **(c)** Rotational electric fields are generated by changing magnetic fields $\nabla \times \boldsymbol{E} = -\partial \boldsymbol{B}/\partial t$. **(d)** Magnetic fields are created by both changing electric fields and moving charges (i.e., currents) $\nabla \times \boldsymbol{B} = \mu_0 \cdot \boldsymbol{J} + \varepsilon_0 \mu_0 \partial \boldsymbol{E}/\partial t$

Figure 26.2 provides a pictorial interpretation of these equations. Figure 26.2a,b defines the source structure of the fields: while the electric field has sources, the magnetic field does not. Figure 26.2c,d describes the rotational structure. The electric field only has a rotational component if changing magnetic fields are present. The magnetic field is entirely rotational and may be generated either by currents or by changing electric fields.

The lines drawn in Fig. 26.2 are called *field lines*. An electrical field line indicates the direction of the force acting on a positively charged particle. The meaning of the magnetic field line is more complicated as the force on a charge is always perpendicular to the line. Later in this chapter, it will be shown that a magnetic dipole, like a compass needle, is subject to a torque that tries to align the dipole with the magnetic field lines.

26.1.3 Integral Formulation of Field Generation

Maxwell's equations may also be formulated in integral form. To do so, the following convention shall be adopted: an integral over a closed path (around an area) or around a finite volume shall be denoted by a $\oint$ symbol with a subscript indicating the path or surface. So $\oint_{\partial V}$ is an integral over the surface enclosing the volume V, and $\oint_{\partial A}$ is a line integral along the enclosure of the surface A.

Applying Gauss's theorem [26.7] to the two equations describing the source structure of the electric field and the magnetic field yields

$$\begin{aligned} &\nabla \cdot \varepsilon_0 \boldsymbol{E} = \rho && \to \oint_{\partial V} \varepsilon_0 \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{A} = Q\,, \\ &\nabla \cdot \boldsymbol{B} = 0 && \to \oint_{\partial V} \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{A} = 0\,. \end{aligned} \tag{26.11}$$

Equations (26.11) are usually referred to as *Gauss's laws for the electric and magnetic field*. The most surprising aspect of these equations is that they connect the property of a volume (ρ is the density of charges within the volume) to an integral that is taken at the surface only. For the integrals in (26.11), the distribution of charges within the volume enclosed by the surface ∂V is completely irrelevant. Surface integrals of the type appearing in (26.11) are called *fluxes of a vector field* (here $\boldsymbol{E}$ or $\boldsymbol{B}$) through a surface $\boldsymbol{A}$. The reason is their formal similarity with physical fluxes such as water through the cross section of a river, grains through a tube, or charges through a wire as in (26.7).

Gauss's law for the electric field is an up-to-date version of a law found by Coulomb in 1784/1785. It states that the force between two point-like carriers of charges Q_1 and Q_2, situated at radii $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$ as shown in Fig. 26.3, is

$$\boldsymbol{F_2} = -\boldsymbol{F_1} = \frac{Q_1 Q_2}{4\pi\varepsilon_0} \cdot \frac{\boldsymbol{r_2} - \boldsymbol{r_1}}{|r_{21}|^3}\,. \tag{26.12}$$

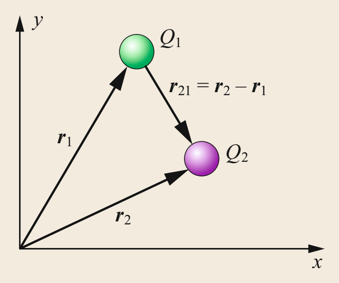

Fig. 26.3 Setup for the definition of the Coulomb force: two oppositely charged balls with centers placed at different locations $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$

being know as *Coulomb's law*. Today, (26.12) is understood to be a consequence of Gauss's law for the electric field, together with the Coulomb force (see also (26.8)). In (26.12), $\boldsymbol{F_2}$ is the force acting on the carrier of Q_2. If the two charges have opposite signs, they attract each other; if not, the force is repelling. The force decreases with the square of the distance between the carriers of charge. The factor $1/(4\pi r_{21}^2)$ can be traced back to Gauss's law, as it is the inverse of the surface of a sphere or radius r_{21}.

If the forces on a positive charge are drawn as little arrows, one finds that these line up like strings ending on the charges. These lines are called *field lines*. Figure 26.4 shows the lines between a positive and a negative charge, and Fig. 26.5 the lines associated with two positive charges.

The *law of induction*, also known as the *law of Faraday and Henry* [26.8], is to some extent a consequence of the rotational part of the electric field. It is crucial for the functioning of a very wide range of electrical devices, as engines, generators, transformers, and solenoid antennas cannot operate without it:

$$\begin{aligned} U_{\text{ind}} &= \oint\limits_{L} \boldsymbol{E}\cdot \mathrm{d}\boldsymbol{\ell} = -\frac{\mathrm{d}}{\mathrm{d}t}\int\limits_{A} \boldsymbol{B}\cdot \mathrm{d}\boldsymbol{A} \\ &= -\frac{\mathrm{d}\Phi_{\mathrm{B}}}{\mathrm{d}t} \,. \end{aligned} \tag{26.13}$$

The translation of this formula into plain English is: for a closed loop L of a conductor, placed in a magnetic field $\boldsymbol{B}$, the induced tension is given by the change of the magnetic flux Φ_{B} through any area $\boldsymbol{A}$ enclosed by this loop.

One reason for this wide variety of applications is the fact that the single formula (26.13) contains two completely different effects, which emerge if the derivative of the flux is worked out in detail. The terms remaining for the magnetic field are

$$\frac{\mathrm{d}}{\mathrm{d}t}\int\limits_{A} \boldsymbol{B}\cdot\mathrm{d}\boldsymbol{A} = \int \left[\frac{\partial \boldsymbol{B}}{\partial t} - \nabla\times(\boldsymbol{v}\times\boldsymbol{B})\right]\mathrm{d}\boldsymbol{A}\,. \tag{26.14}$$

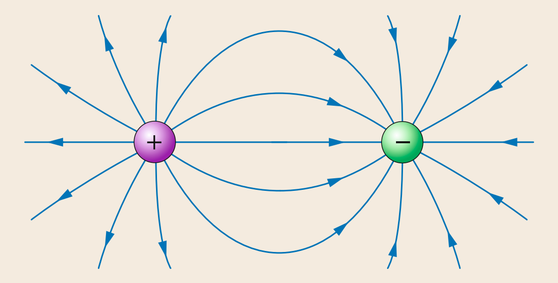

Fig. 26.4 Field lines between a positive and a negative charge

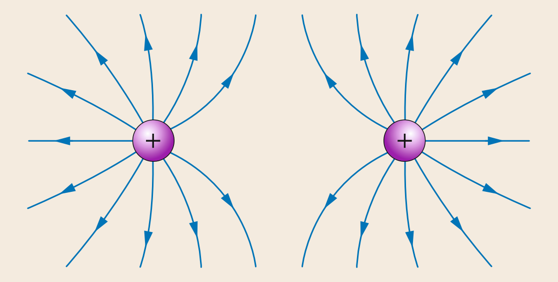

Fig. 26.5 Field lines associated with two positive charges

The mathematical origin of the velocity $\boldsymbol{v}$ is the contribution of the change of area to the total derivative. Its meaning is the speed of those points for which the flux elements are calculated. The physical meaning of (26.14) can be explored by translating the line integral in (26.13) into a surface integral $\oint_{\partial A} \boldsymbol{E}\cdot\mathrm{d}\boldsymbol{\ell} = \int(\nabla\times\boldsymbol{E})\mathrm{d}\boldsymbol{A}$. So, one may conclude by comparison

$$\nabla\times\boldsymbol{E} = -\frac{\partial\boldsymbol{B}}{\partial t} + \nabla\times(\boldsymbol{v}\times\boldsymbol{B})\,. \tag{26.15}$$

The outer terms may be interpreted as new descriptions of the Lorentz force:

$$\begin{aligned} \boldsymbol{F} &= Q\boldsymbol{v}\times\boldsymbol{B} \\ \rightarrow \boldsymbol{E} &= \boldsymbol{v}\times\boldsymbol{B} \\ \rightarrow \nabla\times\boldsymbol{E} &= \nabla\times(\boldsymbol{v}\times\boldsymbol{B})\,. \end{aligned} \tag{26.16}$$

Equation (26.16) should, however, not be misunderstood as describing the appearance of an electric field. It simply means that, for a charge carrier traversing a magnetic field, there will be a force accelerating it in a direction perpendicular to its velocity and perpendicular to the magnetic field, and that the strength of this force is just the same as for the Lorentz force. In view of the theory of special relativity, this is more than a coincidence. According to Einstein, which fraction of an electromagnetic field is magnetic and which is electric varies with velocity. A force attributed to a magnetic field in one frame of reference may be entirely attributed to an electric field in another frame.

A practical result of (26.16) is that all forces appearing in electrical generators with permanent magnets can be calculated either using the Lorentz force on the charge carriers or using (26.13). The results must be identical.

The two terms on the left-hand side of (26.15) are crucial for the functioning of transformers. An oscillating magnetic field can be used to generate a tension in a closed conducting loop. These terms already appeared as one of Maxwell's equations in (26.10).

The fourth macroscopic law to be inspected is the law of *Ampère* and *Maxwell*:

$$\oint_{\partial A} (\mu_0^{-1}\boldsymbol{B}) \cdot \mathrm{d}\boldsymbol{\ell} = I + \int \frac{\partial(\varepsilon_0 \boldsymbol{E})}{\partial t} \mathrm{d}\boldsymbol{A} \, . \tag{26.17}$$

It describes the generation of a magnetic field $\boldsymbol{B}$ by both a current I and a time-dependent electric field $\boldsymbol{E}$, as shown in Fig. 26.2. In (26.17), $\mathrm{d}\boldsymbol{\ell}$ is a small element of any line enclosing a current I, and $\mathrm{d}\boldsymbol{A}$ is a small element of any surface that is completely encircled by this line. Equation (26.17) can be identified with one of the known differential equations by converting all terms into surface integrals:

$$\int \left[\nabla \times (\mu_0^{-1}\boldsymbol{B})\right] \mathrm{d}\boldsymbol{A} = \int \left(\boldsymbol{J} + \frac{\partial(\varepsilon_0 \boldsymbol{E})}{\partial t}\right) \mathrm{d}\boldsymbol{A} \, . \tag{26.18}$$

Since this equation is to be valid for an arbitrary surface $\boldsymbol{A}$, the integrands must be identical

$$\nabla \times (\mu_0^{-1}\boldsymbol{B}) = \boldsymbol{J} + \frac{\partial(\varepsilon_0 \boldsymbol{E})}{\partial t} \, , \tag{26.19}$$

which is identical to the fourth equation in (26.10).

In the limit $\partial(\varepsilon_0 \boldsymbol{E})/\partial t = 0$, the above is often called *Ampère's law*. However, this naming convention may easily lead to misinterpretations, as the *law* is only valid in the absence of time-dependent electric fields, and if a magnetic field varies with no current present, the Ampère–Maxwell law requires the presence of a time-dependent electric field. So, what is referred to as Ampère's law is actually only an approximation that applies to the static case.

Application Example: Whistling Capacitors

Gauss's law for the electric field may be used to determine the forces between the electrodes of a capacitor. If the material between the plates is compressible, the plates will change their separation at the same frequency as the charging and discharging of the plates. If this frequency is below 10 kHz, such vibration can be heard, and the capacitor can thus be identified as being of poor quality.

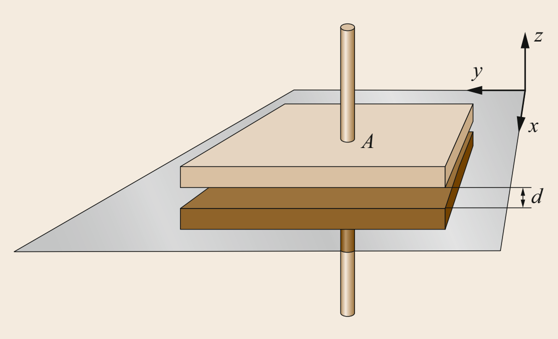

Fig. 26.6 Two electrodes forming a capacitor. All the electric field crosses the plane indicated between the electrodes, mostly traversing the area between the electrodes

Figure 26.6 shows the principal setup of a capacitor. It consists of two rather flat electrodes of area A, separated by a distance d. If a charge Q_1 is placed on the top electrode and a charge $Q_2 = -Q_1$ on the bottom, the electric field will be concentrated almost entirely in the region between the electrodes. For any surface enclosing the top electrode, the only significant contribution to the integral in Gauss's law will be the plane shown in gray in Fig. 26.6. The plane can be either close to one of the electrodes, or in the middle. One gets

$$Q_1 = \oint_{\partial V} \varepsilon_0 \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{A} \approx \varepsilon_0 \boldsymbol{E} \cdot \boldsymbol{A} = \varepsilon_0 E_z A \, ,$$

assuming an oriented area with $\boldsymbol{A} = (0, 0, A)$ and $E_z = -|\boldsymbol{E}|$ because the top plate is positively charged. The magnitude of the force $\boldsymbol{F}_2$ acting on the bottom plane is then

$$\boldsymbol{F}_2 = Q_2 \boldsymbol{E} = \varepsilon_0 Q_1 Q_2 \boldsymbol{A} = \varepsilon_0 Q^2 \boldsymbol{A} \quad (= -\boldsymbol{F}_1) \, .$$

If the electrodes carry oppositely signed charges, they will always attract each other, irrespective of which of the two electrodes is positively charged. Consequently, if the electrodes are charged with an alternating current of a certain frequency, the frequency at which the electrodes vibrate will be twice the frequency of the current. Actually, something similar can be heard close to old electrical locomotives. In that case, it is the oscillation of the magnetic field of the transformers that results in an audible mechanical frequency that is twice the frequency of the locomotive's power supply.

Table 26.1 Correspondence between electrostatics and gravitation

Electrostatics	Gravitation
$\boldsymbol{F} = Q\boldsymbol{E}$	$\boldsymbol{F} = m(0, 0, -g)$
Charge	Mass
Field strength $\lvert\boldsymbol{E}\rvert$	Acceleration g
Potential V	Height h
Tension U	Difference in height Δh

26.1.4 Potentials and Tensions

A charged particle in an electric field will gain potential energy when moving against the force of the field and loose potential energy when moving with the field. In other words: work is to be done if a charged body is to be moved along a path C through the field:

$$W = \int_{\mathrm{C}} -\boldsymbol{F} \cdot \mathrm{d}\boldsymbol{s} = -Q \int_{\mathrm{C}} \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{s} \quad \text{(always)} . \qquad (26.20)$$

If the electric field is static, it is bound to have no rotational component. Then, the dependence on the path vanishes and the value of the integral

$$\begin{aligned} W &= \int_a^b -\boldsymbol{F} \cdot \mathrm{d}\boldsymbol{s} \\ &= -Q \int_a^b \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{s} \quad \text{(static field)} \end{aligned} \qquad (26.21)$$

depends only on the starting point a and end point b. The minus sign in (26.21) indicates that energy is gained if the movement is *against* the direction of force.

For a given field $\boldsymbol{E}$, the work to be done only depends on the charge Q and the location of the two points a and b. Therefore, in the presence of a static electric field, one can assign a new property to each point in space called the *electric potential*, denoted by V. If one knows the potential at all points in space, one can easily calculate the energy gained by moving from one point to any other: it is the potential difference, multiplied by the charge, or just $V(b) - V(a) = W/Q$. The relation between $\boldsymbol{E}$ and V reveals a major practical use in its differential form. The equation

$$\boldsymbol{E} = -\nabla \cdot V \quad \text{(static field)} \qquad (26.22)$$

shows that, if the (nondirectional) potential is known at every point in space, differentiation may be used to find both the strength and direction of the electric field.

The electric potential is only defined up to a constant, just as any indefinite integral. Physically, this freedom of choice simple means that one is free to choose an arbitrary starting point for the calculation of a potential. This freedom is lost when the difference of two potentials is calculated, as the constant of integration cancels out. The potential difference between two points is such an important quantity that it is given a name of its own, the *tension*

$$\begin{aligned} U_{ab} &= V(a) - V(b) \\ &= -\int_a^b \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{s} \quad \text{(static field)} \end{aligned} \qquad (26.23)$$

between two points.

The potential of a static electric field has a one-to-one correspondence with the height in the gravitational field of the Earth. Lines of equal potential correspond to lines of equal altitude in maps of mountains, as shown in Table 26.1.

As soon as the fields vary with time, the electric and magnetic fields are bound to appear together. In this case, it is helpful to define a *magnetic potential* $\boldsymbol{A}$, from which the magnetic field can be derived according to

$$\boldsymbol{B} = \nabla \times \boldsymbol{A} . \qquad (26.24)$$

The electric field can now be written as

$$\boldsymbol{E} = -\nabla \cdot V + \frac{\partial \boldsymbol{A}}{\partial t} \quad \text{(always)} , \qquad (26.25)$$

in a form that reveals the difference between electrostatics and electrodynamics: The second term in (26.25) accounts for the generation of a rotational electric field according to the Faraday–Henry law. It vanishes in the static case, leaving (26.22) only.

The major practical implication of (26.25) is that the idea of a tension between two points has to be dropped in the presence of varying magnetic fields. In fact, the stronger the rotational component of the electric field, the more important the path between two points becomes. Electric generators extensively use this fact by inducing high tensions (colloquially voltages) into wires that form a long path C in a time dependent magnetic field.

The instruction to “please switch off all electronic devices until the plane has reached its parking position” may be understood as a consequence of (26.25), as the functioning of electronic circuit boards relies on $\partial\boldsymbol{A}/\partial t$ being negligible.

26.2 Capacitors, Resistors, and Inductors

The functioning of the majority of basic circuit elements relies on the interaction between matter and electromagnetic fields. In this section, the influence of matter on electromagnetic fields will be analyzed and optimal geometries determined for resistors, capacitors, and inductors. The behavior of these circuit elements will be determined for direct-current (DC) and alternating-current (AC) networks. Finally, deviations from ideal circuit element properties are discussed.

26.2.1 Matter and Fields

Atoms and molecules consist of negatively charged electrons and positively charged nuclei. Therefore, even electrically neutral bodies will react to the presence of electric and magnetic fields. Figure 26.7 shows how an external field $\boldsymbol{E}_{\text{free}}$ applied to a neutral molecule will lead to the appearance of a new field $\boldsymbol{E}_{\text{bound}}$. Electrons that may move long distances (those that are *free* to leave molecular or atomic orbits) are held responsible for $\boldsymbol{E}_{\text{free}}$. In contrast, electrons and nuclei responsible for $\boldsymbol{E}_{\text{bound}}$ cannot move further than distances of about one nanometer, since they are *bound* to atoms. Both sets of carriers of charge are well distinct. The criterion is the value for the binding energy of electrons: those that are tightly bound will not be able to leave their orbit, while the others can. Usually, there is an energy gap between bound and free electrons. In mathematical terms, the bound and free electrons form distinct, nonoverlapping sets. Magnetization suggests a similar distinction for free and bound currents. In combination with a key characteristic of Maxwell's equations, viz. the principle of linear superposition, these distinctions form the basis for a practically usable formulation of fields in matter.

The principle of linear superposition states that the field resulting from two different sources may be calculated as their vector sum. Figure 26.8 shows the application of this principle to free and bound charges.

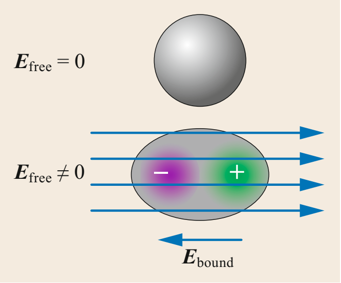
$\boldsymbol{E}_{\text{free}} = 0$
$\boldsymbol{E}_{\text{free}} \neq 0$
−
+
$\boldsymbol{E}_{\text{bound}}$

Fig. 26.7 Sketch of polarization: an external field $\boldsymbol{E}_{\text{free}}$ leads to a separation of charges bound within an atom or molecule. These charges produce a field $\boldsymbol{E}_{\text{bound}}$

The mathematical formulation is then

$$\begin{aligned} \rho &= \rho_{\text{free}} + \rho_{\text{bound}} \quad \text{and} \quad \boldsymbol{E} = \boldsymbol{E}_{\text{free}} + \boldsymbol{E}_{\text{bound}} \,, \\ \boldsymbol{J} &= \boldsymbol{J}_{\text{free}} + \boldsymbol{J}_{\text{bound}} \quad \text{and} \quad \boldsymbol{B} = \boldsymbol{B}_{\text{free}} + \boldsymbol{B}_{\text{bound}} \,. \end{aligned} \tag{26.26}$$

Maxwell's equations will hold for either subset, for example

$$\begin{aligned} \nabla \cdot \varepsilon_0 \boldsymbol{E}_{\text{free}} &= \rho_{\text{free}} \,, \\ \nabla \times (\mu_0^{-1} \boldsymbol{B}_{\text{free}}) &= \boldsymbol{J}_{\text{free}} + \frac{\partial}{\partial t} \varepsilon_0 \boldsymbol{E}_{\text{free}} \,. \end{aligned} \tag{26.27}$$

Inserting (26.26) into (26.27) yields

$$\begin{aligned} \nabla \cdot \varepsilon_0 (\boldsymbol{E} - \boldsymbol{E}_{\text{bound}}) &= \rho_{\text{free}} \,, \\ \nabla \times [\mu_0^{-1} (\boldsymbol{B} - \boldsymbol{B}_{\text{bound}})] &= \boldsymbol{J}_{\text{free}} + \frac{\partial}{\partial t} \varepsilon_0 \boldsymbol{E}_{\text{free}} \,. \end{aligned} \tag{26.28}$$

Equations (26.28) are usually presented in a different notation and are widely known as *Maxwell's equations in matter* in the form

$$\begin{aligned} \nabla \cdot (\varepsilon_0 \boldsymbol{E} + \boldsymbol{P}) &= \nabla \cdot \boldsymbol{D} = \rho_{\text{free}} \,, \\ \nabla \times (\mu_0^{-1} \boldsymbol{B} - \boldsymbol{M}) &= \nabla \times \boldsymbol{H} = \boldsymbol{J}_{\text{free}} + \frac{\partial}{\partial t} \boldsymbol{D} \,. \end{aligned} \tag{26.29}$$

This formulation has historical roots. In the 19th century, the entire world was assumed to be full of a strange substance called *ether* (not referring to the gas C_2H_6O). Electromagnetic waves would be transported by this substance, and it was assumed that charges would *displace* this substance. $\boldsymbol{D}$'s full name was the displacement current of the ether, while $\boldsymbol{H}$ (a little later) was assumed to be a stimulation of the ether. The name *magnetic field* for $\boldsymbol{H}$ is even older. It originates in symmetries which are now known to apply only to static

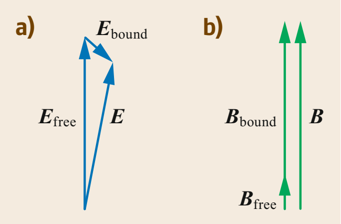
a)
$\boldsymbol{E}_{\text{bound}}$
$\boldsymbol{E}_{\text{free}}$
$\boldsymbol{E}$
b)
$\boldsymbol{B}_{\text{bound}}$
$\boldsymbol{B}$
$\boldsymbol{B}_{\text{free}}$

Fig. 26.8a,b In the presence of matter, the fields $\boldsymbol{E}$ and $\boldsymbol{B}$ are calculated as the vector sum of two contributions (taken from [26.6]) In (**a**), the case of an electrically anisotropic material is shown, for (**b**), a ferromagnetic substance is assumed

Table 26.2 Physical interpretation of traditional field quantities [26.6]

Traditional symbol	Meaning	Cause	Cause symbol
$\boldsymbol{H}$	$\mu_0^{-1}\boldsymbol{B}_{\text{free}}$	Free currents	$\boldsymbol{J}_{\text{free}}$
$\boldsymbol{D}$	$\varepsilon_0\boldsymbol{E}_{\text{free}}$	Free charges	ρ_{free}
$\boldsymbol{M}$	$\mu_0^{-1}\boldsymbol{B}_{\text{bound}}$	Bound currents	$\boldsymbol{J}_{\text{bound}}$
$\boldsymbol{P}$	$-\varepsilon_0\boldsymbol{E}_{\text{bound}}$	Bound charges	ρ_{bound}

fields. Over the last 100 years, the interpretation of $\boldsymbol{H}$ and $\boldsymbol{D}$ has gradually developed from *fields* [26.9], via *not so fundamental fields* [26.10] to *contributions to the fields* $\boldsymbol{B}$ *and* $\boldsymbol{E}$ [26.6].

Comparing the last two sets of equations leads to a physical interpretation of the traditional field quantities. These are summarized in Table 26.2. This table also explains why the equations found in the 19th century are still valid: there is a one-to-one correspondence between the quantities defined then and those which result from the present understanding of atoms and molecules.

For an understanding of electromagnetism in the context of current technological developments, it is important to realize that, when manufacturing techniques allow the creation of structures with the size of a single atom, $\boldsymbol{H}, \boldsymbol{D}, \boldsymbol{M}$, and $\boldsymbol{P}$ are no longer applicable. At such distances, the distinction between free and bound electrons becomes ambiguous. And if nuclear distances are probed, the distinction becomes completely irrelevant as the energies involved are much larger than atomic binding energies.

In Fig. 26.8, the modification of $\boldsymbol{E}_{\text{free}}$ to give the measurable $\boldsymbol{E}$ was achieved by adding field vectors from bound charges. But there is also a very popular alternative: the influence of matter may be formulated as a mapping of the fields from free charges onto the measurable ones. In the most general case, this mapping will include a rotation, and an elongation or shortening. It is written in the form

$$\begin{aligned} \boldsymbol{B} &= \mu_{\text{r}}\boldsymbol{B}_{\text{free}} \,, \\ \boldsymbol{E} &= \varepsilon_{\text{r}}^{-1}\boldsymbol{E}_{\text{free}} \,, \end{aligned} \tag{26.30}$$

where μ_{r} is called the *relative permeability* and ε_{r} is called the *relative permittivity*. In general, μ_{r} and ε_{r} are tensors depending on the field strength. In most cases, however, they are just numbers. A large value for μ_{r} means a *strengthening* of the magnetic field, while a large value for ε_{r} means a *weakening* of the electric field by matter.

Equation (26.30) can be used to derive a simple rule for the incorporation of matter into Maxwell's equations:

$$\begin{array}{lcl} \text{always} & & \text{in matter} \\ \rho & \rightarrow & \rho_{\text{free}} \\ \boldsymbol{J} & \rightarrow & \boldsymbol{J}_{\text{free}} \\ (\varepsilon_0\boldsymbol{E}) & \rightarrow & (\varepsilon_0\varepsilon_{\text{r}}\boldsymbol{E}) \\ (\mu_0^{-1}\boldsymbol{B}) & \rightarrow & [(\mu_0\mu_{\text{r}})^{-1}\boldsymbol{B}] \,. \end{array} \tag{26.31}$$

This substitution rule relates the measurable free charges and currents to the measurable electromagnetic fields. A consequent use of this rule allows a formulation of the whole of classical electromagnetism without ever mentioning the quantities $\boldsymbol{H}$ and $\boldsymbol{D}$, as done in [26.11]. When using (26.31) in integral equations, it is important to keep the quantities in brackets together, as shown.

In the presence of matter, Gauss's law for the electric field can now be written as

$$\oint_{\partial V} \varepsilon_0\varepsilon_{\text{r}}\boldsymbol{E}\cdot \text{d}\boldsymbol{A} = Q \,, \tag{26.32}$$

while the law of Ampère and Maxwell in (26.17) may be written as

$$\oint_{\partial A} \left(\mu_0^{-1}\mu_{\text{r}}^{-1}\boldsymbol{B}\right)\cdot \text{d}\boldsymbol{\ell} = I_{\text{free}} + \int \frac{\partial(\varepsilon_0\varepsilon_{\text{r}}\boldsymbol{E})}{\partial t}\text{d}\boldsymbol{A} \,. \tag{26.33}$$

Gauss's law for the magnetic field and the law of Faraday and Henry remain unchanged in the presence of matter.

Currents flow if electrons are moving. If this happens within matter, the electrons will randomly collide with atoms, thus giving some of their kinetic energy to the body they traverse and being slowed down at the same time. In the case of the presence of an external electric field, the electrons will be accelerated between two collisions. It turns out that this ongoing sequence of collisions and acceleration leads to an average velocity of the electrons $\boldsymbol{v}_{\text{e}}$ that is proportional to the field strength. Therefore, the current density $\boldsymbol{J}$ due to electrons which have a density n_{e} is proportional to the field strength applied. The corresponding formula

$$\boldsymbol{J} = -n_{\text{e}}e\boldsymbol{v}_{\text{e}} = \sigma\boldsymbol{E} \tag{26.34}$$

is know as *Drude's law* [26.12]. The factor of proportionality σ is known as the *specific conductance*. Its inverse $\rho = 1/\sigma$ is called the *specific resistance*; due to the limited number of characters available, it is represented by the same letter as the charge density.

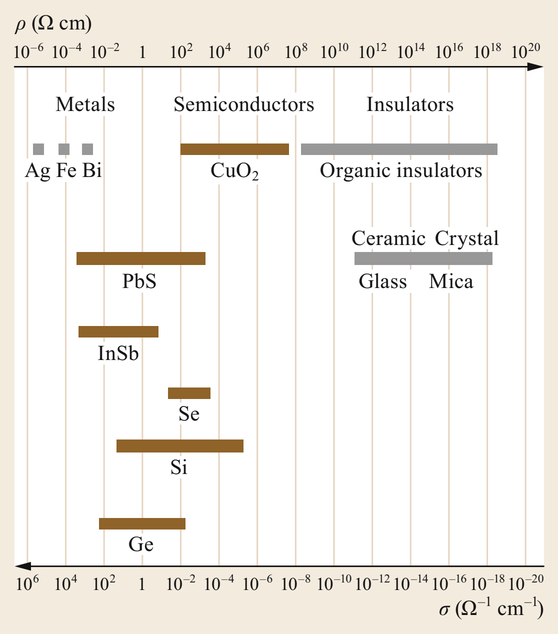

Fig. 26.9 Specific resistance of various materials

Materials are classified as *conductors* (usually metals), *semiconductors*, or *insulators*, depending on their specific resistance. Examples of classes of materials can be found in Fig. 26.9.

26.2.2 Resistors

Sending a current through a resistor is a bit like pulling a brake: the current becomes smaller the more resistors there are, and the resistors themselves become warm if current passes. The main use of resistors is to limit currents and to generate heat.

A resistor is called an *ohmic resistor* if the current is proportional to the voltage applied. Such behavior is a consequence of Drude's law. This will be shown for the example of a wire, as indicated in Fig. 26.10. An electric field $\boldsymbol{E}$ applied to a wire of length L and cross section $\boldsymbol{A}$ will provoke a current density $\boldsymbol{J} = \sigma \boldsymbol{E}$. As the tension between the two ends of the wire is $U = |\boldsymbol{E}|L$, one finds that the current

$$I = \boldsymbol{J} \cdot \boldsymbol{A} = \sigma \frac{A}{L} U \tag{26.35}$$

is indeed proportional to the tension U applied. The proportionality factor is called the *conductance* G, and its inverse has the name *resistance*

$$R = \frac{L}{\sigma A} = \frac{1}{G} \, . \tag{26.36}$$

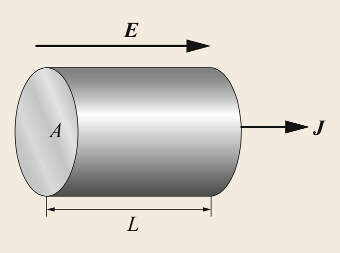

Fig. 26.10 A piece of conducting material of length L and cross section A exposed to an electric field $\boldsymbol{E}$ will have a current density $\boldsymbol{J}$ given by Drude's law

Equation (26.35) can then be written as

$$U = RI \quad \text{or} \quad I = GU \, , \tag{26.37}$$

and in this form, it is known as *Ohm's law* [26.13].

Ohm's law is useful to describe circuit elements with fixed geometry, such as resistors *off the shelf*. In contrast, Drude's law may be applied to complicated geometries. Therefore, (26.34) is often referred to as the *differential form of Ohm's law*.

The resistance can be expressed in terms of the *electron mobility* $\mu_n = v/E$, which relates the field strength to the velocity v and density n of the electrons. The equation

$$R = \frac{1}{AEn\mu_n} \tag{26.38}$$

shows very clearly that, the more electrons there are and the faster they move, the better the conductance becomes.

Resistors come in a variety of shapes, materials, and sizes. The size is determined by the heat produced: the more heat is produced, the larger a resistor needs to be in order to ensure that it will neither melt nor evaporate. The heating power generated for a constant field $\boldsymbol{E}$ parallel to the length L is, according to (26.20),

$$P_{\text{heat}} = \frac{\mathrm{d}W}{\mathrm{d}t} = \frac{-\mathrm{d}Q}{\mathrm{d}t} EL = -UI \, . \tag{26.39}$$

The minus sign indicates that the heating power is *generated*. Clearly, this power must come from the circuit, i.e.,

$$P_{\text{circuit}} = UI \, , \tag{26.40}$$

which means that a resistor transforms electric energy into heat energy. Unless otherwise mentioned, P is used for P_{circuit}.

Table 26.3 Temperature coefficients for selected materials [26.14]

Material	Temperature coefficient α ($°C^{-1}$)
Gold	0.0034
Copper, aluminum, lead	0.0039
Manganin	0.000002
Carbon (amorphous)	−0.0005
Silicon	≈ -0.075
Germanium	≈ -0.048

This is a general rule: a positive value for the power means that the circuit loses electrical energy to other forms of energy (for example, in a resistor), whereas a negative value indicates that the circuit gains energy (for example, from a battery).

Using (26.40), it can be deduced that the power of a resistor rises with both the square of the tension and the current

$$P = P_{\text{circuit}} = I^2 R = \frac{U^2}{R}\ . \tag{26.41}$$

The material used depends on price as well as the required accuracy and temperature independence of the resistance. For most conductors, the value of R increases with temperature. This may be understood as a consequence of the fact that the chance of an electron colliding with an atom increases as the atom moves (vibrates) faster when the temperature θ increases. For most purposes, it is accurate enough to describe this effect in a linear manner, starting at room temperature:

$$R(\theta) \approx R(25\,°\text{C})\,[1 + \alpha(\theta - 25\,°\text{C})]\ . \tag{26.42}$$

The factor α is called the *temperature coefficient*. It describes how the resistance increases with temperature.

Table 26.3 shows that, over a range of $\Delta\theta = 100\,°\text{C}$, the resistance of a metal wire may increase by more than one-third. It also shows that the conductivity of semiconductors increases as the temperature does.

Carbon composite resistors are the most common type, as they are quite robust and easy to manufacture. Figure 26.11 shows a cross section. The value for the resistance is determined by the fraction of carbon in the mixture and by the geometry. The colored rings indicate the value for R as well as the accuracy of this value. Carbon composite resistors have a very small parasitic inductance, thus making them ideally suited for high-frequency applications. Their main drawback is a lack of accuracy in the manufacturing process.

In contrast, *thin-film resistors* may be produced with high precision, often better than ±1%. They are produced by depositing a thin layer of metal onto a nonconducting substrate (Fig. 26.12). After deposition, a helix is engraved into the film. This increases both the resistance and parasitic inductance (see below).

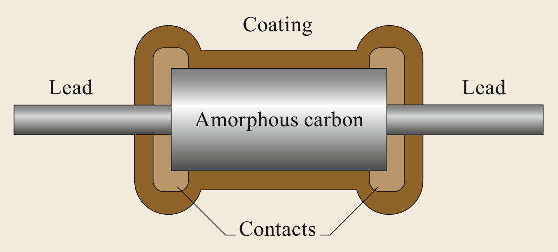

Fig. 26.11 Cross section through a carbon composite resistor

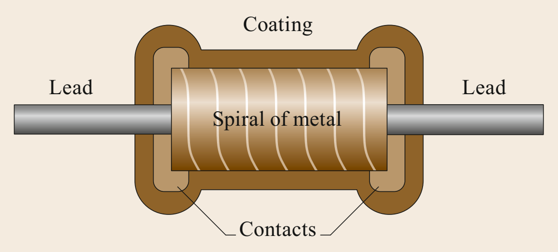

Fig. 26.12 Cross section through a thin-film resistor

For high-power, low-frequency applications, resistors can be made by winding wires to form a helix. These resistors are particularly suitable if low values ($< 100\,\Omega$) for the resistance are desired.

Some applications require resistors with a negative temperature coefficient. These can be made from impure semiconductor materials, as seen in Table 26.3.

26.2.3 Capacitors

Capacitors are used to store energy using an electric field. This is achieved by charge separation. A capacitor consists of two electrodes with a thin layer of a nonconducting substance called a dielectric between them, as shown in Fig. 26.6. A tension U between the two electrodes having an overlap area A and which are separated by a dielectric of thickness d with a relative permittivity ε_r makes one of them acquire a charge

$$Q = \varepsilon_0 \varepsilon_r \frac{A}{d} U\ . \tag{26.43}$$

The other electrode will have the same amount of charge, but with the opposite sign. This shows that an entire capacitor *separates charge*, rather than storing it.

The factors between U and Q in (26.43) are usually combined into a single factor called the *capacitance*

$$C = \frac{Q}{U}\,, \tag{26.44}$$

which for the setup shown in Fig. 26.6 means $C = \varepsilon_0 \varepsilon_r A/d$. Computing the time derivative of (26.44) yields

$$I = C\frac{\mathrm{d}U}{\mathrm{d}t}\,, \tag{26.45}$$

which reveals a typical feature of capacitors in circuits: the faster the tension changes, the more current will flow through a capacitor, while direct currents are not conducted at all.

The amount of energy stored can be calculated by computing how much work is needed to charge a capacitor. The result is

$$W = \frac{1}{2}CU^2\,, \tag{26.46}$$

also demonstrating that the withstand voltage $U_{\max}$ of a capacitor is an important factor with regard to the maximum energy that can be stored. The electric field and the dielectric material can be identified as the place where the energy is stored. Comparing (26.46) with the strength $E = U/d$ of the electric field yields an energy density of

$$w = \frac{W}{V} = \frac{W}{Ad} = \frac{\varepsilon_0 \varepsilon_r}{2}E^2\,, \tag{26.47}$$

which turns out to be a shorthand notation for two energy contributions. For $\varepsilon_r = 1$, the entire energy is stored in the field. For larger values, the elastic deformation of the dielectric molecules uses up the additional energy; for example, with $\varepsilon_r = 3$, the molecules contain twice as much energy as the field itself.

Capacitors may be connected in series or in parallel. In both cases, they can be treated as a single capacitor having a capacitance C_{total}. For a series of two capacitors with capacitances C_{s1} and C_{s2}, one gets

$$\frac{1}{C_{\text{total}}} = \frac{1}{C_{s1}} + \frac{1}{C_{s2}} \quad \text{(series)}\,, \tag{26.48}$$

while a connection of C_{p1} and C_{p2} in parallel gives

$$C_{\text{total}} = C_{p1} + C_{p2} \quad \text{(parallel)}\,. \tag{26.49}$$

The simplest form of a capacitor, the *film capacitor*, is achieved by rolling up two metal foil electrodes with a nonconducting substance placed in between. As the film between the electrodes is often made of plastic, these are also called *plastic film capacitors*, *film dielectric capacitors*, or *polymer film capacitors*. Such dielectric substances have relative permittivities ranging from $\varepsilon_r = 2.2$ for polypropylene (PP) to $\varepsilon_r = 3$ for polyethylene terephthalate (PET), i.e., polyester. In modern production lines, a plastic foil is given a metal coating by means of vacuum deposition and then wound up [26.15]. As both electrodes of such a capacitor form a helix, they will also produce a rather large parasitic inductance. For high-frequency applications, stacks of metal film layers are preferred despite their higher production costs.

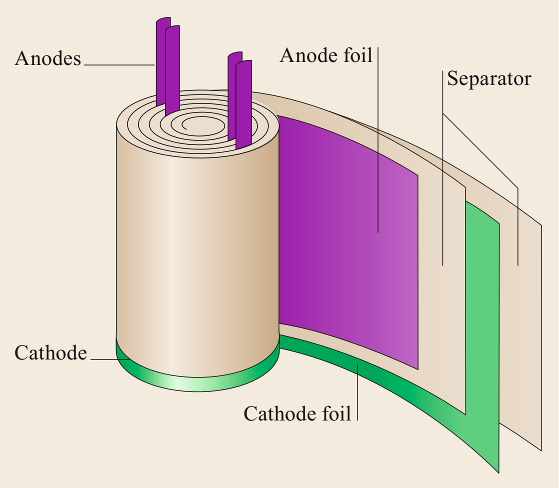

Fig. 26.13 Schematic illustration of geometry of an electrolytic capacitor

Substantially higher values for the capacitance may be achieved by *electrolytic capacitors*. The price is that these are *unipolar*, meaning that their cathode always needs to be connected to a potential that is lower than that of the anode. If the anode and cathode of an electrolytic capacitor are interchanged, the device may explode and poisonous substances may be released. Therefore, electrolytic capacitors have to be used in a careful manner.

Figure 26.13 shows the geometry of an electrolytic capacitor. Production details may be found in manufacturers' information [26.15]. Such a capacitor is formed by putting together a metal electrode, i.e., the anode, with a conducting liquid (the *electrolyte*) that has the property of forming a thin oxide layer on the surface of the metal. This layer is used as the dielectric. As the thickness of this layer is as little as a few atomic diameters, the C values of electrolytic capacitors are substantially larger than those of thin-film capacitors. They are, however, not suitable for high-frequency applications.

The top end of capacitor energy densities is marked by *double-layer capacitors*, branded as *gold caps*, *super* *caps*, or *ultra caps*. Their features are discussed in the section on electrical energy storage.

26.2.4 Inductors

Inductors make use of the following relations between magnetic fields and currents. Time-dependent currents produce time-dependent magnetic fields. Also, time-dependent magnetic fields produce rotational electric fields in the wire conducting the current, thus actually reducing the strength of the current that produces the magnetic field. Therefore, in exact opposition to the behavior of capacitors, inductors conduct the better, the less the current changes. An ideal inductor acts like a short circuit in direct-current networks.

Inductors are made from helix-shaped wires that may or may not be placed around a core. This almost universal construction can be understood as ensuring the optimal use of the Faraday–Henry law (26.13), which states that the tension induced in a closed loop wire is given by the rate of change of the magnetic flux through the area spanned by this loop, and of the Ampère–Maxwell law (26.17), which states that the strength of the magnetic field along a field line is proportional to the current enclosed by this line. Each extra turn of a wire thus increases both the magnetic flux and the tension for a given change of flux. Ten times more turns will increase the effect by a factor $10^2 = 100$. Introducing extra core material may be used to increase the strength of the magnetic field even further.

The calculation of the magnetic fields from the current for a given wire geometry is a difficult exercise and hardly ever produces an exact result. In most cases, only numerical tools give satisfactory results. It is, however, known that, whatever the result, the induced tension must be proportional to the rate of change of the current. Therefore, all the factors that may be hard to determine can be summarized in a single quantity called the *inductance* L, and the equation

$$U = \pm L \frac{\mathrm{d}I}{\mathrm{d}t} \tag{26.50}$$

is correct irrespective of the geometrical details and uncertainties. In some parts of the literature, the induced tension is referred to as the *electromotive force*. This name will be omitted for the sake of consistent use of the term “force.”

It may be irritating that, in the literature, (26.50) is written as often with a + sign as with a − sign. However, this is merely convention: If both the current and tension are given with respect to the same direction, the + sign applies. Otherwise, as usually indicated by the current arrow pointing in the opposite direction to the tension arrow, the − sign applies.

The amount of energy stored in the magnetic field of an inductor can be calculated from the energy that is needed to increase the current from zero to a certain value I. Starting from the definition of power $P = \mathrm{d}W/\mathrm{d}t = IU$, one gets

$$\begin{aligned} &\frac{\mathrm{d}W}{\mathrm{d}t} = IU = IL\frac{\mathrm{d}I}{\mathrm{d}t} \\ \rightarrow\quad &\mathrm{d}W = LI\mathrm{d}I \\ \rightarrow\quad &W = \frac{1}{2}LI^2 \,. \end{aligned} \tag{26.51}$$

The fact that $|\boldsymbol{B}| \approx I$ and $W \approx I^2$ suggests that $W \approx |\boldsymbol{B}|^2$, and in fact, a thorough analysis gives the result

$$w = \frac{W}{V} = \frac{1}{2\mu_0\mu_\mathrm{r}}\boldsymbol{B}^2 \,, \tag{26.52}$$

which not only seems natural but may also be severely misunderstood. At first glance, it seems to suggest that a large value of μ_r will lead to a reduction of the energy density of an inductor. However, can it be that putting iron into a coil will result in a reduction of the energy density? Of course, this is not the case, as can be shown as follows. In the presence of matter, the law of Ampère and Maxwell, (26.17), reads

$$\oint_{\partial A} (\mu_0^{-1}\mu_\mathrm{r}^{-1}\boldsymbol{B}) \cdot \mathrm{d}\boldsymbol{\ell} = I + \int \frac{\partial(\varepsilon_0\varepsilon_\mathrm{r}\boldsymbol{E})}{\partial t}\mathrm{d}\boldsymbol{A} \,, \tag{26.53}$$

indicating that, for a given inductor cross section A and in the absence of electric fields, $|\mu_\mathrm{r}^{-1}\boldsymbol{B}| \approx I$. Hence, doubling μ_r will increase $\boldsymbol{B}^2$ by a factor of 4 and consequently increase the energy density of the magnetic field according to (26.52) by a factor of 2. So, inspecting all the factors together, one finds that the energy density rises in proportion to μ_r.

As in the case of capacitors, networks of inductors may be treated as single inductors; for example, two inductors with inductances $L_{\mathrm{s}1}$ and $L_{\mathrm{s}2}$, placed in series, behave like a single one with an inductance

$$L_\mathrm{s} = L_{\mathrm{s}1} + L_{\mathrm{s}2} \quad \text{(inductors in series)} \,. \tag{26.54}$$

This may be understood as a consequence of the fact that the current is the same through both while the tensions add. Calculating the sum

$$\begin{aligned} U = U_1 + U_2 &= L_{\mathrm{s}1}\frac{\mathrm{d}I}{\mathrm{d}t} + L_{\mathrm{s}2}\frac{\mathrm{d}I}{\mathrm{d}t} \\ &= (L_{\mathrm{s}1} + L_{\mathrm{s}2})\frac{\mathrm{d}I}{\mathrm{d}t} \end{aligned} \tag{26.55}$$

shows the validity of (26.54).

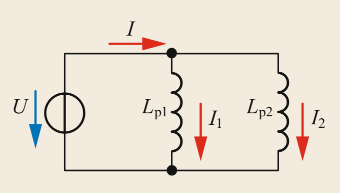

Fig. 26.14 Two inductors placed in parallel

If two coils are placed in parallel, they are both subject to the same tension U while, as shown in Fig. 26.14, their currents I_{p1} and I_{p2} add up to the total current I. If one eliminates I_{p1} from all the equations for the common tension, namely

$$U = L_{p1}\frac{dI_1}{dt} = L_{p2}\frac{dI_2}{dt} = L_p\frac{d}{dt}(I_1 + I_2) , \qquad (26.56)$$

one gets

$$\frac{1}{L_p} = \frac{1}{L_{p1}} + \frac{1}{L_{p2}} \quad \text{(inductors parallel)} , \qquad (26.57)$$

which also demonstrates that inductors and capacitors add in a complementary manner: inductances add when placed in series, whereas capacitances add when placed in parallel.

When choosing an inductor for a certain application, the diameter of the wire and the core material matter. The larger the current, the thicker the wire has to be for sufficient conductivity. At frequencies in or beyond the MHz region, two further effects matter. The *skin effect* describes the fact that the current is pushed towards the wire's surface, making its center useless for charge transport. Therefore, at high frequencies, bundles of insulated thin wires, called *Litz wires*, are used instead of single thick wires. Meanwhile, the *proximity effect* is observed if the conductivity of a wire is reduced by the presence of a high-frequency current close by. This effect can be counteracted by using special wiring geometries, often far from being solenoidal.

The core material is then mainly determined by the frequency of the application. At low frequencies (60 Hz, for example), iron cores are used to boost the magnetic inductance L by a factor of 1000 or even more. However, if a conduction material such as iron is exposed to an oscillating magnetic field, so-called *eddy currents* (or *Foucault currents*) are induced. The Faraday–Henry law describes the formation of rotational electric fields when magnetic fields oscillate. In a conductor, these rotational fields will push the electrons in the conductor to move along the field lines by means of the Coulomb force. In this manner, eddy currents are formed. Because the conductance of a material such as iron is finite, these eddy currents will produce heat: the larger the inductor's current and the larger the frequency, the more heat will be produced.

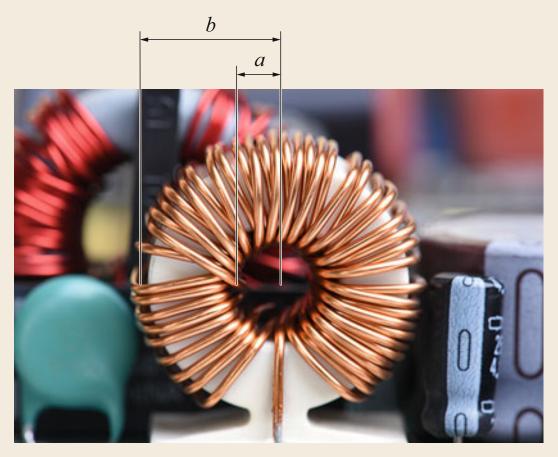

Fig. 26.15 Toroidal inductor with inner radius a and outer radius b. (Photo: © salita2010/stock.adobe.com)

The appearance of eddy currents can be reduced or even eliminated. If the iron core is laminated, i.e., if the core is produced as a stack of thin layers of oxidized iron, the currents cannot traverse the oxide layers. In this manner, the eddy currents are limited geometrically. However, for radio frequencies, this reduction is no longer sufficient. Instead of iron, nonconducting materials with large values of μ_r, i.e., *ferrites*, are used as the core material. These are ceramic compound materials of oxidized metals such as Fe_2O_3 (hematite, rust) and Fe_3O_4 (magnetite) or oxides including nickel or tin in addition to iron. Because they are insulators, there will be no eddy currents inside. At the high-frequency end of the spectrum, small values for the inductance are used and inductors do without cores.

Application Example: The Inductance of a Toroidal Inductor

Toroidal inductors are of special interest, as such devices may be (in fact, are) used to determine the values of μ_r of core materials experimentally.

For simplicity, the core is assumed to have a square-shaped cross section, as shown in Fig. 26.15. The inductance follows from the magnetic flux, which follows from the magnetic field. Hence, the field must be calculated first.

According to the law of Faraday and Henry, only the magnetic field enclosed by the current loops will contribute to the inductance (the field outside being negligible, anyway). The magnetic field lines must form circles with radii ranging from $r = a$ to $r = b$ and enclosing N turns of wire carrying a current I. According to the Ampère–Maxwell law, the strength of the field is

$$B(r) = \frac{\mu_0 NI}{2\pi r}\,. \tag{26.58}$$

The magnetic flux may then be calculated as

$$\Phi_\mathrm{B} = (b-a)\int_a^b B(r)\mathrm{d}r = \frac{(b-a)\mu_0 NI}{2\pi}\ln\left(\frac{b}{a}\right)\,. \tag{26.59}$$

Now, the tension due to the change of the current can be computed. Every current loop adds a tension $U_\mathrm{ind}(1\text{ turn}) = \mathrm{d}\Phi_\mathrm{B}/\mathrm{d}t$, so that all the turns in series produce

$$U_\mathrm{ind} = N\frac{(b-a)\mu_0 N}{2\pi}\ln\left(\frac{b}{a}\right)\frac{\mathrm{d}I}{\mathrm{d}t}\,. \tag{26.60}$$

Comparing this result with the definition of the inductance shows that the factor preceding the time derivative is the inductance L. If a core with μ_r is placed between the windings, the inductance becomes

$$L = \frac{(b-a)\mu_0\mu_\mathrm{r}N^2}{2\pi}\ln\left(\frac{b}{a}\right)\,. \tag{26.61}$$

Equation (26.61) paves the way towards a measurement of μ_r. Its value for a given core material can be calculated as $\mu_\mathrm{r} = L\,(\text{with core})/L\,(\text{without core})$. Finally, if the torus is narrow, i.e., if the diameter $D = b - a \ll a$, then application of $\ln(1+x) \approx x$ yields

$$L \approx \mu_0\mu_\mathrm{r}\frac{N^2D^2}{2\pi a}\,, \tag{26.62}$$

which also shows that, in this case, the shape of the individual turns no longer matters; indeed, the inductance depends only on the area D^2 and the radius of the torus a.

26.2.5 Alternating-Current Behavior

For a tension oscillating according to $u = \hat{U}\sin(\omega t)$, one has (by convention, a small u and a small i indicate sinusoidal forms of $U(t)$ and $I(t)$)

$$\begin{aligned}
R: u = Ri \quad &\rightarrow \quad i = \frac{1}{R}\hat{U}\sin(\omega t)\\
&\rightarrow \quad \varphi_U - \varphi_I = 0\,,\\
C: i = C\frac{\mathrm{d}U}{\mathrm{d}t} \quad &\rightarrow \quad i = \omega C\hat{U}\cos(\omega t)\\
&\rightarrow \quad \varphi_U - \varphi_I = -90^\circ\,,\\
L: u = L\frac{\mathrm{d}i}{\mathrm{d}t} \quad &\rightarrow \quad i = \frac{1}{\omega L}\hat{U}[-\cos(\omega t)]\\
&\rightarrow \quad \varphi_U - \varphi_I = 90^\circ\,.
\end{aligned} \tag{26.63}$$

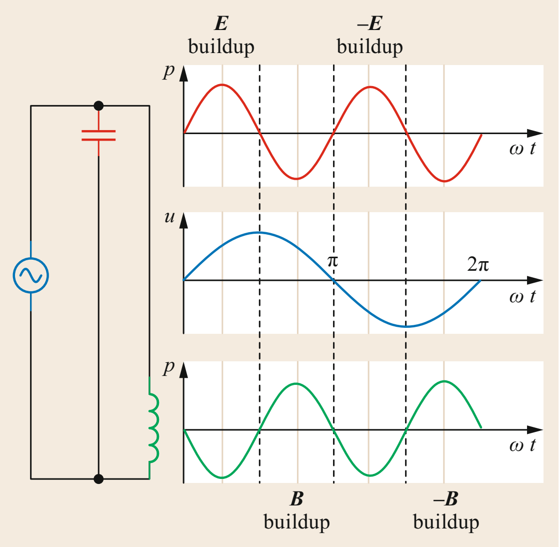

Fig. 26.16 Instantaneous power of an ideal capacitor (*top*) and an ideal inductor (*bottom*), compared with the instantaneous tension (*middle*)

So, ideally, this should be $\varphi_U - \varphi_I = \pm 90^\circ$ or zero. Such angles have surprising consequences for the time development of the instantaneous power $p(t) = i(t)u(t)$, as shown in Fig. 26.16. On average, the power is zero, meaning that there is a balanced flow of power in an out of the elements. When $p(t)$ is positive, energy is taken from the circuit and used to create fields; when it is negative, the fields give the energy back to the circuit. This behavior is in maximal contrast to that of a resistor, for which $p(t)$ is always positive, resulting in an average power of $\langle p\rangle = 1/2\hat{U}\hat{I}$, which is turned into heat by the resistor.

The calculations can be carried out in the most straightforward manner if the AC behavior of the circuit elements is considered in the complex plane. The reason is the simple behavior of the complex exponential function under differentiation and integration. Obviously, the three steps to take are:

- Transfer all equations into the complex plane, for example $\hat{U}\sin(\omega t + \varphi_u) \rightarrow \hat{U}\,\mathrm{e}^{\mathrm{j}\varphi_u}\mathrm{e}^{\mathrm{j}\omega t}$,
- Solve the complex equations, maybe with a result of the form $\hat{B}\,\mathrm{e}^{\mathrm{j}\beta}\mathrm{e}^{\mathrm{j}\omega t}$,
- Extract the result by taking the *imaginary part* of the complex result $\hat{B}\sin(\omega t + \beta) = \mathrm{Im}(\hat{B}\,\mathrm{e}^{\mathrm{j}\beta}\mathrm{e}^{\mathrm{j}\omega t})$.

A sinusoidal voltage is then represented as

$$\underline{u}(t) = \hat{U}\mathrm{e}^{\mathrm{j}\varphi_u}\mathrm{e}^{\mathrm{j}\omega t} = \underline{U}\mathrm{e}^{\mathrm{j}\omega t}\,, \tag{26.64}$$

where the *complex amplitude* $\underline{U} = \hat{U}\mathrm{e}^{\mathrm{j}\varphi_u}$ is the product of all the factors that do not depend on time. Note that

$u = \mathrm{Im}(\underline{u}) = \hat{U}\sin(\omega t + \varphi_u)$ and that complex quantities are indicated by being underlined.

The power of complex calculus is highlighted when applied to the model (26.45) and (26.50) for capacitors and inductors. By computing the derivatives

$$C: \quad \underline{i} = C\frac{\mathrm{d}\underline{u}}{\mathrm{d}t} \rightarrow \underline{u} = \frac{1}{\mathrm{j}\omega C}\underline{i}\,,$$

$$L: \quad \underline{u} = L\frac{\mathrm{d}\underline{i}}{\mathrm{d}t} \rightarrow \underline{u} = \mathrm{j}\omega L\underline{i}\,, \tag{26.65}$$

one may deduce that, in the complex plane, inductors and capacitors may be treated as *complex resistors*. And as a formal add-on, due to $\mathrm{e}^{\mathrm{j}\pi/2} = \mathrm{j}$, (26.65) contain exactly the same phase relations as (26.63).

This idea has been generalized by introducing a complex resistance called the *impedance* $\underline{Z}$ and a complex conductivity called the *admittance*, $\underline{Y}$ as follows:

$$\begin{aligned} &\text{Original} && \rightarrow \text{Complex generalization} \\ &u = Ri && \rightarrow \underline{u} = \underline{Z}\,\underline{i} \\ &i = Gu && \rightarrow \underline{i} = \underline{Y}\,\underline{u}\,. \end{aligned} \tag{26.66}$$

Sometimes, the real and imaginary parts are given separately

$$\begin{aligned} \underline{Z} &= R + \mathrm{j}X\,, \\ \underline{Y} &= G + \mathrm{j}B\,. \end{aligned} \tag{26.67}$$

The factors for the imaginary parts are called the *reactance* X and *susceptance* B. According to (26.66), $\underline{Z} = 1/\underline{Y}$, leading to the following relations:

$$\begin{aligned} R &= \frac{G}{G^2+B^2}\,, & X &= \frac{-B}{G^2+B^2}\,, \\ G &= \frac{R}{R^2+X^2}\,, & B &= \frac{-X}{R^2+X^2}\,. \end{aligned} \tag{26.68}$$

The *power* may also be generalized to suit the complex plane. Recalling that the average power used by a resistor is $\langle P\rangle = R\langle i^2\rangle = R\hat{I}\langle\sin^2(\omega t)\rangle = 1/2R\hat{I}^2$, the following definitions for the *apparent power* $\underline{S}$, the *real or average power* P, and the *reactive power* Q have proven to be useful:

$$\begin{aligned} &\text{Original} && \rightarrow \text{Complex generalization} \\ &P = \langle ui\rangle && \rightarrow \quad \underline{S} = P + \mathrm{j}Q \\ & && \rightarrow \quad P = \mathrm{Re}(\underline{S}) = |\underline{S}|\cos(\varphi_U - \varphi_I) \\ & && \rightarrow \quad Q = \mathrm{Im}(\underline{S}) = |\underline{S}|\sin(\varphi_U - \varphi_I)\,. \end{aligned} \tag{26.69}$$

To make a long story short, P is exactly what is called the power in mechanics, while Q is a measure of the energy oscillation in an out of a circuit element.

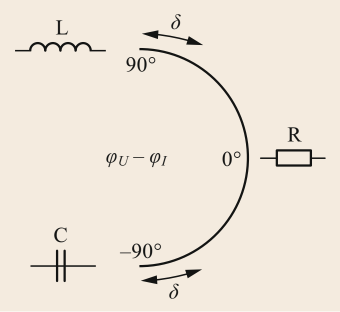

Fig. 26.17 The angle between tension and current in AC networks for the three basic circuit elements R, L, and C. The angular difference δ from the ideal values of ±90° is called the loss angle

26.2.6 Parasitics

The behavior of real i.e., nonideal circuit elements is a mixture of what it should be and deviations due to *parasitic elements*. The finite conductivity of connecting wires will force any element to retain a little bit of resistor behavior, the helical shape of wound capacitors will add some inductive behavior, and the closeness of the wires in a coil will add a small capacitance to an inductor.

In alternating-current networks, the effect of such imperfections is most conveniently described by the deviation of the angle between the current and tension. Any deviation from the angles given in (26.63) can be interpreted as a result of imperfections or parasitic elements. The difference from the ideal values for the angle is called the *loss angle* and appears as δ in Fig. 26.17. Often, the *dissipation factor* D or the *quality factor* Q are given instead of δ. Because of

$$D = \frac{1}{Q} = \tan\delta\,, \tag{26.70}$$

all these terms stand for the same. The larger the quality Q, the less the element behaves like a resistor.

Calculations based on this description of the AC behavior in the complex plane can help to understand why D and δ are called the dissipation factor and loss angle. The effect of the finite conductivity of the materials used in a capacitor can be approximated by introducing a resistance R_R in series with an ideal capacitor C_R. Then, the (*real*) capacitor's impedance is the sum

$$\underline{Z} = R_\mathrm{R} + \frac{1}{\mathrm{j}\omega C_\mathrm{R}} \quad \text{(capacitor with ohmic parasite)} \tag{26.71}$$

which may be represented by an addition of vectors in the complex $\underline{Z}$ plane, as shown in Fig. 26.18. This figure also demonstrates that the tangent of the loss angle δ is the ratio between the resistance R_R and the reactance $X_C(C_R)$. For an inductor, exactly the same analysis may be carried out, yielding

$$\underline{Z} = R_R + j\omega L \quad \text{(inductance with ohmic parasitic)}\,, \tag{26.72}$$

leading to a physical interpretation of the dissipation factor D.

The dissipation factor is the ratio of the ohmic power to the reactive power

$$D = \frac{P}{Q}\,. \tag{26.73}$$

It turns out that the definition in (26.73) remains correct even if other models of parasitic elements are used, while (26.71) and (26.72) are tied to a single ohmic resistor in series.

Finally, at the top end of the frequency spectrum, parasitic capacitances dominate the admittance of inductors, while parasitic inductances begin to dominate the impedance of capacitors. In brief: capacitors become inductors and vice versa.

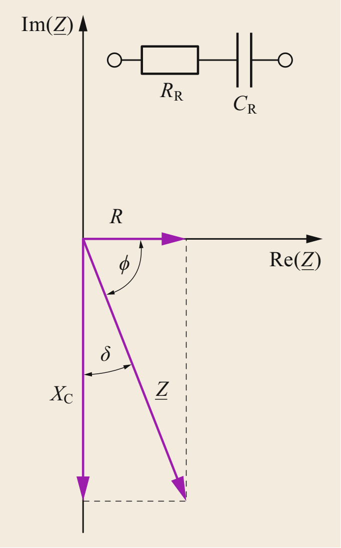

Fig. 26.18 Model of a *real* capacitor and a sketch of the corresponding addition of its impedance contributions in the complex plane, also indicating the meaning of the loss angle δ

26.3 Semiconductor Devices

All key elements of modern electronics are made from semiconductors. The features of most semiconductor devices are based on the special properties of highly purified monocrystalline silicon. Slices of these crystals, called *wafers*, are usually contaminated at the subthousandth level. This process is called *doping*. Most semiconductor functions can be traced back to an interaction of silicon areas with different doping materials and concentrations. In this section, the formation of diodes and bipolar transistors is discussed. Then, *metal–oxide–semiconductor (MOS) devices* are introduced as core elements of digital networks. These may be connected to electrical machines using *power semiconductor devices*.

26.3.1 Semiconductors

Semiconductors are materials that behave almost like insulators [26.16]. At 0 K, all their electrons are tightly connected, either to individual atoms or in binding orbitals. However, in contrast to an insulator, at room temperature, a tiny fraction of the electrons in binding orbitals in a semiconducting material are sufficiently energetic to move over large distances within the crystal.

The quantum-mechanical interpretation of this behavior is sketched in Fig. 26.19. As two orbits of two silicon atoms approach each other, new orbits are created, one with a slightly increased electron energy and another with a reduced energy. The latter will host both electrons. The more atoms that take part in this new orbit creation process, the more energy states will appear above and below the original energy level. It turns out that these energy states form two *energy bands*, the *valence band* and the *conduction band*, with a distinct *energy gap* between them. Energy bands denote regions of energy, where many energy states gather. While at 0 K the valence band hosts all the electrons, the energy gap to the conduction band is small enough for a tiny fraction at the top end of the electron's thermal energy spectrum to reach into the conduction band. Consequently, the number of conducting electrons increases with temperature. The energy that is halfway between the highest-energy state hosting electrons at 0 K and the lowest-energy empty state is called the *Fermi energy*.

Once an electron has reached the conduction band, it may follow the electric fields, just like an electron in a conductor. However, in a semiconductor, there is an additional type of conductance: whenever an electron *jumps* into the conductance band, it is missing from the valence band. In other words, it leaves a binding orbital with one electron too few. This deficit is called a *hole*, and it behaves similar to a carrier of positive charge: under the influence of an external field, the orbital may be refilled by an electron from an adjacent orbital, follow-

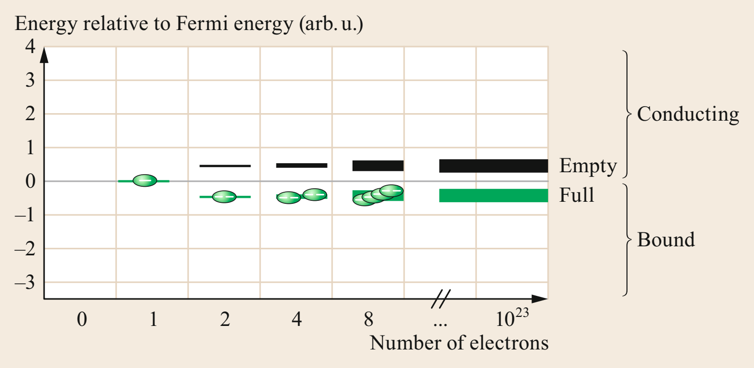

Fig. 26.19 Emergence of energy bands *from left to right*: Two energy states of electrons in silicon atoms combine to form two new states with slightly different energies. At 0 K, both electrons are found in the lower-energy orbit. The more atoms and electrons are combined, the more combined states result. These states turn out to form *energy bands*

ing that field. The net effect is that the positive charge has moved to the location where the refilling electron came from. If this process is repeated many times, the set of all refilling electrons moving in one direction have the same effect as a single positive carrier of charge moving in the opposite direction. This process is called *hole conduction*, and it is a unique feature of semiconductors. Consequently, the resistance of a piece of semiconductor material with a cross section A is determined by the concentrations n, p and mobilities μ_n, μ_p of negative electrons (n) and positive holes (p). The equation

$$R = \frac{1}{Ae} \frac{1}{p\mu_\mathrm{p} + n\mu_\mathrm{n}} \tag{26.74}$$

reduces to the expression used for conductors in (26.38) when p is set to zero.

Apart from silicon, germanium and mixed crystals such as gallium arsenide may also be used as semiconductor substrates. Their advantage is a better electron mobility. However, for a single reason, silicon alone accounts for more than 90% of semiconductor sales. Oxidizing silicon gives quartz, one of the best insulators known. And it is exactly this property that allows efficient fabrication of metal–oxide–semiconductor (MOS) transistors. In this way, all modern processors are not only based on the properties of silicon itself, but also on its ability to form a well-insulating oxide.

26.3.2 Doping, the Key to Top Performance

While the structure of silicon forms the backbone of semiconductor devices, most electrons and holes responsible for the currents come from tiny impurities, introduced via a process called *doping*. In fact, when silicon is doped with a pentavalent substance such as phosphorus, the density of mobile electrons is almost identical to the density of doping atoms. For this reason, these substances are called *donors* (as they donate electrons). Meanwhile, if silicon is doped with a trivalent substance such as aluminum, the number of holes is almost equal to the number of aluminum atoms. These materials are called *acceptors*, as they *accept electrons* from the valence band. According to (26.74), the resistance can thus be adjusted by choosing the concentration of doping atoms.

The effect of doping is shown in Fig. 26.20. A pentavalent atom replacing one silicon atom will place four of its five outer electrons into the binding orbit, leaving one electron that does not fit into the structure. It turns out that the potential energy of this electron is slightly lower than the lower limit of the conduction band. Therefore, at room temperature, almost all the extra electrons will have sufficient thermal energy to be housed by the conduction band. The energies of all the pentavalent doping atoms together form a small band just below the conduction band. A piece of silicon having this extra band is called *n-doped*.

Holes will emerge due to doping with a trivalent substance. As the electrons in the neighborhood of an aluminum atom are not as tightly bound as those of a silicon atom, this type of doping will produce a small band of electron energies just above the valence band. Consequently, at room temperature, this new band will be almost completely filled with electrons from the valence band, leaving one hole per doping atom. Figure 26.21 shows the energetic structure of doped silicon.

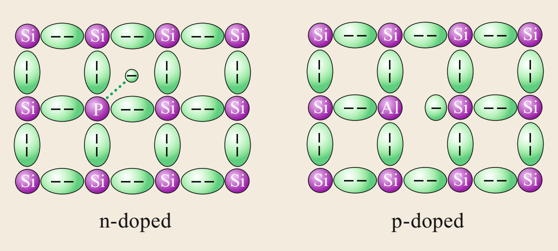

Fig. 26.20 Two-dimensional sketch of doping effect. An n-doped piece of silicon has some of its atoms replaced by pentavalent atoms. These will contribute extra electrons. In p-doped silicon, trivalent atoms mean that adjacent binding orbitals are filled by only a single electron

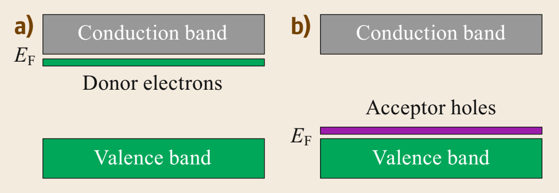

Fig. 26.21a,b Energy structure of n-doped silicon (**a**) and p-doped silicon (**b**). At 0 K, the donor electron band is full, while the acceptor hole band is empty

Donor doping affects the concentration of holes as well as the number of electrons. Their relation is given by an equilibrium condition described by the *law of mass action*, which will be derived next. In pure silicon, the number of holes equals the number of electrons. This density of electron–hole pairs is called the *intrinsic density* n_i. The number is tiny, but grows strongly with temperature. The value of n_i is the result of a dynamic equilibrium that is reached when as many electron–hole pairs are thermally created as are eliminated by electrons dropping into holes. Now, if one doubles the density of electrons n by doping, the chance of a hole to be filled will double as well. So, the density of holes p will be halved. Tripling n will reduce p by a factor of there, and so on.

This reasoning can be generalized as follows: The law of mas action states that the product of the density of electrons and the density of holes equals the intrinsic density squared

$$np = n_i^2 \,. \tag{26.75}$$

Due to this law, an increase of the density of charge carriers of one type is always accompanied by a decrease of the density of the opposite type. Therefore, the corresponding holes or electrons are also called *majority carriers* and *minority carriers*.

26.3.3 Diodes

Diodes, bipolar transistors, and thyristors use the special properties that semiconductors acquire at the junctions between p regions and n regions. An understanding of the properties of this so-called p–n *junction* is thus crucial to understand these devices.

A *diode* consists of a p region adjacent to an n region, with both regions connected to the outside world, separately. Figure 26.22 indicates what happens if a p-doped region comes into contact with an n-doped region. At room temperature, both electrons and holes are in thermal motion. They move almost freely within the silicon crystal. This randomly orientated movement is called *diffusion*. If an electron happens to diffuse into the p region, it will all of a sudden be exposed to a large number of holes to recombine with—and so it will. In this manner, an extra negative charge will be placed in the p region. This charge will also be missing from the n region. Similarly, a hole crossing the p–n junction will find many electrons to recombine with, thus placing an extra positive charge in the n region. In this manner, a region with locally bound positive charges will form in the n region next to the junction, while a region with negative charges will appear in the p region. As almost all freely moving charges close to the junction are trapped, such regions are called *depletion regions*.

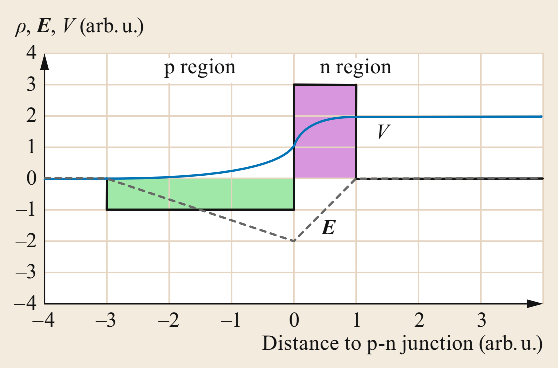

Fig. 26.22 Charge density, electric field, and electric potential near a p–n junction. In this example, the concentration of donor electrons is three times as large as the concentration of acceptor holes. The region with a nonzero electric field is called the *depletion region*

According to Gauss's law, the locally bound charges will produce an electric field. A more elaborate analysis would show that the field produces a tension

$$U_D = \frac{kT}{e} \ln\left(\frac{n_A n_D}{n_i^2}\right) \tag{26.76}$$

between the regions. It is called the *diffusion voltage*. The factor kT, known from thermodynamics, shows that it is a result of thermal movement. The appearance of n_i and the doping densities n_A and n_D indicates that the value of U_D will result from a statistical equilibrium.

The equilibrium condition can be understood on the basis of the forces on the charge carriers in the depletion region. Figure 26.22 shows that, for electrons approaching the p–n junction from the n region, the local charges produce a potential barrier. The higher the tension, the smaller the fraction of electrons whose thermal energy suffices to cross the entire barrier. At the same time, holes from the n region, i.e., the minority carriers, will find themselves on a potential slide. All holes entering the depletion region from the n-doped side will be dragged into the p region. The complementary behavior will be found

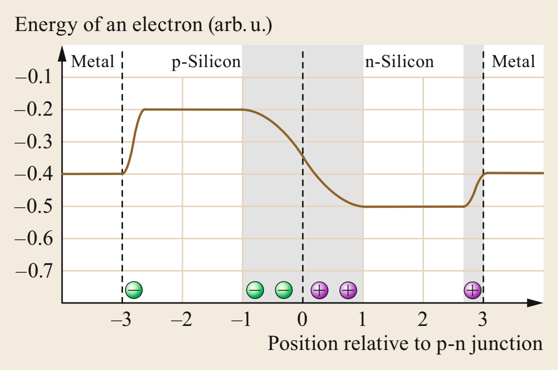

Fig. 26.23 Potential energy of an electron in doped silicon and adjacent metal. Energy conservation requires that the potentials at both ends be identical. Depletion regions (*gray*) and locally bound charges may occur at any of the junctions, depending on the materials chosen

when analyzing the p side of the junction. Equilibrium is reached when the potential barrier for the majority carriers is so high that the number of these that can cross it equals the number of minority carriers moving in the same direction. As majority carriers outnumber minority carriers by many orders of magnitude, one finds the following rule: depletion regions reflect majority carriers, but they are transparent for minority carriers.

Finally, the value of U_D can never be measured directly. A voltmeter connected to a p–n junction (or a diode) will show 0 V. The reason is that the potential barrier at the p–n junction is exactly compensated by the potential jumps at the external connections of the silicon, i.e., the *contact potentials*. Figure 26.23 shows that this compensation must be exact due to energy conservation. An electron moving along a closed loop of metal → p-silicon → n-silicon → metal may neither gain nor lose energy.

If an external voltage U is applied to the p–n junction, the situation described in Fig. 26.23 may change significantly. A positive voltage applied to the p region relative to the n region reduces the width w of the depletion region. For an abrupt change of doping concentrations, one can determine the width of the depletion zone to be

$$w = \sqrt{\left(\frac{n_A + n_D}{n_A n_D}\right) \frac{2\varepsilon}{e}(U_D - U)} . \tag{26.77}$$

At the same time, the height of the potential barrier for majority carriers is reduced—with drastic consequences: the number of majority carriers having enough energy to cross the potential barrier increases exponentially with the voltage applied. This is a consequence of the fact that all thermodynamic distribution functions have an exponentially falling *tail* at the upper end of the energy spectrum. Therefore, the current from majority carriers diffusing into the depletion region, named the *diffusion current*, increases exponentially with the voltage applied. The diode is then said to be *forward biased*. Figure 26.24 shows the density of charge carriers near the p–n junction for this case.

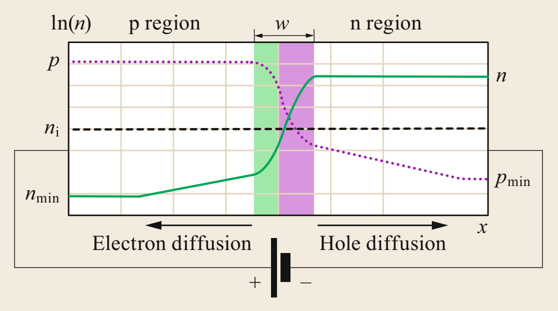

Fig. 26.24 Charge concentration in a forward-biased diode. The dropping of the densities in the depletion region is due to the electric field. The exponential decrease down to the level of the minority carriers is a result of diffusion and recombination

If the n region is given a higher potential than the p region, the diode is said to be *reverse biased*. Then, even fewer majority carriers cross the depletion region than in the equilibrium state. Quite in contrast, the current due to the minority carriers will remain constant, as all of them will move *down the potential slide*, irrespective of its height. As the carriers follow the field, this current is called the *field current*. The net effect is a reverse-bias current which, for large voltages, tends towards a constant value. This value is reached when the current due to majority carriers is negligible.

All together, the current of a diode is approximately given by the *Shockley equation*

$$I = I_S \left(e^{U/U_T} - 1\right) , \tag{26.78}$$

where the *temperature voltage* U_T is a common abbreviation for $U_T = kT/e$. I_S is the current that should be measured in reverse-bias connection. Figure 26.25 shows that this is an approximation that is only valid for low voltages. If the diode is reverse biased, electron–hole pair creation due to thermal movement and impurities will dominate, while for large currents, the ohmic resistance of the silicon reduces the current. For almost all practical purposes though, the Shockley equation suffices. In the context of digital networks or power electronics, it is even sufficient to regard a diode as a device that conducts well above $U \approx 0.6$–0.7 V and not at all below this region.

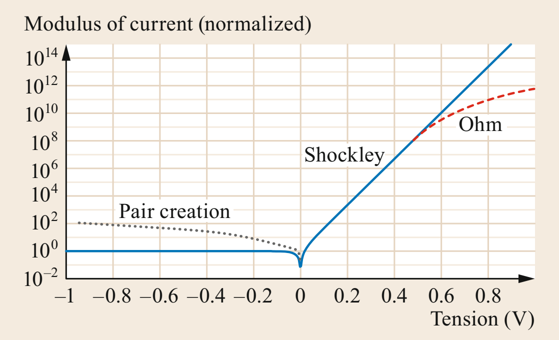

Fig. 26.25 The voltage–current characteristic of a diode, given approximately by the Shockley equation. Deviations are due to the creation of pairs in the depletion region and the ohmic resistance of the silicon

As a result of the steep increase of the current with voltage, the depletion zone around a p–n junction never disappears completely. A vanishing of the depletion region would be equivalent to the formation of a short circuit.

For certain semiconductor materials, efficient pair creation from incoming light is the design goal. Pairs produced in the depletion region will separate under the influence of the electric field, thus contributing to the field current. The number of pairs created is proportional to the number of incoming photons. As shown in Fig. 26.26, this can be used in either *solar cells* or *photodiodes*. In the region where the product $P = UI$ is negative, the diode acts as an electric power generator. In the reverse-biased region, it can be used as a light detector. In *light-emitting diodes* (LEDs), this process is reversed. Finally, there is a variety of types of diode apart from those presented here.

26.3.4 Bipolar Transistors

Bipolar transistors are used when large currents must be steered by small currents. They can be found wherever amplification is needed. Figure 26.27 shows that a transistor is a device with two p–n junctions, corresponding to two diodes connected *back to back*. Their terminals are called the *emitter* (E, emitting electrons), *base* (B, from fabrication history), and *collector* (C, collecting electrons). Two features turn this structure into an amplifying device:

1. The doping concentrations obey $n(\text{emitter}) \gg p(\text{base}) \gg n(\text{collector})$. The first $\gg$ sign ensures that, in forward-bias mode, most of the current through this junction is carried by electrons. So, when forward biased, the emitter emits electrons and the base receives them. On the contrary, a forward-biased junction between the collector and base will mainly have a hole-based current.
2. The p region in the middle (the base) is made so thin that an electron coming from the emitter has little chance to recombine before entering the depletion zone between the base and collector. With electrons being minority carriers in the base, whenever they reach the p–n junction to the collector, the field in the depletion zone will drag them into the collector, thus forcing it to collect electrons.

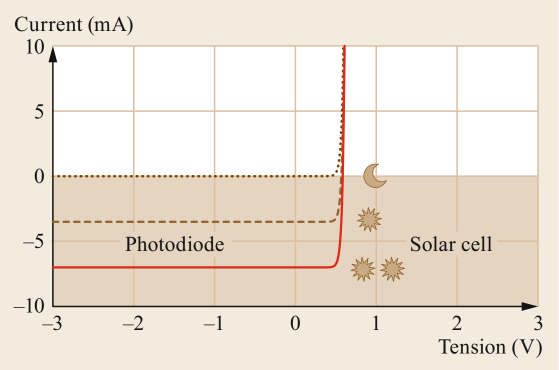

Fig. 26.26 Current–voltage characteristic of a light-sensitive diode. In the low-voltage forward-bias region, it is used as a solar cell. In the reverse-bias region, it acts as a photodiode

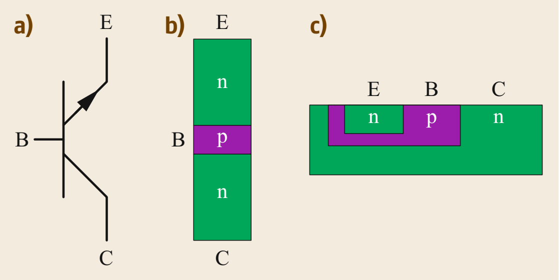

Fig. 26.27a–c The npn bipolar transistor symbol **(a)**, its doping structure **(b)**, and its geometry in planar technologies for integrated circuits **(c)**. The *arrow* indicates the position of the emitter

The design of a transistor is optimized for a mode of operation called *forward active*. This mode is characterized by a forward-biased base–emitter junction but a reverse-biased base–collector junction. In Fig. 26.28, the tensions would then obey $U_{\text{BE}} \approx 0.6\,\text{V}$ and $U_{\text{CB}} > -0.6\,\text{V}$. A small current I_{B} entering the base is physically equivalent to a certain number of electrons entering the base from the emitter. Holes play a minor role here due to the concentration gradient. One would expect them to recombine in the base and thus close the

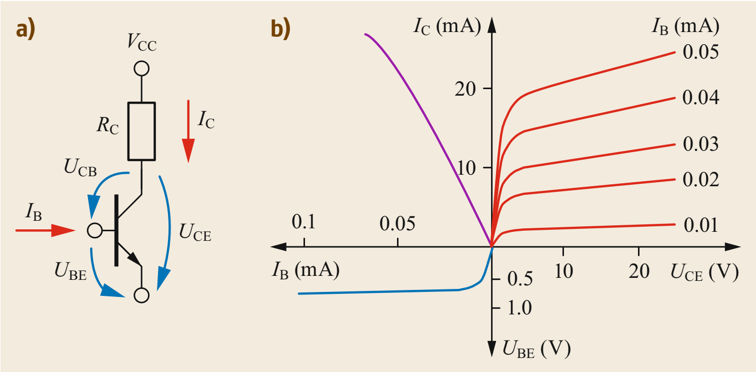

Fig. 26.28a,b Schematic of a transistor and a resistor **(a)** and the corresponding characteristics **(b)**

electrical circuit with the emitter. However, the special geometry of a transistor prevents the electrons from doing so. The base is so thin that most of the electrons will enter the depletion zone adjacent to the collector. Only one or a few percent of the current will make it to the base terminal. With the percentage of electrons being given by the geometry, the small fraction of the current passing through the emitter terminal and coming from the base terminal is also almost constant. In other words, a small value for the base–emitter current I_B determines a much larger value for the collector–emitter current I_C. Colloquially, this phenomenon is called *current amplification*, although *current steering* would be more precise.

The behavior of a transistor is usually summarized as shown in Fig. 26.28b. Starting with U_{BE} and moving clockwise, one finds the typical diode characteristics between U_{BE} and I_B (with axes interchanged). The top left quadrant shows the $I_B \leftrightarrow I_C$ characteristics. A constant current amplification would correspond to a straight line there. The ratio $B_f = I_C/I_B$ is called the *forward current amplification*, while the slope of the line, $\beta_f = \Delta I_C/\Delta I_B$, is called the *differential forward current amplification*. Finally, each value for the collector current depends on the collector emitter voltage U_{CE} as well as the base current I_B. Figure 26.28 shows that, for a fixed value of the base current, there is a separate characteristic $I_C \leftrightarrow U_{CE}$. In this quadrant, a flat line corresponds to constant amplification. Figure 26.29 shows that this is often a good approximation. The *Ebers–Moll model*

$$\begin{aligned} I_B &= I_{B,S}\left(e^{U_{BE}/U_T} - 1\right) \approx I_{B,S}e^{U_{BE}/U_T}\,,\\ I_C &= B_F I_B \approx B_F I_{B,S} e^{U_{BE}/U_T}\,,\\ I_E &= I_B + I_C \end{aligned} \tag{26.79}$$

reproduces the measured currents quite well. It is the first-choice starting point for the development of bipolar circuits.

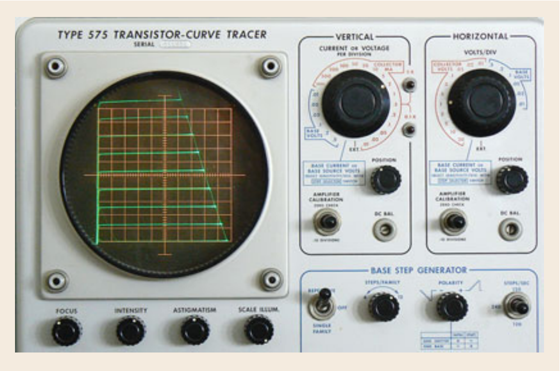

Fig. 26.29 Photo of a transistor tracer. Here, the collector current is displayed as a function of the collector–emitter tension. Each of the almost horizontal lines corresponds to a fixed value of the base current I_B

Nevertheless, as shown in Fig. 26.29, there is always a slight increase of the current with U_{CE}. This increase is due to the fact that, as the collector–base tension increases, the depletion region adjacent to the collector zone becomes broader and reaches further into the base. As a result, the path along which electrons may recombine becomes shorter. The collector current increases relative to the base current. This phenomenon is called the *Early effect*.

Apart from the active mode, there are three more modes of operation. If both p–n junctions are forward biased, the transistor is said to be *in saturation*. This mode is characterized by almost constant voltages $U_{BE} \approx -U_{CB} \approx 0.6\ldots0.7\,\mathrm{V}$ and a current amplification less than the value for B_f given in the transistor's datasheet. Designers try to avoid this mode whenever speed is crucial. If both p–n junctions are reverse biased, no currents flow. The transistor is in the *cut-off* mode. Finally, if the base–collector diode is forward biased while the other junction is not, the transistor is in the *reverse active* mode. The amplification in this mode may have values well below one. For digital electronics

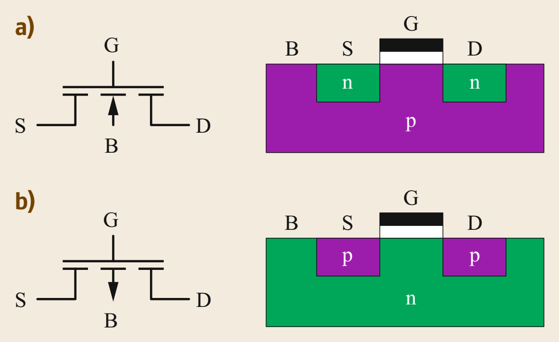

Fig. 26.30a,b Schematic symbol of **(a)** an NMOS transistor and its doping profile cross section and **(b)** the same for a PMOS transistor. The terminals are labeled B for *body*, S for *source*, G for *gate*, and D for *drain*

based on bipolar transistors (such as the 7400 series), a value close to zero is crucial for proper operation.

26.3.5 Metal–Oxide–Semiconductor Transistors

Metal–oxide–semiconductor (MOS) transistors are the workhorses of the digital world. They may be counted by the millions inside computers, smart phones, etc. for exactly one reason: reduced heat production. The steering of a MOS transistor does not require any permanent current. And a certain type of transistor circuit called CMOS (Fig. 26.33), once switched into a certain state, will keep it without further consumption of current. Hence—and this is the important point—there will be no further ohmic heat production. Modern processors have heating power densities exceeding those of electric cookers. So, a reduction of the heat production is crucial for an increase of transistor densities. And due to the finite value of the speed of light, a high density is a requirement for high clock rates. Without MOS, there would be no gigahertz processors.

MOS transistors come in two types with complementary doping profiles, the *n-channel (NMOS) transistor* and the *p-channel (PMOS) transistor*. Their schematic symbols and cross sections through the silicon setup are shown in Fig. 26.30. A MOS transistor has four terminals. The *body* insulates the two other doped areas, the *source* and the *drain*, from each other and from the rest of the chip. The insulation is the result of a reverse-biased (or 0 V) p–n junction forming a depletion region. An up-to-date transistor geometry is shown in Fig. 26.31. It minimizes the size of the p–n junctions and thus minimizes the (useless) current due to minority carriers.

The current through a MOS transistor is determined by a voltage rather than a steering current. The voltage is applied to a very thin layer of silicon dioxide. This layer is situated below a piece of metal connected to the *gate* terminal. It is called the *gate oxide*. If the potential at the gate of an NMOS transistor is larger than the potential of the body, the electric field through the oxide pushes the majority carriers of the body, the holes, away from the gate. A new depletion zone is thus created below the gate oxide. Above a certain value of the voltage, viz. the *threshold voltage* U_{Th}, the field is strong enough to form of a thin conducting layer of minority carriers (electrons) just below the gate oxide. This is called the *inversion layer*. It is formed if an electron loses more energy by moving to the gate oxide than is needed to leave its place in the valence band. Now, electrons are also the majority carriers of the adjacent n-doped regions named the source and drain. So, the emergence of this inversion layer leads to a conducting channel between the two n-doped regions. And this is what MOS transistors are about: the source and drain terminals are either connected or not, depending on the voltages involved. Technologically, there is no difference between the source and drain. Which of the two n-doped regions is the source is entirely determined by the tensions applied, The terminal with the lower potential is the source of an NMOS transistor, while that with the higher potential is the drain. Also, in some circumstances, they may even swap roles. In any case, the absolute value of the voltage between the gate and source will alway be larger than or equal to the value between the gate and drain. And, by definition, any current through an NMOS transistor will always flow from the drain to source. Therefore, the source is *the source of electrons* while the drain is *the drain of electrons*.

If the gate–source voltage is larger than the threshold voltage, and the gate–drain voltage is too, the tension across the gate oxide will suffice to form an inversion layer under the entire gate oxide. The transistor is then said to be in the *triode mode*. In this mode, the

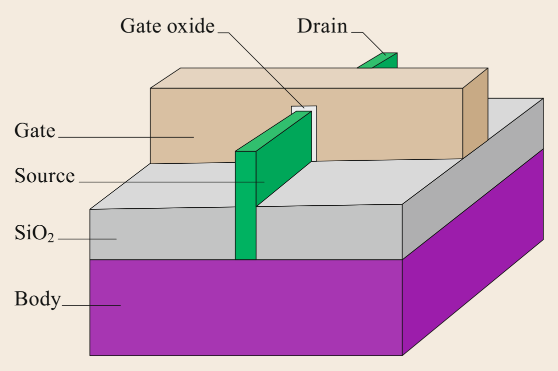

Fig. 26.31 A modern MOS transistor called a 3-D transistor by Intel [26.17]

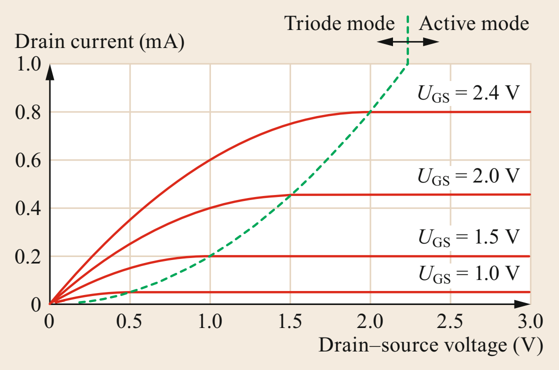

Fig. 26.32 Drain current of an NMOS transistor as a function of the drain–source voltage for various fixed values of the gate–source voltage U_{GS}

current from the drain to source is determined by the gate–source voltage U_{GS}, the drain–source voltage U_{DS}, and a technological factor β_N. The result

$$I_{DS} = \beta_N \left[(U_{GS} - U_{Th,N}) U_{DS} - \frac{1}{2} U_{DS}^2 \right] \qquad (26.80)$$

shows that the current plotted as a function of U_{DS} will give a parabola with a positive maximum. So, this formula accounts for the upper left part of the characteristics shown in Fig. 26.32. Clearly, (26.80) cannot account for the entire characteristic, as it predicts negative currents for large values of U_{DS}, which is not the case, as the discussion below suggests. For a given U_{GS}, the drain current reaches its maximum at $U_{GS} - U_{Th,N} = U_{DS}$. At this point, $U_{GD} = U_{Th,N}$; in other words, the maximal current is reached when the channel is about to be cut off from the drain. If U_{DS} is increased any further, the channel no longer reaches the drain. In this case, the conditions leading to (26.80) are no longer met.

For larger values of U_{DS}, the electrons reaching the end of the channel find themselves in a similar situation to those approaching the p–n junction of a bipolar transistor. They enter a depletion zone with a field dragging them towards the drain. As a result, all electrons entering the channel from the source side eventually reach the drain. Therefore, an incomplete channel goes along with an approximately constant current for a given value of U_{GS}, and the value of the current is given by the maximum that can be derived from (26.80).

$$I_{DS} = \frac{\beta_N}{2} (U_{GS} - U_{Th,N})^2 \,. \qquad (26.81)$$

The transistor is now said to be in the *active mode*, behaving almost like a current source between the drain and source terminals.

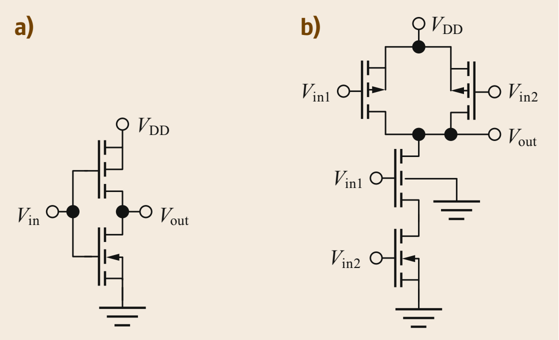

Fig. 26.33a,b Schematics of **(a)** a CMOS inverter and **(b)** a CMOS NAND (= not and) gate

Along the line $I_{DS} = \beta_N U_{DS}^2/2$, both (26.80) and (26.81) give the same current. When crossing this line, the mode changes from triode to active. In Fig. 26.32, this is shown by a dashed line.

Equation (26.81) can only be an approximation. The shorter the channel under the gate oxide becomes, the stronger the lateral field within the channel becomes (d gets smaller in $E = U/d$). Therefore, there is a slight current increase. Usually, a linear model of this increase, i.e.,

$$I_{DS} = \frac{\beta_N}{2} (U_{GS} - U_{Th,N})^2 (1 + \lambda_N U_{DS}) \,, \qquad (26.82)$$

is sufficiently accurate. The phenomenological factor λ_N is referred to as the *channel length modulation*.

Most digital electronic devices use a design technique in which there is one PMOS transistor complementing every NMOS transistor and vice versa. The result is called *complementary MOS* (CMOS). This is the most effective design method as far as energy consumption is concerned. Figure 26.33 shows the two most basic gates of this type, an inverter and a NAND gate. The key to understanding the way these gates work is to realize that, for each pair of complementary transistors connected to the same input potential V_{in}, only one is in a conducting mode. For $V_{in} = 0$, the PMOS transistor will be conducting. When the input changes from V_{DD} to 0, the PMOS switches from cut-off via active to triode mode while the NMOS is in a cut-off state. If $V_{in} = V_{DD}$, only the NMOS transistor conducts, and $V_{out} = 0$. So, if the voltages are associated with logical ones and zeros, the inverter will turn ones into zeros and vice versa, as the following table shows:

$$\text{Inverter:} \quad \begin{aligned} V_{in} &= V_{DD} & &\rightarrow & V_{out} &= 0 \,, \\ V_{in} &= 0 & &\rightarrow & V_{out} &= V_{DD} \,. \end{aligned}$$

The way the NAND gate works can be understood on the basis that a connection is established if either

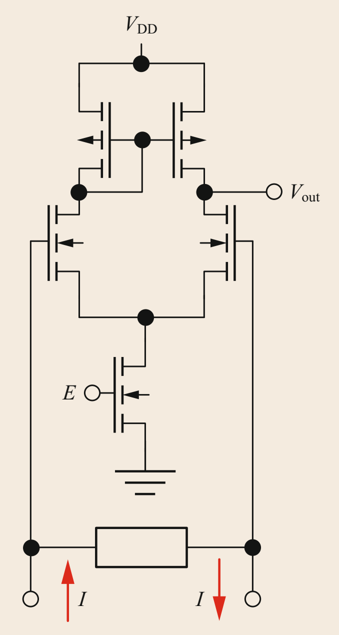

Fig. 26.34 A PCI Express receiver (body connections not shown). With $E = V_{DD}$, it is enabled. When the current I is positive, the resistor produces a potential difference, which is turned into $V_{out} \approx V_{DD}$ by the differential CMOS amplifier; if I is negative, $V_{out} \approx 0$

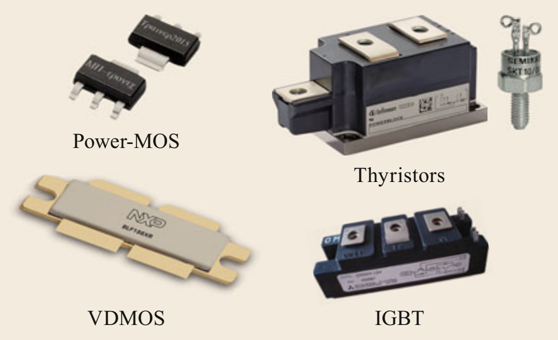

Fig. 26.35 Different types of power electronic devices

one of two parallel transistors conducts, or if both transistors in series conduct. The result

NAND gate:	V_{in1}	V_{in2}	$\to$	V_{out}
	0	0		V_{DD}
	0	V_{DD}		V_{DD}
	V_{DD}	0		V_{DD}
	V_{DD}	V_{DD}		0

may be interpreted as $V_{out} = \text{not } (V_{in1} \text{ and } V_{in2}) = V_{in1}$ NAND V_{in2}.

Whereas originally MOS transistors were used for digital circuits, the goal of integrating as many functions as possible onto a single chip led to increasing demand for analog CMOS circuits. An extensive discussion of these applications can be found in [26.18].

Application Example: A PCI Express Receiver

Peripheral Component Interconnect (PCI) Express is a point-to-point interface based on current loops. It can be found in almost any personal computer produced after 2010. A logical 1 is transmitted if the current flows in one direction, while a 0 corresponds to a current in the opposite direction. Figure 26.34 shows that, at the receiver end, a resistor turns the current into a voltage difference. And that is precisely what the transistor circuit can amplify well. Consider first the case of no current through the resistor: for symmetry reasons, both currents then have to be equal in both branches of the transistor circuit. The top left PMOS transistor has $U_{GD} = 0$ due to its wiring. It is therefore in the active mode, and so has to be the top right PMOS transistor, as it has the same U_{US} and the same current as the other one. Now assume that the current I in Fig. 26.34 reaches a small positive value. Then, the left NMOS transistors gets a larger U_{GS} than the right one. The NMOS transistor will therefore force the current in the left branch to be larger than that in the right branch. At the same time, the two PMOS transistors will keep the currents equal, *as long as the voltages in both branches are similar*. Because of the almost flat I_D versus U_{GD} characteristic of the PMOS transistors in the active mode (Fig. 26.32), a tiny decrease of I_D is necessarily accompanied by a large decrease of U_{GD}. But this is necessary, as the NMOS transistor connected to ground forces the sum of the currents through both branches to be constant. As a consequence, a little bit more current through the left branch will make V_{out} rise strongly. Consequently, $V_{out} \approx V_{DD}$ for $I > 0$ and $V_{out} \approx 0$ for for $I < 0$. This circuit is particularly popular, because V_{out} is stable, even if the surrounding is noisy.

26.3.6 Power Semiconductor Devices

Power semiconductor devices connect logical circuits to electrical machines. In this way, they turn strong machines into intelligent strong machines. Electromobility is unthinkable without power semiconductor devices. Figure 26.35 shows that power semiconductor devices are rather large. It may happen that an entire silicon wafer is turned into a single power device, rather than 100 microprocessors. The most popular power components are *thyristors*, *vertically double-diffused MOS (VDMOS)* transistors, and *insulated-gate bipolar transistors (IGBTs)*.

Thyristors behave like diodes that may be switched on or off. Once switched off, they may withstand several kV, whereas while being in the *on state*, they may conduct currents of several kA. Once a thyristor is in the on state, it will keep it until either the tension

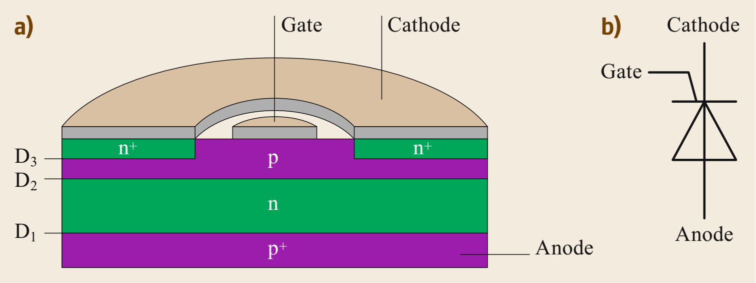

Fig. 26.36a,b Cross section through a thyristor **(a)** and schematic symbol **(b)**

approaches 0 V or the gate gets a reversed polarity. Figure 26.36 shows a cross section through a thyristor. Its four-layer npnp structure may be regarded as a series of three diodes. Either the outer two diodes (D_1 and D_3) are forward biased, or the one in the middle (D_2) is. One might therefore expect that a thyristor would never conduct at all. However, the setup may be turned into a switchable device by using the same techniques as known from bipolar transistors: the uppermost n layer is doped with a much higher concentration than the upper p layer. And because this p layer is thin, a small gate-to-cathode current will pass, along with the much larger current between the two n layers. So, the p–n junction labeled D_2 will be traversed by a large current despite being reverse biased. The result can be seen in the thyristor's characteristics, shown in Fig. 26.37. The lowermost junction serves as a current stabilizer, as can be understood in the following manner: because the lowermost p layer has a much higher doping concentration, any current traversing the D_1 p–n junction will be predominantly carried by holes. A large fraction of them will traverse the adjacent n layer and, being minority carriers there, also the depletion zone of the reverse-biased diode D_2. In this way, the current from the anode to cathode becomes quite independent of the current through the gate. In other words: a small gate-to-cathode current leads to an *ignition of the thyristor* by setting it into a permanent on state.

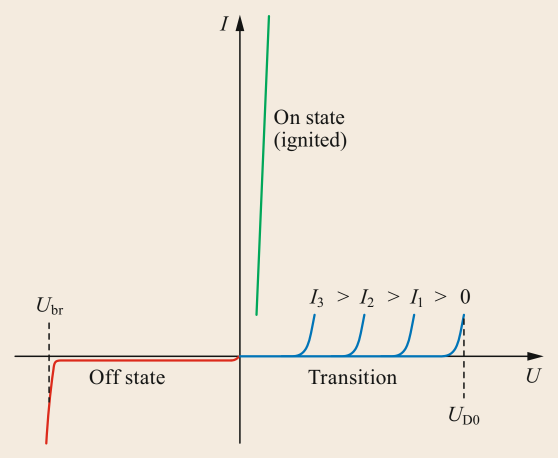

Fig. 26.37 Thyristor current as a function of the voltage between anode and cathode. If the reverse-bias voltage is smaller than the breakdown voltage U_{br}, the current is negligible. The larger the gate currents ($I_1, I_2, \ldots$), the lower the tension needed to ignite the device

In alternating-current networks, thyristors can only be in their on state for at most 50% of the time. For this case, which includes all home applications, the triode for alternating current (TRIAC) has been developed. Figure 26.38 shows that TRIACs can be regarded as antiparallel pairs of thyristors. The thyristor shown on the right of Fig. 26.38 will conduct current from terminal A_1 to terminal A_2, while the left one is responsible for the other direction.

The ignition of a thyristor is a process that cannot be influenced from the outside; in particular, it cannot be sped up by injecting more current into the gate. When timing is crucial, MOS solutions are preferred.

Power MOS transistors can operate at frequencies above 500 Hz, making them particularly suitable for switching power supplies. Figure 26.39 shows a cross section through the most powerful of these, an enhancement n-type vertical double-diffused metal–oxide–semiconductor (VDMOS) field-effect transistor. Seen from the top, the gate would appear as a disc, surrounded by a circular source. When U_{GS} is sufficiently large, an n-type inversion channel forms in the p region underneath the gate, thus forming a conducting channel to the central n region. Below that region, a highly doped n region helps to keep the contact resis-

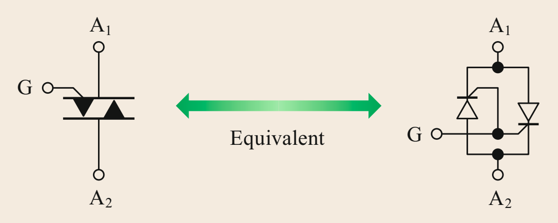

Fig. 26.38 The schematic symbol (*left*) and internal structure (*right*) of a TRIAC

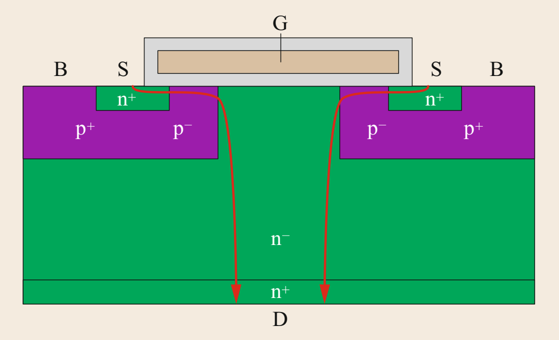

Fig. 26.39 Cross section through a VDMOS transistor with gate (G), source (S), body (B), and drain (D) terminals. The *arrows* indicate the path of electrons moving towards the drain

tance to the drain terminal at a low level. The length of the channel is not determined by the size of gate, but rather by the geometry of the underlying diffusion regions. Usually, the body and source are attached to the same terminal, thus ensuring that there will always be a sufficiently large depletion zone between them.

As can be seen in Fig. 26.40, the characteristic of a power MOS device resembles that of a MOS transistor, the only exception being the larger saturation currents due to the junction areas' being many orders of magnitude larger.

When used for switching a large amount of power, the gate potential must change very quickly. In technical terms, if $P_{\text{loss}} = U_{\text{DS}}^2/R$ is to be small, then U_{DS} must reach a small value as quickly as possible. However, as long as the transistor is in the off state, there will be a large tension between the source and drain. If the power losses within the device are to be kept at a tolerable level, this tension has to drop fast, allowing the transition from the cut-off mode to active mode (incomplete channel) to triode mode to take place as fast as possible. Finally, in the triode mode, the remaining resistance $R_{\text{DS}}(\text{on}) = U_{\text{DS}}/I_{\text{D}}$ is mainly determined by the channel resistance and the conductivity of the low-doped drain region.

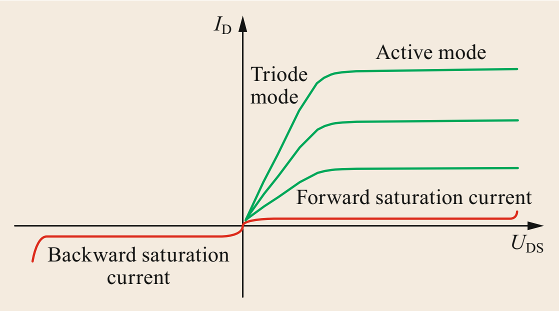

Fig. 26.40 Drain current as a function of the drain–source voltage of a VDMOS power transistor. The forward saturation current corresponds to $U_{\text{GS}} < U_{\text{Th}}$, and each line in the first quadrant corresponds to a fixed value of U_{GS}

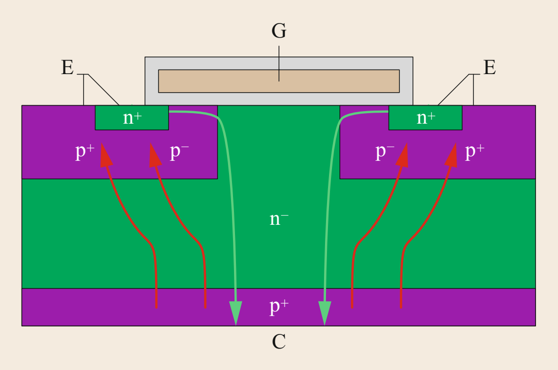

Fig. 26.41 Cross section through an IGBT with emitter (E), gate (G), and collector (C) terminals

Often, entire silicon wafers are turned into massive parallel transistor arrays with densities of up to 800 000 transistors per cm^2. All the transistors together act like a single large transistor. Its power is limited by the power density in the channel, which is small, but has to carry the entire current. This limitation is overcome by the next technological advance.

A small technological modification turns a VDMOS transistor into another powerful semiconductor device, the *insulated-gate bipolar transistor* (IGBT). Figure 26.41 shows that it may be derived from a VDMOS transistor by replacing the lowermost n-doped area by a p-doped area. In this manner, another forward-biased p–n junction is introduced. This extra junction works like the afterburner of a jet engine. If, for example, the extra p layer has a doping concentration that is ten times higher than that in the n region above, for purely statistical reasons, each electron crossing the junction will pass about ten holes moving in the opposite direction. And because the number of electrons passing is entirely determined by the gate–source–body setup, the extra p–n junction will increase the current by a factor of $1 + 10 = 11$.

With the central n region having a sufficiently low doping concentration, most of the holes will reach the body, which is connected to the emitter terminal. The extra current due to holes coming from the collector terminal is particularly helpful because it does not have to pass through the channel underneath the gate. And a current distributed over a larger volume is equivalent to less concentrated heat production.

However, there is a price to pay for the efficient use of silicon in the production of IGBTs. The increased switch-off time compared with VDMOS devices is the first drawback of an IGBT. It is imposed by the large number of minority carriers in the central n region in the on state. As all p–n junctions are transparent, an IGBT is only switched off after all the holes have left this region. The second drawback is the voltage drop across the forward-biased extra diode. A drop of 0.7 V exposed to a current of 1000 A will create a power loss of 700 W in the device. Nevertheless, for many power supply applications, IGBTs combine the best of the bipolar world with the best of the MOS world.

26.4 Networks

In this section, different types of networks are introduced. An explanation of technical terms used in networking is followed by a discussion of reference directions and the corresponding convention with respect to power. Kirchhoff's rules are presented in the context of conservation laws. It is shown how these rules form the basis of various network analysis methods. Finally, the properties of three-phase alternating-current networks are discussed.

26.4.1 Network Terminology and Reference Directions

Understanding networks is the first step towards understanding electrical and electronic systems. With up-to-date electronic systems having more than one billion components, computing techniques that can be automated are needed. In this section, the underlying rules are presented as well as some of the techniques themselves.

Schematics of electrical systems not only show the components and connections, they often show little arrows (often in red or blue) which seem to indicate directions. Such arrows are *not vectors*. They indicate a *reference direction*. Mathematically, they are simply a one-dimensional coordinate system. Thus, a red arrow (commonly used for currents) pointing to the right indicates that, if $I > 0$, the current also flows from left to right, while if $I < 0$, the current also flows in the opposite direction. A blue arrow (used for tensions) pointing from top to bottom corresponds to higher potential being at the top if $U > 0$.

There may be situations in which, at the same component, the reference directions for current and voltage are antiparallel. Physically, this is a modeling mistake: it means that two quantities (current and tension) are calculated in different coordinate systems. However, as shown in Fig. 26.42, there is an easy way to get it right again. Whenever the reference directions are antiparallel, the corresponding element equations need to be given an extra minus sign.

One might argue that these sign swaps are an artificial complication of the matter, as one might establish the rule that all reference directions should always be the same. Figure 26.43 may help to understand why electrical engineers have followed a different route. Each circuit gains energy from somewhere (battery, generator, ...) and loses it somewhere else (resistor). The gain of energy is always connected to a movement *against a force*. In mechanical examples, this is obvious. However, when capacitors or batteries are charged, the same principle applies to charge carriers: If a component delivers energy to a circuit, the current passing through will flow from the lower potential to the higher potential, i.e., *antiparallel to the tension*. Therefore, electrical engineers have adopted the convention that, for energy generators, the reference directions for current and tension should be antiparallel. If this convention is followed, the values for both the current and tension are positive in an energy source. However, as long as the rule indicated in Fig. 26.42 is followed, any combination of reference directions will also give correct results.

Networks are described using the technical terms illustrated in Fig. 26.44. Tensions may be measured between pairs of *terminals*. A *node* is characterized by having a certain potential. The connection between two nodes is called a *branch*. If several nodes are connected in series, the result is still a branch, unless the connection ends in a closed loop, as shown in Fig. 26.45. If (and only if) this *loop* does not contain another loop, it is called a *mesh*.

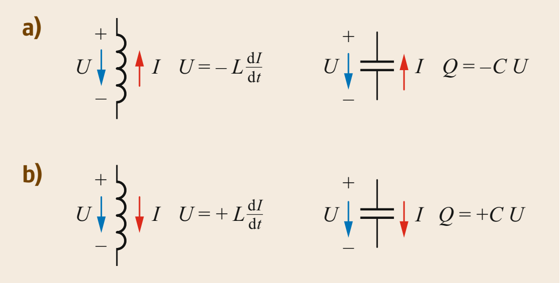

Fig. 26.42a,b The influence of reference directions of tension and current on the signs of element equations. Here, the arrows indicate the reference directions, representing one-dimensional coordinate systems for currents and tensions. If current and tension have the same reference direction, the equations have a + sign. **(a)** Generator, **(b)** consumer

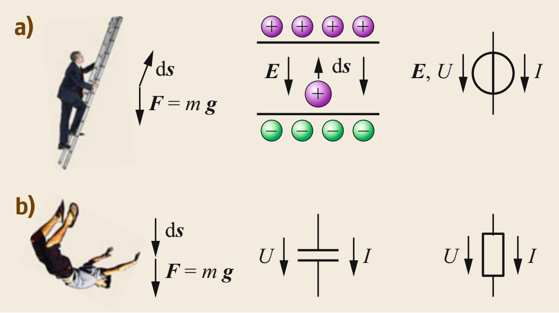

Fig. 26.43a,b Three examples for the transfer of energy. Potential energy is gained if there is a movement *against* a force. **(a)** Energy generation, gain, **(b)** energy consumption, loss

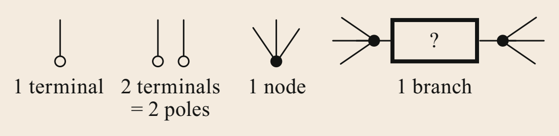

Fig. 26.44 Illustration of the meaning of electrical terminals, nodes, and branches

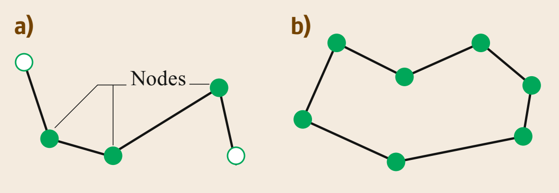

Fig. 26.45a,b Representation of a branch **(a)** and a mesh **(b)** made from nodes, terminals, and branches

26.4.2 Kirchhoff's Rules

The rules established by Gustav Robert Kirchhoff in the 19th century can be traced back to elementary conservation laws. They still form the basis of all methods to determine currents and potentials in a network.

Conservation of charge is the basis of Kirchhoff's first rule, the *node rule*. It is illustrated in Fig. 26.46, and it states that the sum of all currents flowing into a node must be zero:

$$\sum_i I_i = 0\,. \tag{26.83}$$

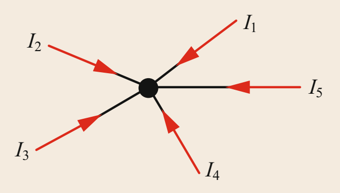

Fig. 26.46 Illustration of the node rule. If all the reference directions point into the node, the sum of all the currents must be zero

Clearly, this rule assumes that all reference directions point into the node (they might also all point outwards). It also implies that, at each node, there must be at least one current whose value is negative. A slight reformulation helps to find the relation to charge: For every node, the incoming currents must be balanced by the outgoing ones—just like a river has to carry all water that is supplied from smaller rivers at a river junction (Fig. 26.47). In this manner, it is guaranteed that neither water nor charge is lost or added.

Generally, in an open network with a number of k nodes, each node will have its own equation. Hence, the node rule will produce k equations. The majority of networks, however, are either closed or balanced with respect to the currents coming in and out (Fig. 26.48). This boundary condition makes one of the equations obsolete, leaving $(k-1)$ independent equations. Luckily, it is completely irrelevant which of the k node equations is left out for a determination of the currents in a closed network.

Conservation of energy is the basis of Kirchhoff's second rule, the *mesh rule*. As the static electric field is a conservative one, any charge traveling from one point to another and back again on a different path must have the same energy as at the beginning of the journey. As a consequence, any closed-loop movement back to the starting point may neither add nor subtract energy from the carrier of charge. Recalling that the electric potential is the potential energy per charge and that tension is

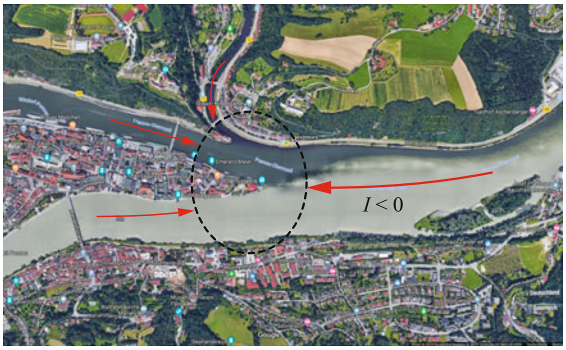

Fig. 26.47 Joining of three rivers in Passau, Germany. Analogously to the node rule, the value for the water flow on the *right* must be the negative sum of the flow of the three supplying rivers that can be seen on the *left* (©2019, Google, map data ©GeoBasis-DE/BKG (©2009))

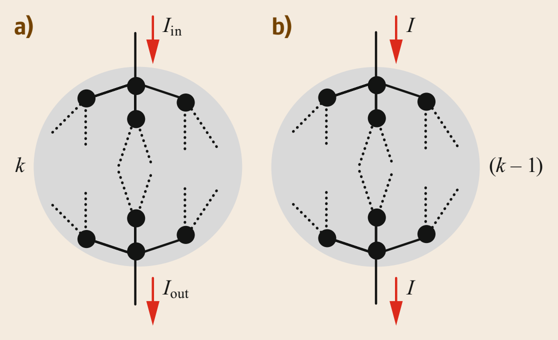

Fig. 26.48a,b Explanation of the number of node equations in a network. **(a)** In general, k nodes give k equations. **(b)** For networks with balanced input and output currents, there is one equation fewer. The latter includes closed networks ($I = 0$)

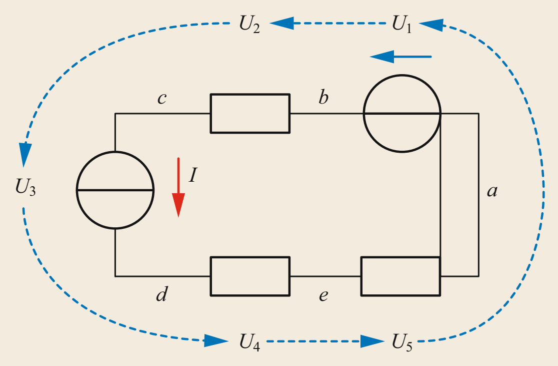

Fig. 26.49 Illustration of the mesh rule. If the reference directions for tensions form a closed loop, the sum of all tensions has to be zero

defined to be the difference of potentials, Kirchhoff's second rule seems self-evident, as it states that the sum of all tensions in a mesh is zero:

$$\sum_i U_i = 0\,. \tag{26.84}$$

An illustration of the rule is shown in Fig. 26.49. In a network with m meshes, the rule supplies m equations. This number strictly refers to the meshes, not to all loops.

While the validity of the node rule is limited by the alternating-current behavior of parasitic capacitors (e.g., long parallel wires), the use of the mesh rule is limited to the case where no strong alternating magnetic fields are present (see (26.25) and its explanation). The latter indicates a severe problem: electrical and electronic systems are developed by assuming that the mesh rule applies. However, in the case of strong electromagnetic wave interference, the potentials start to oscillate, and this may lead to a malfunction or even destruction of the exposed device. This is why "please switch off all electronic devices" is an instruction commonly heard before takeoff.

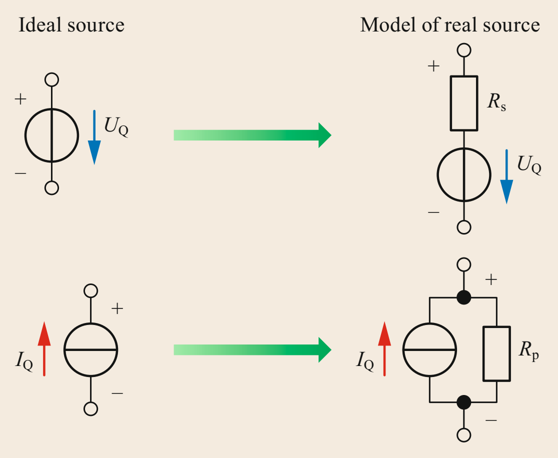

Fig. 26.50 Model of a nonideal voltage source (*top, right*) and model of a nonideal current source (*bottom right*)

In summary, one may conclude that, for a network with k nodes and m meshes, Kirchhoff's rules provide $k + m$ equations. If the network is closed, $(k - 1) + m$ equations remain.

26.4.3 Nonideal Voltage and Current Sources

An *ideal voltage source* is a component that will deliver a constant voltage irrespective of the current passing through it. In reality, this can only be an approximation. The more current that flows, the smaller the tension will be. In almost all cases, it is sufficient to model such a nonideal voltage source as a series combination of an ideal source and an ohmic resistor, as shown in Fig. 26.50. Similarly, a real current source can be approximated by an ohmic resistor placed in parallel to an *ideal current source*. The voltage and current at the terminals of the real (nonideal) sources are then

$$\begin{aligned} U &= U_Q - R_s I \qquad \text{(real voltage source)}\,, \\ I &= I_Q - \frac{U}{R_p} \qquad \text{(real current source)}\,. \end{aligned} \tag{26.85}$$

As both real source models produce a linear dependence between voltage and current, it is always possible to exchange the two, as shown in Fig. 26.51. In fact, with

$$R_p = R_s \quad \text{and} \quad U_Q = R_S I_Q\,, \tag{26.86}$$

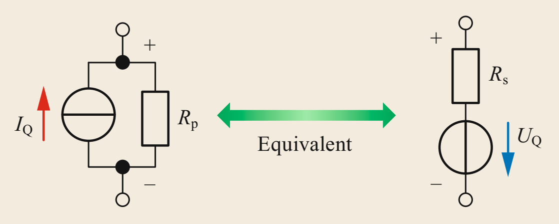

Fig. 26.51 Illustration of the fact that, for every real (linear) voltage source, a current source with the same behavior can be found

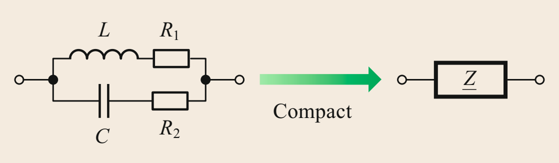

Fig. 26.52 Compaction of an electrical branch

one gets exactly the same behavior at the terminals of both sources. And in both cases, the value of the internal resistance R_i may by determined in the same manner. If the source is exposed to a variable resistor, a simultaneous measurement of the current and tension of either source will reveal its value, because

$$R_i = R_p = R_s = \frac{\Delta U}{\Delta I} = \frac{dU}{dI} \qquad (26.87)$$

is a direct consequence of (26.86).

When analyzing networks, it is always a good idea to replace internally complicated branches not only by real sources, but also by impedances, including the ones of inductors and capacitors. An example of such a compaction is shown in Fig. 26.52. For the analysis within a larger network, an entire branch can be replaced by a single impedance $\underline{Z}$. When the current and tension through $\underline{Z}$ are known, its components may by looked at in detail. In the case shown, one would have $\underline{Z} = (j\omega L + R_1) \parallel [1/(j\omega C) + R_2]$.

Application Example: Replacing Any Two-Terminal Network by Real Sources

A two-terminal network consisting of only sources and resistors has a linear current–voltage characteristic. It can therefore always be replaced by a real current source or a real voltage source. Figure 26.53 indicates a method to find these sources. The values for I_Q and U_Q can simply be determined by analyzing the behavior of the circuit for the two cases where the terminals are connected (short circuit) or disconnected (open circuit). This may be understood by a glance at Fig. 26.50. The behavior of the open circuit is independent of R_s, while R_p has no influence in the short-circuit current source. The method to determine the value for the internal resistance R_i can be understood to be a consequence of (26.87). An ideal current source has an infinite internal resistance. For the determination of R_i, it is replaced by an open switch, i.e., no connection. An ideal voltage source has $R_i = 0$ and is therefore replaced by a connection (short circuit). Both replacements together give the circuit shown on the top right of Fig. 26.53. An analysis of this circuit gives the value of R_i.

For the circuit shown, one thus gets

$$\begin{aligned} R_i &= (R_1 \parallel R_2) + R_3 \,, \\ I_Q &= \frac{1}{R_3}\left(I_B + \frac{U_B}{R_1}\right)(R_1 \parallel R_2 \parallel R_3) \,, \\ U_Q &= \frac{U_B + R_1 I_B}{1 + (R_1/R_2)} \,. \end{aligned} \qquad (26.88)$$

26.4.4 Network Analysis Algorithms

Using Kirchhoff's rules will always produce enough equations to determine all the currents and tensions in a given network. However, as networks get larger, the path towards the solution may become rather clumsy. For a network having k nodes and m meshes, a system of $m + k - 1$ equations has to be solved. And so the repeated emergence of the strategic question "which variable should be eliminated next?" may turn out to be rather nerve-racking.

There are two popular methods that allow the number of equations to be reduced. And, as long as linear elements such as R, L, C, and sources are the only components, these methods help to write down the equations in a manner that produces a system of equations that can be solved using matrix inversion. In the following, the methods will be presented with sources and resistors being the only components. All methods may, however, also be applied to alternating-current networks. In this case, impedances $\underline{Z}$ have to be used instead of resistances R.

The first algorithm, the technique of *mesh analysis*, also known as the *mesh current method*, reduces the number of equations to m by writing down all the mesh equations in a manner that makes the use of the node equations obsolete. Its principal idea is to interpret all the currents in a network as sums of so-called *mesh currents*. As shown in Fig. 26.54, the current through any component is written down as the sum of the currents of all meshes of which it is a part. So, the current through the resistor R_C in Fig. 26.54 would be $I_{M1} + I_{M2}$. Closer inspection also reveals that computing this sum is in fact the incorporation of the node rule. If the component is

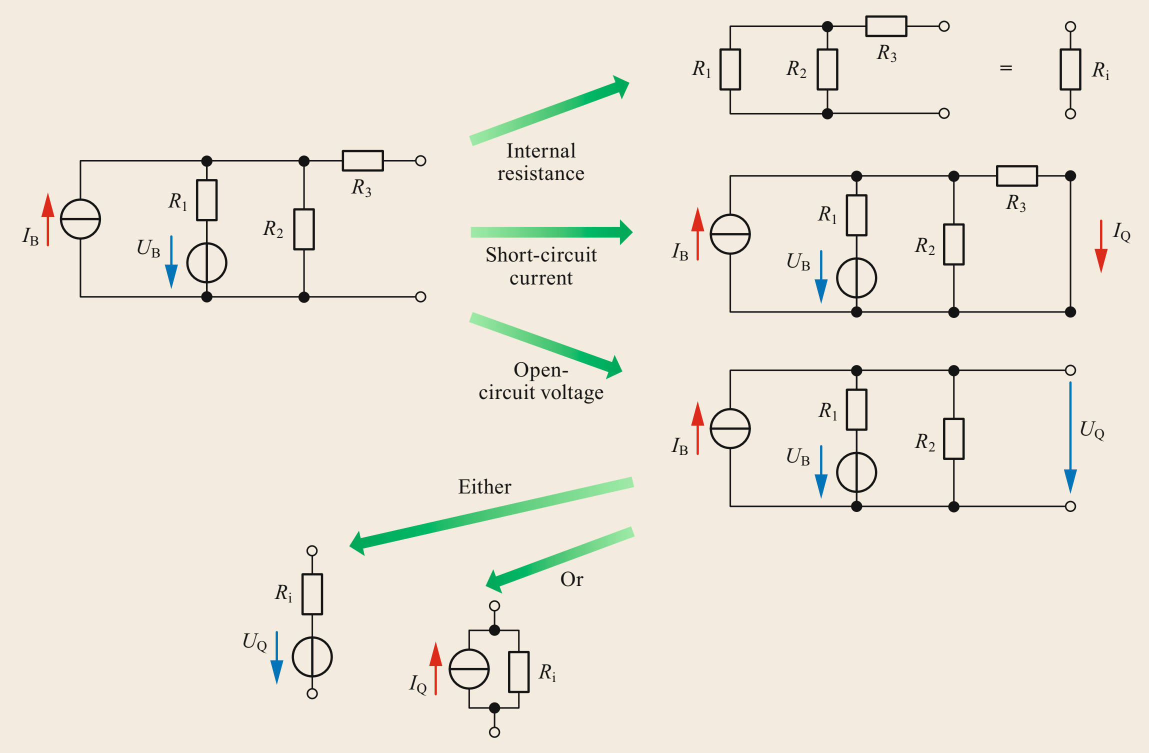

Fig. 26.53 Example for the replacement of a two-terminal network by either a real voltage source or a real current source

part of one mesh only (R_1 and R_2 in Fig. 26.54), the current of the mesh and the current through the component are identical. Clearly, if all the mesh currents are known, the currents through all the components are known as well.

Applying the mesh rule to the complete network is done in two steps. First, the meshes are considered one by one, then in a second step, all voltages from adjacent meshes are added. Hence,

$$(R_1 + R_C + R_2)I_{M1} + R_C I_{M2} = U_Q \tag{26.89}$$

describes the voltage balance for mesh M_1 in Fig. 26.54. The next step is to write down the equations for all the meshes in such a manner that they may easily be transformed into a matrix equation such as the following one:

$$\begin{pmatrix} R_{11} & R_{12} & \dots \\ R_{21} & R_{22} & \dots \\ \dots & \dots & \dots \end{pmatrix} \cdot \begin{pmatrix} I_{M1} \\ I_{M2} \\ \dots \end{pmatrix} = \mathbf{R} \cdot \begin{pmatrix} I_{M1} \\ I_{M2} \\ \dots \end{pmatrix} = \begin{pmatrix} U_{Q1} \\ U_{Q2} \\ 0 \\ \dots \end{pmatrix}, \tag{26.90}$$

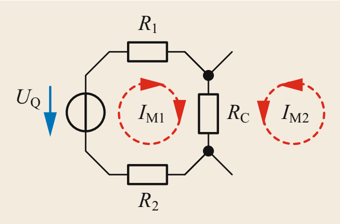

Fig. 26.54 The principle of mesh analysis. First, all the *mesh currents* (I_{M1}, I_{M2}) are computed. Then, the currents through the components are known to be the sums of their mesh currents

i.e., the *resistance matrix* **R** *times the vector of mesh currents equals the vector of all the source voltages.* The indices in the resistance matrix refer to the meshes of which the resistors are part. Clearly, this preparation requires all voltage sources to appear on the right-hand side of the equations, and that all the mesh currents be written down in the same order in all the equations. Meshes with no voltage source have a 0 on the right. Each diagonal element of the resistance matrix is the sum of all the resistors in a mesh. The off-diagonal elements are the coupling resistors. If a mesh M_x has no common resistor with a different mesh M_y, one has $R_{xy} = 0$.

All elements of the resistance matrix may also be copied directly from the schematic of a network. The diagonal elements are the sums of all the resistances in a mesh. The off-diagonal elements are either zero, or the coupling resistances, or—if the two mesh currents through a resistor are antiparallel—the negative resistance. A check of correctness can then be made by symmetry, i.e., by confirming that $R_{ij} = R_{ji}$.

All mesh currents may now be computed via inversion of the resistance matrix $\mathbf{R}$. Multiplication of (26.90) by $\mathbf{R}^{-1}$ from the left leaves

$$\begin{pmatrix} I_{\mathrm{M}1} \\ I_{\mathrm{M}2} \\ I_{\mathrm{M}3} \\ I_{\mathrm{M}4} \\ \dots \end{pmatrix} = \mathbf{R}^{-1} \cdot \begin{pmatrix} U_{\mathrm{Q}1} \\ U_{\mathrm{Q}2} \\ U_{\mathrm{Q}3} \\ 0 \\ \dots \end{pmatrix} ; \tag{26.91}$$

i.e., all mesh currents can be read directly from the result.

There is, however, one set of networks that cannot be handled by such mesh analysis, viz. all networks with at least one ideal current source. Real current sources are no problem, as they may be replaced by real voltage sources. The method shown next can do so, albeit at the price of not being able to handle ideal voltage sources.

The branch current method, also known as *nodal analysis* or *node-voltage analysis*, reduces the number of equations from $m+k-1$ to $k-1$ by automatically incorporating all the mesh equations. Figure 26.55 illustrates the idea. First, the current balance for each node is written down separately. Then, all the resulting equations are put in an order that allows them to be transformed into a matrix equation of the type *conductance matrix times vector of unknown tensions equals vector of given current sources*.

The behavior of a network does not change if all the potentials are augmented by the same, fixed amount. Only potential differences, i.e., voltages, matter. The easiest way to use this freedom is to set one of the potentials to zero (*mass*) and give it the node number zero. For the node shown in Fig. 26.55, the appropriate form to write down the node rule would then be

$$\begin{aligned} & I_1 + I_2 + I_3 + I_{\mathrm{Q}} = 0 \\ & \to \frac{U_{10} - U_{x0}}{R_1} + \frac{U_{20} - U_{x0}}{R_2} + \frac{U_{30} - U_{x0}}{R_3} + I_{\mathrm{Q}} = 0 \\ & \to -\frac{1}{R_1} U_{10} - \frac{1}{R_2} U_{20} - \frac{1}{R_3} U_{30} \\ & \quad + \left(\frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} \right) U_{x0} = I_{\mathrm{Q}} \\ & \to -G_{x1} U_{10} - G_{x2} U_{20} - G_{x3} U_{30} \\ & \quad + (G_{x1} + G_{x2} + G_{x3})\, U_{x0} = I_{\mathrm{Q}} \, . \end{aligned} \tag{26.92}$$

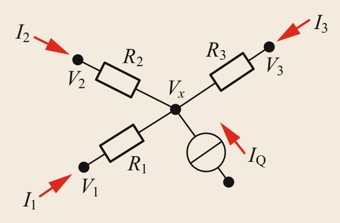

Fig. 26.55 Basic principle of nodal analysis: For each node, the current balance is written down separately. Later, all the equations for all the nodes are combined to form a matrix equation

Here, the indexes of the conductances indicate which nodes they connect; e.g., G_{x1} connects V_x and V_1. All currents entering a node are transformed into potential differences multiplied by the inverse resistances of the components on the branch pointing to the node (second line in (26.92)). Then, all the terms are ordered according to the node numbers. The current sources are moved to the right-hand side of the equation. Finally, for the sake of having easier-to-read formulas, the inverse resistances are replaced by conductances (fourth line in (26.92)). These rearrangements thus result in a form that can be interpreted as being one line of a matrix equation. A closer look at this line

$$\begin{aligned} & (-G_{x1},\ -G_{x2},\ -G_{x3},\ [G_{x1} + G_{x2} + G_{x3}]) \cdot \begin{pmatrix} U_{10} \\ U_{20} \\ U_{30} \\ U_{x0} \end{pmatrix} \\ & = I_{\mathrm{Q}} \end{aligned} \tag{26.93}$$

again reveals a certain structure of the conductance matrix. The diagonal elements are the sums of all the conductances along the branches connected to the nodes (i.e., $[G_{x1} + G_{x2} + G_{x3}]$ in (26.93)). The off-diagonal elements are the conductances between the nodes, equipped with an extra minus sign. If all but one (i.e., $k-1$) of the node equations (the ground node gets no equation) are written down in this manner, they can be summarized into the following matrix equation:

$$\begin{aligned} & \begin{pmatrix} G_{11} & G_{12} & \dots & G_{1(k-1)} \\ G_{21} & G_{22} & \dots & G_{2(k-1)} \\ \dots & \dots & \dots & \dots \\ G_{(k-1)1} & G_{(k-1)2} & \dots & G_{(k-1)(k-1)} \end{pmatrix} \cdot \begin{pmatrix} U_{10} \\ U_{20} \\ \dots \\ U_{(k-1)0} \end{pmatrix} \\ & = \begin{pmatrix} I_{\mathrm{Q}1} \\ I_{\mathrm{Q}2} \\ \dots \\ I_{\mathrm{Q}(k-1)} \end{pmatrix} . \end{aligned} \tag{26.94}$$

Inversion of the conductance matrix then gives then all the voltages relative to the ground potential. Multiplica-

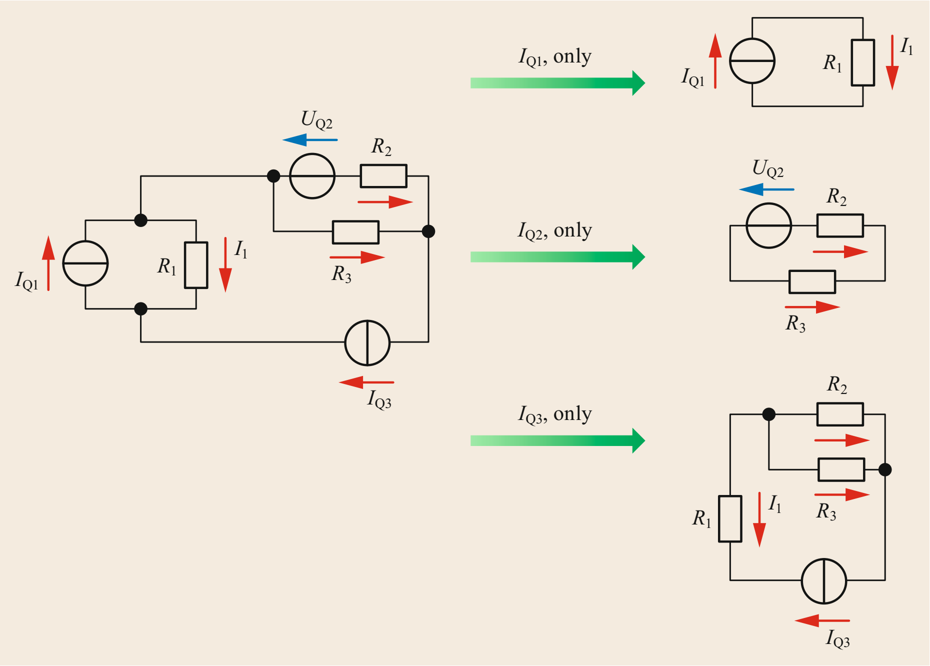

Fig. 26.56 Example for the determination of currents by separating and then adding the contributions from all sources. *Left*: original circuit, *right*: the three circuits used for the determination of the current contributions

tion of (26.94) with $\mathbf{G}^{-1}$ does this, as

$$\begin{pmatrix} U_{10} \\ U_{20} \\ \dots \\ U_{(k-1)0} \end{pmatrix} = G^{-1} \cdot \begin{pmatrix} I_{Q1} \\ I_{Q2} \\ \dots \\ 0 \\ 0 \end{pmatrix} \qquad (26.95)$$

shows.

Finally, both nodal analysis and mesh analysis can be combined with a further method, the *superposition of current contributions*. This method is particularly helpful if there are many sources in a network. Figure 26.56 illustrates the principle. For each source, the currents through all the components are calculated separately. For a given component, the total current is then the sum of the currents thus determined. This method is similar to the replacement of two-terminal networks by real sources: an ideal current source not under consideration is replaced by an open switch (no connection), while an ideal voltage source is replaced by a short circuit (connection). In this way, the three sources in the network in Fig. 26.56 yield the three networks shown on the right of this figure. The currents may then be read from the schematics of the networks:

$$\begin{aligned} I(R_1) &= I_{Q1} + 0 \qquad\qquad - I_{Q_3} \,, \\ I(R_2) &= 0 \quad + \frac{U_{Q2}}{R_2 + R_3} + \frac{I_{Q3}(R_2 \parallel R_3)}{R_2} \,, \\ I(R_3) &= 0 \quad - \frac{U_{Q2}}{R_2 + R_3} + \frac{I_{Q3}(R_2 \parallel R_3)}{R_3} \,. \end{aligned} \qquad (26.96)$$

The current contributions in (26.96) are ordered in the same manner as the networks in Fig. 26.56. The more sources there are, the more drastic the simplifications become.

Table 26.4 presents a comparison of the network analysis algorithms. The compaction of two-terminal networks includes their replacement by real sources. It may be combined with node analysis or mesh analysis. The same holds for the method of current superposition.

Table 26.4 Algorithms for analysis of linear networks

Method	Advantages	Disadvantages
Kirchhoff's rules	Always work	Many equations $(m+k-1)$
Mesh analysis	Few equations (m)	No ideal current sources
Node analysis	Few equations $(k-1)$	No ideal voltage sources
Current superposition	Simplified networks	One network computation per source
Two-terminal compaction	Simplified networks	Not always a great help

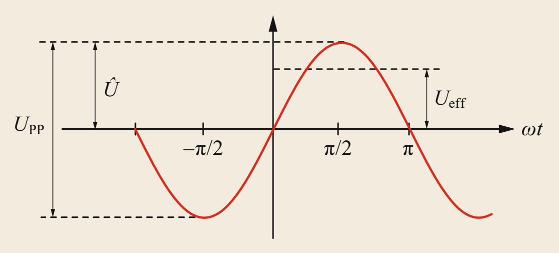

Fig. 26.57 Commonly used variables to describe alternating voltages

26.4.5 Alternating-Current Networks

All the algorithms presented so far may also be applied to alternating-current networks. In this manner, they include analyses of networks with capacitors and inductors. For these applications, the resistances R_i should be replaced by impedances $\underline{Z}_i$ (see below). For simplified calculations, alternating sources should also be represented using complex variables.

There are a variety of representations for alternating currents and voltages. Figure 26.57 shows the most popular ones. Digital oscilloscopes will determine the *peak-to-peak voltage* U_{PP} by subtraction of the smallest value from the largest one. Ideally, this is twice the value of the amplitude $\hat{U}$, known as the factor before the sine function. Books on AC analyses often use the *effective voltage*

$$U_{eff} = \frac{\hat{U}}{\sqrt{2}} , \tag{26.97}$$

because an alternating voltage with $U_{eff} = x\,\mathrm{V}$ will heat up a resistor to exactly the same temperature as a direct-current voltage of $x\,\mathrm{V}$. Many books on AC networks leave out the suffix $_{eff}$, leaving some uncertainty as to whether or not factors of $\sqrt{2}$ have been included. For this reason, in this book, effective voltages are always labeled U_{eff}.

Within AC networks, resistors, inductors, and capacitors are characterized by their impedances:

$$\begin{aligned} \underline{Z}_R &= R && \text{(resistor)} , \\ \underline{Z}_L &= \mathrm{j}\omega L && \text{(inductor)} , \\ \underline{Z}_C &= \frac{1}{\mathrm{j}\omega C} && \text{(capacitor)} , \end{aligned} \tag{26.98}$$

which happen to be either real (R), positive imaginary (L), or because of $1/\mathrm{j} = -\mathrm{j}$ negative imaginary (C). These characteristics form the basis of a rather illuminating way of adding impedances. Each of the components is represented as a vector in one of the complex planes for $\underline{Z}$ or $\underline{Y}$, as shown in Table 26.5. A series connection is then represented by the addition of vectors in the $\underline{Z}$ plane, while a parallel connection is represented by an addition of vectors in the $\underline{Y}$ plane. Figure 26.58 shows the example $\underline{Z} = \underline{Z}_R + \underline{Z}_L + \underline{Z}_C$.

Table 26.5 Impedance vectors for linear components

Component	Resistor	Inductor	Capacitor
Vector in the $\underline{Z}$ plane	$(R, 0)$	$(0, \mathrm{j}\omega L)$	$(0, -\mathrm{j}/(\omega C))$
Vector in the $\underline{Y}$ plane	$(1/R, 0)$	$(0, -\mathrm{j}/(\omega L))$	$(0, \mathrm{j}\omega C)$

Whenever large amounts of energy are to be transported, losses along lines such as those shown in Fig. 26.65 must be minimized. AC currents may produce losses even if no energy is transferred at all. This apparent contradiction results from the fact that, for a component with a 90° angle between current and tension, on average, the energy transfer into the component is exactly balanced by the transfer out of the component. At the same time, the current on the power line is nonzero. For such a component, i.e., one with a 90° angle between current and tension, the current is

$$\begin{aligned} \underline{i} &= \frac{\underline{u}}{\mathrm{j}X} = \frac{\hat{U}}{\mathrm{j}X} \mathrm{e}^{(\omega t - \pi/2)} \\ \rightarrow \quad i &= -\frac{\hat{U}}{X} \cos\left(\omega t - \frac{\pi}{2}\right) , \end{aligned} \tag{26.99}$$

i.e., not zero for most times. Interestingly, the average magnitude of the current

$$\langle |i| \rangle = -\frac{2\hat{U}}{\pi X} \tag{26.100}$$

is exactly as large as the corresponding average for an ohmic resistor of equal magnitude: $\langle |i| \rangle = -2\hat{U}/(\pi R)$. In fact, it turns out that the angle between the voltage and current does not matter at all for the loss in supply lines.

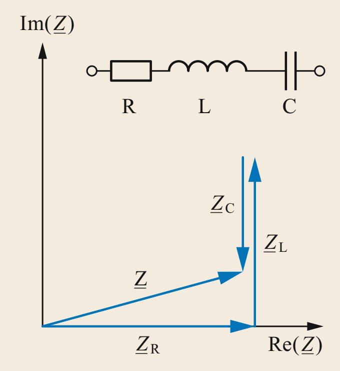

Fig. 26.58 Example of addition of impedances in the $\underline{Z}$ plane, showing the graphical determination of the impedance of an RLC oscillator circuit

Fig. 26.59 Electric cooker. Good ones are equipped with reactive power compensation capacitors (photo: © Oleksandr Delyk/stock.adobe.com)

The *reactive power* Q is a measure of the power that oscillates back and forth on a power line without ever ending. Hence, only P suits the purpose of a power transmission line. For this reason, electrical energy engineers will try to minimize the reactive power. The method they use is called *reactive power compensation*.

Application Example: Reactive Power Compensation in a Cooker

The heating spiral of an electric cooker has an ohmic resistance of $R = 20\,\Omega$ and an inductance of $L = 30\,\text{mH}$. It is attached to a 60 Hz network. How can the reactive power be reduced to zero?

Clearly, the only way to compensate the reactive power of the inductance is to add a capacitor, as shown in Fig. 26.60. The task is now to find a value for the capacitance that makes the reactive currents vanish, i.e., to find $\text{Im}(\underline{Y}) = 0$ for

$$\underline{Y}_{\text{including }C} = \frac{1}{R + \text{j}\omega L} + \text{j}\omega C\,.$$

This is achieved by choosing

$$C = \frac{L}{(R^2 + \omega^2 L^2)} = 57\,\mu\text{F}\,.$$

Then, one has

$$\begin{aligned}\underline{Y}_{\text{including }C} &= \text{Re}(\underline{Y})\\ &= \frac{R}{R^2 + \omega^2 L^2}\,.\end{aligned}$$

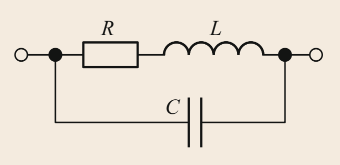

Fig. 26.60 Example of reactive power compensation. A heating resistor R with a parasitic inductance L will have its reactive power compensated by a capacitor of appropriate capacitance C

Now, the reduction of current can be computed. The factor by which the current is reduced is the ratio of the moduli of the acceptances with and without the capacitor:

$$\begin{aligned}\frac{\hat{I}_{C=0}}{\hat{I}_{\text{including }C}} &= \sqrt{\frac{|\underline{Y}_{C=0}|^2}{|\underline{Y}_{\text{including }C}|^2}} = \frac{1}{\sqrt{1+\frac{\omega^2 L^2}{R^2}}}\\ &= 0.87\,,\end{aligned}$$

so that the current is reduced by 13%, and there is no price to pay in terms of performance. As long as the losses on the powering network are negligible, the heating power remains unchanged by the capacitor. Otherwise, there may even be a tiny increase of power.

26.4.6 Transformers

The success of AC applications is for exactly one reason: there is an easy way to transform the voltages of AC systems. Figure 26.61 shows that the basic setup of a *transformer* is simple: two coils are wound around a common iron core. Essentially, one of the coils (the *primary coil*) is connected to an alternating voltage. Its alternating current produces a magnetic field that is strengthened and shaped by the iron core. The *secondary coil* is thus exposed to an alternating magnetic

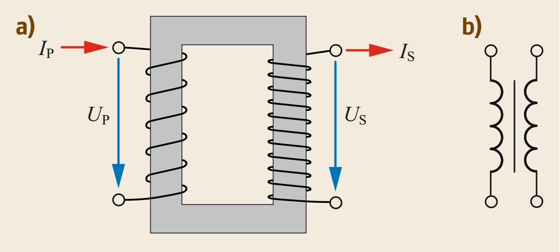

Fig. 26.61a,b A transformer consisting of two coils wound around a common core (a) and its schematic symbol (b)

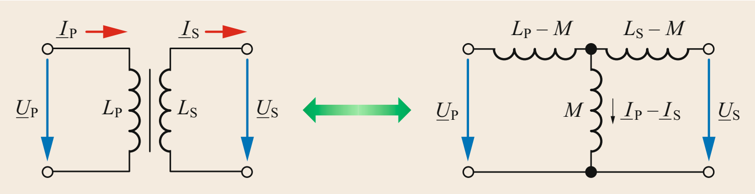

Fig. 26.62 A lossless transformer and a four-terminal network with the same AC behavior

field. Therefore, an alternating voltage will be induced in the secondary coil.

For simplicity, the operation of a transformer is first derived while assuming no losses anywhere. Losses are then introduced in a second step. If the iron core has the same diameter everywhere, each loop carrying the same current I will also produce the same magnetic flux Φ_B:

$$\Phi_B(\text{one loop}) = \Lambda I\,, \tag{26.101}$$

where Λ is some constant. A coil having N loops will then have an inductance

$$L = N^2 \Lambda\,. \tag{26.102}$$

For the setup shown in Fig. 26.61, the magnetic flux will be the sum of the flux Φ_P from the primary coil and that due to the secondary coil, Φ_S:

$$\Phi_B = \Phi_P + \Phi_S = N_P \Lambda I_P + N_S \Lambda I_S\,. \tag{26.103}$$

A change of the magnetic flux will induce a tension $U_{\text{ind}} = \pm \mathrm{d}\Phi_B/\mathrm{d}t$ in each loop. The plus sign applies to the primary coil because the reference directions for current and voltage are the same. On the secondary side, they are antiparallel, so the minus sign applies there. The induced voltages are then given by

$$\begin{aligned} U_P &= N_P \Lambda \left(N_P \frac{\mathrm{d}I_P}{\mathrm{d}t} - N_S \frac{\mathrm{d}I_S}{\mathrm{d}t} \right)\,, \\ U_S &= N_S \Lambda \left(N_P \frac{\mathrm{d}I_P}{\mathrm{d}t} - N_S \frac{\mathrm{d}I_S}{\mathrm{d}t} \right)\,, \end{aligned} \tag{26.104}$$

a set of equations that is known as the *transformer equations*. Because the terms in brackets are the same for U_P and U_S, dividing the two equations gives an equally simple and useful result

$$U_S(t) = U_P(t) \frac{N_S}{N_P}\,; \tag{26.105}$$

in plain terms: the ratio of the voltages is determined by the ratio of the number of turns in the coils. The term $N_P N_S \Lambda = \sqrt{L_P L_S}$, which appears in both equations of the system (26.104), is called the *coupling inductance*, M. Using this quantity, the transformer equations can be written in the form

$$\begin{aligned} U_P &= L_P \frac{\mathrm{d}I_P}{\mathrm{d}t} - M \frac{\mathrm{d}I_S}{\mathrm{d}t}\,, \\ U_S &= M \frac{\mathrm{d}I_P}{\mathrm{d}t} - L_S \frac{\mathrm{d}I_S}{\mathrm{d}t}\,. \end{aligned} \tag{26.106}$$

In general, (26.106) are hard to calculate. However, within the framework of complex alternating-current calculus, derivatives may be replaced by factors. Because $\mathrm{d}/\mathrm{d}t\mathrm{e}^{\mathrm{j}\omega t} = \mathrm{j}\omega\mathrm{e}^{\mathrm{j}\omega t}$, the transformer equations simplify to

$$\begin{aligned} \underline{u}_P &= L_P \mathrm{j}\omega \underline{i}_P - M \mathrm{j}\omega \underline{i}_S\,, \\ \underline{u}_S &= M \mathrm{j}\omega \underline{i}_P - L_S \mathrm{j}\omega \underline{i}_S\,. \end{aligned} \tag{26.107}$$

In this way, complex calculus turns a system of differential equations into a system of linear equations.

Most surprisingly, there is a four-terminal network consisting of three inductances which has exactly the same alternating-current behavior. This network is shown in Fig. 26.62. If a two-terminal network with arbitrary impedance $\underline{Z}_S = \underline{u}_S/\underline{i}_S$ is connected to the secondary side of the transformer, its tension and current may be eliminated from the transformer equations, leaving

$$\frac{\underline{i}_P}{\underline{u}_P} = \frac{1}{\mathrm{j}\omega L_P} + \frac{L_S}{L_P} \frac{1}{\underline{Z}_S}\,, \tag{26.108}$$

i.e., a very simple formula for the admittance of the system. Equation (26.108) shows that, seen from the primary side of the transformer, the latter behaves like a parallel connection of the primary inductor and a load with a modified impedance $\underline{Z}_S(L_P/L_S)$. Finally, the current on the secondary side comes out as

$$\underline{i}_S = \frac{N_P}{N_S} \left(\underline{i}_P - \frac{\underline{u}_P}{\mathrm{j}\omega L_P} \right)\,. \tag{26.109}$$

The average power P_P used by a lossless transformer on the primary side is exactly the same amount as the power delivered to the secondary side: $P_S = -P_P$. However, the reactive powers differ.

In literature, $\underline{i}_S = \underline{i}_P(N_P/N_S)$ may be found instead of (26.109) and referred to as describing an

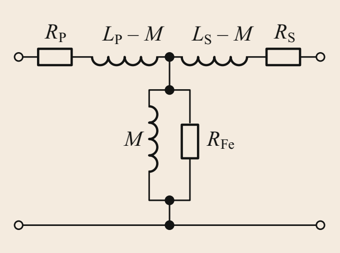

Fig. 26.63 Four-terminal network describing a lossy transformer

ideal transformer. The difference between these two equations results from an additional assumption. The ideal transformer is assumed to have $\mu_r \to \infty$, corresponding to $L_P \to \infty$. If no load is connected to the secondary side, the ideal transformer behaves like an open switch on the primary side, while the lossless transformer behaves like a single inductor with inductance L_P.

A *real transformer* has losses, mainly due to eddy currents and repeated magnetizations in the yoke (iron core), as well as losses in the windings. These can be incorporated into the circuit description as shown in Fig. 26.63. At first glance, it may be irritating that losses within the iron are modeled as a resistor to ground. This resistor should be understood as a feature of the coupling inductor, as it affects both sides of the transformer. Also, as long as the frequency remains unchanged, a suitable value for R_{Fe} can always be found (see also Fig. 26.52). Losses due to stray magnetic fields are usually small compared with the losses just described.

The power that can be transfered by a real transformer is mainly limited by two properties of the iron yoke: remanence and saturation. Both properties become visible if the magnetic field in the iron is plotted as a function of the current in the coil. From Fig. 26.64 one can deduce that an ever-growing current will not produce an equally ever-growing magnetic field. This effect is called *saturation*. It is commonly explained by all of the iron atoms becoming aligned with the field produced by the coil. Saturation has a very disadvantageous consequence, as it produces an almost flat voltage–current characteristic. So, if a transformer is in saturation, a small increase in voltage will provoke a very large additional current. The worst case is then evaporation of the coils. For this reason, transformers are equipped with iron yokes that are large enough to avoid saturation (the larger the yoke, the more atoms to be aligned, and the later saturation will occur).

Remanence describes the fact that a piece of iron remains magnetized even if it is no longer exposed to an external magnetic field. Depending on the direction of the former external field, the remanent field may point in one direction or the other. In physical terms, this means that some of the iron atoms retain their orientation until the external field is strong enough to literally turn them around. This rotation of atoms produces frictional losses in addition to eddy currents. These cannot be avoided, but as in the case of inductors, the eddy currents can be minimized by using stacks of slices of iron whose thicknesses limit the diameter of the eddy currents.

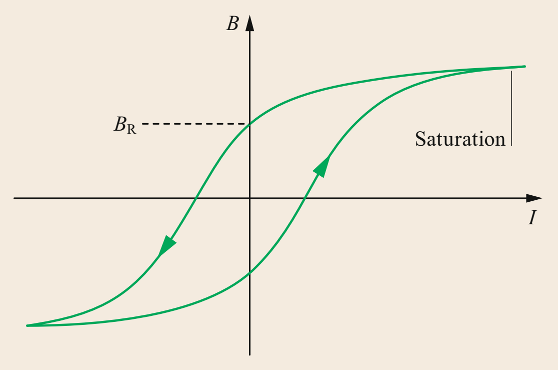

Fig. 26.64 Magnetic field within an iron yoke as a function of the current in the coil around the yoke

26.4.7 Three-Phase Alternating-Current Systems

Power lines often come in bundles of three, as shown in Fig. 26.65, because three-phase AC systems have been proven to be ideally suited for the transport of large amounts of energy. The backbone of each there-phase AC network is a triplet of power lines whose voltages are shifted by an angle of 120°, as shown in Fig. 26.66. Colloquially, each of the lines is called a *phase*.

The amplitudes of the three phases can be written as

$$\underline{U}_1 = \hat{U}\mathrm{e}^0\,, \quad \underline{U}_2 = \hat{U}\mathrm{e}^{\mathrm{j}2\pi/3}\,, \quad \underline{U}_3 = \hat{U}\mathrm{e}^{\mathrm{j}4\pi/3}\,. \tag{26.110}$$

A striking feature of this combination of lines is the fact that the tension between each pair of wires is larger than the tension between each phase and the ground potential. Calculating the instantaneous tension $u_2 - u_1$, for example, gives the tension indicated in red in Fig. 26.67. The corresponding amplitude can be calculated from (26.110). The result

$$\underline{U}_2 - \underline{U}_1 = \hat{U}\left(\mathrm{e}^{\mathrm{j}2\pi/3} - \mathrm{e}^0\right) = \sqrt{3}\,\hat{U}\,\mathrm{e}^{\mathrm{j}5\pi/6} \tag{26.111}$$

shows that the tension between each pair of lines is larger than the tension to ground by a factor of $\sqrt{3}$. The

Fig. 26.65 Masts with 2×3 380 kV power line phases in the upper part, and 2×3 20 kV line phases in the lower part. At the top, a lightning conductor can bee seen. (Photo: © okanakdeniz/stock.adobe.com)

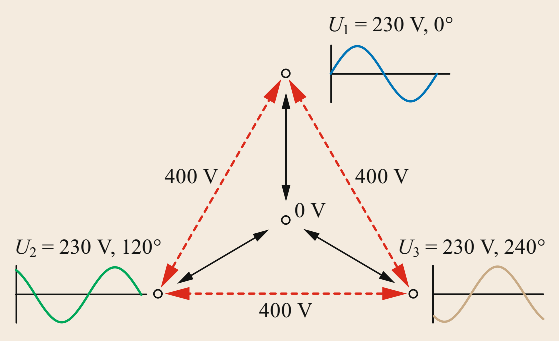

Fig. 26.66 Three power lines forming a three-phase AC network. If each line has an effective tension $V_{\text{eff}} = 230\,\text{V}$, then the effective tension between each pair of lines is $V_{\text{eff}} = 400\,\text{V}$

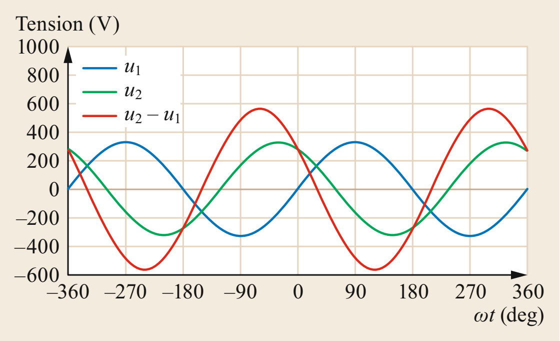

Fig. 26.67 Instantaneous tension between two phases of a three-phase network

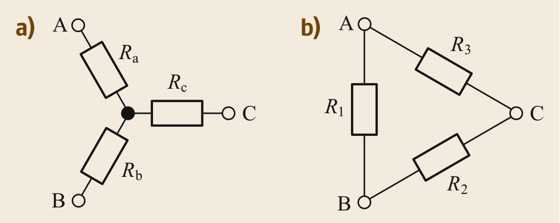

Fig. 26.68a,b Two options to connect three-phase supply lines. The *star* type network (**a**) connects all lines to a ground line, while the *triangle*-type network (**b**) only connects the phases among each other

instantaneous tension is also shifted by an amount of $\omega t = -5\pi/6 = -150°$ relative to $\underline{u}_1$.

This increased tension between two phases can be used in a *triangle network*, as shown on the right in Fig. 26.68. This type of network is particularly popular for high-power devices. Since in a triangle network, there is no connection to ground, the entire energy flow is on the networks with no energy load on the ground. This network is also popular for heating, as the increased tension allows the generation of a given heating power with a reduced current. In this manner, losses on the lines are reduced.

In a home environment, most devices are connected between one of the phases and ground. The typical network of an entire home is therefore rather of the *star* type shown in Fig. 26.68a.

26.5 Electrical Machines

This section starts with a discussion on the interactions between current loops and magnetic fields, i.e., the Lorentz force and induction. It is shown how these interactions can lead to a conversion of mechanical energy into electrical energy or vice versa. A classification of electrical machines follows, then the features of direct-current machines, asynchronous machines, and synchronous machines.

26.5.1 How Wires Are Forced to Move

Electrical machines use the close relations between moving charges (currents) and magnetic fields. Currents produce magnetic fields, and moving charges are subject to the Lorentz force in magnetic fields. An electrical machine is called a *motor* if it converts electrical energy into mechanical energy, and a *generator* if it converts

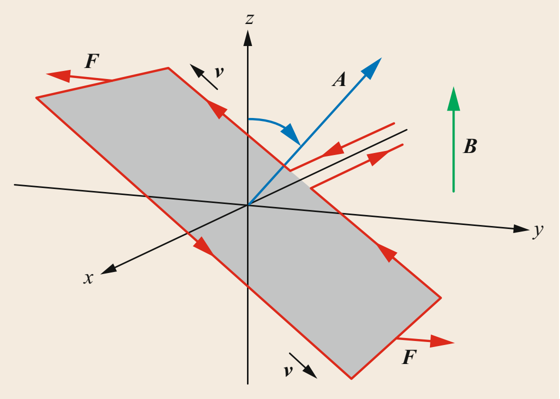

Fig. 26.69 Current loop enclosing an oriented surface $\boldsymbol{A}$ in a magnetic field $\boldsymbol{B}$. Positive charges moving with a velocity $\boldsymbol{v}$, will feel a Lorentz force $\boldsymbol{F}$. This results in a nonzero torque as long as the surface vector is not parallel to the magnetic field

mechanical energy into electric energy. Most electrical machines may be used in either way, i.e. as motors or generators.

The function of electrical machines can be traced back to the behavior of a current loop in a magnetic field, as shown in Fig. 26.69. A rectangular current loop, rotatable around the x axis and placed in a magnetic field $\boldsymbol{B} = (0, 0, B)$ is subject to the Lorentz force. This force is suitably calculated for the four straight elements of the loop, *front*, *back*, *top*, and *bottom*, one by one. If the current loop is made from a wire with a constant diameter, all carriers of charge will move with the same velocity $|\boldsymbol{v}|$. With the magnetic field vector pointing in the z direction, the Lorentz force for any moving charge will point in a direction within the xy-plane. The force on a sample charge ΔQ in the bottom part of the loop will point along the y-axis: $\Delta \boldsymbol{F} = \Delta Q \boldsymbol{v} \times \boldsymbol{B} = (0, \Delta F, 0)$. For a current I traversing the bottom element of length l, summing all the sample charges gives a force $\boldsymbol{F} = (0, BlI, 0)$. The force on the top element of the loop will be $(0, -BlI, 0)$, i.e., equally large, but with the opposite orientation.

The forces on the remaining elements of the loop are of similar size. However, they have little effect, as the loop is mounted such that it can only be rotated around the x-axis. So, the forces can merely *try* to stretch the loop.

The torque $\boldsymbol{M}$ can now be calculated according to $\boldsymbol{M} = \boldsymbol{r} \times \boldsymbol{F}$. It reaches its maximum value when the current loop and force form a right angle. Then, the distance $|\boldsymbol{r}|$ between the charges and the axis of rotation is half as large as the length of the front element of the current loop. Also, because the force itself is proportional to the length of the bottom element, the torque turns out to be proportional to the area enclosed by the loop:

$$\boldsymbol{M} = I\boldsymbol{A} \times \boldsymbol{B} . \quad (26.112)$$

The vector on the left of the × sign is called the *magnetic dipole moment* of the loop

$$\boldsymbol{\mu} = I\boldsymbol{A} . \quad (26.113)$$

A more general investigation would show that (26.112) and (26.113) may be used for current loops of arbitrary shape. And

$$\boldsymbol{M} = \boldsymbol{\mu} \times \boldsymbol{B} \quad (26.114)$$

serves as the definition of the magnetic moment $\boldsymbol{\mu}$, because $\boldsymbol{M}$ and $\boldsymbol{B}$ are easily measurable quantities.

If this torque is used to rotate the current loop, Fig. 26.69 shows the most elementary form of an electrical motor—and also of a generator, since there is no action without reaction. If a current in a field forces the angle to change, a change of angle will also force a current to change. So, if a current loop is forced to rotate, the setup shown in Fig. 26.69 represents the most elementary form of a generator. Its function is then governed by the Faraday–Henry law (law of induction, (26.13)), which states that the tension induced in a closed loop, U_{ind}, is determined by the rate of change of the magnetic flux through the surface enclosed by the loop: $\Phi_{\text{B}} = \boldsymbol{B} \cdot \boldsymbol{A}$. In this case,

$$U_{\text{ind}} = \frac{\mathrm{d}\Phi_{\text{B}}}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{B} \cdot \boldsymbol{A} = B \cdot A \frac{\mathrm{d}\cos\theta}{\mathrm{d}t} , \quad (26.115)$$

where θ is the angle between the magnetic field $\boldsymbol{B}$ and the surface vector $\boldsymbol{A}$.

As electrical machines transform mechanical energy into electrical energy or vice versa, according to

$$p_{\text{mech}} = -p_{\text{el}} , \quad (26.116)$$

the sign in front of the powers determines the direction of the energy transfer. Wherever the sign is positive, energy is used up, while a negative power indicates a gain of energy (see also Fig. 26.43). For the setup shown in Fig. 26.69, the mechanical energy used to rotate the current loop from the xy-plane back into it $(\Delta\theta = 180°)$ is the work to be done against the Lorentz force, $\Delta W = -F_y \Delta y$. For a half-turn, summing the forces of both wire elements contributing to the torque gives $W_{180°} = -2BIA$. If one assumes that—by some clever mechanism—the direction of the field is flipped, the same amount of energy will be needed for the next 180°. With power being the energy transfer per unit time, for a frequency f of complete turns, then

$$p_{\text{el}} = -p_{\text{mech}} = 4BAIf \quad (26.117)$$

follows.

The electrical power generated by the above setup can be calculated by assuming that the current induced by the rotation is lead through a resistor R. The rotor is assumed to be forced to have an angle changing according to $\theta(t) = 2\pi ft = \omega t$. The power p_R used by the resistor must be the same as the electrical power generated. According to (26.115), therefore

$$p_{\text{el}}(t) = -p_R = -\frac{U_{\text{ind}}^2}{R} = -\frac{\omega BA}{R}\sin^2(\omega t)\,, \tag{26.118}$$

which shows that the power rises quadratically with the frequency. The fact that electrical machines can be used to transfer both mechanical energy into electrical energy and vice versa is one of their most outstanding features.

The above formulae assumed a current loop in a constant magnetic field. In this case, the corresponding machine is called a *DC machine* (also called a *commutator machine*). So, DC machines have a constant magnetic field and a *rotor* inside. However, there are also other types of electrical machines, namely *synchronous AC machines* and *asynchronous AC machines*. AC machines use rotating magnetic fields which are produced by coils powered by alternating currents. If the rotor rotates with the same angular velocity as the field, the machine is a synchronous one, otherwise it is called asynchronous.

Figure 26.70 shows a comparison of the efficiencies achieved by synchronous, asynchronous, and direct-current machines. For powers less than 1 kW, machines with permanent magnets have good efficiencies. As the power of the machine increases, permanent magnets no longer offer sufficient field strengths. Therefore, all machines with powers exceeding 10 kW use electromagnets. Induction machines are popular for electric trains. The highest-power machines are synchronous generators in power plants.

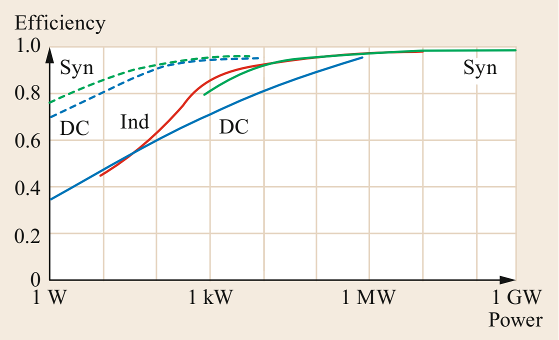

Fig. 26.70 Efficiencies of electrical machines: Syn for synchronous machines, Ind for induction (asynchronous) machines, and DC for direct-current (commutator) machines. *Dashed lines* refer to machines with permanent magnets, while *solid lines* indicate electromagnets

26.5.2 DC Machines

Direct-current motors offer a wide range of frequencies of operation, in no way limited by the frequency of the powering network. This feature guarantees their production in large numbers. In contrast, DC generators can only be found in niches.

Direct-current machines or commutator machines have a stator that delivers a magnetic field which always points in the same direction. If the field is made by permanent magnets, it is even constant. Figure 26.71 shows a typical rotor that may be placed inside a stator of a DC motor. The copper plates seen on the left are connected in pairs. They connect the rotor to the electrical power supply. Each pair of plates is connected to one coil. In Fig. 26.71, the coils are made from copper wires. The copper plates rotate between one (static) pair of *brushes* in such a way that one coil at a time is connected. In this way, the current is forced to *commute* between the coils of the rotor. Accordingly, the ring of copper plates is called a *commutator* and the motor as a whole is referred to as a *commutator machine*. The other metal parts of the rotor are made from ferromagnetic material, thus increasing and guiding the magnetic field. A good quality of these parts requires the metal to be laminated. In this way, eddy currents are minimized. For the following reason, the rotor should fit into the stator as precisely as possible: if μ_r is the relative permeability of the metal, according to Ampère's law, a fraction of $1/\mu_r$ of the magnetic field traversing air will halve the magnetic field strength. For electrical engines, fitting thus refers to accuracies well below 1 mm.

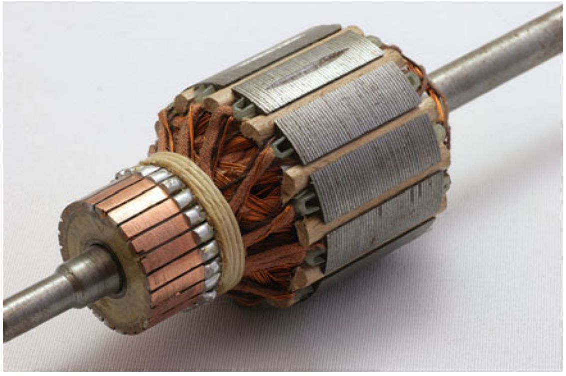

Fig. 26.71 Rotor of a DC machine; the commutator is on the *front left side*; copper wires form the coils. Laminated iron between the coils amplifies the magnetic field (photo by Sebastian Stabinger/CC BY 3.0)

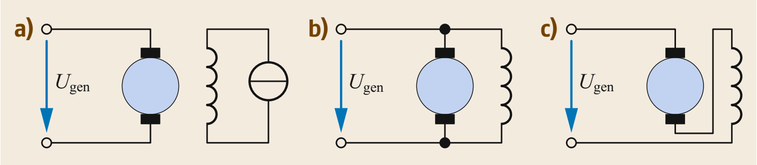

Fig. 26.72a–c Three options for powering the stator magnets of DC engines: **(a)** external; **(b)** parallel; **(c)** series. Case **(a)** includes motors with permanent magnets. U_{gen} is the voltage either generated by the machine or supplied by an external generator if the engine is operated as a motor. The inductances represent the coils of the stator

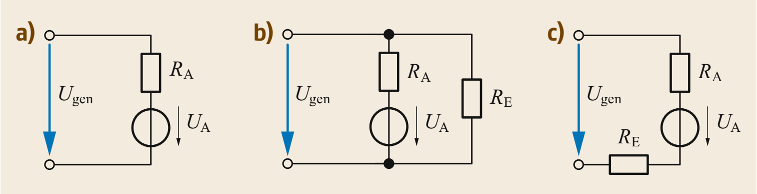

Fig. 26.73a–c Equivalent circuits for the three types of DC machines shown in Fig. 26.72. The inductances are not drawn because they have no influence on the DC behavior. R_{A} is the resistance of the rotor coil, and R_{E} that of the stator coil

There are three options to provide the energy to the stator to produce the magnetic field. Figure 26.72 shows that the simplest one (Fig. 26.72a) has a field supplied by an external current source or by permanent magnets. The stator field can also be powered by the same supply as the rotor. Its coil is then connected either in parallel (Fig. 26.72b) or in series (Fig. 26.72c). These three options can be represented by the equivalent circuits shown in Fig. 26.73. For the two options (a) and (b), the torque $M = |\boldsymbol{M}|$ can be calculated for a fixed voltage U_{gen} at different rotation frequencies f by using the relations between the torque and the current in the rotor ($M \approx I$), this current and the tension induced in the rotor coil ($I \approx (U_{\text{gen}} - U_{\text{A}})$), and this tension and the frequency ($f \approx U_{\text{A}}$). As all these relations are linear, thus so is the dependence of the torque. The straight line in the M versus f plane can be fixed at the end points corresponding to the idle speed $f_0 = f(M = 0)$ and the starting torque $M_{\text{start}} = M(f = 0)$. The result

$$M = M_{\text{start}} \left(1 - \frac{f}{f_0}\right) \quad \text{(DC, parallel)} \qquad (26.119)$$

shows that the torque for this type of machine has a maximum at $f = 0$.

For the motor with the stator coils connected in series (Fig. 26.72c), one has $M \approx I^2$, because increasing the current increases the dipole moment of the rotor as well as the strength of the magnetic field of the stator. The other proportionalities remain as before, giving

$$M = M_{\text{start}} \left(1 - \frac{f}{f_0}\right)^2 \quad \text{(DC, series)}\,. \qquad (26.120)$$

Obviously, both types of DC motors have a maximum torque at $f = 0$. They are easy to start.

The rotor field influences the stator field, just as the stator field influences the rotor field. The first effect is wanted, while the second is not. Therefore, stators are usually equipped with *compensation coils* which are connected in series with the rotor. The fields of these coils can almost exactly cancel the field that is imposed on the stator.

Finally, two further points are worth noting. The maximum of the torque goes together with a maximum current through the machine. Therefore, large DC motors need an extra resistor placed in front of the rotor connection. This resistor limits the current while the motor is being started. When DC machines are used as generators, care has to be taken that some stator field is present whenever the rotor is forced to move. If no field is present, nothing will retard the rotor and only destruction will limit its rotational speed.

26.5.3 Asynchronous Machines

All DC engines need brushes, and these are subject to wear and tear. Designers of reliable machines may thus try to avoid the use of brushes altogether. AC machines like the one shown in Fig. 26.74 offer methods to do so. They can be found in electric cars, high-speed trains, and power stations. And their efficiencies are unmatched when powers exceed 10 kW (see Fig. 26.70).

The key idea leading to an asynchronous AC machine is the following: if a static field influences a rotating magnetic dipole, then a rotating field should influence a resting dipole. In fact, it should make it

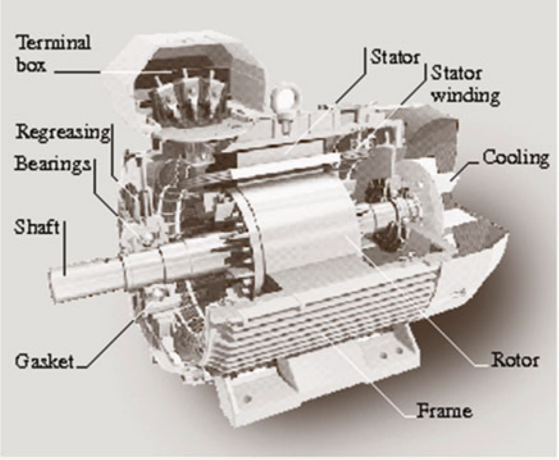

Fig. 26.74 Induction motor (asynchronous machine) (courtesy ABB)

move. Therefore, the key ingredient of an AC machine is a stator that can create a rotating field. Figure 26.75 shows that a three-phase network connected to three coils suffices to achieve this goal. The field will rotate with the same frequency as the AC supply currents.

A current loop placed in such a field will have a tension induced. When the loop is closed, this tension will lead to a current in the loop. The current will produce a magnetic dipole moment, and this will be subject to a torque due to the rotating field from the stator. The net effect is that there is a torque on an element that is not in any way electrically connected to the outside world. The current in the rotor is a result of induction. Therefore, this type of machine is also called an *induction machine*. Clearly, if the speed of the rotor coincides with the rotational speed of the field, there will be neither induction nor any forces. Because the function of the machine relies on the speeds being different, it is called an *asynchronous machine*. No brushes are needed for such a device.

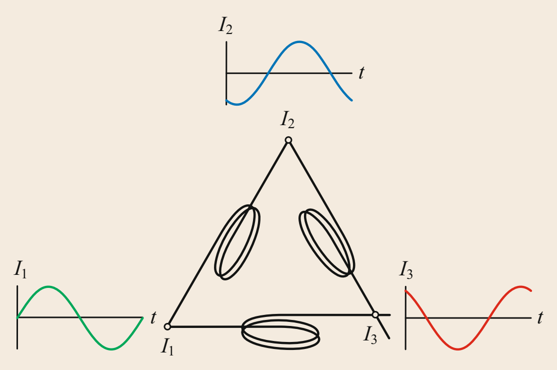

Fig. 26.75 Sketch of three coils in an AC network that produce a rotating magnetic field

As in the case of DC machines, the maximum torque can be achieved by superimposing various current loops with different azimuthal angles around the rotor axis. Joining them makes a metal cage, and such a cage can be found in almost all asynchronous AC machines.

The change of the magnetic flux $\mathrm{d}\Phi_\mathrm{B}/\mathrm{d}t$ only depends on the relative speed between the field and the rotor. The normalized difference of the angular speeds

$$s=\frac{f_\text{field}-f_\text{rotor}}{f_\text{field}}=\frac{\omega_\text{field}-\omega_\text{rotor}}{\omega_\text{field}} \qquad (26.121)$$

is called the *slip*. Therefore, the tension induced in a current loop may be expressed as a function of the slip and the voltage $U_{\text{ind},0}$ which is induced when the rotor is locked (not rotating): $U_\text{ind}(s)=sU_{\text{ind},0}$.

The cage has an ohmic resistance R_A as well as a reactive resistance $L_\mathrm{A}(\omega_\text{field}-\omega_\text{rotor})$, according to the cage inductance L_A. Therefore, the current is related to the induced voltage via

$$U_\text{ind}=\sqrt{R_\mathrm{A}^2+(\omega_\text{field}-\omega_\text{rotor})^2L_\mathrm{A}^2}\,I_\mathrm{A}\,. \qquad (26.122)$$

The current may now be expressed as a function of the slip. The result

$$I_\mathrm{A}=\frac{sU_{\text{ind},0}}{\sqrt{R_\mathrm{A}^2+\omega_\text{field}^2s^2L_\mathrm{A}^2}} \qquad (26.123)$$

shows that, the closer the circular frequency of the rotor approaches that of the stator field, the smaller the current becomes.

The relation between the current I_A and the torque is complicated by the fact that the reactive resistance of the cage introduces an angle between I_A and U_ind. If that angle is 90°, the magnetic dipole moment of the rotor is parallel to the field of the stator and the torque is zero. In order to filter out the fraction of the rotor current that does contribute to the torque, I_A has to be multiplied by the cosine of the phase angle $\varphi=\arctan(s\omega_\text{field}L_\mathrm{A}/R)$. According to (26.123), the torque then has the characteristic

$$M\approx\frac{s}{\sqrt{1+\left(\frac{s\omega_\text{field}L_\mathrm{A}}{R}\right)^2}}\cos\left[\arctan\left(\frac{s\omega_\text{field}L_\mathrm{A}}{R}\right)\right]\,, \qquad (26.124)$$

which is also shown in Fig. 26.76. The value of the torque at $\omega_\text{rotor}=0$ is referred to as the *locked rotor torque*. Its nonzero value guarantees an easy start of such engines. The torque increases with the increasing circular frequency of the rotor, reaching a maximum called the *breakdown torque*, and then approaching zero for $\omega_\text{rotor}=\omega_\text{field}$. The *rated torque* (as stated in the

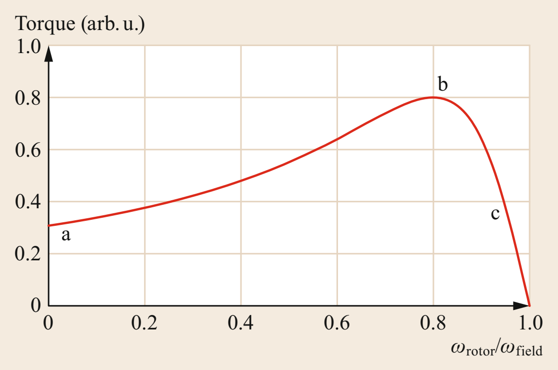

Fig. 26.76 Typical characteristics of the torque of an induction motor according to (26.124) in arbitrary units. Special values are the *locked rotor torque* (a), the *breakdown torque* (b), and the *rated torque* (c), which is about half the breakdown torque

adverts) of the machine is in some cases half the breakdown torque, and in other cases one-third of it.

As in the case of DC machines, the magnetic field in all the parts has contributions from both the stator and the rotor. According to Lenz's rule, the rotor will always weaken the stator field. When the rotor is locked, the tension induced in the rotor reaches a maximum for a given circular velocity of the stator field. The magnetic field from the current in the rotor then almost suffices to cancel out the field from the stator. When ω_{rotor} approaches ω_{field}, the induced tension approaches zero and the field of the stator is hardly weakened.

Induction motors have excellent efficiencies, and are very reliable and easily maintained. As their torque changes direction when the slip exceeds the value one, they can also serve as electrical brakes to transform kinetic energy back into electrical energy. This makes them the first choice for high-speed trains, trams, and electric cars.

Application Example: Mechanical Construction of a Rotor

The construction of the rotor for an asynchronous machine is shown in Fig. 26.77. Obviously, it does not resemble any other devices which use magnetic fields. There are no windings. Instead, there is a cage embedded in laminated iron (only three slats are drawn in Fig. 26.77). So, the magnetic interaction entirely relies on a single current loop. The currents in the rotor are thus bound to be large. The reason for the absence of windings is the need to have as little inductance as possible in the rotor. According to (26.124), the torque approaches its maximum as the phase angle $\varphi = \arctan(s\omega_{\text{field}}L_{\text{A}}/R)$ tends to zero. Therefore, in this case, one loop is better than many loops. A closer look at Fig. 26.77 also reveals that the bars forming the cage are skewed with respect to the rotor's axis. This is a measure to minimize the mechanical oscillations of the cage. The source of these oscillations is the oscillating magnetic attraction of the cage bars by the stator coils. The bars are embedded in laminated iron. The lamination is such that the magnetic field does not traverse the oxide between the sheets. In this manner, the magnetic field strength is maximized while the eddy currents are minimized.

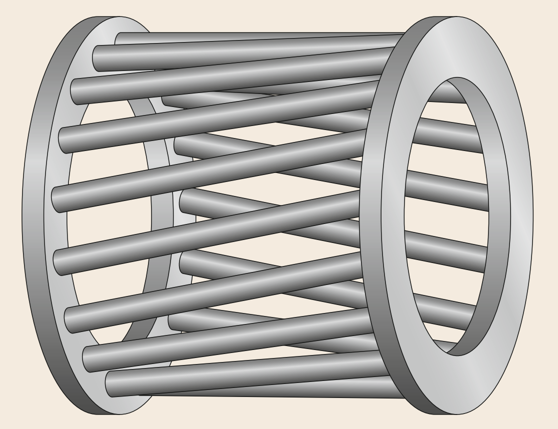

Fig. 26.77 Construction of a squirrel cage rotor: copper or aluminum bars are embedded in laminated iron and short-circuited by rings at the end. Usually, there are more bars than shown here

The manufacturing process starts by fixing a bundle of laminated iron (dynamo sheet metal) on an axis. Each of the sheets has a circular shape with notches at the rim. At the end, the bars of the cage have to be placed in the notches. This may be achieved by casting of liquid aluminum or by inserting copper bars (casting copper is difficult). Casting gives more magnetic flux, while copper bars have better conductance.

26.5.4 Synchronous Machines

Some 99% of electric power comes from synchronous generators. In synchronous machines such as that shown in Fig. 26.78, the frequency of the current is a multiple of the angular frequency of the rotor. Therefore, synchronous motors are used when the mechanical speed is to be determined electrically, while synchronous generators are used in the complementary case. For this very reason, the largest electrical engines known, AC generators in power stations, are synchronous ones. Synchronous motors are used in industrial applications and in the French high-speed train *train à grande vitesse* (TGV).

The dipoles rotating in an AC generator need to be powered by a current of almost constant strength. This

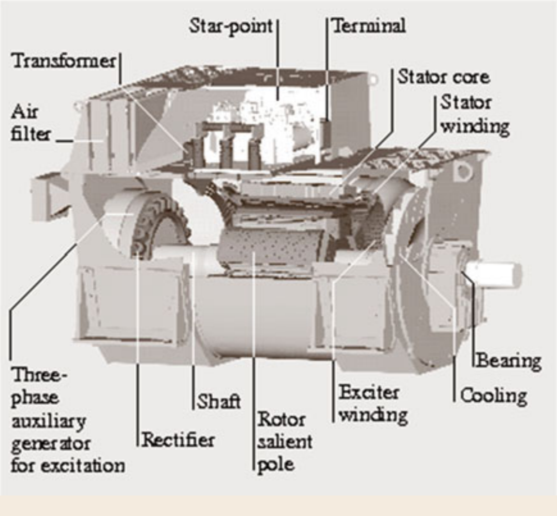

Fig. 26.78 Synchronous generator (courtesy ABB)

can be achieved using mechanical contacts, or by induction. For large machines, the latter is preferred as it causes no frictional losses. The current induced in the rotating coils needs to be rectified by power electronics.

The magnetic flux through a current loop placed next to a rotating dipole oscillates with the dipole's frequency of rotation. An alternating current will thus flow in the loop according to the law of induction. Synchronous generators use this effect by placing windings adjacent to the rotor. The currents through such windings are those delivered by large power stations. Clearly, these windings need to be placed inside laminated iron. In this way, the magnetic flux is maximized and eddy currents are minimized.

The number of dipoles (or pole pairs) of the stator usually equals the number the rotor dipoles. The smaller the angular velocity of the rotor, the larger the number of magnetic poles. In a hydroelectric power station, there may be up to 40 dipoles on the rotor, while generators attached to a steam turbine often have only one dipole. Rotors with two dipoles are found in nuclear power stations due to the mechanical limitations of the materials used. Turbines delivering up to $P = 1.3\,\text{GW}$ of power would simply disintegrate at 3000 rpm (50 Hz current) or even 3600 rpm (60 Hz current). The mechanical stability of the turbine blades is also a limiting factor of the frequency instabilities of supply networks. It is turbines that really need stable frequencies rather than the users of the currents.

The Steering of a Synchronous Generator

In a synchronous generator, the tension and the current and the phase angle between them may be varied during operation. The way this is done is identical for any number of dipoles. For simplicity, a single dipole is assumed in this section. Also, it is assumed that the magnetic fields of the rotor and the stator may simply be added (linear superposition). Usually this is a very good approximation.

The parameter that may be influenced in the most obvious manner is the tension. Suppose there is no load on the coils next to the rotating dipole. Then, the magnetic field in both the rotor and the stator is entirely determined by the rotor. Starting from a dipole coil with no current, the tension induced will rise in proportion to this current until the iron reaches the state of saturation. From then on, the magnetic field and consequently the tension induced in the stator coils rises with a slope which is less steep, 1% of that for small voltages. Such saturation is thus to be avoided under all circumstances.

As soon as there is a load on the generator, the magnetic field acquires a contribution from the stator coils in addition to that from the rotating dipole. The rotor induces a harmonically oscillating tension in the stator coils. It thus acts like an AC voltage source, because both the frequency and magnitude of the voltage are determined by the rotor and not by the stator. As the currents through the stator coils oscillate, they produce an oscillating magnetic field as well. This field will influence the coil as any other coil by introducing an inductance. This situation is sketched in Fig. 26.79. Also shown in the figure is a resistance R to account for the finite conductivity of the wire the stator coil is made of.

According to Fig. 26.79, the tension delivered by one stator coil

$$\underline{U}_{\text{gen}} = \underline{U}_{\text{ind}} - (R + \text{j}\omega L)\underline{I}_{\text{L}} \tag{26.125}$$

contains the *synchronous reactance* $X_{\text{d}} = \omega L$.

The power delivered by one coil can be calculated from (26.125) for a given load impedance as shown in Fig. 26.80. $\underline{Z} = R_{\text{L}} + \text{j}X_{\text{L}}$. The result for the apparent power,

$$\begin{aligned}\underline{S} = \frac{1}{2}\underline{U}_{\text{gen}}\underline{I}_{\text{L}}^{*} &= \frac{1}{2}\hat{U}_{\text{ind}}^{2}\left(\frac{R_{\text{L}} + \text{j}X_{\text{L}}}{(R + R_{\text{L}})^2 + (\omega L + X_{\text{L}})^2}\right)\\ &= U_{\text{ind,eff}}^{2}\left(\frac{R_{\text{L}} + \text{j}X_{\text{L}}}{(R + R_{\text{L}})^2 + (\omega L + X_{\text{L}})^2}\right)\end{aligned} \tag{26.126}$$

can be separated into the power P and the reactive power Q. Neglecting the ohmic resistance of the stator coil, one gets

$$\begin{aligned}P &\approx U_{\text{ind,eff}}^{2}\left(\frac{R_{\text{L}}}{R_{\text{L}}^{2} + (X_{\text{d}} + X_{\text{L}})^2}\right)\\ Q &\approx U_{\text{ind,eff}}^{2}\left(\frac{X_{\text{L}}}{R_{\text{L}}^{2} + (X_{\text{d}} + X_{\text{L}})^2}\right)\,,\end{aligned} \tag{26.127}$$

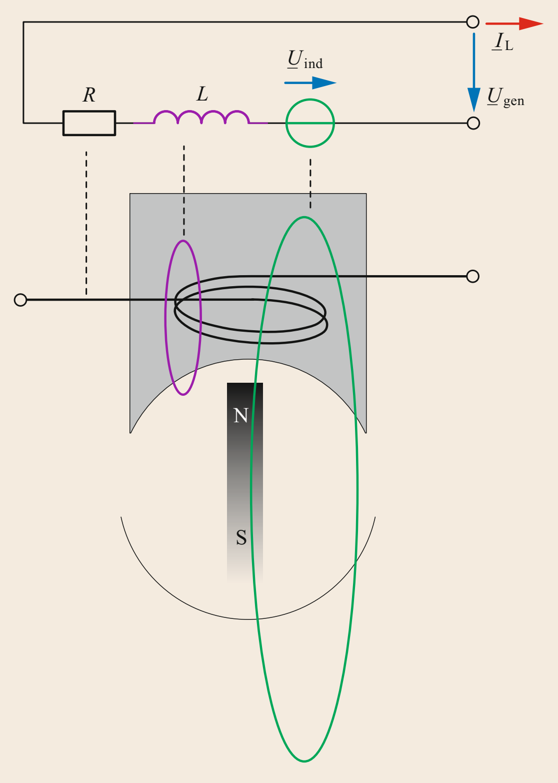

Fig. 26.79 Contributions to the circuit diagram describing one phase of a synchronous generator. The load current is $\underline{I}_L$, while $\underline{U}_{gen}$ is the voltage seen at the terminals

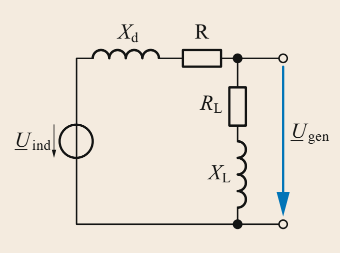

Fig. 26.80 Replacement circuit diagram of the synchronous generator shown in Fig. 26.79 when connected to a load impedance $R_L + X_L$

where, again, $U^2_{ind,eff}$ is a shorthand notation for $\hat{U}^2_{ind}/2$. For a generator with three equally loaded stator coils, the power P calculated by (26.127) is one-third of the mechanical power taken by the generator.

Figure 26.79 shows that the flux through the stator coil reaches a maximum when it is faced by the north pole of the rotor dipole. Then, most of the field lines traverse the stator coil. This situation resembles that shown in Fig. 26.69. Accordingly, the induced voltage depends on the angle θ_P between the dipole and the field of the stator. Because the value of this angle turns out to be crucial for the modes of operation of a synchronous machine, it has been given a name of its own. It is called the *polar wheel angle*.

The field produced by a stator coil is strongest when the current is at its maximum. The polar wheel angle may therefore be regarded as a measure of the delay Δt of the current maximum with respect to the point in time when the dipole field points at the stator coil: $\theta_P = \omega \Delta t$.

The modes of operation of a synchronous machine can now be related to the value of the polar wheel angle as follows:

- $\theta_P = 0$: There are no forces between the rotor and stator. Hence, neither is there any transfer of mechanical power into electric power or vice versa. The machine appears to be idle. It may, however, change the phase angle between the voltage and the current.
- $0 < \theta_P < \pi/2$: The stator field slows down the rotor. In this manner, mechanical energy is transformed into electrical energy. The machine operates as a generator. The larger the polar wheel angle, the larger the torque, and the greater the power transfer.
- $\pi/2 < \theta_P < \pi$: The field of the stator slows down the rotor. However, the slightest disruption will weaken the retarding force. Even the connection of another electrical consumer may be perturbative enough. If the mechanical power input remains unchanged, the angular speed of the rotor will rise in a manner that can no longer be controlled. In the case of a large power input, the result may well be complete destruction of the machine.
- $\theta_P = \pi$: This is the most unstable mode of operation. It corresponds to attempting to balance a compass needle with its north pole pointing downwards somewhere close to the north pole of the Earth.
- $\pi < \theta_P < 3\pi/2$: The machine is in an unstable motor mode.
- $3\pi/2 < \theta_P < 2\pi$: The field of the stator is slightly ahead of the rotor's field. The stator field drags the rotor. The machine thus acts as a motor. A slight additional mechanical load will increase the electromagnetic torque, as required for stable operation.

Power stations are usually run with polar angle values below $\theta_P < 80°$. In this manner, extra electrical loads will not make the machine leave the stable mode of operation.

Although the value of the polar wheel angle is crucial for the machine's operational mode, it is very difficult to measure. Next it will be shown which angle can be measured directly at the terminals of the machine, and what it means, viz. the angle between the tension and the current, $\varphi_{U_{\text{gen}}} - \varphi_I$.

Comparing the voltages $\underline{U}_{\text{ind}}$ and $\underline{U}_{\text{gen}}$ as shown in Fig. 26.80 gives the phase angle between the two voltages,

$$\begin{aligned}&\varphi_{U_{\text{gen}}} - \varphi_{U_{\text{ind}}}\\&\quad = \arctan\left(\frac{X_{\text{L}}(R+R_{\text{L}}) - R_{\text{L}}(X_{\text{L}}+X_{\text{d}})}{R_{\text{L}}(R_{\text{L}}+R) + X_{\text{L}}(X_{\text{L}}+X_{\text{d}})}\right).\end{aligned} \tag{26.128}$$

Usually, the ohmic resistance R of the stator winding can be neglected. The resulting approximation

$$\varphi_{U_{\text{gen}}} - \varphi_{U_{\text{ind}}} \approx \arctan\left(\frac{-R_{\text{L}}X_{\text{d}}}{R_{\text{L}}^2 + X_{\text{L}}(X_{\text{L}}+X_{\text{d}})}\right) \tag{26.129}$$

shows that ohmic loads correspond to small phase angles. Inductive loads give negative values for $\varphi_{U_{\text{gen}}} - \varphi_{U_{\text{ind}}}$, while large capacitive loads ($(X_{\text{L}}+X_{\text{d}}) < 0$) give positive values.

The phase angle $\varphi_{U_{\text{ind}}}$ may also be related to the phase angle of the current. According to Fig. 26.80, one has

$$\varphi_Z = \varphi_{U_{\text{ind}}} - \varphi_I = \arctan\left(\frac{X_{\text{d}}+X_{\text{L}}}{R+R_{\text{L}}}\right) . \tag{26.130}$$

The phase angle that can be measured can now be expressed as a function of the impedances of the circuit, because the equation

$$\varphi_{U_{\text{gen}}} - \varphi_I = (\varphi_{U_{\text{gen}}} - \varphi_{U_{\text{ind}}}) + (\varphi_{U_{\text{ind}}} - \varphi_I) \tag{26.131}$$

can be worked out using (26.129) and (26.130).

Powering Three-Phase Alternating-Current Networks

If the stator contains three coils symmetrically placed around the rotor, the tension induced in these coils will have phase angles shifted by 120° with respect to each other. Connecting them in series to form a triangle-type network as sketched in Fig. 26.68 will ensure equal current magnitudes. In this case, the stator produces a rotating $\boldsymbol{B}$ field of almost constant magnitude and angular velocity. These properties are key ingredients for the safe operation of high-power electrical generators. A constant field strength implies a constant torque on the rotor. And a $P = 10^9$ W steam turbine cannot stand anything else.

For the same reason, connecting a power generator to a network requires that the following conditions be met:

- The frequency of the generator must have almost the same value as the network frequency. In fact, it should be slightly higher, as it will be reduced as soon as a load is applied. This can only be achieved by steering the mechanical part of the generator.
- The amplitude of the generator voltage $\hat{U}_{\text{gen}}$ must match the network voltage. This can be achieved by varying the current in and thus field strength of the rotor.
- The phase angle between the generated voltage and the network voltage must be zero.

Once the generator has been connected to the network, the current in the rotor and the load on the stator coils can be increased. Increasing the load current strengthens the stator field, while the latter increases the retarding torque on the rotor. Nevertheless, the frequency of rotation will not change. Instead, because of $\boldsymbol{M} = \boldsymbol{\mu} \times \boldsymbol{B} \rightarrow |\boldsymbol{M}| \approx \sin\theta_{\text{P}}$, the polar wheel angle will increase. The value for the polar angle can be influenced during operation by varying the current that forms the rotor's dipole current: the stronger the dipole, the smaller the polar wheel angle. The polar angle may never approach 90°, because then the machine would become unstable.

State-of-the-art generators have efficiencies beyond 95%. So, the electrical power generated almost equals the mechanical power used. Thus, increasing the mechanical torque will increase the electrical power in almost the same manner as the mechanical power $-P_{\text{electric}} \approx P_{\text{mech}} = M\omega$.

A special feature of synchronous generators appears if the generator is attached to a large network with many power stations. Then, the tension at the terminals of the generator is determined by the network, rather than by the generator. This has a useful practical implication, as can be seen by dividing (26.125) by $\text{j}\omega L$. The result

$$\underline{I}_{\text{L}} = \frac{1}{\text{j}\omega L}\left(\underline{U}_{\text{gen}} - \underline{U}_{\text{ind}}\right) \tag{26.132}$$

shows that the sign of the imaginary part of the current may be chosen by having $\underline{U}_{\text{ind}}$ larger or smaller than the tension fixed by the network, $\underline{U}_{\text{gen}}$. In other words, the sign of the angle between the current and the tension may be chosen by choosing an appropriate rotor current. This characteristic introduces a new application of synchronous generators, i.e., to reduce the reactive power oscillating in supply networks, because their operation can be chosen to be inductive or capacitive. Supplying only reactive power may turn it into a pure *phase shifter*. Phase shifter operation can be achieved by powering the generator with an synchronous motor attached to the same network.

Finally, a sudden change of the electric load current will give a little *kick* to the polar wheel angle. If no countermeasures are imposed, the consequence will be a polar wheel angle oscillating around the value for which the electrical torque matches the mechanical torque. An oscillating polar angle corresponds to a ripple in the rotor's movement. For a gigawatt machine, this is a highly undesirable effect. For this reason, shorted extra windings are placed on the rotor. These extra windings act like a small asynchronous machine. They produce torque if, and only if, there is a mismatch between the stator's frequency and the rotor's frequency. In this way, they damp any oscillations of the polar wheel angle.

When used as motors, synchronous machines are difficult to start, as the operation begins from a not at all synchronous state. Therefore, a synchronous motor attached to an AC network needs an auxiliary starter motor. Electronically steered variable-frequency networks offer an alternative by ramping up the AC frequency. The combination of the motor and the electronic control is called an *electronically controlled motor* and is becoming increasingly popular.

26.6 Energy Storage

In this section, various techniques for storing energy are discussed, including the use of pumped water, double-layer capacitors, lead–acid accumulators, and zinc–air batteries. With hydrogen being a candidate for the storage of large amounts of energy, the production of electricity in a fuel cell will mark the end of this section.

26.6.1 Introductory Remarks

Storing electrical energy is a key challenge for the development of portable devices (smart phones, etc.), electric mobility, and the introduction of renewable sources into power supply chains. Consequently, there is hardly any field of engineering that is as dynamic as this one. Also, it requires cooperation between mechanical and electrical engineers, as well as chemists and physicists. Figure 26.81 shows that a large variety of techniques are presently under investigation, including purely mechanical ones such as spinning wheels (*flywheel energy storage*) or compressed air.

Some storage techniques are directly related to electrical engineering. Double-layer capacitors are particularly suitable for applications in which frequent and

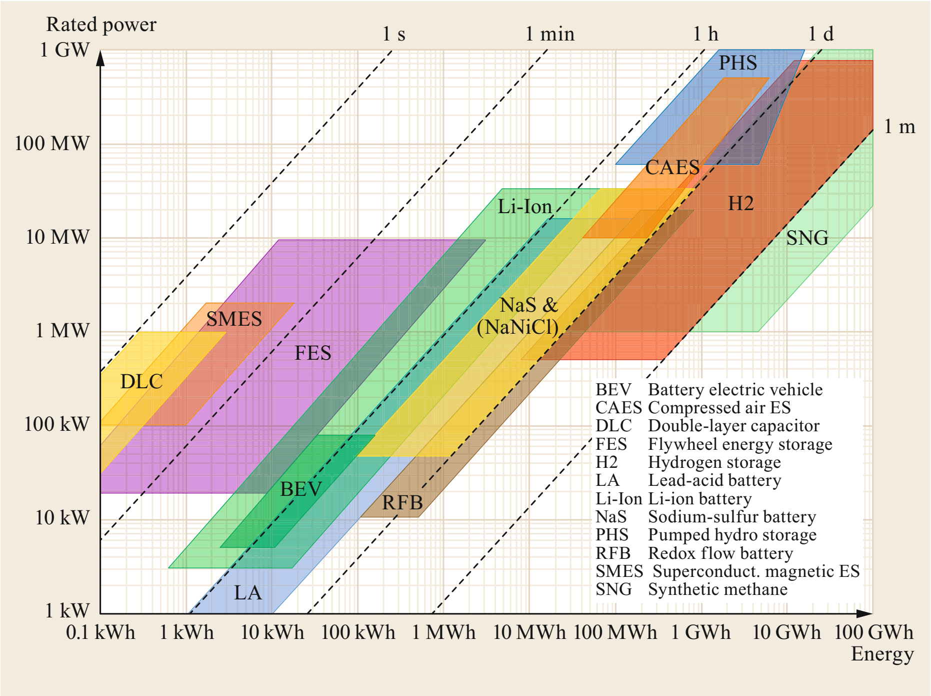

Fig. 26.81 The potential of various storage techniques according to [26.19]. The vertical axis shows the rated power, while the horizontal axis indicates the total amount of energy that may be stored in the future. (Figure by Tom Smolinka, Fraunhofer Institut für solare Energiesysteme)

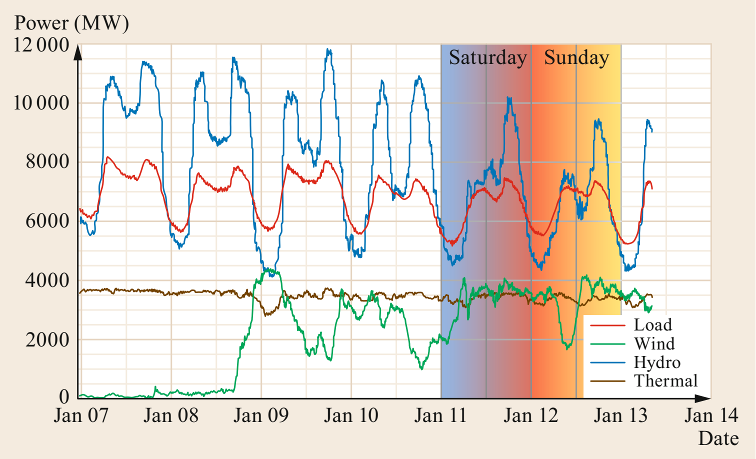

Fig. 26.82 An example of the balance of power being used (*red*) and produced by wind (*green*), water (*blue*), and conventional power houses (brown) (data published by Bonneville Power Administration, www.bpa.gov)

fast charging and discharging is needed. Batteries and their rechargeable variants (accumulators) are the most common devices when mobility is required. Today, almost every car is equipped with a lead acid accumulator (labeled LA in Fig. 26.81).

At the top end of the amount of storable energy, electromechanical techniques such as pumped hydro storage (labeled PHS in Fig. 26.81) or electrochemical techniques can be found. Hydrogen production via steam reforming from hydrocarbons, electrolysis, or thermolysis combined with fuel cells (H2 in Fig. 26.81) has enormous potential. But the cost-effective storage of large amounts of hydrogen is an outstanding problem. A possible solution is to extend hydrogen production by a second step: the synthesis of methane via

$$2H_2 + CO_2 \rightarrow CH_4 + O_2 \qquad (26.133)$$

Methane produced in this manner is easy to store, and the infrastructure already exists. Gas supply networks may thus serve as storage devices for energies in the 100 GWh regime. Unfortunately, at present, the process (26.133) runs at energetic efficiencies near the 1% level. So, the area marked SNG (synthetic methane) in Fig. 26.81 indicates a hope for the future.

26.6.2 Mechanical Storage: Water

This storage technique is both old and up to date. It is old because it consists of components that have been around for more than a century. And it is up to date because, to this day, its capacity is unmatched. Driving supply networks without water turbines is unthinkable, as can be deduced from Fig. 26.82. While power generation by thermal engines needs to be kept constant in time for good efficiency, both the load and contribution of wind turbines suffer from fluctuations with very limited control options. These fluctuations are most easily compensated by changing the contribution from hydro power.

The amount of energy that can be stored by pumping water is limited by the shape of the surface of the Earth and by the number of acres one is willing to dedicate to such energy storage. Figure 26.83 shows a storage plant close to Hohenwarte, Germany. Up to $V = 3\times10^6\,\mathrm{m}^3$ of water can be pumped to a height of 304 m. This corresponds to a gain of $\Delta W = mgh \approx 9\times10^{12}\,\mathrm{J}$, or $\Delta W = 2.5\,\mathrm{GWh}$. The maximum power of 320 MW can thus be delivered for roughly 7.5 h.

This example shows that such plants are suited to overcome energy shortages in the network that last for periods of several hours. Their commercial use is mainly to deliver power at peak usage times. Discussions are also ongoing concerning the use of such plants as buffers for energy from renewable sources, such as

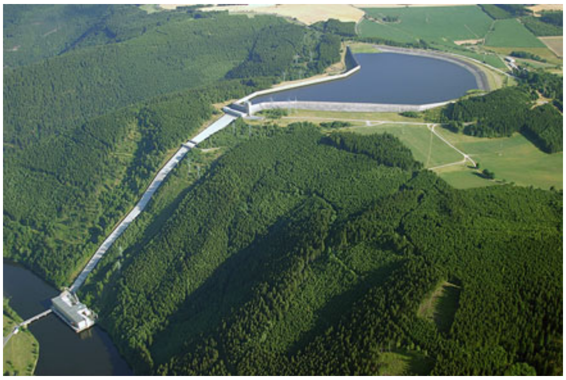

Fig. 26.83 Energy storage plant near Hohenwarte, Germany. The lake is filled with water if there is more power in the network than is needed. During times of shortage, the water is given back to the River Elbe, which can be seen at the bottom of the image (Photo: Vattenfall)

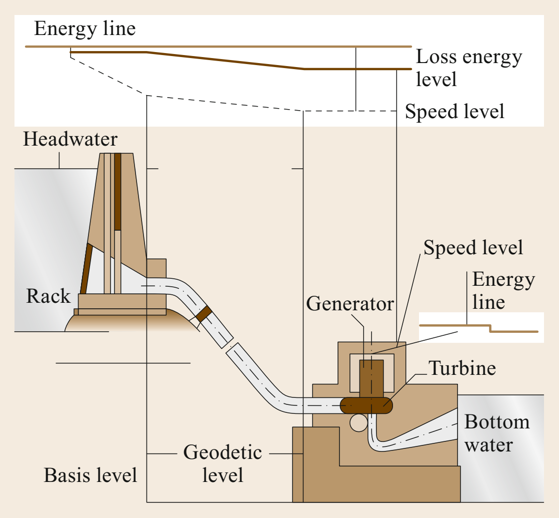

Fig. 26.84 Cross section through a water storage plant

wind energy or energy from solar cells. The time span for the delivery of power would then increase from hours to weeks. The energy to be stored would have to rise by at least three orders of magnitude, meaning that entire valleys in mountainous areas would have to be turned into water storage reservoirs. The limits on the amount of energy stored in this manner are not technical but rather set by the acceptance of the associated environmental impact.

26.6.3 Electric Storage: Supercapacitors

Supercapacitors, also branded as *gold caps*, have energy densities in the range up to 30 Wh per kg. These large values of capacitances are achieved for two reasons. Firstly, their surface area is made extremely large by using active carbon as the electrode material. Secondly, the distance between the positive and negative charge may be shrunk to less than the diameter of a single atom.

Supercapacitors are double-layer capacitors, where a capacity is formed between the anode and a electrolyte, and another between the electrolyte and the cathode. Short circuits are avoided by placing a separator between the electrodes. Ions can pass through this separator.

Figure 26.85 shows the situation at the surface of the cathode if the capacitor is charged. Negative charges (i.e., electrons) find their way right beneath the surface of the cathode. Many of them attract water molecules, as these are electrical dipoles. Some negative charges make positively charged ions (cations) form contacts directly with the surface of the cathode. In this case, they may be loosely bound by van der Waals forces, chemisorption, or simply electrostatics. Often, the positive ions remain surrounded by water molecules, even if adsorbed by the cathode. They are not mobile, but the adsorbing forces are even smaller. The term "adsorption" is used whenever a binding process is limited to taking place at a surface. If the material under the surface is involved, the term "absorption" is used. The plane above the cathode formed by the immobile ions is often referred to as the *outer Helmholtz layer* or the *Stern layer*. It is indicated by a dashed line in Fig. 26.85.

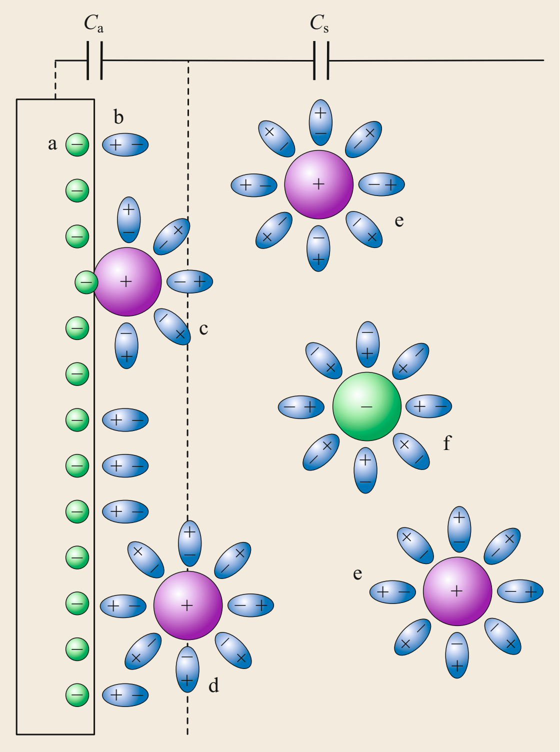

Fig. 26.85a–f Schematic view of the cathode of a supercapacitor. The negative charges **(a)** are directly placed under the surface of the cathode material (carbon). They adsorb water molecules **(b)**, cations **(c)**, and cations surrounded by water molecules **(d)**, thus forming a capacitor C_a. The sum of all these charges is still negative. It forms one electrode of another capacitor C_s with the cations solved in the liquid **(e)**. At large distances, these are balanced by negative ions **(f)**

If all the negative charges were balanced by ion adsorption, one would expect a rather constant value for the capacitance. The similarity to a capacitor with electrons moving on both electrodes seems obvious. In experiments, however, supercapacitors show a strong dependence of the capacitance on the ion concentration, surface charge, and temperature. This can be explained by assuming that only some fraction of the charges on the cathode are neutralized by adsorption of positively

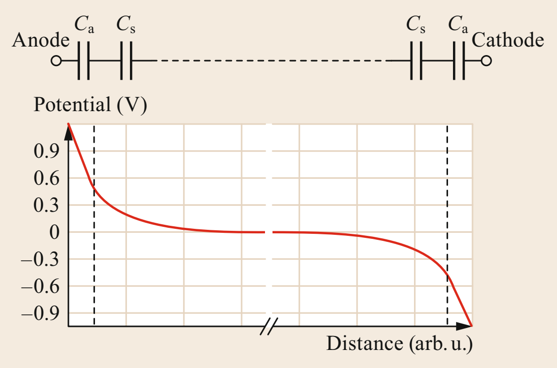

Fig. 26.86 Potential characteristics of a double-layer supercapacitor and the corresponding circuit schematic diagram. The steepest part of the characteristic is between the electrodes and the outer Helmholtz layers (*dashed line*). The shallowing off reveals the effect of diffusion within the liquid

charged ions. Other ions have sufficient thermal kinetic energy to move away from the cathode. So, only at a distance which is much larger than the distance to the outer Helmholtz layer are all the negative charges balanced by cations. The charge imbalance between the outer Helmholtz layer and the liquid far from the cathode will grow the more charge there is on the cathode. This effect is often modeled using a *pseudocapacitor* C_s between the immobile charges and the charges moving about within the solution. Here, the term "pseudocapacitor" refers to the fact that charge is separated (i.e., as in a capacitor), but not strictly, only statistically due to an equilibrium of Brownian motion and electrostatics (i.e., not really like a capacitor). The potential characteristic is sketched in Fig. 26.86, together with the corresponding circuit diagram. The steepest part of the characteristic is in the region where the ions are attached to the electrode's surfaces, either directly or via water bridges. The less steep parts are the result of Brownian motion and electrostatic attraction. Supercapacitors are not to be used for AC applications. Also, with self-discharge times measured in weeks, they cannot replace batteries if a constant supply is needed for years. They are, however, the first choice if large amounts of power are needed for a limited time (see the example from Formula 1 racing below).

Application Example: A Formula 1 Energy Recovery System (ERS)

Formula 1 cars use *energy recovery systems* (ERSs) to increase their energetic efficiency. According to the rules of the sport, such systems are limited to deliver a power of $P_{\max} = 120\,\mathrm{kW}$ for at most 33 s. The energy may be stored in supercapacitors, as shown in

Fig. 26.87 Supercapacitors (Photo: Capcomp GmbH)

Fig. 26.87. Each of them has a withstand voltage of 2.7 V, a nominal capacitance of $C = 3\,\mathrm{kF}$, and a weight of 535 g. For the design of the car, it is important to know for how long the energy of one capacitor suffices and the extra weight needed for a given time at the maximum power.

According to (26.46), the energy stored in a single capacitor is $W = 0.5 \times 3000\,\mathrm{F} \times (2.7\,\mathrm{V})^2 \approx 22\,\mathrm{kJ}$. Because $P = W/t$, delivering a constant power of $P = 120\,\mathrm{kW}$ will empty the capacitor after a time

$$t = \frac{W}{P} \approx \frac{22\,\mathrm{kJ}}{120\,\mathrm{kW}} \approx 0.18\,\mathrm{s}\,.$$

This value is quite realistic, because two factors approximately cancel. The huge variation of the values of C due to production uncertainties, usually quoted as $-20\% + 80\%$, allows one to select the best capacitors from large samples (as long as money does not matter). At the same time, the tension at the connections of the capacitor tends to zero as it is emptied, making the last 30% (as a rule of thumb) of the total energy unusable. Thus, to store sufficient energy for 10 s of full ERS power, more than 50 capacitors with a weight of about 30 kg are needed. Achieving the maximum allowed time of 33 s would require an extra weight of roughly 100 kg.

Heating up is a major issue. For an F1 car, one will therefore store the energy in stacks of capacitors connected in series, increasing the voltage and simultaneously reducing the total capacitance according to (26.48). As the power is $P = UI$, such a stack delivers the power with the smallest possible current, thus minimizing losses in wires, connections, etc. The maximum voltage allowed in an ERS is limited to $U = 1\,\mathrm{kV}$.

26.6.4 Electrochemical Storage: Batteries, Accumulators, and Fuel Cells

The basis of all batteries is the *galvanic cell*, in which redox reactions are used to make electrical current flow. In brief: ions transport electrons from one electrode

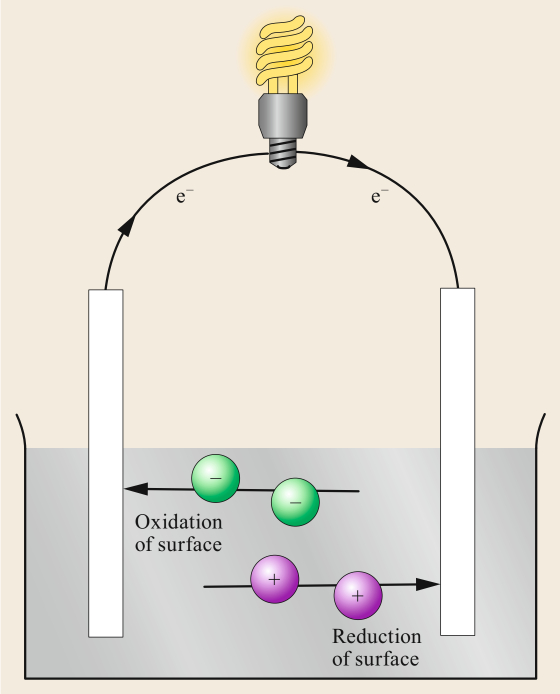

Fig. 26.88 Galvanic cell. In the most general case, there are positive and negative ions in a solution that react with solid bodies. Where the negative ions oxidize, a cathode forms, while the reduction by positive ions makes an anode. When connected, a current flows until the ions are used up

placed in an electrically conducting liquid to another one. If the electrodes are connected by a conductor, as shown in Fig. 26.88, the electrons will find their way to the other electrode. The electrode emitting electrons into the electric circuit is the cathode, while the other one is the anode. The voltage between the electrodes can be calculated in two steps; first the potential of one electrode and the liquid is calculated, then the other. The voltage of the galvanic cell is the difference of both potentials. The potential of an electrode that is reduced by accepting electrons from the liquid is given by a potential V_0 (often called E^0 in books on chemistry) as measured under standard conditions, several constants, and the concentration α of active molecules relative to a concentration of 1 mol/L. For temperatures near 20 °C and normal pressure, it may be calculated as

$$V = V_0 - \frac{0.05916\,\mathrm{V}}{n} \log_{10}\left(\frac{\alpha_{\text{products}}}{\alpha_{\text{reactants}}}\right) , \qquad (26.134)$$

where n is the number of electrons transferred per molecular reaction. In (26.134), known as the Nernst equation, α is set to 1 mol/L for solid bodies and for water by convention (for details, see [26.20]). The value 0.05916 comes from multiplying various natural constants.

Galvanic cells produce electrical energy, whenever the chemical reactions taking place between the electrodes and the solved substances are exothermic. Most of the energy released by the reaction is then transformed into electrical energy. The rest heats up the cell. There are also cells that use endothermic reactions. They are called *electrolytic cells*. These cells require electrical energy to be used by the cell, so the light bulb in Fig. 26.88 would have to be replaced by some kind of electrical generator. If a cell can be used as both a galvanic cell and an electrolytic cell, it may serve as an *accumulator*.

Lead–Acid Battery

The lead–acid battery is, despite its name, an accumulator, as it can be recharged. It is the workhorse of the automotive world. It offers energy densities of $W/m \approx 30\,\mathrm{Wh/kg}$ and is suited for high currents. It relies on the simultaneous oxidation of lead to form lead sulfate and the reduction of lead oxide, also to form lead sulfate. A fully charged battery has a lead cathode, lead oxide as its anode, and a high concentration of sulfuric acid. A completely discharged battery has both electrodes consisting of lead sulfate and a very low concentration of sulfuric acid.

Before the redox reactions can start, sulfuric acid has to deliver ions according to

$$2\,\mathrm{H_2SO_4} \rightarrow 4\,\mathrm{H^+} + 2\,\mathrm{SO_4^{2-}} \qquad (26.135)$$

Oxidation of lead produces two electrons according to

$$\mathrm{Pb} + \mathrm{SO_4^{2-}} \rightarrow \mathrm{PbSO_4} + 2\,\mathrm{e^-} \quad \text{(oxidation)} \qquad (26.136)$$

thus turning this piece of lead into a *cathode*. Since one atom of lead replaces two hydrogen atoms in $\mathrm{H_2SO_4}$, each lead atom loses two electrons in this reaction ($\mathrm{Pb} \rightarrow \mathrm{Pb^{2+}}$). At the same time, reduction of lead oxide according to

$$\begin{aligned}&\mathrm{PbO_2} + \mathrm{SO_4^{2-}} + 4\,\mathrm{H^+} + 2\,\mathrm{e^-}\\ &\quad \rightarrow \mathrm{PbSO_4} + 2\,\mathrm{H_2O} \quad \text{(reduction)}\end{aligned} \qquad (26.137)$$

needs exactly those two electrons, thus turning the lead oxide into an *anode*. Because oxygen has a valence of 2, the lead oxide gains two electrons ($\mathrm{Pb^{4+}} \rightarrow \mathrm{Pb^{2+}}$). If the lead and lead oxide are connected by a wire, electrons will pass through this wire and the current may be

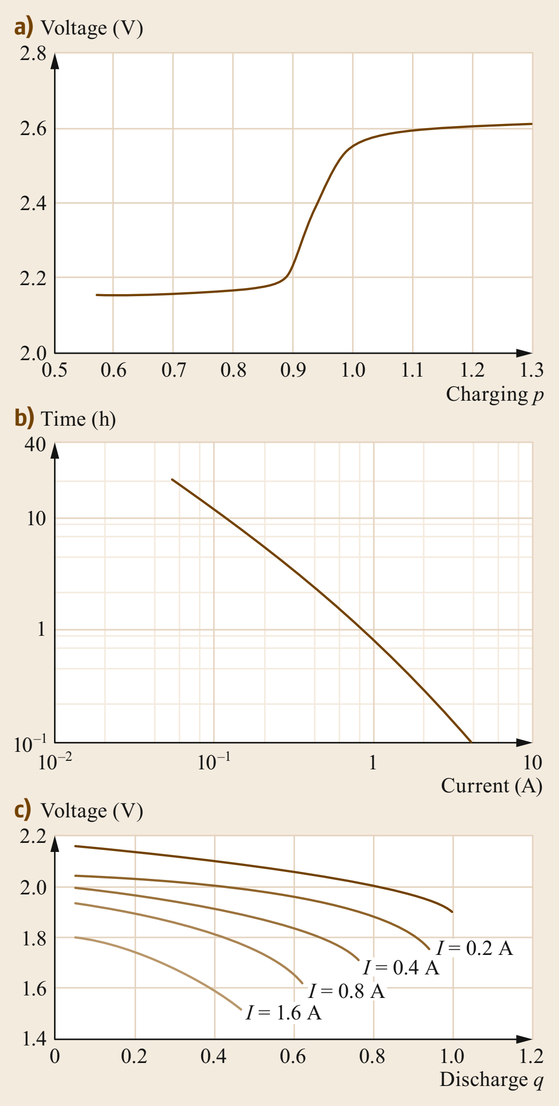

Fig. 26.89a–c Characteristics of lead–acid batteries: **(a)** charge process; **(b)** discharge time of one cell related to discharge current; **(c)** discharge characteristic of a cell at 15 °C

used. If all substances are given in a concentration of 1 mol/L, the log terms in (26.134) vanish, and one gets a cell voltage [26.21] of

$$\begin{aligned} V_{\text{cell}} &= +V_0(\text{reduction}) - V_0(\text{oxidation}) \\ &= 1.69\,\text{V} - (-0.36\,\text{V}) = 2.05\,\text{V}\,. \end{aligned} \quad (26.138)$$

One may look at the overall reaction

$$\begin{aligned} &\text{Pb} + \text{PbO}_2 + 2\,\text{SO}_4^{2-} + 4\,\text{H}^+ \rightarrow 2\,\text{PbSO}_4 + 2\,\text{H}_2\text{O} \\ \text{or}\quad &\text{Pb} + \text{PbO}_2 + 2\,\text{H}_2\text{SO}_4 \rightarrow 2\,\text{PbSO}_4 + 2\,\text{H}_2\text{O} \end{aligned} \quad (26.139)$$

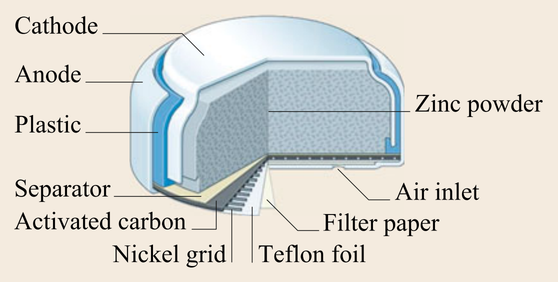

Fig. 26.90 Sketch of a zinc–air battery as used in hearing aids

to determine the voltage for different concentrations

$$V = 2.05\,\text{V} - \frac{0.05916\,\text{V}}{2}\log_{10}\left(\frac{\alpha_{\text{PbSO}_4}^2\alpha_{\text{H}_2\text{O}}^2}{\alpha_{\text{Pb}}\alpha_{\text{PbO}_2}\alpha_{\text{H}_2\text{SO}_4}^2}\right). \quad (26.140)$$

When setting the activities of solids and water to one, the result

$$\begin{aligned} V &= 2.05\,\text{V} - \frac{0.05916\,\text{V}}{2}\log_{10}\left(\frac{1}{\alpha_{\text{H}_2\text{SO}_4}^2}\right) \\ &= 2.05\,\text{V} + 0.05916\,\text{V}\log_{10}\left(\alpha_{\text{H}_2\text{SO}_4}\right) \end{aligned} \quad (26.141)$$

shows that the voltage depends only on the concentration of the acid. This simple relation offers an easy test of the state of charge. Therefore, the concentration of sulfuric acid is a measure of the charge left in a lead–acid battery. It can be checked by measuring the specific weight of the liquid: the heavier the liquid, the more charged the battery. Commercially available lead–acid accumulators are said to have concentrations up to 6 mol/L, corresponding to a density of roughly 1.34 kg/L.

If the lead–acid accumulator is connected to a generator, the reactions (26.136) and (26.137) are reversed, enhancing the concentration of sulfuric acid and the energy content of the cell. The charging process can be seen in Fig. 26.89. The electrolyte gas starts to appear at above 2.4 V, and charging should be stopped at 2.65 V per cell. During discharging, the minimal voltage should not be below the cut-off of 1.8 V. The usable capacity is a function of the discharge current. Figure 26.89b shows the discharge time as a function of the current for a single cell, on a logarithmic scale. The discharge characteristics of a single cell as a function of the discharge rate with the discharge current as a parameter are presented in Fig. 26.89c. Lead–acid storage batteries are sometimes used as storage devices in the field of electricity supply. Then the battery is used for load leveling, frequency control, provision of instantaneous reserve, or voltage control.

Lead–acid accumulators are rather heavy. A significant weight reduction can be achieved by using

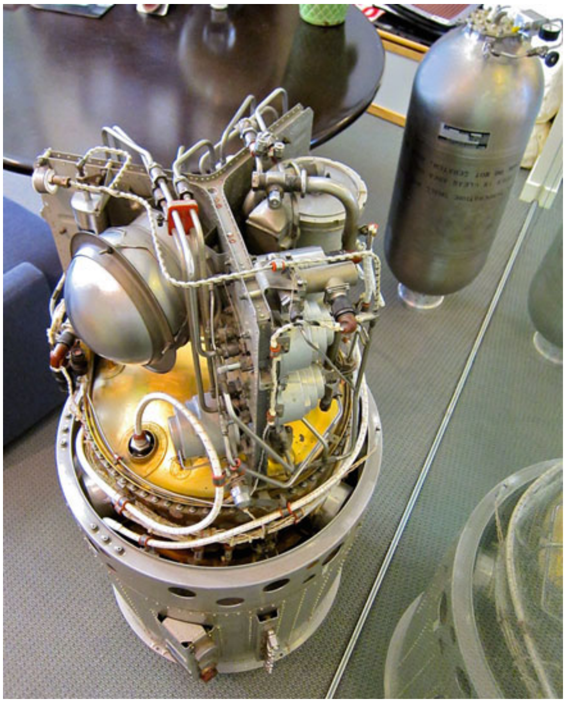

Fig. 26.91 The fuel cell used in the Apollo program (©2011 Steve Jurvetson, flickr.com)

other metals and oxygen in the air. Figure 26.90 shows a widely used variety of this type of battery: the zinc–air battery. It is hoped that, in the near future, lithium–air batteries with a similar construction will achieve very high energy densities.

Fuel Cells

In a fuel cell, chemical energy is transformed into electrical energy by means of a steady material input. The zinc–air battery might also be called a fuel cell, rather than a battery, as its production of electricity stops when the air inlet (Fig. 26.90) is closed. One of the early applications, shown in Fig. 26.91, was electricity generation in the spaceships of the Apollo missions. For an interesting historical overview, see [26.22]. The overall reaction is

$$2\,H_2 + O_2 \rightarrow 2\,H_2O \qquad (26.142)$$

thus producing clean water as a collateral benefit. The two parts forming this reaction can be deduced from Fig. 26.92. They are [26.20]

$$\begin{aligned} &2\,OH^-(\text{sol}) + H_2(\text{gas}) \\ &\quad \rightarrow H_2O(\text{gas}) + H_2O(\text{liq}) + 2\,e^- \quad (\text{cathode}) \\ &2\,H_2O(\text{liq}) + O_2(\text{gas}) + 4\,e^- \\ &\quad \rightarrow 4\,OH^-(\text{sol}) \quad (\text{anode}) \end{aligned} \qquad (26.143)$$

So, the electrolyte has to be a solution with a high concentration of OH^- ions. In the Apollo missions, concentrated potassium hydroxide (KOH) was used at a pressure of 20–40 bar and a temperature of 200 °C. The electrodes were made of porous nickel powder, providing a large surface for catalytic reactions. The tension produced by one cell is then

$$\begin{aligned} V_{\text{cell}} &= V_0(\text{reduction}) - V_0(\text{oxidation}) \\ &= 0.4\,\text{V} - (-0.83\,\text{V}) = 1.23\,\text{V}\,. \end{aligned} \qquad (26.144)$$

With hydrogen being quite difficult to store, a large variety of materials and operating conditions have been tried out during the last decades. An easy-to-read introduction to the more widely used kinds of fuel cells can be found in [26.23].

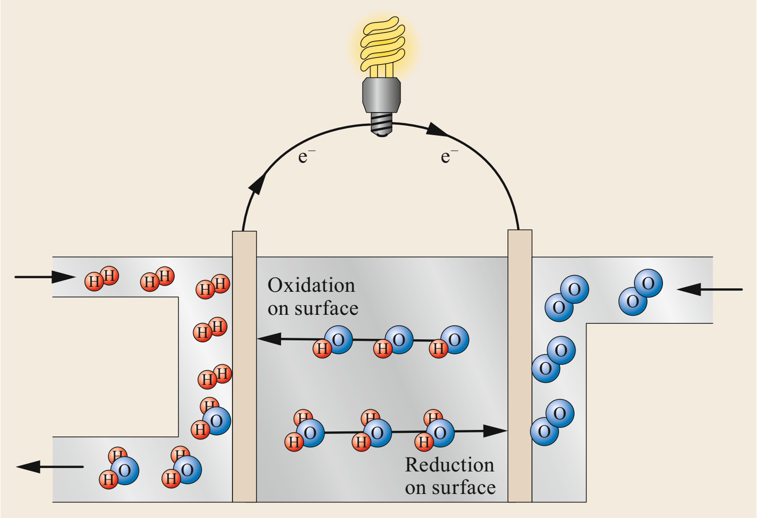

Fig. 26.92 An alkaline fuel cell; at the cathode (*left*), hydroxide ions and hydrogen combine to water. At the anode, oxygen and water form hydroxide ions

References

26.1 A. Einstein: Zur Elektrodynamik bewegter Körper, Ann. Phys. **322**(10), 891 (1905)

26.2 J. Clerk Maxwell: A dynamical theory of the electromagnetic field, Philos. Trans. R. Soc. **155**, 459 (1865)

26.3 M. Poppe: Exclusive hadron production in two-photon reactions, Int. J. Mod. Phys. A **1**(3), 545 (1986)

26.4 R. Van Dyk, P. Ekstrom, H. Dehmelt: Axial, magnetron, cyclotron and spin-cyclotron-beat frequencies measured on single electron almost at rest in free space (geonium), Nature **262**, 776 (1976)

26.5 R.S. Van Dyck Jr., P.B. Schwinberg, H.G. Dehmelt: Precise measurements of axial, magnetron, cyclotron, and spin-cyclotron-beat frequencies on an isolated 1-meV electron, Phys. Rev. Lett. **38**, 310 (1977)

26.6 M. Poppe: *Die Maxwellsche Theorie* (Springer, Heidelberg 2015)

26.7 C.F. Gauss: *Observationes cometae secundi, observatorio Gottingensi factae, adjectis nonnullis adnotationibus circa calculum orbitarum parabolicarum* (Dieterich'sche Univ.-Buchdruckerei Kaestner, Göttingen 1813)

26.8 M. Faraday: Experimental researches in electricity series XI, Philos. Trans. R. Soc. **128**, 1–40 (1838)

26.9 W.K.H. Panofsky, M. Phillips: *Classic Electricity and Magnetism* (Addison Wesley, Boston 1990)

26.10 R.P. Feynman: *The Feynman Lectures on Physics: The Definitive and Extended Edition*, Vol. 2 (Addison Wesley, Boston 2005)

26.11 P.A. Tipler, G. Mosca: *Physics for Scientists and Engineers*, 6th edn. (Freeman, London 2007)

26.12 P. Drude: *Lehrbuch der Optik* (Hirzel, Leipzig 1906)

26.13 G.S. Ohm: *Die galvanische Kette* (Riemann, Berlin 1827)

26.14 Georgia State University: Hypherphysics, http://hyperphysics.phy-astr.gsu.edu (2016)

26.15 EPCOS AG: Technical library, https://en.tdk.eu/tdk-en/180390/tech-library/publications/capacitors (2016)

26.16 M. Grundmann: *The Physics of Semiconductors*, 3rd edn. (Springer, Heidelberg 2016)

26.17 Intel.com: Standards 22nm-3d-tri-gate-transistors presentation, http://www.intel.com/content/www/us/en/silicon-innovations/standards-22nm-3d-tri-gate-transistors-presentation.html?wapkw=transistor (2016)

26.18 R. Gregorian, G.C. Temes: *Analog MOS Integrated Circuits for Signal Processing* (Wiley, New York 1986)

26.19 Energiespeicher – Forschungsinitiative der Bundesregierung: http://forschung-energiespeicher.info/wind-zu-wasserstoff (2016)

26.20 P.W. Atkins, J. de Paula: *Atkins' Physical Chemistry*, 9th edn. (Wiley, Hoboken 2009)

26.21 S.-C.S. Wang: Advanced secondary batteries and their applications for hybrid and electric vehicles, http://sites.ieee.org/clas-sysc/files/2012/05/Wang-Battery-and-EV.pdf (2011)

26.22 The Smithsonian Institute: Fuel cells, http://americanhistory.si.edu/fuelcells (2016)

26.23 F. Barbir: *PEM Fuel Cells* (Elsevier, Amsterdam 2013)

Martin Poppe

Electrial Engineering and Computer Science
Muenster University of Applied Sciences
Steinfurt, Germany
poppe@fh-muenster.de

Martin Poppe is Professor of Electronics and Prototyping at Muenster University of Applied Sciences. He is a Rhodes Scholar who read Physics at the University of Oxford (DPhil) and worked on experimental quantum electrodynamics at DESY and CERN. The results, published in 1986, are still in use by research collaborations. He is coauthor of several textbooks.