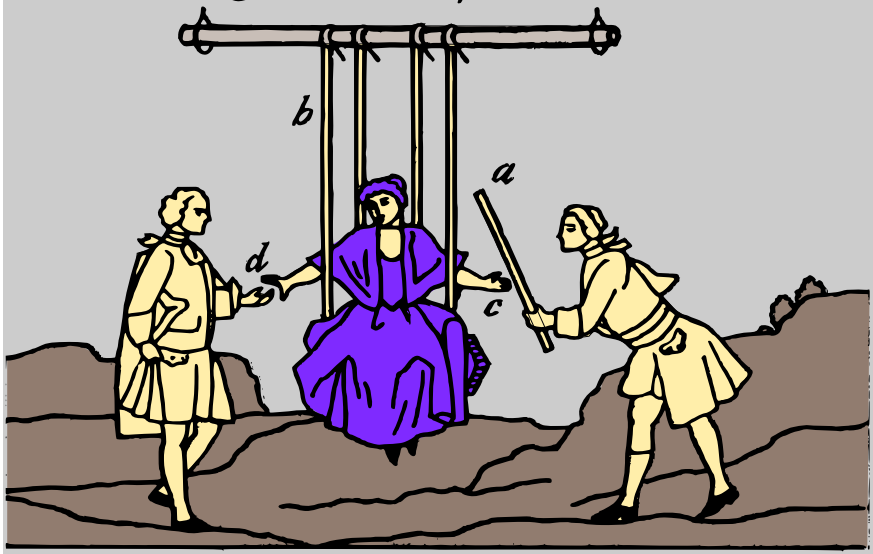
Physics for Everyone

Book 3

A.I.Kitaigorodsky



ELECTRONS

Mir Publishers Moscow

Physics for Everyone

Book 3

The fourth Russian of Physics for Everyone by L. Landau and A. Kitaigorodsky was published in 1978 as two separate books: Physical Bodies (Book 1) and Molecules (Book 2). They were published in English in 1979. This is the first publication of Book 3 of this series. It is called Electrons and was written by A. Kitaigorodsky as a sequel to Physics for Everyone.

This book deals with physical phenomena in which our attention is focussed on the next level in the structure of matter - the electrical structure of atoms and molecules. Electrical and radio engineering, without which the existence of today's civilization is inconceivable, are based on the laws governing the motion and interaction of electrical particles, primarily electrons, the quanta of electricity.

The main subjects of this book are electric current, magnetism and electromagnetic fields.


ИЗДАТЕЛЬСТВО МИР

Physics for Everyone

Book 3

A. I. Kitaigorodsky

ELECTRONS

Translated from
the Russian
by Nicholas Weinstein

Mir Publishers Moscow

Физика для всех
Книга 3
А. И. Китайгородский
Электроны
Издательство «Наука»

First published 1981

На английском языке

PREFACE

The first book of the series *Physics for Everyone* dealt with the laws of motion of large bodies and with gravitational forces. The second is about the molecular structure of matter and molecular motion.

The present, third, book of the series discusses the electrical structure of matter, electric forces and electromagnetic fields.

The next, fourth, book is concerned with photons, structure of the atomic nucleus and nuclear forces.

Thus, the four books of the series contain information on all the basic concepts and laws of physics. Specific facts presented in the series have been selected to illustrate physical laws in the clearest possible way, to demonstrate the techniques most commonly used by physicists in investigating phenomena, to give the reader an idea of the evolution of physical theory and, finally, to substantiate, in a most general way, the fact that physics is the foundation of all natural science and engineering.

Physics has drastically changed in the lifetime of a single generation. Many of its chapters have grown into independent branches of science whose applications are of tremendous significance. In my opinion, one cannot consider that his education is complete today if he has mastered only the fundamentals of physics. *Physics for Everyone* is intended as a series of books enabling one to acquire some specific knowledge of the principles of physics and to find out what new advances have been made in the physical sciences during the last decades.

This series should, of course, prove to be of greatest interest as a teaching aid and as a supplement to the textbook for physics students.

I remind the reader again that this is not a formal textbook. It was written for the layman, and its purpose is to render physics intelligible to the nonspecialized reader. The amount of space devoted to any subject in a textbook depends upon the difficulty with which the student understands the presented material. A book on science written for the general reader does not follow this rule. Hence, various pages do not read with equal ease. Another essential difference is that we can permit ourselves to expound certain traditional chapters in a less detailed manner, condensing older material to make room for new developments.

A few words about the present book, *Electrons*. Somewhat unusual use has been made of the necessity for reminding the reader of the definitions for the simplest concepts employed to describe electrical phenomena. I have tried to give an idea of the phenomenological approach to physics.

Two out of the six chapters deal with applied physics. Electrical engineering is presented as a summary. A detailed description would require us to resort to drawings and diagrams. It was considered feasible, therefore, to limit the text to a presentation of only the basic principles of electrical engineering and of important facts that everyone should know.

The same is true of the chapter on the radio. The small size of the book restricted the material to a brief history of discoveries and developments, and an account of the fundamentals of radio engineering.

A. Kitaigorodsky

CONTENTS

1. Electricity

Electric Current

Using the study of electricity as an example, it is feasible (and necessary) to acquaint readers showing an interest in physics with the phenomenological approach to the investigation of nature. The word "phenomenon", from a Greek word meaning "to appear", is defined by the Webster Dictionary as "any fact, circumstance, or experience that is apparent to the senses and that can be scientifically described or appraised". The approach mentioned above consists in the following. The investigator is not interested in the "nature of things". He uses words only to tell about facts. His aim is not to "explain", but only to describe phenomena. Almost all the terminology that he introduces is meaningful to him only if he can indicate some way to numerically evaluate the corresponding concepts.

He resorts to certain auxiliary names only to facilitate a verbal account of the facts. But the role of these names is absolutely secondary, they could just as well be replaced by other names or by simply saying "something" or "thingumajig".

The phenomenological method plays an immense role in natural science. Electrical phenomena are exceptionally suitable examples for explaining the essence of this method to the reader.

At the end of this chapter I shall review briefly the actual sequence of events in the history of electricity theory. At the present, I wish to present a certain ideal-

ized outline for the evolution of the phenomenological theory of electrical phenomena.

First, let us combine into a single mythical personage such scientists as Charles Augustin de Coulomb (1736-1806), Alessandro Volta (1745-1827), Georg Simon Ohm (1787-1854), André Marie Ampère (1775-1836), Hans Christian Oersted (1777-1851), Heinrich Friedrich Emil Lenz* (1804-1865) and others of that ilk. Imagine that our corporate investigator is endowed with a capacity for modern scientific thought, and let us equip him with a complete set of up-to-date terminology. This composite scientist is to play the chief role in our further discussion.

He begins to conceive a phenomenological theory of electricity by carefully examining a storage battery. The first feature he takes notice of is that the battery has two "poles". Touching them with his two hands, he instantly learns that this is something to avoid (the shock may be quite severe). But after this first experiment, it probably occurs to him that something passed through his body; we shall name this "something" electricity.

Being extremely careful not to suffer another shock, he begins to connect the poles (actually terminals) by various wires, rods and cords. He then discovers the following: of the articles brought into contact with the two poles, some are intensely heated, others become only warm and, in some cases, no heating is observed.

Selecting appropriate words to describe his discovery, our investigator decides to speak of it as follows: "When I connect the poles with a wire, electricity flows through the wire. I shall call this phenomenon an electric current. Experiments indicated that items of different materials are differently heated by the current. Those which

---

* Known as Emil Christianovich Lenz in his native Russia.

are only slightly heated evidently "conduct" electricity poorly or offer high resistance to the current. They can be called *insulators* or *dielectrics.*"

Next our investigator begins experimenting with liquids. Again he finds that different substances behave differently. Finally, he makes an interesting discovery: taking a solution of copper sulphate and immersing carbon electrodes (as we call the items connected to the poles) into the bath, our scientist finds a reddish copper deposit on one of the carbons.

At this stage, our investigator becomes convinced that the phenomenon he is studying has to do with the flow of some kind of fluid. It obviously makes sense to speak of the direction of flow. Let us agree upon marking the electrode on which the copper is deposited with a minus sign and consider the other electrode to be positive. Since it is longish and inconvenient to say "negative electrode" and "positive electrode" each time, the terms *cathode* and *anode* are suggested instead. Current passage is from the plus to the minus, i.e. from the anode to the cathode.

But the value of the discovery is far from being exhausted at this point. It is further established that an equal mass of copper is deposited on the cathode each second. Evidently, the copper atoms carry the electric fluid. Therefore, the investigator introduces two new terms. First, he supposes that the mass M of the copper is proportional to the quantity q of electricity flowing through the circuit, i.e. he introduces the equation

$$q = kM$$

where k is the proportionality factor. Next, he proposes that the quantity of electricity passing along the circuit in unit time be called the *current strength* or, simply, *current:*

$$I = \frac{q}{\tau}$$

Our investigator has become substantially enriched. He can now characterize the current by two measurable quantities: by the amount of heat evolved by a definite portion of the circuit in unit time, and by the current, the quantity of electricity flowing in unit time.

This leads to a new opportunity: he can compare the currents produced by various sources. He measures the current I and then the energy Q generated in the form of heat by a single piece of wire. Repeating the experiment with various conductors, the investigator finds that the ratio of the amount of heat generated to the current in the wire differs for various current sources. It remains to invent a term for this ratio. It was called the *voltage*. The higher the voltage, the more the heat generated.

This last statement can, to some small extent, justify our choice of terms. The more the tension required to pull a loaded wagon (and tension is sometimes used to denote voltage, e.g. high-tension current), the hotter we get. Denoting the voltage by V, as is customary today, we obtain

$$V = \frac{Q}{q}, \text{ or } Q = VI\tau$$

Thus, the first steps have been taken. Two phenomena have been discovered. An electric current deposits matter in passing through certain liquids, and an electric current generates heat. Heat is something we are capable of measuring. The method for measuring the quantity of electricity has been devised, i.e. a *definition of this concept has been given*. Definitions have also been given for *derived* concepts: the current and the voltage.

A number of simple equations have been written, but notice that they cannot be called laws of nature. For

instance, our investigator *called* the ratio Q/q the voltage, instead of *finding* that Q/q is equal to the voltage.

Now, he begins to search for the law of nature. Two quantities can be measured for the same conductor: the current and heat evolved, or the current and voltage (which, in principle, mean the same).

An investigation of the dependence of the current on the voltage leads to the discovery of an important law. The vast majority of conductors comply with the law:

$$V = IR$$

The quantity R can be called the *resistance*; this fully agrees with the initial qualitative observations. The reader has, of course, recognized this equation; it is *Ohm's law*. Substituting the value of the current from Ohm's law into our preceding formula, we obtain

$$Q = \frac{V^2}{R} \tau$$

You will not be confused, I hope, by the possibility of writing the expression for the energy evolved by a conductor in the form of heat in a different way:

$$Q = I^2 R \tau$$

It follows from the first of the last two formulas that the amount of heat is inversely proportional to the resistance. This is true if we add: at constant voltage. This is the case we had in mind when we first used the term "resistance". The second formula, contending that the heat is directly proportional to the resistance, requires the condition: at constant current.

These two equations are evidently recognized by the reader as the law named after James Prescott Joule (1818-1889) and Lenz (i.e. the *Joule-Lenz law*).

Finding thus that the voltage and current are proportional, thereby enabling the resistance of a conductor

to be determined, our investigator naturally poses the question: How is this important quantity related to the shape and size of the conductor and to the material of which it is made?

Experiments lead to the following discovery. It is found that

$$R = \rho \frac{l}{A}$$

where l is the length of the conductor and A is its cross-sectional area. This simple equation is valid when we deal with a linear conductor of constant cross section along its whole length. Resorting, if necessary, to more complex mathematical operations, we can write the resistance formula for a conductor of any shape. What, here, is the factor ρ? It characterizes the material of which the conductor is made. The value of this quantity, named the *resistivity*, varies in an extremely wide range. The resistivity of various substances may vary by thousands of millions of times.

Let us carry out several formal transformations that will prove useful further on. Ohm's law can be written as

$$I = \frac{VA}{\rho l}$$

A frequently employed quantity is the ratio of the current to the cross-sectional area of the conductor. It is called the *current density* and is usually denoted by the letter j. Then the same law can be written as

$$j = \frac{1}{\rho} \frac{V}{l}$$

At this point, it seems to our investigator that he has found everything related to Ohm's law. If he has at his disposal an unlimited number of conductors of known

resistance, our investigator can reject the cumbersome technique of determining voltage by means of a calorimeter: he now knows that the voltage equals the product of the current by the resistance.

Our investigator soon finds, however, that this statement is in need of refinement. Using the same current source, he connects its poles through various resistances. The current naturally differs in each experiment. But, he finds, the product IR of the current and resistance does not remain constant. When he studies this, as yet unexplained, phenomenon, the investigator finds that with an increase in the resistance the product IR tends to a certain constant value.

Denoting this limit by $\mathscr{E}$, we derive a formula that does not coincide with the one established when measuring the voltage and current. The new formula is

$$\mathscr{E} = I(R + r)$$

What a strange contradiction!

After some thinking, we come to the conclusion that there is, of course, no real contradiction. When we directly measured the voltage with a calorimeter, we were concerned only with the conductor that connected the storage battery terminals. It must be clear, however, that heat is also evolved in the battery itself (we can make sure by touching the battery with our fingers). The storage battery has its own resistance. The meaning of quantity r, found in the new formula, is obvious, it is the internal resistance of the current source. Quantity $\mathscr{E}$ requires a special name. One cannot contend that the name selected is particularly appropriate. Quantity $\mathscr{E}$ is called the *electromotive force (emf)*, though it has neither the meaning nor the dimensionality of force.

Both formulas continue to be called Ohm's law (observing, so to speak, historical fairness), only the first is

called Ohm's law for a portion of a circuit, and the second, Ohm's law for the whole circuit.

Now, everything seems to be cleared up. The laws of direct current have been established.

But our investigator is still unsatisfied. Even without direct measurement of the voltage with a calorimeter, its determination remains cumbersome. Imagine weighing the cathode with its copper deposit each time! You must admit that this is extremely inconvenient, to say the least.

One truly fine day, our investigator quite accidentally placed a compass near a conductor in which there was a current. This turned out to be a great discovery. The magnetic needle turned violently when current passed through the conductor, turning in the opposite direction when the current was reversed.

The moment of force acting on the magnetic needle can be readily determined. Hence, a measuring instrument can be devised on the basis of this observed phenomenon. All that is required is to establish the dependence of the moment on the current. Our investigator solves this problem and designs excellent pointer-type instruments for measuring the current and voltage.

Our account of the research carried out by our composite investigator during the first half of the nineteenth century in studying the laws of direct currents would not be complete if we did not mention that he discovered the interaction of currents. Parallel conductors with currents travelling in the same direction attract each other. They repel each other when the currents travel in opposite directions. Of course, this phenomenon can also be used to measure the current.

I shall certainly not limit myself to the last paragraphs in presenting the laws of electromagnetism; a whole chapter is devoted to this subject. But I had to remind the reader of these important facts to achieve the aim

of the present chapter, which is to show how the basic quantitative concepts and units of measurement describing electrical phenomena—current, charge and field—are introduced.

Stationary Electricity

We shall assume that our idealized investigator has a comprehensive knowledge of the wide variety of phenomena that were said to be electrical in the remote past. The special properties of amber, or a glass rod rubbed with fur, the initiation of sparks jumping between two bodies that have been electrified were studied (or, rather, were used for striking demonstrations) for a sufficiently long time. Naturally, our investigator studying electric currents asked himself: Is the fluid passing through a wire and the one that can exist in a stationary state on some body until the body is "discharged" one and the same "something"?

Even if we digress from the information that has been accumulated previously, is it not necessary to pose the question: If electricity is "something" that flows through a conductor like a liquid, can it be "poured into a glass"?

If our investigator wants a direct answer to this question, he should proceed as follows. First he obtains a current source of sufficiently high voltage (so far we have not mentioned units of measurement, and so the reader should wait a while before asking what we mean by a high voltage, a heavy current, etc.). One of the battery poles is grounded and a tiny hollow bead, made of very thin aluminium foil, is placed on the second pole. This bead is suspended by a silk thread. The same is done with another such bead.

Next the two tiny beads are brought close together (say, to a distance of 2 mm between their centres).

To his extreme delight or amazement (or any other epithet you care to use), our investigator finds that the beads repel each other. From the angle between the threads and the known mass of the beads, he can compute the force acting between the beads.

The investigator also establishes the following fact: if the beads are charged by touching the same pole of the battery, they repel each other. If one bead obtains electricity from one pole and the other bead from the other pole, they attract each other.

This experiment confirms the right to speak of electricity as something behaving like a liquid and shows that we can deal both with electricity in motion and at rest.

Since our investigator can determine the quantity of electricity by weighing the mass of the copper deposited on the cathode, he can find "how much liquid was poured into the glass", i.e. the quantity of electricity taken by the bead from the battery electrode.

First of all, the investigator observes the following. If a charged bead is earthed, or grounded, i.e. connected to the earth by a wire or other conductor, the bead loses its charge. Next, he can show that the charge "drains" through the wire, i.e. that a current passes through the wire. Finally, he can measure the amount of copper deposited on a cathode in an instrument with an electrolyte (conducting solution), inserted into the path of the charge to the earth. This, in other words, is a measurement of the quantity of static (stationary) electricity that was on the charged bead.

This quantity of electricity is called the *charge* on the bead by our investigator who ascribes a sign—positive or negative—depending upon which electrode the electric fluid was taken from.

Now we can commence our next series of experiments. Various quantities of electricity can be obtained by beads or balls of various sizes from the poles of various

batteries. Locating the balls at different distances from one another, he can measure the force of interaction between them. This leads him to the following important law of nature:

$$F = K\frac{q_1 q_2}{r^2}$$

which states that the force of interaction is directly proportional to the product of the charges on the balls and inversely proportional to the square of the distance between them. The reader recognizes it as the expression of *Coulomb's law* which was established differently than we have described. But our investigator is no historical character.

Electric Fields

Our investigator knows of forces of two different kinds. One kind is developed upon direct contact of one body with another. Such is the case in pulling or pushing. With respect to forces that act at a distance, so far he knew only of the force of gravity or, in a wider aspect, the force of universal gravitation.

Thus a new force was added to the remotely acting ones: the force of Coulomb attraction or repulsion between two charged bodies. It closely resembles the gravitational force. Even the formulas resemble each other.

The force of gravity exerted by the earth on a body does not lead to any particular difficulties in calculations. As for Coulomb, or, as they are also called, electrostatic, forces, we may encounter cases in which electric charges are distributed in space in some complex and, even worse, unknown manner.

But we can manage without knowing the way in which the charges are distributed. We know that these charges "sense" one another at a distance. Why not say that the

charges set up an *electric field?* It may seem at first that difficulties arise from the fact that we cannot see the electric field. "But I think," says our investigator, "that the electric field should not be regarded as mathematical fiction for facilitating calculations. If a charge located at some point is subject to a force, this means that this point (in space) is in a special state. An electric field is a physical reality, i.e. it exists by itself even though we cannot see it." Since he made this guess at the beginning of the nineteenth century, our investigator could not prove it. But the future showed that it was correct.

Coulomb's law establishes a formula by means of which we can determine the effect of one small ball on another. One of the balls can be fixed and the other moved to various points in space. At all points the movable (test) ball is subject to a force. Today, this fact is differently formulated: a ball charged with electricity sets up a field of electric forces or, more concisely, an electric field about itself.

The source of an electric field may be charged bodies of any shape. Coulomb's law is not valid in this case, but, using a test ball, we can measure the electric field surrounding a charged body and specify this field in an entirely comprehensive way by indicating the magnitude and direction of the forces. To avoid the dependence of the specification of an electric field on the charge of the test ball, an electric field is characterized by its intensity:

$$E = \frac{F}{q}$$

where q is the electric charge of the test ball.

There is a visual way of describing an electric field by lines of force. Diagrams showing these lines of force may differ greatly depending upon the shape of the charged bodies and their mutual location. The simplest diagrams

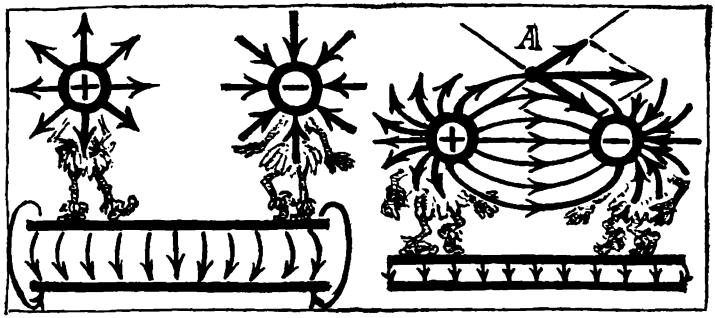


Figure 1.1

of electric fields are illustrated in Figure 1.1. Their meanings are the following. A tangent at any point to a line of force indicates the direction of the electric force at that point. The number of lines per unit of area perpendicular to the lines of force is entirely arbitrary provided that it is proportional to quantity E. When, however, we speak of the number of lines of force without using a diagram, it is assumed that this number is simply equal to E.

If a free electric charge is placed in an electric field, it moves along the lines of force, unless other forces, for instance gravity, interfere.

Of simplest appearance are force fields set up by bodies of spherical shape. If two spheres or two charges that can be regarded as points approach each other, their fields are superimposed. Their field intensities are added by the parallelogram method. We can determine the direction of the line of force at any point A and find the field intensity by constructing the parallelogram of forces as shown in Figure 1.1.

If the charged bodies are in the form of plates, the field resembles that shown below in the drawing. By

bringing the plates closer together and increasing their area, we can obtain almost ideal field uniformity, with only a slight edge, or fringe, effect. It can be said of two closely located plates that they condense the field. Accordingly, such a device is called a condenser. It is now generally known as a capacitor.

As we know, work done in moving a body subject to a force is equal to the product of the force by the length of path. To transfer a charge from one plate of a capacitor to the other along a line of force, the work required is equal to qEl. The work required to carry over a unit quantity of electricity equals El.

Now let us connect the two plates of a capacitor with a conductor. The amount of energy generated when quantity q of electricity is passed through the conductor equals qV. Since we already realize that there is no difference in principle between the motion of a charged ball in an electric field and the passage of the electric "fluid" along the metal conductor, we can equate these two expressions for the energy expended by the field. Thus

$$qEl = qV$$

The validity of this equation can be readily checked by moving the plates of a capacitor farther apart and measuring the force exerted on a test charge.

This measurement can be done by an elegant procedure without resorting to a charged ball suspended by a silk thread.

Everybody knows that light bodies fall considerably slower than heavy ones. It will be recalled that this is what led wise men of antiquity and the Middle Ages to suppose, previous to Galileo's experiments, that the velocity of motion (and not the acceleration) of a body is proportional to the force acting on it. This point of view was proved false only after it had been demonstrated that a piece of paper and a metal ball drop side by side

in a vertical glass tube from which the air has been pumped out. It was found that all bodies gain velocity in falling at the same rate, i.e. fall to the earth with the same acceleration. For our purposes, however, we can utilize the influence of air resistance to make our light hollow metal ball (or bead), used to demonstrate Coulomb's law, fall very slowly.

If this ball is dropped between the plates of a capacitor, we can vary the voltage between the plates until the electric field that is set up stops the ball from falling further. Equilibrium is achieved when the force of gravity equals the field force: $mg = qE$. This equation can be used to determine the field intensity and to prove the validity of our theoretical reasoning.

The number of lines of force passing through any imaginary or real surface in an electric field is selected so that in our chosen units of measurement it equals the electric flux. Can we find the electric flux passing through a closed surface that encompasses charged bodies?

Let us first consider the simplest case. The field is set up by one small ball. We describe a sphere around the ball. If the radius of the sphere is R, then the intensity at any point of the surface on the sphere is Kq/R^2. The area of the sphere is $4\pi R^2$. Thus the electric flux through the sphere is $4\pi Kq$. It is clear, however, that the flux remains the same regardless of the kind of surface we consider.

Next we complicate the picture by assuming that the field is set up by a large number of charged bodies of any shape. These bodies, however, can be imagined to be broken down into tiny portions, each being equivalent to a point charge. We enclose the system of charges with a surface of arbitrary shape. The flux from each charge equals $4\pi Kq$. It is quite natural to assume that the fluxes can be added together arithmetically. Hence, the total flux through any closed surface, encompassing all

the charges, is proportional to the sum of the charges of all the bodies enclosed in this surface.

This statement is the basic law governing electric fields (one of Maxwell's four equations, see Chapter 5).

Note that we neither derived nor proved this formula. We simply guessed that this is the way things are and not differently. This means we are dealing with a general law of nature whose validity is established by an experimental confirmation of any consequences that follow from the general law.

It is vitally important to know the general rule that holds for any system. On the basis of the written law, an electronic computer requires only seconds to calculate the required data on an electric field set up by a most complex system of charged bodies. We shall restrict ourselves to the modest problem of deriving a formula of practical importance for determining the capacitance of a capacitor (demonstrating, at the same time, techniques of theoretical physics applied to an elementary case).

First we define this well-known quantity. The capacitance of a capacitor is the ratio of the charge that accumulates on its plates to the voltage applied across the plates. Thus

$$C = \frac{q}{V}$$

In a capacitor, the lines of force do not pass out beyond the edges of the plates to any appreciable extent. They emerge from the positive plate and enter the negative one. Neglecting distortion of the field at the edges of the capacitor, we can write the product EA for the flux. It follows from the general law that

$$EA = 4\pi Kq$$

from which the field intensity between the plates is

$$E = 4\pi K \frac{q}{A}$$

On the other hand, the field intensity of a capacitor can be written as

$$E = \frac{V}{d}$$

where d is the distance between the plates. Equating the last two equations and recalling that $C=q/V$, we obtain the formula for the capacitance of a capacitor:

$$C = \frac{A}{4\pi K d}$$

In their actual design, capacitors are strips of metal pressed against the faces of strips of mica or paraffin-impregnated paper. The last two are insulating materials. What do we gain by inserting a dielectric between the capacitor plates? Experiments show that the capacitance C of a capacitor is related to its capacitance C_0 when it has no separating medium by the formula $C=C_0\varepsilon$.

The constant ε is called the *permittivity*, or *dielectric constant*. For air, mica, water and Rochelle salt, the permittivities are equal to 1, approximately 6, 81 and 9000.

What Is Basic?

Ohm's law and the Joule-Lenz law relate energy, current, voltage and resistance. We can say that the voltage is equal to the product of the current by the resistance. But we can also say: the current is the voltage divided by the resistance. Both definitions are found in textbooks and both have the shortcoming that they

prove convenient only in cases when Ohm's law is valid. And, as mentioned before, this law is not always true. Consequently, it is better to proceed as we did, namely, to assume that the derived quantity is the resistance of the conductor, and that it is defined as the ratio of the voltage across the ends of the conductor to the current passing through it.

Since the energy of an electric current can be measured on the basis of the energy conservation law—by the thermal or mechanical effects of the current—it is clearly expedient to define the current or voltage as quantities derived from the energy. The most natural procedure is to determine the current from electrolysis phenomena and the voltage across the ends of a portion of a circuit as the quotient of the generated energy divided by the quantity of electricity.

The reader must clearly perceive, however, that this is not the only possible system of determining these electrical quantities. Instead of electrolysis, i.e. deposition of metal on an electrode, the current can be determined on the basis of any other of its effects, such as the effect of a current on a magnetic needle or on another current.

There is nothing wrong in principle with another procedure: we select a certain standard current source and then the voltage of any other current source is determined as the number of equal standard elements. Such a proposal was made, and the standard element is called the Weston cell.

Still another possibility is to devise the system of definitions and units of measurement by selecting a certain standard resistance. Then, as above, we find how many standard resistances can substitute for the given conductor. At one time, a column of mercury of specified length and cross section was used as such a unit of resistance.

It is helpful to understand that the sequence in which physical concepts are introduced is a random matter. It does not in any way, of course, alter the laws of nature.

So far we have dealt with phenomena concerned with direct electric current. Even within this group of phenomena it is feasible to devise various systems of concepts and, consequently, various systems of units of measurement. Actually our choice is even more extensive because electrical phenomena are not at all restricted to those dealing with direct electric current.

Many physics textbooks, including ones recently published, define the concept of the quantity of electric charge (or the quantity of electricity, which is the same thing) from Coulomb's law. Then the textbooks treat of voltage and only later, after finishing the discussion of electrostatics, do they introduce the concepts of the current and electrical resistance. As you have seen, we went about this matter in a different way.

There is even more freedom in selecting the units of measurement. The investigator has the right to proceed as he finds convenient. He should never forget, however, that the selection of the units of measurement affects the proportionality factors he must use in various formulas.

There is nothing wrong (again, in principle) with selecting the units of current, voltage and resistance independently of one another. But this requires introducing into Ohm's law a factor with specified dimensions. Until quite recently, before the habitually familiar calorie had been banished forever from physics by a stern international committee, the formula of the Joule-Lenz law had a numerical factor. This was due to the fact that the units of measurement for current and voltage were defined quite independently of the selected units for energy (heat and work).

In the preceding paragraphs, I have written only

two formulas in the form of proportionalities, all the rest are equations without numerical coefficients. One of these two formulas relates the mass of the substance deposited on the electrode with the quantity of electricity, and the other is Coulomb's law. This was not done by accident, but because physicists seem to be unwilling, so far and to some extent, to go over to the International System of Units, SI, accepted as a law, and continue to employ the absolute physical system (cgs units) in their work (less and less, it is true, due to the pressure exerted on them by the editors of their articles and books). In the absolute physical system, the quantity K in the Coulomb formula for the interaction of charges in a vacuum is set equal to unity. When this is done, the value of the so-called "absolute" unit of the quantity of electricity is predetermined (the charge equals unity if two identical charges, located at unit distance from each other, interact with unit force).

After measuring the mass in grams, it would be necessary, to remain consistent, to calculate the value of factor k in the law of electrolysis, indicating how much matter is deposited on the electrode in the passage of one absolute unit of charge. But it is of no avail to look through your textbooks in search of such a value of this factor. Knowing the obstinate reluctance of engineers to reject the ampere and coulomb, physicists made use in the law of electrolysis of a number indicating the mass of matter deposited when one coulomb of electricity passes through the electrolyte. Thus, two units were given in books for the same quantity. It was also clear that it was convenient to employ one or the other in entirely different cases, because the coulomb is equal to three thousand million absolute units.

It is convenient, of course, to have K equal to unity, but engineers drew attention to the fact that formulas for the electric flux, capacitance of a capacitor and others

include the unwanted coefficient 4π. They contended that it would be a good thing to eliminate it.

As is usual in such controversies, victory was celebrated by those closer to practice than to theory. The system accepted today proceeded along the course that engineering has long followed. Supporters of the SI system insisted on using a single unit of energy in all branches of science and also required that the current, or current intensity, be accepted as the only basic electrical concept.

Thus, we begin the study of electricity with the *joule* as the unit of energy. We select the *coulomb*, equal to the ampere-second, as the unit of the quantity of electricity. We propose that the ampere be defined from the force of interaction of currents. This definition (given on p. 108 in the chapter devoted to electromagnetism) is devised so that factor k in the electrolysis formula remains the same one that everybody has long become accustomed to. Still, we must understand that in the SI system of units this factor *does not determine* the value of the coulomb. With the increase in the accuracy of measurement, we will be obliged to revise this value so as to retain the definition of the ampere (I do not think such a time will come, because I cannot conceive that the measurement of electrodynamic forces can be more precise than the measurement of mass).

The SI system next follows the course I plotted for our composite investigator. First came the unit of voltage, called the *volt*, equal to one joule divided by one coulomb; then the unit of resistance, called the *ohm*, equal to one volt divided by one ampere; the unit of resistivity, one ohm multiplied by one metre, etc.

Now we have reached Coulomb's law and find that we no longer have the right to dispose of factor K. The force is measured in newtons, the distance in metres and the charge in coulombs. Factor K acquires dimensionality and has a certain value determined experimentally.

Coulomb's law is rarely required and the equation for the capacitance of a capacitor is the working formula in many engineering calculations. To eliminate factor 4π in formulas for the electric flux, capacitance of a capacitor and others, engineers replaced factor K by the expression $1/4\pi\varepsilon_0$ a long time ago. For readily understandable reasons, the quantity ε_0 can be called the *electrical permittivity of vacuum*. It equals

$$\varepsilon_0 = 8.85 \times 10^{-12} \frac{\text{coulomb}^2}{\text{newton-metre}^2}$$

Hence, the flux of lines of force is expressed by

$$\frac{1}{\varepsilon_0}(q_1 + q_2 + \quad .)$$

and the capacitance of a capacitor is written as

$$C = \frac{\varepsilon\varepsilon_0 A}{d}$$

The unit of capacitance, one *farad*, equals one coulomb divided by one volt.

Evolution of Electricity Theory

The actual evolution of electricity theory was not at all along the lines followed by our mythical composite investigator.

Electrostatic phenomena were well known in ancient times. It is difficult to say today whether the ancient Greeks knew what bodies in addition to amber (the Greek word for amber, *elektron*, is the source of such words as electricity, electron, etc.) acquire special properties after being rubbed and attract small bits of straw. Not until the beginning of the 17th century did Sir William Gilbert (1540-1603) show that this strange property was possessed by diamond, sealing wax, sulphur, alum and

many other substances. This outstanding English physicist, who was also court physician to Queen Elizabeth I, was evidently the first to devise instruments that could be used to observe the interaction of electrified bodies. By the 18th century it was already known that certain bodies are capable of retaining charges, while the charges "leak away" from other bodies. Few scientists of this century doubted that electricity was something resembling a fluid. The first electrostatic machines were built. They produced sparks and could "shock" a line of people holding hands when one of the end persons touched a terminal of an operating electrostatic machine. It was considered good taste for the courts of many countries to visit the laboratory of a scientist as they would go to a circus. And the scientists, in their turn, made every effort to dramatize the phenomena being demonstrated.

In the 18th century we can already speak of electrostatics as a science. A great many different kinds of electroscopes were invented and made; Coulomb began his quantitative measurements of the interaction forces between charges.

In 1773, the Italian physician and physicist Luigi Galvani (1737-1798) began to investigate the contraction of the leg muscles of a dead frog when a voltage was applied across them.

Continuing Galvani's experiments, Volta came to the conclusion at the end of the 18th century that an electric fluid passed through the frog's muscles. The next remarkable step forward was the invention of the first current source, the galvanic cell, and later, the voltaic pile.

At the very beginning of the 19th century, Volta's discoveries were known by the whole world of science. This started off extensive investigations of electric current. New discoveries followed, one after the other.

A number of investigators studied the thermal effects of current. This is what Oersted was engaged in when, entirely by chance, he discovered the effect of a current on a magnetic needle.

The brilliant investigations of Ohm and Ampère were conducted at about the same time: in the twenties of the nineteenth century.

Ampère's work quickly won him worldwide fame. Ohm had no such luck. His scientific articles, combining neat and elegant experiments with precise calculations, distinguished by their rigorous substantiation and systematic introduction of phenomenological concepts, paying absolutely no attention to the "nature" of things, were not appreciated by his contemporaries. When anyone mentioned Ohm's works, it was only to ridicule the "morbid fantasy of the author, who is trying to belittle the dignity of nature". (These words belong, evidently, to the physicist De la Rive, who made no contribution whatsoever to science.)

It is extremely difficult to read the original works of physicists written in those days. Experimental discoveries were described in language alien to us today. It is even impossible, in many cases, to understand what the author meant by certain words. The names of famous scientists are retained in the memory of posterity only due to the efforts of historians of science.

2. Electrical Structure of Matter

Minimum Quantity of Electricity

For many years, all the information on electrical phenomena known to physicists consisted in the certainty that electricity was something like a liquid. The following joke was still enjoying success at the turn of the century. Wishing to make a laughing-stock of a poorly prepared student, the examiner says, "Well, since you can't answer all my other questions, try this simplest one: What is electricity?"

The student answers, "On my word of honour, I knew, Professor Jones, but I have forgotten."

At this the examiner exclaims, "What a loss for mankind! There was only one person in the whole world who knew what electricity is, and he has forgotten!"

The first hints that electricity consists of special particles instead of being a continuous fluid, and the certainty that these electrical particles are related in some way to atoms were obtained in studying electrolysis.

In conducting experiments on the dissociation of substances dissolved in water when a current is passed through the solution, the English scientist Michael Faraday (1791-1867) found that the same electric current deposits various amounts of substances on the electrodes depending on what chemical compound is dissolved in the water. Faraday discovered that 96 500 coulombs pass through the electrolyte to deposit one gram-atom of a monovalent substance, and that twice as much is required to deposit one gram-atom of a bivalent substance.

Maybe you think that when Faraday obtained these results he cried, "Eureka!" and announced that he had discovered the essential character of electricity? Not at all. The gifted experimenter permitted himself no such illusions. Faraday, in any case, with respect to electric current, behaved like our mythical investigator of the preceding chapter. He considered it feasible to make use of only those concepts that can be characterized by numerical values.

"How so?" the reader may ask. It has been shown that 6.023×10^{23} atoms (you recall that this is Avogadro's number) carry over 96 500 coulombs of electricity. Consequently, if we divide the second figure by the first, we obtain the quantity of electricity carried by any monovalent atom. The quotient is 1.6×10^{-19} coulombs. This then is the minimum quantity of electricity, or the "atom of electricity", or the "elementary charge".

But Avogadro's number was not determined until 1870.

It was only then (just think of it; a mere hundred years ago) that physicists who like to devise hypotheses (their frame of mind and mentality greatly differentiate them from the investigator who tries to keep within the bounds of the phenomenon being studied) decided that the following assumption is highly probable. Along with electrically neutral atoms, particles exist that carry one or several elementary charges of electricity (positive or negative). Atoms carrying a positive charge (*cations*) are deposited on the cathode during electrolysis; those carrying a negative charge (*anions*) are deposited on the anode.

Molecules of salts dissolved in water are dissociated into anions and cations. A molecule of common salt, sodium chloride, for instance, dissociates into a positive ion of sodium and a negative ion of chlorine, rather than into an atom of sodium and an atom of chlorine.

Ion Flow

Naturally, electrolysis only suggests to the investigator the idea that electrical particles exist.

Many procedures were proposed at the turn of the last century for converting molecules into charged fragments (a phenomenon called ionization). Methods were found for producing directed streams of charged particles and, finally, procedures were worked out for measuring the charge and mass of ions. Physicists gained their first knowledge of ion flow when they connected a glass tube with rarified gas into a d-c circuit. At a low voltage across the electrodes sealed into the glass of the tube, no current is observed. But it turned out to be quite simple to convert the rarified gas into a conductor. The gas can be ionized by the effects of X-rays, ultraviolet light or radioactive radiation. We can even manage without such special measures if we apply a higher voltage to the tube terminals.

Gas thus becomes a conductor of electric current. We can assume that the molecules are broken up into anions and cations. The anions travel toward the positive electrode and the cations toward the negative one. An important breakthrough in investigating this phenomenon was the production of a stream of particles. This is done by making a hole in one electrode and accelerating the ions of a single sign passing through the hole with an electric field. By means of a diaphragm, we can obtain a narrow beam of anions or cations travelling at considerable velocity. If such a beam is directed to a screen of the type on a TV picture tube, we see a bright spot. By passing the ion beam through two mutually perpendicular electric fields and varying the voltage over the capacitors setting up these fields, we can make the spot wander over the screen.

Using a similar device, we can determine one of its most important parameters: its charge-to-mass ratio.

In an accelerating field, ions gain energy equal to the work done by the electric forces, i.e.

$$\frac{mv^2}{2} = eV$$

We know the applied voltage, and the velocity of the particles can be measured by various essentially differing ways. We can, for instance, measure the deviation of the spot of light on the screen. It is clear that the longer the path of the particle and the lower its initial velocity, the greater the deviation. This problem can be quite rigorously solved. It resembles the calculation of the path of a stone thrown horizontally.

There are also methods for direct measurement of the time it takes the ion to travel over its whole path.

Hence, we know the voltage and the ion velocity. What can we calculate from the results of this experiment? It follows from the equation that we can find the ratio of the charge of the particle to its mass. It is a pity, however, that we cannot separate the charge from the mass no matter how we change the conditions of the experiment, or what deviations and accelerations of the particles we use. Only on the basis of data obtained by chemists, and the value of the elementary charge obtained in electrolysis, can we come to the reliable conclusion: the charges of all monovalent ions are the same, the charges of bivalent ions are twice as much, those of trivalent ions are three times as much, etc. The differences in the charge-to-mass ratios, which can be measured with exceptional precision, indicate a method for measuring the mass of the ion.

This is why the instrument, of exceptional significance in chemistry and chemical technology and based on the

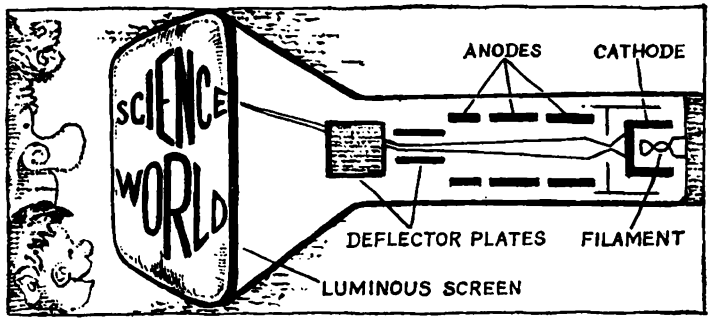


Figure 2.1

principle of the simple experiment described above, is called the mass spectrograph (see Book 4), though, in fact, it measures the charge-to-mass ratio of ions.

Electron Beam

We shall not trace the zigzag course of historical events that led physicists to the unshaken conviction that a minimum portion of electricity not only exists, but has a material carrier, which they called the electron. Instead we shall describe an experiment that is demonstrated in school lessons in physics.

The apparatus used in this experiment was once called a cathode, or cathode-ray, tube. Now it is called an electron-beam tube or an electron gun or an oscillograph. If your school days are far behind and you had no opportunity to become acquainted with this instrument, no matter, you have seen and see plenty of them now: an electron-beam tube is the main component of your television set. On the somewhat flattened end, or screen, of the tube, the electron beam draws the moving pictures that we sometimes view with pleasure and which always enable us to kill time.

Let us return, however, to the school experiment. A diagram of such a tube is illustrated in Figure 2.1. The tube is ideally evacuated; all molecules that could dissociate have been pumped out. However, after a current (called the cathode current) heats the filament to incandescence, the filament heats the cathode. Then, when the cathode and anode are connected to the corresponding terminals of the voltage source, you can observe a bright spot on the fluorescent screen. You can now establish by means of a measuring instrument that an electric current passes from the anode to the cathode. It is quite naturally called the anode current.

Since the current passes through a vacuum, you must come to the conclusion that the incandescent cathode emits negatively charged particles. This is called *thermoelectron*, or *thermionic*, *emission*. Any incandescent body has this property.

These particles (we shall not hide from the reader that they are electrons) travel toward the anodes, which are designed as cylinders closed at one end and with a tiny round hole in the centre of the bottom. The electrons are emitted as a narrow beam, which can be investigated in the same way as described above for a beam of ions.

Convinced by the spot on the luminous screen that the cathode is emitting electrons, we proceed to find their charge-to-mass ratio by means of deflector plates. The results are the following. This ratio for the electron is 1840 times greater than for the lightest ion, namely, the hydrogen ion. This leads us to the conclusion that the hydrogen ion is 1840 times heavier than an electron. Consequently, the mass of the electron is 9×10^{-28} grams.

Here our reader may rightly object, saying that we are in too much of a hurry. On the face of it, you cannot come to the conclusion that the mass of an electron is less than that of an ion only by measuring their charge-

to-mass ratio. Maybe the charge of the positively charged ion and that of the electron are entirely different?

The charge-to-mass ratio of the electron was determined way back at the end of last century by the brilliant English physicist Sir Joseph John Thomson (1856-1940). His friends called him J. J. (This abbreviation, so often found in memoirs and biographies, is due less to the fact that the English like abbreviations so much as to the fact that another famous British physicist of the same name lived and worked during the 19th century. He was William Thomson (1824-1907). A peerage was bestowed upon him for his scientific merits and he became Lord Kelvin.) The cathode tube used by J. J. Thomson was, of course, not as highly perfected as an up-to-date oscillograph. Thomson was perfectly aware that his measurements only showed that the discreteness of electric charge was quite feasible and that a minimum portion of electricity could exist.

Strange as it may seem, even though many physicists had observed the behaviour of cathode and anode beams, there were still many supporters of the hypothesis that these beams were of a wave nature. These investigators found no necessity for admitting that the currents passing through a metal conductor, through a liquid, and through a gas or vacuum are near of kin. They insisted on more direct proofs. We can, of course, understand their point of view: circumstantial evidence is insufficient to convert a hypothesis into a fact.

Hence, it was essential, first of all, to substantiate the validity of the hypothesis by direct measurements of the size of the electric charge of particles. These measurements, far from unsuccessful, were begun by Thomson and his colleagues in the early years of the 20th century. The most precise measurements were made in 1909 by the American physicist Robert Andrews Millikan (1868-1953).

Millikan's Experiment

The idea of the discreteness of electricity seems quite bold, and the determination of the elementary charge, with an account of which we began this chapter, can be treated in a different way. What is wrong with contending that anions actually exist and that negative electricity is a liquid entrained by the positive ion? One ion picks up one amount of this liquid, another ion picks up another amount, and the experiment indicates some average value. This is a reasonable explanation, is it not?

As mentioned above, Thomson's experiments were a powerful, but not decisive, argument in favour of the existence of the electron. No need then to prove the vital importance to physics of an experiment that could demonstrate the existence of the elementary charge of electricity so obviously that all doubts were instantly discarded. Such an experiment was devised in 1909 by Robert Millikan. I shall not mention here the other works of this famed scientist, but this single investigation was sufficient to put his name in all physics textbooks.

The principle of this ingenious experiment is based on a simple fact. Just as a glass rod rubbed by a piece of fur acquires electrical properties, other substances behave in a similar way. This is called electrification by friction. But, strictly speaking, what reason have we to think that this property is inherent in only solid bodies? If we spray tiny droplets of oil into a chamber, will they become electrified by the friction as they pass the orifice of an atomizer? We find that they will. This can be demonstrated by what is, in principle, quite simple apparatus. We spray a stream of fine oil droplets into the space between two horizontal capacitor plates enclosed in a chamber. Then we arrange a suitable microscope to enable us to observe the droplets. Until an

electric field is set up, the droplets naturally fall downward by gravity. The droplets are very light and the force of gravity is almost immediately counterbalanced by the air resistance so that they drift downward at uniform velocity. As soon as we apply a voltage across the plates, we observe a definite change in the behaviour of the droplets. Their motion is either accelerated or decelerated, depending upon the direction of the electric field. Millikan chose the direction that slowed down the droplets. By gradually increasing the intensity of the field, he could, so to speak, hold the droplet at rest between the plates. For hours, he observed a single oil droplet. Varying the field, he could control its motion or stop it at will.

What can we calculate by means of this experiment? First we shall discuss the data obtained in observation before setting up the electric field. The equality of the force of gravity and the air resistance can be expressed by the equation

$$mg = av$$

The density of the oil can be determined by an independent experiment; the diameter of the droplet is measured by the microscope. After this the mass of the droplet is readily calculated. The droplet drifts slowly downward and by engraving scale divisions on the microscope lens we can measure the velocity v of the droplet with a stopwatch with a sufficient degree of accuracy. Then, substituting these quantities into the preceding equation, we obtain the resistance coefficient a.

Next we switch on the field. It proves most convenient to select a field intensity for which the droplet begins to rise at uniform velocity. We have thus added a third force to the two previous ones; it is the force exerted by an electric field whose intensity E is known (the ratio of the voltage across the plates to the distance between

them). Upward motion at uniform velocity means that the three forces counterbalance one another. The condition for such equilibrium is given by

$$qE - mg = av'$$

The new value v' of the velocity can be measured with the same microscope and stopwatch. All the quantities in the equation are known now except the charge of the droplet. We next calculate the charge and enter the value in a journal for experimental data of the kind kept by all scrupulous experimenters.

Now we have come to the main idea of the oil-drop experiment. The current in an electrolyte, reasoned Millikan, is carried by ions of different signs. But ions can be formed in a gas as well. Air can be ionized by various procedures. We can, for instance, place our whole arrangement near an X-ray tube. Then the X-rays ionize the air. This was well known in 1909. But if a droplet is charged it attracts ions of the opposite sign. As soon as an ion becomes attached to a droplet, the charge of the latter is changed, and, with its charge, its velocity is also changed. This new velocity is immediately determined by another measurement.

Observations proved the correctness of the idea. When the X-ray tube was switched on, every now and then various droplets would abruptly change their velocity. Keeping his eye on a single droplet, the observer measured the difference in velocities before and after the X-rays were switched on. From the formula given above, the value of q was readily found.

Have you not guessed yet what all this is being done for? Think again. If there is such a thing as an elementary electric charge, then the measured value should equal it if a single monovalent ion joins the droplet or be a multiple of the elementary charge if several ions become attached to the droplet.

Conducting his experiment with droplets of oil, water, mercury and glycerine, and reversing the charge of the droplets, Millikan filled his data journal with hundreds of values of q, and they all were equal to or a multiple of a single value, the same that had been found by the investigators of electrolysis.

When Millikan published his results, even his most skeptical opponents were convinced that the electric charge is found in nature in discrete portions. Strictly speaking, however, Millikan's experiments do not directly prove the existence of the electron as a particle.

But hypotheses foreshadow facts. Some scientists were sure that electricity is of a granular nature as far back as the beginning of the 19th century. The charge of the ion was first calculated in 1881 by the Irish physicist George Johnstone Stoney (1826-1911) and ten years later he first proposed the term *electron*, not for the particle, but for the charge of a monovalent negative ion. J. J. Thomson's experiments compelled the great majority of physicists to believe in the existence of the electron as a particle. The German physicist Paul Karl Ludwig Drude (1863-1906) was the first to unambiguously define the electron as a particle carrying an elementary charge of negative electricity.

Thus the electron had been generally recognized before it was "seen".

Direct proof of the existence of electrons was obtained later by precise experiments. A weak beam of these particles was directed onto a screen where they can be counted one by one. Each electron causes a flash on a luminescent screen. Luminescent screens have long been replaced for this purpose by special counters named after their inventor, the German physicist Hans Wilhelm Geiger (1882-1945). In a word, the idea of such a counter is that a single electron, like the trigger of a revolver, initiates a strong current pulse, which can be readily

registered. This enables the physicist to establish the number of electrons entering a trap per second. If this trap is a metallic bulb into which the electrons fly, the bulb is gradually charged with a quantity of electricity large enough to be precisely measured. To find the charge of the electron it is sufficient to divide the quantity of electricity by the number of captured electrons.

Only after this can we contend that the existence of the electron is no longer a hypothesis; it is a fact.

We have reviewed at racing-car speed the discoveries that laid the foundations of modern physics. Such, however, is their fate. New matters crowd out the old, and even cardinal events, occurring during the construction of the cathedral of Science, become only material for historians.

Now, perhaps, we can answer the question: What is electricity? An electric fluid is a current of electrical particles. A body is electrically charged when the number of particles of one sign exceeds that of the other sign.

"Well, what a feeble explanation," retorts our indignant reader. "And what then is an electrical particle?"

"Isn't it perfectly clear? A particle is said to be electrical if it interacts according to Coulomb's law."

"And is that all?" asks the puzzled reader.

"All that concerns the answer to your question," answers the physicist. "But ahead of us are answers to many other interesting questions. We have not yet mentioned the cases in which we shall find elementary particles of positive electricity. We shall also find out that electrical particles are characterized by other properties besides their charge and mass."

First, however, we shall discuss the structure of the atom.

Model of the Atom

How is an atom built up of electrical particles? The answer was obtained by means of rays emitted by radium. We shall discuss this wonderful substance and the extensive family of natural and artificial radioactive elements in Book 4 of this series. For the time being it is sufficient for us to know that radium constantly emits penetrating electromagnetic radiation (gamma rays), a stream of electrons (formerly called beta rays) and alpha rays, which consist of doubly charged ions of helium atoms.

In 1911, the eminent New Zealand-born physicist Sir Ernest Rutherford (1871-1937) proposed the so-called planetary model of the atom, based on his careful investigations of the scattering of alpha particles by various substances. Rutherford conducted his experiments with thin gold leaf, only a tenth of a micron in thickness. He found that out of 10 000 alpha particles only one was deflected through an angle exceeding 10°.

In these strikingly simple experiments, the passage of each separate particle was recorded. Up-to-date techniques, of course, enable such measurements to be made entirely automatically.

It becomes clear then that the atom consists mainly of empty space! The rare head-on collisions should be interpreted as follows: inside the atom there must be a positively charged nucleus. This nucleus is surrounded by electrons. They are extremely light and therefore present no appreciable obstacles to the alpha particles. The electrons slow down the alpha particles, but a collision with each separate electron cannot deflect the particle from its path.

Rutherford assumed that the forces of interaction between the atomic nucleus and alpha particle, which are of like charge, are Coulomb forces. Assuming further

that the mass of the atom is concentrated in its nucleus, he calculated the probability of the particle being deflected through the given angle and obtained splendid coincidence of theory with experimental data.

That is how physicists check the models they have devised.

"So the model predicts the experimental results?"

"Yes."

"Which means that it represents reality?"

"Why so categorical? A model explains a series of phenomena and so is a good one. Its refinement is a matter for the future."

The results of Rutherford's experiments swept aside all doubts as to the validity of the following statement: electrons are subject to Coulomb forces and travel about the nucleus.

The theory also led to certain quantitative estimates which were later confirmed. The size of the very smallest atomic nuclei turned out to be about 10^{-13} cm while the size of the whole atom is of the order of 10^{-8} cm.

Comparing the experimental results with the calculations, physicists were able to estimate the charges of colliding nuclei. These estimates played a vital, if not the main, role in interpreting the periodic law of the structure of the elements.

Thus, we have built the model of the atom. But the following question is immediately posed: Why don't the electrons (negatively charged particles) fall into the nucleus (positively charged)? Why is the atom stable?

"There is nothing strange about this," reasons the reader. "The planets do not fall on the sun. A force of electrical origin, like the gravitational force, is centripetal and provides for circular motion of the electrons about the nucleus."

But the point is that the analogy between the planetary system and an atom is only superficial. As we shall

find out further on, an atom must emit, from the point of view of the general laws of electromagnetic fields, electromagnetic radiation. But this is something we can ascertain without a knowledge of the theory of electromagnetism. Matter, i.e. atoms, is capable of radiating light and heat. Since this is so, the atom loses energy and the electron must finally fall into the nucleus.

What is the way out of this dilemma? It is quite simple. We must reconcile ourselves to the facts and elevate them to the rank of a law of nature. This step was taken in 1913 by one of the greatest physicists of our century, Niels Henrik David Bohr (1885-1962).

Quantizing Energy

Like all first steps, this step was relatively timid. We shall now present the new law of nature that not only saved Rutherford's atom, but forced us to the conclusion that the mechanics of large bodies is inapplicable to particles of small mass.

Nature is arranged so that a number of mechanical quantities, such, for instance, as angular momentum or energy, cannot have a continuous series of values for any system of interacting particles. On the contrary, the atom, now being discussed, or the atomic nucleus, to be discussed later, has its own sequence of energy levels inherent in only the given system. There is the lowest level (ground state). The energy of the system cannot be less than this value. In the case of the atom this means that there is a state in which the electron is at a certain minimum distance from the nucleus.

Changes in the energy of an atom can occur only in jumps. If the jump is "upward", it means that the atom absorbed energy; if it is "downward", the atom radiated energy.

We shall see further on how nicely, on this basis, the

radiation spectra of various systems can be deciphered. The formulated law is called the *law of the quantization of energy*. We can also say that energy is of a quantum nature.

It should be noted that the law of quantizing is of an entirely general nature. It is applicable not only to the atom but to any item consisting of thousands of millions of atoms. When we deal with large bodies, however, we may not "notice" the quantizing of energy. The point is that for an item consisting of a million million million atoms the number of energy levels are increased, roughly speaking, by a million million million times. These energy levels are so close to one another that they practically merge into a single band. Consequently, we cannot observe the discreteness of the energy values. Thus, the mechanics we discussed in Book 1 remains practically unchanged when we deal with large bodies.

In Book 2 we found that energy can be transferred from one body to another in the forms of work and heat. Now we are in a position to explain the difference between these two forms of energy transfer. In mechanical action (for example, in compression), the energy levels of the system are displaced. This displacement is very slight and can be detected only by precise experiments if the pressure is high enough. As to heat action, it consists in converting a system with a lower energy level to a higher one (in heating) or one with a higher energy level to a lower one (in cooling).

The quantizing of energy, like that of other mechanical properties, is a general law of nature. A great variety of consequences that follow strictly from this law can be confirmed experimentally.

Maybe you wonder why energy is quantized? There is no answer. Nature just happens to be made that way! Any explanation is a reduction of a particular fact to a more general one. We know today of no statement

sufficiently general for the quantizing of energy to follow from it as a consequence. This does not imply, of course, that at some future date, in principle at least, sufficiently broad and comprehensive laws will be discovered for the principles of quantum mechanics to be their corollaries. However that may be, the law of quantizing is today one of the few great laws of nature that require no logical substantiation. Energy is quantized because—it is quantized.

This law was established in such a general form during the years 1925-1927 in the works of the French physicist Louis Victor Pierre Raymond, Prince de Broglie (b. 1892) and the Austrian and German physicists Erwin Schrödinger (1887-1961) and Werner Karl Heisenberg (1901-1976). The branch of physics based on the quantizing principle (I forgot to mention that the word quantum means a specified quantity or portion) was named *quantum*, or *wave*, *mechanics*. But why wave? This will be explained later on.

Mendeleev's Periodic Law

In 1869, the famed Russian chemist Dmitri Ivanovich Mendeleev (1834-1907) published the periodic law he discovered in the order of the chemical elements. We shall not give Mendeleev's periodic table here; it can be found in any school chemistry textbook. We need recall here only that when he arranged the elements in the order of their atomic weights Mendeleev noted that their chemical properties and certain physical features vary periodically in accordance with their atomic weight.

In the table devised by Mendeleev, each element belongs to one of nine groups and to one of seven periods. He arranged the elements of the groups in columns so that the elements whose symbols are arranged one below the other possess the same chemical properties. He found

that this could be done only if he assumed that certain as yet undiscovered elements existed. He left empty spaces (gaps) for them in his table. The exceptional foresight of this great scientist also led him to put the nickel atom in its "proper" place following cobalt, notwithstanding the fact that the atomic weight of cobalt is somewhat higher.

Some of the gaps were filled during Mendeleev's life. This brought him world renown because it became clear that the compilation of this table was the discovery of a great law of nature rather than a mere formality.

The essence of the atomic number assigned by the table to each chemical element became evident after the physicists had no more doubts about the validity of Rutherford's planetary model of the atom and of the quantizing of energy. What does this number mean? The answer turned out to be extremely simple. The number of an element in the periodic table, its atomic number, is equal to the number of electrons rotating about the nuclens. We can say that the atomic number of an element is the positive charge of its nucleus expressed in units of charge of its electrons.

Mendeleev's periodic law, the principle of energy quantization and a study of the characteristic optical and X-ray spectra of atoms (to be discussed further on) cleared up the reason for the same chemical behaviour of atoms in any one column of Mendeleev's table.

The energy of an atom is the energy of interaction between the electrons and the nucleus. Since energy is quantized, it is logical to assume that the electrons of each atom can be arranged in a series according to their energies. The first electron is most strongly bonded to the nucleus, the second, more weakly, and the third, still more weakly, etc. Thus the electrons of an atom are arranged in energy steps. We have not been misled by our logical reasoning, but investigations have more ac-

curately defined this concept. In the first place, each energy step can be occupied by two rather than a single electron. True, these electrons are identical; they differ from each other in a property known as *spin*. This is a vector quantity. Hence, people who prefer a more visual concept can imagine that a filled step has two "tiny dots" with arrows, one arrow pointing "downward" and the other "upward".

The word "spin" is of the following origin. One of its definitions is to rotate swiftly, like a top. To explain the difference between two electrons occupying the same step, one was to imagine one electron as rotating clockwise and the other counterclockwise about their axes. This model was suggested by the superficial resemblance between the atom and the planetary system. If an electron is something like a planet, then why not allow it to rotate about its axis? Again I must distress my readers: it is quite impossible to visualize the spin of an electron. But it can be measured, as we shall see in the next chapter.

This, however, is not the only significant conclusion that we can reach (by a careful examination of atomic spectra). A second conclusion is that the energy steps can be at unequal distances from one another and can be divided into groups.

Following the first step, called the *K*-level, is an energy gap. This is followed by a group of eight electrons, denoted by the letter *L*, and then a group of 18 electrons, denoted by the letter *M*, etc. We shall not describe here the locations of the levels and the order in which they are filled for all atoms. It is not a simple picture and its description would require much space. Details play no appreciable role in our small book. I mentioned the energy steps only to explain the similarity of atoms located in a vertical column of Mendeleev's table. It was found that they have an equal number of electrons in the upper group of steps.

This clears up the chemical concept of the valency of an atom. Thus, lithium, sodium, potassium, rubidium, cesium and francium have a single electron in the upper group of steps. Beryllium, magnesium, calcium, etc. have two electrons in the upper group of steps. The valence electrons are the most weakly bonded in the atom. Therefore, in ionizing atoms of the first column, singly charged particles are most easily formed. Ions of beryllium, magnesium and others usually carry two charges, etc.

Electrical Structure of Molecules

Chemists call the molecule the tiniest representative of a substance. Most physicists use this word only when this tiniest representative really exists as a separate small body.

Do molecules of common salt exist? "Of course!" answers a chemist and writes the formula NaCl. Common salt is sodium chloride. A molecule consists of one atom of sodium and one of chlorine. This answer, however, is only formally valid. Actually, we cannot find pairs of these atoms behaving as a single whole neither in crystals of common salt, nor in a solution of salt in water, nor in the vapour of sodium chloride. As we mentioned in Book 2, each atom of sodium in a crystal is surrounded by six chlorine neighbours. All of these neighbours possess equal rights and there is no way to find which one "belongs" to the sodium atom.

Let us dissolve common salt in water. We find that the solution is an excellent conductor of electric current. Rigorous experiments, discussed above, can be conducted to demonstrate that the electric current is a stream of negatively charged atoms of chlorine travelling in one direction and a stream of positively charged atoms of sodium travelling in the opposite direction. Hence, in

a solution, atoms of chlorine and sodium do not form strongly bonded pairs of atoms.

After we have established the model of the atom, it becomes clear that an anion of chlorine is an atom of chlorine with an "extra" electron. A cation of sodium, on the contrary, has a "shortage" of one electron.

Can this lead us to the conclusion that a solid body is also built up of ions rather than atoms? Yes, it can. This can be proved by many experiments on whose description we shall not dwell here.

What about the vapour of sodium chloride? Here again we find no molecules. The vapour of sodium chloride consists of ions or of various extremely unstable groups of ions. We can speak of the molecules of ionic compounds only in the chemical meaning of the word.

Ionic compounds all dissolve in water. Such solutions, classical examples being simple metal salts, such as sodium chloride, have good conductivity and are therefore called strong electrolytes.

We can give several examples of substances built up of genuine molecules, molecules in the physical meaning of the word. They are oxygen, nitrogen, carbon dioxide, hydrocarbons, carbohydrates, steroids, vitamins, etc. This list could be continued to great length.

All classifications are always somewhat arbitrary. Therefore, I must warn the reader that sometimes we find cases in which the substance consists of physical molecules when it is in one state of aggregation, but does not when it is in other states. These substances include such vital ones as water. Water vapour molecules are undoubtedly separate bodies. But it is a difficult matter to "outline" a single molecule in ice crystals and to contend that these atoms of hydrogen are bonded only to this particular atom of oxygen.

Be that as it may, the class of molecular crystals is vastly extensive.

In Book 2 we have already discussed the structure of molecular crystals. Recall that in a crystal of carbon dioxide (CO_2) the carbon atom has two very close oxygen neighbours. In all other cases, in investigating the structure of a molecular crystal, we can readily see that it is possible to break down the crystal into intimately arranged groups of atoms.

If they are so closely located, they must be bonded by high forces. This is true. The forces bonding atoms belonging to a single molecule are, roughly speaking, a hundred or even a thousand times greater than the forces acting between atoms of neighbouring molecules.

Just what is this intramolecular bond? It is sufficiently clear that we shall not be able to manage by using only concepts of the mutual attraction of electrically charged negative and positive ions. We know that the molecules of oxygen, nitrogen and hydrogen consist of identical atoms. It is impossible to assume that one atom loses and the other atom gains an electron. Why on earth should an electron prefer to stay near either one of two identical atoms?

The principle of the intramolecular bond was grasped only with the evolution of quantum mechanics. We have just mentioned that the energy of any system is quantized and that two electrons of opposite spin can occupy a single energy level. In addition, an interesting consequence follows from one of the basic hypotheses of quantum mechanics. It was found (and this is no hypothesis, but a strict mathematical derivation which we do not give because of its complexity) that the lowest energy value assumed by an electron depends upon the size of the region within which it travels. The greater the size, the lower the energy of this "zero level".

Now imagine that two atoms of hydrogen approach each other. If they merge into a single system, the "apartment" for each electron is approximately doubled. Two

electrons with opposite spin can peacefully coexist in a single "apartment". Consequently, such life together is expedient. The region in which the electrons exist has appreciably increased. This means that the total energy of the system has decreased after the two atoms merge together. By now we have completely mastered the fact that any system, if there are any such possibilities, tends to go over to the state with minimum energy. For the same reason, a ball, left to itself, rolls downhill.

Thus, the formation of a chemical bond implies the collectivization of the electrons. There is a certain number of electrons (said to be inner-shell electrons) that rotate about the nuclei of the atoms, but some (outer-shell) electrons include at least a pair of nearest atoms in their motion or even may travel among all the atoms of the molecule.

We can recognize a substance built of molecules by its electrical properties. A solution of such a substance conducts no current. The molecules do not break down into parts and a whole molecule is electrically neutral. In liquids and vapours, the molecules retain their structure; the whole group of atoms travels as a single whole, with translatory or rotary motion. Atoms belonging to such molecules can only vibrate about their equilibrium positions.

A neutral molecule carries no electric charge. But do not hasten to the conclusion that such a molecule does not set up an electric field. If the molecule is asymmetric, the centres of gravity of its positive and negative charges almost certainly do not coincide. It is intuitively clear that the centres of gravity of the two charges coincide in such molecules as oxygen and nitrogen, consisting of two identical atoms. Nor is it difficult to understand that in a molecule of a gas such as carbon monoxide (CO) these centres may be displaced with respect to each

other. If there is such a displacement, the molecule is said to have a dipole moment.

This term has the following origin: a dipole molecule behaves like a system of two point charges (one point is the centre of gravity of the negative charges and the other is the same for the positive charges). A dipole is specified by the size of the charge and the dipole arm, i.e. the distance between the centres of gravity.

Do not require proof that an asymmetric molecule has an electric dipole moment. We can dispense with a theoretical discussion because the existence of a permanent (or, as they also say, rigid) dipole moment can be readily demonstrated by experiments.

Dielectrics

We can place equality signs between a dielectric, a nonconductor of current and an insulator.

Dielectrics include molecular gases, molecular liquids and solutions of solids made up of molecules. Solid dielectrics include glass, both organic and nonorganic (silicate, borate, etc.); polymer substances, built up of giant molecules; plastics, molecular crystals, as well as ionic crystals.

We reminded the reader in Chapter 1 that the capacitance of a capacitor increases when we insert any dielectric into the space between the plates. Imagine that the capacitor was connected to a source of constant voltage. The capacitance increased even though the voltage remained the same. This means that an additional charge approached the capacitor plates. It would seem that the field intensity must also increase. It has not changed, however, because it is the quotient of the voltage divided by the distance between the plates. What is the way out of our quandary? The only way is to assume that

an electric field of the opposite direction is set up in the insulator. This is called the *polarization* of the dielectric.

What are the peculiar charges that are produced inside a dielectric? How can we interpret the failure of attempts to draw the charge of a dielectric off to the earth? Even with no knowledge of the electrical structure of matter we can say that these charges are "bound" instead of being free as in a metal. But when we have at our disposal information on the structure of the molecule, we are capable of comprehensively explaining polarization phenomena and the mechanism of formation of the counter field. All other things being equal, the higher the permittivity, or dielectric constant, ε, the stronger this field.

Before going any farther, we must answer the question: What can an electric field do to an atom or a molecule? An electric field can shift the electrons of a neutral atom or ion in the direction opposite to that of the field. The atom or ion is thereby converted into a dipole and sets up a field in the opposite direction. Thus, the polarization of a substance results from the polarization of the atoms, ions or molecules of which it is made.

The polarization mechanism just described is called the process of producing induced dipoles. If there is no field, there are no dipoles. The stronger the field, the greater the displacement of the centre of gravity of the electrons, the higher the induced dipole moment, and the more the polarization.

The formation of induced dipoles cannot depend upon the temperature. Experiments indicate that there are dielectrics which are not influenced by temperature. Consequently, the above-described mechanism is valid for them.

Well, and what can we do about the cases in which the permittivity distinctly depends upon the temperature? A careful investigation of the relationship between

the structure of molecules and the behaviour of the given substance in an electric field, as well as the nature of the temperature dependence of ε (polarization always drops with a rise in temperature) lead us to the following conclusion. If molecules have a dipole moment even in the absence of a field (permanent dipole) and can change their orientation, then this explains the temperature dependence of permittivity.

As a matter of fact, the molecules are oriented haphazardly when there is no field. The dipole moments are added vectorially. Hence, the resultant moment for a volume containing many molecules equals zero. An electric field seems to "comb" the molecules, making them face mainly in one direction. They are subject to two antagonistic forces: thermal motion, introducing disorder in the arrangement of the molecules, and the ordering effect of the field. It can be readily understood that the higher the temperature, the more difficult it is for the field to "handle" the molecules. From this it follows that the permittivity of such substances drops with an increase in temperature.

These ideas can be more easily visualized and remembered by studying the drawings in Figure 2.2. The upper drawing shows that the polarization of an atom consists in the displacement and deformation of its electron shells. The farther an electron from the atomic nucleus, the greater the effect of the field on this electron. The layers represented in these schematic drawings by dots show where the electrons are located. It must be borne in mind that the drawings are extremely rough approximations because various electrons have differently shaped regions in which they exist in molecules (cf. p. 122).

The middle drawing illustrates the behaviour of a symmetric diatomic molecule. In the absence of a field it has no moment. The field induces an electric moment. This moment may vary in magnitude in accordance with

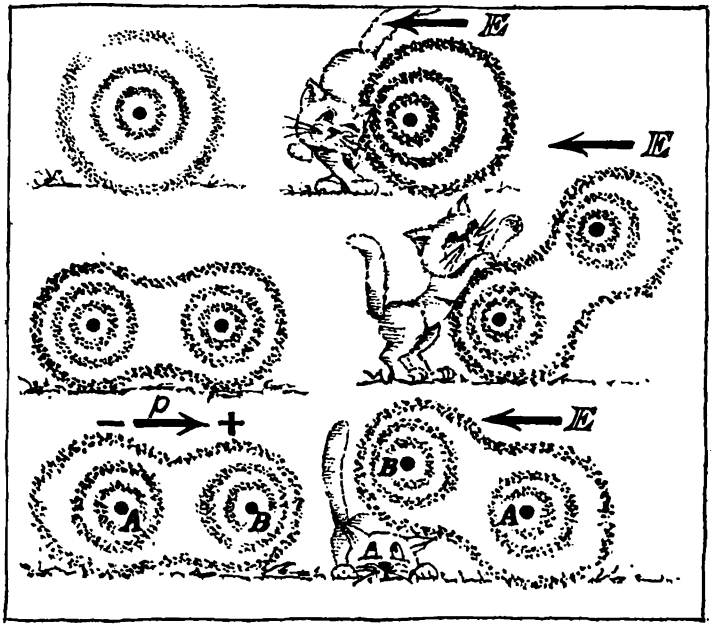


Figure 2.2

the angle of the molecule with respect to the field. The moment is formed due to the deformation of the electron shells.

Finally, the lower drawing shows the behaviour of a molecule which has a dipole moment even when there is no field. In our case, the molecule was simply turned when a field was applied. In the general case, however, both mechanisms of polarization are found in substances whose molecules have a dipole moment in the absence of a field: in addition to rotation of the molecule its electrons may also be displaced. These two effects can

be readily distinguished by making measurements at very low temperatures where the thermal motion has practically no influence on the molecules.

If our model is valid, we should observe no temperature dependence in the permittivity of substances having symmetric molecules, for example, molecules of oxygen or chlorine. On the other hand, if the diatomic molecule consists of two different atoms, for instance, a molecule of carbon monoxide (CO), the permittivity ε should be temperature-dependent. This is actually the case. Nitrobenzene is one of the substances whose molecules have a very high dipole moment.

What will happen to ordinary dielectrics when the strength E of the applied electric field is increased? Evidently, the polarization of the substance is increased as well. This takes place due to stretching of the dipoles: in an atom this is a displacement of the electron cloud with respect to the nucleus; in a molecule it may be due to the pulling apart of two ions. This naturally poses the question: Up until what point does an electron, pulled far away from the nucleus by the applied field, remain an electron of the given atom, or do two ions, pulled sufficiently far apart, still form a molecule? Such a limit doubtlessly exists, and, at a sufficiently high field strength, or intensity, E, the so-called breakdown of the dielectric occurs. The order of magnitude of this field intensity is several thousand kilovolts per metre. In any case, such breakdown implies the release of electrons or ions, i.e. the production of free current carriers. The dielectric ceases to be a dielectric; it conducts electric current.

We most frequently find such a breakdown when a capacitor is disabled in a television or radio set. We can, however, observe another kind of breakdown: an electric discharge in a gas. Electric discharges in gases are to be specially discussed later. For the present, we shall become

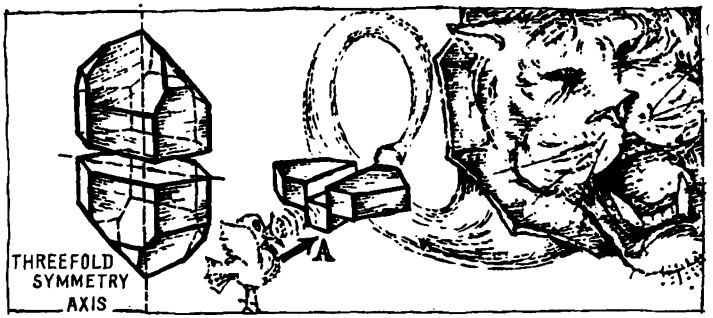


Figure 2.3

acquainted with two distinguished members of the family of dielectrics: piezoelectric and ferroelectric crystals.

Quartz is the chief representative of the class of piezoelectric crystals. Members of this class (which includes, besides quartz, such substances as sugar and tourmaline) must have a definite symmetry. Shown in Figure 2.3 is a crystal of quartz. The principal axis of this crystal is a threefold symmetry axis. Three twofold symmetry axes lie in a perpendicular plane.

A plate about 2 cm thick is cut out of this crystal as shown in the drawing. We see that it is perpendicular to the principal axis and that the twofold symmetry axes lie in a plane of this plate. Then a thin wafer, about 0.5 mm thick, is cut out of the thick plate in a direction perpendicular to one of the twofold symmetry axes. Interesting experiments can be conducted with the thin piezoelectric wafer obtained in this manner (it is shown displaced downward in the drawing in the middle of Figure 2.3).

Let us compress the wafer in the direction of arrow A, perpendicular to the symmetry axes, and connect an electrometer (an instrument used to detect electric charges)

to the side surfaces of the wafer. These surfaces, actually faces of the wafer, must be silver-plated to ensure proper electric contact. We shall find that compression has induced unlike charges on the faces of wafer. If tension is applied instead of compression, the signs of the charges are reversed: where a positive charge appeared in compression, a negative charge appears in tension and vice versa. This phenomenon—the induction of electric charges by applying pressure or tension—is called *piezoelectricity*.

Piezoelectric quartz crystal devices are exceptionally sensitive: electric devices enable us to measure charges induced on the quartz by extremely low forces that could not be measured in any other way. A piezoelectric quartz crystal is also capable of detecting extraordinarily rapid variations in pressure, a feature unattainable with other instruments. Consequently, the effect described above is of immensely practical significance as a method of electrically detecting all kinds of mechanical action, including sounds. It is sufficient to blow softly at a piezoelectric quartz wafer, and the electric indicating instrument instantly responds.

Piezoelectric quartz wafers are used in medicine to listen to murmurs of the heart. They are also employed in a similar manner in engineering to test the operation of machinery in which they can detect all "suspicious" noises.

Quartz is used as a source of the piezoelectric effect in the tone arms (pickups) of record players. The motion of the needle in the groove of the record leads to compression of the piezoelectric crystal, which, in turn, produces the electric signal. The electric current is amplified and is fed to a dynamic loud-speaker to be converted into sound.

So far we have discussed substances that are electrically polarized by an electric field and (in certain cases) by mechanical deformation. When the external effect is

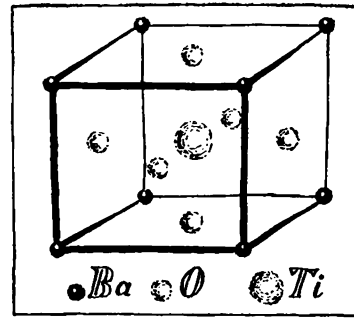


Figure 2.4

removed, the substance becomes electrically neutral again. But, along with this widespread behaviour, we find certain bodies possessing a total electric moment in the absence of external forces. We obviously cannot find such substances among liquids and gases because thermal motion, which is not withstood by the ordering effect of a field, must inevitably lead to disorder in the arrangement of the dipole molecules. Crystals, however, can be conceived of in which the atoms are arranged so that the centres of gravity of the anions and cations within each unit cell are identically displaced. Then all the dipole moments face the same direction. Such substances could be expected to possess maximum possible polarization and, consequently, a permittivity of huge value.

Such crystals do exist. The phenomenon was first discovered using Rochelle salt and this class of substances are called *ferroelectrics* (or *ferroelectric crystals*).

Of greatest practical value among the ferroelectrics is barium titanate. Using it as an example, we shall discuss the exceptionally singular behaviour of this class of substances.

The unit cell of a barium titanate crystal is illustrated in Figure 2.4. We have chosen the barium atoms for the corners of the cell. The small light-coloured spheres are

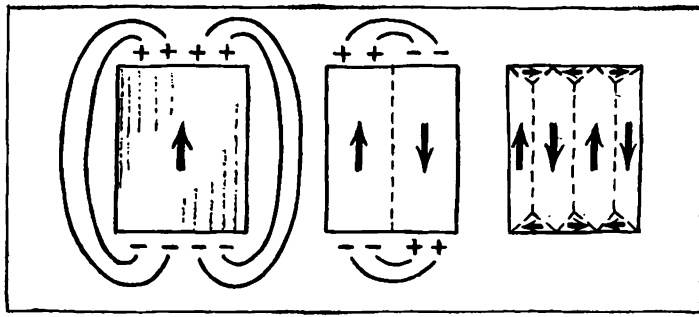


Figure 2.5

anions of oxygen, and the large sphere in the centre is a cation of titanium.

The drawing looks as if the cell were a cubic one. A strictly cubic cell of the substance actually does exist, but only at temperatures above 120 °C. Obviously, a strictly cubic cell is symmetrical and can have no dipole moment. Consequently, the special properties of barium titanate disappear above this temperature, which is called the *Curie point.* Above this temperature the substance behaves as an ordinary dielectric.

When the temperature is lowered below 120 °C, the ions of oxygen and titanium are displaced in opposite directions by an amount of the order of 0.1 angstrom. At this, the cell acquires a dipole moment.

Note the following especially important fact. This displacement can with equal success take place in any of three directions: along the three axes of the cube. Displacement leads to distortion of the cell. Hence, not all ways of dividing up the crystal regions within which all the dipole moments are arranged in a single direction turn out to be equally expedient.

Feasible ways of dividing a crystal into ideally polarized regions (called domains) are shown in Figure 2.5.

Along with the case in which the whole crystal is a single domain—when the maximum electric field is obtained—there may be less expedient versions and even one (at the extreme right) in which the external field equals zero.

How does a ferroelectric behave when an external field is applied to it? It was found that the polarization mechanism consists in the growth of a domain that faces the "required" direction by spreading its boundaries. Domains oriented with their dipole moment at an acute angle to the field "devour" domains oriented at an obtuse angle. When located in very strong fields, inversion of the domains may also be observed.

Barium titanate is the main industrial ferroelectric. It is obtained by the firing of two powders: titanium dioxide and barium carbonate. The result is a kind of ceramic material.

Ceramic ferroelectrics are extensively applied in electrical and radio engineering. Their main feature is that they drastically increase the permittivity of capacitors. Moreover, as is clear from the described polarization mechanism, the value of ε increases with the intensity of the electric field. A capacitor is thus converted into a varactor, or variable capacitor, by means of which frequency modulation can be most easily accomplished. This is a process that takes place in any radio or TV set.

For many purposes, ferroelectric ceramics have replaced quartz. They can be used to produce more powerful sounds. In the same way, the gain is higher for ultrasounds as well. The only field in which quartz has no competitors is the stabilization of radio frequencies.

The great majority of textbooks begin their chapters on electricity by describing electric charges produced on glass or hard rubber rods by rubbing them. The explanation of this phenomenon is usually evaded. Why?

First we should emphasize the fact that the electri-

fication of dielectrics by friction is not related (at least, directly) to the polarization of insulators that we have just discussed. As a matter of fact, polarization consists in the formation of bound electric charges whose special feature is that they cannot be "drawn off" the dielectric. The charges produced on the glass or hard rubber by rubbing them with cat fur are obviously free charges and, of course, that means electrons.

The general picture of what takes place is more or less clear, but not entirely. Apparently, the scanty amount of free electrons possessed by the insulator are bound to its molecules by forces of various magnitudes for various dielectrics. Hence, if two bodies are held in tight contact, electrons may pass over from one to the other. At this, electrification occurs. But tight contact here means to bring the surfaces within distances equal to interatomic ones. Since no atomically smooth surfaces exist in nature, rubbing helps to eliminate all kinds of projections and increases the area, so to say, of true contact.

Transfer of electrons from one body to another may occur for any pair of bodies, whether they are metals, semiconductors or insulators. But only insulators (or nonconductors) can be electrified because only in such bodies do the charges remain at the places to which they were rubbed off the other body.

I cannot express especially full satisfaction with this theory. It does not explain the advantages of using hard rubber, glass and cat fur for this purpose. We can pose heaps of more questions that have no convincing answers.

Conduction in Gases

If we fill a glass tube with a gas, seal electrodes into the tube and apply a voltage across them, we obtain a simple device for studying the conduction of electric-

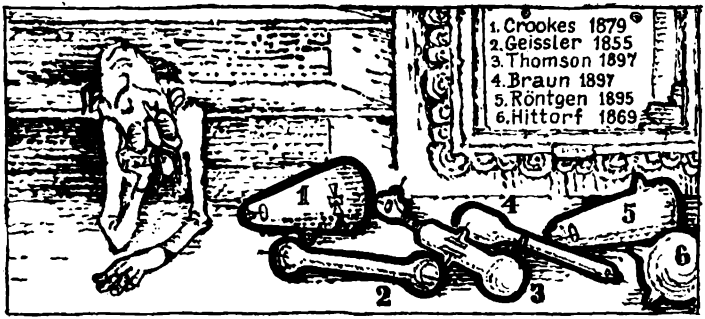


Figure 2.6

ity in gases. We can change the substance through which the current passes, we can vary the pressure of the gas and the applied voltage.

Investigations of the conduction in gases played an immense role in developing our concepts of the electrical structure of matter. The main part of this research was conducted in the nineteenth century.

Tubes of various shapes, used to study the phenomena we are discussing, are illustrated in Figure 2.6. Since all ancient statues and paintings of the old masters have long since been purchased and repurchased, dealers in antiques are now offering connoisseurs obsolete laboratory equipment. In curiosity shops in Western Europe and America, you can buy (usually at a good price) one of the rare items shown in Figure 2.6.

An electric current is initiated in a gas because the neutral molecules are broken down into anions and cations. Moreover, an electron may break away from molecules and atoms. The current is produced by a beam of positive ions and beams of negative ions and electrons travelling in the opposite direction.

To make a gas conduct a current it is necessary to convert the neutral molecules or atoms into charged particles. This may be accomplished either by an external ionizer, or by collision of the particles. As mentioned above, external means of ionization may be ultraviolet, X-, cosmic or radioactive rays. High temperature also leads to the ionization of a gas.

The passage of a current through a gas is often accompanied by some light effect. The character of the glow varies with the gas used, its pressure and the voltage applied. Investigations of this glow played a vital role in the history of physics, particularly as a source of information on the energy levels of the atoms and the laws of electromagnetic radiation.

Conduction in a gas does not obey Ohm's law. It is characterized by a curve showing the dependence of the current on the voltage. This curve is called the *current vs voltage characteristic* (not only for gases, but for any conducting systems that do not conform to Ohm's law).

Let us consider the phenomena, typical of any gas, that occur when voltage applied to a gas-discharge tube is increased. This behaviour of the gas, now to be described, is found in a wide pressure range. We shall except only such low pressures for which the free path of the molecules becomes commensurate with the dimensions of the gas-discharge tube. Our discussion will also not refer to such high pressures at which the density of gases approaches that of liquids.

We begin by applying a low voltage across the tube. If we do not use some ionizer, there is no current through the tube. If an ionizer is available, it produces charged particles—ions and electrons—in the gas. As soon as a field is set up in the tube, the particles are directed by the field to the electrodes of the tube. The velocity at which the charged particles travel toward the elec-

trodes depends upon many circumstances and primarily on the field intensity and gas pressure.

Chaotic motion is superimposed on the ordered motion of the ions and electrons resulting from the constant electric force. A particle accelerated by the field travels only a short distance. Its short run inevitably ends in a collision. At low velocities, these collisions conform to the law of elastic collisions.

The mean free path is determined first of all by the pressure of the gas in the tube. The higher the pressure, the shorter the mean free path and the lower the average velocity of ordered motion of the particles. Voltage applied across a gas-discharge tube has the opposite effect: higher voltage raises the average velocity of ordered motion of the particles.

If no voltage is applied across the tube, the following events occur in the gas. The ionizer produces ions and then ions of opposite signs join each other when they meet. This is called ion recombination. Since it takes a pair of particles to recombine, the rate of this process is proportional to the square of the number of particles.

When the ionizer operates continuously, equilibrium is set up between the two processes: ionization and recombination. This is what happens in the ionosphere surrounding our earth. Depending upon the time of day and the season, the number of ionized particles in a cubic centimetre varies from a million to hundreds of millions of ions and electrons. Thus, the degree of ionization is a quantity of the order of one per cent (recall the number of air molecules in unit volume at high altitudes).

But let us return to the ionized gas in the tube subject to a voltage. Equilibrium is upset, of course, between ionization and recombination because a part of the ions reach the electrodes before they are recombined. As the voltage is raised, a larger and larger part of the ions produced in unit time reach the electrodes, increasing

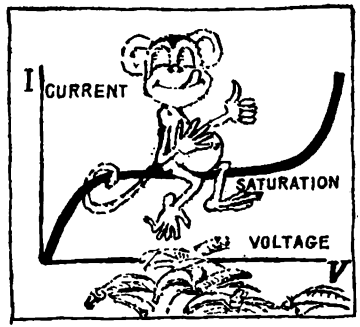


Figure 2.7

the current through the gas. This continues until there is no time left for recombination and all the ions produced by the ionizers reach the electrodes. It is obvious that no further increase in voltage can increase the current (this is called the *saturation current* and is represented by the horizontal portion of the current-voltage characteristic in Figure 2.7).

The less the density of the gas, the lower the field voltage at which the saturation current is reached.

The saturation current is equal to the charge of the ions formed by the ionizer in one second throughout the volume of the tube. The saturation current is usually not very large: of the order of a microampere or less. Its magnitude depends, of course, on the number of destructive missiles with which the ionizer bombards the gas.

When we operate the tube at a point on the current-voltage characteristic within the limits of the saturation current and protect the gas from the effects of an external ionizer, the current ceases to flow. This is said to be a non-self-maintained gas discharge.

If the voltage is raised again, new phenomena occur. At a certain instant, the velocity of the electrons becomes

high enough for them to knock electrons out of the neutral atoms and molecules. For this the voltage across the tube should reach a value at which an electron acquires energy sufficient in its free path to ionize a molecule. The initiation of collision ionization has a marked effect on the current-voltage characteristic: the current begins to increase again because an increase in voltage leads to an increase in the velocity of the electrons. An increase in velocity, in its turn, raises the ionizing capacity of the electrons so that a greater number of ion pairs is produced and the current increases. The current-voltage characteristic curve turns sharply upward. In comparison with the saturation value, the current increases by hundreds and thousands of times and the gas begins to glow.

Now, if we eliminate the action of the external ionizer, the current continues to flow. We have gone over to the region of a self-maintained gas discharge. The voltage at which this qualitative change occurs is called the *breakdown voltage* or the *ignition voltage of a gas discharge.*

The sharp increase in current in passing over this critical limit is due to the avalanche-type increase in the number of charges. One released electron destroys a neutral molecule and produces two charges of such high energy that they are capable of breaking down another pair of molecules that they encounter. Thus, two charges form four, four form eight, etc. You must agree that the use of the word "avalanche" is fully justified.

A quantitative theory has been proposed that predicts the shape of the current-voltage characteristic curves for gases with a fair degree of accuracy.

Self-Maintained Discharge

There are many types of discharges in gases. We shall discuss only a few.

Spark discharge. A spark passing through the air between two electrodes can be readily observed in the most elementary experiments. Sparks can be produced simply by bringing two wires, across which a voltage has been applied, sufficiently close together. What do we mean by "sufficiently"? If air is the medium through which the spark is to pass, the required field intensity is 30 thousand volts per centimetre. This means that a potential difference of 3000 volts is sufficient for a small gap of one millimetre. Any reader has observed such small sparks in everyday life when repairing faulty electric wiring or when bringing two leads from the terminals of a storage battery close together (here the gap between the ends of the leads must equal the thickness of a safety razor blade).

The breakdown voltage depends upon the density of the gas. It also depends upon the shape of the electrodes.

A spark can pass through dielectric liquids and solids as well as through a gas. It is important for an electrician to know the breakdown voltage of all the materials he uses in his work.

Today it seems quite obvious to us that lightning is a spark jumping between clouds charged with electricity of opposite sign. This was not always the case, however, and Mikhail Vasilievich Lomonosov (1711-1765) and Benjamin Franklin (1706-1790) made great efforts to prove this statement. Georg Wilhelm Richmann (1711-1753), a Russian physicist that collaborated with Lomonosov, lost his life in an attempt to ground lightning through a twine that conducted electricity from a kite that he flew during a thunderstorm.

Interesting data are available on the spark discharge

in a bolt of lightning. The voltage between the cloud and the earth is about 10^8 to 10^9 volts and the current ranges from tens to hundreds of thousands of amperes. The diameter of the spark channel ranges from 10 to 20 cm.

The duration of a flash of lightning is extremely short: of the order of a microsecond. We can readily estimate that the quantity of electricity passing through the lightning channel is comparatively small.

These sparks from the heavens have been studied in detail by means of high-speed motion picture photography. A bolt of lightning quite frequently consists of a series of spark discharges following a single path. Lightning has a sort of "leader" which pierces the most convenient, always freakishly branched way for the electric discharges.

Also observed to some extent is ball lightning. This type is not fully understood and cannot, unfortunately, be reproduced under laboratory conditions. It is an incandescent sphere of gaseous plasma, 10 to 20 cm in diameter. Ball lightning moves slowly and is sometimes even stationary. It exists several seconds, or even minutes, and then disappears with a powerful explosion. It must be admitted that so far no comprehensive theory of this interesting phenomenon has been proposed.

Arc discharge. An arc discharge was first obtained by the famed Russian physicist and electrical engineer Vasily Vladimirovich Petrov (1761-1834) as far back as 1802. He struck the arc by bringing into contact two pieces of carbon across which a powerful voltage had been applied from a source of electric energy. Then he moved the carbons (electrodes) apart. This procedure is still being employed today. True, special carbon rods of pressed graphite powder are used as the electrodes. The positive rod burns away more rapidly than the negative one. By just looking at the electrodes we can immediately find to which one the positive pole is connected: a

hollow, called the crater, is formed at its end. The temperature in the crater at standard pressure may reach 4000 °C. If the pressure is increased, the temperature can be raised almost 6000 °C, i.e. the temperature at the surface of the sun. An arc struck between metal electrodes has a flame with a considerably lower temperature.

A low voltage, of the order of 40 to 50 volts, is sufficient to maintain the arc. The current may reach hundreds of amperes because the resistance of the incandescent gaseous column is low.

How do we explain the high electrical conductivity of the gas at such low voltages? The molecules are not accelerated to high velocities, and their collisions cannot play much of a role in producing a heavy current. The explanation is as follows: at the first instant the arc is struck, a large quantity of heat is evolved at the point of contact, drastically raising the temperature at the ends of the electrodes. This initiates the process of thermal electron (thermionic) emission in which the cathode ejects an immense number of electrons. From this it follows, by the way, that only the high temperature of the cathode is of importance; the anode can remain cold.

The mechanism of an arc discharge of this type is entirely different from that of a spark discharge.

No need, evidently, to remind the reader how important this phenomenon is in engineering practice. Arc discharges are widely employed in welding and cutting metals, as well as in electrometallurgy.

Glow discharge. This kind of self-maintained discharge is also of great practical value, since it takes place in gas-discharge tubes or, as they are also called, daylight lamps. The tube is designed and is filled with gas (at a pressure substantially below atmospheric pressure) so that it operates at a voltage exceeding the ignition voltage. The electric current in gas-discharge lamps is produced by the ionization of the gas molecules by

electrons, and also because electrons are knocked out of the cathode. A gas-discharge lamp does not ignite instantaneously. This is evidently due to the fact that the first impulse must be obtained from the small amount of charged particles that are always present in any gas.

Corona discharge. A corona is observed at atmospheric pressure in a highly nonuniform field, for example, in the vicinity of a wire or sharp point. The field intensity should be high: of the order of a million volts per metre. It makes no difference to which pole the sharp point is connected. Thus, there are both positive and negative coronas. Since the field intensity is reduced as we depart from the sharp point, the corona disappears at a short distance away. We can contend that a corona discharge is an incomplete breakdown of the gas gap. A corona is produced by electron avalanches travelling either toward the point or from the point to external space. Besides electrons, the corona region contains negative and positive ions—products of disintegration of the neutral air molecules. The corona glows only in the small space near the point in which there is an electron avalanche.

The initiation of a corona depends on the atmospheric conditions and, primarily, the moisture.

The atmospheric electric field may cause the tops of trees and tips of ship's masts to glow. Formerly, this phenomenon was called St. Elmo's fire. It was supposed to be an evil omen. This has a reasonable explanation. It may well be that such a glow is observed just before a storm breaks.

We can learn a moral lesson from an event that took place quite recently. Two amateur investigators, Mr. and Mrs. Kirlian, had spent some years studying the following phenomenon. A person lays a hand, connected to one terminal of a high-voltage power supply, on a photographic film that is separated by a layer of insulation from the other electrode of the same power cir-

cuit. When the voltage is switched on, a blurred image of the palm and fingers is obtained on the film. This image is due to a corona discharge. Naturally, the voltage must be less than that required for a spark breakdown of the insulation.

These experiments drew the attention of experts in the field of so-called parapsychology, the general name for a group of "theories" that the great majority of physicists and psychologists consider to be pseudo-scientific. This attention was due to the fact that the Kirlians and their followers related the kind of photograph obtained to the "psychic" state of the subject.

As a result of the extensive publicity received by such an extravagant interpretation of these experiments, a group of physicists and psychologists from various universities in the USA decided to check the experiments more carefully. They searched for a simpler explanation of the undoubted fact that the appearance of the photograph thus obtained really does differ for different persons or even for the same person if the photograph is taken under different conditions.

The investigators reached the following conclusion: "Photographic images obtained by the Kirlian technique are principally a record of corona activity during an exposure interval. Most of the variations in the images of the corona of a living subject who is in contact with the photographic film can be accounted for by the presence of moisture on or within the subject's surface. During exposure, moisture is transferred from the subject to the emulsion surface of the photographic film and causes an alteration of the electric charge pattern on the film, hence the electric field at the surface of the subject."

The investigators proposed further application of this technique, which they preferred to call "corona-discharge photograph", "for the detection and quantification of

moisture in animate and inanimate specimens through the orderly modulation of the image due to the various levels of moisture".

This interesting piece of information, published in the December 1976 issue of the journal *Scientific American* with the title "Sweaty Palms", leads to two conclusions. In the first place, any real phenomenon deserves attention and it is quite possible that it may prove useful in practice. Secondly, when an investigator discovers a new fact, he should begin by overcoming the temptation to interpret it in a way that does not fit in with up-to-date scientific concepts. The discovery can be made public and presented for judgement by specialists only after it has been comprehensively shown that existing theories are incapable of explaining the new fact.

Real facts, which are falsely interpreted and explained, can be called (after an old joke) cockroach experiments. According to the joke, the legs are pulled off one cockroach which is then placed on a table side by side with a cockroach having all his legs. The "investigator" then knocks on the table. The uninjured cockroach runs away while the "cripple" remains. This proves, contends the "investigator", that a cockroach hears with its legs.

Every year publications appear that describe such "cockroach experiments". I consider it good policy to warn my readers against them.

Matter in the Plasma State

The term Plasmenzustand was proposed as far back as 1939 by two German scientists whose article was translated into the Russian by the author for the Soviet journal *Advances in Physical Sciences*. This seems to be a suitable name. As a matter of fact, plasma is neither a

solid, nor a liquid, nor a gas. It is a special state of matter.

Thermal ionization of gases, i.e. stripping the atoms of their electrons and the disintegration of neutral molecules into ions, begins at temperatures exceeding 5 or 6 thousand degrees. Is it worthwhile, then, to discuss this problem? No materials exist that can withstand a higher temperature.

It certainly is worthwhile! The great majority of celestial bodies, like our sun, are in a plasma state. An example of plasma is the ionosphere. Plasma can be confined in a limited volume by properly shaped magnetic fields, so-called magnetic bottles, under laboratory conditions. We can, in addition, speak of the plasma of a gas discharge.

The degree of ionization of a gas depends on the pressure as well as the temperature. Hydrogen at a pressure of the order of 1 mm of mercury column is practically completely ionized at a temperature of 30 thousand degrees. Under such conditions, there is only one neutral atom per 20 thousand charged particles.

Hydrogen in the plasma state consists of a mixture of chaotically travelling and colliding particles of two gases: proton "gas" and electron "gas". Plasma formed of other substances is a mixture of many "gases". In such a plasma we find electrons, bare nuclei, various ions, as well as a negligible quantity of neutral particles.

A plasma with a temperature of tens or hundreds of thousands of degrees is said to be cold. A hot plasma has a temperature of millions of degrees.

But one must be careful with the concept of a plasma temperature. As the reader knows, temperature is uniquely determined by the kinetic energy of the particles. In a gas consisting of heavy and light particles, a state of equilibrium is set up only after the heavy and light particles have acquired the same average kinetic energy.

This means that in a gas existing for some time under stable conditions, the heavy particles travel slowly and the light particles, rapidly. The time required to set up equilibrium depends on what we had "at the beginning". Other conditions being equal, the greater the difference between the masses of the particles, the longer the time required to attain equilibrium.

These are exactly the conditions we find in a plasma. The mass of the lightest nucleus is almost two thousand times greater than that of an electron. In each collision, an electron transmits only a small part of its energy to a nucleus or ion. Thus the average kinetic energies of all the particles of plasma are equalized only after an immense number of collisions. Such a plasma is said to be *isothermal*. It is the kind existing in the interior of the sun and other stars. The time required to reach equilibrium in a hot plasma ranges from fractions of a second to some seconds.

The plasma in a gas discharge (spark, arc, etc.) is a different matter. Here the particles not only travel chaotically, but also produce an electric current. In their path between the electrodes, the rapid electrons do not have the opportunity to transmit any appreciable part of their energy to the leisurely paced ions. This is why the average velocity of the electrons in a gas discharge is much higher than that of the ions. Such a plasma is said to be *nonisothermal* and must be specified by two temperatures (or even three if we take the neutral particles into account). Naturally, the electron temperature is substantially higher than the ion temperature. Thus, in an arc discharge, the electron temperature is from 10 to 100 thousand degrees, while the ion temperature is close to 1000 degrees.

The behaviour of the particles in a plasma can be described by means of the same quantities that are used in the kinetic theory of gases. Many techniques have

been devised for directly or indirectly determining the mean free path of the particles, the mean free time, and the concentrations of particles of various kinds.

To povide the reader with an idea of the orders of magnitude found in plasma physics, we present certain data describing a high-concentration hydrogen plasma (10^{20} ions per cubic metre). It was found that in a cold plasma (at a temperature of 10 000 °C), the mean free path equals 0.03 cm and the mean free time is 4×10^{-10} s. If the same plasma is heated to a hundred million degrees, the respective figures are 3×10^{6} cm and 4×10^{-4} s.

In citing such data, it is absolutely necessary to specify additionally the kind of collisions we have in mind. The data given are for encounters between electrons and ions.

It is evident that a volume containing many particles is electrically neutral. But we may be interested in the behaviour of the electric field at some definite point in space. The field varies rapidly and drastically because ions and electrons alternate in rushing past the point. The rapidity of these variations can be calculated, as can the average intensity of the field. Plasma complies with the condition of neutrality with exceptionally high precision. Strictly speaking, we should employ the term "quasi-neutrality", i.e. near-neutrality. But what do we mean by this "near"?

Rather simple calculations indicate the following. Consider a length of one centimetre in the plasma. Calculate the electron and ion concentrations for each point in this length. Quasi-neutrality means that these concentrations should be "nearly" equal. Next, let us imagine that in one cubic centimetre we have an "extra" amount of electrons that are not neutralized by positive ions. If this is so and the particle density is equal to the air density at the earth's surface, a field with an intensity of about 1000 V/cm is set up along the length being con-

sidered, even if the difference in ion and electron concentrations equals only one thousand millionth of one per cent! Here is what the word "near" signifies.

But even this negligible lack of equality of positive and negative charges lasts only for an extremely short instant. The field that forms excludes all the superfluous particles. This automatism is already operable for regions measured by thousandths of a centimetre.

We shall return to plasma in magnetic bottles again in Book 4. The reader has undoubtedly seen accounts, and perhaps descriptions, of installations of the Tokomak type (in the USSR). A whole army of scientists are working on their improvements. The point is that if a high-temperature plasma could be produced and properly handled, it would lead to controlled fusion of light atomic nuclei, which would be accompanied by the generation of titanic amounts of energy. Physicists have learned to realize this process (uncontrolled fusion) in bombs. Will it be possible to produce a plasma with a sufficiently high temperature and sufficiently long duration to initiate a chain reaction of the kind accomplished in a nuclear reactor? There is no answer yet to this question.

Metals

The subdivision of solids into various classes according to their electrical resistance is based on the mobility of their electrons.

An electric current is a stream of moving charged particles. When we deal with streams of ions or electrons, we literally visualize an electric current. An electric current also reveals itself distinctly in passing through liquids because matter is deposited on the electrodes. But as for solids, there is only circumstantial evidence of the passage of an electric current.

A series of facts are available that enable us to make the following statements. No displacement of the atomic nuclei occurs in solids. An electric current is produced by electrons. The electrons travel due to the energy supplied by the current source. This source sets up an electric field inside the solid.

The equation relating the voltage and the electric field intensity is valid for any conductors. Therefore, combining the equations on pp. 14 and 22, we can write Ohm's law for a solid conductor in the form

$$j = \sigma E$$

where $\sigma = 1/\rho$ is called the *electrical conductivity*.

The electrons of a solid can be divided into bound and free electrons. The bound electrons belong to definite atoms; the free electrons form a kind of electron gas. These electrons move around in the solid. When no voltage is applied to the solid, the electrons have random motion. The more the motion of the free electrons is impeded, the oftener they collide with fixed atoms and with one another, the higher the electrical resistance of the body.

The vast majority of electrons in a dielectric have an owner that is either an atom or a molecule. The number of free electrons is negligible.

In metals each atom donates one or two electrons for common use. This electron gas is the current carrier.

On the basis of a roughly approximate model we can estimate the electrical conductance and thereby check the model.

In exactly the same way as when we discussed a molecular gas, we shall assume that each electron manages to travel a path of length l without collisions. The distance between the atoms of a metal equals several angstroms. It is logical, then, to assume that in order of magnitude the mean free path of the electrons equals 10 Å, i.e. 10^{-7} cm.

In its motion the electron is subject to the accelerating force eE during the time l/v, where v is its velocity. The chaotic velocity of electrons can be estimated on the basis of data obtained in investigations of thermionic (thermal electron) emission. This velocity is of the order of 10^8 cm/s.

To determine the velocity of ordered motion of electrons, i.e. the velocity of the motion that produces a current, the acceleration eE/m is to be multiplied by the mean free time. This assumes that each collision discontinues motion of the electron, after which it begins to pick up speed again. Multiplying, we obtain the velocity of the electrons that produce the electric current:

$$u = \frac{eEl}{mv}$$

Next, we attack the problem of determining the resistivity of a metal. If we obtain the correct order of magnitude, then we can presume that our model "works".

We shall leave to our readers the task of showing that the current density j can be written as the product of the number of electrons in unit volume by the charge of the electron and by the velocity of ordered motion (drift) of the electrons. Thus $j=neu$. Substituting into this equation the drift velocity, we obtain $j = \frac{ne^2l}{mv}E$. Then the electrical conductivity is

$$\sigma = \frac{ne^2l}{mv}$$

If we assume that each atom contributes one electron for common use, then we find that a conductor has a resistivity of the order of 10^{-5} ohm-m. A very reasonable value! It confirms both the validity of our roughly approximate model and the proper choice of the values of the parameters in our "theory". I place the word

theory in quotation marks only because it is a crude approximation and elementary. This example, however, illustrates the typical physical approach in interpreting phenomena.

According to the theory of a free electron gas, the electrical resistance should decrease with a drop in temperature. But do not hasten to relate this circumstance with the change in the velocity of chaotic motion of the electrons. This velocity has nothing to do with the matter because it depends only slightly on the temperature. The reduction in resistance is due to the reduced amplitude of vibration of the atoms. As a result, the mean free path of the electrons increases.

This fact can also be expressed in other words: upon an increase in the amplitude of vibration of the atoms, the electrons are scattered to a greater degree in various directions. Consequently, the component velocity in the direction of the current is reduced, i.e. the resistance should increase.

The increase in electron scattering also explains the increased resistance of a metal (and not only a metal) when impurities are added to it. The impurity atoms act like defects in crystal structure and therefore facilitate electron scattering.

Electric energy, as we know, is transmitted by wires. Owing to their resistance, the wires draw energy from the current source. Such energy losses are enormous and their prevention is one of the most vital of engineering problems.

There is hope that this problem can be solved on the basis of the remarkable phenomenon known as *superconductivity*.

In 1911, the Dutch physicist Heike Kamerlingh Onnes (1853-1926) found that at temperatures close to absolute zero certain bodies suddenly lose practically all of their electrical resistance. If a current is induced in a

ring-shaped superconductor, it continues to flow for days without diminishing. Of pure metals, the highest temperature at which superconductive properties are found is about 9 K, the metals being niobium (8.9 K) and technetium (9.3 K). No need to mention what a vast army of scientists are engaged in the search of superconductors that would acquire this wonderful property at a higher temperature. So far they have not been any too successful, though an alloy has been discovered that is claimed to become superconductive at about 20 K.

There is reason to believe, however, that this limit can be raised (perhaps even to room temperature). The search is being made among special polymer substances and among complex lamellar materials in which a dielectric alternates with a metal. The significance of this problem can hardly be overvalued. I take the liberty of regarding it to be one of the cardinal problems in modern physics.

The search for superconductors that acquire this property at sufficiently high temperatures especially gained in scope after a theory had been proposed for this phenomenon. This theory suggested new courses along which the required materials might be found.

It is characteristic that much time passed between the discovery of the phenomenon and its explanation. The theory was advanced in 1957. It might be well to point out that the laws of quantum physics, on which the theory of superconductivity is based, were established as far back as 1926. It follows that the explanation of the phenomenon is far from simple. In this book I can only start, as you might say, from the middle of the story. It seems that with the slowing down of vibrations in the atomic lattice, certain electrons become "paired" Such a "pair" behaves in coordination with each other. When the pairs are scattered by the atoms (and this scattering, as mentioned above, is the cause of resis-

tance), the rebound of one of the members of the pair to one side is compensated for by the behaviour of its "friend". This compensation is in the sense that the total momentum of the electron pair remains constant. Thus, electron scattering does not disappear but no longer influences the passage of the current.

In addition to paired electrons, a superconductor also carries ordinary electron gas. Hence, two fluids seem to exist simultaneously: one is ordinary and the other is superconductive. As the temperature of the superconductor is raised from absolute zero, thermal motion breaks apart more and more electron pairs, and the percentage of the ordinary electron gas increases. Finally, at the critical temperature, all the paired electrons disappear.

In Book 2 we made use of the two-fluid model, with one ordinary fluid and one special fluid, to explain superfluidity observed in liquid helium. These two phenomena are closely related: superconductivity is superfluidity of the electron fluid.

Each electron pair, of the kind mentioned above, has a total spin equal to zero. Particles with a spin equal to zero or a whole number (i.e. with integral spin) are called *bosons*. Under certain known conditions, large amounts of bosons can occupy the same energy level. In such cases their motions become ideally coordinated and nothing can impede their displacements. We shall return to this phenomenon again in Book 4.

Electron Emission from Metals

Since a part of the electrons behaves like a gas of rapid particles, it is natural to expect that electrons are capable of emerging to the surface of a metal. For the electron to leave the metal entirely, it must overcome the attrac-

tive forces of the positive ions. The work done by the electron for this purpose is called the *work function.*

The higher the temperature of the metal, the greater the kinetic velocity of the electrons. If the metal is heated to incandescence, an appreciable number of electrons will be able to escape from it.

Thermionic emission, as the ejection of electrons from a metal is called, can be investigated by a simple experiment. An additional electrode is sealed into an electric light bulb. A sensitive instrument can be used to measure the current set up by the part of "evaporating" electrons that reach the new electrode (a part, and not all, because the electrons fly out of the incandescent filament in all directions).

To evaluate the work function we resort to a "barrier" voltage, i.e. we connect the sealed-in electrode to the negative pole of a battery. Gradually raising the voltage, we reach a value at which the emitted electrons can no longer arrive at the electrode.

The electron work function for tungsten is about 5 electron volts. Special coatings can, if required, reduce the work function to one electron volt.

What is this unit of work called the *electron volt*? It is, as the name implies, the energy acquired by an electron in travelling over a portion of its path with a voltage of one volt applied across this portion. One electron volt equals 1.6×10^{-19} joules. The thermal velocity of electrons is quite considerable, but their mass is very small. Therefore, the given barrier height is extremely large. Theory and experiments show that electron emission depends drastically on the temperature. A temperature rise from 500 to 2000 K increases the emission current a thousandfold.

Owing to thermal motion the emission of electrons from a metal is, so to speak, a natural process. But electrons can also be knocked out of a metal.

In the first place, this can be done by bombarding the metal with other electrons. This is called *secondary (electron) emission.* It is made use of to multiply the number of electrons in certain engineering instruments.

Vastly more essential is the extraction of electrons from solids by light. This phenomenon is called the *photoelectric effect.*

Thermoelectric Phenomena

Very long ago (in the evolution of mankind this time is a mere instant, but in the development of science it is almost as long as eternity), over 150 years in the past, a simple fact was discovered. If an electric circuit is made up of a piece of copper wire and a piece of bismuth wire by soldering them together at two junctions, current flows through this circuit. This happens only when the temperature of one junction is higher than that of the other. This is the *thermoelectric effect.*

What makes the electron travel along our combined circuit? The explanation of this phenomenon is not at all simple. The electromotive force is due to two factors. In the first place, we have a contact electric field, secondly, we have a temperature electric field.

We have just mentioned that work is required to remove an electron beyond the limits of a metal. It is natural to assume that work function A differs for various metals. Hence, there is a voltage across the junction of the two metals equal to

$$\frac{1}{e}(A_1 - A_2)$$

The contact voltage can be detected experimentally. By itself, however, this voltage cannot set up a current in a closed circuit. Such a circuit obviously consists of two junctions and the contact voltages oppose and cancel

each other. But why does the difference in the temperatures of the junctions produce an electromotive force? The answer is the only logical one. Evidently the contact voltage depends upon the temperature. Heating one of the junctions makes the voltages unequal and sets up a current. But here we must take another phenomenon into account. It can naturally be assumed that there is an electric field between the ends of a conductor if these ends have different temperatures because the electrons travel faster at higher temperatures. This being the case, diffusion of the electric charges begins and continues until a field is set up that counteracts the tendency to uniform distribution.

Experiments leave no doubt of the fact that both phenomena are simultaneously present and both are to be taken into account in proposing a theory.

Thermoelectromotive forces are small: of the order of a millivolt for a temperature difference of 100 degrees. Such voltages are readily measured. Consequently, the thermoelectromotive effect is employed for measuring temperatures. You cannot, of course, insert a liquid-in-glass thermometer into molten metal. For such purposes a *thermocouple* (as the thermoelement used for measuring temperatures is called) is an excellent instrument. A thermocouple has, in addition, many other advantageous features. How vitally important it may be to measure temperatures at great distances. And the exceptional sensitivity. Electrical measurements are always precise, and it was found that differences in temperature as small as a millionth of a degree can be sensed by a thermocouple.

This high sensitivity enables thermoelements to be applied for measuring heat flow from extremely remote objects. The reader can himself estimate the possibilities of a thermoelement. Suffice it to say that a tenth of an erg per second is no limit.

Like storage cells, thermoelements are frequently assembled into banks to form a thermal battery. If the power requirements are not very high, such a battery can serve as a generator of energy and can find application in radio communications.

Semiconductors

A great many substances, both elements and chemical compounds, have conductivities filling the wide range between conductors and insulators. Such substances were first discovered a long time ago. But a mere twenty years ago it was hardly probable that anyone could foresee that semiconductor physics would give rise to a powerful branch of industry whose importance in world economics cannot be overrated. Without semiconductors, up-to-date electronic computers, TV sets and tape recorders would be unfeasible. Radio engineering is inconceivable today without semiconductors.

Insulators (nonconductors) have a conductivity ranging between 10^{-8} and 10^{-18} $ohm^{-1}\text{-}m^{-1}$, the range of conductivity of metals in the same units is from 10^{2} to 10^{4}. The conductivity of semiconductors lies between these two ranges. As we shall see, not only their resistance is of interest in dealing with semiconductors.

Like in metals, no chemical changes occur in semiconductors when we pass an electric current through them. This indicates that the ions of these substances, forming the frame of their crystal lattice, are not moved around by the action of the electric field. Therefore, as in metals, the motions of the electrons are responsible for electric conduction.

Though this seems obvious, physicists, in the early stages of semiconductor research, decided, in any case, to find which charges were the current carriers. For sol-

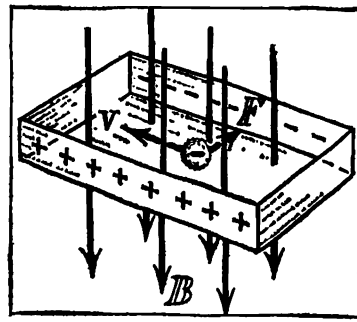


Figure 2.8

ids, this can be done by means of the Hall effect, discovered in 1879 by Edwin Herbert Hall (1855-1938).

In the next chapter I shall remind you that a magnetic field deflects positive and negative particles in different directions. If a solid in which charges are travelling is made in the form of a plate and is placed in a magnetic field of the proper direction, a voltage appears across the edges of the plate. This arrangement is shown schematically in Figure 2.8.

The physicists were surprised to find that certain bodies, when investigated in this manner, behaved sometimes as if positive particles were travelling along the conductor, and at other times, as if the carriers of electric charge were negative. We can readily find names for this behaviour. We call the first case positive (*p*-type) conduction and the second, negative (*n*-type) conduction. The point, of course, is not in the name, but in an explanation of the matter. There is no doubt that electrons are travelling inside the semiconductor. How can we reconcile this contradiction? How can we explain the positive conduction?

Imagine a formation of athletes. For some reason, one man drops out of line. This leaves an empty space. Though it doesn't sound any too aesthetic, we can say that a

"hole" is formed. To dress the ranks, the commanding officer tells the neighbour of the "hole" to move over one place. It is absolutely clear that this forms a new empty space. It can be filled by commanding the next man to move into the "hole". If the athletes shift, one by one, to the left, the "hole" moves to the right. This is the mechanism that explains positive conduction of semiconductors.

The concentration of free electrons is very low in semiconductors. Hence, the definition itself of conduction (recall the formula we derived a few pages back for the current density) should suggest that the majority of atoms in a semiconductor are not in the form of ions, but are neutral atoms. Still, a semiconductor is not an insulator. Consequently, a small number of electrons are freed. These electrons travel as in a metal and are responsible for the negative, i.e. electron, conduction. But a positive ion, surrounded by neutral atoms, is in an unstable state. As soon as an electric field is applied to the solid, a positive ion tries to "lure" an electron from its neighbour; the next atom proceeds in exactly the same way. A positive ion is quite similar to a "hole". This interception of electrons may overcome the motion of free electrons. This is how positive, or hole, conduction occurs.

Maybe you do not care for this model? I can suggest another. As we have mentioned, the energy of a particle is quantized. This is a fundamental law of nature. All phenomena occurring in semiconductors can be explained if we assume that the electrons are distributed among energy levels in a solid as they are in the atom. But since there are so many electrons in a solid, the levels merge into *energy bands*.

If there is only weak interaction between the electrons, the bands are quite narrow. Therefore, the fact that atoms are a part of a solid has practically no effect on the inner electrons.

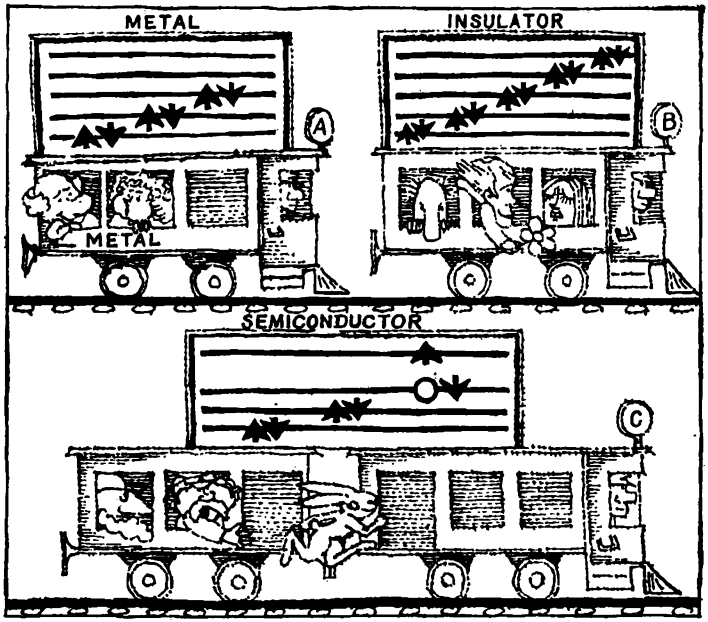


Figure 2.9

Outer electrons are a different matter. Their levels form the bands. The width of these bands and the "distance" between them differ in different bodies (actually we should call them *energy gaps*; the word "distance" in this context is simply physical shoptalk or jargon).

This picture clearly explains the division of solids according to their electrical conductivity into metals, semiconductors and insulators (Figure 2.9). When a band is completely filled with electrons and the width of the gap to the next higher empty band is large, the body is an insulator, or nonconductor. If the upper band is only

partly filled with electrons, we have a metal because even the weakest electric field can kick an electron to a slightly higher energy level. Typical of a semiconductor is the fact that its upper band is separated from the next lower one by a narrow gap. In contrast to nonconductors and metals, thermal motion is capable in semiconductors of transferring electrons from one band to another. In the absence of a field, the number of such transitions upward and downward are the same. A rise in temperature only increases the electron concentration in the upper band.

But what happens when a field is applied to semiconductor?

Now a free electron in the upper band begins to move and contributes to the negative conduction. But the equilibrium of upward and downward electron transitions is violated. Therefore a hole is formed in the lower band and begins to move, due to the field, in the opposite direction. Such semiconductors are called *conductors with mixed (positive-negative) conduction.*

The band theory of semiconductors is an orderly and consistent one. The reader should not consider the described model to be artificial or far-fetched. It simply and clearly explains the principal difference between a metal and a semiconductor, namely, their particular behaviour with a change in temperature. As mentioned in a preceding section, the electrical conductivity of metals drops with a rise in temperature because the electrons more frequently collide with obstacles. An increase in the temperature of a semiconductor leads to an increase in the number of electrons and holes, and a consequent increase in conductivity. Calculations indicate that this effect substantially exceeds the loss in conductivity due to collisions with obstacles.

Of prime importance in engineering are semiconductors with impurities. By adding impurities, called *doping*, we

can produce a body that has only negative or only positive conduction. The idea is extremely simple.

The most extensively employed semiconductor materials are germanium and silicon. These elements are tetravalent. Each atom is bound to four neighbours. Ideally pure germanium is a semiconductor of the mixed type. The number of holes and free electrons per 1 cm^3 is very small, namely, about 2.5×10^{13}. This comes to about one free electron and one hole per thousand million atoms.

Now let us replace an atom of germanium with an atom of arsenic. Arsenic is pentavalent (having five valence electrons). Four of its electrons bind it to atoms of the host—germanium—and the fifth is free. This material possesses electron (negative) conduction because the addition of an arsenic atom does not lead to the formation of holes.

Even if only a trace of arsenic is added, one atom per million atoms of germanium, the conductivity of the germanium increases by a thousandfold.

It should be clear now what is needed to convert the germanium into a *p*-type conductor. This can be done by replacing a germanium atom with a trivalent atom, for example, indium.

This leads to the following situation. An atom of germanium adjacent to the guest is converted into a positive ion because it must form a bond with the indium, which has one electron too few. But we already know that a positive ion plays the role of a hole. Owing to the field, the holes move and there is motion of free electrons.

No wonder that the semiconductor industry has greatly influenced the techniques of growing pure crystals. It could not be otherwise when even a trace of impurities, as small as a millionth part, makes all the difference.

It would be incorrect to suppose that there is no hole conduction in *n*-type semiconductors. There are holes,

of course, but their number is substantially less than the number of free electrons. In *n*-type semiconductors, the electrons are the *majority carriers* of current, while the holes, constituting a minority, are called *minority carriers.* On the contrary, the majority carriers in *p*-type semiconductors are holes, and the minority carriers are electrons.

p-n Junction

Now that we know what *p*- and *n*-type semiconductors are, we can look into an interesting effect that is vitally important in up-to-date electronics. This effect occurs in the transition region between *p*- and *n*-type semiconductors tightly joined together. This region is commonly called a *p-n* junction, though the word *transition* has served as the basis for naming a whole class of devices (called transition-region devices) operating on the *p-n* junction principle. What happens when we take two blocks of the same cross section, one made of Ge doped with In (*p*-type semiconductor) and the other of Ge doped with As, grind one face on each block to an exceptionally high degree of smoothness and flatness, and join the ground faces very tightly together? We shall have, actually, a single crystal of germanium, but one half has an excess of free electrons and the other, an excess of holes.

To keep the explanation as simple as possible, we shall, for the moment, forget about the minority current carriers. At the initial instant of time (see upper drawing in Figure 2.10), both halves of the crystal are electrically neutral. But the *n*-type half has (notwithstanding its electrical neutrality) "surplus" electrons (black dots) and the *p*-type half has "surplus" holes (circles).

Both the electrons and holes can freely pass through the boundary. The reason for such transitions is the same

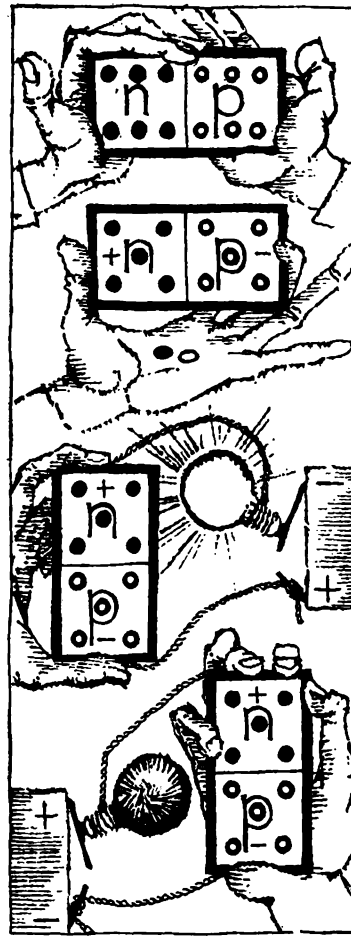


Figure 2.10

as in mixing two gases when their containers are connected together. But, in contrast to gas molecules, electrons and holes are capable of recombining.

At first we had six black dots at the left and six circles at the right. As soon as the transition begins, a circle and a dot annihilate each other. The next drawing shows that there are less electrons left in the left half than are required for it to remain electrically neutral. The right half has one circle less. Depleting the left half of one electron, we give this half a positive charge. For the same reason, the right half acquires a negative charge.

Transition through the boundary of the next holes and electrons becomes difficult. They have to move against the electric fields set up by the charges. Transition continues for some time, as long as the thermal motion is able to surmount the constantly increasing barrier. Finally, dynamic equilibrium is reached.

Now, what happens if we apply a voltage across our composite *p-n* crystal as shown in the third drawing in Figure 2.10? Evidently, we submit additional energy to the current carriers of an amount sufficient to enable them to clear the barrier.

On the contrary, if we connect the positive pole to the *n*-type half, it remains impossible for electrons and holes to surmount the barrier.

In short, a *p-n* junction allows current to flow in only one direction, thereby possessing rectifying properties.

Today, rectifiers (for which valves and diodes are synonyms) find most extensive application in many different fields of engineering. Their principle has just been described.

Our explanation, however, was extremely crude. In no detail did it take into account the behaviour of the holes and electrons capable of surmounting the barrier without recombination. The main drawback was our disre-

gard for the minority carriers that do not allow a *p-n* crystal to completely rectify a current. Actually, a weak current does flow when we apply the voltage as shown in the lower drawing of Figure 2.10.

Let us consider in more detail the events that occur at the boundary, or junction, when dynamic equilibrium is reached.

We first discard the simple assumption we made above and recall the existence of minority carriers.

In establishing dynamic equilibrium, the situation is as follows. Approaching the boundary from the depth of the *p*-type crystal, the hole current gradually increases. Contributing to this current are holes that manage to reach the *p-n* junction and jump across it without recombining with electrons.

Of course, these holes must also possess sufficient energy to overcome the potential barrier.

In passing through the transition region, this current gradually fades due to recombination with electrons. At the same time, a hole current gradually increases from the depth of the *n*-type crystal, flowing in the opposite direction. There are much fewer holes in this part, but they do not have to surmount the barrier to get into the *p*-type part. We can say that the barrier is arranged so that the forward and reverse currents compensate each other.

The above also concerns electron current. True, the hole and electron currents may differ greatly in magnitude because the *p*- and *n*-type parts are differently doped with impurities and possess, consequently, different amounts of free carriers. If, for instance, there are a great many more holes in the *p*-type part than electrons in the *n*-type part, the hole current is much higher than the electron current. Such a *p*-type part is called the *emitter* of free current carriers, and the *n*-type part, the *base*.

After this more detailed account of the events hap-

pening at the *p-n* boundary, we can understand why the current cannot be completely rectified.

As a matter of fact, if the positive pole is connected to the *p*-type crystal (or half), the barrier is lowered. The voltage drives the electrons ahead. If the positive pole is connected to the *n*-type part, the electric field set up by the power source coincides in direction with that of the barrier. The field in the junction is increased. Now the number of electrons capable of surmounting the barrier, as well as the number of holes capable of moving in the opposite direction, are reduced. This raises the resistance in the junction region, leading to the so-called asymmetric volt-ampere characteristic.

As we can now see, a more thorough consideration clearly explains why the rectification in the junction layer cannot be complete.

3. Electromagnetism

Measure of Magnetic Field Intensity

The interaction between bars and needles made of certain kinds of iron ores has been observed since ancient times. These items have a certain peculiar property: one end of the needles (or bars) points to the north. Two poles, north and south, can be ascribed to such needles. It can be readily shown that like poles repel and unlike poles attract each other.

The behaviour of such special bodies was comprehensively investigated by William Gilbert (1540-1603). He explained the laws of their behaviour at various regions of the earth and the rules of interaction between one another.

On July 21, 1820, a Danish physicist Hans Christian Oersted (1777-1851) published and widely acclaimed his discovery in an article with the extremely strange name: "Experiments Concerning the Effect of an Electrical Conflict on a Magnetic Needle". The small article, only four pages long, announced to the reader that Oersted (to be more exact, one of the members of the audience at his lecture) noticed that a magnetic needle is deflected when placed near a wire through which an electric current is passed.

This was soon followed by another discovery. The eminent French physicist André Marie Ampère (1775-1836) found that electric fields interact.

Thus, magnets affect other magnets and currents, and currents affect other currents and magnets.

These interactions as well as electrical phenomena can be conveniently described by introducing the concept

of a field. We shall contend that electric currents, and natural and artificial magnets set up magnetic fields.

It is worth emphasizing here that the reality of electric and magnetic fields or, in other words, the fact that a field is a form of matter, can be proved only by investigating variable fields. For the time being, a field is simply a convenient concept and nothing more. As a matter of fact, the sources of a magnetic field can be hidden behind a screen, and we can make sure that it exists in space by observing the actions it performs.

In the presence of a magnetic field the same systems react that set up the field, i.e. a magnetic field affects magnetic needles and electric currents. The first question that arises before an investigator studying magnetism is the "probing" of the space in which the magnetic field exists. When we specified an electric field, we determined at each point of the field the force acting on unit charge. How do we go about describing a magnetic field?

In the general case, a tiny magnetic needle behaves in a quite complex way. It turns in a definite manner, but sometimes moves in a straight line. To characterize a magnetic field, we should not permit motion of the needle. First of all, we should find out the direction its north pole points to (i.e. the end that points north when there are no currents or magnetic objects in the vicinity).

We explained above that a convenient graphical way of describing an electric field is by means of lines of force. The direction of the electric lines of force indicates which way the positive charge is deflected. The density of the lines corresponds to the magnitude of the force. We can proceed in a similar way in describing a magnetic field. The end of a freely rotating magnetic needle indicates the direction of the lines of force.

What are we to take as the measure of the "intensity"

of a magnetic field? By means of a simple device, we can, of course, measure the moment of force acting on the magnetic needle. It is worthwhile, however, to search for a better way. A magnetic needle is a kind of "thing in itself". In conducting experiments with a magnetic needle, we must simultaneously search for the measure of "intensity" of the magnetic field and a measure characterizing the needle. Physicists prefer to avoid such situations. As the saying goes: If you run after two hares, you will catch neither. It is better to chase one hare first and then the other.

Hence, for the time being, we shall restrict the function of the magnetic needle to an indication of the pattern of the lines of force. To find a qualitative measure of the "intensity" of a magnetic field, we shall consider one of Ampère's experiments. As far back as 1820 Ampère discovered that a current loop behaves much as a magnetic needle. Namely, a current loop turns in a magnetic field so that a normal to the plane of the loop points in the same direction as a magnetic needle would, i.e. along the lines of force. The role of the north pole is played by the face of the loop you are looking at when the current is flowing counterclockwise in the loop.

In contrast to a magnetic needle, a current loop is not an item we find difficulty in characterizing. The properties of a current loop are uniquely specified by the current, loop area and the direction of the normal to this area. We can expect that such a loop should be a suitable instrument for probing a magnetic field.

We decide, then, to use the torque acting on a current loop as a measure of the "intensity" of a magnetic field. There is no reason to suppose that this instrument is less convenient than a magnetic needle. A resourceful investigator can make a loop of tiny area and devise a simple way to counterbalance the rotation of the field by compressing a calibrated spring.

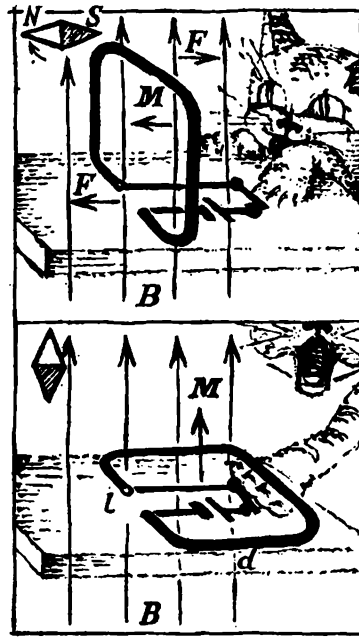


Figure 3.1

First we examine the behaviour of various test loops at some definite point in a constant magnetic field. These investigations yield the following result: the moment of force is proportional to the product of the current by the area of the loop. This means that the test loop is characterized by the product of the current by the area and not by each separately.

In addition to this product, we must also know how the normal to the loop is located with respect to the direction of the field. The loop behaves similar to a magnetic needle, as mentioned above. Hence, if it is located so that its positive normal, i.e. the vector emerging from

its north face, is directed along the lines of force, the loop remains in this position (the moment of force being equal to zero) as shown below in Figure 3.1. If the loop is located with its normal perpendicular to the lines of force (above in Figure 3.1), the moment of force reaches its maximum value.

It follows from the foregoing that it would be expedient to introduce a new concept at this point. This concept, as we shall see, proves to be of prime significance. We shall characterize the current loop by the vector M, which we shall call the *magnetic moment* (see Figure 3.1). The magnitude of the magnetic moment is taken equal to the product of the current I by the loop area $A = ld$. Thus

$$M = IA$$

Vector A has the direction of the positive normal to the plane of the loop.

Now we have an instrument that can be used to measure a field. It is most convenient to measure the maximum moment of force acting on the test loop.

Going from one point of the field to another or changing the field by moving its sources or varying the currents setting up the field, we obtain different values of the moment of the couple of forces F acting on the test loop. The maximum moment of force can be written as

$$N = BM$$

where B is a quantity that we take as the measure of the field. It is called the *magnetic induction*. Thus, the magnetic induction is equal to the maximum moment of force acting on a test loop with unit magnetic moment.

We take the density of lines of force, i.e. their number per unit area, proportional to value B. Vector B is directed along the lines of force.

The magnetic moment, magnetic induction and our old friend the moment of force are all vectors. After thinking it over, we must admit that these vectors differ from the vectors of displacement, velocity, acceleration, force, etc. As a matter of fact, the velocity vector, for example, of a body indicates the direction of motion of the body, the vectors of acceleration and force indicate the direction of attraction or repulsion. In these cases, the arrowhead at the end of the line symbolizing the vector has an entirely objective and real meaning. As to our new acquaintances and the moment of force, these are horses of another colour. The vectors are directed along the axis of rotation. It is clear that an arrowhead put at one or the other end of a line symbolizing an axis of rotation is of a wholly arbitrary nature. It is necessary, however, to agree upon the direction of the vectors. An arrowhead at the "end" of an axis of rotation is meaningless. But the direction of rotation has an objective meaning, and this is what we should try to specify by the arrowhead. It has been agreed that the arrowhead is put on the axis in such a manner that when we face the arrowhead, rotation is either clockwise or counterclockwise. Physicists are accustomed to the latter.

These two types of vectors have expressive names that speak for themselves: *polar and axial vectors.*

After measuring the fields of various systems, we can formulate the following rules. Magnets always have two poles: a north pole near which each line of force starts and a south pole near which it ends. Naturally, we cannot determine by such experiments what happens to the lines of force inside the magnet.

With respect to the magnetic fields of currents (Figure 3.2), we discover the following law: the magnetic lines of force are circular and surround the current conductor. If we look along the conductor in the direction of the current, the lines of force have the clockwise

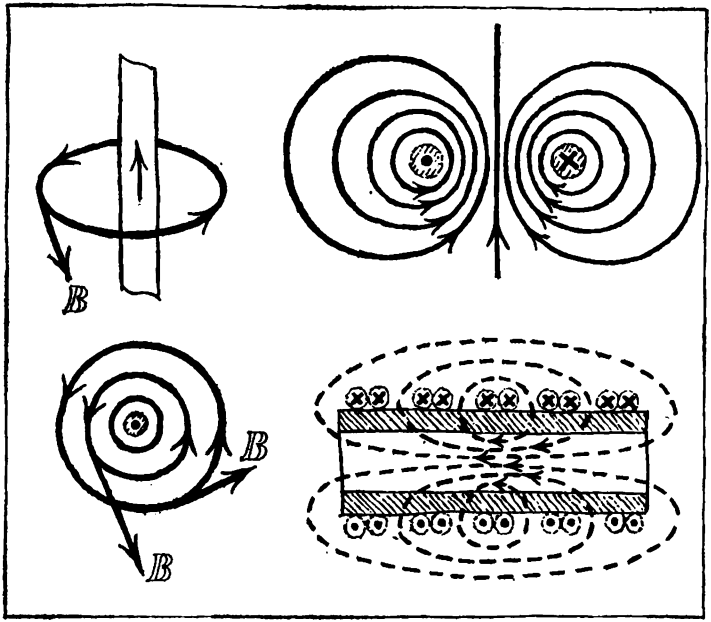


Figure 3.2

direction. A point or an X in diagrams or drawings indicates (and this is universally accepted) that the current flows toward or away from us.

As is clear from the formula, the magnetic moment is measured in amperes multiplied by the loop area in square metres.

The unit of magnetic induction is the *tesla*. One tesla (T) is equal to 1 kg/(A-s^2).

Magnetic fields are set up by currents and permanent magnets. Magnetic fields affect currents and permanent

magnets. If for some reason the investigator does not wish to resort to the concept of a magnetic field, he can classify all kinds of interaction in which magnetic fields participate into four groups. These are: magnetic, i.e. the action of a magnet on another magnet; electromagnetic, i.e. the action of a current on a magnet; magnetoelectric, i.e. the action of a magnet on a current; and, finally, electrodynamic, i.e. the action of a current on a current.

This terminology is mainly used in engineering. An instrument is said to be magnetoelectric, for instance, if the magnet is fixed and the current-carrying loop is movable.

The electrodynamic interaction is the basis for the modern definition of a unit of current. This definition is as follows: an *ampere* is the equivalent of a constant current which, in passing along two parallel and straight conductors of infinite length and negligible cross section, located in a vacuum at a distance of one metre from each other, would develop a force between these conductors equal to 2×10^{-7} newton per metre of length.

In the International System of Units (SI), accepted all over the world, the ampere is one of the fundamental units. Correspondingly, the coulomb is defined as an *ampere-second.* I must admit to the reader that I prefer the system in which the quantity of electricity is the fundamental unit and is expressed in terms of the mass of deposited silver in electrolysis. But metrologists must know best. Evidently, the above definition must have some merits, though, it seems to me that in practice a measurement of the electrodynamic force with high accuracy is a far from simple matter.

Since he now knows how to determine the direction of a magnetic field, as well as the rules for finding the direction of the force exerted on the current by a magnetic field (to be discussed a little further on), the reader

can deduct for himself that currents flowing parallel in the same direction attract one another, and those flowing in opposite directions repel one another.

Effects of a Uniform Magnetic Field

A magnetic field is said to be *uniform* if its effect on any device indicating its presence is the same at various places in the field. Such a field can be set up between the poles of a magnet. Naturally, the closer to each other the poles are located and the larger the flat end faces of the magnet, the more uniform the field.

We have already discussed the effect of a uniform magnetic field on a magnetic needle and on a current loop: if there is no counterbalancing spring, then they locate themselves in the field so that their magnetic moment coincides with the direction of the field. Their "north pole" faces the "south pole" of the magnet. This fact can be expressed in the words: the magnetic moment becomes aligned with the lines of force of the magnetic field.

Let us consider the effect of a magnetic field on moving charges.

It is utterly simple to show that such an effect exists and that it is no small one. We just take a plain horseshoe magnet of the kind used in school physics classes and bring it near an electron beam produced by an electron gun. The bright spot on the fluorescent screen is displaced and moves from place to place as we move the magnet.

From a qualitative demonstration of this phenomenon we can pass on to a quantitative investigation. We find that the force exerted by a magnetic field of magnetic induction B on an electron travelling at the velocity v at right angles to the lines of force is

$$F = evB$$

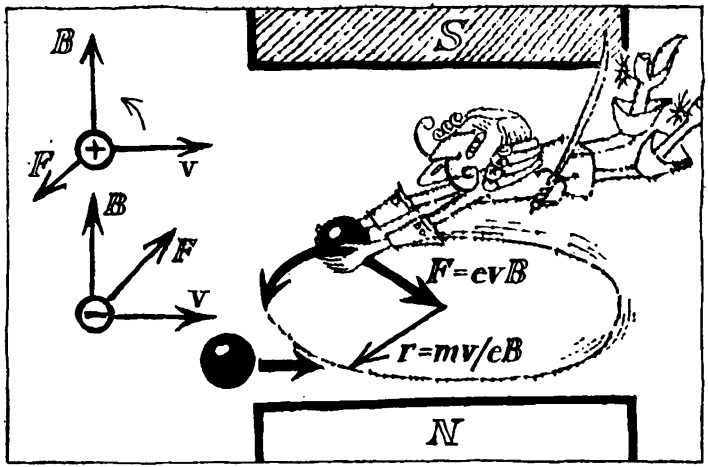


Figure 3.3

where e is the charge of the particle (this law is valid, not only for electrons, but for any charged particles).

When, however, a charged particle travels along a line of force in a magnetic field, the field has no effect on the particle. Readers familiar with trigonometry should have no difficulty in writing an equation for the force exerted on a particle travelling at some angle to the direction of the field. We shall not clutter up the text with equations that will not be required further on.

But we have not yet said anything about the direction of the exerted force, though this is of prime importance. Experiments show that the force is perpendicular both to the direction of motion of the particle and to the magnetic induction. In other words, the force is perpendicular to a plane passing through vectors v and B. But this is not all. Every medal has two sides. How do

they differ in our case? In the direction we must rotate one vector so that it coincides with the other. If we see the rotation of vector v toward vector B through the angle less than 180° as being counterclockwise, we are on the side of the positive normal.

The simple vector diagrams at the left in Figure 3.3 indicate that a positively charged particle is deflected by the field in the direction of the positive normal. An electron is deflected in the reverse direction.

Consider now the interesting result that follows from this law for an electron flying at right angles into a constant magnetic field (at the right in Figure 3.3). Try to figure out the path described by the electron in the field. It will, of course, travel in a circle. The force exerted by the field is a centripetal force and we can readily calculate the radius of the circle by equating mv^2/r and evB. Thus, the radius of the path is

$$r = \frac{mv}{eB}$$

Note that we could calculate the properties of a particle from its behaviour. But again we find ourselves in the same predicament as when we investigated the motion of particles in an electric field. We cannot determine the electric charge and the mass of the particle separately! Here as well we determine the ratio e/m.

Thus, a particle travels along a circle if its velocity is directed at right angles to the magnetic field; a particle travels by inertia if its velocity is directed along the magnetic field. What does it do in the general case? Your answer is ready, of course. The particle travels along a helix whose axis is a line of force. The helix consists of tightly or loosely wound coils, depending upon the initial angle of entry of the electron into the magnetic field.

Since a magnetic field acts on a moving particle, it

should also exert a force on each piece of wire carrying a current. Consider a portion of length l of an electron beam. Assume that there are n particles in this portion. The force exerted on a wire of the same length, along which the same number of particles travel at the same velocity, is equal to $nevB$. The current is equal to the total charge passing through the wire in unit time. The time τ during which the electrons being discussed travel a path l equals

$$\tau = \frac{l}{v}$$

This means that we can write the equation for the current as

$$I = \frac{ne}{\tau} = \frac{nev}{l}$$

Substituting the velocity from this equation

$$v = \frac{Il}{ne}$$

into the formula for the force acting on a portion of an electron beam, we obtain the force exerted on a conductor of length l. Thus

$$F = IlB$$

This is valid only when the wire is perpendicular to the field.

The direction that a wire carrying a current is deflected can be determined by the diagram shown in Figure 3.3.

In deference to the investigators working in the 19th century, I have included Figure 3.4. As a matter of fact, this drawing is not of only pure academic interest. It can serve as an aid in remembering the rule for the deflection of currents in magnetic fields. The drawing shows how the field set up by a current (flowing away from us)

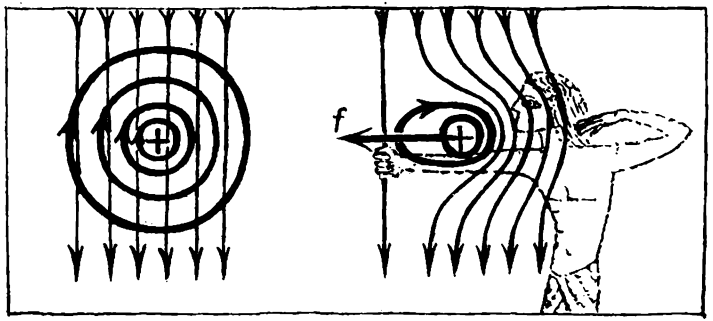


Figure 3.4

combines with the external field. The result of this combination is illustrated at the right. If we conceive of lines of force as being tensioned ether and having material properties (a widespread point of view last century), the direction the conductor is displaced can be visually interpreted: the conductor is simply pushed out by the field.

Next we shall demonstrate that the effect of a magnetic field on a moving charge and on a piece of current-carrying conductor is the same phenomenon with which we began our discussion of the effects of a magnetic field.

Let us return to Figure 3.1. It shows the forces acting on a current-carrying loop. No forces are exerted on the sections of the wire that are aligned with the field. The force couple acts on the other two sections. We can see from the drawing that the moment of this couple is precisely equal to the product of the force by the arm:

$$N = IlBd = IAB = MB$$

Thus the expression for the moment of force as the product of the magnetic moment of the loop by the

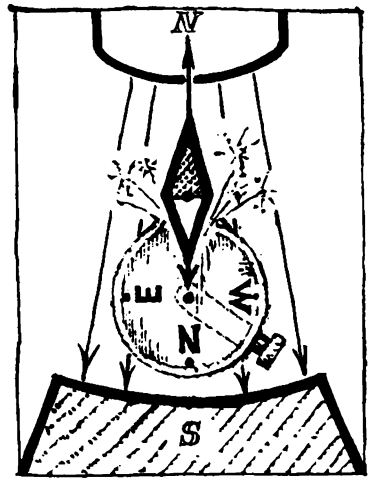


Figure 3.5

magnetic induction follows directly from the equation for the force exerted on a charge.

The formula $F=evB$ with which we began this section is called the *Lorentz equation* after Hendrik Antoon Lorentz (1853-1928), the Dutch physicist who proposed it in 1895.

Effects of a Nonuniform Magnetic Field

It is quite easy to set up a nonuniform magnetic field. We can, for instance, impart curved shapes to the faces of the poles as shown in Figure 3.5. Then the paths of the lines of force are as illustrated.

Assume that the poles are sufficiently far away from each other and place a magnetic needle near one of the poles. As mentioned before, such a needle can move in a straight line, in the general case, as well as rotate. We observe rotary motion alone of a needle or current loop

only if the magnetic field is uniform. Both kinds of motion occur in a nonuniform field. The needle turns so that it is aligned with the lines of force and is attracted toward a pole (see Figure 3.5). It is pulled into the region where the field is stronger. (The artist has, of course, overdone it; it is improbable that even an extremely strong field can split the compass in two.)

To what is this behaviour due? Evidently, to the fact that not only a force couple acts on the needle in a nonuniform field. The "forces" exerted on the north and south poles of a needle placed in a nonuniform field are not the same. The end located in a stronger region of the field is subject to a larger force. Therefore, after the needle turns, the arrangement of the forces is as shown in the drawing. There is an excess force exerted toward the side where the field is stronger.

True, a current loop of exceptionally thin wire behaves in exactly the same way. Thus, in beginning with a model of a needle with two "poles", my aim was only to provide a more pictorial representation.

What, then, is the law of nature pertaining to our phenomena? What is the force equal to? Experiments and calculations indicate that, for any system having the magnetic moment M, this force is the product of the moment of the system by the slope of the curve representing the increase in the field strength (magnetic induction).

Assume that the magnetic needle aligns itself with a line of force. The magnetic induction values differ at the places where the north and south poles of the magnetic needle are located. We plot a curve of the field along the line passing through the poles. For the sake of simplicity, we replace the section of the true field curve between the poles with a straight line. The shorter the needle, i.e. the closer its poles to each other, the more accurate our approximation. The slope of the curve, i.e. the tangent

of the angle our straight line on the diagram makes with the horizontal axis, is equal to the quotient of the difference in the values of the field strength divided by the length of the needle. Thus

$$F = M \frac{(B_N - B_S)}{l}$$

where l = length of the needle; and B_N and B_S = field strengths at the north and south ends of the needle. (Do not be surprised that the tangent of an angle turns out to be a dimensional quantity.)

If we replace the fraction by the value of the tangent of the angle of inclination of the straight line tangent to the curve representing the field at the point where the particle interesting us is located, the "poles disappear" and the formula is valid for any particle or system of particles.

In conclusion, we can say that in a nonuniform field a system or a particle with a magnetic moment is attracted to the poles of the magnet or repelled by them in accordance with the direction of the magnetic moment: either along with or opposite to the lines of force.

But can the magnetic moment align itself against the direction of the field? It certainly can! We shall discuss below the cases in which this occurs.

Ampèrian Currents

Right up to the nineteenth century it was no difficult matter to advance a physical theory. If a body becomes heated, that means it contains more caloric. If a medicine makes you fall asleep sooner, it contains a soporific power. Certain rods, made of iron ores, point to the north. A strange behaviour, but immediately understood if we say that such rods and needles possess a magnetic spirit. We know that magnetic needles have ren-

dered good service to seafarers since ancient times. But sometimes they played tricks. So what; no problem: evil spirits are to blame! Likewise, it was no surprise that iron, steel and certain alloys could be magnetized. These are simply bodies that a magnetic spirit (or soul) can easily be imparted to.

After the discoveries made by Oersted and Ampère, it became clear that a bridge could be constructed between electrical and magnetic phenomena. At one time, two theories enjoyed equally wide support. From one point of view, everything became clear if we assume that the wire along which an electric fluid flows is converted into a magnet. The other point of view was expounded by Ampère. He contended that the magnetic spirit of iron ores consists of microscopic electric currents.

Ampère's point of view seemed to be more logical. No particular importance was attached to this theory because, in the first half of the nineteenth century, no one even dreamt of the feasibility of really discovering such currents. They even doubted that the world is built up of atoms and molecules.

But in the twentieth century, when physicists conducted a whole series of ingenious experiments that proved that the world around us consists of atoms and that atoms consist of electrons and atomic nuclei, the existence of Ampèrian currents was finally taken to be a real fact that could be employed in an effort to understand the magnetic properties of substances. Most scientists agreed that the "molecular currents" proposed by Ampère are formed by the motion of electrons about atomic nuclei.

It seemed possible to explain magnetic phenomena by means of these conceptions. As a matter of fact, an electron travelling around a nucleus can be likened to an electric current; we have the right to ascribe a magnetic moment to this system and relate it to the angular momentum of the moving charged particle.

It is extremely simple to prove this last statement.

Assume that an electron revolves in a circle of radius r. Since the current equals the charge carried in unit time, the revolving electron can be likened to a current $I=Ne$, where N is the number of revolutions per second. The velocity of the particle can be related to the number of revolutions per second by the equation $v=N\times 2\pi r$. Then the current equals

$$I = \frac{ve}{2\pi r}$$

It is natural to call the magnetic moment of an electron revolving about a nucleus its *orbital moment*. It equals

$$M = IA = \frac{ve}{2\pi r}\pi r^2 = \frac{1}{2}evr$$

We remind the reader (see Book 1) that the angular momentum of a particle is $L=mvr$ and find that between the angular momentum and the magnetic moment we have the following relationship that is of great significance in atomic physics:

$$M = \frac{e}{2m}L$$

It follows that atoms must have magnetic moments.

By various procedures, on which we shall not dwell here, we can obtain the atomic gas of various substances. Using two slits in a gas chamber we can produce beams of neutral atoms of hydrogen, lithium, beryllium, etc. They can be passed through a nonuniform magnetic field and traces of the beam can be observed on a screen. The question we put to nature is the following: Will the stream of atoms be deflected from a straight line, and if so, how?

The atom has an orbital moment and, consequently, behaves similar to a magnetic needle. If the magnetic

moment is directed along the field, the atom is deflected toward the region of a strong field. In antiparallel arrangement of the moment and field, the atom is deflected toward the region of a weak field. The amount of deflection can be calculated by a formula similar to the one given on p. 116 for the force acting on a magnetic needle.

The first thing that occurs to us is that the magnetic moments of atoms are arranged at random. If this is so, we should expect that the trace of the beam is blurred. But experiments yielded entirely different results. The beam of atoms is never blurred; it splits into two, three, four or more components, depending upon the kind of atoms. This splitting is always symmetrical. Sometimes the components of the beam include an undeflected beam, sometimes there is no undeflected beam, and sometimes the beam does not split up at all.

It follows from this experiment, one of the most important ever conducted in physics, firstly, that the motion of electrons about an atom can really be likened to a closed-circuit electric current. It can be likened in a narrow and quite definite sense: like closed-circuit currents, atoms can be ascribed a magnetic moment. Further, the magnetic moments of the atoms can make only certain discrete angles with the direction of the magnetic induction vector. In other words, the projections of the magnetic moments on this direction are quantized.

A great triumph for theoretical physics was the fact that these results had been predicted in elaborate detail. It follows from theory that the angular momentum and the magnetic moment of the electron, due to the motion of atomic electrons in the field of the nucleus (these moments are said to be orbital*), are antiparallel and their

* This name is of historical origin: the development of atomic theory began with the supposition that an atom resembles the solar system.

projections on the direction of the field can be written in the form

$$L_z = m\frac{h}{2\pi} \quad \text{and} \quad M_z = m\mu_B$$

where m=a whole number that can take the values 0, 1, 2, 3, ...; $h/2\pi$=the smallest value of the projection of the angular momentum; and μ_B=the smallest value of the projection of the magnetic moment. The values of h and μ_B are determined from experiments:

$$h = 6.62 \times 10^{-27} \text{ erg-s} \quad \text{and} \quad \mu_B = 0.93 \times 10^{-20} \text{ erg/Gs}$$

We may add that these important constants of physics have been named after the great scientists that laid the foundations of quantum physics: h is called *Planck's constant* after the German physicist Max Karl Ernst Ludwig Planck (1858-1947) and μ_B is the *Bohr magneton* after the Danish physicist Niels Henrik David Bohr (1885-1962).

The postulates of quantum mechanics were not sufficient, however, to provide a comprehensive understanding of the different ways that the beams of various atoms are split. Even the simplest of atoms, those of hydrogen, behaved unexpectedly. It became necessary to add another exceptionally vital hypothesis to the laws of quantum mechanics. We have already mentioned it once. It consists in ascribing its own (intrinsic) angular momentum, called *spin*, and the corresponding intrinsic magnetic moment to the electron (or to any elementary particle, as was found later). To understand why it is inevitable to liken the electron to a magnetic needle, we must first investigate in more detail the motion of atomic electrons.

Electron Cloud of the Atom

It is impossible to observe the motion of an electron. And what is more, we cannot even hope that the advance of science can ever provide the opportunity of seeing an electron. The reason is sufficiently clear. To "see" something, we must first "illuminate" it. But "illumination" means to subject the electron to the energy of some kind of ray. The electron is of such tiny mass that any interference, by means of an instrument for observing it, inevitably causes the electron to leave the place it was previously located at.

Not only the meager information about the structure of the atom which we are about to convey to the reader, but the whole consistent doctrine of the electronic structure of matter, is the result of theoretical rather than experimental investigations. We are, however, sure of its validity because of the innumerable amount of effects observed in experiments and rigorously derived from theory by logical reasoning. We establish the picture of electronic structure, invisible to us, with the same degree of assurance that Sherlock Holmes established the picture of a crime from the clues left by the criminal.

Primarily, a vast source of confidence in this theory is the fact that the picture of electronic structure is predicted by means of laws of quantum mechanics that were established by other experiments.

We have already mentioned that the atomic number of a chemical element in Mendeleev's periodic table is none other than the charge of its nucleus or, what is the same thing, the number of electrons belonging to a neutral atom. An atom of hydrogen has one electron, an atom of helium has two, lithium has three, beryllium has four, etc.

How do all these electrons travel? This question is far from simple and its answer is of an approximate nature.

Difficulties arise from the fact that the electrons interact with one another and not only with the nucleus. Fortunately, the mutual repulsion (avoidance) of the electrons plays a smaller role than the motion that is due to the interaction of the electron with the nucleus. Only this circumstance enables us to draw conclusions on the nature of motion of electrons in various atoms.

Nature has allocated to each electron a spatial region in which it travels. According to their shape these regions of the electrons are divided into categories denoted by the letters *s*, *p*, *d*, and *f*.

The simplest is the "apartment" of the *s*-electrons. It is a spherical layer. We know from theory that most of the time the electron is at the middle of the spherical layer. Hence, any talk about a circular orbit of such an electron is a crude simplification.

The region of space through which the *p*-electron travels is entirely different. It resembles a dumbbell used for physical exercise. Other categories of electrons have even more complex regions of existence.

Electronic theory can indicate (though not without resorting to experimental data) how many electrons of each kind each element of Mendeleev's table contains.

Is this distribution of electrons according to their types of motion related to their distribution among the *K*, *L*, *M*, ... energy levels discussed in the preceding chapter? It is, and most directly. Theory and experiments show that electrons belonging to the *K*-level can only be of the *s*-type; to the *L*-level, of only the *s*- and *p*-types; to the *M*-level, of the *s*-, *p*-, and *d*-types; etc.

We shall not discuss the electronic structure of atoms in any particular detail, restricting ourself to an account of the structure of the first five elements of the table. Atoms of hydrogen, helium, lithium and beryllium have only *s*-type electrons. The boron atom has four *s*-electrons and one *p*-electron.

The spherical symmetry of the region of space in which the *s*-electron travels casts a doubt on our discussion of the magnetic moment of an atom containing a single electron. As a matter of fact, since the angular momentum can take on identical values directed to all sides with equal probability, the average rotational moment and, consequently, the magnetic moment of such a system should equal zero. Quantum physics also reaches this same natural conclusion: atoms containing only *s*-electrons cannot have a magnetic moment.

But if this is so, the beams of atoms of the first four elements of Mendeleev's table should not be deflected in a nonuniform magnetic field. Is this what we observe? It was found that these predictions are upset for the atoms of hydrogen and lithium. Beams of these atoms behave in an exceptionally strange manner. In both, the beam of atoms is split into two components deflected in opposite directions the same distance from the initial direction. Incomprehensible!

Magnetic Moments of Particles

Electron spin made its first appearance on the scene in 1925. The necessity of including it in the participants of the events taking place in the microcosm was first revealed by Samuel Abraham Goudsmit (b. 1902) and George Eugene Uhlenbeck (1900-1974). Proposing that the electron has its own (intrinsic) angular momentum, these two investigators showed that all the confusion that had accumulated by that time in the interpretation of atomic spectra could be readily eliminated by this new concept.

The experiments mentioned above for splitting atomic beams were performed somewhat later. When it became clear that here also only the concept of spin could pro-

vide a complete explanation of the observed facts, all physicists finally accepted it.

A short time elapsed and it was found that intrinsic angular momentum, or spin, is a property possessed by all elementary particles and not only electrons.

We have already mentioned that the name spin is witness to a natural tendency toward visualization. Since the angular momentum was first introduced in physics as a property of a rotating solid, certain physicists immediately resorted to the graphic picture of a particle rotating about its axis when it was found necessary to ascribe a certain value of the angular momentum to elementary particles in order to save the conservation laws. This naive concept holds no water: we have no more right to speak of an elementary particle rotating about its axis than we would of a mathematical point.

The supporters of visualizability were able to assess the size of the electron on the basis of certain circumstantial evidence. To be more exact, they established that if this concept is applicable to the electron, its size must be less than a certain definite value. The value of spin is known; we shall give it on the following page. Assuming a shape for the electron, we can calculate with what velocity "points on its surface" rotate. This velocity was found to be greater than that of light. Thus, if we are persistent in advocating such particle rotation, we must throw overboard the theory of relativity.

Perhaps, the most irrefutable argument against visualizability is the fact that the neutron, which carries no electric charge, possesses spin. Why is this argument decisive? Judge for yourself.

If a particle can be conceived of in the form of a charged sphere, its rotation about its axis should produce something resembling an Ampèrian current. But if a neutral particle also has angular momentum as well as a magnetic moment (we shall mention these properties of the neu-

tron briefly in Book 4), any analogy with an Ampèrian current is out of the question.

It does not pay, of course, to pose as a prophet and state that spin and magnetic moment of elementary particles will never be made clear, even on the basis of some more general, as yet undiscovered, law. This problem has been partly solved by the theory of the brilliant English physicist Paul Adrien Maurice Dirac (b. 1902). But we cannot give our reader even a general idea of this theory; it is so abstract. But, as for today, we must consider the "arrows" representing the angular momentum and magnetic moment of a particle to be primary concepts (not reducible to something simpler).

About fifty years ago, the majority of physicists upheld the point of view of Einstein, who wrote: "Every physical theory should be such that it can be illustrated, apart from any calculations, by means of simple images." Alas, this opinion of the great thinker turned out to be mistaken. For many years, physicists have been calmly applying theories containing measurable quantities that we cannot associate with any visual image.

The electron and other elementary particles have no "poles". In many cases, we confidently speak of them as point particles, agreeing that the idea of shape is not applicable to elementary particles. Nevertheless, we are obliged to ascribe to them two vector properties, an angular momentum (spin) and a magnetic moment. These two vectors always lie along a single line and are parallel in some cases and antiparallel in others.

Experiments show that the general formulas for the projections of the angular momentum and the magnetic moment, given on p. 120, are also valid for the intrinsic moments as well. All experiments, both in spectral analysis and in splitting atomic beams in a nonuniform magnetic field, can be irreproachably interpreted if quantity m in the formula for the projection of the angular

momentum of an electron is allowed to take two values: $\pm 1/2$. As for the formula for the projection of the magnetic moment, quantity m can also have two values: ± 1.

Electron spin has the numerical value (1/2) $(h/2\pi)$ and may be disposed only in two directions: along the field or against the field. As to the magnetic moment of the electron, it imitates the spin, also having only two orientations in the field. Its numerical value equals one Bohr magneton.

Now let us return to the experiments with atomic beams. We shall show how easily all the specific features of atomic beam splitting can be explained by the concept of spin.

Indeed, how can we explain the fact that beams of helium and beryllium atoms do not split? As follows. The electrons of these atoms have no orbital moment because they are of the *s* "kind". As to the spins of the electrons, they are in opposite directions. As a matter of fact, this statement does not follow from anything, though intuitively seems quite natural. The principle, according to which a pair of electrons in an atom establish themselves so that their spins are opposed, is called *Pauli's exclusion principle*, after the Austrian-Swiss physicist Wolfgang Pauli (1900-1958).

So many hypotheses! Yes, quite a few. But taken all together they form the shapely structure of quantum physics from which so many consequences follow that there is not the least uncertainty but that spin must be ascribed to the electron, that the value of 1/2 must be given to the spin quantum number and that the spins of a pair of electrons must obey the Pauli principle. Not even a single physicist has a shadow of a doubt on these matters. The sum of these hypotheses represents the structure of the microcosm.

Let us return to our atomic beams. We have just

explained why no splitting is observed in beams of helium and beryllium atoms.

But why do hydrogen and lithium behave differently?

The hydrogen atom has a single electron. Its orbital moment equals zero because it is an *s*-electron. The projection of its spin can have only two values: plus 1/2 and minus 1/2, i.e. the spin can be aligned either opposite to or along the direction of the magnetic field. This is why an atomic beam splits into two components. The same occurs with lithium atoms because two of its electrons "compensate for" their spin and the third behaves like the single electron of the hydrogen atom.

The atoms of other elements that have a single unpaired electron in their outer shell behave in exactly the same way.

To explain why the atomic beams of other elements split into a large number of components it would be necessary to cite several other theorems without giving their proof. They are proved in quantum physics. By taking into account the facts that only *s*-electrons have no orbital moment and that the spin of an electron is manifested only when the electron is alone at its energy level, physicists managed to comprehensively explain the behaviour of atomic beams of all kinds. After studying this fascinating chapter of physics even the staunchest skeptic becomes convinced that all the unproved assumptions accepted in quantum physics are general laws of nature.

I fear that many readers may remain unsatisfied by these statements. Experiments on the deviation of atomic beams in a nonuniform magnetic field are insufficient in themselves, of course, to introduce such a strange concept as spin. This book is too small, however, for me to cite the vast number of facts that demand full rights for spin among the authentic phenomena of the physical sciences.

For example, how worthy as such proof is the phenomenon of magnetic resonances, having nothing in common with the aforesaid. Radio waves of the centimetre range are absorbed by a substance if they have to flip (invert) the spin. No difficulty is encountered in calculating the energy of the interaction between the magnetic moment of an electron and the constant magnetic field into which the substance is placed in magnetic resonance experiments. This energy constitutes the difference between two energies (parallel and antiparallel arrangement), which is equal to a quantum of the absorbed electromagnetic wave. We can determine the value of the wave frequency in this experiment with exceptionally high precision, verifying the fact that it absolutely coincides with the frequency we calculated on the basis of the known induction of the field and the value of the electron's magnetic moment.

This phenomenon forms the foundation for a large branch of science: the study of electron resonance. It is remarkable that the same events, but, naturally, in another wavelength range, are observed for atomic nuclei. Nuclear magnetic resonance is the most essential method used in studying the chemical structure of substances.

Before going on, it will, perhaps, be helpful to sum up all the facts concerning systems that set up magnetic fields and respond to the presence of a magnetic field.

First of all, we must emphasize again that Ampère's hypothesis was only partly substantiated: magnetic fields are set up not only by moving electric charges. Other sources of magnetic fields are elementary particles, primarily electrons, which have an intrinsic magnetic moment. The technical classification of the kinds of interaction, given on p. 108, turns out to be inexact. Magnetic fields are set up by natural and artificial magnets, by electric currents (including streams of electrical

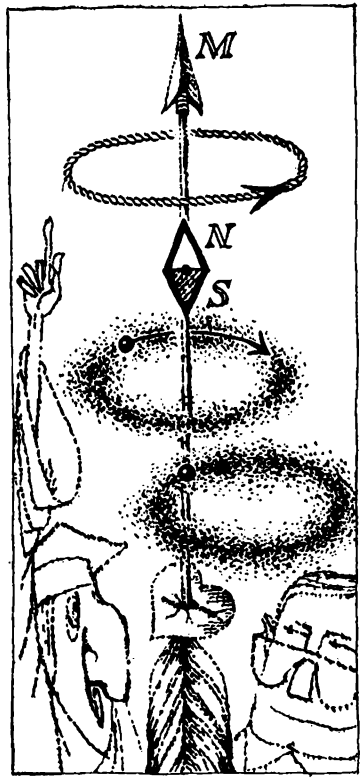


Figure 3.6

particles in a vacuum), as well as elementary particles. The same systems, as well as the particles, respond to the action of magnetic fields.

The main quantity characterizing a magnetic field and its effects is the magnetic moment vector. For currents, this vector depends upon the shape of the current loop. The moment of a needle is related in a complex manner to the atomic structure of the substance, but it

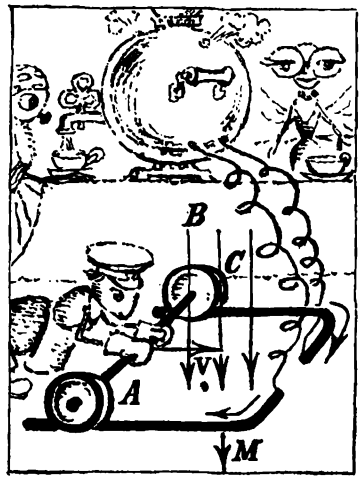


Figure 3.7

can be readily measured. Electrons travelling in the field of the nucleus have an "orbital" magnetic moment as if (note, please, this "as if") their motion about the nucleus generated an electric current. Finally, the intrinsic magnetic moment is a primary property that characterizes elementary particles.

Figure 3.6 should help you to remember this fundamental information. This drawing represents the sum of our knowledge today about the "magnetic soul" or, if you please, the magnetic heart. The French for a magnet is "aimant" (from the verb "aimer"—to love). The drawing underlines the fact that macroscopic current, a bar magnet, the orbital motion of the electron and the electron itself are all characterized by a single physical concept.

Electromagnetic Induction

Experiments show that a beam of electrons travelling in a magnetic field is deflected from a straight line. As

mentioned on p. 109, the force responsible for this deflection, called the *Lorentz force*, is directed perpendicular to the magnetic lines of force and to the velocity vector of the electrons. It is determined by the formula $F=evB$. This is the simplest expression for the Lorentz force and is valid when the direction of the electron velocity makes a right angle with the direction of the magnetic field.

If to this fact we add our certainty that a metal conductor contains free electrons, we can, by simple reasoning, come to the conclusion that upon certain motions of a conductor in a magnetic field an electric current should be produced in the conductor.

This phenomenon, which, we might say, is fundamental for all modern engineering, is called *electromagnetic induction*. We shall proceed to derive its formula.

Illustrated in Figure 3.7 is a current-conducting loop consisting of rod AC of length l, which rolls along metal wires between the poles of a magnet without breaking the circuit of the loop. If the rod is rolled in the direction perpendicular to the lines of force, its electrons are subject to a force and an electric current flows along the loop circuit.

We reach a conclusion whose importance cannot be overestimated: an electric current can be produced in a conductor of a closed circuit even if the circuit contains no storage battery or other current source.

We can calculate the emf, i.e. the work required to carry unit charge along the closed loop. Work is the product of a force by the path. It is performed only on the part of the loop moving in the field. The length of the path is l and the force per unit charge is vB.

This developed electromotive force is called the *induced emf*. Its value is determined by the formula

$$\mathcal{E}^{\text{ind}} = vBl$$

It is desirable to generalize this formula so that it is suitable for any motion of any conducting loops. We derive this generalization as follows. During time τ the rod-type conductor travels the distance x, its velocity v being equal to x/τ. The area of the conducting loop is reduced by the amount $A = xl$. Then the formula for the induced emf takes the form

$$\mathscr{E}^{\text{ind}} = \frac{BA}{\tau}$$

What does the numerator in this formula signify? This is sufficiently clear: BA is the amount of change in the magnetic flux (the number of lines of force) through the loop.

Our proof has, of course, been carried out for a very simple case. The reader will have to take my word that an entirely rigorous proof can be carried out for any case. The formula obtained is of most general application and the law of electromagnetic induction can be formulated as follows: an induced emf is always developed when the number of lines of force through the loop is changed. Here the value of induced emf is numerically equal to the change of magnetic flux per unit time.

There may be movements of a loop in a magnetic field that induce no current. There is no current when the loop moves in a uniform field parallel to the lines of force. If the loop is rotated in a uniform magnetic field, a current is induced. This also occurs when the loop is moved away from or toward a pole of a bar magnet.

Experiments indicate that our generalization is even more significant than we have found so far. We considered cases in which the current loop and the source of the magnetic field changed their relative positions. The last formula we derived says nothing about any

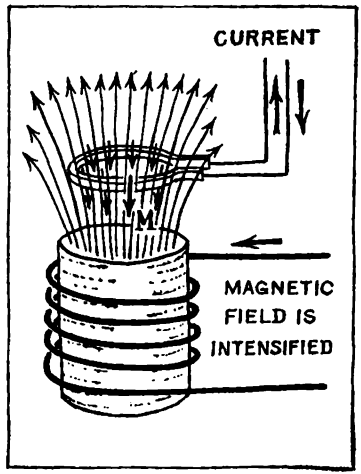


Figure 3.8

motion. The only factor is the change of magnetic flux. But a change in the magnetic flux through a conducting loop does not necessarily involve a mechanical displacement.

As a matter of fact, we can use, as the source of the magnetic field, not a permanent magnet, but a loop or, even better, a coil through which a current from any outside source is passed. By means of a rheostat or in some other way we can vary the current in this primary coil which is the source of the magnetic field. Then the magnetic flux through the loop changes even though the source of the magnetic field and the conducting loop are stationary (Figure 3.8).

Will our generalization be valid in this case as well? This question can be answered by an experiment, and the answer is "yes". Regardless of how the number of lines of force is changed, the formula for the induced emf, given on the preceding page, is valid.

Direction of Induced Current

Next we shall demonstrate that a simple universal rule exists for the direction of induced currents. Let us consider several examples from which we shall draw a general conclusion.

Returning to Figure 3.7 we can note the following. When we reduce the area of the loop, the magnetic flux through the loop is reduced. The direction of the current shown in the drawing is such that the magnetic moment of this induced current is directed along the lines of force. This means that the intrinsic field of the induced current is directed so as to "hinder" the reduction of the magnetic field. We reach the same conclusion in the reverse case. When the area of the loop is increased, the flux through the loop is also increased. But now the magnetic moment of the loop is directed against the lines of force. Again we find that the field of the induced current hinders the action that causes it.

Another example. Assume that the loop is located between the poles of the magnet in such a manner that the flux through it is equal to zero. Then we begin to turn the loop clockwise and counterclockwise. Both cases are illustrated in Figure 3.9. Solid lines show the projection of the loop in the initial position; dash lines show the projections in the turned positions when the current is induced. Using the left-hand rule, we find the direction of the induced current in each case. In our drawing, the north pole is shown at the left. Consequently, when the loop is turned clockwise, the magnetic moment of the induced current faces downward; when the loop is turned counterclockwise, it faces upward. As the angle of rotation is increased, the intrinsic magnetic field of the loop (in either case) reduces the field causing the induction more and more. Again the same rule is valid.

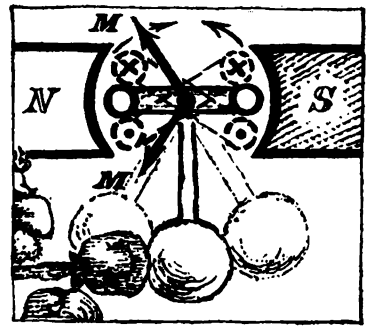
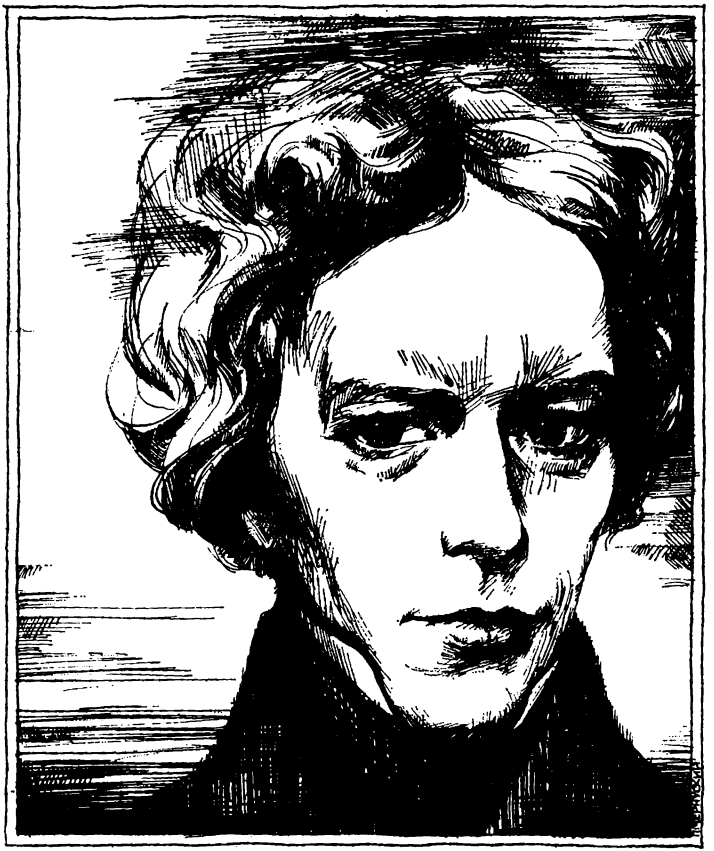


Figure 3.9

Next we shall see how our loop behaves in nonuniform fields. Return to Figure 3.8. Assume that the current of the electromagnet is constant. What happens when we displace the loop? When we move the loop toward the north pole, the magnetic moment is directed against the lines of force. If the loop is moved away from the pole, the intrinsic field of the induced current strengthens the field. We can predict such behaviour by again resorting to the left-hand rule.

How about magnetic fields set up by alternating currents? Any increase or decrease of the current in the primary coil leads to a change in flux. An emf is induced in the loop (see Figure 3.8 again).

How can we determine the direction of the current? We can no longer use a hand rule because there is no motion. This is where our generalization comes in handy. It was found that in this case as well the direction of the current, induced by reducing or increasing the number of lines of force through the loop, obeys the same rule: the induced current sets up a field in the direction that compensates for the change in the magnetic field causing the induction.



Michael Faraday (1791-1867)—famous English physicist; he discovered electromagnetic induction (in 1831). This was no chance discovery; Faraday searched for it. Faraday's laws of electromagnetic induction are the foundation of electrical engineering.

It is difficult to overestimate the importance of the laws of elec-

Discovery of the Law of Electromagnetic Induction

The discovery of electromagnetic induction belongs to those rare events that have a decisive influence on the progress of mankind. It would therefore be unpardonable not to dwell to some extent on the history of this discovery. It was made a long time before the investigation of the behaviour of an electron beam in a magnetic field. The historical course of events does not coincide in any way with the sequence we chose in the preceding sections for expounding this subject. Logic and the continuity of thought from item to item are not at all obliged to proceed in parallel with the historical succession of events.

Up to the time when Michael Faraday began his experiments that led to the discovery of electromagnetic induction, the situation in the science of electric and magnetic fields was the following.

By this time, the production of a direct current and the laws of its behaviour in electric circuits posed no serious problems to physicists. The effects of current

---

trolysis established by Faraday. This great scientist introduced and explained such terms, so widely used today, as anode, cathode anion, cation, ion and electrolyte.

Faraday proved that the intermediate medium influences electrical interaction.

Mention must be made of the discovery of magnetic rotation of the plane of polarization. The fact that all bodies belong either to paramagnetic or to diamagnetic materials was also established by Faraday.

The world had never known a more gifted experimental physicist than Michael Faraday.

on a permanent magnet and the interaction of currents had been established. It became clear that a direct current sets up a magnetic field surrounding the conductor. This field can be measured either with a magnet or by means of another current. This raised the question of the inverse phenomenon: Can a magnetic field produce a current in a conductor?

In an entry made in 1821 in his diary, Faraday set himself the task of converting magnetism into electricity. It took this great scientist ten years to achieve this aim. He had failed for so many years because he tried to obtain a current by placing a conductor in a constant field. But in 1831, his persistent efforts finally succeeded. The quotation given below, from an article written by Faraday in 1831, is the first description of this phenomenon.

"10. Two hundred and three feet of copper wire in one length were coiled around a large block of wood; other two hundred and three feet of similar wire were interposed as a spiral between the turns of the first coil, and metallic contact everywhere prevented by twine. One of these helices was connected with a galvanometer, and the other with a battery of one hundred pairs of plates four inches square, with double coppers and well charged. When the contact was made, there was a sudden and very slight effect at the galvanometer, and there was also a similar effect when the contact with the battery was broken. But whilst the voltaic current was continuing to pass through one helix, no galvanometric appearances nor any effect like induction upon the other helix could be perceived, although the active power of the battery was proved to be great, by its heating the whole of its helix and by the brilliancy of the discharge when made through charcoal."

The discovery of electromagnetic induction was the first stage in the next twenty years of Faraday's inves-

tigations whose aim was to find a unified relationship between all electrical and magnetic phenomena.

In discussing electromagnetic induction, we should also mention the names of other prominent physicists. The American physicist Joseph Henry (1797-1878) discovered self-induction in 1832. If the current in a coil changes, the magnetic field set up by this current also changes, thereby changing the flux of the field passing through this same coil and inducing an emf in its "own" circuit.

And who discovered the law governing the direction of the induced emf? The most comprehensive answer to this question can be found in the works of Lenz. Lenz's law determines the direction of the induced current. "If a metal conductor is moved near a current or magnet, a galvanic current is induced in the conductor. The direction of this current is such that if the wire were at rest, it would begin to move in the direction directly opposite to the actual movement. It is assumed that the wire can move both in the direction of its actual motion and in the opposite direction."

After 1840, a unified concept of electromagnetism was developed, step by step. The discovery of electromagnetic waves was the last and, evidently, most brilliant step.

Induced Eddy Currents

If currents can be induced in wire conductors, it would seem quite natural that they can also be induced in heavy solid pieces of metal. Each piece of metal contains free electrons. If the metal moves in a constant magnetic field, the free electrons are subject to the Lorentz force. The electrons describe circular paths, i.e. form eddy currents. This phenomenon was first observed in 1855 by the French physicist Jean Bernard Léon Foucault (1819-1868).

The laws of electromagnetic induction are equally valid whether the magnetic flux changes due to relative displacement of the metal and the source of the field or whether the magnetic field changes due to motion of the electric current setting up the field. Consequently, eddy currents are induced when the magnetic field changes with time and not only when there is a relative motion. The most convincing experiment illustrating the latter is dropping a coin between the poles of a strong magnet. It drops as if in viscous oil rather than with the usual acceleration. The idea of this experiment is obvious: eddy currents are induced in the coin. Their direction, according to Lenz's law, is such that their interaction with the primary magnetic field brakes the motion that causes the induction.

Among the useful applications of eddy currents are the following. In the first place, they are used in the so-called induction furnaces for heating to high temperatures and even melting metals. Secondly, they provide for "magnetic damping" in many indicating instruments.

An ingenious invention (and this is thirdly) is the electric meter. You have, of course, noticed that its main component is a rotating disk. The more lights you turn on or electric appliances you plug in, the faster the disk rotates.

The principle of this device consists in having two currents. One is in a circuit parallel to the load and the other is the load current itself. These currents flow in coils wound on iron cores. The alternating current magnetizes the iron cores. Since we have an alternating current, the poles of the electromagnets are continually being reversed. A sort of running magnetic field is set up between their poles. The coils are arranged so that the running field formed by both coils induces eddy currents in the body of the disk. The direction of these

eddy currents is such that the running magnetic field pulls at the disk, rotating it.

The speed of rotation depends upon the currents in both coils. As can be shown by exact calculations, this speed is proportional to the product of the current by the voltage and by the power factor (cosine of the phase angle) or, in other words, to the consumed power. We shall not dwell on the simple mechanical transmission connecting the rotating disk to the counter for indicating the watt-hours of energy used.

In the majority of cases, however, efforts are made to eliminate eddy currents. This is one of the concerns of designers of electric machinery. As all other currents, eddy currents utilize some energy of the system. These energy losses may reach such high values that it becomes necessary to resort to all kinds of contrivances. The simplest method of combatting eddy current losses is to replace large solid pieces of metal in electric machines with laminated sheet stock. In sheet metal, the eddy currents lack sufficient "elbowroom", their magnitudes are substantially lower and we obtain a corresponding drop in heat losses.

The reader has undoubtedly noticed that transformers become heated. At least one-half of the heating is due to the effect of eddy currents.

Inductive Surge

Highly perfected methods of measuring magnetic fields can be devised by making use of electromagnetic induction. Thus far we proposed using a magnetic needle for this purpose, or a test loop with a known direct current. The magnetic induction was determined by the magnitude of the moment of force acting on the test loop or needle whose magnetic moment is equal to unity.

Now we proceed in a different manner. After connect-

ing a tiny current loop to a measuring instrument, we locate it perpendicular to the lines of force and then, with a rapid motion, turn the loop 90° As the loop is turned, a current is induced in it and a quite definite quantity Q of electricity, which can be measured, flows through the loop circuit. In what way is this quantity of electricity related to the strength of the field at the point where we located the test loop?

The necessary calculations are sufficiently simple. According to Ohm's law, the current I is the quotient of the induced emf by the resistance, i.e.

$$I = \frac{1}{R} \mathcal{E}^{\text{ind}}$$

If we make use of the expression for the law of electromagnetic induction $\mathcal{E}^{\text{ind}} = BA/\tau$ and recall that $Q = I\tau$, then the magnetic induction is

$$B = \frac{\mathcal{E}^{\text{ind}}\tau}{A} = \frac{I\tau R}{A} = \frac{QR}{A}$$

We repeat again that this formula is valid, of course, when in the final position the lines of force do not pass through the loop, and in the initial position they intersect the area of the loop at right angles. As a matter of fact, it does not matter whatsoever which we call the initial and which the final position. Only the direction of the current is changed, but not the quantity of electricity flowing in the loop.

The sensitivity of this method of measurement is increased n-fold if we use a coil instead of a single turn (loop). The quantity of electricity is proportional to the number of turns n. Good experimental physicists manage to wind coils only one millimetre in size, enabling them to examine a field in great detail by the inductive surge method.

Probably the most expedient application of this meth-

od is for measuring the magnetic permeability of iron bodies. We shall now discuss this important property of iron.

Magnetic Susceptibility of Iron

We found in the preceding chapter that atoms have magnetic properties. Single electrons have a magnetic moment, and orbital magnetic moments are developed by the movement of electrons about the nucleus. The nuclei of atoms have magnetic moments. Therefore, when we put a body in a magnetic field, this must affect the field in some way. The opposite is also true: the presence of a magnetic field affects the behaviour of solid, liquid and gaseous bodies to a more or less extent.

Iron has absolutely outstanding magnetic properties, as have certain of its alloys and some substances akin to iron. This small class of substances is said to be *ferromagnetic*. We can, for instance, conduct the following experiments. We suspend a small rod, about the size of a wooden match, free to turn, by a thread and bring a magnet near to it. Whatever other substances, besides iron, we make the rods from: wood, glass, plastics, copper, aluminium, etc., we cannot reveal the magnetic properties of these substances by bringing a magnet up to them. To prove that any substance has magnetic properties, it is necessary to perform precise, careful experiments, which are to be discussed below.

But iron bodies behave themselves in an entirely different manner. They move obediently to follow even the weakest bar magnet found in school physics laboratories.

To convince the reader of the sensitivity of iron bodies to the presence of a magnetic field, I wish to relate the following true story, instructive in every sense, of which I was the main character.

Several years ago I was asked to become acquainted with the experiments of a Czech "magician" He had won world fame and was called the "Czech Merlin" by American reporters who have a weakness for sensational news. The act of this wizard included several dozens of experiments that supposedly could not be explained in any rational way. The Czech Merlin attributed the results of these experiments to his psychic powers.

One of his leading items was to magnetize a wooden match. First he showed that the wooden match, suspended by a thread, is not deflected by a magnet. After this, he began to "hypnotize" the match, making certain mysterious passes with his hands. As an indispensable element of this performance, he brought the wooden match into contact with a metal idol which, as Merlin explained, was the receptor of his psychic energy.

After several weeks of work, I was able to show that all the experiments, without exception, of this Czech magician could be rationally explained by known facts of science. But how did he manage to magnetize the match? After making the passes and touching the idol he suspended the match again from the same thread. Now the match began to obediently follow a magnet he held and moved around. How could this be?

Merlin's "psychic power" had the following explanation. When he touched the metal idol, a negligible amount of fine iron dust was transferred to the end of the match. I demonstrated that one thirty-millionth of a gram of iron is sufficient to impart appreciable magnetic properties to the match. Here we have another case of "cockroach experiments".

This striking example demonstrates with sufficient clarity that, in the first place, we should not believe in "miracles" that contradict the laws of nature, and, secondly, and this is what interests us at the moment,

that the magnetic properties of iron are particularly extraordinary.

The classical experiment characterizing the magnetic properties of iron is performed in the following way. An electric circuit is wired that consists of two coils, one inside the other. The primary coil is connected to a storage battery circuit and the secondary coil is connected to an instrument that measures the quantity of electricity. If the primary circuit is closed, the magnetic flux passing through the secondary coil varies from zero to a certain limiting value Φ_0. This magnetic flux can be measured to great accuracy by the inductive surge method.

The magnetic properties of substances are investigated by means of the device just described. A rod is made of the substance and is inserted into the coils. The results of the two measurements, with and without the rod, are compared. If the rod is made of iron or some other ferromagnetic material, the quantity of electricity measured by the instrument is increased by several thousand times.

The ratio of the magnetic fluxes measured with and without a rod can be taken as an indication of the magnetic properties of the rod material. This ratio $\mu = \Phi/\Phi_0$ is called the *magnetic susceptibility* of the substance.

Thus, an iron body drastically increases the flux of lines of force. This can only have a single explanation: the iron body adds its intrinsic magnetic field to that set up by the electric current in the primary coil.

The difference $\Phi - \Phi_0$ is usually denoted by the letter J. Thus, $J = (\mu - 1)\,\Phi_0$ is the additional magnetic flux produced by the substance itself.

After we have completed the experiment for measuring the magnetic susceptibility and the rod is pulled out of the coils, we find that an iron rod retains its mag-

netization. It is less than J, but is still quite considerable.

The remaining, or remanent, magnetism of the iron rod can be eliminated. This is called *demagnetization* and is done by inserting the rod again into our experimental rig, but so that the intrinsic field of the metal and the magnetic field set up by the primary coil electric current are opposed. We can always select a primary current such that an inductive surge in the opposite direction eliminates the magnetic properties of the iron and returns it to the initial state. For historical reasons that we shall not go into, the strength of the demagnetizing field is called the *coercive force*, or *coercivity*.

This peculiar property of ferromagnetic materials, by means of which they retain magnetism in the absence of a current, and the possibility of eliminating this remanent magnetism by an electric current of the proper direction, is called *hysteresis*. What is the origin of this word? It comes from the Greek word *hysteros*, meaning behind or later. But what has this to do with the phenomenon we are discussing? One cannot know beforehand what the susceptibility μ of a definite piece of iron is equal to. It depends on previous events—whether the specimen was magnetized previously, and if so how strongly. In short, the magnetic permeability depends on the history of the specimen. If we plot a curve of the magnetization versus the magnetizing field strength and reverse the magnetizing field, we find that the two S-curves for the two directions of magnetizing field do not coincide, with the magnetization lagging behind, and thereby the curves form a loop, known as the *hysteresis loop*. This *lagging behind* gave the name to the hysteresis phenomenon.

Engineering requirements may specify ferromagnetic materials with various properties. In the magnetic alloy Permalloy, the magnetic susceptibility μ approaches

100 000; the maximum value for soft iron is only one-fourth as much.

The feasibility of increasing the flux of magnetic lines of force an immense number of times, by inserting an iron body into a wire coil, enables electromagnets to be produced. The capacity of an electromagnet, i.e. its capability of attracting and holding iron items of great weight, increases, of course, with the current passed through its winding. This process, however, is not unlimited; there is such a phenomenon as magnetic saturation, though it is not easy to reach saturation when we are concerned with heavy powerful magnets.

In recent years exceptionally strong magnetic fields are being set up by employing a superconductive winding. Engineers face formidable technical difficulties in working at extremely low temperatures. But in this range, we can be sure that we obtain from ferromagnetic materials all that they are capable of because μ drops with an increase in temperature.

In heating, the ferromagnetic properties disappear when we reach a certain temperature limit, for instance, 767 °C for iron and 360 °C for nickel. At this the magnetic permeability approaches unity as for all other bodies. This limit was found in 1895 by the French physicist Pierre Curie (1859-1906) and is called the *Curie point* of the particular substance.

Domains

The main feature of ferromagnetic materials is their domain structure. A *domain* is a region that has been magnetized to the limit. Inside the domain, all the atoms are aligned so that their magnetic moments are parallel to one another.

The behaviour of magnetic (ferromagnetic) domains is exactly the same as that of ferroelectric domains in fer-

roelectric materials. The linear dimensions of ferromagnetic domains are not especially small, namely, of the order of 0.01 mm. Therefore, using a simple contrivance, domains can be observed in an ordinary microscope.

To render visibility to domains, a drop of colloidal suspension, consisting of a carefully crushed ferromagnetic substance of the magnetite type, is applied to the polished surface of a ferromagnetic monocrystal. The colloidal particles concentrate along the boundaries of the domains because the magnetic fields are especially strong at these places (in the same way that ordinary magnets accumulate magnetic particles in the regions adjacent to their poles).

As in ferroelectric substances, the domains in ferromagnetic materials exist even when the material is unmagnetized and not only when an external magnetic field is applied.

The arrangement of the domains in an unmagnetized monocrystal is such that the total magnetic moment of the crystal equals zero. But this does not imply that the domains are located haphazardly. Again, as an exact analogy to what was mentioned on p. 64, the nature of the crystal structure dictates certain directions in which the magnetic moments are most easily aligned. Crystals of iron have a cubic unit cell and the axes of the cube are the directions of easiest magnetization. In other ferromagnetic metals, the moments are aligned along the diagonals of the cube. In any case, there is complete order in the arrangement of the domains in an unmagnetized crystal. There are just as many domains with magnetic moments pointing in one direction as there are with magnetic moments pointing in the opposite direction. Examples of domain structure have already been illustrated in Figure 2.5.

Magnetization, like polarization, consists in the

"devouring" of domains whose magnetic moments are oriented at an obtuse angle with the field.

The struggle between the tendencies toward order and toward disorder in atomic arrangement is a necessary feature of any state of matter. This has been discussed in detail in another book by the author, published by the same publishers. The book is called *Order and Disorder in the World of Atoms.*

As we found in Book 2 of this series, the tendency toward order is the tendency to a state of minimum energy. If there is not much thermal motion, the particles, left to themselves, form that marvel of atomic architecture, the crystal. The crystal is the symbol of ideal order in the world of atoms. The tendency toward disorder is dictated by the law of degradation of energy, or increase of entropy.

When the temperature is raised, entropy tendencies gain the upper hand and disorder becomes the prevalent form of existence of matter.

In dealing with ferromagnetic materials we have the following. As the temperature is raised, the magnetic moments begin to swing from side to side. First these oscillations keep in time without disturbing the established order. Then first one and another atom swivel into an "incorrect" position. The number of such atoms that drop out of the ordered ranks continually increases. Finally, at a strictly definite temperature (the Curie point) magnetic order completely falls apart.

It is difficult in a book of this size to explain why so few substances have ferromagnetic properties. What particular details in the structure of their atoms have put these substances in an exclusive class? I feel that the reader would be too demanding if he wanted to find answers to all questions in this small book for the layman.

Let us discuss the behaviour of other substances.

Diamagnetic and Paramagnetic Bodies

It has already been mentioned that, with the exception of ferromagnetic materials, all other substances have a magnetic permeability very close to unity. Substances for which μ is slightly higher than unity are said to be *paramagnetic*; substances for which it is less than unity are said to be *diamagnetic*. Examples of both classes of substances and their magnetic susceptibilities are listed in the following table:

	μ		μ
Aluminium	1.000 023	Silver	0.999 981
Tungsten	1.000 175	Copper	0.999 912
Platinum	1.000 253	Bismuth	0.999 824

Even though the values differ only slightly from unity, extremely precise measurements can be made. Generally speaking, we could use the inductive surge method with which we began our discussion on the magnetic measurements of the properties of substances. But the most exact values are obtained by means of a magnetic balance.

A hole is made in one of the pans of a microbalance (which, as is known, is capable of measuring forces with an accuracy within one ten-millionth of a gram). Suspended by a thread passing through this hole is the specimen, which hangs between the poles of a magnet. The pole shoes of the magnet are designed to set up a nonuniform field. Then the specimen is either attracted into or repulsed from the region with a strong field. It is pulled in when the magnetic moment of the specimen tends to align along the field; it is pushed out in the reverse case. The formula for the force is given on p. 116.

The specimen is counterbalanced by weights when there is no magnetic field. When the field is switched on, equilibrium is upset. If we are measuring paramagnetic substances, weights will have to be added to restore equilibrium. With diamagnetic substances, some of the weights must be removed. We can readily calculate that we can cope with our difficult task if we use a good balance because (in the easily achieved case with a field non-uniformity of the order of some hundredths of a tesla per centimetre) the force acting on 1 cm^3 of the substance will equal about one milligram.

Both kinds of properties, paramagnetic and diamagnetic, are sufficiently simply explained.

Diamagnetism is the direct result of the fact that in a magnetic field each electron of the substance describes a circle. These circular currents develop their own magnetic moments directed against the field that causes the rotation.

Diamagnetism is a property common to all substances.

Paramagnetism and, to an even greater extent, ferromagnetism "suppress" the diamagnetic properties of substances.

Paramagnetic materials are ones whose atoms or ions have a magnetic moment. This moment may be due to orbital motion of the electrons, to spin of a single electron or to both causes acting together.

Atoms of diamagnetic substances have no magnetic moment in the absence of a magnetic field. Atoms of paramagnetic substances have magnetic moments, but, due to thermal motion, they are arranged in complete disorder, exactly like those of ferromagnetic bodies above the Curie point. When a field is applied, a struggle begins between the ordering forces of the field and the disorder introduced by thermal motion. As the temperature drops, more and more atoms locate themselves so that their magnetic moment makes an acute angle with the direc-

tion of the field. This makes it quite clear why the magnetic susceptibility of paramagnetic bodies increases with a drop in temperature.

Earth's Magnetic Field

People today are accustomed to the fact that any instrument is devised on the basis of some physical theory. When the instrument has been developed, engineers concern themselves with it; physicists have no more to do. The nature of the phenomenon, on which the principle of the instrument is based, was well understood even before it had been developed.

Matters were entirely different in the case of the compass. It was probably developed in China in the 11th century and was used as the chief nautical instrument for hundreds of years before somebody really understood the principle of its operation. Why does one end of the needle always point to the north? Most wise men of that time thought that the behaviour of the needle was due to extraterrestrial forces, for example, attraction of the needle tip by the North Star (Polaris).

The brilliant work of William Gilbert (1540-1603) called *About the Magnet, Magnetic Bodies and a Great Magnet—the Earth* was published in 1600. A strictly scientific approach enabled this English physicist to come up close to an understanding of magnetic phenomena. Gilbert shaped a piece of magnetic iron ore, called a lodestone, into a sphere and thoroughly investigated the orientation of a magnetic needle suspended over various parts of the sphere. He found complete analogy with the orientation of a compass needle at various points on the earth. He drew the conclusion that the action of a compass can be excellently explained if we assume that the earth is a spherical permanent magnet whose axis is directed along the earth's axis.

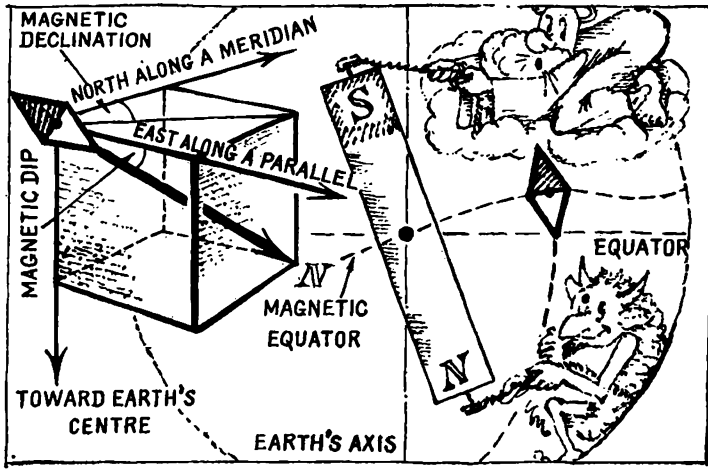


Figure 3.10

From this moment, the study of geomagnetism was raised to a new level. More precise investigations showed that a magnetic needle does not point exactly north. The deviation of a compass needle from the meridian passing through the given point is called the *magnetic declination*. The magnetic poles are displaced with respect to the earth's axis by 11.5° (Figure 3.10). The needle is not exactly in a horizontal plane but points downward (in the northern hemisphere) through an angle, varying at various latitudes, called the *magnetic dip*. After measuring the magnetic dip at various points on the earth, we come to the conclusion that the magnetic "dipole" is deep within the earth. It sets up a nonuniform field, the nonuniformity reaching 0.6×10^{-4} T at the magnetic poles and 0.3×10^{-4} T at the equator.

What is this "magnet" inside the earth? The magnetic "dipole" is in the earth's core, which consists mainly of

molten iron. Even in the molten state iron remains a good conductor of electricity. On this basis, a model constituting a sort of "magnetic dynamo" has been proposed to explain the earth's magnetic field. We shall not describe this model. It is sufficient to point out that the "terrestrial magnet" is the result of currents passing through the molten iron core.

The earth's magnetic field varies. The magnetic poles are continually being displaced at a rate of 5 or 6 km per year. This is a negligible displacement on the scale of the whole earth and is a phenomenon that can be observed only in the course of hundreds of years. This is why it has been called the secular variation of the earth's magnetic field.

It is needless to demonstrate how vitally important it is to have an exact knowledge of all the elements of terrestrial magnetism at any point on our planet. The magnetic compass still serves navigators. This being the case, they must be furnished with maps showing the magnetic declinations and dips. Near the poles, as is shown in Figure 3.10, the north end of a magnetic needle no longer points toward the north. Near the equator, it is also difficult to manage without a magnetic map. The magnetic equator does not coincide at all with the line of zero latitude (true equator).

A precise knowledge of the earth's magnetic field is also of immense interest on land, as well, because it can be of service in geological surveys. But we cannot dwell on these problems. Geological physics, or geophysics, is a significant and extensive field of science and deserves a special discussion.

We shall devote a few words to so-called paleomagnetic investigations that give us an idea of the earth's magnetic field in ancient and prehistoric times. These investigations are based mainly on a study of remanent (residual) magnetization of rock, etc.

The following is the essence of methods employed for the prehistoric period. Bricks and clay pottery have a small remanent magnetism that is developed in the hot clay when it is being fired. The direction of the magnetic moment coincides with that of the magnetic field at the time the item was fired and cooled. Sometimes it is possible to determine the position of the item during its manufacture.

Another example of similar investigations is finding the geographic direction of the magnetic moment of an ore, its age being determined by the amounts of radioactive isotopes.

Paleomagnetic investigations are the most rigorous proof of continental drift. It was found that the magnetism of iron ore deposits, formed several hundreds of millions of years ago on the various continents, can be directed along the lines of force of the earth's magnetic field if the continents are brought together into a single vast continent Panganea which split into two supercontinents Laurasia and Gondwanaland. Later these continents were further split up, dividing Gondwanaland, for instance, into Africa, Australia, Antarctica and South America, which then gradually drifted apart.

So far we have mentioned only the intraterrestrial origin of magnetism, and this is actually its main source. Certain changes occur, however, in the magnetic field of the earth due to charged particles arriving from space. These are chiefly streams of protons and electrons emitted by the sun. The charged particles are carried by the field to the magnetic poles and are rotated there in a circle by the Lorentz forces. This leads to two phenomena. In the first place, the moving charged particles set up a supplementary magnetic field consisting of magnetic storms. Secondly, they ionize the molecules of atmospheric gases, producing the aurora borealis, commonly called the northern lights (this is in the Northern Hemisphere:

the lights seen in the Southern Hemisphere are called the aurora australis). Severe magnetic storms occur periodically (after 11.5 years). This period coincides with the periods of intensive solar activity.

Direct measurements by means of spacecraft indicate that the bodies nearest to the earth—the moon, and the planets of Venus and Mars—do not have their own magnetic field similar to that of the earth. Of the other planets of the solar system, only Jupiter and, evidently, Saturn have their own magnetic fields. A field of a strength up to 10 Gs and a number of typical phenomena (magnetic storms, synchrotron radio-frequency radiation, etc.) were discovered on Jupiter.

Magnetic Fields of the Stars

Magnetism is found, not only on planets and extinct stars, but on incandescent heavenly bodies as well. Since the sun is our nearest star, we know more about its magnetic field than that of other stars. The magnetic field of the sun can be visually observed during solar eclipses. Particles of solar material that have a magnetic moment align themselves along the lines of force, thereby sketching a picture of these lines. The magnetic poles are clearly seen and the strength of the magnetic field can be estimated. In regions of a size of the order of tens of thousands of kilometres, this field is of a strength a thousand times greater than that of the earth's field. These regions are called *sunspots*. Since the spots are darker than the rest of the sun's surface, the temperature here must be lower, namely, 2000 degrees below the "normal" temperature of the sun.

Without doubt, the lower temperature and the stronger magnetic field are related in some way. But no proper theory relating these two facts has yet been proposed.

What about the other stars? The advances in astro-

physics have been so amazing in recent years that it is now feasible to establish the existence of magnetic fields on the stars. It was found that these "stellar magnetic spots" have a temperature of about 10 000 °C and can change their position or even disappear entirely in the course of several months. It is simpler to explain these changes if we assume that the whole star rotates rather than have spots change their location.

The presence of magnetic fields is indicated by the anomalous intensity of certain spectral lines. It would seem that magnetic stars have an increased iron content on their magnetic equator.

Magnetic fields in space are very weak (a millionth of a gauss). This needs no explanation because an extremely high vacuum reigns in space. When stars are formed of atoms scattered throughout the universe, the condensing of the stellar material is accompanied by a "condensing" of the magnetic field. Why, then, do not all stars have a magnetic field?

The earth has existed for thousands of millions of years. It follows that the magnetic field of the earth is continually maintained by the electric currents flowing in its depths. Certain stars, having no magnetic field, have evidently cooled to an extent that they no longer have electric currents inside them. It is improbable, however, that this explanation is a universal one.

4. Summary of Electrical Engineering

Sinusoidal Emf

Storage batteries and dry batteries are sources of direct current. Electric power mains, on the other hand, provide us with alternating current. The words "direct" and "alternating" refer to the values of the voltage, emf and current. If these quantities remain unchanged as the current passes through a circuit, the current is said to be *constant*, or *direct (d-c)*; if they vary, the current is said to be *alternating (a-c)*.

An electric current may vary in different ways with time in accordance with the device producing the current. The curve described by the change in electric current can be obtained by means of a cathode-, or electron-beam, tube. The electron ray is deflected by the fields of two mutually perpendicular parallel-plate capacitors. Applying various voltages across the capacitor plates, we can make the luminous spot produced by the ray wander all over the screen.

To obtain a picture of an alternating current, we proceed as follows. A so-called sawtooth voltage is applied across one pair of plates. The curve of this voltage is illustrated in Figure 4.1. If the electron beam is subject only to its action, the spot travels uniformly across the fluorescent screen and then returns with a jump to its starting point. The position of the spot provides information on the instant of time. When the alternating voltage being investigated is applied across the second pair of plates, the spot is given an up-and-down motion as it sweeps across the screen, producing a waving line

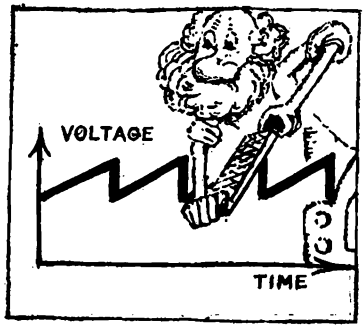


Figure 4.1

in the same manner that mechanical vibrations could be made visible by a simple device as shown in Book 1.

It was no slip of the tongue when I mentioned "vibrations" (or "oscillations"). Most of the quantities characterizing an alternating current oscillate according to the same harmonic or sine-curve law that a pendulum complies with when deviated from its equilibrium position. This can be readily shown by connecting a-c city mains to an oscillograph.

Either the current or the voltage may be plotted along the vertical direction. The characteristics of the current are the same as parameters of mechanical vibrations. The interval of time after which the picture of current change is repeated is called the period T. Current frequency ν, the reciprocal of the period, is usually equal to 50 or 60 cycles per second.

When we examine a single sine curve (as the curve obtained for alternating voltage or current is called), the selection of the initial instant of time is of no consequence. But if two sine curves are superimposed, as illustrated in Figure 4.2, we must specify the fraction of a cycle that they are shifted in phase. The *phase* is expressed by the angle $\varphi = 2\pi t/T$. Thus, if the curves are

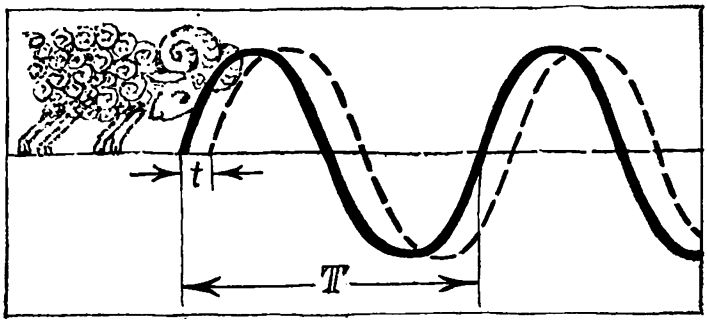


Figure 4.2

displaced with respect to each other by one-fourth of a period, we say they have a phase shift of 90°; if by one-eighth of a period, the phase shift is 45°, etc.

When we are discussing several sine curves, shifted in phase, engineers speak of current or voltage vectors. The length of the vectors represents the amplitude of the sine curve, and the angle between them represents the phase shift. Many engineering devices provide a current whose curve is the sum of several sine curves displaced with respect to one another, rather than a simple sinusoidal current.

We shall show that a simple sinusoidal current is produced when the conducting loop rotates at constant speed in a uniform magnetic field.

Upon an arbitrary direction of the rectangular loop with respect to the lines of force, the magnetic flux passing through the loop equals

$$\Phi = \Phi_{max} \sin \varphi$$

where φ is the angle between the plane of the loop and the direction of the field. This angle varies with the time according to the law $\varphi = 2\pi t/T$.

The law of electromagnetic induction enables us to calculate the induced emf. Let us write the equations of the magnetic fluxes for two instants, differing by the extremely small time interval τ:

$$\Phi = \Phi_{max} \sin \frac{2\pi}{T} t \text{ and } \Phi = \Phi_{max} \sin \frac{2\pi}{T} (t + \tau).$$

The difference between these two equations is

$$2\Phi_{max} \cos \frac{2\pi}{T} \left(t + \frac{\tau}{2}\right) \sin \left(\frac{2\pi}{T} \frac{\tau}{2}\right)$$

Since τ is very small, the following approximate equations are valid:

$$\sin \left(\frac{2\pi}{T} \frac{\tau}{2}\right) \cong \frac{2\pi}{T} \frac{\tau}{2} \quad \text{and} \quad \cos \frac{2\pi}{T} \left(t + \frac{\tau}{2}\right) \cong \cos \frac{2\pi}{T} t$$

The induced emf is equal to this difference referred to the time. Then

$$\mathscr{E}^{ind} = \frac{2\pi}{T} \Phi_{max} \cos \frac{2\pi}{T} t = \frac{2\pi}{T} \Phi_{max} \sin \left(\frac{2\pi}{T} t - \frac{\pi}{2}\right)$$

We have shown that the induced emf is expressed by a sine curve shifted in phase with respect to magnetic flux sine curve by 90°. As for the maximum value of induced emf, its amplitude, it is proportional to the product of the magnetic flux amplitude by the frequency of rectangular loop rotation.

The law for the current is obtained by dividing the induced emf by the resistance of the circuit. But we shall make a crude error if we equate the resistance to alternating current, in the denominator of the equation

$$I_{a\text{-}c} = \frac{\mathscr{E}^{ind}}{R_{a\text{-}c}}$$

to the ohmic resistance, i.e. the quantity we have dealt with previously. It turns out that $R_{a\text{-}c}$ is determined not only by the ohmic resistance but also by two more param-

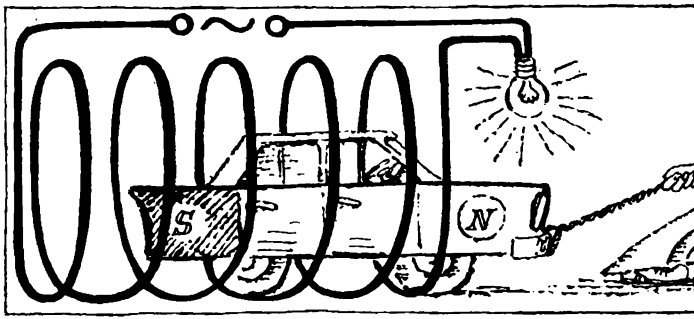


Figure 4.3

eters of the circuit: its inductance and the capacitances connected into the circuit.

The fact that Ohm's law is more complicated for alternating current than for direct current is demonstrated by the following simple experiment. Illustrated in Figure 4.3 is an electric circuit with the current passing through an electric lamp and a coil into which an iron core can be inserted. First we connect the lamp circuit to a d-c source. Next we move the core into the coil and pull it out. No effect whatsoever! The resistance of the circuit remains constant and so does the current. We repeat the experiment for the case when the circuit is connected to an a-c source. The result is spectacular, is it not? Now the lamp burns brightly when the core is not inserted into the coil, and gradually dims as it is inserted.

Thus, at a constant external voltage, and a constant ohmic resistance (depending only on the material, length and cross section of the wire), the current varies with the position of the iron core in the coil.

What does this mean?

Recall that an iron core drastically increases the magnetic flux passing through the coil (by thousands of times). For an alternating current, the flux varies con-

tinually. But if it varied without the iron core from zero to some arbitrary unit, with the core it varies from zero to several thousand units.

As the magnetic flux varies, the lines of force thread through the turns of their "own" coil. This produces a current of self-induction in the coil. According to Lenz's law, the induced current is in the direction that weakens the effect that causes it: the external emf meets with an obstacle that did not exist when the current was direct. In other words, an alternating current is subject to an additional resistance because the magnetic field, threading through the wires of its own circuit, produces a back electromotive force, called the *emf of self-induction*, which weakens the average current. This additional resistance is called *inductive reactance*, to distinguish it from a true (ohmic) resistance.

Experiments indicate (and this will seem quite natural to the reader) that the magnetic flux threading through the coil (or, more generally speaking, the whole current-carrying circuit) is proportional to the current: $\Phi = LI$. As for the proportionality factor L, called the *inductance*, it depends upon the geometry of the circuit or coil and what kind of cores it encompasses. As is obvious from the formula, the numerical value of the inductance is equal to the magnetic flux when the current equals one ampere. The unit of measurement of L is the *henry* (1 H = 1 ohm-s).

We can theoretically derive and prove by experiments that the inductive reactance R_L is expressed by the formula

$$R_L = 2\pi\nu L$$

If the ohmic resistance (with which we are already acquainted) and the capacitive reactance (which we shall discuss below) are small, the current in an a-c circuit equals

$$I = \frac{\mathscr{E}}{R_L}$$

To be able to judge what is meant by "small" or "large", we can calculate the inductive reactance for the frequency of ordinary city power supply current and an inductance of 0.1 H. We obtain about 30 ohms.

Now, what would a coil with an inductance of one henry be like? The following formula, which we give without derivation, is used to assess the inductance of coils and chokes (coils with an iron core):

$$L = \mu_0 \mu \frac{n^2}{l} A \quad \text{and} \quad \mu_0 = 4\pi \times 10^{-7} \frac{\text{joule}}{\text{ampere}^2\text{-metre}}$$

where n = number of turns; l = length of the coil; and A = cross-sectional area. Thus, an inductance of 0.002 H can be obtained, for instance, with a coil having the following parameters: $l = 15$ cm, $n = 1500$ and $A = 1$ cm^2. If an iron core with $\mu = 1000$ is inserted into this coil, its inductance equals 2 henries.

An emf of any origin and, consequently, an emf of self-induction, performs work. This work, as we know, is equal to $\mathscr{E}I$. If the current is alternating, the values of $\mathscr{E}$ and I vary each instant. Let their values equal $\mathscr{E}_1$ and I_1 at the instant t and $\mathscr{E}_2$ and I_2 at the instant $(t+\tau)$. The magnetic flux threading the turns of a coil of inductance L is equal to LI. At the instant t, its value is LI_1 and at the instant $(t+\tau)$, it is LI_2. What amount of work is required to increase the current from I_1 to I_2? The emf is equal to the change in magnetic flux divided by the time during which the change took place:

$$\mathscr{E} = \frac{L(I_2 - I_1)}{\tau}$$

This equation is to be multiplied by the time and the current, i.e. $\mathscr{E}I\tau$, to obtain the amount of work. But by

what current? The average value: $(I_1+I_2)/2$. Hence, the work of the emf of self-induction equals

$$\frac{L}{2}(I_2+I_1)(I_2-I_1)=\frac{L}{2}I_2^2-\frac{L}{2}I_1^2$$

This arithmetical result can be expressed in words: the work performed by the emf equals the difference in the values of $LI^2/2$ at two instants of time. This means that energy is not dissipated by an inductive reactance; it is not converted into heat as in circuits with ohmic resistance, but is transferred into the "reserve". For this reason, we can rightly call the quantity $LI^2/2$, the *magnetic energy of the current.*

Now let us discuss the effect of introducing a capacitor into an alternating current circuit.

If we include a capacitor in a d-c circuit, there is no current. Introducing a capacitor is the same thing as breaking the circuit. But the current is not interrupted by a capacitor in an a-c circuit.

The difference is what interests us; what is it? The explanation is simple enough. When the circuit is connected to an a-c power source, an electric charge begins to accumulate on the plates of the capacitor. The positive charge goes to one plate and negative to the other. Assume that the inductive reactance and ohmic resistance are low. Charging continues until the voltage across the capacitor plates reaches a maximum and is equal to the emf of the power source. At this instant, the current equals zero. Then the voltage of the source begins to drop and the capacitor is "discharged".

If we measure the current with some instrument in a circuit containing a capacitor, we shall find that the current varies in accordance with two quantities. In the first place, it can be proved (both by experiments and by theoretical reasoning) that the current decreases with the frequency. This means that the *capacitive reactance*

is inversely proportional to the frequency. This result is quite natural because the lower the frequency, the more alternating current, so to speak, approximates direct current.

If we change the geometrical parameters of the capacitor, i.e. the distance between the plates and their area, we shall find that the capacitive reactance is also inversely proportional to the capacitance of the capacitor.

The formula for capacitive reactance is of the form:

$$R_C = \frac{1}{2\pi\nu C}$$

At the frequency of city power supply current, i.e. 50 cps (cycles per second), a capacitor with a capacitance of 30 microfarads constitutes a capacitive reactance of about 100 ohms.

I do not intend to relate how the total resistance (called the *impedance*) is calculated in complex circuits consisting of ohmic resistances, and inductive and capacitive reactances. I only wish to warn the reader that the impedance is not the sum of the separate resistances and reactances.

The electric current in a part of a circuit that includes an ohmic resistance, a capacitor and an inductive coil, and the voltage across this part can be measured in the ordinary way by means of an oscilloscope (electron-beam tube). We see both the current and the voltage on the screen in the form of sine curves. It should not surprise us to find that these sine curves are shifted with respect to each other by a certain phase angle φ. (The reader readily comprehends that this must be so if he recalls that, for instance, the current equals zero in a circuit with a capacitor when the voltage across the capacitor reaches its maximum value.)

The value of the phase shift φ is of prime importance.

The power transmitted equals the product of the current by the voltage. If the current and voltage sine curves coincide, the power has its maximum value. But if they are shifted as they would be in a circuit including only capacitive or only inductive reactance, the power equals zero. This can easily be demonstrated by plotting two sine curves with a phase shift of 90°, multiplying their ordinates together at each point and adding together these products over a single period. It can be strictly demonstrated that the average power per period of alternating current equals

$$P = IV \cos \varphi$$

One of the main concerns of the electrical engineer is to increase $\cos \varphi$ (called the *power factor*).

Transformers

You have just purchased an electric refrigerator. The salesperson warned you that the refrigerator is designed for a mains voltage of 220 V. But in your home the mains voltage is 127 V. Is this a hopeless situation? Not at all. You must simply spend a little more and buy a suitable transformer.

A *transformer* is a very simple device enabling you to either raise or lower the voltage. It consists of a laminated iron core on which two coils have been wound. Each coil, or winding, has a different number of turns.

When we plug one of the windings into a socket with the mains voltage, we find with a voltmeter that the voltage across the ends of the second winding differs from that of the power mains. If the primary winding has w_1 turns and the secondary, w_2 turns, the ratio of the voltages is

$$\frac{V_1}{V_2} = \frac{w_1}{w_2}$$

Thus, a transformer steps up the voltage when the primary voltage is connected to the winding with the smaller number of turns, and steps down the voltage in the opposite case.

Why is this so? The fact is that practically the whole magnetic flux is within the iron core. This means that both windings are linked by an equal number of lines of force. A transformer can operate only with an alternating voltage across the primary winding. A sinusoidal current variation in the primary winding induces a sinusoidal emf in the secondary winding. Each turn of the primary and secondary windings are in identical conditions. The emf per turn of the primary winding is equal to the emf of the power mains divided by the number of turns of the primary winding, i.e. V_1/w_1, and the emf of the secondary winding is equal to the product of V_1/w_1 by the number w_2 of turns of the secondary winding.

In principle, each transformer can be employed either as a step-up or a step-down device depending upon the winding to which the primary voltage is applied.

One frequently deals with transformers in everyday life (Figure 4.4). Besides the transformers we employ willy-nilly because some electric appliance is designed for one voltage and the city mains for another, we may also have to deal with the spark coil of an automobile. The spark coil is a step-up transformer. To produce the sparks that ignite the fuel mixture in the cylinders, we require a high voltage. It is obtained from the storage battery of the automobile, after converting its direct current into alternating current by means of an interrupter (contact-breaker) and then stepping up the voltage with a spark coil.

It can readily be seen that when the voltage is stepped up, the current is reduced and vice versa, with an accuracy within the energy losses due to the heating of the transformer.

Figure 4.4

Welding apparatus requires step-down transformers. An exceptionally heavy current is required for welding

operations, and the transformer has only a single output turn.

You have probably noted that the core of a transformer is made of thin sheets of steel. This is done to prevent energy losses in converting voltages. As mentioned previously, eddy currents have a considerably weaker effect on sheet stock than on solid material.

In the home you deal with small low-power transformers. Powerful transformers are huge structures. In them the core with its windings is inserted into a tank filled with cooling oil.

Machines that Produce Electric Current

Machines that convert mechanical motion into electric current were first developed only some hundred and fifty years ago.

The first generator of electric current was the machine built by Michael Faraday and consisted of a rectangular loop of wire rotating in a field set up by permanent magnets. Soon somebody (but not Faraday) got the idea of replacing the single loop with a whole coil, thereby adding all the emf induced in all the turns. Only in 1851 were the permanent magnets replaced by electromagnets, i.e. coils wound on an iron core. Originated with these devices was the expression "exciting the machine" because it was necessary first to "vitalize" the electromagnet before an electric current could be produced. The early generators were separately excited from some outside power source.

The next stage in generator development was the discovery of the principle of self-excitation, which eliminated the necessity of a separate power supply to excite the electromagnets. It proved sufficient to connect the excitation, or field, winding of the generator in some manner to the main winding. By the end of the eighties

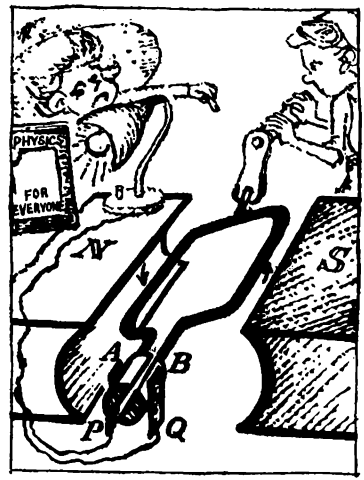
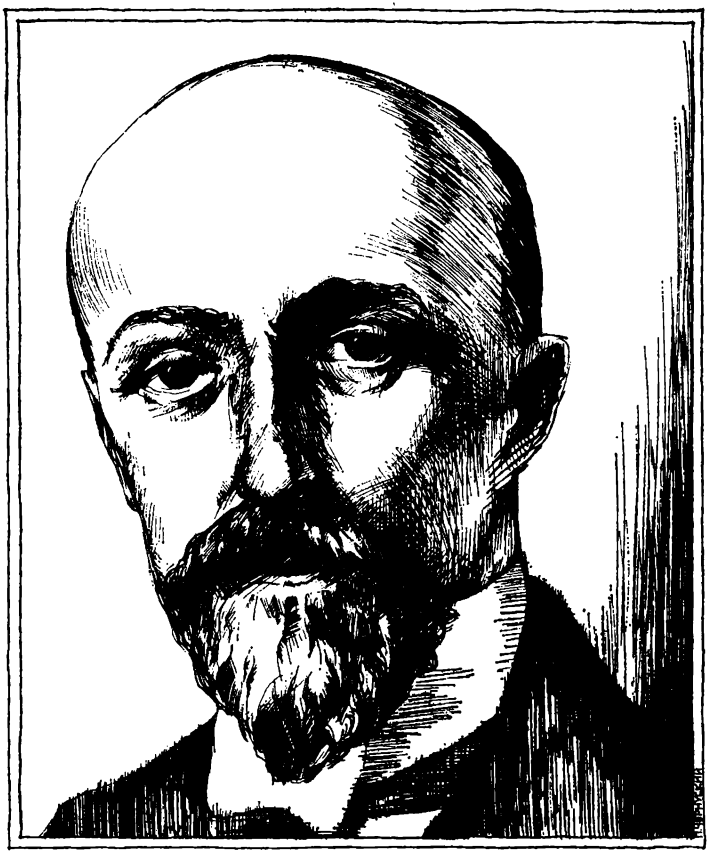


Figure 4.5

of last century, electric machines had already acquired the principal features they have today. The simplest model of a d-c generator is illustrated in Figure 4.5. If the loop is rotated in a field of permanent magnets, a sinusoidal emf is induced in it.

If you wish to obtain a direct current from the alternating current in the loop, it will be necessary to equip the generator with a special device called a split-ring commutator. This commutator consists of two half-rings, A and B, insulated from each other and mounted on a common cylinder (see Figure 4.5). This cylinder rotates together with the rectangular loop. Applied to the half-rings are contacts P and Q (called brushes), by means of which the current is conducted to the external circuit. Upon each half-revolution of the loop, its commutator half-rings pass from one brush to the other. Hence, notwithstanding the reversal of the current in the loop itself, the current in the external circuit has only a single



Mikhail Osipovich Dolivo-Dobrovolsky (1862-1919)—distinguished Russian scientist and engineer; he developed the three-phase current systems, which are the foundation of all modern electrical engineering. He worked out all the elements of three-phase circuits with alternating currents. In 1888 he built the first three-phase a-c generator with a rotating magnetic field.

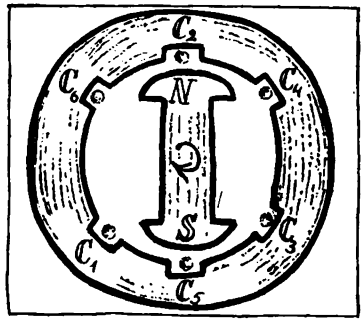


Figure 4.6

direction. Since the rotating component (rotor) of a real machine consists of a great number of loops, or sections, displaced through a definite angle with respect to one another, and the commutator consists of the corresponding number of segments, we obtain a practically constant emf.

D-c generators are being built today with a power rating from fractions of a kilowatt to several thousand kilowatts. Powerful generators are employed for the power supply in electrolysis in the chemical industry and in nonferrous metallurgy (aluminium and zinc production). They are designed for a large current and a relatively low voltage (120 to 200 V and 1000 to 20 000 A). D-c generators are likewise used for electric welding.

But d-c generators are not the main producers of electric power. Alternating current with a frequency of 50 Hz (or cps) is applied in the USSR for the production and distribution of electric energy. An a-c generator, called an alternator, is designed so that it simultaneously produces three emf's of the same frequency, differing in phase by the angle $2\pi/3$.

Such a three-phase generator is shown schematically in Figure 4.6. In the drawing each coil is represented by

a single turn. The wire of one turn is denoted by C_1-C_4, the second, C_2-C_5 and the third, C_3-C_6. If the current enters C_1, it returns through C_4, etc. (Of course, at instants corresponding to various relative positions of the rotor and stator, any of these leads, or wires, may be points of entry or exit of the current.) The emf in the stationary turns of the stator is induced because they cut across the magnetic field of the rotating electromagnet, the rotor. As the rotor rotates at constant speed, a periodically varying emf is induced in the windings of each phase of the stator. The three induced emf's are of the same frequency and differ from one another in phase by the angle 120° as a result of their spatial displacement.

The three turns of the winding can be either *star* or *delta connected.* These circuits were worked out and put into operation in the nineties of last century by Mikhail Osipovich Dolivo-Dobrovolsky (1862-1919). In a star connection, the ends of all three windings of the generator stator, C_4, C_5 and C_6, are connected together at a single point, which is called the neutral, or zero, point. Four leads connect the generator to the consumer of the energy: three line wires, from the starts of the windings, C_1, C_2 and C_3, and a neutral wire from the neutral point of the generator. This is called a *four-wire system.*

The voltage between the neutral point and the start of a phase is called the *phase voltage.* The voltage between the starts of the windings is called the *line voltage.* These voltages are related by the equation

$$V_l = \sqrt{3}\, V_{ph}$$

If the loads (*I*, *II* and *III*) in the three phases are the same, the current in the neutral wire equals zero. In this case, the neutral wire can be done away with, using what is called a *three-wire system.* Diagrams of star connections are illustrated in Figure 4.7.

A delta connection also allows a three-wire system to be

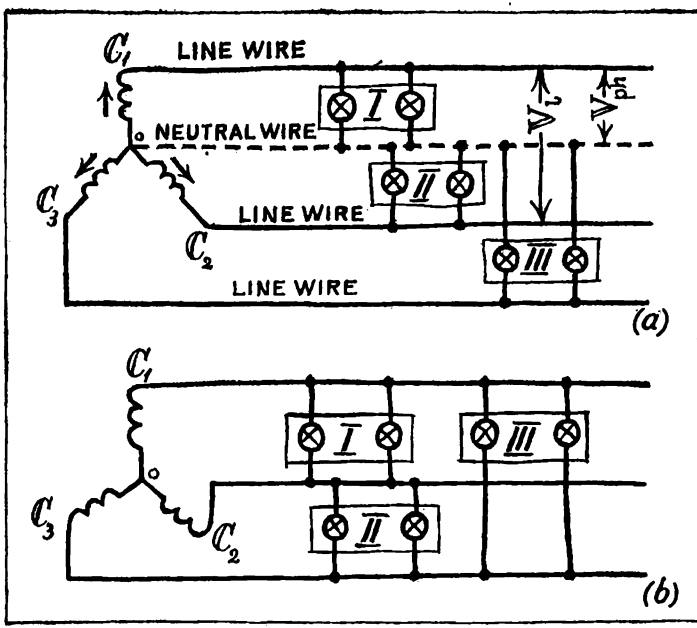


Figure 4.7

employed. Here the end of each winding is connected to the start of the next one so that the windings form a closed triangle. The line wires are connected to the vertices of this triangle. Here the line voltage equals the phase voltage and the currents are related by the equation

$$I_l = \sqrt{3}\, I_{ph}$$

Three-phase circuits have the following advantages: more economical energy transmission than in single-phase circuits, and the feasibility of obtaining two voltages—phase and line—from a single installation.

The a-c generator, or alternator, described above belongs to the class of synchronous machines. These are machines in which the speed of rotor rotation coincides with the speed of rotation of the magnetic field set up by the stator.

Synchronous generators are the main producers of electric energy. They exist in several design versions, depending on the method of driving the rotor.

The reader may ask: If certain machines are said to be synchronous, are there asynchronous machines as well? Yes, there are! But they are employed as motors and we shall discuss them in the next section. There we shall also consider why the magnetic field rotates in a three-phase a-c machine.

Electric Motors

More than half of all the electric energy produced is converted by electric motors into mechanical energy to satisfy the needs of industry, agriculture, transportation and the household. The most extensively used is the simple, reliable, inexpensive, easily maintained asynchronous, or induction, electric motor, invented in 1889 by the same talented Russian engineer Dolivo-Dobrovolsky. Its main features have been retained up to the present time. Motors of this type are used for the drives of various machine tools, pump and compressor installations, forging and pressworking machinery, materials handling and transportation equipment and other mechanisms.

The prototype of the induction motor was the model devised by the French physicist Dominique François Jean Arago (1786-1853). In 1824, Arago demonstrated before the Paris Academy of Sciences a phenomenon which he called "magnetism of rotation". He showed that a magnetic needle suspended over a rotating copper

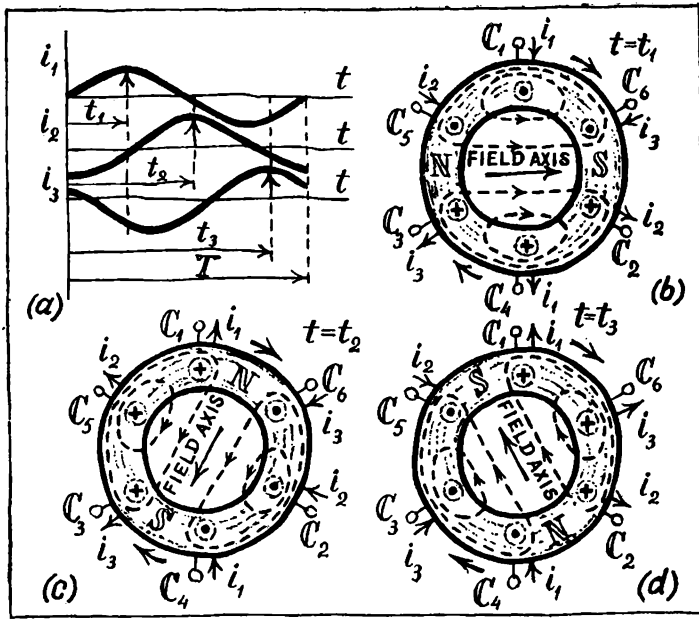


Figure 4.8

disk rotates with the disk. The inverse is also true: a disk mounted on a pivot rotates if it is in the field of a rotating permanent magnet. This idea was brilliantly utilized by Dolivo-Dobrovolsky, who combined it with the features of a three-phase system of currents. This enabled rotation of the magnetic field to be obtained without using any additional devices.

Let us examine the schematic diagrams in Figure 4.8. For the sake of extreme simplification only three turns are shown (actually, of course, electric machines have windings with a great number of turns). The cross and black dot indicate whether the current flows away from us

or toward us in each turn at a definite instant of time. The three turns make angles of 120° with one another. Figure 4.8*a* shows the phase relations of the three currents i_1, i_2 and i_3 flowing in the turns. Of interest to us is the resultant magnetic field set up by the three "coils". Figure 4.8*b* shows the lines of force of the resultant field for instant t_1 (current flowing into C_2, C_3 and C_4). Similar diagrams are shown in Figure 4.8*c* and *d* for instants t_2 and t_3. Hence, as we can see, the field we are interested in rotates (note the positions of the crosses), and rotates in the full sense of the word! The axis of the field in the middle of the system is aligned along the axis of the turn (phase) in which the current is maximum at the given instant of time.

The diagrams we have just discussed give an idea of how a three-phase a-c winding is arranged in the stator of a three-phase induction (asynchronous) electric motor. The rotor (Figure 4.9), set into motion by the rotating magnetic field, is short-circuited, i.e. we see neither the start nor the end of the winding. Such a rotor, also called an armature, resembles a squirrel cage. It consists of copper bars joining two copper rings. Compare it with a d-c electric machine. How much simpler this is! We need only to connect a three-phase a-c power supply to the stator. This sets up a rotating magnetic field in the machine. The magnetic lines of force of this field link the rods of the rotor, inducing currents in the rods and giving the motors their common name. As a result of the interaction between the rods, along which currents flow, and the magnetic field, the rotor begins to rotate at a speed near to that of the field, but does not reach this speed. This is just as needs be as otherwise the rods of the rotor would not cut the magnetic lines of force of the rotating field of the stator, and the rotor would stop. This is why such machines are said to be asynchronous. The lag of the rotor is called slip.

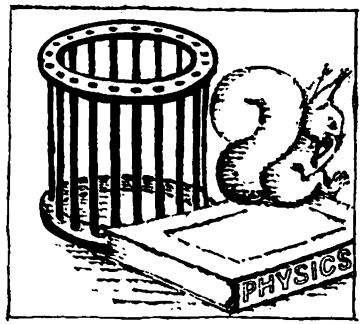


Figure 4.9

Induction electric motors are available in a large power range: from fractions of a watt to hundreds of kilowatts. There are even more powerful induction motors: up to 6000 kW, operating on a voltage of 6000 V.

Asynchronous micromachines are used in automatic control systems as actuating mechanisms, or power units, to convert the input electric signal into mechanical motion of a shaft. They can also be used as tachogenerators for converting rotation into an electric signal.

The synchronous machines previously considered and d-c machines can also be electric motors. This follows from the obvious principle of convertibility of electric machines: any electric machine can operate either as a generator or as a motor.

For example, the Kiev hydraulic power system on the Dnieper River includes a hydraulic accumulator station equipped with convertible units that can operate either as pumps or as turbines. When there is excess electric energy in the power system, the units pump water into the storage basin. At this time, the synchronous machines coupled to the units operate as drive motors. Upon peak consumers' load, the units are operated by the stored water as turbines and generators.

Synchronous motors are used in metallurgical plants, mines and refrigerators to power pumps, compressors, fans and other mechanisms that operate at constant speed. Miniature synchronous motors are extensively applied in automatic control systems. They have a rating from fractions of a watt to several hundred watts. Since the speed of these motors is rigidly coupled to the frequency of the power supply mains, they are most expediently employed where a constant speed of rotation is required. Such applications include: electric clock mechanisms, tape-moving mechanisms of recording instruments, film-advancing devices of motion-picture projectors and cameras, radio apparatus, programming devices, as well as systems of synchronous communication, where the speed of rotation of the mechanisms is controlled by varying the frequency of the supply voltage.

In the principle of its design, a d-c motor in no way differs from a d-c generator. The machine has a stationary system of poles whose excitation (field) winding is connected in one or the other way with the armature winding (either in series or in parallel). The machine can also be separately excited from a power supply. The armature has a winding properly distributed in its slots and connected to a source of direct current. D-c motors like d-c generators have a commutator whose purpose is to "straighten out" the torque, i.e. to make the machine, if it is a motor, rotate continuously in one direction.

Series-wound d-c electric motors (with the field winding in series) are especially suitable for electric traction, cranes and hoisting machinery. Such installations require the speed of the drive motor to drop drastically at heavy loads and its torque to increase substantially. Series-wound d-c motors are distinguished particularly for these features.

The first experiments with electric traction in Russia were conducted by Fyodor Apollonovich Pirotsky (1845-

1898). As far back as 1876, he adapted ordinary railway tracks for transmitting electric energy, and in 1880 he started to run an electric streetcar on an experimental line of the horsecar railway near Rozhdestvensky Park in St. Petersburg. As cars for the first streetcar line, he used the double-decker horsecars, mounting an electric motor under the body.

The first electric trolley-car line in Russia was opened in Kiev for the general public in 1892. Its electric motor was supplied from an overhead contact wire. The municipal construction committee agreed to build the streetcar line only after they had been convinced by calculations of the technical advantages of electric traction over horse traction on the steep hilly streets of Kiev. These streets proved to be beyond the capacity of either horse or steam traction.

The first experiments in "electric navigation" were carried out in 1838 by Boris Semyonovich Yakobi (1801-1874). He demonstrated an electric boat, holding fourteen passengers, on the Neva River. It was powered by a 550-watt electric motor. Yakobi supplied the motor from 320 galvanic batteries. This was the first time in history that an electric motor was used for traction.

The term "turbo-electric ship" appears more and more often in the press in recent years. It simply means that steam is used to drive powerful d-c turbogenerators (turbodynamos) and that the screws are mounted on the shafts of electric motors. But isn't this too complicated? Why not mount the screws directly on the turbine shafts?

The point is that a steam turbine generates maximum power only at strictly definite speeds. Powerful turbines run at 3000 rpm. If the turbine is slowed down, the generated power drops. If screws were mounted directly on the turbine shafts, a ship powered in this way would have poor seafaring properties. A d-c electric motor, on the other hand, has an excellent traction performance

curve: the higher the forces of resistance, the higher the tractive effort it develops. Moreover, such a motor can develop higher power at low speed at the moment the ship is just getting under way.

Thus a d-c generator and a d-c motor, arranged between the turbine and screw of a turbo-electric ship, operate like an automatic stepless (infinitely variable) gearbox, developed to a high state of perfection. It may seem that such a system is somewhat bulky, but at the high power ratings of modern turbo-electric ships, any other would take up just as much space and would be less reliable.

The power installation of a turbo-electric ship can be substantially improved in an entirely different way. It proves highly efficient to replace cumbersome steam boilers by a nuclear-power reactor. There is a huge saving in the volume of the fuel required for each voyage. The well-known Lenin nuclear-power icebreaker of the Soviet arctic fleet is the first of its kind. The nuclear-power installation of this turbo-electric ship provides for voyages as long as a year without refueling.

D-c electric motors are installed on main line electric locomotives, suburban electric trains, streetcars and trolley buses. They are supplied by energy from stationary electric power stations. In the USSR, electric traction is powered by either direct current or by single-phase alternating current with a commercial frequency of 50 Hz. Silicon rectifiers are being widely used in the traction substations of streetcar, trolley bus and subway lines. As to electric railways, the current may be rectified either in substations or in the electric trains themselves.

5. Electromagnetic Fields

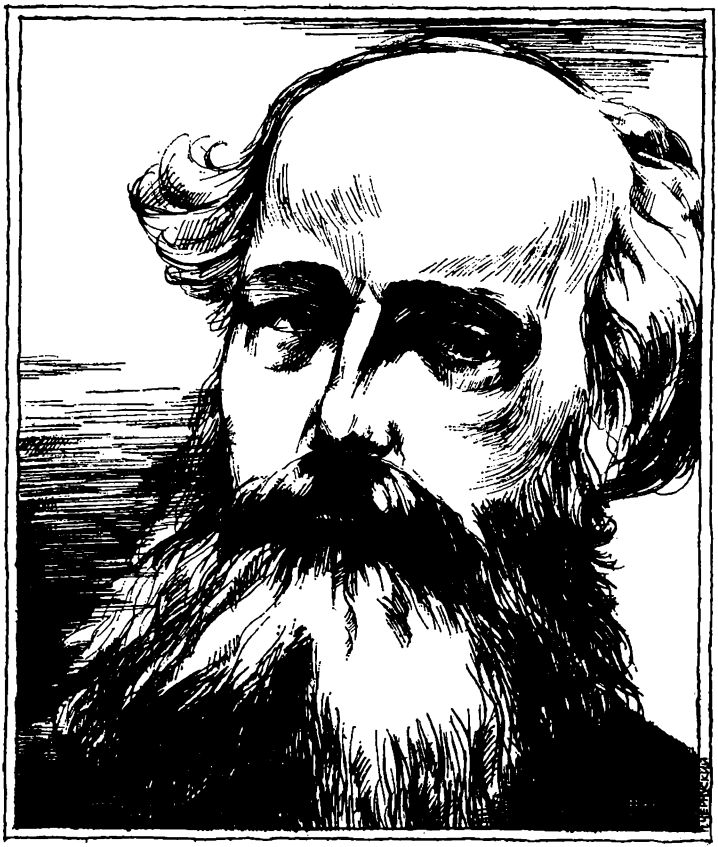
Maxwell's Equations

By the fifties of last century, much information had accumulated on electricity and magnetism. It was disconnected in the main, however, sometimes contradictory and, in any case, did not fit into a single orderly system.

But a great deal was known. In the first place, physicists knew that electric charges at rest set up an electric field; secondly, that electric currents set up magnetic fields; and, thirdly, the results of Faraday's experiments were published and generally recognized. In them he proved that a variable magnetic field generates an electric current.

In those times, a number of scientists, and primarily Faraday, doubtlessly became convinced that some kind of events occur in the space surrounding electric currents and charges. This group of investigators supposed that magnetic and electric forces are transmitted from point to point. Attempts were widely made to draw diagrams, similar to a system of meshing gears, which could visually demonstrate the mechanism of the transmission of electric energy. But certain men of science advocated the theory of long-range interaction. They denied the physical process of transmitting electric and magnetic forces. They said that the concepts of a field and lines of force are to be treated as only geometric images that do not represent any possible kind of reality.

As has frequently occurred in the history of science, the truth turned out to be somewhere in the middle. Attempts to reduce electromagnetic phenomena to the motion of a special kind of matter—the ether—were



James Clerk Maxwell (1831-1879)—famed Scottish physicist; he was the founder of theoretical electrodynamics. Maxwell's equations describe the behaviour of electromagnetic waves and an electromagnetic field regardless of its origin. Maxwell is the author of the electromagnetic theory of light. The theoretical value of the velocity of light follows automatically from his equations,

found to be unsound. But, on the other hand, investigators that proposed that electromagnetic interaction is transmitted instantaneously from one charge or current to another were also wrong.

The famed Scottish physicist James Clerk Maxwell (1831-1879) published a work called *On Faraday's Lines of Force* when he was only 26. As a matter of fact, this work already contained the laws he discovered. Several more years were required, however, for him to discard his mechanistic conceptions and to formulate the laws of an electromagnetic field in a form that requires no naive graphical illustration.

In this connection Maxwell once said that for the benefit of people with different kinds of minds scientific truth should be presented in various forms and be regarded as equally scientific whether it is presented in the clear form and lively colours of physical illustration or in the simplicity and pallor of symbolic expression.

Maxwell's laws belong to the fundamental general laws of nature. They are not derived by logical reasoning and mathematical calculations. Fundamental laws of nature are generalizations of our knowledge. Laws of nature are discovered, found or ascertained. It is a matter of great interest to historians of science and to psychologists to follow out the chain of guesses and creative perception by means of which a genius discovers a law of nature. This, however, is a subject for a special book.

Maxwell's theory is the basis for the relationship between the permittivity and the refractive index, the orthogonality of the electric and magnetic vectors in a wave, and the existence of light pressure.

Of no less importance was Maxwell's contribution to the kinetic theory of gases. Maxwell worked out the velocity distribution law for gas molecules,

As for the reader and myself, nothing is left but to analyze a certain outline of consecutive assumptions that lead to Maxwell's laws.

What was known to Maxwell when he set himself the task of expressing in concise symbolic form the laws governing the behaviour of electric and magnetic fields?

In the first place, he knew that any point in space in the vicinity of an electric charge can be specified by a vector of electric force (intensity), and any point near an electric current, by a vector of magnetic force.

But are stationary electric charges the only source of an electric field? And are electric currents the only source of a magnetic field?

Maxwell answered these questions in the negative and proceeded in his search for the laws of an electromagnetic field along the following series of guesses.

Faraday showed that an electric current is induced in a wire loop which is threaded by an alternating flux of magnetic lines of force. But a current is produced when an electric force acts on electric charges. This being so, we can express Faraday's law by the following phrase: an electric field is set up in a wire loop through which an alternating magnetic flux passes.

But is it essential that the magnetic flux is encompassed by a wire loop? Is it not all the same to the electric field where it is set up: in a metal conductor or in empty space? Let us assume that it is all the same. Then the following statement is valid: a closed electric line of force appears near a variable flux of magnetic lines of force.

Thus the first two of Maxwell's laws concerning an electric field have been formulated. We contend that an electric field is set up in two ways: by electric charges (in which case the lines of force start at positive and end at negative charges) and by a variable magnetic field (in which case the electric line of force is closed and encompasses the variable magnetic flux).

Next we attempt to find the laws concerning a magnetic field. A magnetic field is set up by currents; this Maxwell knew. Direct (constant) current is the source of a constant magnetic field, and alternating (variable) current sets up a variable magnetic field. But we know that an alternating current is produced in wire by a variable electric field. And what if there is no wire, but only a variable electric field existing in empty space? Is it not logical to assume that a closed magnetic line of force is produced near a variable flux of electric lines of force? Its symmetry makes this picture very attractive: a variable magnetic flux sets up an electric field, and a variable electric flux, a magnetic field.

Thus, the two laws concerning electric fields are supplemented by two more that govern the behaviour of magnetic fields. That a magnetic field has no sources (there being no magnetic charges) is the third law, and that a magnetic field is produced by electric currents and a variable electric field is the fourth law.

Maxwell's four laws can be written with exceptional elegance in the form of mathematical equations. With respect to them and quoting a line from Goethe, the great Austrian physicist Ludwig Eduard Boltzmann (1844-1906) wrote: "Was it a god who wrote these lines... ." Unfortunately, I cannot acquaint the reader with the meaning expressed by these equations. This would require from him a knowledge of mathematics at a level above that used in this series (Figure 5.1).

Maxwell's laws indicate that a variable magnetic field cannot exist without an electric field, and a variable electric field, without a magnetic field. This is why the two adjectives, electric and magnetic, are not separated by a comma. An electromagnetic field is a single entity.

As we move away from the charges that are the sources of an electromagnetic field, we deal with electromagnetic matter, so to speak, in the pure form. It is not necessary

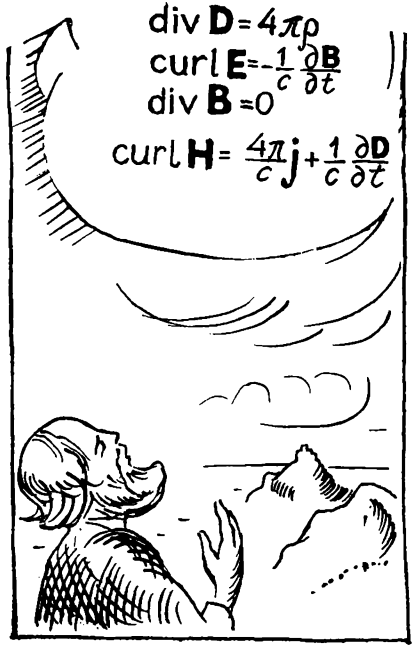


Figure 5.1

to consider pencils of lines of force. Maxwell's laws can be written so that they are applicable to a point in space. Then they sound especially simple: at a point where the electric vector varies with time, there exists a magnetic field vector that also varies with time.

"But is not all the aforesaid pure imagination?" asks the reader. In practice, as a matter of fact, it is absolutely impossible to measure the magnitudes of rapidly

varying electric and magnetic field vectors at a point in space.

True! But the grandeur of laws of nature is assessed by the consequences that follow from them. The consequences, in this case, are legion. It is no exaggeration whatsoever to contend that all of electrical and radio engineering is contained in Maxwell's laws.

It is imperative, however, to tell about one vital conclusion based on Maxwell's equations. It can be shown by faultlessly rigorous calculations that electromagnetic radiation must exist.

Assume that there are charges and currents in some bounded region of space. Various energy transformations may occur in this system. Mechanical or chemical sources produce electric currents, and the currents, in their turn, drive mechanisms and evolve heat that is liberated by the wires. Now let us calculate all the debits and credits. But they do not tally! The calculations show that some fraction of the energy of our system has escaped into space.

Has theory anything to say about this "radiated" energy? It seems that it has. The solution of the necessary equation is of quite complex form near the source, but at distances substantially exceeding the size of the "radiating" system, the picture becomes highly clear-cut and, what is most important, lends itself to experimental proof.

At great distances electromagnetic radiation, as we call the energy deficiency in our system of moving charges, can be characterized at each point in space by its direction of propagation. In this direction the electromagnetic energy travels at a velocity of about 300 000 km/s. This figure follows from the theory! The second conclusion of the theory is that the electric and magnetic vectors are perpendicular to the direction of wave propagation and perpendicular to each other. Thirdly, the

intensity of electromagnetic radiation (energy per unit area) decreases inversely proportional to the square of the distance.

Since it was known that light travels with about the same velocity that had been calculated for electromagnetic radiation, and sufficiently complete information on the polarization of light had been acquired to indicate that luminous energy has some "transverse" properties, Maxwell came to the conclusion that light is a form of electromagnetic radiation.

Ten years after Maxwell's death, at the end of the eighties, the famous German physicist Heinrich Rudolf Hertz (1857-1894) experimentally substantiated all the conclusions from Maxwell's theory. Following these experiments, Maxwell's laws and their equations were firmly established forever and ever as one of the few cornerstones supporting the structure of modern physics.

Mechanical Models of Radiation

Mechanical models are contrasted with mathematical models. Mechanical models can be contrived by using balls, springs, strings, rubber cords, etc. A mechanical model helps to make some phenomenon visual. By building a mechanical model and demonstrating its operation we help a person to understand the phenomenon by saying: "Such and such a quantity behaves similar to such and such a displacement." Far from all mathematical models can be contrasted with a mechanical one.

Before we take up electromagnetic radiation, a fact established by innumerable experiments and following by strict logic from Maxwell's equations, we should discuss the feasible mechanical models that can be contrived for radiation.

There are two such models: corpuscular and wave. We can make a toy which "radiates" streams of tiny

particles—peas or poppyseed—in all directions. This is the *corpuscular model* because the word "corpuscle" means a particle.

A particle flying at a certain velocity and having a certain mass should conform to the laws of mechanics. Particles are capable of colliding, changing their directions of motion, but only if the collision obeys the laws of conservation of energy and momentum. Some bodies may prove impenetrable to the particles. In such cases the particles must bounce off the bodies according to the law: the angle of incidence equals the angle of reflection. Particles can be absorbed by the medium. If particles can travel easier in one medium than in another, we can readily explain refraction. After passing through a hole in an opaque screen, a stream of particles emerging from a point source should travel within a cone. True, a small amount of scatter is possible because a small fraction of the particles are reflected by the edges of the hole. Of course, these reflected particles can only be chaotic and cannot produce any regular pattern that extends outside the geometric shadow.

The *wave model* is usually demonstrated by a water bath. The water can easily be made to oscillate at some point. From this point, as from a stone dropped into water, concentric waves spread outward in ever-widening circles. We can see the wave-like surface of the water. Energy is propagated in all directions and a wooden chip, floating at some distance, begins to oscillate with the frequency of the point to which we supply energy.

Sound vibrations are somewhat more difficult to see. But we can conduct absolutely convincing experiments to show that sound is transmitted by the mechanical displacement of the medium from point to point.

A great many phenomena are equally well explained by both the wave and corpuscular models. Both models are equally suitable, however, only when an additional

condition is complied with: a wave acts just like a stream of particles if the obstacles and holes it encounters are much smaller than the wavelength.

As we can readily calculate from the main formula, used to describe the wave model, $c = \nu\lambda$, the average frequency of a human voice of 1000 Hz corresponds to a wavelength of 30 cm. Such a wave can go around a corner if it has to pass through holes one metre in size. But if the hole has a size of the order of one centimetre, we can contend that a sound beam passes through it only when a straight line connecting the source and receiver of the sound meets no obstacles.

Suppose that a radio set is speaking in a room with open windows high up on the wall. A man sitting outside on a bench under the window can hear what is going on. If the windows are tightly closed and the walls are thick, the sound can pass only through the keyhole of the door. Now, even the most sensitive detector receives a signal only if the source of the sound, the keyhole and the detector are on a straight line. Here sound energy is transmitted like a stream of particles.

It can readily be shown by both reasoning and experiments in a water bath that the law of reflection from a wall of a roughness less than the wavelength is complied with by the wave model as well.

The reader knows quite well how sound or any other wave is reflected by a smooth plane surface. Interesting problems may be posed when the reflecting surface is of curved shape.

One such problem follows. What kind of surface is required to collect a wave, emitted by a point source, back into a single point? The shape of this surface must be such that rays from one point, incident to the surface at various angles, are all reflected to another single point. What is this surface?

The reader may recall one remarkable property of the

curve called an ellipse. The distance from one focus of the ellipse to any point of the curve plus the distance from this point to the other focus is the same for all points on the ellipse. Now imagine that the ellipse is rotated about its major axis. The rotating curve describes a surface which is said to be an ellipsoid. (Its shape resembles a symmetrical egg.) Another property of an ellipse is the following. If an angle is drawn with its vertex at a point of the ellipse and its sides passing through the foci, the bisector of this angle is normal to the ellipse at the point. This means that if a wave or a stream of corpuscles is emitted at one focus of the ellipsoid, it is reflected by the internal surface and arrives complete only at the other focus.

As far as sound waves are concerned, the walls and ceiling are sufficiently smooth. If the ceiling is vaulted, a special case of sound reflection can be observed. Since the shape of the arched ceiling is close to that of an ellipsoidal surface, sound emitted at one focus is reflected by the ceiling to the other focus. This property of vaulted ceilings was known in ancient times. In the Middle Ages, during the inquisition, it was made use of to eavesdrop. Two persons, disclosing their thoughts to each other in a quiet voice had no idea that they were being overheard by the dozy monk sitting at the other end of the tavern (Figure 5.2).

The corpuscular and wave models are equally suitable to explain this phenomenon. But a phenomenon like the collision of billiard balls cannot be explained by the wave model.

There are, on the other hand, several vital facts that the corpuscular model cannot cope with.

Primarily, this concerns *interference*, i.e. a summation in which the sum may turn out to be less than the addends or even equal to zero. If two waves arrive at a point and are added together, of prime importance is their differ-

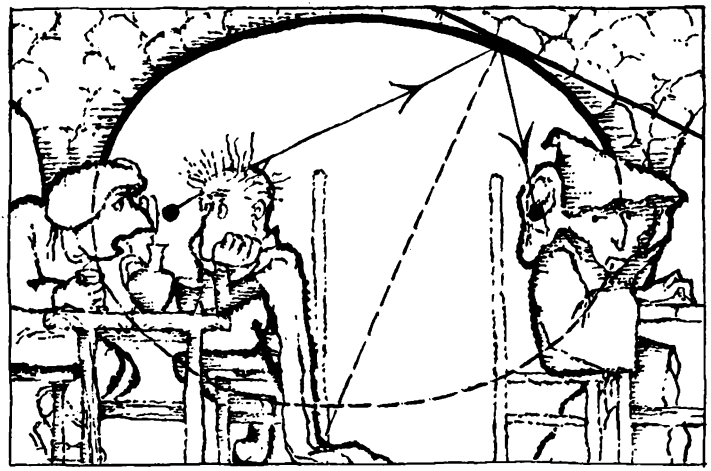


Figure 5.2

ence in phase at this point. If the crest of one wave coincides with the crest of the other, the waves add up. But if the crest of one wave coincides with the trough of the other and their amplitudes are the same, addition has a zero result. The waves arriving at a single point cancel out each other. When two wave fields are superimposed, at certain places the waves are added together and at others they are subtracted one from the other. This is interference. This is the first phenomenon that is absolutely impossible to deal with using the language of streams of particles. If radiation behaves like a stream of peas, then the superimposed fields must intensify each other at all places and at all times.

The second important phenomenon is *diffraction*, i.e. bending around corners, or passing around obstacles. A stream of particles cannot act in this way, but a wave

must behave in exactly this way. In school, diffraction is demonstrated by generating waves in a small bath of water. Then a partition with a hole is put into the path of the waves and the bending around corners becomes visible to the naked eye. The cause of this phenomenon is entirely natural. The particles of water begin to vibrate in the plane of the hole. Each point lying in the plane of the hole produces a wave, all the points having the same rights as the initial source of radiation. Nothing hinders this secondary wave from "bending around a corner".

Interference and diffraction can be demonstrated without difficulty if the condition mentioned above is complied with: the wavelength must be greater than or at least commensurable with the size of the obstacles or holes. We shall refine this condition and deal with diffraction and interference in more detail in Book 4.

Next we shall discuss the change in the frequency of waves observed when the source of radiation is in motion. That this phenomenon is a necessary consequence of the wave model was shown by the Austrian physicist Christian Johann Doppler (1803-1853) in the early days of theoretical physics.

We shall derive the equation of the *Doppler effect*, as it is called; it shall prove useful later. Assume, for the sake of clearer representation, that an automobile is approaching a marching brass band. The number of periodic compressions of the air reaching the driver's ear per unit time is greater, than if the automobile were at rest, by the ratio $(c + u)/u$, where c is the velocity of propagation of the waves and u is the relative velocity of the source of the waves and their detector (the driver in our case). Hence

$$\nu' = \nu\left(1 + \frac{u}{c}\right)$$

This means that the heard frequency ν' is higher when the automobile and band approach each other (the pitch is higher when $u > 0$) and drops when they are moving away from each other (the pitch is lower and $u < 0$). Getting a little ahead of our story, we can say that for light waves this conclusion is the following: a "red shift" occurs when the source is receding. The reader will appreciate the importance of this conclusion when we discuss the observation of the spectra of distant stars.

From a long time in the past and down to the twenties of our century, philosophers and scientists have often argued whether some case of energy transmission is of a wave or a corpuscular nature. Experiments indicated that any kind of radiation has two aspects. Only a combination of these two aspects properly represents reality. Theory has elevated this fact to the rank of a principal law of nature. Wave mechanics, quantum mechanics and quantum physics are equivalent names for the modern theory describing the behaviour of fields and particles.

Two Aspects of an Electromagnetic Field

In certain phenomena electromagnetic radiation behaves like waves and in others like a stream of particles.

In this sense, Maxwell's laws are prone to "error"; they only outline the wave aspect of electromagnetic radiation.

In complete agreement with experimental investigations, a solution of Maxwell's equations leads to the conclusion that we can always conceive of electromagnetic radiation as being the sum of waves of various lengths and intensities. If the radiating system is an electric current of fixed frequency, the radiation consists of a "monochromatic" (single-colour) wave.

An electromagnetic wave is illustrated in Figure 5.3. To understand the changes that take place in space in

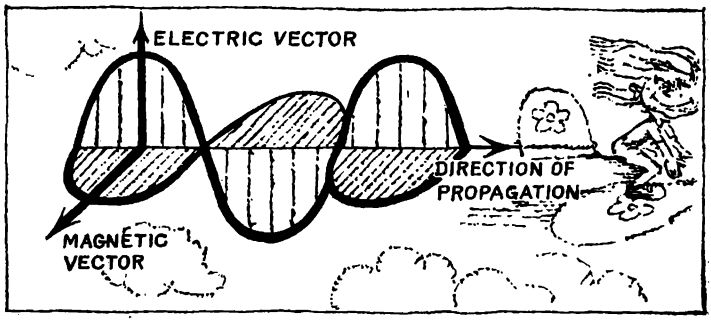


Figure 5.3

the transmission of electromagnetic energy, we must "pull" our figure as a rigid whole along the axis of abscissas.

This picture is obtained as a solution to Maxwell's equations. It is what enables us to speak of electromagnetic waves. In employing this term and in resorting to the analogy between an electromagnetic wave and one spreading outward in water when we drop a stone, we must proceed with extreme caution. Pictorial representations can easily mislead us. A wave on water is only a model of an electromagnetic wave. This means that electromagnetic waves and waves on water behave the same only in certain respects.

But are not Figure 5.3, illustrating an electromagnetic wave, and a water wave, which alternately raises and lowers a wooden chip thrown into the sea, alike as two peas? Nothing of the kind! Consider carefully the essence of the drawing. Plotted along the vertical axis are the electric field vectors and not some kind of spatial displacements!

The magnitude of the ordinate at each point indicates the force that an electric charge would be subject to if

it were placed at that point. Strictly speaking, nothing is moved from its place during the travel of an electromagnetic wave. It is absolutely impossible in practice to conduct an experiment showing the variation in value of an electromagnetic wave at some point, even for very slow vibrations.

Hence, the concept of an electromagnetic wave is of a theoretical nature. We confidently speak of the existence of electromagnetic waves because we listen to the radio. We have no doubt that an electromagnetic wave has a certain frequency because we must tune our radio set to a definite frequency for the reception of a definite broadcasting station. We are sure that the concept of length is applicable to an electromagnetic wave. Not only can we measure the wave velocity and calculate the wavelength by formula $c = \nu\lambda$, which relates the frequency, wavelength and velocity of propagation, but we can obtain an idea of the length of electromagnetic waves from diffraction, i.e. bending around a corner or an obstacle. The principle involved in this last kind of measurement is the same as for waves travelling on water.

It is absolutely necessary to warn the reader that he should not attempt to work out a visualizable conception of an electromagnetic wave because, as mentioned at the beginning of this section, electromagnetic radiation not only "resembles" waves, but in many cases "reminds" us of the behaviour of a stream of particles. To try to imagine something that simultaneously resembles a stream of particles and a wave is entirely contrary to reason. We are discussing physical processes that cannot be represented by chalk on a blackboard. This does not imply, of course, that we cannot obtain exhaustive knowledge of an electromagnetic field. We must keep in mind that graphic pictures are only a teaching aid, a method for more easily memorizing symbols. But we should not

forget that the wave picture is only a model of electromagnetic radiation, and nothing more! Where feasible, we employ this model, but we should not be surprised that this model can only lead us into error in some cases.

In exactly the same way, the corpuscular aspect of an electromagnetic field is not always observed. It would, of course, be simpler if the conditions under which these two aspects are manifested were mutually exclusive. But no, this is not so. Even in describing one and the same experiment, it is often necessary to speak in two languages.

It is nevertheless simpler (or, it is better to say, it previously was simpler) to observe the corpuscular aspect of electromagnetic radiation for short waves. In an ionization chamber and other similar instruments, we can observe the collision of particles of electromagnetic radiation with an electron or other "honest" particles. The collision may resemble that of billiard balls. It is absolutely impossible to understand such behaviour if we only resort to the wave aspect of electromagnetic radiation.

Using the language of Maxwell's theory, let us consider the production of electromagnetic radiation. A system of charges oscillates with a certain frequency. The electromagnetic field varies in time with these oscillations. The frequency ν of oscillations of the field divided by the velocity of propagation (300 000 km/s) gives the wavelength of the radiation.

When we deal with the same phenomenon in the language of quantum physics, it is described as follows. We have a system of charges characterized by a system of discrete energy levels. For some reason the system got into an excited energy state, but did not exist long in this state and went over to a lower energy level. The energy $E_2-E_1=h\nu$ evolved in this transition is radiated in the form of particles called photons. We are already

acquainted with constant h (see p. 120). It is Planck's constant.

If the energy levels of the system are very close together, the photon has low energy, low frequency and, consequently, a long wavelength. Here the quantum corpuscular aspect of the electromagnetic field is hardly noticeable and manifests itself only in absorption phenomena, which are associated with extremely small changes in the energy of electrons or atomic nuclei (magnetic resonance). In the case of waves of long length we do not succeed in observing collisions of a photon with particles, like the collisions of billiard balls.

We shall briefly mention facts that, so to speak, drove physicists into a corner, forcing them to agree that wave theory (which they had believed in for many years as the complete and comprehensive truth) is incapable of explaining *all* the facts concerning electromagnetic fields. There are many such facts but, for the time being, we shall restrict ourselves to a phenomenon called the photoelectric effect. After the reader has consented that a comprehensive picture of the electromagnetic field cannot be created without its corpuscular aspect, we shall discuss Hertz's wonderful experiments on which all of radio engineering is based. We shall show how the wave aspect of the electromagnetic field was depicted, not only in its general form, but in all its details as well.

Photoelectric Effect

The fine sonorous name "photon" appeared on the scene somewhat later than the product of Planck's constant h by the frequency ν of an electromagnetic wave. As mentioned above, the transition of a system from one energy state to another is accompanied by the absorption or emission of a portion of energy equal to $h\nu$. The first to come to this conclusion, at the turn of the century,

was the famous German physicist Max Karl Ernst Ludwig Planck (1858-1947). He showed that this was the only way that the radiation of incandescent bodies could be explained. The argument concerned electromagnetic waves produced by a method having nothing in common with radio engineering. It had not yet been proved at that time and was not recognized by all physicists that what is valid for light is true for radio waves as well. This was the situation even though Maxwell's laws definitely proved that there is no difference in principle between radio waves and any other electromagnetic waves, including light. An understanding and the experimental proof of the universal validity of Planck's statement came later.

Planck's work concerned the radiation of light in discrete portions, i.e. in *quanta*. The work did not note, however, that the quantum nature of radiation made it inevitable to introduce the corpuscular aspect of the electromagnetic field into the discussion. A field was said, at that time, to radiate in portions, but a portion was thought to be a certain wave train.

The most important step, i.e. the recognition of the fact that the radiated portion of energy $h\nu$ is the energy of a particle which was immediately named the *photon*, was made by Einstein. He showed that only corpuscular concepts could explain the *photoelectric effect* (*photoeffect*), i.e. the ejection of electrons from solids under the influence of light.

Shown schematically in Figure 5.4 is the apparatus employed at the end of last century to begin a detailed study of the phenomenon called the *extrinsic photoelectric effect*.

In 1888, Heinrich Hertz reported that light in some way affects the electrodes of a vacuum tube. He was evidently the first to notice this phenomenon. Working at about the same time, the Swedish chemist Svanté

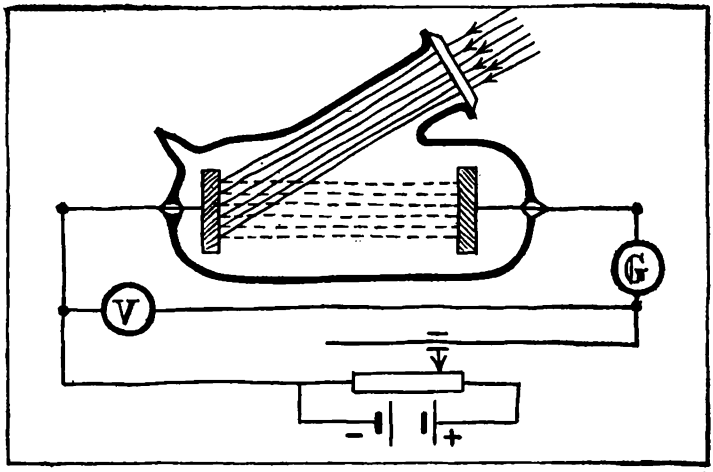


Figure 5.4

August Arrhenius (1859-1927), the German physicist Wilhelm Ludwig Franz Hallwachs (1859-1922), the Italian physicist Augusto Righi (1850-1921) and the outstanding Russian physicist Alexander Grigoryevich Stoletov (1839-1896) showed that illumination of the cathode produces an electric current. If no voltage is applied across the tube shown in Figure 5.4 (it is called a *photoelectric cell*), only a negligible part of the electrons, ejected from the cathode through the impact of light, reach the opposite electrode. A weak accelerating voltage (with a negative photocathode) increases the current. Finally, as the voltage is increased, the current reaches its saturation value: all the electrons (whose number is quite definite at the given temperature) reach the anode.

The magnitude of the photocurrent is strictly proportional to the light intensity. The light intensity is

uniquely determined by the number of photons. It immediately occurs to us (and this is confirmed by rigorous calculations and experiments) that each electron is ejected by one photon.

The energy of the photon is expended in liberating an electron from the metal and imparting a velocity to it. This is expressed by an equation first derived by Albert Einstein in 1905. Here it is:

$$h\nu = \frac{mv^2}{2} + A$$

where A is the work function (see p. 88).

The energy of the photon must at least exceed the work function, i.e. the minimum energy it is necessary to expend to eject an electron from the metal. This means that for the photon each energy (and the energy is uniquely related to the "chromaticity") has its threshold value of the photoelectric effect.

Photoelectric cells, based on the extrinsic photoeffect, are extensively used in photorelays, television and sound motion pictures.

The sensitivity of a photoelectric cell can be increased by filling it with gas. Here the current is increased because the liberated electrons break up the neutral molecules of gas and unite them with the photocurrent.

The photoelectric effect, true, not the one we have described, but the so-called *internal photoelectric effect*, occurring at the boundary of the *p-n* layer in semiconductors, is of exceptional importance in modern engineering. To avoid interrupting our discussion, however, we shall postpone the treatment of the application of the photoelectric effect until Book 4. Here we mentioned this phenomenon only to demonstrate the inevitability of recognizing that an electromagnetic field has corpuscular properties.

For many years photons were the forlorn stepsons of

physics. Proof of the existence of photons and the investigation of the laws of the photoelectric effect anticipated the founding of quantum physics by about 20 to 30 years. Only by the end of the twenties, when these laws had been established, it became clear why the same numerical constant, Planck's constant h, appears both in the formula for the energy of a photon and in the formula, discussed on p. 120, that determines the possible values of the angular momentum of particles.

The value of this constant is determined from many different kinds of experiments. The photoelectric effect, the Compton effect (change in the wavelength of X-rays in scattering), the radiation produced upon the annihilation of particles and many other experiments yield the same numerical value.

Hertz's Experiments

Now let us consider the way in which the hypotheses concerning the wave aspect of the electromagnetic field were proved.

Logic and mathematics extract many consequences from Maxwell's laws. These consequences may turn out to be valid or they may not be confirmed by experiments. A physical theory becomes a part of science only after being experimentally checked. The establishment of electromagnetic field theory—from separate facts to a general hypothesis, from the hypothesis to the consequences and to the final stage, an experiment that provides decisive proof—is the only proper procedure for a naturalist. This highway to scientific truth is especially defined in the case of the laws of the electromagnetic field.

For this reason we shall discuss Hertz's experiments in more detail. They are still being used by teachers and lecturers to demonstrate to students of schools and

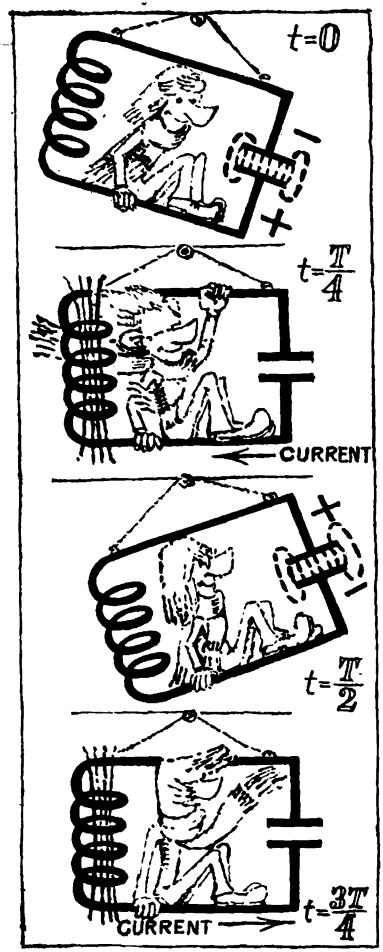


Figure 5.5

institutes how a scientist gains confidence in the validity of laws of nature.

We must begin our story in 1853, when the famous British physicist Lord Kelvin proved mathematically that electric oscillations occur when a capacitor is discharged through an inductance: the charge on the capacitor plates, the voltage at any point in the circuit and the current all vary according to the law of harmonic oscillations. If we assume that the resistance of the circuit is negligible, these oscillations will continue forever.

The pictures in Figure 5.5 demonstrate the phenomena that occur in a so-called *oscillatory circuit*. At the initial instant of time the capacitor is charged. Current flows as soon as the circuit is closed. After one-fourth of a period, the capacitor is completely discharged. Its energy, $q^2/2C$, is converted into the energy of the coil's magnetic field. The current is at a maximum at this instant. Continuing in the same direction, the current gradually decreases. After a half-period, the current drops to zero, the magnetic energy, $LI^2/2$, disappears, the capacitor is completely charged again and its energy is restored. But the sign of the voltage is reversed. Then the process is repeated in the reverse direction. After a certain time T (the period of oscillation), everything returns to the initial state and the process begins again.

Electric oscillations would continue to infinity except for the inevitable resistance to current flow. Energy is lost in each period due to resistance. As a result, the oscillations decrease in amplitude and are soon damped.

The obvious analogy with a load suspended by a spring enables us to dispense with an algebraic analysis of the process and to reason out the period of oscillations of such a circuit. (The reader should refresh his memory of the corresponding pages in Book 1.) It is sufficiently evident, as a matter of fact, that the electric energy of the capacitor is equivalent to the potential energy

of the compressed spring, and the magnetic energy of the coil, to the kinetic energy of the weight.

Associating analogical quantities, we "derive" the formula for the period of oscillations occurring in the circuit: $q^2/2C$ is the analogue of $kx^2/2$, $LI^2/2$ of $mv^2/2$, k of $1/C$ and L of m. Hence the frequency of oscillations is $\nu = 1/2\pi\sqrt{LC}$ because the corresponding formula for mechanical oscillations is of the form

$$\nu = \frac{1}{2\pi}\sqrt{\frac{k}{m}}$$

Now let us try to guess the train of thought of Hertz, who made it his aim to prove the existence, without leaving his laboratory, of electromagnetic waves propagating at a velocity of 300 000 km/s. For this purpose, an electromagnetic wave is required with a wavelength of the order of 10 m. If Maxwell was right, the electric and magnetic vectors should oscillate with a frequency of 3×10^8 hertzes; pardon me, cycles per second. At that time Hertz did not know that his name would be perpetuated as the unit of frequency.

"What shall I begin with?" he probably asked himself. In the first place, since the oscillations are damped, it is necessary to devise an arrangement that recommences the process when the current stops. This is not a difficult task. The required circuit is illustrated in Figure 5.6. An alternating voltage is applied across the primary winding of transformer T. As soon as it reaches the breakdown voltage between the metal spheres connected to the secondary winding, a spark is initiated. This spark is what closes the oscillatory circuit OC, acting like a switch key, and dozens of oscillations, with a decreasing amplitude and a more or less high frequency, pass through the circuit.

But a high frequency is required. What can be done to obtain one? We can reduce the self-induction and the

Heinrich Rudolph Hertz (1857-1894)—brilliant German physicist; using his "oscillator" and "resonator" he proved in experiments that an oscillatory discharge produces electromagnetic waves in space. Hertz showed that electromagnetic waves are reflected, refracted and subject to interference, thereby confirming

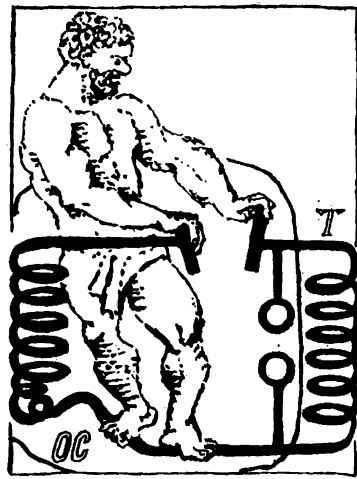


Figure 5.6

capacitance. But how? We replace the coil by a straight wire, and we spread the capacitor plates apart and reduce their area. Into what does the oscillatory circuit degenerate? It is converted into two simple rods ending in metal spheres between which sparks jump.

Thus Hertz got the idea for his oscillator which can serve both as a source and as a detector of electromagnetic waves.

It was difficult for Hertz to predict beforehand the inductance and capacitance of this unusual "circuit" of which almost nothing remained. The inductance and capacitance of the oscillator are not concentrated in

---

Maxwell's theory. Hertz's experiments laid the foundation of radio engineering. In his first radio broadcast in 1895, Alexander Stepanovich Popov, inventor of the radio, transmitted two words: "Heinrich Hertz".

some definite place in the circuit but are distributed along the rods. Some other theory is required.

A discussion of this new approach to electric circuits with currents of very high frequency would take us too far from our main subject. The reader can believe our word that oscillations of high-frequency current really do occur in Hertz's oscillator.

The "transmitter" and "receiver" of waves, used by Hertz, were practically the same. In the "transmitter" the waves were produced by sparks periodically jumping the gap between the spheres in accordance with the operation of the transformer that applied voltage to the oscillator. The width of the spark gap could be regulated by a micrometric screw. Serving as the receiver, or detector, was either a rectangular turn of wire interrupted by a spark gap or two small rods that could be adjusted as required to obtain a gap within a fraction of a millimetre.

In his first work, published in 1885, Hertz showed that oscillations of extremely high frequency can be obtained by the above-described method and that these oscillations actually do set up an alternating field in adjacent space. The existence of this field, he wrote, can be demonstrated by sparks jumping across a gap in the "receiver". Hertz called this receiving oscillator a resonator. He immediately grasped the principle of detecting an electromagnetic field. This principle is the basis of modern radio engineering.

Incidentally, we may note that neither in Hertz's works nor in others published in the next few decades do we find the terms "electromagnetic waves" or "radio waves". They speak only of electric waves or of waves of electrodynamic force.

In his next work Hertz showed that, in accordance with the requirements of Maxwell's theory, the dielectric medium (a bar of sulphur or paraffin) affects the

frequency of the electromagnetic field. When he received this article, Helmholtz, the editor of the scientific journal, sent Hertz a postcard with the message: "Your manuscript has been received. Bravo! I shall send it to the press on Thursday."

Hertz's work in which he showed that electromagnetic waves can be reflected made a tremendous impression on the physicists of his day. The waves were reflected by a zinc screen, 4×2 m in size. The oscillator was 13 m from the screen and at a height of 2.5 m above the floor. The tuned resonator was arranged at the same height and could be moved between the oscillator and screen. Setting the resonator at various distances from the screen and observing the intensity of the sparks, Hertz established that there were maximums and minimums typical of the interference phenomena that set up standing waves. The wavelength was found to be approximately 9.6 m.

I should like to emphasize the fact that at that time no one could say what material could serve as a mirror for electromagnetic waves. Today we know that waves of such length do not penetrate metals, but are reflected from them.

In trying to obtain additional proof of Maxwell's theory, Hertz reduced the geometrical size of his apparatus until he reached a wavelength of 60 cm. In 1888 he published the work *On the Beams of Electric Force*. He was able to concentrate electromagnetic energy by means of parabolic mirrors. The rods of the oscillator and resonator were arranged at the foci of the mirrors. Employing the mirror-type transmitter and receiver, Hertz showed that electromagnetic waves do not pass through metals, but wooden screens do not check them.

Figure 5.7 shows how Hertz demonstrated the polarization of electromagnetic waves. He arranged grating C, made of parallel copper wires, in the path of the elec-

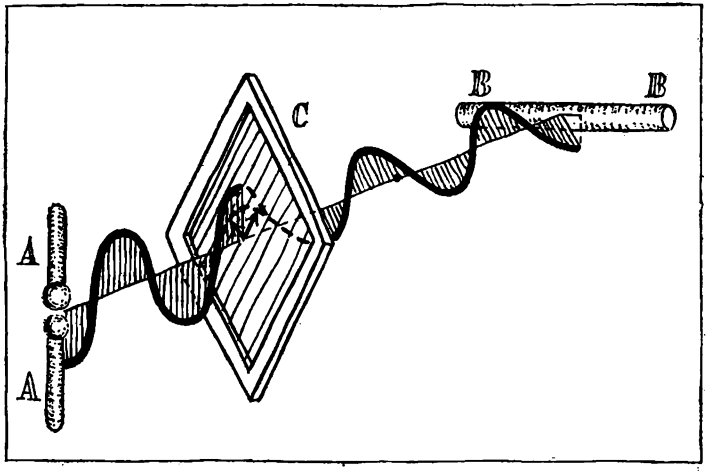


Figure 5.7

tromagnetic beam produced by oscillator *AA*. Hertz showed that when he turned the grating, the intensity of the spark in resonator *BB* was varied. When the grating wires were parallel to the electric vector and perpendicular to the axes of the oscillators, the beam did not pass through the grating. Thus, the transverse character of an electromagnetic wave was proved.

Finally, to study the refraction of the waves, Hertz made a prism weighing over one tonne from an asphalt compound. The refractive index of the asphalt for waves 60 cm in length could be measured with high accuracy. It was found to equal 1.69.

The proof of the existence of electromagnetic waves, measurement of their length, and the establishment of the laws of their reflection, refraction and polarization are all the results of only three years of research! Truly a capacity for work to be admired.

Classification of Electromagnetic Radiation

Physicists deal with electromagnetic radiation over a huge range. The electromagnetic radiation of current of city mains frequency is absolutely negligible. Practically, the feasibility of detecting electromagnetic radiation begins with frequencies of the order of dozens of kilohertzes, i.e. at wavelengths equal to hundreds of kilometres. The shortest waves have a length of the order of ten thousandths of a micron, i.e. a frequency of the order of a thousand million gigahertzes.

Radio waves are in the range of electromagnetic radiation produced by electrical engineering means, i.e. as a result of the oscillation of electric currents. The shortest radio waves have a wavelength of hundredths of a millimetre.

Beginning with several hundred microns and smaller is the wavelength range of radiation due to energy transitions inside molecules, atoms and atomic nuclei. As we can see, this range substantially overlaps the radio range.

Visible light occupies a narrow range. Its limits are 0.38 to 0.74 micron. Radiation of longer wavelength, not obtained by radio engineering techniques, is said to be *infrared*; that of shorter wavelength is called *ultraviolet* and extends down to a wavelength of about 0.1 micron.

Electromagnetic radiation produced in X-ray tubes overlaps the range of ultraviolet waves and extends down to a wavelength of 0.01 micron, where it overlaps the gamma ray range. *Gamma rays* are produced in nuclear decay, in nuclear reactions and in collisions between elementary particles.

The principal characteristic of any electromagnetic radiation is its spectrum. The *spectrum* is a graph on which

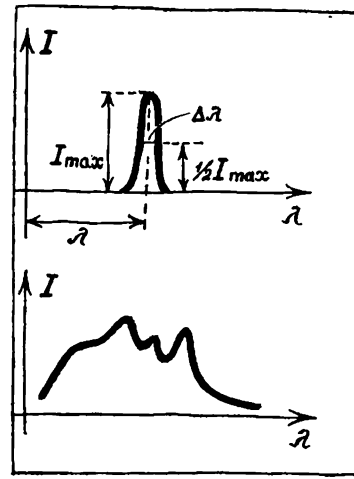


Figure 5.8

the intensity (i.e. the energy per unit time per unit area) is plotted in the vertical direction and the wavelength or frequency in the horizontal direction. The simplest spectrum is that of *monochromatic* ("single-colour") *radiation.* Its graph consists of a single line of extremely narrow width (Figure 5.8, above). The degree of monochromaticity of the line is specified by the ratio $\lambda/\Delta\lambda$. A broadcasting station produces almost purely monochromatic radiation. For a short-wave station, for instance, operating in the 30-m range, $\lambda/\Delta\lambda$ equals about 1000.

Excited atoms, for example, atoms of gases in daylight lamps (the excitation being due to the collisions of positively and negatively charged particles travelling toward the anode and cathode), produce a spectrum consisting of a great number of monochromatic lines with a relative width of $100\,000^{-1}$. Lines are observed in magnetic resonance with a relative width as small as 10^{-7}.

Strictly speaking, no continuous spectra exist. How-

ever, if the lines overlap, the experiment yields an intensity curve of the kind shown below in the same illustration.

Information on an electromagnetic spectrum can be obtained either by investigating radiation or by studying absorption. In general, both experiments provide the same information. This is evident from the principal law of quantum physics. In radiation, the transition of the system is from a higher to a lower energy level; in absorption, it is from a lower to a higher level. But the difference in energy, which determines the frequency of radiation or absorption, is exactly the same. Which spectrum we investigate, radiation or absorption, is simply a matter of convenience.

In characterizing a radiation spectrum, we can, of course, use either wave or corpuscular terminology. Employing the wave aspect of radiation, we state that the intensity is proportional to the square of the wave amplitude. Regarding radiation as a stream of particles, we count up the intensity as the number of photons.

We repeat again that we should in no way be disturbed by this alternating application of the two aspects of radiation. Radiation resembles neither waves nor a stream of particles. Both concepts are only models that can be conveniently applied in explaining various phenomena.

We have not presented a scale of electromagnetic waves, but have mentioned with sufficient clarity that the names of its various ranges are, to some degree, arbitrary. In any case, we find occasions in which waves of the same length are differently called, depending on the way in which they were produced.

Today, the scale of electromagnetic waves is continuous. There are no ranges that have not been obtained by one or another method.

But the overlapping of the infrared and radio waves,

the gamma and X-rays, etc. was discovered a relatively short time ago. For a long time there was a gap between the short radio waves and the infrared range. Waves with a length of 6 mm were first obtained by the brilliant Russian physicist Pyotr Nikolayevich Lebedev (1866-1912) and heat (infrared) waves of a length to 0.34 mm, by the German physicist Heinrich Rubens (1865-1922).

In 1922 this gap was closed by the Russian physicist Alexandra Andreyevna Glagoleva-Arkadyeva (1884-1945) when she obtained electromagnetic waves in a range from 0.35 mm to 1 cm by nonoptical techniques.

At the present time, waves of this length are obtained by radio technicians without any trouble whatsoever. Glagoleva-Arkadyeva had to employ much ingenuity and resourcefulness to develop the required apparatus, which she named the mass radiator. The source of the electromagnetic radiation was metal filings in suspension in transformer oil. A spark discharge was passed through this mixture.

6. Radio

Some History

Just as Faraday had no idea that the discovery of electromagnetic induction would lead to the founding of electrical engineering, and Ernest Rutherford considered idle talk and rank ignorance the feasibility of extracting energy from the atomic nucleus, so Heinrich Hertz, after discovering electromagnetic waves and showing how they can be detected at distances of several metres, had no notion of radio communication and even denied its possibility. Three amusing facts, are they not? But we shall leave their discussion to the psychologists. Therefore, restricting ourselves to a mere statement of this remarkable circumstance, we shall see how events developed after Hertz's early death in 1894.

Hertz's classical experiments, which we have described in such detail, attracted the attention of scientists all over the world. Professor N. G. Yegorov of the St. Petersburg University accurately repeated these experiments. The spark in the resonator was barely visible. It could be seen only in complete darkness and then only with a magnifying glass.

Alexander Stepanovich Popov (1859-1906), an unassuming lecturer in electrical engineering at the Kronstadt Military Academy, set to work in 1889, at the age of 30, to improve Hertz's experiments. The sparks that he obtained in his resonators were much fatter than those other investigators succeeded in producing.

In 1894, a fall issue of the English journal *Electrician* published an article by the well-known English physi-

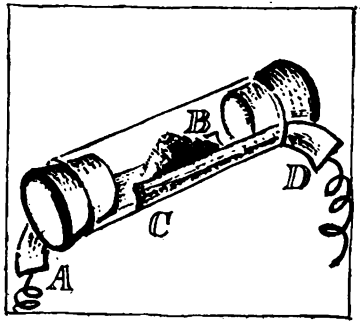


Figure 6.1

cist Sir Oliver Joseph Lodge (1851-1940) in which he claimed that Hertz's resonator could be improved by using the Branley tube. The French scientist Édouard Branley (1844-1940) was engaged in research on the conductivity of metal filings. He found that such filings do not always offer the same resistance to electric current. Loosely packed metal filings in a tube have practically infinite resistance, but if the tube is placed in the vicinity of an operating Hertz resonator, the resistance drops drastically. The explanation is that the small filings cohere owing to the welding action of the tiny sparks produced between them by the electromagnetic wave. When lightly tapped or shaken, the resistance of the filings is restored.

This property of metal filings was made use of by Lodge. He wired a circuit consisting of the Branley tube (which came to be known as a *coherer*), a battery and a sensitive galvanometer. The hand of the instrument was deflected at the instant electromagnetic waves passed by. Lodge succeeded in detecting radio waves at distances up to 40 m.

This system was inconvenient, however, because the coherer immediately stopped operating. It was necessary to find a way to restore the cohering (welded) filings

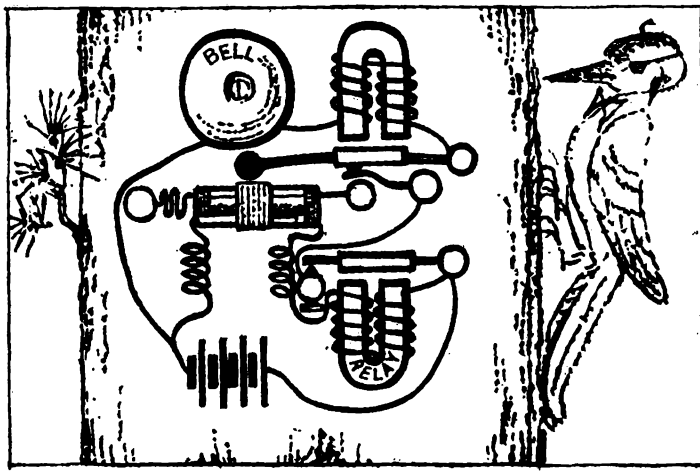
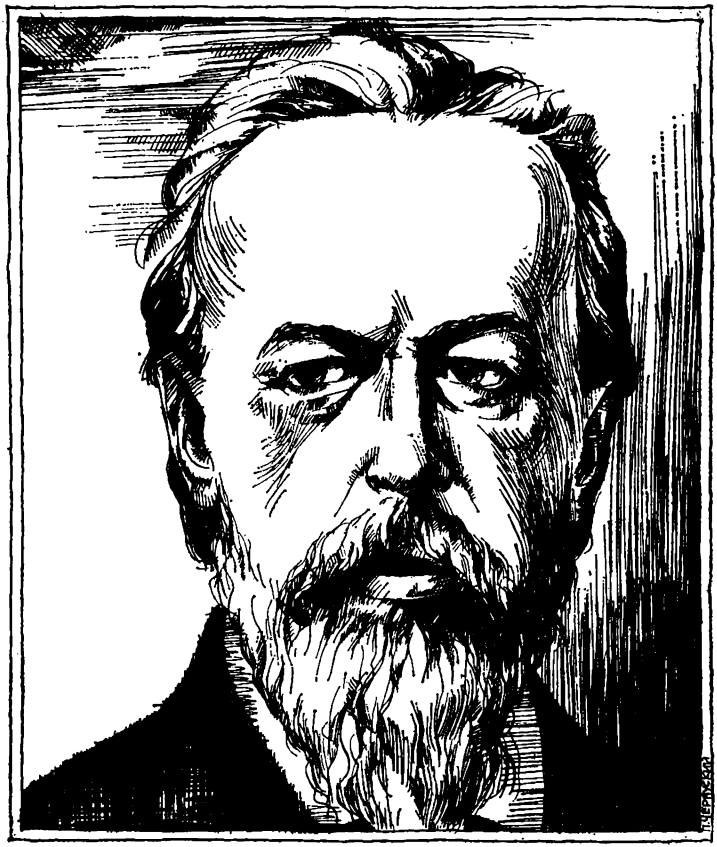


Figure 6.2

to their initial state, with the required device designed so that the shaking occurs automatically.

This problem was solved by Popov. He tried many kinds of coherers and finally decided to use one designed as follows. "Two strips of thin sheet platinum, AB and CD, are glued on the inside wall of a glass tube and extend over almost the whole length of the tube. One strip emerges at one end of the tube and the other strip at the other end. The strips are 8 mm wide and are arranged with a space of about 2 mm between them. The inner ends, B and C, of the strips do not reach the corks that plug the tube so that filings jammed between the cork and the tube cannot form conducting chains that cannot be broken up when the tube is shaken. This happened in some of the earlier models. The length of the whole tube is from 6 to 8 cm and its diameter is about



Alexander Stepanovich Popov (1859-1906)—great Russian physicist and electrical engineer; invented the radio. Popov's scientific work was highly appraised by his contemporaries. In 1900 he won a Gold Medal at the World's Fair in Paris for his invention.

1 cm. In operation the tube is arranged horizontally so that the strips are at the bottom and are covered by metal filings. Best operation is achieved when the metal filings fill not over one-half of the space in the tube."

Popov's coherer, just described in his own words, is shown in Figure 6.1. He used iron or steel powder in it.

The main problem, however, was not to improve the coherer, but to invent a method of restoring its initial state after detecting an electromagnetic wave. In Popov's first receiver, whose wiring diagram is shown in Figure 6.2, this job was performed by an ordinary electric bell. The striker of the bell replaced the galvanometer hand and its hammer struck the glass tube when the striker returned to its initial position.

What a simple solution of a puzzling problem! And really simple. Do you realize, dear reader, that this elementary arrangement, which neither Hertz nor Lodge hit upon, was the first application of what engineers now call a relay circuit? The negligible energy of radio waves is not directly detected, but is used to control a current circuit.

In the spring of 1895, Popov set up his apparatus outside, in an orchard. He began to move the receiver farther and farther away from the oscillator. At 50 m the bell responded to the spark of the oscillator; at 60 m the apparatus still worked, but at 80 m the bell refused to ring. Then Popov took a coil of copper wire, threw one end up into a tree and connected the other end to the coherer. The bell rang. This is how the first antenna was devised.

In the USSR, the 7th of May is celebrated each year as Radio Day. On this day in 1895 Popov read a paper with the unassuming name "On the Relationship of Metal Powders to Electric Oscillations" at the regular session of the Russian Physico-Chemical Society. Many of those present had watched a demonstration, several years

previously, of Hertz's experiments in which the tiny sparks could be seen only through a magnifying glass. But, when they heard the brisk clangour of Popov's receiver, they realized that they were witnessing the birth of the wireless telegraph, that a new and more efficient method of transmitting signals over long distances had been invented.

On March 12, 1896, Popov transmitted the world's first radiogram. By pressing a key for short and long intervals, the words "Heinrich Hertz" were transmitted over a distance of 250 m from one building to another and were recorded on telegraph tape.

By 1899, the range of radio communication between the training ships of a mine-layer detachment had already reached 11 km. The practical importance of the wireless telegraph was no longer doubted even by the most skeptical officials.

The Italian electrical engineer and inventor Guglielmo Marconi (1874-1937) began his experiments somewhat later than Popov. Marconi carefully followed all the advances in the fields of electrical engineering and the study of electromagnetic waves. He skilfully employed them to improve the quality of radio reception and transmission. His especial contribution was more in the line of organization and management rather than the strictly technical aspect. This is no small matter, however, and Marconi's fame is well deserved. We should not forget that the priority in the discovery of the radio, based on the paper read on May 7, 1895, belongs to the unpretentious Russian scientist Popov who always refused to put his knowledge and research at the disposal of any country except his native land.

Marconi did not mention Popov in his articles and lectures. But not everyone knows that in 1901 he offered Prof. A. S. Popov a position in the commercial company that Marconi had founded and was the president of.

Figure 6.3

The range of radio reception grew at a rapid rate. In 1899, Marconi established radio communications between France and England, and in 1901 a radio signal was sent from Europe to America.

What technical innovations facilitated these successes and made radio broadcasting feasible?

Beginning with 1899, radio engineering no longer based reception on the coherer. Instead of detecting radio waves by the drop in the resistance of a circuit due to the effect of an electromagnetic wave, an entirely different technique can be made use of. A rectified pulsating electromagnetic wave can be detected by the clicks heard in an ordinary telephone receiver.

This began a search for rectifiers. The extensively used contact detector, applied right up to the twenties of our century, consisted of a crystal with one-way conduction. Such crystals had been known since 1874. They include metal sulphides, copper pyrites and hundreds

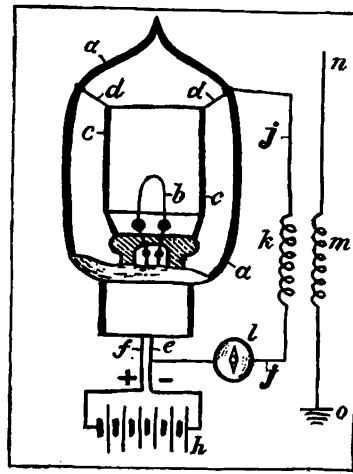


Figure 6.4

of different minerals. People of my age remember such radio sets and the irritating procedure of searching for "good contact" with a whisker (contact spring). Such contact was achieved when the point of the whisker had found a "suitable" spot on the crystal (Figure 6.3). By that time many broadcasting stations were in operation so that the set had to be tuned to the required wave. This was done with a multicontact switch for a small number of stations, or by steplessly varying the capacitance of the capacitor, which is also employed in up-to-date apparatus.

It was extremely difficult, or even impossible in some cases, to operate at high power with the spark-gap broadcasting stations because the spark-gap device became overheated. Such stations were soon replaced by one operating on the principle of an oscillating electric arc or a high-frequency alternator. After this the power ratings reached hundreds of kilowatts.

The real revolution in radio communications, enabling speech and music to be transmitted instead of only telegraph code, came with the development of the vacuum tube.

In October 1904, the English electrical engineer Sir John Ambrose Fleming (1849-1945) showed that high-frequency current could be rectified by a vacuum tube consisting of a filament heated by the current and surrounded by a metal cylinder. Its diagram is illustrated in Figure 6.4. Fleming realized the value of the vacuum-tube diode for converting electric oscillations into sound (he called this device a "valve" because it opened and shut the gate to the flow of electricity), but he could not achieve wide application of his detector.

The fame of inventing the electron lamp was won by the American scientist Lee de Forest (1873-1961). In 1906, he converted the two-electrode tube (diode) into a triode by adding a third element. The tube was called an *audion* (from the Latin word *audire*, meaning *I hear*). De Forest's vacuum tube radio set received signals on the grid (third element) in the lamp, rectified them and enabled them to be heard as telegraph code in headphones.

The feasibility of using the electron tube as an amplifier was evident to the American scientist. But only after seven years had passed, in 1913, a triode was first applied in a generator circuit by the German radio engineer Alexander Meissner (1883-1958).

Attempts to transmit speech, i.e. the modulation of an electromagnetic wave, had been made before the electron tube was used as a generator. But the difficulties were very great: the band of frequency modulation could not be made wide enough. Speech could be transmitted with some little success, but not music. Only in the twenties did radio transmitters and receivers, operating with electron tubes, demonstrate the truly inexhaustible capabilities of radio communications for transmitting the whole range of audio frequencies.

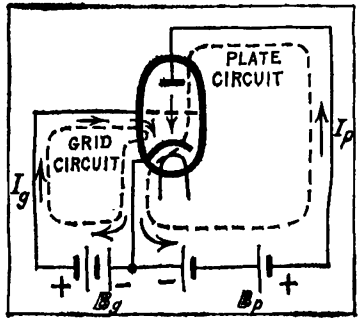


Figure 6.5

The next revolutionary breakthrough occurred not long ago, when semiconductor elements superseded electron tubes in radio circuits. This established a new branch of applied physics dealing with the huge complex of problems concerning the input, transmission and storage of information.

Vacuum-Tube Triode and Transistor

Vacuum-tube triodes incited an upheaval in radio engineering. But technology ages faster than people do. Today, the electronic tube has become a pensioner or, as they say in the USA, a senior citizen. Not so many years ago you could hear impatient prospective customers in shops selling TV sets demanding transistorized models, i.e. sets in which semiconductors have replaced electronic tubes.

But age demands respect. Moreover, the principles underlying the two fundamental applications of tubes and transistors, namely, the amplification and generation of waves of a definite frequency, can be more simply described by using the electronic tube as an example.

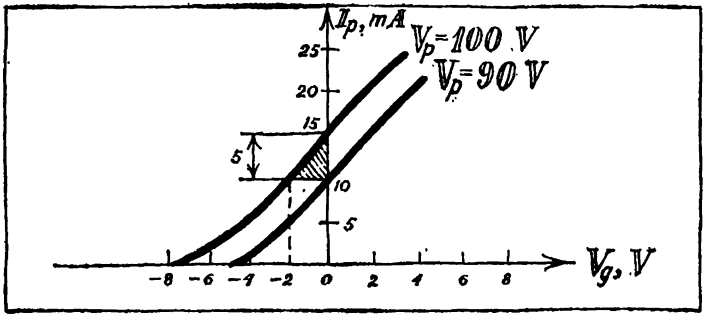


Figure 6.6

We shall therefore cover its operation in more detail than that of the transistor.

Besides the plate (anode) and heated filament (cathode), the bulb of a three-electrode tube has a sealed-in third electrode, called the grid. The electrons freely pass through the grid. Its openings are as much larger than the electron as the earth is larger than a dust speck. Figure 6.5 illustrates how the grid enables the plate current to be controlled. It is evident that a negative voltage impressed on the grid reduces the plate current and a positive voltage increases this current.

Let us conduct a simple experiment. First we apply 100 volts across the cathode and anode (filament and plate). Then we begin to vary the grid voltage as shown, for instance, in Figure 6.6 in a range from minus eight volts to plus five. Using an ammeter, we measure the current in the plate circuit. From this data we can plot the upper curve shown in Figure 6.6. This is called the *tube characteristic*. Next we repeat the experiment with a plate voltage of 90 V. We obtain a similar curve (lower curve in Figure 6.6).

Take note of the following outstanding feature. As is

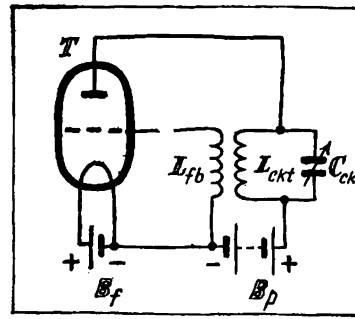


Figure 6.7

obvious from the hatched triangle, we can increase the plate current by 5 milliamperes in two different ways: either by increasing the plate voltage by 10 volts or by increasing the grid voltage by 2 volts. The introduction of the grid makes an amplifier out of the vacuum-tube triode. The amplification factor in our example equals 5 (ten divided by two). In other words, the effect of the grid voltage on the plate current is five times that of the plate voltage.

Now, we shall discuss how a triode enables us to generate waves of a definite length.

This is illustrated by the extremely simplified circuit diagram given in Figure 6.7. When the plate voltage is applied (switched on), capacitor C_{ckt} of the oscillatory circuit is charged through the tube. The lower plate is positively charged. The capacitor is immediately discharged through the inductor L_{ckt}. This produces free (natural) oscillations that would be damped if there were no continuous energy input from the tube. How can we ensure that this energy is supplied at the proper times so that the oscillatory circuit builds up in the same way as the amplitude of a swing when you push it at the right times? This requires what is called *feedback*. The current of the

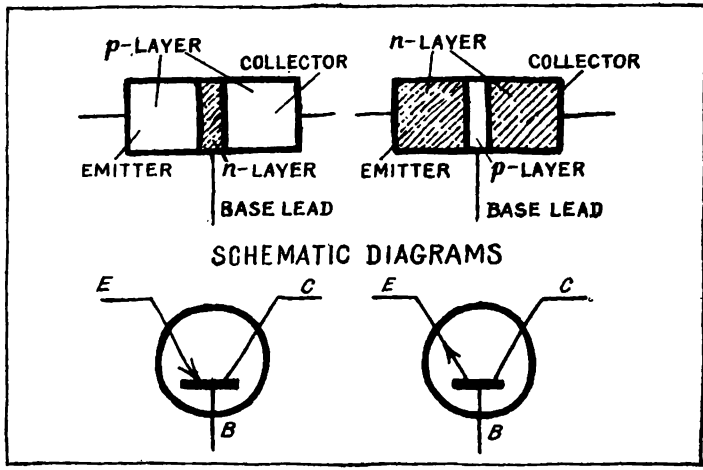


Figure 6.8

oscillatory circuit induces an emf in coil L_{fb} of the same frequency as that of free oscillations. Thus the grid produces a pulsating current in the plate circuit; this current builds up the circuit with its natural frequency.

The two ingenious principles described above are the basis on which radio engineering and allied fields have been established. The electronic tube has become obsolete, making way for the transistor, but the idea of amplifying and generating electromagnetic oscillations has remained the same.

As in a vacuum-tube triode, low power in the input circuit of a transistor can control high power in the output circuit. There is a difference in the way control is accomplished. The plate current of a tube, as we have seen above, depends upon the grid voltage. The current

of a collector in a transistor depends on the emitter current.

But we have not described a transistor yet. It has three electrodes. The emitter corresponds to the cathode, the collector to the plate (anode) and the base to the grid. The lead from the emitter is the input and that from the collector is the output.

As shown in Figure 6.8, the transistor consists of two *p-n* junctions. At the left is *p-n-p* transistor with the *n*-layer in the middle between two *p*-layers. We can also have a *p*-layer in the middle, in which case we have an *n-p-n* transistor (at the right).

We always bias the emitter positively so that it can produce a large number of majority carriers of charges. When the low-resistance emitter circuit changes the current in the high-resistance collector circuit, we obtain amplification.

The ways in which transistors are connected into circuits and employed as amplifiers and generators are similar, in the main, to the principles of the vacuum-tube triode. But we shall not discuss this extremely vital branch of up-to-date physics here.

Radio Transmission

The kinds of radio transmission can be classified on the basis of the power rating of the broadcasting station. Large stations transmit signals at powers ranging up to a megawatt. A miniature transmitter of the walkie-talkie type, employed by a traffic cop to notify his colleague that a Lada car with the license number 31416P has just jumped a stop light and is to be held up, radiates about one milliwatt. Even less power is sufficient for some purposes.

There are essential differences in the layout and design of stations operating with waves over several metres

long and transmitters radiating ultrashort waves with a length of dozens of centimetres or even only fractions of a centimetre. Within each of the wavelength and power ranges, the engineer designing the station can apply any of a huge number of circuits and layouts that may be dictated by the locality, specific aims, economic considerations or, simply, engineering intuition.

The basic unit of a radio transmitter is the radio wave generator, or oscillator. What kind will you employ? You have at least five choices. You can use a vacuum-tube oscillator. Its range is exceptionally wide. Power ratings may vary from fractions of a watt to hundreds of kilowatts, and frequencies from dozens of kilohertzes to several gigahertzes. But if your power requirement is small, of the order of tenths of a watt, only a transistor oscillator will suit you. On the contrary, you will have to reject transistors for the time being (probably not for long) if your power requirement exceeds several hundred watts. If, however, the power rating is such that both types of oscillators can be efficiently applied, the designer will evidently prefer the transistor version. Such an engineering solution is doubtlessly more elegant. A transistorized transmitter occupies considerably less space and, if necessary, can much more easily be designed as a portable model than a transmitter with a vacuum-tube oscillator.

Magnetron and klystron oscillators have more specialized application. The former can be extremely useful in sending pulses of several megawatts into space. The frequency band for which magnetron oscillators can be used is much narrower: it lies approximately between 300 megahertzes and 300 gigahertzes.

Klystron oscillators are used for the same range of ultrashort waves. But they find application only in low power installations: not exceeding several watts in the centimetre and several milliwatts in the millimetre bands.

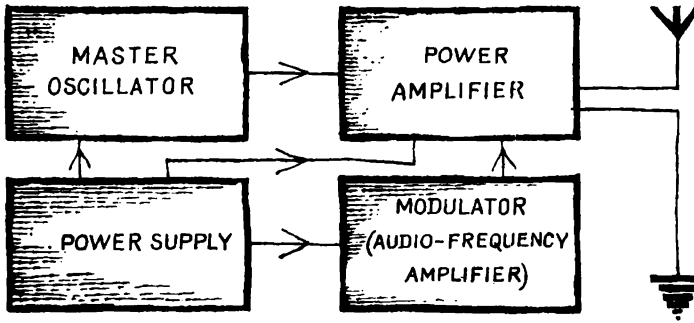


Figure 6.9

The last two types of oscillator, as well as the fifth type, the quantum oscillator, are highly specific and require a special discussion. As to transistor and vacuum-tube transmitters, they resemble each other. There is a clear-cut engineering rule enabling one to replace a vacuum tube with an equivalent transistor.

The choice of the generator of electromagnetic oscillations is by far not all that is required for designing a transmitter. You must decide to how amplify the power produced by the primary (or, as they say, master) oscillator. You must also select a method for modulating the carrier wave by the audio frequencies. There are also many ways for transmitting power to the antenna field. The arrangement of the antenna field itself provides abundant opportunity for exercising engineering ingenuity.

Radio engineers frequently resort to so-called *block diagrams*. Such a diagram consists of several rectangles with captions. The contents of each rectangle is cleared up as required. A block diagram of a radio broadcasting station is illustrated in Figure 6.9. The master oscillator generates continuous, almost harmonic, oscillations of the same frequency and wavelength to which you tune

your radio receiver if you wish to listen to the given station. The second unit is the power amplifier. Its name speaks for itself and we shall not describe its design. The task of the unit called the modulator is to convert sound vibrations into electric oscillations and superimpose them on the carrier wave of the broadcasting station.

Modulation can be accomplished in various ways. The simplest kind to explain is frequency modulation. In many designs the microphone is a capacitor whose capacitance is varied by the sound pressure because the capacitance depends upon the distance between the plates. Imagine now that such a capacitor is connected into the oscillatory circuit which generates the wave. Then the frequency of the wave varies with the sound pressure.

Since we have "invaded" the oscillatory circuit with our microphone, a band of frequencies is transmitted into space rather than a strictly definite frequency. It is sufficiently evident that ideally this spreading should include the whole audio interval of frequencies which, as we know, equals about 20 kHz.

If the station is broadcasting on long waves, corresponding to a frequency of the order of 100 kHz, the passband is about one-fifth of the carrier frequency. It is clear that long waves cannot provide for a large number of non-overlapping broadcasting stations. Short waves are an entirely different matter. For a frequency of 20 MHz, the band width is only a fraction of one per cent of the carrier frequency.

There is probably not a single home in the USSR that has no plug socket for listening to the radio. You receive these broadcasts from the so-called rebroadcasting network. It is also called wire broadcasting.

The first single-program rebroadcasting network was established in Moscow in 1925. The broadcast was transmitted simultaneously through 50 loudspeakers.

Single-program rebroadcasting is carried out on audio

frequencies. From the broadcasting station, the program is transmitted by wire to the central amplifying station. From the central station it is transmitted, again by wire, to the control points, where it is again amplified and transmitted along the trunk feeder mains to the transformer substations. From each substation the wires branch off to substations of the next lower rank. Depending upon the size of the city or region, the number of steps in the network and, consequently, the number of times the voltage is stepped down, may vary. In the subscriber's lines the voltage is equal to 30 V.

Since 1962, three-wire rebroadcasting is being installed in the cities of the USSR. The transmission of two additional programs is accomplished along independent networks by amplitude modulation on carrier frequencies of 78 and 120 kHz. These two broadcasts are demodulated (i.e. the sound is separated out and the high frequency is filtered out) in the home by turning the knob of your Mayak wire-broadcasting plug-in receiver or some other model.

Thus, in three-program rebroadcasting, a single wire carries three programs simultaneously: the main program is on audio frequencies and two are nondemodulated. Therefore, the broadcasts do not interfere with one another. A simple idea, but what excellent results! Economy, reliability and high fidelity of the broadcasts are factors that indicate that wire broadcasting has a great future, including the installation of wire networks for television.

Radio Reception

Radio receiving sets are available in innumerable designs. The field of radio electronics is being developed at an exceptionally rapid rate, so that radio sets soon become obsolete and new items, better than the previous models, are available each year in the shops.

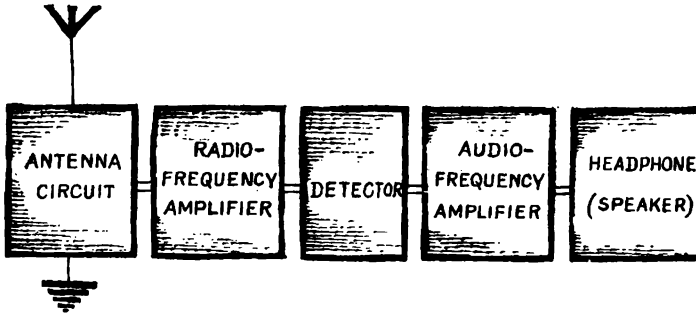


Figure 6.10

What do we mean by "better" with respect to receivers? Each and every reader knows the answer even if he does not understand the physics involved. A good receiving set must separate out of the chaos of radio waves that reach the antenna only those signals that are required. This property is called *selectivity*. A radio set must also be as sensitive as possible, i.e. it should be able to receive even the weakest signals. Finally, it must have high fidelity, i.e. it must reproduce the music and speech broadcasted from the tuned-in station without any distortion whatsoever.

Thus, sensitivity, selectivity and fidelity. We could, perhaps, add one more requirement: the set should operate well on all wave bands.

The block diagram of a receiver with straight amplification is sufficiently clear (Figure 6.10). It is necessary, first of all, to separate out the required wavelength and to amplify the radio-frequency oscillations produced in the antenna by the wave of the broadcasting station. Next, it is necessary to accomplish rectification, or demodulation, which is the name of the process of "discarding" the carrier frequency and sifting out the infor-

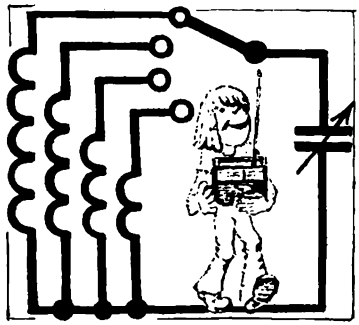


Figure 6.11

mation carried by the sound from the electric current. Finally, it is necessary to provide one more amplifier, but for audio frequencies. The concluding stage is to convert these electric oscillations into sound. This is done by means of a dynamic loudspeaker or by headphones. The latter are used by considerate people who do not wish to be a nuisance to their neighbours.

The antenna of the receiving set is usually inductively linked to the oscillatory circuits of several frequency bands. When we turn the band switch knob, we perform the operation shown schematically in Figure 6.11. Within the limits of each band we usually tune the set by varying the capacitance of the oscillatory circuit capacitor. The capacity of the radio set to select the frequency in the optimal way is determined by the resonance curve of the oscillatory circuit.

I am looking at the specifications in the instruction book of an automobile radio set. Its selectivity is within 9 kHz of resonance for the long- and medium-wave frequency bands. This, of course, is not the limit that can be reached.

The sensitivity of a radio receiver is characterized by the minimum emf in its antenna that enables us to hear

the broadcast with sufficient clarity (I cannot say that this definition is very exact). In the automobile radio set the sensitivity for long waves is better than 175 muV and for ultrashort waves, better than 5 muV.

The sensitivity depends upon the amplification factor and on set noise. The amplification factors of radio sets vary from 10^5 to 10^8. This means that the broadcasting station I want to listen to must produce an induced emf of at least 10^{-8} muV in the antenna of the receiver.

Radio-Wave Propagation

The simplest case is the propagation of radio waves in free space. At a relatively short distance away the radio transmitter can be regarded as a point. If so, then the radio wave front can be assumed spherical. If we imagine several spheres concentrically surrounding the transmitter, it will be clear that, in the absence of absorption, the energy passing through the spheres remains constant. As we know, the surface of a sphere is proportional to the square of the radius. Hence, the wave intensity, i.e. the amount of energy transmitted by the wave in unit time through unit area perpendicular to the direction of wave propagation, decreases as we move away from the source in inverse proportion to the square of the distance.

This important rule is applicable, of course, if no special measures have been taken to obtain a narrow directed beam of radio waves.

Various techniques exist for producing directed radio beams. One way is to use the proper beam, or antenna, array. The antennas should be arranged so that the waves they transmit are sent in the required direction, "hump to hump". Reflectors of various shapes are also employed for this purpose.

Radio waves travelling in space may deviate from the

straight-line direction of propagation by being reflected, dispersed or refracted if they meet with an obstacle commensurate with the wavelength.

Of greatest interest is the behaviour of waves travelling near the surface of the earth. Each case may prove to be entirely unique, depending upon the wavelength.

Of cardinal importance are the electrical properties of the earth's surface and of the atmosphere. If the surface can conduct a current, it does not "release" the radio waves. Electric lines of force of an electromagnetic field enter a metal (or, for that matter, any conductor) at a right angle.

Now just imagine that the radio transmission is near the surface of the sea. Sea water contains dissolved salts and is therefore an electrolyte. Sea water is an excellent current conductor. Consequently, it "holds" the radio waves, making them travel along the surface of the sea.

Plains and timbered areas are also good conductors for currents of not especially high frequency. In other words, plains and forests behave like metals with respect to long waves.

For these reasons, long waves are contained by the whole surface of the earth and are capable of circumventing the globe. Incidentally, the velocity of radio waves can be measured in this way. Radio engineers know that it takes a radio wave 0.13 second to go around the earth. How about mountains? As far as long waves are concerned, mountains are not so high, and a wave a kilometre long can go around a mountain with more or less ease.

The feasibility of long-range reception of short waves is based on the presence of the ionosphere. The sun's rays are capable of breaking up the molecules of air in the higher regions of the atmosphere. The molecules are converted into ions and form several charged layers at altitudes from 100 to 300 km. Thus, the space through

which waves of short length travel is a layer of dielectric sandwiched between two conducting surfaces.

Since plains and forest lands are poor conductors for short waves, they are incapable of holding them. Short waves depart on their journey into free space but encounter the ionosphere which reflects them like the surface of a metal.

The ionosphere is nonuniformly ionized and this varies from day to night. Therefore, short radio waves may travel along various paths. They may reach your radio set after being reflected time after time by the ionosphere and the earth. The actual fate of any short wave depends upon the angle it makes with the ionosphere layer. If this angle is close to a right angle, no reflection occurs and the wave goes through and off into outer space. More frequently, however, total internal reflection takes place and the wave returns to the earth.

The ionosphere is transparent to ultrashort waves. Radio reception of such waves is therefore possible only along a line of sight (with the receiver antenna within sight of the transmitting antenna) or with the aid of communications satellites. If we direct our wave at the satellite, we can pick up the signals reflected from it at enormous distances.

Satellites ushered in a new era in radio communications techniques by making radio and TV reception feasible on ultrashort waves.

Transmission by centimetre, millimetre and submillimetre waves presents extremely interesting opportunities. Waves of this length can be absorbed by the atmosphere. It was found, however, that there are "windows" and, if the wavelength is properly selected, we can make use of waves that are already within the optical band. The advantages of such waves are well known: in a narrow wave band we can "fit in" a huge number of nonoverlapping broadcasts.

Radar

The principles of radar (*r*adio *d*etecting *a*nd *r*anging) are sufficiently simple. We transmit a signal, it is reflected by the object that interests us and returns to our receiver. If the object is 150 m away, the signal returns in 1 microsecond, if the distance is 150 km, it returns in 1 millisecond. The direction in which the signal returns is along a line from the point where the airplane, rocket or automobile was at the instant it met the radio beam.

Naturally, the radio wave must be a pencil beam; the angle of the beam, within which the greater part of its power is concentrated, should be of the order of one degree of arc.

The principle is really not at all complicated, but the apparatus required is far from simple. To begin with, exceptionally high requirements are made to the oscillator. Vacuum-tube oscillators are used for the metre and decimetre bands (longer waves are obviously inapplicable), and klystrons and magnetrons, for the centimetre band.

The pulsed system of operation is evidently the most natural one. Very short pulses are periodically transmitted into space. The duration, or length (of time), of the pulse ranges from 0.1 to 10 microseconds in up-to-date radar transmitters. The frequency of the pulses must be selected so that the echo signal has time to return and be received in the pause before the next pulse is sent.

The maximum range at which an airplane or rocket can be detected is limited only by the condition that the object must be on a line of sight from the radar set. The reader knows, without doubt, that up-to-date radar installations are capable of picking up signals reflected from any planet of our solar system. The waves they

use must, of course, pass unimpeded through the ionosphere. Fortunately, shortening the wavelength also has a direct effect on the range of radar operation because this range is proportional to the frequency of radiation and not only to the energy of the transmitted pulse.

Traces of the transmitted and received pulses are visible on the screen of an oscilloscope (cathode-ray tube). If the airplane is flying toward the observer, the trace of the echo signal moves toward the trace of the transmitted pulse.

Radar installations do not necessarily have to operate with a pulsed system. Assume that the airplane is flying toward the transmitter antenna with the velocity v. The radio beam is being continuously reflected from the airplane. Due to the Doppler effect, the frequency of the received waves is related to that of the transmitted waves by the equation

$$\nu_r = \nu_{tr}\left(1 + \frac{2v}{c}\right)$$

Frequency values are determined with great precision by radio engineering methods. By tuning into resonance, we determine the value of ν_r' and from it we calculate the airplane's velocity v. If, for instance, the frequency of the transmitted signal equals 10^9 Hz and the airplane or rocket is approaching the radar antenna at a velocity of 1000 km/h, the received frequency exceeds the transmitted frequency by 1850 Hz.

The reflection of a radar beam from an airplane, rocket, steamship or automobile is not the same as from a reflector. The wavelength is commensurate with or substantially smaller than the size of the reflecting object which, moreover, is of a complex shape. As the rays are reflected by various points of the object, they (the rays)

interfere with one another and are scattered to the sides. Owing to these two phenomena, the effective reflecting surface of the object differs considerably from its true surface. Calculations here are extremely complicated and only the skill and experience of the radar operator enable him to determine what kind of object is encountered by the radar beam.

You have, of course, seen radar antennas: large spherical reflectors of wire lattice structure, always in motion, surveying space. A great many different kinds of motion can be imparted to the radar reflector. For example, the beam can be moved so that it scans space in lines or circles. With this mode of operation, the path of flight of the airplane can be followed besides determining its range.

This technique is used to home aircraft into an airport under conditions of a total lack of visibility. This job may be done either by a person or even by an automatic device.

A radar set can be "deceived". In the first place, the object can be covered by a material that absorbs radio waves. Coal dust or rubber can be used for this purpose. Moreover, to reduce the reflection factor, the coating may be corrugated so that the major part of the radiation is disorderly scattered in all directions. If packages of aluminium foil strips or metallized fibres are thrown overboard from the airplane, the radar set is completely disoriented. This trick was first used by the English flying forces during World War II. A third way is to fill space with false radio signals.

Radar is an interesting field of engineering and has found extensive applications for peaceful purposes. It is impossible to conceive of defense today without the use of radar.

A competitor of radar is the laser (*l*ight *a*mplification by *s*timulated *e*mission of *r*adiation). The principles of

location and ranging with a laser in no way differ from those described above.

Communication between spacecraft and the earth is based on the principles of radar. Radio telescopes are located so that they keep the spacecraft in view. The antennas of these telescopes are of huge size, some hundreds of metres in diameter. Such large antennas are required so that they can transmit powerful signals and pick up extremely weak signals from a radio transmitter. Of vital importance, naturally, is a narrow radio beam. If the antenna operates with a frequency of 2.2 thousand million oscillations per second (the wavelength being about 1 cm), the beam diverges only to a diameter of 1000 km over the distance to the moon. True, when the beam reaches Mars (300 million km away), its diameter is already equal to 700 000 km.

Television

Since 99 out of 100 readers daily spend an hour or two watching some TV program, it is only fair to say a few words about this wonderful invention. We shall discuss only the principles involved in television broadcasting and reception.

The idea of sending pictures over some distance consists in the following. The picture to be transmitted is divided into small squares. A physiologist can tell us how small the square must be for the eye not to be able to distinguish variations of brightness within it (the square). The luminous energy of each portion of the picture can be converted by the photoelectric effect into an electric signal. Next a way was devised for reading these signals. This is, of course, done in a strictly definite sequence as in reading a book. The electric signals modulate the electromagnetic carrier wave in exactly the same manner as in radio transmission. What happens further

on is also identical to radio communications. The modulated oscillations are amplified and detected. The TV receiver reconverts the electric pulses into a visible image.

The televisor tube in the transmitter is called an image iconoscope, superorthicon or a vidicon. By means of a lens in the TV camera, the image is focussed on a photocathode. The most extensively used are cesium-oxide or cesium-antimony cathodes. The photocathode is mounted in an evacuated tube together with the target screen.

It is possible, in principle, to transmit an image by successively projecting the luminous flux from each element of the image. In this case, the photocurrent should flow only during the short time each element is being transmitted. Such operations would be inconvenient, however, and the camera tube contains a large number of photocells, rather than a single one, this number equalling the number of elements into which the image being transmitted is divided. This receiving screen is called the target and is of mosaic design.

The mosaic target is a thin plate of mica with a large number of grains of silver, insulated from one another and coated by cesium oxide, applied on one side. Each grain is a photocell. The other side of the mica plate is coated by a metallic film. A small capacitor is formed between each grain of the mosaic and the metal, and is charged by electrons emitted from the photocathode. It is evident that the charge of each small capacitor is proportional to the brightness of the corresponding spot on the image being transmitted.

Thus, a latent electric image of the object is produced in the metallic plate. How do we take it from the plate? By means of an electron beam which is made to scan the plate in the same way as the eye passes along the lines of print on the page of a book. The electron beam plays the role of a key switch which, for an instant, closes the electric circuit through a microcapacitor. At the

instant the circuit is closed, the current is uniquely related to the brightness of the image.

Each signal can and should be amplified many times by the ordinary means applied in radio engineering. In transmitting the image, the eye should not be able to distinguish the fact that the electron beam is successively scanning the various spots of the luminous screen. A complete image, obtained on the screen of the kinescope in the TV receiver, during one cycle of electron beam motion, is called a frame. The frames must change at a frequency sufficiently high so that persistence of vision eliminates flickering of the picture being watched.

What no-flickering frame frequency should be chosen? It must be a number related to the current frequency in the mains. The fact is that the pulsating voltage applied to the grid of the cathode-ray tube produces dark and bright lines on the screen. These lines are stationary and imperceptible only if the frequency with which the frames change is equal to or a multiple of the power mains frequency. Continuous motion is seen when the frames change with a frequency of at least 20 Hz; in television a frequency of 25 Hz is taken but a small flickering is still noticeable. It is undesirable to change the frames with a frequency of 50 Hz and therefore TV designers resorted to what is called *interlacing* in the scanning process. The frequency is left at 25 Hz, but the electron beam (also called the *exploring spot*) first scans the odd numbered lines and then the even numbered lines. Thus, the frequency with which the semiframes change equals 50 Hz and any flickering in the brightness of the image becomes unnoticeable.

The frequency with which the frames change and the line-by-line scanning should be strictly synchronized. The technical details of this synchronization are beyond the scope of our book. Consequently, we shall not explain why the number of lines must be odd and consist of

several whole factors. In the USSR the frame is divided into 625 lines, i.e. 5^4. Since 25 frames are changed per second, lines are scanned with a frequency of 15 625 Hz. From this condition follows the width of the band of TV signal frequencies.

The lowest frequency of 50 Hz is the semiframe frequency. The highest frequency is determined by the time required to transmit a single element.

Simple calculations, which we shall not carry out here, indicate that it is necessary to take 6.5 MHz as the higher frequency. It follows that the carrier frequency must be at least 40 or 50 MHz, because the frequency of the carrier wave must be at least 6 to 7 times the bandwidth of the transmitted frequencies. Now we understand why only ultrashort waves can be used in television transmission and why the range is limited to the line of sight from the transmitting antenna.

This, however, is a slip; I should have written: was limited. A breakthrough, enabling TV reception to be carried out at any distance, is the use of communications satellites. The USSR first employed satellites for this purpose. Today the whole of the Soviet Union is covered by communications accomplished by making use of a number of satellites.

Without going into the design of powerful TV stations, we shall only cite some interesting figures that characterize the huge possibilities possessed by up-to-date radio engineering in the amplification of signals. Before amplification, an ordinary video signal has a power up to 10^{-3} W; the power amplifier increases this power by a million times. This pulse of 10^3 W is fed to a parabolic antenna about 30 m in diameter. This antenna produces a narrow directed beam which is reflected by the satellite. After the electromagnetic wave has travelled about 35 000 km to the satellite, its power equals only 10^{-11} W.

The amplifier mounted on the repeater satellite increases the power of this exceptionally weak signal to about 10 W. When this signal, reflected from the satellite, returns to the earth, its power is 10^{-17} W. Amplification returns the power of this video signal to its initial 10^{-3} W.

I think that ten years ago not even the most optimistic engineer would have believed these figures.

Microelectronic Circuits

It is impossible to end this chapter devoted to radio engineering without at least a few words about the new revolution that we are witnessing in this field.

We have in mind the fantastic miniaturization of all electronic devices and apparatus. This became feasible in going over from apparatus built up of separate elements, such as resistors, transistors, etc., connected together by wiring, to electric circuits "drawn" by special techniques on a piece of silicon several millimetres in size.

This new technology (at least one of its versions) consists in using various kinds of masks and a series of chemical compounds enabling *p*-type and *n*-type impurities to be introduced at the required places of a crystal of silicon or germanium. Ion-beam treatment is used for this purpose.

An electric circuit consisting of tens of thousands of elements (!) is arranged on an area with linear dimensions of about two millimetres. When we mentioned the "drawing" of the circuit, the reader may have gotten the impression that the problem involved only some surface treatment of a piece of semiconductor material. This is not so, however. The techniques are much more complicated. Each electronic element is of three-dimensional structure. Several layers containing different

amounts of impurities must be built up on a tiny area of silicon.

How is this done? First a layer of oxide is applied to the surface of the silicon. This is coated with a light-sensitive material. The obtained layer-cake item is exposed to ultraviolet light through a mask of the designed shape. After developing, recesses are formed in the surface of the piece of silicon only at the places where the light passed through the mask.

The next stage consists in treating the electronic circuit being manufactured with hydrofluoric acid. The acid removes the silicon dioxide but has no effect on either the primary surface (i.e. silicon) or the light-sensitive layer. In the final stage another solvent is used to remove the light-sensitive layer. As a result, a piece of semiconductor material is coated with an insulating layer —silicon dioxide—wherever required by the design. The recesses of the required shape are bare silicon. These recesses are treated with an ion beam to dope the silicon with the required amount of impurities.

The manufacture of microelectronic circuits is one of the most rapidly advancing fields of engineering today.

New ideas and new discoveries in semiconductor physics indicate that the fantastic results that have already been achieved are only the beginning.