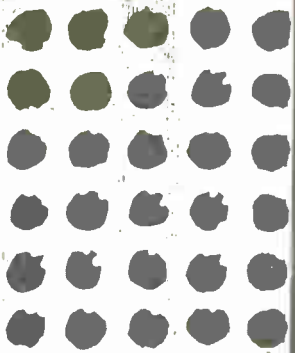


$4.60

FUNDAMENTALS OF

SEMICONDUCTORS

M. G. SCROGGIE, B. S. E. E.

GERNSBACK LIBRARY, Inc., New York 11, N. Y.

Library of Congress Catalog Card No. 60-10767

contents

chapter

page

introduction

Most people are by now aware that transistors are displacing tubes in many applications. But it may still not be fully realized that transistors are only one of the ways in which semiconductors are revolutionizing science and industry. Already semiconductor devices have invaded almost every department thereof, but it is clear that this is only a beginning. Manufacturing difficulties still hold back many developments, but sooner or later these will be overcome, and there is no foreseeable limit to the future of semiconductors.

So much literature is appearing on them that it is impossible to keep pace with it all. Much of it is unintelligible without an advanced knowledge of modern physics and mathematics. At the other extreme there are popular explanations that do not get one very far and insufficiently prepare the mind for more advanced reading. Many people will have to study semiconductors seriously if their potentialities are to be realized.

One purpose of this book is to provide enough theory in a simple way to make it possible to understand more advanced literature. The other purpose is to explain how the special properties of semiconductors are being applied in useful devices of many kinds. These are being invented and developed so fast that the reader must be prepared to keep up to date by means of the technical papers.

It should be understood that not all the devices described here are on the market, and some of them may never be. The law of "survival of the fittest" rules. For example, a major problem in transistor production is to make them work at high radio frequencies. Many solutions have been devised, but it is not yet clear which of those now in the field will win the race, or whether in fact some so-far undisclosed competitor may not succeed.

The purity of materials needed for semiconductor devices far exceeds anything previously considered for manufacture or even scientific experiment. Whereas "99.99% pure" used to be reckoned nearly perfect, in the semiconductor factory it would be thoroughly dirty. Purity of the order 99.99999999% is a practical commercial requirement.

Another difficulty is that the dimensions of high-frequency semiconductor devices are far smaller than anything hitherto considered practical in mass production. Parts have to be made precisely to a few millionths of an inch. This fact must not lead one to suppose that semiconductor techniques are concerned solely with the microscopic; some of the products handle thousands of amperes and volts, and the heavy branches of electrical engineering are likely to be very considerably influenced before long.

After a general survey in the first chapter, the next two start on the theory, which means getting down to the basic ingredients of the universe — atoms and energy. These concepts are next applied to a typical semiconductor material to explain its peculiar behavior. Combinations of two slightly different kinds of semiconductor are then considered, still on a theoretical basis. The remaining four chapters show how these properties are being applied, in rectifiers, transistors, light-sensitive cells, and a host of other devices.

M. G. Scroggie
London, England

chapter 1

a wide-open prospect

THE use of semiconductors is expanding right now almost explosively. No limits can be seen to future developments. Already transistors have largely or altogether displaced vacuum tubes in hearing aids, portable radios and electronic computers, and that is only a beginning. All-transistor television receivers have been produced, and will no doubt be commercialized when manufacturing difficulties have been overcome. Similarly with tape recorders.

Just as the original vacuum tube has developed into a multiplicity of -odes and -trons difficult even for the electronic engineer to keep up with, so have transistors — only faster. Besides displacing tubes they are opening up new possibilities of their own. For instance, they can be used to step up the voltage of dc, as an alternative to rotary machines. They make possible the production of portable versions of test gear and measuring instruments. And their ruggedness and small size and power economy are invaluable in aircraft and, especially, guided missiles.

But transistors are far from being the only semiconductor devices. Rectifiers are less glamorous, but essential in most electronic equipment. They come into heavy electrical engineering, too; electric railroads, for example. Modern research into semiconductors has resulted in button-sized rectifiers handling power that would have seemed incredible only a few years ago.

Then there are photoelectric or light-sensitive devices. Here again semiconductors are tending to displace and supplement tubes. Most people are familiar with them in photographic ex-

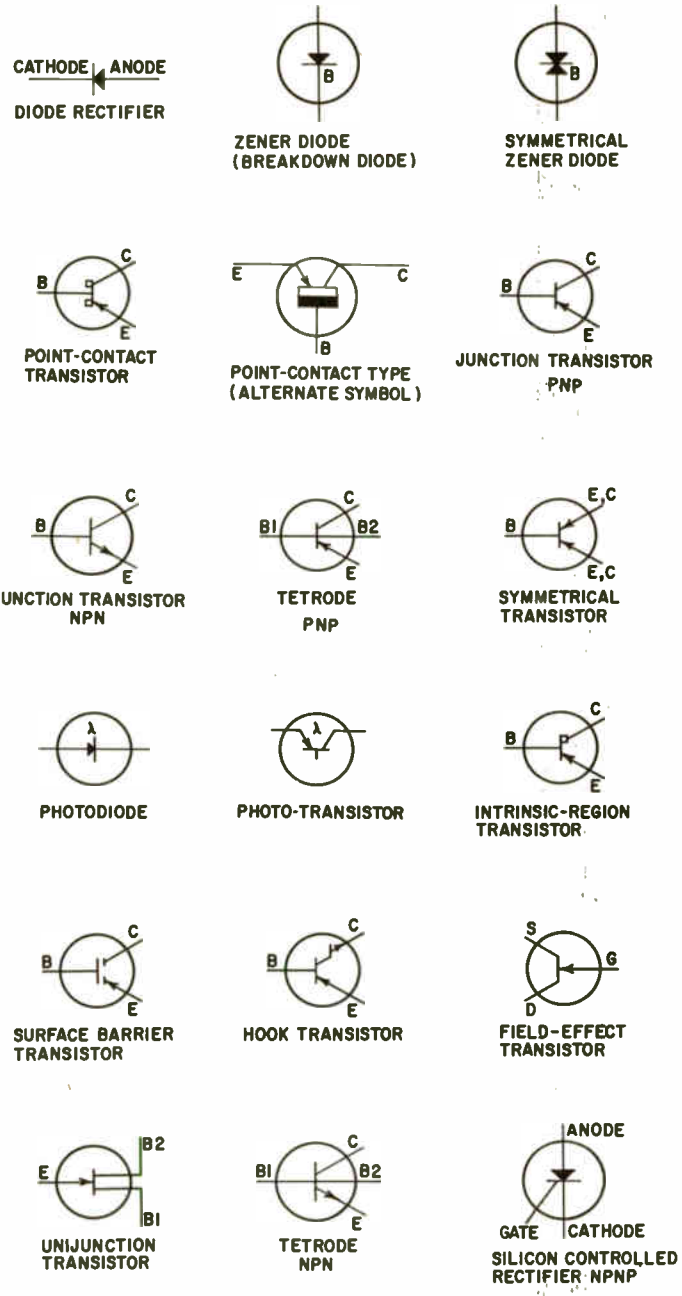


Fig. 101. *These are some of the types of semiconductors that have been developed.*

posure meters. A bare list of their other applications in the armed forces, industry and research would take up too much space. Drawing power from the sun to energize radio transmitters and other equipment in space rockets, and "seeing in the dark",

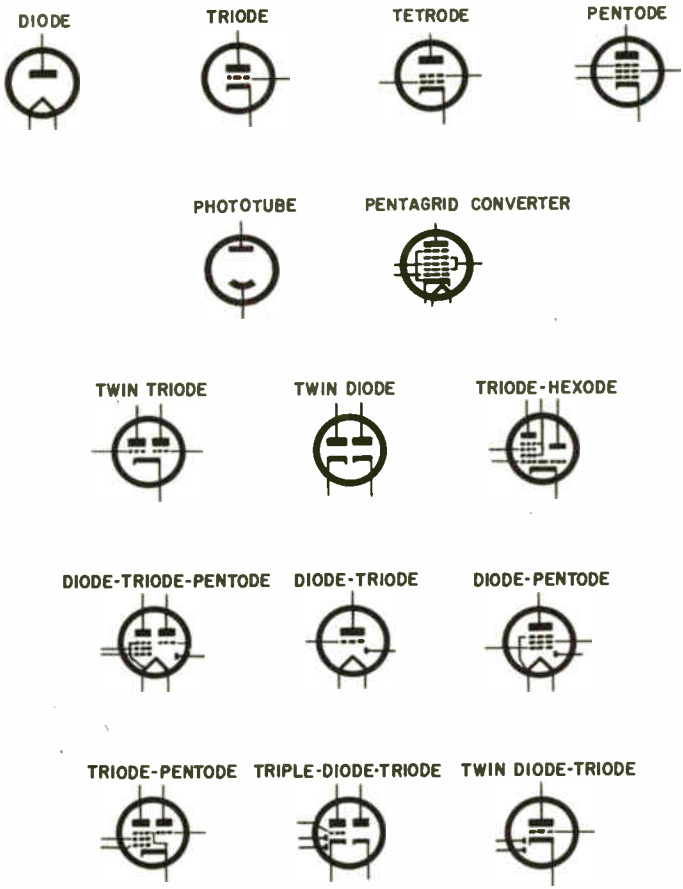


Fig. 102. *Multielement tubes have the diode as a common ancestor.*

are just two of the more spectacular applications. Even in television, with its high signal frequencies, semiconductors have begun to displace photoemissive surfaces in camera tubes, where the light patterns of the scene are converted into electrical signals.

Lastly, there are semiconductor devices falling into none of these three classes. There are thermistors and varistors, with a growing list of applications. There are Hall-effect devices; among

dozens of uses, they make possible supersensitive navigational compasses. Other semiconductor properties can be used for refrigeration, for generating electricity directly from heat, and for increasing the range of radio telescopes. Electroluminescence is yet another phenomenon with increasing uses. And even electronic tubes themselves have to rely on semiconductors for phosphors and low-temperature emitters.

Semiconductors are not new

All this may sound very modern. And certainly it is mainly since World War II that semiconductors have been studied and developed on a large scale. But they are not new. The rectifying properties of certain solid substances were discovered as long ago as 1835, by Munk Af. Rosenshold. This seems to have been forgotten and had to be rediscovered, which F. Braun did in 1874. Another interesting date is 1873, when a technical assistant of Willoughby Smith, testing underwater telegraph cables, found that the high resistance he was using for the purpose varied considerably according to the amount of light shining on it. This resistor, requiring many megohms of resistance, had been made of a semiconductor — selenium. And so it was discovered that selenium conducts electricity better in light than in darkness. Selenium cells were thus among the wonders demonstrated by Victorian exponents of popular science. Another interesting property of selenium is rectification, first noticed in 1876 but not utilized until much later.

Coming to the twentieth century we take note of 1904, because it saw the radio race between vacuum tubes and semiconductors start almost evenly. Fleming produced his thermionic diode as a detector, while J. C. Bose patented a semiconductor crystal diode for the same purpose. The Fleming diode was displaced a few years later by the de Forest triode, but returned to favor in the 1930's and is commonly used to this day. Meanwhile, however, the crystal detector was for a time even more popular, during the early years of broadcasting. It, too, staged something of a comeback. Improved for better reliability, it became a key component in radar receivers during World War II and has held that position ever since. The particular semiconductor used is one associated with the latest types of transistor — silicon. However, a silicon radio detector was invented by Pickard back in 1906. He called it the "Perikon" detector — a name that will be remembered by old-timers in connection with a better combination of minerals he patented in 1909.

These were all what we now call point-contact rectifiers. The first junction diode is dated 1941, and these types are now commonly used in radio and TV receivers, among other things. In view of the fact that only three years elapsed between the introduction of the Fleming diode tube and its elaboration into a triode, it is surprising that the corresponding development of the semiconductor diode had to wait 34 years. Then, in 1948, Bardeen and Brattain of the Bell Telephone Laboratories produced the first crystal triode and named it a transistor. Because of the amplifying properties of a triode — either tube or transistor — it can generate oscillations. All readers will be aware of that, but some may be surprised to learn that these things were accomplished with semiconductor *diodes* in 1924 by a Russian, Lossev*. In fact, the use of a crystal diode as an oscillator goes back even further, to Eccles in 1909.

Meanwhile, semiconductor rectifiers for purposes other than radio detectors were having a history of their own. In 1926, Grondahl's work resulted in the copper-oxide rectifier, still used though largely superseded by selenium. Selenium, in turn, will probably give way to silicon. Germanium is a material linked in the mind with transistors, but in fact was used for rectification from 1925.

Further outlook

Those were just a few of the landmarks in the history of semiconductors in the service of man. One thing is certain: this history will run at an even more rapid pace in the next 10 or 20 years. The natural trend of most things is toward the "bigger and better". And certainly there has been a steady increase in the size and consequent power-handling capacity of transistors. But perhaps an even more promising trend is in the opposite direction. One of the great attractions of a transistor as compared with a tube is that it can be made smaller. This feature has given deaf people hearing aids concealed in spectacle frames and other convenient and unobtrusive places.

Transistors fit well into the current development of printed circuits. They are usually on flat surfaces like cards. But already there is a move towards "solid" circuits, in which the one-piece wiring is in three dimensions, formed with the components into a more or less solid block. In this way, a very large number

* Victor Gabel, "The Crystal as Generator and Amplifier", *Wireless World*, Oct. 1, 1924, page 2.
"The Crystodyne Principle", *Radio News*, Sept. 1924, page 294.

of units or functional "cells" can be built into a small space. In fact, the structure of electronic equipment is getting more like that of living organisms, and one can foresee the production of compact electronic brains and robots, not as science fiction but as a discernible extension of current trends.

Fig. 101 gives an indication of the "evolution" of transistors. Compare this with the development of multielement tubes (see Fig. 102) starting with the diode.

An enormous expansion in the use of semiconductors is assured, then. That means a lot of people will have to know a lot more about them. For a start, what exactly *are* semiconductors? And why do they work the way they do? To understand this, it will be necessary to inquire into the basic structure of solid materials. The following chapters point the way.

chapter 2

beginning with the atom

A FAINT idea of the variety of semiconductor products and performance can be gathered from the previous chapter. To give you a full idea, it would have to be supplemented by innumerable data sheets, specifications and application reports. Yet most semiconductor devices are very simple in construction. How can they provide such variety of behavior?

The answer lies in the peculiar ways in which semiconductors conduct electricity. The name suggests — quite correctly — that they conduct less than good conductors (metals) and more than insulators. Actually, a typical semiconductor conducts about a million times less than metals and a quadrillion (10^{15}) times more than insulators. But that alone certainly would not account for the results obtainable. Anyway, why *do* some materials conduct well, some hardly at all and some intermediately? What is electrical conduction?

Simple story not enough

So long as semiconductors could be left out of it, quite a simple story would do. It goes something like this. Solid substances are made up of atoms which occupy fixed positions in the material. Each atom has a number of electrons buzzing around it, like satellites. Nearly all the atoms of insulating materials have their teams of electrons permanently assigned to them, so that hardly any electrons are free to roam through the material in response to an applied electric field. Metals, on the

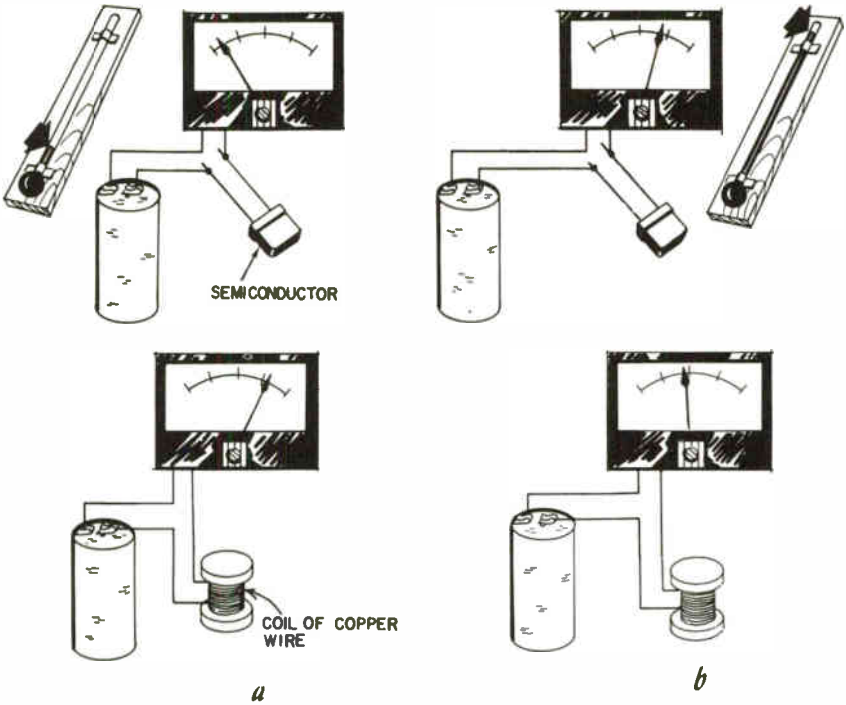
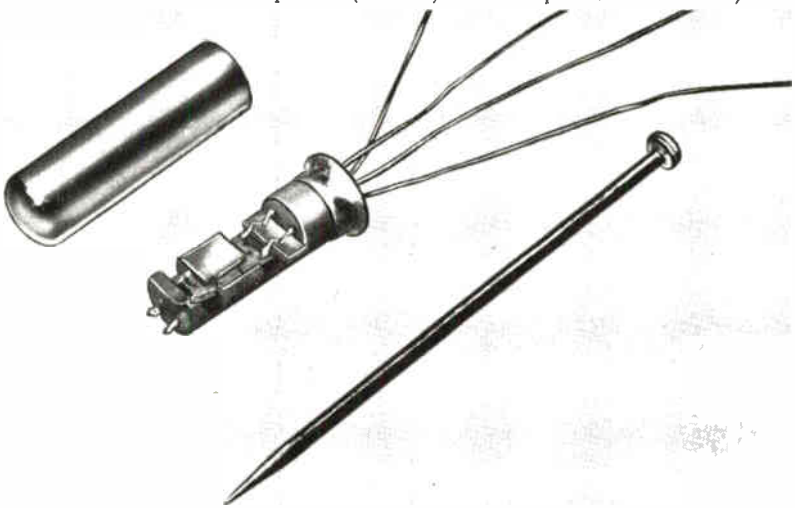


Fig. 201. *At a low temperature (a) the semiconductor conducts very little and the conductor freely. When the temperature is raised (b) the semiconductor conducts much more and the conductor slightly less. The temperature sensitivity of a semiconductor can have a useful application, as in the germanium resistance thermometer shown in the photo.* (Courtesy Bell Telephone Laboratories)



contrary, have one or two electrons in each team free to move from atom to atom through the material. This movement of electrons is an electric current. In vacuum tubes, the situation is even simpler, because the electrons are entirely on their own.

That is about all the fundamental electronic physics many professional electrical engineers have to this day, and they get

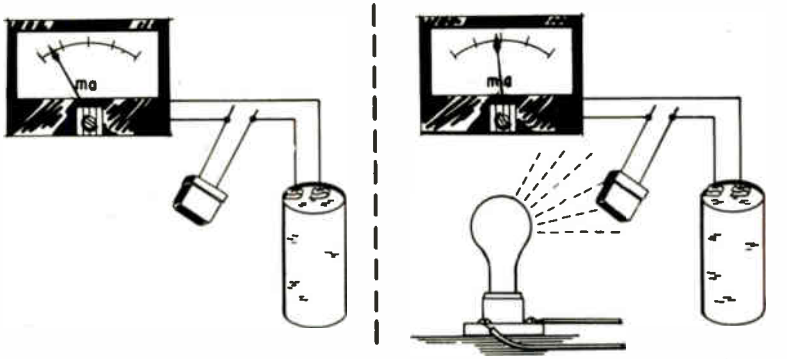


Fig. 202. *Many semiconductors conduct more freely when light shines on them.*

along quite well on it even though it is only partly true. But semiconductors cannot just be tacked on to this story as a supplement by saying they have fewer movable electrons. That again may be partly true, but it is quite insufficient to account for the behavior of semiconductors; for example, why their conductivity varies so much more (and in the opposite direction) with temperature than that of metals (Fig. 201), why it depends on light (Fig. 202) and why simple combinations of semiconductors rectify (Fig. 203). We will have to go into the subject in more detail and learn much that may seem theoretical and even far-fetched. But this is necessary not only for understanding semiconductors but many other branches of modern applied science, so the effort will be worth while.

An example of how the simple story fails is the fact that copper conducts electricity about 10^{22} (that is, ten thousand million million million) times more than polyethylene, but polyethylene needs nothing remotely like that much higher voltage to detach its electrons from their atoms.

Molecules

If a giant Martian were to land on earth and start pulling a house apart to find out what it was made of, he would discover

the answer was mainly bricks and planks of timber. He might conclude that they were the "atoms" of which all these strange terrestrial growths were made.

We, being smaller creatures, can pulverize the bricks and timber and reduce them to fine dust. But, except for size, they would still be the same materials.

If the process could be carried much further still by microscopic creatures, they would reach a stage where any further pulling apart would change the nature of the substance itself. Table salt, for example, would cease to be salt altogether; some of the pieces would make a soft shiny metal and other pieces a poisonous gas. The smallest possible pieces before drastic changes like this take place are called *molecules*.

Such submicroscopic examination would reveal that apparently smooth continuous solids like glass are made up of individual molecules. And that even liquids and gases are not what they appear to be but consist of molecules. The differences are that in solids the molecules maintain fixed positions like soldiers in a guard of honor, in liquids they are still close together but free to move like the same soldiers streaming into the canteen, and in gases they are still freer to spread out into the whole space available like soldiers on annual leave.

Atoms

Molecules can be broken down into a number of subassemblies called *atoms*. Some molecules consist of only two atoms. Salt is an example of this; each of its molecules consists of one atom of the metal sodium and one atom of the gas chlorine. Most of the "chemicals" have less than a dozen atoms to a molecule, but the materials forming living creatures are often more complicated. Some of their molecules have hundreds or even thousands of atoms. It is a bit of luck for us that semiconductor molecules are among the simplest, the most important of them having merely one atom per molecule.

The word "atom" means "indivisible", and, until nearly the twentieth century, that was just what an atom was thought to be. As such, it was naturally permanent and unchangeable; there was no hope of realizing the old dream of changing base metal into gold. It is now known that atoms themselves can be divided into still smaller parts, and that these parts can be rearranged to form different atoms. Changing one metal into another (though not lead into gold!) is now commonplace. On that discovery rests the whole science of nucleonics and the release of atomic

energy. These are right outside the scope of this book, and for our present purpose we can regard every atom as an unchanging central core or nucleus surrounded by a variable number of electrons.

The nucleus

The most important thing about any atom is the number of units of positive electrical charge it has in its nucleus, because — surprisingly — it is that number which decides what it is an atom of.

An atom with only one of these positive units or particles, called *protons,* is the smallest possible quantity of the substance we know as hydrogen — a very light and inflammable gas. Atoms with 13 protons each are recognizable as the metal aluminum; those with 15 are the inflammable solid phosphorus; those with 32 the semiconductor germanium; those with 80 the liquid metal mercury, and so on. This decisive number of protons per atom is

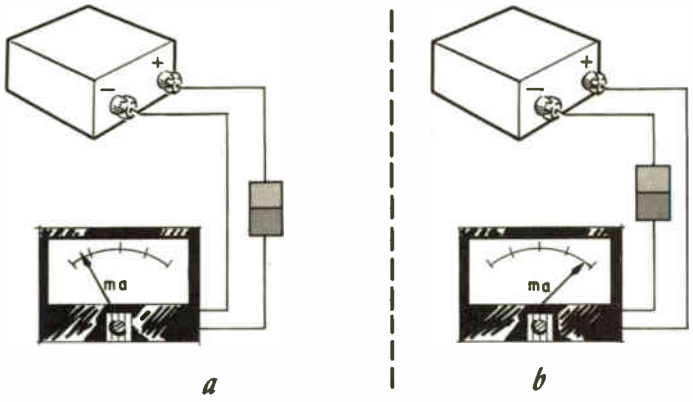


Fig. 203. *Contacts between two semiconductors, or a semiconductor and a metal conduct current much more freely in one direction (b) than the other (a).*

called the *atomic number.* Atoms having nearly every atomic number from 1 through 92 have been discovered in nature; the few missing ones and also several above 92 have been made artificially.

Substances consisting of only one of these kinds of atoms are called *elements.* Obviously very few out of the thousands of known different substances can be elements, and of the one hundred elements the majority are quite rare. Most substances are combinations of a comparatively small number of elements,

such as carbon, oxygen, hydrogen, silicon and calcium. Atoms of two or more different elements often club together to form molecules of what are called compounds, and in doing so they completely lose their own identities, as for example, sodium and chlorine when they form salt.

Although things can be mixed together in any proportion, they can only combine — if at all — in certain fixed proportions. Air is just a mixture of oxygen and nitrogen; they do not combine. Oxygen and hydrogen can also be mixed together to remain a mixture in any proportion. But a lighted match or spark starts a violent combining action in which each oxygen atom unites with two hydrogen atoms to form a water molecule. If these were not the proportions in the mixture, some oxygen or hydrogen would be left over.

Often two — and occasionally three — atoms of the *same* element join to form a molecule of that element. The oxygen in the air we breathe is an example. In those cases, the substance is still

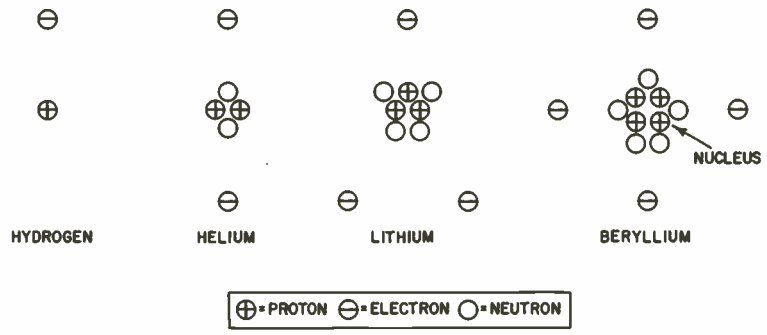


Fig. 204. *These diagrams show the parts that go to make the atoms of the first four elements.*

basically the same, but may differ in less obvious respects.

Besides protons, the nucleus of an atom includes a roughly equal number of electrically neutral particles called *neutrons*. The number of neutrons in an atom can vary within fairly wide limits; this makes no difference to the nature of the substance except that atoms with certain numbers of neutrons are unstable and tend to rearrange themselves violently, like a woman with a spider down her back. We say these substances are radioactive. The single proton in the hydrogen atom's nucleus is sometimes accompanied by one or two neutrons, but most often by none (Fig. 204). Although for some purposes the number of neutrons per nucleus is vitally important, so far as this book is concerned

we can forget about neutrons. The essentials about the nucleus are its atomic number — the number of protons or electrically positive particles it contains — and the fact that nearly all the atom's weight is concentrated there.

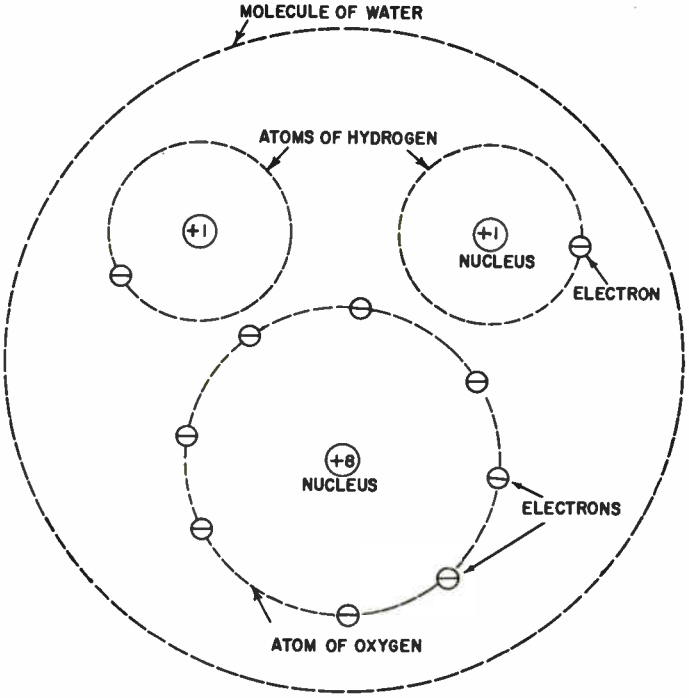


Fig. 205. *Diagram—not to scale—of the parts in the smallest possible quantity (molecule) of water.*

Electrons

In the normal state of an atom, the electrical effect of these positive charges is exactly cancelled by an equal number of negative charges around the outside of the nucleus (Fig. 204). These negative charges are well-known to us as *electrons,* and together with the nucleus make up the whole atom. Although, electrically, one electron exactly counterbalances one proton, the proton weighs as much as 1,836 electrons. That is why the electrons, equal in number to the protons, are such an insignificant part of the weight of an atom.

Nevertheless, it is the electrons that are responsible for certain combinations of atoms uniting to form molecules. They are re-

sponsible for the variety of characteristics of elements and compounds, such as hardness, color, poisonousness, inflammability, etc. They are responsible for all the electrical effects, for radiation of light and, in fact, for most of what goes on in the universe.

The odd thing is that one or two electrons can be added to an atom, making it electrically negative, or taken away, making it positive, without otherwise affecting the nature of the atom. For example, an atom of oxygen has eight protons in its nucleus, and that is what decides it is oxygen, even though its oxygen properties are due to its electrons, which can be more or less than the normal eight.

Summary

Fig. 205 is a diagram of a molecule of that not-unfamiliar substance, water. It shows that the molecule is made up of lesser structures — atoms — and these in turn are made up of nuclei and

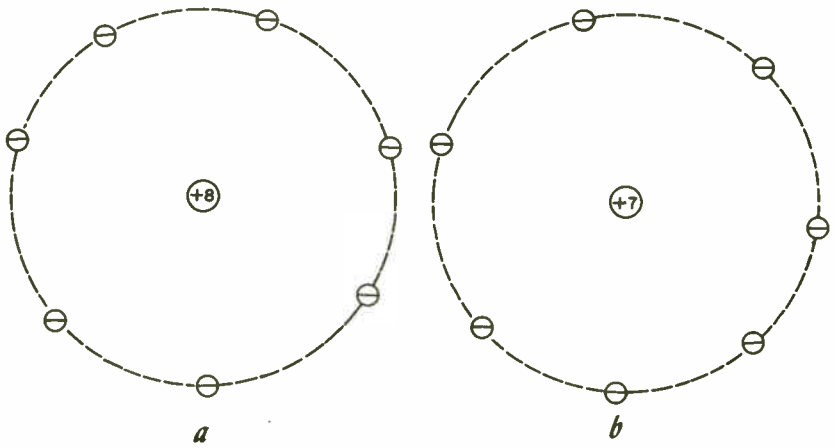


Fig. 206. *Although both these atoms have seven electrons, which for most purposes are the working parts, the way they work is decided by the amount of electric charge on the nucleus. Eight positive units (a) means oxygen, and seven (b) nitrogen.*

electrons. The nuclei could be further subdivided into protons and neutrons as in Fig. 204, but for our purpose only the numbers of protons — which are measures of their positive electric charges — need be shown.

These are the numbers that decide whether, for example, an atom is oxygen or nitrogen — a very vital difference to us as creatures that breathe! Fig. 206-a shows an atom of oxygen from which one electron has been removed — a common occurrence in those upper layers of the atmosphere called the ionosphere. It has

the same number of electrons as the atom of nitrogen shown near it (Fig. 206-b). Yet, although it is the electrons that are responsible for its life-maintaining action in the lungs, it is still oxygen, because its nucleus has eight protons whereas the nitrogen atom has only seven.

Fig. 207. *Three examples of how things "stay up" in spite of attraction: (a) earth round sun, (b) water round man and (c) electron round nucleus.*

Although these diagrams show how many there are of each kind of part, they should not be taken too seriously concerning the exact arrangement of those parts. That is an important and complicated matter, which we will now consider.

Inside the atom

The first and most obvious question is why the electrons are shown floating about at a distance from a nucleus little bigger than themselves. Since positive and negative electric charges attract one another, and the nucleus weighs 1,836 times as much as all its associated electrons together, one would expect these electrons to be sticking closely to a relatively large nucleus like fleas on a dog. But experiments show that atoms even of solid materials are almost entirely empty space. It is as if the nucleus had the weight of a dog but the size of a flea, and the space far around was very thinly occupied by a cloud of gnats.

The problem, then, is to explain what keeps the electrons at such relatively great distances from the nucleus against the force of electrical attraction. On a vastly greater scale, the same situation exists in the solar system, where a number of planets keep their distances from the central sun in spite of gravitational attraction. It is plausible to suggest that the same explanation

holds good for the atom — that the electrons are revolving around the nucleus so fast that centrifugal force exactly balances the attraction (Fig. 207). But modern science teaches that one cannot apply to particles of subatomic size the same principles that hold so accurately in astronomy or even on the scale of grains of sand.

It appears that electrons are quite different from very small grains of sand. They require more difficult mathematical treatment than planets or satellites and their movements cannot be clearly visualized. For a simple study like ours, however, we can stick to the old revolving-planet idea of electrons, as long as we are prepared to accept without question some rather strange additional rules that govern their behavior. To understand them, we must have a clear picture of two fundamental things — energy and matter.

chapter 3

energy and matter

Most people think energy is the thing they haven't much of on Monday mornings. And certainly that fits the school definition of energy as "capacity for doing work". But this sort of work may not be quite what we had in mind. In the scientific sense, it means movement against force or pressure or resistance.

Energy

Lifting a 10-pound weight 5 feet does 50 foot-pounds of work on it (Fig. 301-a). The process gives the weight 50 foot-pounds of potential energy, which means that it has been put in a position to do 50 foot-pounds of work. If it fell back freely — without any resistance — it would very quickly lose all its height and, therefore, all of its potential energy, but this energy would not have been lost. It would have all been converted into an equal amount of kinetic energy — energy of movement.

As soon as it touched the floor, it would move a short distance against considerable resistance, making a dent in the floor or whatever it fell on (Fig. 301-b). The energy would now appear in the form of a small amount of heat. Alternatively, the falling weight could have been coupled to a small dynamo and used to generate electricity. This electrical energy could itself be converted into heat energy in a lamp or resistor, or used to make current flow through the coils of an inductor, storing up energy as a magnetic field. Energy, then, can change its forms, but never goes out of existence. You can change dollars into francs, and then if you do not need them you can change them back into dollars.

You are likely to find that owing to fluctuating rates of exchange you have fewer dollars than you had. But the rates of exchange between different forms of energy are exact and unalterable.

How to launch satellites

In the solar system which on page 21 we took as the large-scale model of an atom, a certain amount of energy has to be given to each planet to get it going in its orbit. It needs both potential and kinetic energy – potential, because of its weight and the distance it would fall to the sun if it were not for its kinetic energy, which it has as a result of its speed around the sun.

We are becoming familiar with this situation, because it is precisely the problem of getting an earth satellite into orbit. To keep itself above the ground at a height of a few hundred miles, the satellite has to circle around it at about 18,000 miles per

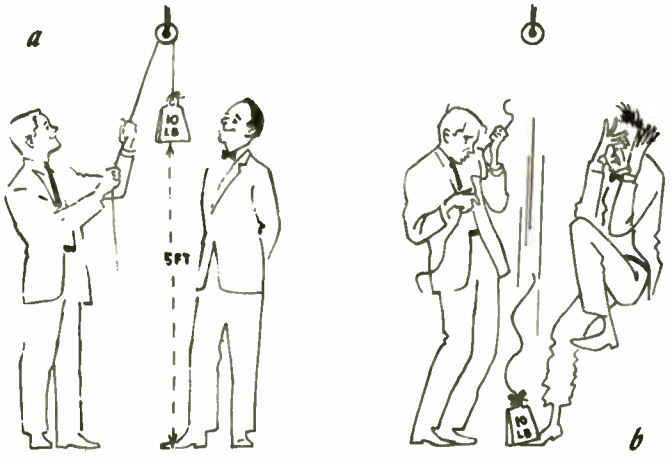


Fig. 301. *(a) A 10-lb. weight, having been raised 5 feet, possesses 50 ft.-lbs. of potential energy. (b) The weight being now at zero height has no potential energy, but the energy has not been lost. It has been transformed into kinetic or motional energy, which is available for producing various results.*

hour. So it has to be given the kinetic energy represented by that speed, plus the potential energy represented by its distance from the earth (Fig. 302)*. The boy in Fig. 207 has essentially the same problem if he is not to get wet.

To lift a satellite from its present orbit into a higher one would

* In practice, the problem is greatly complicated by air resistance on the way up; otherwise it would be an easy calculation!

necessitate giving it an extra amount of energy. There is an exact mathematical tie-up between the gain in energy and the resulting increase in distance. One could put a satellite into any orbit by giving it the right amount of energy. Its height could be increased by a fraction of an inch by giving it a little extra energy, or by 1,000 miles by giving it a lot of extra energy. There is no restriction on choice.

"Of course not," you may say, "Why should there be?" With an earth satellite or a planet, no reason at all; but when we come to atomic satellites — electrons moving around a nucleus — we must remember the warning on page 21 that they are bound by some curious rules.

The book of rules

The first rule is that only certain orbits are allowed. The difference between the two systems is like the difference between a

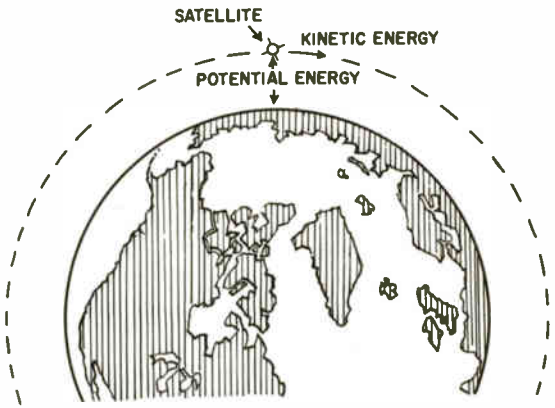


Fig. 302. *Getting a satellite into orbit means giving it sufficient energy, which is partly in potential form and partly kinetic.*

tuning control and a bandswitch: the tuning knob can be turned to any position, but the switch clicks into a limited number of fixed positions — it has to move in whole steps or not at all.

You can see how extraordinary this rule is if you consider the mathematical connection between size of orbit and energy. If only certain sizes are allowed, then the electron cannot accept just any amounts of energy that are offered, but only those amounts that would raise it into one of the permitted larger orbits. In the same way, it cannot give up its energy gradually

or to any extent it has in stock, but only in amounts that would lower it into one of the permitted lesser orbits.

The second rule is that unless the electrons are continually being offered the right amounts of energy to keep them in one of the larger orbits, they tend to fall down into the smallest permitted orbit, nearest the nucleus.

The third rule prevents most of the electrons from obeying the second rule, because it absolutely forbids more than two electrons to occupy the same orbit, and even those two have to be different from one another by spinning like tops in opposite rotations.

You may think these sound like the silly sort of rules they have in some institutions; the kind that seem to be made just to prevent you doing what you want. But in fact the universe and life could not go on for a moment without them. All the electrons would just fall straight into their nuclei, and that would be that.

These rules are at the bottom of all the amazing differences between different substances. The whole subject is tremendously complicated, so we shall study only as much of it as is necessary for the purpose of this book.

An atomic structure

Take a familiar substance — copper. In accordance with the rules just mentioned, the 29 electrons in each of its atoms are arranged around the nucleus in the 15 lowest-energy orbits (Fig. 303). At least, that is the setup when the atom is not being disturbed by energy from outside or by neighboring atoms.

In real life, it doesn't have such a sheltered existence. It is commonly exposed to a whole range of electromagnetic radiations, from radio waves at the lowest frequency end, through heat, light, ultra-violet, X-rays and gamma rays, to cosmic rays. Then in some situations, such as tube electrodes, atoms are liable to be violently bombarded by other electrons. Elsewhere, they are subjected to the usually gentler influences of electric fields. We shall consider the last of these first.

Electrons 1 and 2, nearest the nucleus, are tightly bound to it by the electric attraction of opposite charge, +29 units strong. Only an intense amount of energy could dislodge them over the heads of all the other 27 electrons.

No. 29, in contrast, can hardly feel the attraction of the comparatively distant nucleus, which is largely shielded by the repulsion of the inner 28 electrons. So it needs very little incoming energy to make it drift away. In a piece of copper, consisting

of many atoms, the outermost electron in each is so loosely attached that it cannot really be said to belong to any one atom more than another. It readily accepts even the smallest bribe of energy to transfer its allegiance.

Electrical energy

As we saw at the beginning of this chapter, mechanical work is calculated as the distance a thing is moved multipled by the force it is moved against; for example, in foot-pounds. The same meas-

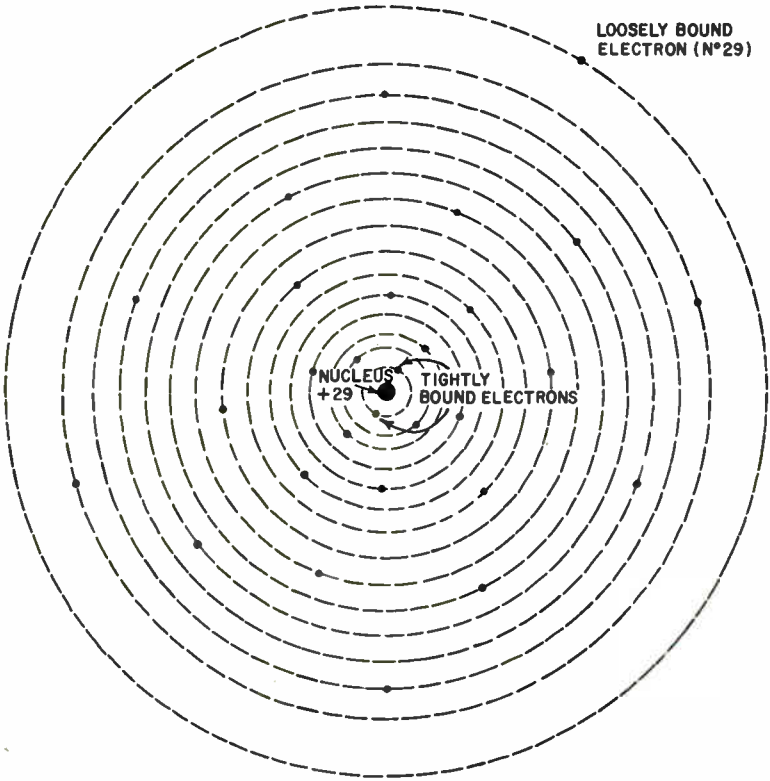


Fig. 303. *This diagram shows the number of electrons around a single neutral atom of copper, and also shows that no more than two can have equal energy at one time. But it must not be regarded as a lifelike picture of an atom.*

ure can be used for energy, in which case it means the amount of work the energy can do.

In electrical work, an electric charge takes the place of the weight in Fig. 301, and potential difference (voltage) takes the

place of difference in height. To say that 1 ampere is flowing in a circuit means that a charge of 1 coulomb of electricity is being moved every second. Suppose the source of the current is a 2-volt battery. What that battery is doing in every second is to lift 1 coulomb 2 volts higher in potential. We could say it is doing 2 coulomb-volts of work per second, by expending energy at that rate. What we are more likely to say, meaning exactly the same thing, is that the battery is working at the rate of 2 watts.

In this book, we will often be concerned with what individual electrons are doing, rather than the 6,240,000,000,000,000,000

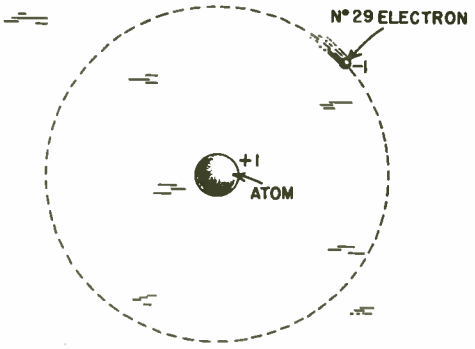


Fig. 304. *For our purpose, all except the outermost electron in Fig. 303 can, for simplicity, be lumped together as a single unit with one net positive charge.*

electrons that make 1 coulomb. So instead of the coulomb-volt, our unit of energy will be the correspondingly smaller electron-volt (eV). It has the great advantage of telling us the energy gained or lost by an electron in terms of its gain or loss in voltage. It gains when it moves against the voltage to a higher potential (compare Fig. 301) and loses when it moves with it.

The innumerable outermost electrons in a copper wire need so little energy to move them that the smallest fraction of a volt distributed among them is enough to set the whole lot drifting towards the positive end. Copper is, as we say, a good conductor.

Mortar for the bricks

Conducting electricity is by no means the only job these 29 electrons do. They hold copper atoms tightly together into a solid mass of metal. You might think that because two electrons, being "like" charges, repel one another, they would not be very

effective agents for welding their respective atoms together—especially as loyalty does not seem to be their strong point.

For simplicity, we can replace the formidable-looking copper atom diagram of Fig. 303, with its nucleus carrying 29 units of positive electric charge and its 29 electrons each carrying one unit of negative charge, by Fig. 304. In any case Fig. 303 is not accurate — the pattern is really far more complicated — and for our purpose all but the outermost electron can be lumped together with the nucleus as "the atom." Electrically, its charge is $+29 - 28 = +1$ unit, which is neutralized by the one electron shown.

When two such atoms come close together, each electron is

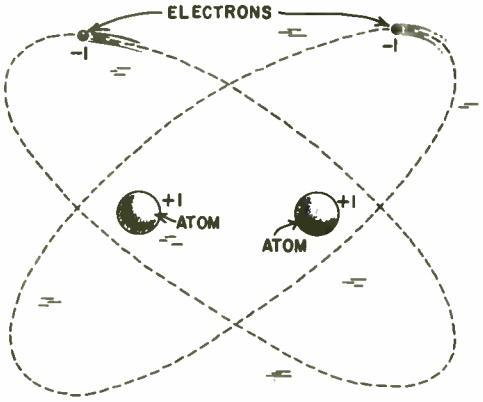


Fig. 305. *In a molecule, two atoms are held together by forces arising from their two valence electrons in orbit around them.*

attracted to some extent by both atoms, and at a certain critical spacing the whole system "clicks" into a single stable structure, with both electrons encircling both atoms (Fig. 305). This is hard to explain or visualize, but you can perhaps imagine a square-dance figure in which two boys form a stable group with two girls even though they are continually changing partners.

The principle applies to copper atoms in bulk; they come together naturally into a regular crystalline structure, like innumerable tennis balls packed tidily in a crate. The crystalline nature of copper can actually be seen by closely examining a broken piece. Of course, one can't see the individual atoms; only the regular flat surfaces of the "crates".

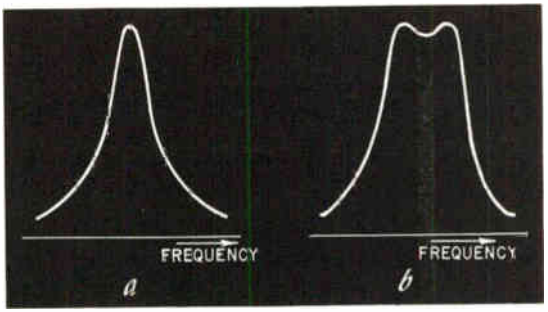


Fig. 306. *A single tuned circuit has a single resonance curve like (a). When two circuits are coupled together, the resonance peak splits into two, as at (b).*

Valence electrons

Still another way in which these outermost electrons work for their living is by joining their atoms in marriage with atoms of another element. If hot copper is exposed to oxygen, the two kinds

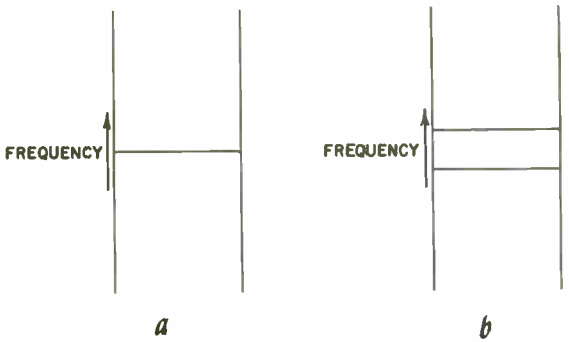


Fig. 307. *The resonant frequencies shown on a horizontal frequency scale in Fig. 306 can alternatively be displayed as lines on a vertical or "thermometer" frequency scale.*

of atoms join together to form molecules of what is called copper oxide, which is a substance quite different from the metal, copper and the gas, oxygen, being a black powder. Its differences are due entirely to the rearrangements that take place in atoms of different kinds when they combine.

We are now in the subject of chemistry, where the number of atoms of hydrogen (or its equivalent) that one atom can persuade to combine with it to form a molecule is called its *valence*. As you may have guessed, it is equal to the number of its active outermost electrons, so they are called valence electrons. Copper,

as we have seen, has only one of them per atom, so is classed as univalent. But some of the materials we are going to be interested in as semiconductors have as many as four.

Energy bands

In examining the structure of copper, we began with a single atom (Fig. 303) — which is not of much practical interest! — and

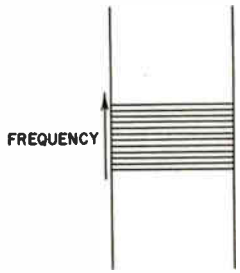


Fig. 308. *If very many equal tuned circuits were coupled, there would be that number of resonance frequencies, spreading out into a band as shown here.*

then went on to masses of atoms bound closely together in regular crystalline formation by the astonishing activity of the nimble valence electrons, one per atom. An important effect of this neighborliness of atoms is that the orbits of all the electrons are upset. It is as if millions of solar systems were brought together at regular intervals of only a few thousand million miles. Their outermost

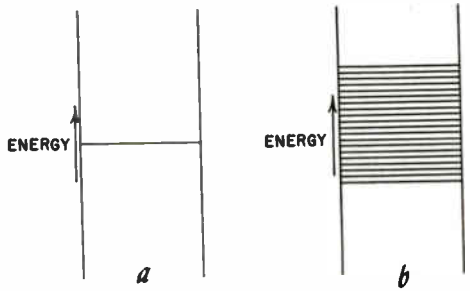


Fig. 309. *The energy possessed by an electron in orbit around an atom can also be indicated on a vertical scale. (a) The single energy level of an electron belonging to a lone atom. When a large number of atoms are close together, as they are in solid matter, the electrons are "coupled" and their equal energies spread out into a band (b).*

planets would then, at times, be closer to one another than to their own suns, so naturally they would distort one another's orbits by their gravitational attractions.

Another way of looking at it is to think of a sharply tuned

circuit, first by itself and then placed close to a number of identical circuits. We all know that one if transformer coil has a single resonance peak (Fig. 306-a) and, when this coil is coupled to

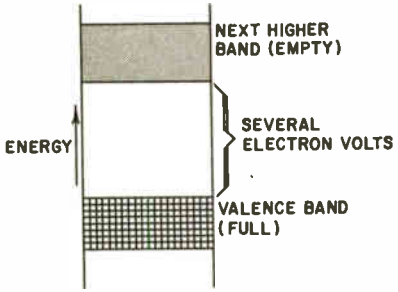


Fig. 310. *Insulating materials have a wide energy gap between the band in which the valence electrons are normally found and the next higher one into which they could be raised by energy coming from outside.*

another coil, the single peak broadens out into two (Fig. 306-b). An alternative form of diagram (Fig. 307) shows the resonant frequencies on a vertical scale. The same principle holds for many coupled coils, which together might be said to cover a whole band of frequencies (Fig. 308).

In a similar way, a molecule consisting of two copper atoms has two closely spaced orbits or energy levels shared between its two valence electrons. And, remarkably, the small chunk of copper made up of a quintillion atoms has no less than a quintillion energy levels, spaced so closely as to cover a whole band of frequencies. Yet, although the spacing is too small to imagine, the "only-two-per-level" law still applies as strictly as ever. The difference between the single energy level of the valence electron in a single atom, and the energy band of the valence electrons in a solid mass, is often shown as in Fig. 309.

Why copper conducts

Suppose you connect a copper wire, say 100 feet long, to the terminals of a battery giving 2 volts. Then (assuming the wire is the same gauge throughout), what you have done is to apply to it an electric field of 0.02 volt per foot. On an atomic scale of size, this field strength offers each valence electron only a very small voltage indeed. But the difference between one energy level and the next is so unimaginably small that any offer is sufficient to

boost the electron into it. This increase in energy speeds the electron towards the positive terminal of the battery. Meanwhile, the same thing is happening to all the other valence electrons. Even though there is only one of them per atom, there are in every cubic inch more than 1,000,000,000,000,000,000,000,000 (10^{24}) atoms, which is plenty. Though their drift may be only as fast as a few inches a minute, it adds up to quite a strong current.

This way of looking at electrical conduction is more difficult to understand and visualize than the Simple Story, and you may be

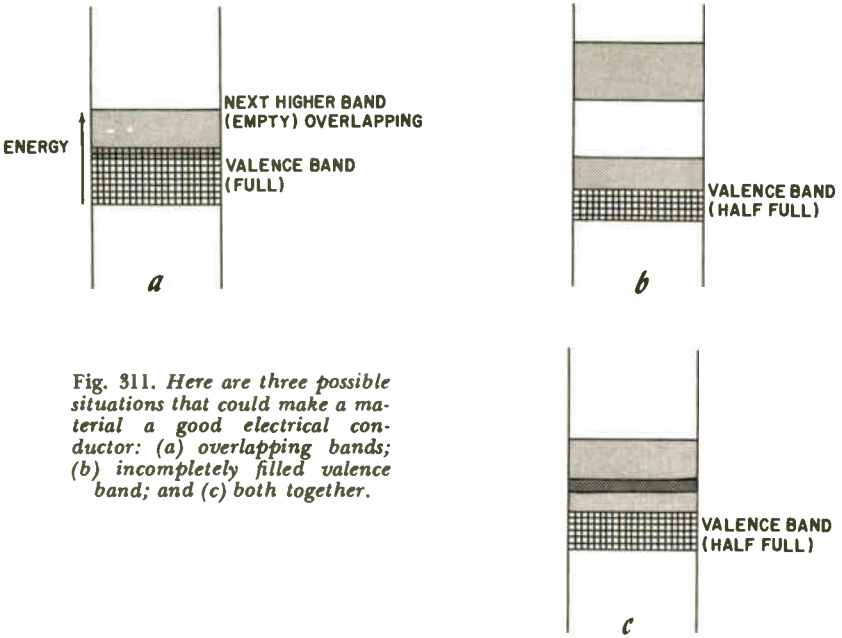


Fig. 311. *Here are three possible situations that could make a material a good electrical conductor: (a) overlapping bands; (b) incompletely filled valence band; and (c) both together.*

wondering why we were not content with that. Why bring in bands of energy levels? The reason may be clearer in a moment or two, when we ask —

Why insulators don't conduct

Contrary to popular ideas, the valence electrons of insulating materials are also free to move from atom to atom. Why, then, don't they conduct?

If every seat in a railroad car is occupied, a passenger cannot accept an invitation to sit nearer one end unless someone else moves toward the other end to make room. So if there are any movements at all they are equally in opposite directions. With

people, who are not all identical, that may yet be useful, but with electrons, which are quite indistinguishable from one another, it cannot be. Opposite movements of electrons cancel.

In an insulator, there are as many valence electrons as there are energy levels for them in what is called the valence band. So no valence electron can accept the boost in energy offered by an electric field unless another one takes a lower place, and that would cancel the current due to the first.

True, there are vacant higher-energy orbits farther out from the center of the atom. In solid (or liquid) material, the mutual coupling of the atoms has spread each of these energy levels also into bands. Insulators are distinguished by the fact that there is a wide energy gap — at least two or three electron-volts — between the valence band and the first vacant band (Fig. 310). To lift electrons across it would need an enormous total voltage. That is why insulators pass negligible current until a very high voltage is applied; then they break down with a bang as the electrons are forced across the energy gap.

Metals, on the other hand, have their atoms so arranged that either there is no gap between a full (two-electrons-per-atom) valence band and the next higher one, (in which case its energy diagram is like Fig. 311-a) or there is only one valence electron per atom, leaving half the band vacant for them to play about in (Fig. 311-b). Or else both conditions apply (Fig. 311-c). Copper comes into the last class. Because the band next above the valence band is normally vacant, allowing plenty of scope for any electrons that reach it to accept the small energy available in electric fields and so form currents, it is often called the conduction band.

Disturbing energy

What about the other sources of energy that were mentioned — heat, light and other radiations? Heat is everywhere, even at the South Pole, for the absolute zero (—273°C or —460°F) is never reached, although it has been closely approached in some laboratories. At quite low temperatures, there is enough heat to stir up at least the valence electrons a little, so that at any given moment some of them are in higher energy orbits than the lowest possible. (They keep on falling back again, so it is not always the same electrons that are up.) This continuous agitation is the cause of what is known as Johnson noise in circuits, which puts a limit to the amount of amplification that can be used.

Insulating materials, which have a wide energy gap between the valence electrons and the conduction band, need quite an

energetic kick to get any of them into it. Low-temperature heat has hardly any kicks hard enough to do it. But if the temperature is raised, the average energy of the heat kicks rises, and sooner or later the number of electrons in the conduction band becomes appreciable. The importance of this is that such electrons have almost unlimited empty "seats" in front of them (Fig. 312) and therefore are wide open to the attraction of any electric field or

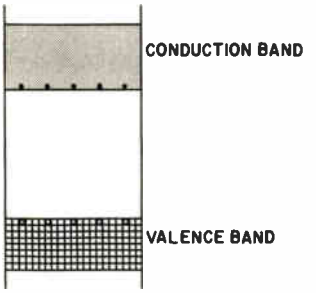


Fig. 312. *Whereas Fig. 310 shows an insulator at low temperature and without any other incoming energy, here heat energy has lifted a few valence atoms into the next higher band, leaving room for movement in both bands.*

electromotive force there may be. So the insulator begins to conduct, though only a little, for these promoted electrons are very few out of so many. But the number increases steeply with temperature. That is why the resistance of insulating materials falls steeply as temperature rises.

A remarkable law

Although you may not have thought so just now, the fact that it is temperature rather than quantity of heat that is responsible for lowering the resistance of insulators is remarkable. The total amount of heat in the Atlantic Ocean is something colossal (as you would find if you had to reduce its temperature to absolute zero!) yet it has no ill effects on the insulation of the cables laid through it. The comparatively negligible amount of heat from a single match, applied to the insulation, would affect it most noticeably *because of its higher temperature.*

Heat travels from place to place as radiation, and is in fact exactly the same thing as radio waves except for being at a higher frequency. The higher the temperature, the higher the frequency. When it is hot enough to be visible, it means that a small proportion of the heat energy has a high enough frequency to come into the light band, which begins at about 400 million megacycles. So it seems that it is *frequency* of energy that counts, rather than total amount of it.

That is one of the most important principles in science, dis-

covered by Planck as recently as 1900. It accounts for the fact that all the heat in the world at a low temperature can't kick many valence electrons up into the conduction band of an insulator, but a very little at a high temperature will do the trick.

How remarkable Planck's principle is can be pictured by imagining that it applied to sea waves as well as electromagnetic waves. If it did, then the heavy stones forming a sea wall, able to stand up to the most violent ocean breakers of long wavelength (low frequency), would be knocked out by much smaller ripples of short wavelength!

Some useful results

At a visible temperature, some of the energy kicks may be hard enough to knock electrons clean past all the vacant levels right out into the open air. This is how the hot cathode in a tube works. We say, more politely, that the electrons are thermionically emitted. The energy picture would look something like Fig. 313.

In some materials, such as tungsten, the energy difference between the normal valence band and the emission level is so large that a white heat is needed to emit a useful number of electrons. Others, used in receiving and TV tubes, have smaller differences and emit at a dull red heat. A few release electrons even when ordinary light is shone on them, without heating. They are the ones used for photoelectric cells.

As the frequency of energy radiation is raised beyond that of light we get in turn ultra-violet, X-rays, gamma rays and, lastly, cosmic rays. Their violence increases in proportion to their frequency, so that while light, with energy of a few electron-volts, can knock out valence electrons, X-rays will knock out the inner electrons, and cosmic rays, with millions of eV, can break up even the nucleus itself.

Electron bombardment

Emitted electrons can be accelerated by positive voltages and made to bombard other substances, such as gas in the tube or electrodes at which they are aimed. Provided the electrons are traveling fast enough, their kinetic energy can knock other electrons out of a gas or solid. In gas, the process is called ionization; in solids, it is called secondary emission. We won't go further into these processes now, for they belong to tube electronics rather than semiconductors, but it is interesting to know how they are related.

In this chapter, semiconductors have hardly been mentioned,

and you may be wondering if we have strayed from the point. But actually we have been accumulating the basic ideas needed to understand why semiconductors work the way they do. Now we have enough to go ahead. But as these ideas are somewhat complicated, we had better run over them again quickly.

Summary

Electrical *work* is calculated as the amount of charge times the number of volts its potential is raised. When the charge moved is the amount carried by each electron, it is convenient to measure the work in electron-volts. The *energy* possessed by anything is the amount of work it can do.

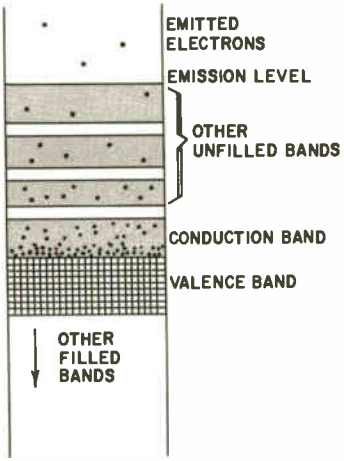


Fig. 313. *In this case the heat energy is sufficient not only to lift electrons into all the higher bands but to throw some of them right out of the material, as emission.*

The possible orbits of an electron around a single atom are arranged at fixed intervals, so that the energy needed to boost the electron into higher orbits can only be accepted in packets of certain sizes.

When many atoms are brought close together to form solid matter, these single energy levels are widened out into bands.

Where there is no disturbance from heat, light, etc., electrons always occupy the lowest orbits or energy levels; i.e., nearest the main body of the atom. Only two can occupy any one level. The outermost electrons, which are mainly responsible for the atom's activities, are called *valence* electrons.

If all the levels in the top inhabited (valence) band are filled, and there is a substantial energy gap between that and the next higher (empty) band, then any rise in energy by any one electron must be cancelled with a fall by another, so there can be no net gain in energy and no electric current. The material is an insulator.

If the valence band is not filled, or there is no gap between it and the next higher one, electrons are free to accept energy from electric fields and be set in motion; the material is a conductor.

Electromagnetic waves (radiation) have amounts of energy that increase in strength in proportion to their frequency.

Heat at moderate temperature kicks a few electrons across the gaps in insulators, making them conduct a very little; at high temperatures they may be made to conduct quite a lot, and some of the electrons are emitted completely from the material and can be used on their own, as in a vacuum tube. A similar result can be obtained by bombarding the material with electrons having sufficient energy.

A warning

The relative energies of electrons in single atoms or bulk material can conveniently be shown by a "thermometer" type of diagram as in Figs. 308–313, but it must be clearly understood that this is not meant to show their relative *positions*. It does happen that the energy of an electron increases with the distance from its nucleus, but valence electrons (which are almost the only ones of any interest to us) can hardly be said to "belong" to particular atoms at all. In Fig. 310, they are all shown massed together in the valence band; this means that they all have nearly the same amounts of energy, *not* that they are concentrated in one part of the material, for actually they are distributed uniformly throughout it.

In Fig. 313, the emission level does look as if it might be a picture of the surface of the material, but the resemblance is (like that of a film character to any living person) purely coincidental.

chapter 4

a typical semiconductor

THE material we studied in the preceding chapter, in terms of its atoms, was copper. It has 29 electrons per atom, but only one of these is a valence electron. The other 28 we put back in the parcel along with the nucleus, and called that the atom. To be precise, it is a positive ion, for it is one electron short and so has one net positive charge. On the one valence electron, which circulates around the rest of the atom, falls the whole responsibility for foreign affairs, such for example as alliances with other copper atoms to form copper crystals, or with other kinds of atoms to form chemical compounds. At the same time, by marvelous juggling, it manages to move from atom to atom as random "noise" or as an organized electric current.

Energy gaps again

This last job can be done very readily by the valence electrons in copper because there are almost unlimited higher-energy levels open to them right at hand. But we noted that many other kinds of atoms are so organized that there is a wide gap between all the energy levels occupied by the valence electrons and the nearest vacant ones. These make insulators. An electric field has to be overpoweringly strong to compel any of them to cross this gap and qualify as current carriers. But, at a suitably high temperature, the heat energy kicks some of them across and current can then flow, though it is severely limited by the scarcity of these mobile electrons.

It has probably occurred to you that some substances might have energy gaps smaller than those in insulators, so that even at ordinary temperatures there would be a medium amount of conduction — more than in insulators and less than in metals. There are, and we call them semiconductors. Because they do have gaps they hardly conduct at all at very low temperatures. A rise in temperature causes increasing numbers of electrons to be kicked across the gap, and the conductance increases very steeply. In this respect, semiconductors are more like poor insulators than metals, which actually conduct slightly less when the temperature is raised. So now we have the explanation of Figs. 201 and 202 as well, if we remember that light is energy of a higher frequency.

Semiconductor crystal structure

The most-used semiconductors are elements — substances with atoms of only one kind — having four valence electrons per atom and therefore called tetravalent. So each atom is able to hold on to

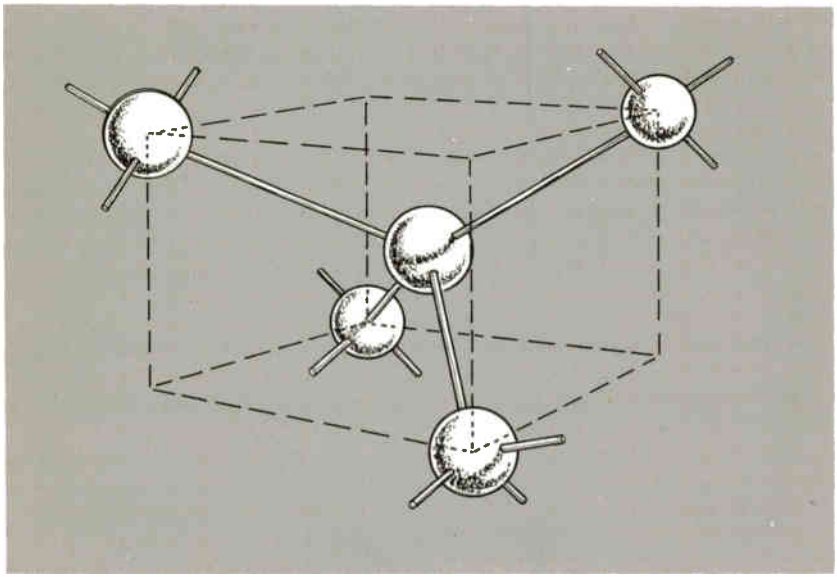


Fig. 401. *Atoms in tetravalent materials such as germanium and silicon are spaced in regular crystalline formation like this.*

four other atoms. This makes the materials crystallize into a characteristic form, as shown in Fig. 401. Of course, a crystal structure is in three dimensions, which is difficult to show clearly on two-dimensional paper. That is the reason for the dotted lines, which

trace the outline of a cube in perspective and help the eye to see how the whole thing would appear in solid form. The atoms are represented by balls, and the forces that hold them together are represented by thick lines. The atom in the center of the cube is linked to the four others in the corners by its four valence electrons. Each of these other atoms must be imagined as linked to four — three besides the center one shown. And so on, to form a continuous structure — a crystal made up of vast numbers of atoms.

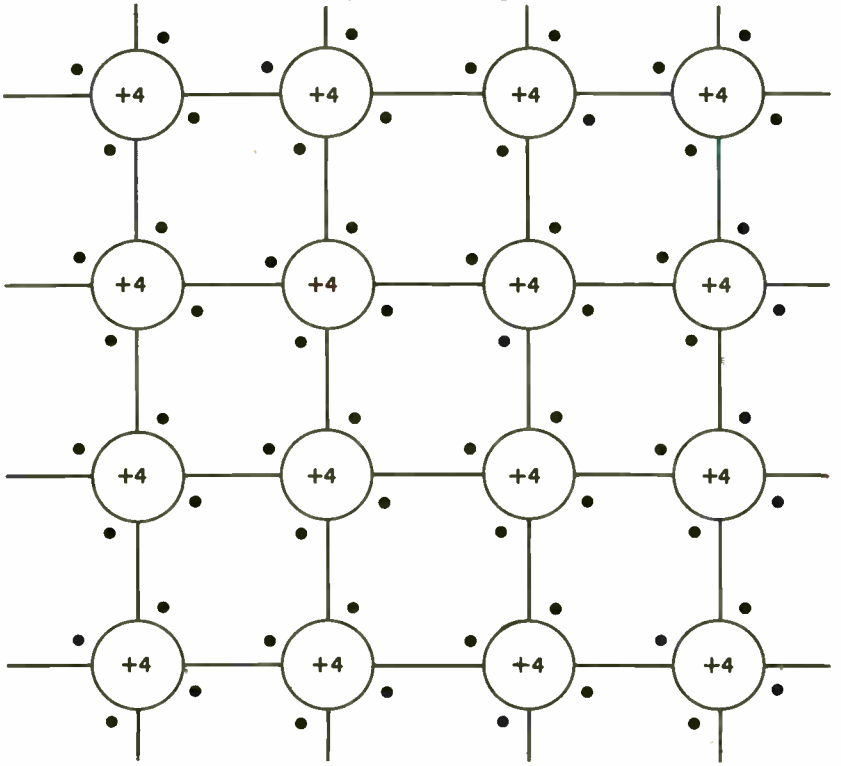


Fig. 402. *If the germanium atomic structure of Fig. 401 is flattened out it looks like this. The four valence electrons in each atom are shown as dots, and the rest of the atom as a circle with "+4" to indicate its positive electrical charge. The lines represent binding forces due to the valence electrons.*

For diagrammatic purposes, it is more convenient to flatten the whole thing into two dimensions, as in Fig. 402. Each atom is shown here as a circle marked "+4," because it includes the nucleus and all except the four outermost or valence electrons, and so carries a net positive charge of four electronic units. The four valence electrons per atom are shown as black dots. And the links of attractive force created by the pairs of valence electrons shown

directly between any two atoms are represented by straight lines, as they were in Fig. 401.

A group of semiconductors

The first tetravalent element on the list is a familiar one — carbon. It has only six electrons altogether, so valence electrons actually outnumber the rest, which is most unusual. Carbon is interesting in another respect: its atoms are so constructed that they can click into more than one stable pattern. One of them is graphite, which is soft and black and has such a small energy gap that at ordinary temperatures it conducts almost as well as a metal. You can easily experiment with it, for it is the "lead" in a pencil. The other form of carbon — much less plentiful! — is diamond, which

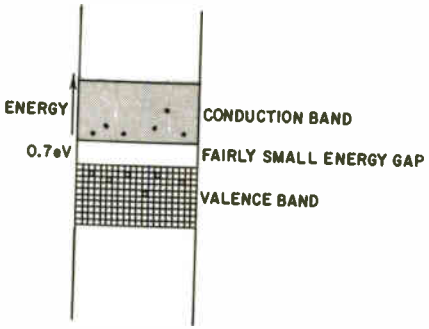


Fig. 403. *Here is the energy diagram for germanium. There is a gap, but a few electrons receive enough heat energy at ordinary temperature to lift them across it.*

is hard and transparent and has such a large gap that it is an insulator. These remarkably different properties of one and the same element are due simply to the ways in which its electrons arrange themselves. However, useful though these alternative crystalline forms are for some purposes, they are obviously no good as semiconductors.

The next tetravalent element is No. 14: silicon. Combined with other elements, it is one of the commonest that make up the earth. It is of great commercial value as a semiconductor. But No. 32 — germanium, a rare element, found chiefly in the flue dust — is still used for the greatest number of semiconductor products, so we shall choose it as our typical specimen. Silicon is very similar except that its energy gap is larger, so at the same temperature it conducts less. The next is tin (No. 50), which also comes in two forms, one of which — the well known one — is metallic, and the

other – gray tin – is technically a semiconductor but its gap is so small that it conducts too well for practical purposes.

Germanium

Fig. 401 gives us some idea of the structure of a piece of pure germanium crystal and the relative positions of its atoms. So to some extent does Fig. 402, even though it has to be distorted to get it into two dimensions on a piece of paper. Fig. 403 shows the narrow energy gap which is the mark of a semiconductor, and (since we will assume the material to be at room temperature) a few electrons which have been lifted into the conduction band. These are free to respond to electric fields produced say by emf's and be moved towards the positive end of the crystal as an electric current. But because only perhaps one in a thousand million of the valence electrons is in this position, the current for a given emf cannot be large. In other words, the conductivity of the material is small.

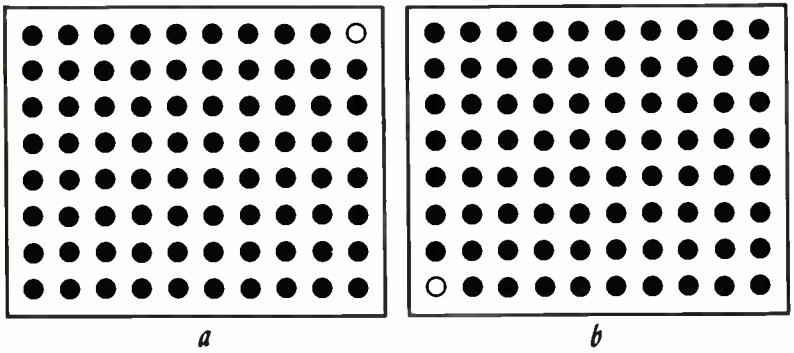


Fig. 404. *The one white circle represents a vacant seat, which appears to move from top right in (a) to bottom left in (b), although actually only people move, from one seat to another.*

Holes

The drawing also shows empty places in the valence band, left behind by the promoted electrons. These vacancies are called *holes*. Absolutely vital to semiconductor action, they are often misunderstood and so are worth spending some time over.

Just as promotions to the head office of a company create vacancies in the branch, enabling promotions to be made there too, so the vacancies in the valence band leave energy room for current conduction. But there is more than one way in which such movements can be described.

Suppose Fig. 404-a represents part of the seating of an auditorium, the black blobs being occupied seats and the single white one a vacant seat. Someone next to this vacant seat could move over into it. Then someone next to the new vacancy could move into that, and so on, until perhaps in the end the vacant seat was the one in the opposite corner, as at Fig. 404-b.

What actually happened was that a lot of people each moved a little distance and the whole operation would be very tedious to describe in full. But the same thing could be described much more simply by saying that one vacant seat moved all the way from one corner to the other. You may object that a vacancy, being nothing,

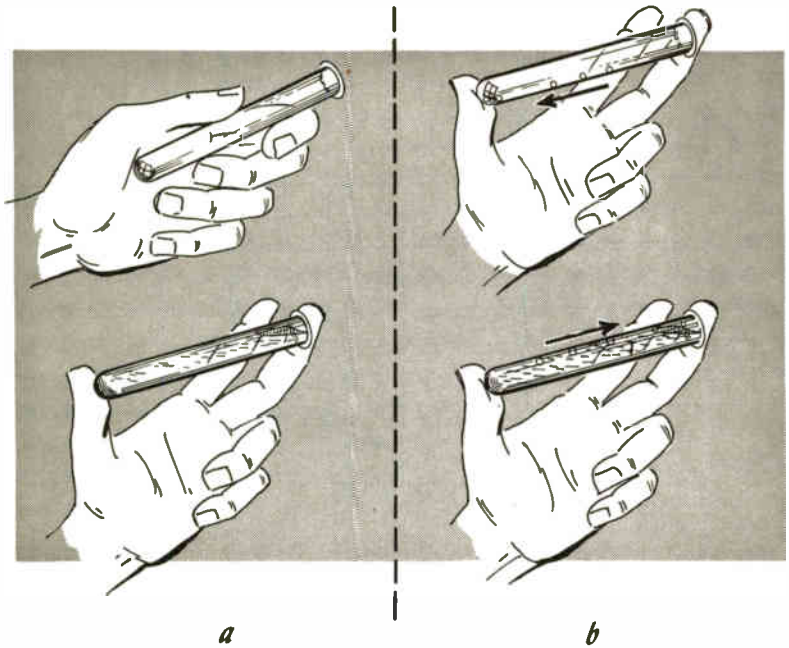


Fig. 405. *If a tube is either completely empty or completely filled with liquid, no movement can take place inside when it is tilted. Transferring a few drops from one to another allows drops to move one way at the same time as bubbles move the other way.*

can't move; in a way that is true, yet all the same it is a convenient method of describing and picturing what took place, and the result is equivalent to the stricter description.

Note particularly that anything which attracts the people in one direction repels the vacant seat in the opposite direction. So if the black blobs in Fig. 404 were now to represent negative charges

(electrons) it would make sense to regard the white circle as a positive charge of the same magnitude. Although we know that the energy exchanges in the valence band in Fig. 403 cause vast numbers of electrons to move a little towards the positive end of the germanium crystal, it is easier to think of what happens as a movement of a few holes – positive charges – toward the negative end. As we saw with Fig. 404, it amounts to the same thing.

If you prefer a different analogy, consider what happens when two glass tubes are tilted, as in Fig. 405. First (*a*) one tube, representing the conduction band, is empty and the other (the valence

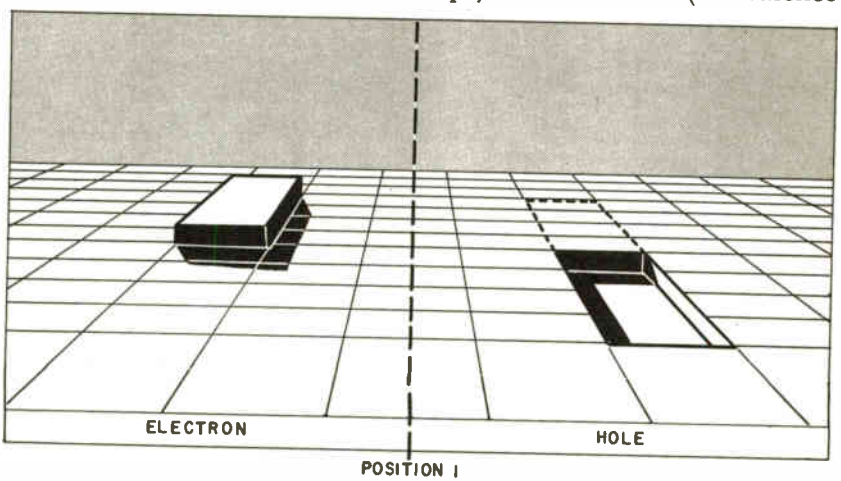


Fig. 406. *Here a square "tile" is in its backward position, making a backward prominence or a forward depression.*

band) is completely filled. The tilting can cause no liquid to move. But if a few drops are transferred from the full to the empty tube (*b*), those drops will move towards the lower end when the tube is tilted. They represent electrons moving towards the positive end when an emf is applied to the crystal. At the same time, all the drops in the lower tube can now move a very little way towards the same end – but it is so much easier to regard this as a few bubbles moving towards the opposite end.

Hole currents

Whichever way you choose to look at it, one thing quite clear is that the vacancies in the valence band result in current over and above that due to the electrons in the conduction band. It is found by experiment (in germanium and silicon, at least) that holes move only about half as easily as electrons. So the result is not quite as much as you might have expected.

Teachers are sometimes so anxious to make clear that holes are not real positive charges but only the equivalent in electron movements that they may make it difficult to understand Hall effect (which we will be coming to in Chapter 9). So before assuming we understand this hole business completely, let us take a look at Fig. 406.

Here a square piece projecting above a flat surface represents a unit of negative charge, an electron. A square depression below the surface represents a unit of positive charge, a hole. In the position shown, an electron is one square's length farther away

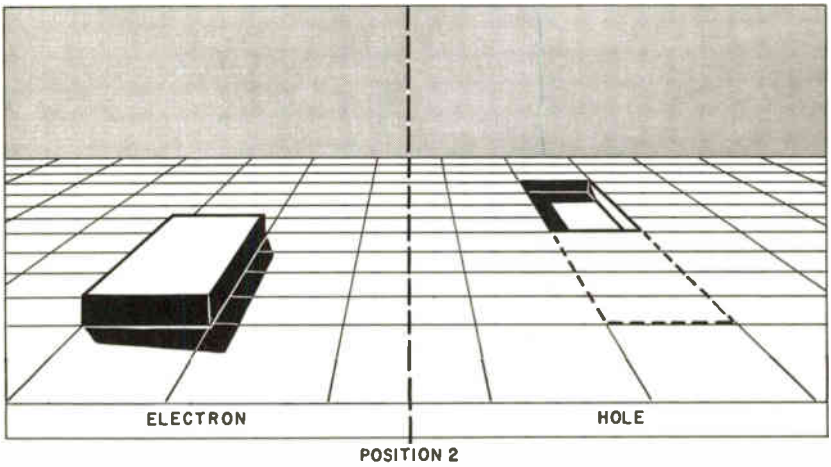


Fig. 407. *Moving the tile forward makes the prominence move forward and the depression move backward.*

than a hole. Now compare Fig. 407. The electron has moved in one direction – nearer – and the hole has moved in the opposite direction. Considering what has brought about this change, you will see that in both cases it is exactly the same – a square piece of material has moved nearer. That corresponds to the view that a hole movement one way is really an electron movement the opposite way. But looking at them as projections or depressions – negative and positive electric charges – you would be bound to say that in one case a negative charge has moved one way and in the other a positive charge has moved the opposite way, so the results are different in spite of the causes being the same. That corresponds to the view that a hole is a real positive charge and not just an electron missing.

Again, don't assume you understand holes completely, for it

must be admitted that this mechanical analogy wouldn't satisfy a scientist. A very difficult study, known as wave mechanics, is needed for that. But it does confirm that for most purposes we are justified in regarding a hole as a real positive particle.

Intrinsic conduction

The piece of germanium we have been studying was specified at the start as pure. That is to say, we assume it to contain no atoms of any other element. But even in the most painstaking laboratory there is no such thing as a perfectly pure sample of any substance, any more than there exists a perfect vacuum. Usually, if there was no more than one part in a million of impurity, we would call that pure. But with semiconductors—no!

As a matter of fact, even if perfectly pure germanium or silicon were obtainable, it would be of no use in semiconductor products such as transistors. As we shall see, their usefulness depends entirely on impurities. The germanium crystal corresponds to the vacuum in a vacuum tube rather than to the active electrons. And the electron and hole current we have been spending several pages on is really no more than an incidental nuisance.

Have you then been having your time wasted? Definitely not. For one thing, this nuisance current is vitally important in the design of power amplifiers using germanium transistors (less so with silicon transistors). For another thing, if the principles of current flow in a pure semiconductor are understood, one is well on the way to understanding impurity conduction.

Before we go on to that, let us take note that the sort of conduction we have been considering – which would be the only kind if the germanium were perfectly pure – is called *intrinsic* conduction, because it is a result of the structure of the semiconductor material itself. Since it depends on electrons being kicked up into the conduction band by packets of energy of at least 0.7 eV each, it would not exist and the material would be an insulator if there were no disturbances of an energy-giving kind.

Effect of temperature

The most important of these disturbances is heat. Light packs even more powerful punches, but they come much less thickly, and anyway, can easily be excluded from such things as transistors by an opaque covering. But apart from expensive and inconvenient refrigeration, we have to accept heat. Moreover, there is additional heat generated by power loss in the semiconductor devices themselves when they are in use.

When the temperature rises, free electrons and holes are created faster, so that soon there are more of them and conduction is greater. But this increase doesn't go on forever; the more electrons and holes there are, the quicker they combine and cancel one another. This is represented in an energy diagram by electrons falling down again and filling holes. Soon a balance is reached, so that the number of free electrons and holes is steady, and so is the conduction. Things work out so that the conductance is about doubled by every 10°C rise in temperature.

At room temperatures, the intrinsic current in germanium devices is appreciable but not a serious embarrassment for most purposes. However, if it is worked hard enough, or the surroundings are hot enough, to bring it near to the temperature of boiling water, the intrinsic current begins to control the situation and may additionally heat the germanium so much that ultimately it is burned out. This regrettable process is called thermal runaway.

The great merit of silicon is that its energy gap is about 1.2 eV. To an electron, 1.2 eV is a lot more difficult to cross on the wings of heat than 0.7 eV. So silicon devices work quite happily well above the boiling point of water. Why use germanium at all? Well, it too has its advantages, one of them at the present being that it is easier to purify.

Impurity conduction

A cubic millimeter of germanium (which is about the size of a pinhead) contains about 5×10^{19} atoms. So if only one part in a million of some impurity were present, there would be in that tiny piece of germanium 50,000,000,000,000 atoms of impurity. That is mentioned at this time just in case someone thinks one part in a million couldn't be important.

Whether it actually is important or not depends on what sort of impurity it is. We are going to bother about only two kinds of impurity — trivalent and pentavalent elements — elements with valences of three or five. That is, one less or more than our tetravalent semiconductor.

The pentavalent group sounds as if it belonged in a detective story — arsenic, antimony and phosphorus. Suppose a few atoms of this kind (the odd fifty trillion per cubic millimeter) have become mixed up with our otherwise pure crystal of germanium. They take their places in the crystal structure (usually called the lattice) but they are misfits. Look at Fig. 408, for instance. You can easily spot the odd man in, because it is marked "+5" —since its five valence electrons must be balanced by an equal

positive charge in the rest of the atom. Four of the valence electrons immediately find jobs holding the lattice together, but the fifth is at a loose end. It is free to be influenced by any emf around.

Now, as we had to learn for conductors and insulators, the mere fact that an electron appears to be free in a position diagram such as Fig. 408 doesn't necessarily mean that it is free to take part in an electric current. Nor does the fact that an electron appears to have

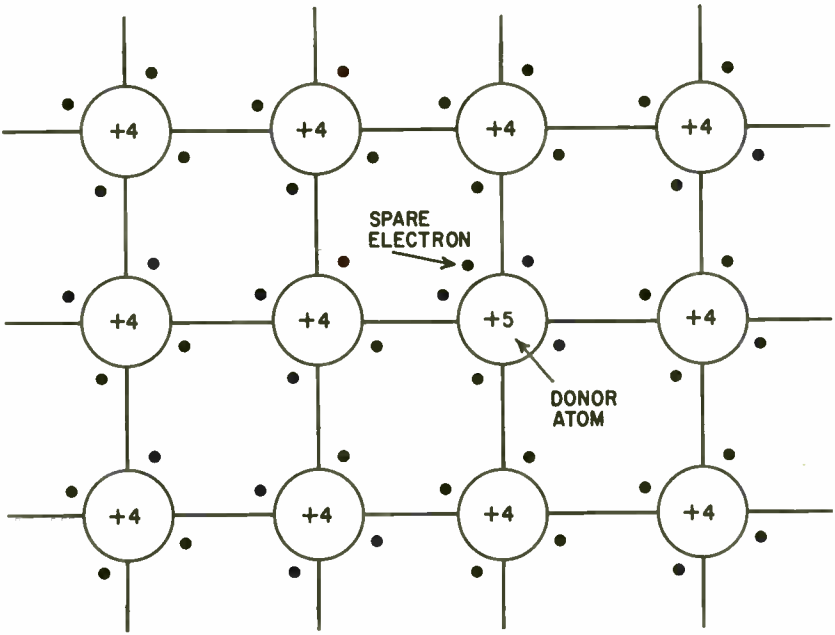


Fig. 408. *In this germanium crystal there is one atom of an impurity having five valence electrons, one of which is easily movable.*

a full-time job holding the crystal together mean that it is debarred from being part of an electric current. We have to know its energy status.

Fig. 409 is the electron energy diagram, which shows that the spare electrons given by pentavalent atoms come only about 0.01 eV below the bottom of the conduction band. This gap is so small that even in the coldest weather the heat energy is enough to keep practically all the spare electrons aloft in the conduction band.

A jumping competition

How do they compare with the number of electrons that ordinary temperature is able to hoist from the million times greater number of germanium atoms 0.7 eV lower down the energy scale?

In spite of the enormous numerical advantage of the germanium atoms, the deterring power of the 70 times bigger gap is such that only about 10,000,000,000 make it. So they are outnumbered

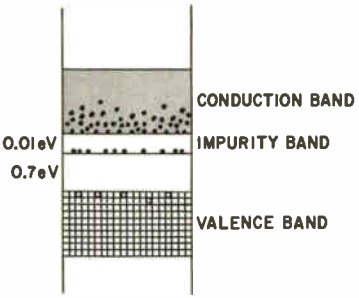


Fig. 409. This is the energy diagram corresponding to Fig. 408, except that it includes a larger piece of crystal with many impurity atoms.

5,000 to 1 by the 50 trillion impurity electrons. When an emf is applied, the intrinsic current is almost negligible compared with that due to the minute trace of impurity.

The situation is like a competition between a whole nation set to jump over 6 feet and a thousand people set to jump 1 foot. Practically all the thousand would succeed, outnumbering the few top athletes in the whole nation.

Obviously, even a much smaller amount of impurity than one in a million would have a measurable effect on the conductance — far less than could be detected by the most sensitive chemical tests. As little as one part in a trillion (10^{12}) can be detected by its electrical effects, and one in ten billion (10^{10}) is commercially important. This is at the rate of one quarter of a person in the entire population of the world! So it is not difficult to imagine why the manufacture of semiconductor devices is so tricky. Methods of purifying materials are described in Chapter 6.

Current carriers

This material, which by ordinary commercial standards is so pure but by electronic standards so impure, is called *n* type, because conduction is mainly by electrons, which are *n*egative particles. Because of its generosity in supplying these spare electrons, pentavalent impurity is called a donor. (If you want something more to remember it by, you will note that "donor" too has an *n* in it.)

What about holes? Wherever there is nothing but holes there are no holes. Try to imagine a sock in the state, or, if that is too big a stretch of the imagination in these days of nylon, turn back to Fig. 405-b and suppose all the liquid in the lower tube transferred to the upper. The lower tube then being occupied exclusively by holes could equally be said to have no holes. Either way, there would be no "conduction."

The impurity band in donor material, being cleaned almost right out by heat, etc., is in this sort of state. So the only holes are the comparatively few (at room temperature) in the valence band. These are therefore called minority current carriers. The impurity electrons, by contrast, are majority carriers.

As a matter of fact, there are even fewer holes than there would be if the germanium were perfectly pure. The reason is that there

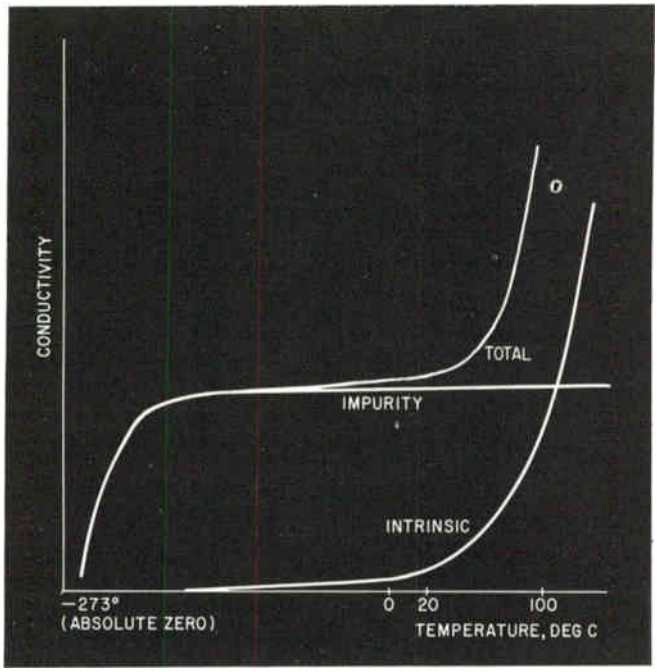


Fig. 410. *The impurity energy gap in Fig. 409 is so small that nearly all the electrons are raised into the conduction band at quite a low temperature. The main gap being much larger, the germanium electrons need a higher temperature to make them conductive. So the total conductivity curve has a double bend in it.*

are so many more electrons ready to fall into them — all the impurity ones as well as those originally belonging. So, on the average, the holes last only a very short time. In a town where there are equal numbers of males and females, there will be on the average a certain number of unmarried females of eligible age. But if the male population were to be increased a hundred fold, that number would presumably decrease.

When temperature rises, the number of holes rises steeply, as we have seen. But even at the most Arctic temperature practically all the spare impurity electrons are already in the conduction band, so the impurity conductance is hardly affected. Actually it goes down a bit, as in a metal.

Fig. 410 summarizes these temperature effects. Because of the

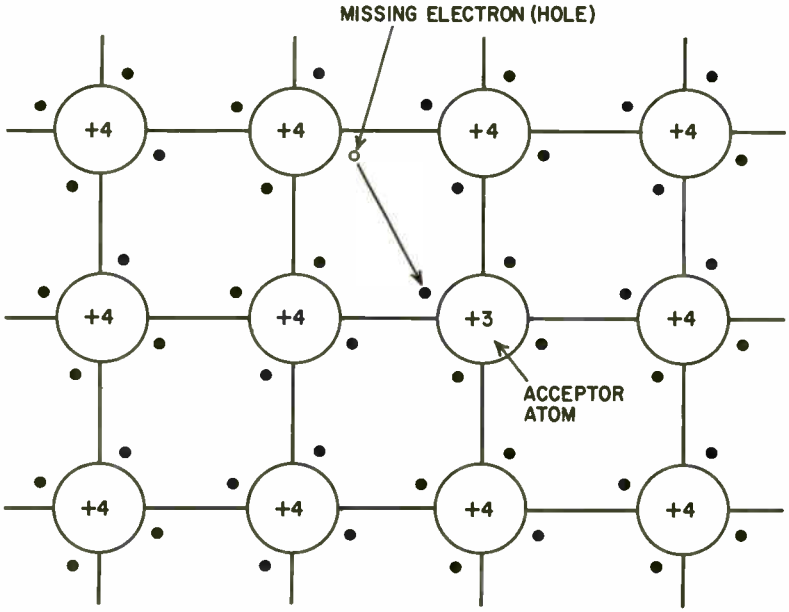


Fig. 411. *Here the one impurity atom has three valence electrons, so is apt to borrow one from a germanium atom, leaving a hole there.*

0.7-eV gap, intrinsic conduction has a slow start, but above room temperature it rises steeply and, because the germanium atoms are so many, it can rise high. The 0.01-eV impurity gap gives an early start, but saturation is soon reached because impurity atoms are so few. The conductance of the impure material is a combination of these two — not exactly simple addition, because of the effect of the many impurity electrons on the few intrinsic holes.

One other important point. Mention of "spare" or "surplus" electrons might give you the idea that a considerable number of negative charges accumulate somewhere. So it is necessary to remember, first, that Fig. 409 is purely an *energy* diagram, and transfer of a lot of electrons from one band to another does not mean a transfer from one *place* to another. Second, all these extra electrons are neutralized by the extra positive charges in the impurity atoms from which they came. The vital difference between these positive and negative charges is that the positive are *fixed* (being part of the crystal structure) and the negative are *mobile.*

Trivalent impurities

What happens when a trivalent impurity is present — such as aluminum, gallium or indium — need not take us so long, for it is the same as for pentavalent in reverse. Fig. 411 shows an atom of this kind of impurity, with its three units of positive charge. Having only three valence electrons, it is unable to man its position in the crystal lattice fully; there is one short, and so it recruits it from one of the neighboring germanium teams. This atom in turn proceeds to do the same to another, and so on.

As we saw a few pages back, this process can be described more economically by saying that the impurity atom creates a hole, which wanders at random around the crystal. If, however, an emf is applied to the crystal, so that its atoms are in an electric field, the tendency is for the electrons which move into holes to be those on the negative side of the hole, so that they move towards the positive pole of the emf.

Because a trivalent impurity atom grabs an electron from its hosts, it is called an acceptor. The "*p*" in this word helps to remind us that it creates holes which are *p*ositive charges, and that the material it helps to form is called *p* type.

Fig. 412 is the counterpart of Fig. 409. This time the impurity atoms have *unoccupied* energy levels only about 0.01 eV above the top of the valence band. Any ordinary temperature gives enough heat energy to fill these levels from the relatively unlimited supply of germanium valence electrons, creating an equal number of holes. This number, therefore, depends on the number of impurity atoms, which is usually comparatively small. But at working temperatures it is still very much greater than the holes created by electrons that have jumped the 0.7 eV into the conduction band.

Note that the impurity band is so jammed with electrons that they have no room for maneuver, so are unavailable for current

flow. Conduction takes place mainly by holes, which now are the majority carriers. The minority carriers are the electrons in the conduction band.

Fig. 410 takes no account of whether conduction is by electrons or holes, so its general character applies to *p*-type material as well as *n*-type.

Summary

Before going any farther, ponder over the conduction process for both types of material, with regard both to position (Figs. 408 and 411) and energy (Figs. 409 and 412), until it is clear. Neither type of diagram or point of view by itself tells the story completely enough for a clear understanding.

Perfectly pure semiconductor is unattainable, so conduction in practice is made up of two parts:

1. The intrinsic conduction of the basic semiconductor, in which nearly all the atoms are involved but are handicapped by the wide

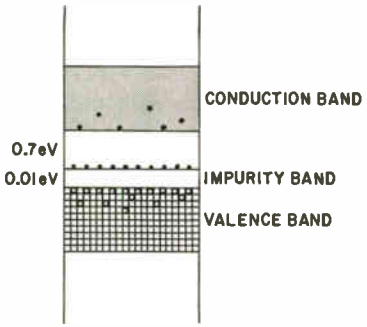


Fig. 412. *Energy diagram corresponding to Fig. 411.*

energy gap to be jumped. So only a comparatively high temperature (depending on the semiconductor used) gives enough heat energy for it to amount to much, its current carriers being in a minority. This is as well, for it is usually a nuisance. The amount of this conduction depends on temperature, and beyond a certain point increases very steeply indeed (Fig. 410). The temperature at which it becomes serious depends on the semiconductor, being often inconveniently low for germanium and higher for silicon. Intrinsic conduction is also reduced by the presence of majority carriers.

2. Impurity conduction, which depends on the proportion of impurity present, usually extremely small by ordinary standards. But this smallness is more than offset by the smallness of the gap

to be jumped, which insures that practically all are available for current conduction. Temperature (except near absolute zero) therefore, has little effect, but the amount of impurity is all important.

Semiconductor material with pentavalent impurity is called n-type because conduction is by negative majority current carriers (electrons) provided by donor impurity. These mobile charges are neutralized (so far as the material as a whole is concerned) by fixed positive charges — the donor atoms minus their spare electrons.

Material with trivalent impurity is called p-type because conduction is by positive majority current carriers (holes) provided by acceptor impurity. These mobile charges are neutralized by the fixed negative charges of the acceptor atoms plus electrons pilfered from the semiconductor atoms.

Compensation

The question you are bound to ask is, what happens when *both* kinds of impurity are present in the same material?

The quick answer is that so far as possible they cancel one another. If germanium had 10 parts per million of arsenic, and you added 6 parts of indium, the result would be as if the material had 4 parts of arsenic.

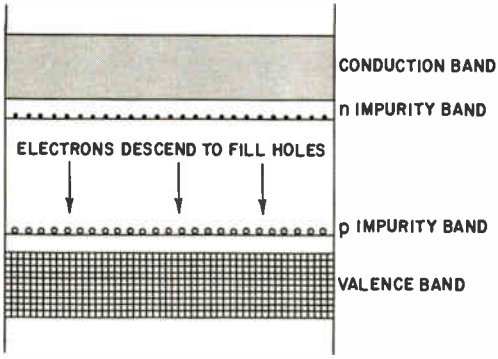


Fig. 413. *When both kinds of impurity are present together, they tend to cancel one another by direct borrowing of electrons one from the other.*

This may seem an easy way of avoiding the troublesome process of purifying crude germanium sufficiently to get the very small proportion of impurity required — or even to get the equivalent of perfectly pure material. It doesn't work like that, for several rea-

sons. First the neutralizing action — called *compensation* — is satisfactory only when the amounts of impurity are quite small. Germanium with exactly 1% each of donor and acceptor impurity would not behave as perfectly pure germanium. Second, even if it did, the problem of insuring that the amount of compensating impurity added was exact enough — say to one in a billion — would probably be no less difficult than straightforward purification. And, third, such a precise balance between two large quantities of impurity would be almost impossible to maintain throughout a bulk supply of material.

So although compensation is useful in practical semiconductor technique, it is a supplement to purification — not a substitute.

In terms of our position diagram, we can visualize the electrons from the donor impurity filling the holes in the acceptor impurity, so that neither lot of carriers is available for current flow. In the energy diagram, there is initially a filled impurity band near the top of the gap and an empty one near the bottom. In accordance with the rule about filling the lowest levels first, the electrons drop from the upper to the lower impurity level, as shown in Fig. 413. (To keep it clear, intrinsic effects have been omitted, and equal amounts of n- and p-impurity assumed.) This obviously prevents either kind of impurity conduction taking place, because the descending electrons can't go up into the conduction band, and they fill the holes into which electrons in the valence band could go.

chapter 5

junctions

By now we know that semiconductor material can be had in three varieties: (1) p type, in which conduction is mainly by holes (positive); (2) n type, in which it is by electrons (negative); and (3) what is sometimes called i type ("intrinsic"), in which there is only a little conduction, by equal quantities of both holes and electrons at once. None of these types by itself looks very useful (except, perhaps, to a limited extent as a photocell — see Chapter 8). It is when different types are in contact that their peculiar modes of conducting lead to the most useful results.

Separate p and n

Just pushing two pieces together is hardly good enough. To get a sufficiently close contact, the material should be in one piece divided into two or more regions or zones having different impurity content.

So as not to have too many things to think about at once, let us ignore intrinsic conduction altogether at first. In fact, in our diagrams we will ignore perhaps 99.9999% of all the material and concentrate on the trace of impurity that makes up the balance. The impurity atoms are scattered at random throughout the solid, so instead of the regular pattern of Fig. 402 we have Fig. 501. This shows two separate pieces of material, one p-type and the other n.

The two kinds of mobile charges or current carriers are denoted by plus and minus signs, and the fixed impurity atoms which have had one electron added or removed are denoted by the opposite signs in circles. Note particularly that the only current carriers in

the p-piece are holes, and the only ones in the n-piece are electrons. The alternate form of diagram underneath shows this more quickly, and also more clearly that holes and electrons are dis-

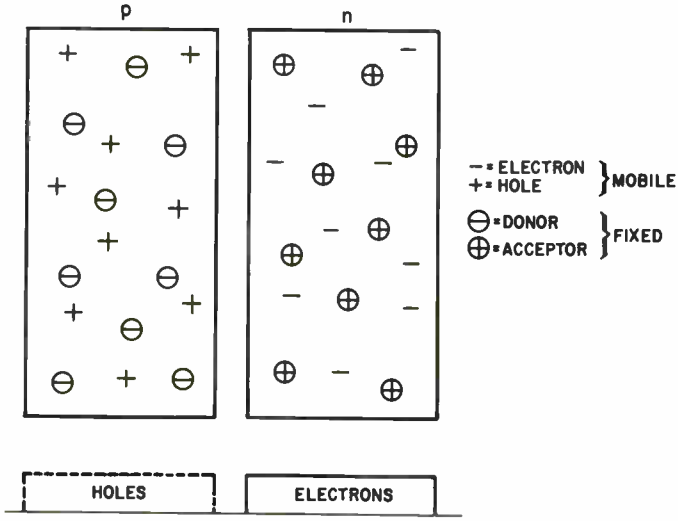


Fig. 501. *This kind of diagram ignores the many germanium atoms and shows only the few impurity atoms scattered about through the material. On the left is an acceptor impurity (p) and on the right a donor impurity (n).*

tributed uniformly, throughout their own pieces, with equal density.

Although the fixed charges are unable to take part as currents, they do serve an essential purpose by neutralizing the mobile charges, so that the material as a whole is not charged. We will see later how helpful this is.

The two in contact

Next suppose these two pieces are both in one crystal, forming what is called a p-n junction. The boundary between the two regions — the junction itself — is marked by a dotted line in Fig. 502. What happens?

To begin with it is very much like what might be expected if a fence separating the playgrounds of a boys' school and a girls' school was removed. The electrons in the n-region stray across into the region that has none, and likewise with the holes. A few of these strays can be seen near the dotted line in Fig. 502, and the changed distribution is shown below. In semiconductor language, this tendency for electrons and holes to spread out so as

to occupy the whole available space uniformly is called *diffusion*.

But then arises something that has no analogy in school playgrounds. A difference of potential appears between the two sides of the junction. Instead of the exact balance of positive and negative in Fig. 501, there is a positive surplus along the n-side and a negative surplus along the p-side. The n-side becomes positively charged with respect to the p. This is so, not only because positive

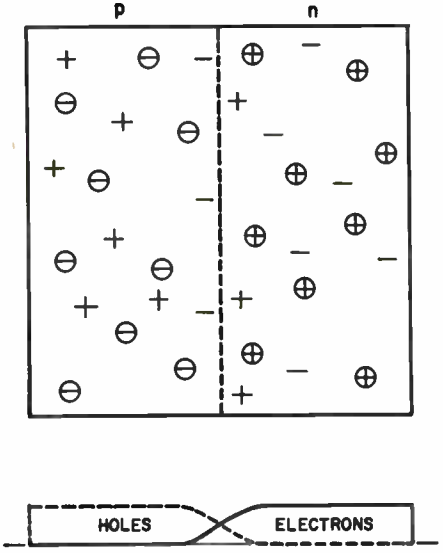


Fig. 502. *Here the p and n materials are in close crystalline contact, and there is some straying of holes and electrons across the boundary, setting up a difference of electric potential between the two sides.*

charges have arrived on the n-side but also because negative charges have gone away and left an equal number of *fixed* positive charges unneutralized. The same thing happens in reverse on the p-side, which goes negative.

Another way of looking at this event is as a flow of current from p to n. (A movement of holes in that direction and a movement of electrons in the *opposite* direction both add up to it.) Because there is no complete circuit, so that this current cannot circulate, the n-side is bound to become positively charged.

A limited aggression

The positive charge repels holes, and soon prevents any more crossing from the p-region; meanwhile the negatively charged

p-side brings to an end the invasion of electrons. So, very quickly, a balance is established between the straying or diffusing tendency of the mobile charges and their natural reluctance to approach their own kind. A potential difference of a few tenths of a volt is enough to balance these conflicting desires.

If you are familiar with vacuum tubes, you will see a likeness

Fig. 503. *Compare an electronic tube, in which there are no holes, but electrons stray across to the anode, charging it negative.*

between this p-n behavior and what happens in a diode when the cathode is heated but it and the anode are not externally connected (Fig. 503). Some of the electrons emitted from the cathode cross the vacuum and land on the anode, charging it negatively. When this charge has built up sufficiently, it prevents any more from following. The difference of potential can easily be measured by connecting a vacuum-tube voltmeter between anode and cathode. It is usually around 1 volt.

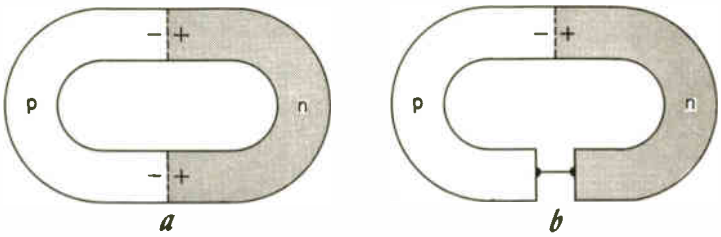
The mystery of the missing voltage

The p-n junction differs by also having positive current carriers emitted from its "anode", but this just adds up to the same result. Another difference is that if you try to measure the potential difference with a vtvm you will get no reading and may wonder if what you have just been told is true. The reason is that the contacts between the metal wires and the two kinds of semiconductor also develop potential differences which together cancel the potential difference at the p-n junction.

That this must be so can be seen by extending the p- and n-material into a complete circuit, as in Fig. 504-a. Obviously there must be a second junction between them, which will develop a potential difference that just cancels the first in one complete trip around the circuit. Putting in a section of metal wire (b) makes two junctions instead of one and, while we may not know the potential difference at each one of them, we do know that together they must be equal to the one they replace in diagram (a). Otherwise, perpetual motion would be a reality, for this circuit has no heater to supply energy as had the tube in Fig. 503. The same principle applies to a more complicated circuit with a voltmeter in it.

One-way current

With Fig. 503 back in view, remember what happens when you pass current through a diode by means of a B-battery. If, as usual, you connect the positive pole of the battery to the anode, as in Fig. 505-a, it will attract the movable charges of opposite sign – the

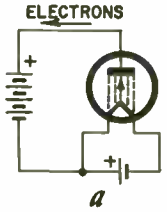


a *b*

Fig. 504. *Whenever there is a complete circuit including a p-n junction, there must be at least one other junction, at which there is a difference of potential canceling the first (a). The same applies when there is more than one other junction (b).*

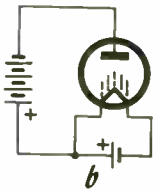
electrons – and keep them moving around the circuit as an electric current.

But if you reverse the battery, as in Fig. 505-b, no current at all flows. The electrons shot off from the cathode by the energy



a

Fig. 505. *In a diode tube, a positive charge on the anode attracts electrons across and sets up a current (a); negative repels and holds them back, so no current (b).*



b

of its high temperature find themselves confronted by an anode of the same polarity (negative), which repels them. So, as Fig. 505-b shows, they never get far from the cathode.

In a p-n junction, the n- half plays the part of the tube cathode by providing movable electrons. So, when the p-part is made positive, it attracts them across the boundary and an electron current flows around the circuit just as with the diode.

Again, in the same way as with the diode, a negative p-region repels the electrons and no current flows. The positive n-region cannot attract electrons across from the p-region, because after the few strays have been mopped up there are none; the p-region is a source of holes only, which are repelled. The only negative charges are fixed atoms.

A p-n junction, then, acts like a vacuum diode with respect to electrons. But, in addition, it has holes. When the p-end is made positive, it repels holes towards the n-end, which, being negative, attracts them. So, while electrons are flowing across the junction from n to p, holes are flowing from p to n, and both lots mean electric current from p to n and then around the circuit back to p.

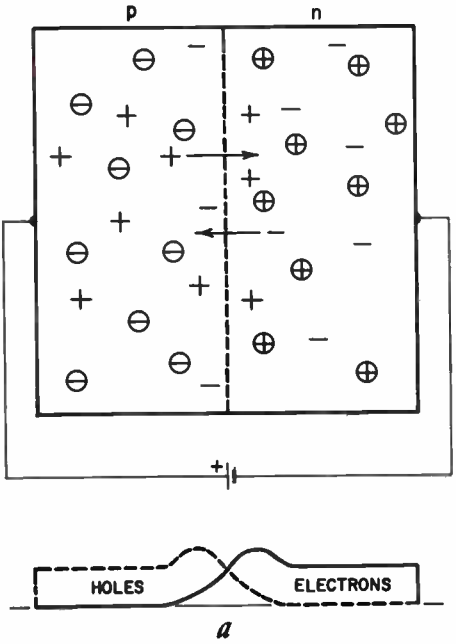


Fig. 506-a. *The situation with a p-n junction is rather similar to Fig. 505-a, except that besides the electrons from n there are holes from p to add to the current.*

On the other hand, a negative p-end cannot attract holes across from the n-end, because none are produced there.

Like a vacuum diode, then, a p-n junction lets current flow in only one direction; in short, it is a rectifier. This is of great practical value, as we shall see in the next chapter. Meanwhile, there are still a few things to note on the theoretical side.

What happens to the holes?

Anyone who knows how a vacuum diode works can easily understand a p-n junction, by remembering that the n-half is a source of electrons, like a heated cathode. The new feature is that the p-half is a source of holes, which by flowing in the opposite direc-

tion to the electrons adds to the current when the positive pole of a battery is connected to p. But, remembering that there are no holes in metals, you may be wondering what happens to the holes

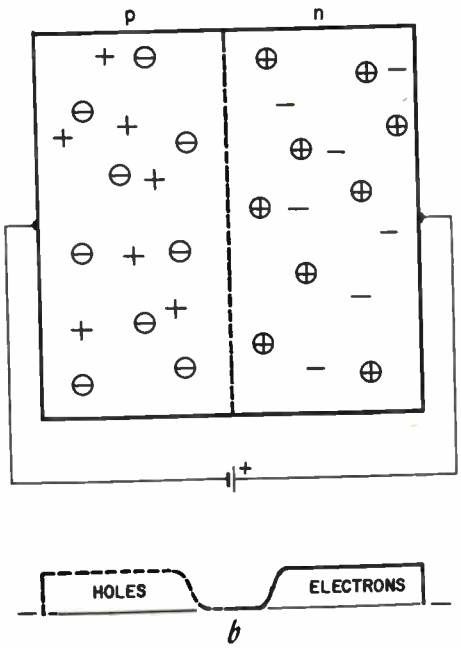


Fig. 506-b. *With the transposition of the battery, the electrons and the holes are held back. Compare this with Fig. 505-b.*

when they have crossed the semiconductor. We must think about the whole thing in a little more detail.

In the tube circuit (Fig. 505-a), the wire and battery contain practically the same number of electrons all the time; what is happening is that those which are free are moving around in the direction of the arrow, and the shortage this would otherwise create on the anode is being made up by those arriving across the vacuum. Meanwhile, a surplus at the cathode is avoided by emission of electrons into the vacuum.

Much the same thing happens in Fig. 506-a, except that when electrons meet holes they fill them. Their two opposite charges then cancel, leaving nothing. This process is called *recombination,* and usually takes places within a few thousandths of an inch after crossing the boundary. Farther away from the junction, almost all the current in the p-half is being carried by holes, and in the n-half by electrons. To make up for the holes continually moving

out of the p-half, new ones are created by withdrawal of electrons at the end held positive by the battery.

Absence of space charge

In Fig. 506-a, the voltage of the battery is used up mainly in driving current through the resistance of the semiconductor. It was not just by accident that only one battery cell appears here, compared with several in Fig. 505-a. As that diagram shows, the space between cathode and anode contains nothing but electrons, so is charged negatively. From the point of view of electrons just leaving the cathode, these come between and weaken the positive attraction of the anode. So, to make a current flow, the voltage there has to be higher than it would have been without the negative space charge.

Now you can see how helpful the fixed atoms in the semiconductor are — they neutralize the charges carried by the mobile electrons and holes, preventing the formation of a space charge. So less voltage is needed to make the same current flow. That is

Fig. 507. *The intrinsic conduction of the germanium, being due to both electrons and holes everywhere, is not dependent on polarity so can be represented by a shunt resistance.*

one reason why semiconductor devices are cheaper to run than vacuum tubes. Another reason, of course, is that they need no cathode heating.

The diagrams so far have shown equal numbers of impurity atoms on each side of the boundary. That was just to make things easy. There is no necessity for it, and in fact the manufactured products usually have unequal percentages of impurity. We shall see why in Chapter 7.

Intrinsic conduction again

Another thing about the diagrams in this chapter so far is that they show only impurity atoms and the electrons or holes detached from them. (A hole detached really means, of course, an electron attached). The germanium or silicon itself has been serving merely as a sort of framework into which the few impurity atoms have fitted themselves here and there. All this is nearly as things actually would be at very low temperatures — hundreds of degrees F below zero. As you can easily believe, the heat energy at such a tempera-

ture isn't sufficiently intense to kick a significant number of germanium electrons 0.7 eV up into the conduction band, much less any silicon electrons up 1.2 eV. But it is enough to drive practically all the impurity atoms up their mere 0.01 eV gap.

However, most of us are more interested in reasonably comfortable temperatures, and perhaps also the rather higher temperatures liable to occur in electronic equipment. These energize a

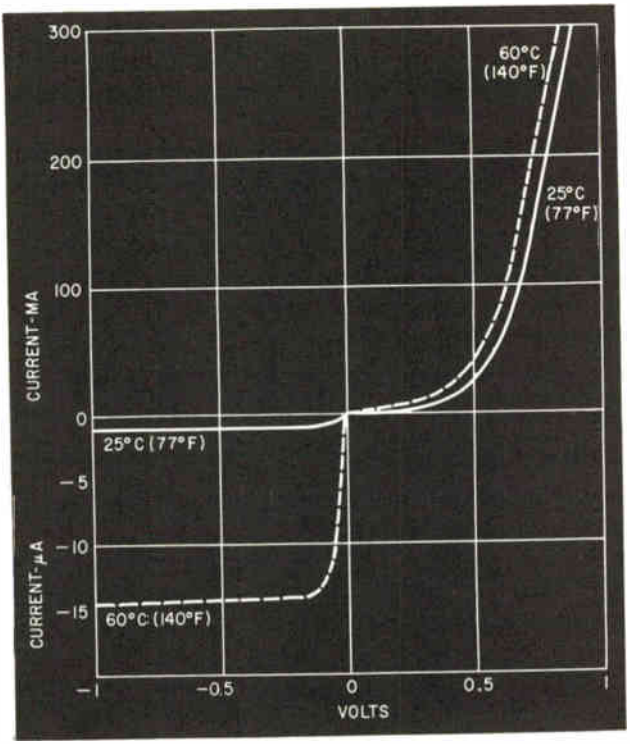


Fig. 508. *Current/voltage characteristic curves for a germanium p-n junction at two different temperatures. Note that to show the small reverse current its scale has to be in microamperes. This reverse current is far more affected by temperature than the forward current, because it is all intrinsic conduction.*

small proportion of the semiconductor atoms, making their electrons free to conduct. The vital difference between them and the energized impurity atoms is that every electron set free creates a hole, which is a movable positive charge — not a fixed positive atom. So opposite current carriers are created in equal numbers throughout the material. We have just seen that rectification

depended on only one kind of current carrier being present at each end. These intrinsic carriers, being present at both ends, respond to voltages either way —they don't rectify.

Impurity conduction plus intrinsic

Because impurity conduction gives perfect rectification, we can represent it by the usual triangular rectifier symbol in Fig. 507. The both-ways intrinsic conduction can then be represented by a resistance shunted across it. But this is not ordinary resistance, in two respects.

First, its resistance falls very rapidly when the temperature goes up. We know why.

Second, the current through it is not in direct proportion to the voltage, as in Ohm's law. The reason is that its few intrinsic carriers — the minority carriers — can't just ignore the crowds of impurity carriers. We have already noted that their presence reduces the number of minority carriers by providing many more opportunities to unite with the opposite kind.

But there is more to it than that. During times when the voltage is "forward" (i.e., making impurity current flow), the intrinsic carriers add to this flow but, being relatively so few, their contribution is usually negligible. When the voltage is "backward," however, they are the only current there is, so they are important. If it were not for the other carriers, lined up on opposite sides of the junction as shown in Fig. 506-b, the applied voltage would be distributed proportionately over the whole crystal and would devote itself exclusively to making these minority carriers flow. Doubling the voltage would double the current. But, as we will see in a moment, the voltage is mostly at the junction itself, in keeping the majority holes and electrons apart. The result is that anything over about 0.1 volt backward causes little increase in the backward current, which is limited by the number of germanium atoms excited at that temperature. This state of affairs is like what happens when the heater of a tube is underrun; above a certain voltage the current "runs into saturation" and cannot be increased without raising the heater temperature. Saturation of the backward current in a semiconductor diode is shown by the left-hand part of the curve in Fig. 508.

Obviously the intrinsic current is an imperfection, and one that grows very rapidly with temperature. The temperature that makes it serious is considerably higher for silicon than for germanium, because of its bigger energy gap. That is why silicon is chosen for semiconductor devices which have to work at high temperatures.

Other imperfections

Fig. 507 is roughly equivalent to a p-n junction with dc. What about ac?

When considering Fig. 506-b, we saw that, when a backward or reverse voltage is applied there is no continuously flowing im-

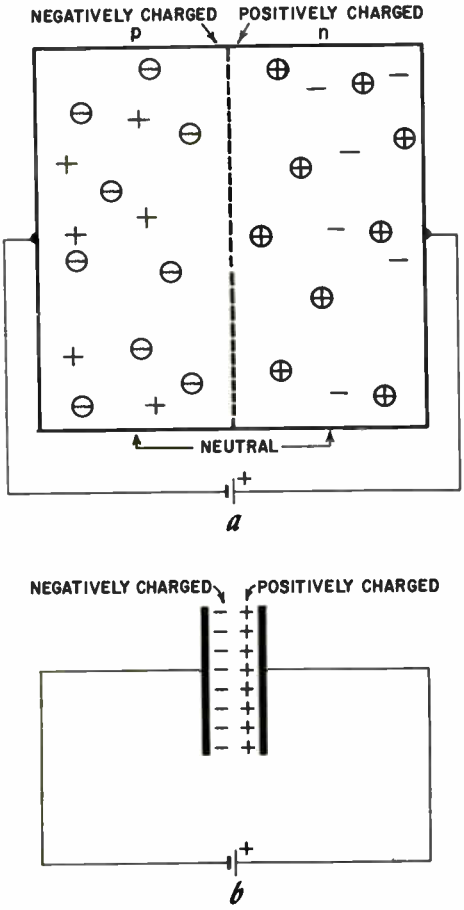


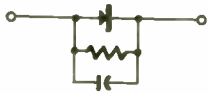
Fig. 509. *Diagram (a) duplicates Fig. 506 (b), and is here compared with the similar situation in a charged capacitor (b).*

purity current. Electrons and holes are attracted a little way towards the positive and negative poles, and stop when enough fixed charges have been uncovered to set up a back voltage equal to that applied. This is shown again in Fig. 509-a. Now compare it with Fig. 509-b, where the same voltage is applied to a capacitor, charging it. In both cases, the reason why there is no current is

that the battery voltage is exactly balanced by an opposite pair of charges. The same sort of thing happens when you push a locked door; the door pushes back at you with exactly the same amount of pressure, which corresponds to voltage in a circuit.

So although a p-n junction doesn't pass any steady current in the reverse direction, (except of course the usually small intrinsic or

Fig. 510. *The capacitance effect explained in Fig. 509 (b) can be represented by adding a shunt capacitor to Fig. 507.*



"leakage" current) there is current as long as the reverse voltage is increasing, just as if the junction were a capacitor. So, when the voltage is alternating, there is some alternating current because of what seems like the capacitance of the junction. For ac, then, we must add a capacitance to Fig. 507, as in Fig. 510.

Because the charged layer is so thin, even a small area of junction has quite a large capacitance. A typical value is 50 μμf per square millimeter! So it has a serious effect on the performance of a junction diode at high frequencies.

But that isn't all. Before there can be a switchover from the current flow during one half-cycle to no current (except for the apparent capacitance) during the other half, the holes and electrons that have strayed a little way beyond their own frontiers, as in Fig. 506-a, have to be pulled back. This withdrawal means that, as soon as the voltage reverses, some current flows momentarily in the wrong direction. Although in this respect like the capacitive current, it differs in not continuing to flow as long as the reverse voltage is increasing.

Only when both sides of the junction have equal proportions of impurity do holes and electrons penetrate beyond the boundary in equal numbers. Often, the p-half is more heavily "doped" with impurity, in which case the effect is due mainly to holes and is called *hole storage*.

A useful imperfection

One way in which the apparent capacitance of a junction differs from that of an ordinary capacitor is that it varies greatly with the voltage applied. The greater the backward voltage, the farther apart the charges are separated (Fig. 509-a). It is as if the plates of the capacitor (b) were pulled a little apart. So the higher the voltage, the *less* the capacitance.

This variation is shown in Fig. 511 for a typical p-n junction. It can be put to good use as a variable reactor in automatic frequency correction and in "mavars" – see Chapter 9.

There is also quite a substantial capacitive effect in the forward direction, but the resulting current is usually unimportant

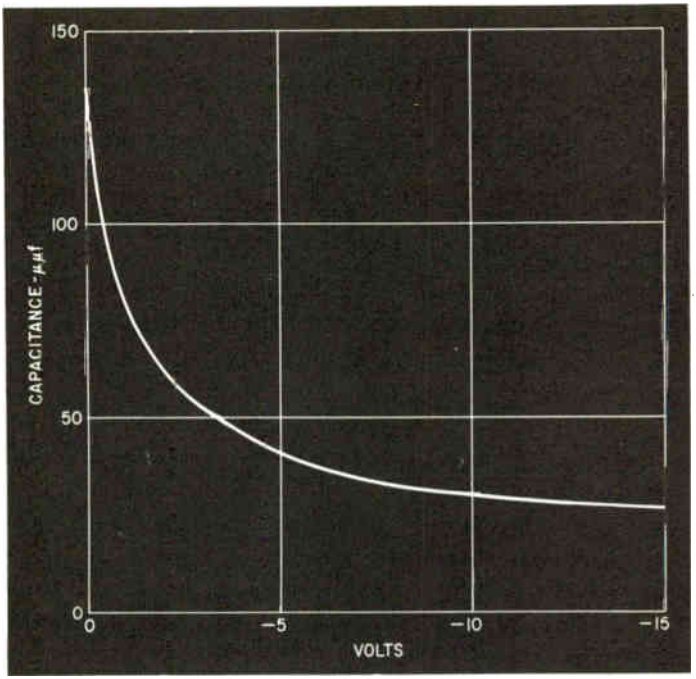


Fig. 511. *The capacitance of a p-n junction varies with the reverse voltage like this.*

in comparison with the large forward current. In other words, the forward capacitance is shunted by the comparatively low forward junction resistance.

Breakdown

If you keep on increasing the voltage across a capacitor, sooner or later it will break down. In Chapter 3, we saw that this happened when the voltage applied was sufficient to overcome the main energy gap (about 0.7 volt per electron in germanium). If you remember that there may be millions of germanium atoms for every impurity atom, you will realize that this will be like the bursting of a dam compared with the trickle of normal working current. The resulting short circuit may be enough to damage the

junction permanently. But if there is enough resistance elsewhere in the circuit to prevent this, the breakdown point (which may be perhaps 100 volts for a typical p-n junction, but depends on the amount of impurity, etc.) may be quite useful, as we shall see in Chapter 9. A junction specially designed for this purpose is called a Zener diode, after the scientist who investigated the effect.

It is believed that breakdown is often started by what is called the avalanche effect. This is caused by the intrinsic carriers (which make up the reverse current) being driven so hard by a high reverse voltage that they knock out more electrons from the germanium crystal, adding to their numbers and starting a chain reaction.

Semiconductor-to-metal junctions

A whole circuit made of semiconductors is hardly a practical proposition (Fig. 504-a was just to bring out a theoretical point). So there are bound to be junctions between semiconductor and metal. They are the other main subject of this chapter. Often it is these junctions, rather than junctions between two different kinds of semiconductors, that do the work. Although they have been used very much longer than p-n junctions, they are still not so well understood. In the old crystal-and-cat'swhisker days, few people even tried to understand them — they were always too busy trying to find a more sensitive spot on the crystal!

In our diagrams, such as Fig. 506, we have already seen some semiconductor-to-metal contacts, without giving them much attention. We did guess, with the help of Fig. 504, that there would be voltage steps at those places; otherwise, a p-n junction would be able to drive a current around a circuit, delivering energy without receiving any — nice work if you could get it! But we didn't think out what might cause such steps.

We also presumed that electrons could flow from a metal into a semiconductor or the reverse in the usual way. And that holes, although there are none in metal, can flow to and from metal. To be quite sure about this last one, perhaps we had better look at it more closely.

Fig. 512 shows a metal and semiconductor in close contact. The metal is thick with minus signs, to remind us that it has about 10^{20} of them per cubic millimeter. This enormous negative charge is neutralized by the atoms from which they have broken loose (not shown). The semiconductor has both electrons and holes, in proportions depending on the impurity content, but in any case comparatively few. The electrons pose no problem, so are not

shown. But we must remember that the holes which are shown are not the same lot all the time – they are continually coming into existence by atoms being stripped of electrons, and then disappearing again by uniting with (probably other) electrons. It is, in fact, like the current list of unmarried film stars – rapidly changing through divorce and remarriage.

Suppose now that the semiconductor is connected positively and the metal negatively. This makes holes move to the left and electrons to the right. When the holes arrive at the junction and

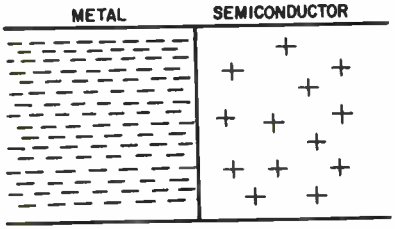


Fig. 512. *In a metal-to-p semiconductor or p-m junction, electrons are vastly more crowded on the metal side than holes on the semiconductor side.*

meet electrons in greater numbers than they had ever imagined in their wildest dreams, they are quickly filled by them. The disappearance of negative charges in the metal is of course instantly made good by the arrival of electrons from the left.

If the polarity is reversed, so that electrons move to the left and holes to the right, this might be expected to create a no-man's land at the boundary, as in Fig. 506-b. As soon as it did, the metal side would be left positive, which would tend to attract electrons from the semiconductor, creating new holes. So the shortages on both sides are continuously replenished. Holes and free electrons stream from the boundary in opposite directions, because they are created there.

Why didn't that happen with our p-n junction in Fig. 506-b? Or, if you prefer, granted that it doesn't happen with the p-n junction, why does it with what we might call a p-m junction (m for "metal")?

An essential difference

That is a tricky question. It involves the relative energy levels on each side of the junction, in a way that is not easy to show. But perhaps the essential difference between the two situations can be put like this. With a p-n junction, when holes and electrons have

been swept away from the zone close to the boundary on both sides — the no-man's land — we are left with a few *fixed* charges (the impurity atoms) standing up here and there in a sea of perfectly ordinary pure semiconductor atoms. And we know that movable bits can't be broken off them by any voltage short of breakdown. So, except for the few broken up by heat or other

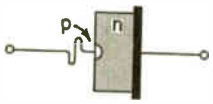


Fig. 513. *A point contact m-n junction is essentially a p-n junction, but there are complications not yet fully understood.*

disturbances, there are no current carriers and therefore no current.

When the n-side is replaced by metal, however, all the atoms on that side have at least one mobile electron each, and there is no appreciable energy gap between them and the semiconductor atoms on the p-side. So when the metal atoms nearest the boundary have their electrons drawn away into the interior, their positive charge easily entices electrons away from the semiconductor atoms on the other side, creating new holes there.

What has just been described is called an ideal ohmic junction, because it passes current equally either way and in proportion to the applied voltage. It doesn't rectify. It would be used for making circuit connections with semiconductors. But unless special methods are adopted, actual contacts between metal and semiconductors seldom behave in this way. There is usually some degree of rectification. That is particularly so when the metal makes contact at a point.

The reasons for this are still not certain, but it does seem clear that in many cases — perhaps most of them — rectification takes place because what looks like a metal junction is really a disguised p-n junction.

Point contacts

The modern germanium and silicon point-contact diodes — the successors of the old crystal detectors — are examples. A small wafer of n-type semiconductor is fitted with a springy wire contact, as in Fig. 513. Before use, the diode is "formed" by passing through the contact a strong pulse of current which has the effect of converting the material nearest the point to p-type. So it is really a p-n junction in which the contact area — and therefore its capacitance — is much smaller than in the ordinary junction type. This is very helpful for high frequencies.

The small contact area means, of course, that it cannot pass

much current without getting too hot at its vital point, but diodes are rarely required to handle much current at very high frequencies.

The first transistors also were of the point-contact type, but are now obsolete. However, point-contact diodes are used very much, especially in television and radar receivers.

Plate contacts

Even when relatively large areas of metal and semiconductor are placed in contact and no special forming is adopted, the possibility of p-n action cannot be ruled out. Unless special precautions are taken, most metals have tarnished surfaces, which may them-

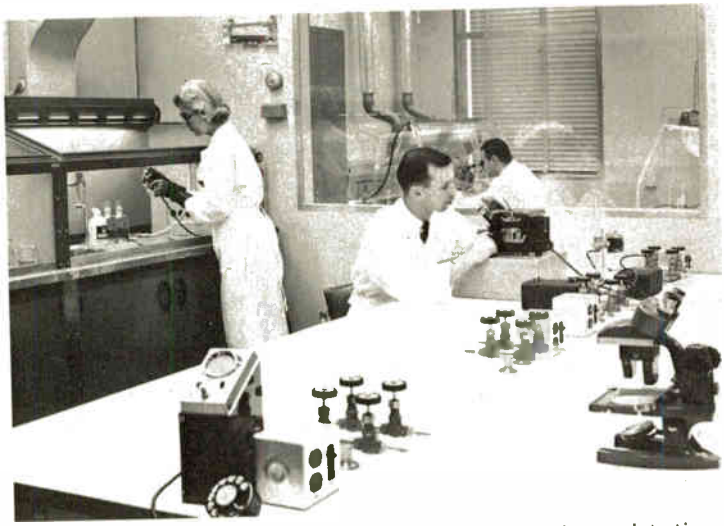


Fig. 514. *Cleanliness is essential in transistor production and testing.* (Bell Telephone Laboratories)

selves be semiconductors. Since, as we have seen, fantastically small changes in the composition of semiconductors are enough to make all the difference electrically, the results obtained by ordinary contacts between metals and semiconductors are largely a matter of chance. That will be no news to the old-timers of the crystal detector days!

However, it was found that by keeping to strict manufacturing techniques, consistent rectifying performance could be obtained with what at least appeared to be contacts between metal and semiconductors, especially selenium and copper-oxide. The details of these will appear in the next chapter, but at this stage it can be said that rectification takes place at what is called the *barrier layer.*

Although it seems that p-n action is not necessarily involved, the tendency is to think that it often is. One thing certain is that the barrier layer is very thin – typically a few millionths of an inch.

Fig. 514 shows the almost-surgical environment needed for testing semiconductors in the laboratory.

Ohmic contacts

Of course, if there were rectifying contacts on *both* sides of the semiconductor, they would stop current flowing either way, and that would be cheaper to achieve by having no contact at all! So at least one contact in every semiconductor device has to be of the "ohmic" description. Like the rectifying contacts, these are more the result of factory know-how than theoretical predictions. Certain soldering and welding techniques, using certain materials, are found to yield such contacts, and that is perhaps all that can be said about it in simple terms. Some details of the practical production of semiconductor materials and junctions are included in the next chapter.

chapter 6

diodes and rectifiers

THE title of this chapter really needs a little explanation. Technical terms have come into use one by one over many years, like the buildings in an old city. The result – the technical language or the city – is very different from what it would have been had it been planned as a whole for present requirements.

A question of name

When electron tubes started developing into many types, they were classified according to the number of electrodes, using names consisting of "ode" preceded by the appropriate Greek numeral; thus, "triode" for a tube with three electrodes, "pentode" for one with five, and "diode" for one with two. For some not very obvious reason, the name "diode" was more often attached to tubes designed for rectifying high-frequency signals (demodulators) than to those for rectifying 60-cycle ac for supplies, which were commonly called rectifiers. For an even less obvious reason, when semiconductor devices began to develop along similar lines, this tendency to distinguish between diodes (for signals) and rectifiers (for power) became more fixed, so that neither term alone is generally understood to mean all one-way two-electrode devices, however designed and used. And these are just what this chapter is about. Any distinction there may be between the two main applications affects the details of construction rather than the principle. In fact, some types may be used for either signals or power. In what follows, then, either name may involve both kinds of device or their use of them.

Two-electrode devices in which rectification is either absent or subsidiary are included in Chapters 8 and 9.

First let us review the uses of rectifiers of all kinds, then see what other kinds there are and how semiconductors compare with them, and finally, get down to the various types and constructions of semiconductor rectifiers and consider which is best for each use.

Rectifiers for power

Perhaps the most obvious use is for converting ac to dc for power supply. Because of the ability to step the voltage up and down efficiently, nearly all electric power is distributed as ac, most often at 50 or 60 cycles. On the other hand, radio and electronic equipment, and many other things such as electroplating, need power as dc. One can, of course, use a rotary machine — a motor generator — to convert from either to the other, but such machines are expensive and make a certain amount of noise. So nearly always the choice is a rectifier. The requirements range from microamperes to many thousands of amperes, from a volt to many thousands of volts, giving power from milliwatts to megawatts.

In a television receiver, for example, one rectifier is needed to supply less than 1 ma at say, 15,000 volts, and another to supply 0.25 ampere at 260 volts. Transmitters need power of quite another order.

Battery chargers for cars are familiar, but do we think about them for trains, aircraft and electrically-propelled vehicles? Incidentally, battery circuits use rectifiers additionally as reverse-current cutouts. Factories need rectifiers for their dc-operated equipment — magnetic chucks, lifting magnets, arc welders and electroplaters. Then there are such things as elevators, movie projector arcs, aircraft starters, magnetizers, cable and capacitor testers, and many other power applications. One of the less familiar ones provides high voltage (e.g., 60,000) for getting rid of dust by charging it electrically.

Fig. 601 shows some of the power rectifier circuits best known in radio and electronics. There are also polyphase systems for high power. One of the most used of all is the simplest — Fig. 601-a. It is the hardest to smooth, but has the advantage that equipment using it works on dc supplies as well as ac. So it is used for most radio and some television receivers.

Signal applications

Every receiver needs at least one rectifier as a demodulator or

"detector" to reduce radio-frequency currents to audio and video frequencies. Most of them need a frequency changer too, and at the highest frequencies there is nothing so effective for this as a diode rectifier.

Demodulators imply modulators, and although the modulators used in most radio transmitters are multi-electrode tubes, diode

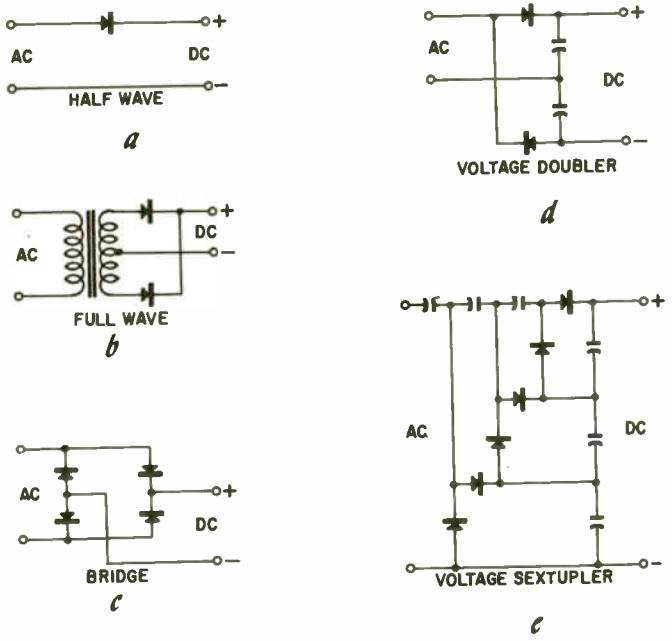


Fig. 601. *Five types of rectifier circuit.*

modulators are used on a large scale in such communications services as carrier-current telephony. Line telephony, too, is making increasing use of rectifiers for incidental purposes such as echo suppressors, surge limiters and frequency doublers.

Electronic circuits have even more numerous incidental uses for diode rectifiers, as limiters, clippers, clamps, and so on, according to the ingenuity of the designer.

Lastly — or is it? — they figure largely for instrumentation. Dc meters and relays in general are so much better than ac types that most of those used for ac are really dc types with a rectifier. We have only to think of the technician's multirange meter.

Semiconductors vs tubes

Apart from rotary machines and vibrators, which have only very limited use and obvious disadvantages, the only serious competitors of semiconductor rectifiers are vacuum and gas-discharge tubes. Of these, the gas tubes resemble vacuum tubes in appearance and construction, but are more like semiconductors in containing charged material (gas) to neutralize the space charge, which is such a hindrance in vacuum tubes, especially for heavy currents at low voltage. But gas tubes share the fragility and limited life of vacuum tubes, and in addition are a little more fussy about such things as temperature and starting conditions. So they are seldom used for domestic equipment, in spite of their much higher efficiency. Tubes of both kinds are usually smaller and lighter than the older kinds of semiconductor rectifiers for equal duties in the higher-voltage power ranges, but the opposite is true of signal diodes and low-voltage rectifiers, and generally for modern silicon rectifiers. Vacuum tubes have the least reverse current of any, and are simple and compact for high-voltage low-current uses.

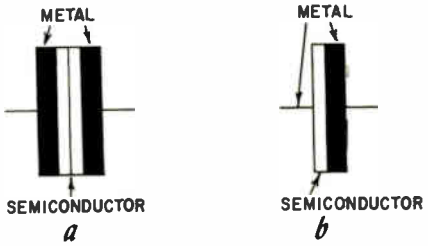
In all other respects – ruggedness, unlimited life (if properly made and used), absence of cathode heating, and adaptability – semiconductor types have the advantage. Price depends largely on commercial considerations outside the scope of this book, but the trend would seem to be in favor of semiconductors. In the foreseeable future, tube diodes will probably be very much the exception.

Semiconductors used in diodes

Many different semiconductors have been and still are being tried for rectification. One has only to think of the innumerable combinations tried – and many patented – during the crystal detector era. At that time, the theoretical basis was virtually nil, and the experimental basis was not far removed from that of pure chance.

As a result of the sudden increase of theoretical and practical knowledge of semiconductors since about 1943, the use of germanium and silicon has vastly increased – for rectifiers as well as transistors and other devices. And a new urge to try other materials has broken out, but on a much more intelligent basis than the urge of 1906–26. However, at the time of writing, the only semiconductors of commercial importance for rectifiers are germanium, silicon, selenium and copper oxide.

Selenium is by far the oldest, for its rectifying properties were discovered in 1876, and at the present moment it may still be rectifying more power than all the rest put together. Silicon came



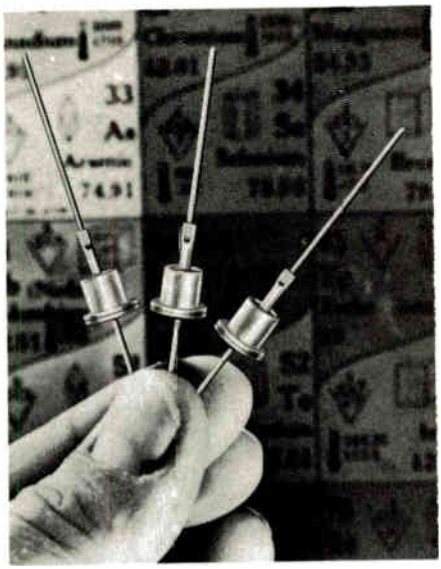


Fig. 602. *In some commercial rectifiers (a) two different types of semiconductor are sandwiched between metal contacts; in others (b) one semiconductor is applied, but a surface layer with other characteristics may form. Silicon-diodes are shown in the photo.* (International Rectifier Corp.)

next, for it dates from 1906 in a crystal detector, but, due to lack of both theoretical and practical knowledge, it is only during the last few years that it has come to the fore. Copper oxide was developed commercially in the 1920's as the first semiconductor rectifier to be manufactured in quantity, but except for special purposes it has been largely superseded by selenium. Germanium

dates from the same time, but was little used until about 1940 and not very much then for some years.

The theory of rectification has been covered in the previous chapters, especially in relation to germanium and silicon, because they are more clearly understood than the other two. There is not much difference in general form; all are sandwiches in which the "bread" is metal and the filling is a piece of semiconductor. This piece, by design or accident, usually has two different layers — either two different semiconductors, or the same semiconductor with different impurities (Fig. 602-a). If so, the metal pieces on each side are there simply for making electrical connection with the semiconductor, and their contacts should both be of the ohmic or non-rectifying kind. However, where only one kind of semiconductor is used, one metal contact should be ohmic and the other rectifying. Though there seems to be some doubt as to whether this type isn't really the other in disguise, from a practical viewpoint there is a distinct dividing line between p-n junction types with their two-layer semiconductor and point-contact types with one (Fig. 602-b) — at least regarding germanium and silicon. (The photo shows silicon diodes.)

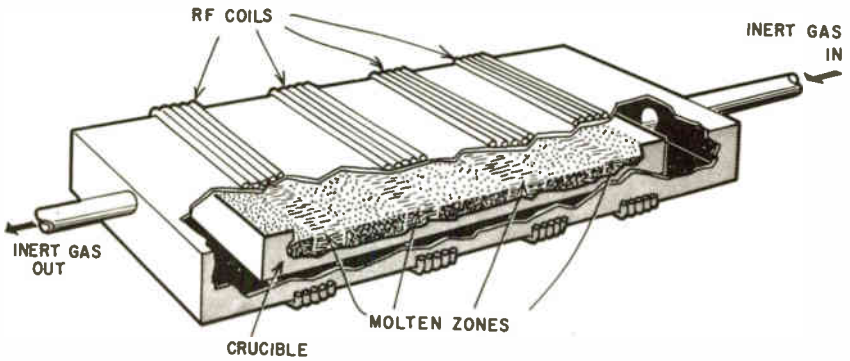


Fig. 603. *Section showing the essentials of zone-refining equipment for purifying germanium or silicon.*

Selenium and copper-oxide rectifiers more closely resemble the conventional three-part sandwich in construction.

Manufacture of semiconductor crystals

As we have seen, the working of germanium and silicon rectifiers depends entirely on the correct kind and quantity of impurity to an exactness beyond chemical analysis. The first procedure, then, is to obtain the germanium (or silicon) in a much higher degree of purity than in normal commercial practice; that is to

say, with less than one part in one billion of impurity. Then the right amounts of the right impurities are added in the right places.

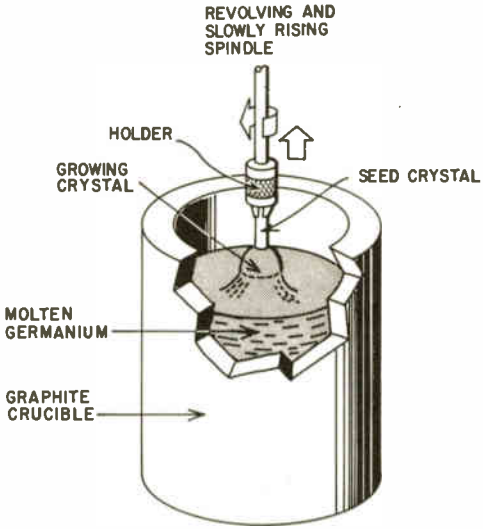


Fig. 604. *The essentials of the usual (Czochralski) method of forming large semiconductor crystals.*

Another requirement is that the material be in the form of a single large crystal, rather than the jumble of small ones that form when crystalline materials are allowed to solidify in the usual

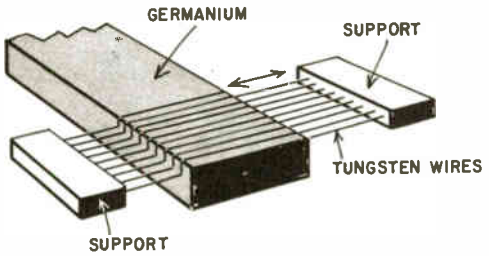


Fig. 605. *Semiconductor crystals are chopped into thin slices by a multiple "saw" consisting of parallel tungsten wires loaded with abrasive.*

way. The reason is that predictable satisfactory performance is obtainable only when there are as nearly as possible no breaks or irregularities in the crystal "lattice" which would release valence electrons for current carrying.

When chemistry has done its best, further purification or refining depends on the fact that when the molten material begins to solidify, the impurities tend to remain in the molten part. The usual process taking advantage of this is called zone refining. A rod of germanium is placed in a graphite or quartz boat-shaped crucible and surrounded by an atmosphere of hydrogen or nitrogen to prevent oxidation. The crucible is ringed around at intervals by high-frequency induction coils, which induce sufficient current in the germanium to melt it at those places, as shown in Fig. 603. By moving the set of coils along the crucible (or the crucible through the coils) these molten zones pass slowly along the germanium and sweep the impurities towards one end, which is finally cut off and mixed in with another batch for refining. The purified part is cut up and the pieces are "etched", i.e., chemically cleaned by removing the surfaces.

Although now pure enough (impurity about 1 part in 10^{10}), the germanium is still a mass of small crystals. It is melted down again in a different sort of furnace (Fig. 604) and a small but perfect crystal, called the seed, is lowered into it. The temperature must be controlled precisely, so that the molten germanium begins to solidify slowly around the seed crystal, enlarging it by adding atoms all with the same pattern and alignment. Meanwhile the seed is kept slowly rotating and rising out of the crucible.

In this way, it is possible to grow single crystals of germanium as much as 2½ inches in diameter by 10 inches long, weighing more than 10 pounds and providing material for many thousands of diodes or transistors.

These pure single crystals are very expensive, and the next problem is to cut the material into slices perhaps less than 1/50 of an inch thick without wasting a lot of it. One method is to saw it with a piece of tungsten wire 0.005 inch in diameter, loaded with abrasive dust to give it "teeth." A whole array of such wires, spaced apart by the desired thickness of the slices, is motor-driven backwards and forwards, as shown in Fig. 605.

Formation of junctions

In one method, called the doped-grown process, the production of junctions is combined with the crystal-growing process just described. A carefully measured amount of impurity, say arsenic, is added to the molten germanium, so that when it solidifies it forms n-type material of the desired concentration. After this has gone on for some time, a pellet of acceptor impurity, say indium, is melted in. The amount is chosen so that the part of

the crystal which then begins to form is p-type. The amount of indium is enough to cancel the arsenic by compensation and leave some over as acceptors. So a p-n junction is formed across the whole area of the crystal, while preserving the same lattice structure throughout.

Later still, by adding arsenic in more than sufficient quantity to cancel the indium, another junction can be formed, and so on.

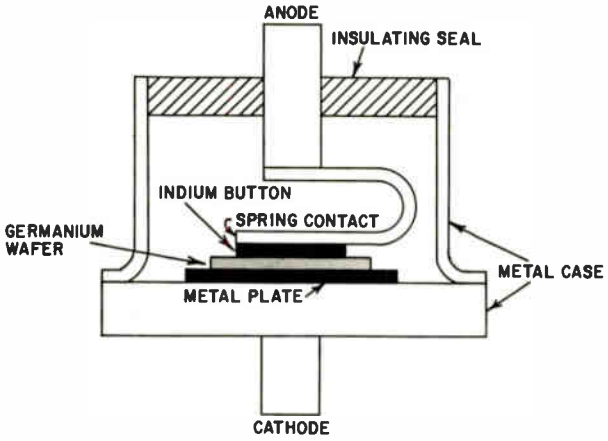


Fig. 606. *Section of a germanium power rectifier.*

But there is a limit to this sort of thing, for the material soon becomes so loaded with both kinds of impurity that it would not work as intended. In practice, only a few reversals can be made, so the crystal is not used in the most economical manner.

To avoid this disadvantage there is a modification, called the meltback process, in which both kinds of impurity are added to the whole crystal, which is then cut up into small bars. If one end of a bar is melted and then allowed to solidify, something rather like zone purification on a small scale occurs, to different extents with the different impurities, and the original proportions are altered. By suitable choice of impurities, one end of the bar finally turns out to be n-type and the other p-type.

An alternative process produces what are called alloy junctions. Usually, n-type germanium is cut up into slices the correct size for rectifiers of the desired rating, and a small piece of indium is placed against one side of each slice. The temperature is raised so that the indium begins to melt and the germanium in contact dissolves in it. On cooling, the germanium recrystallizes with some of the

indium atoms carried along with it, providing a p-layer to form a p-n junction.

The alloy process is a popular one and has displaced the grown type for germanium, but not so much so for silicon, which takes

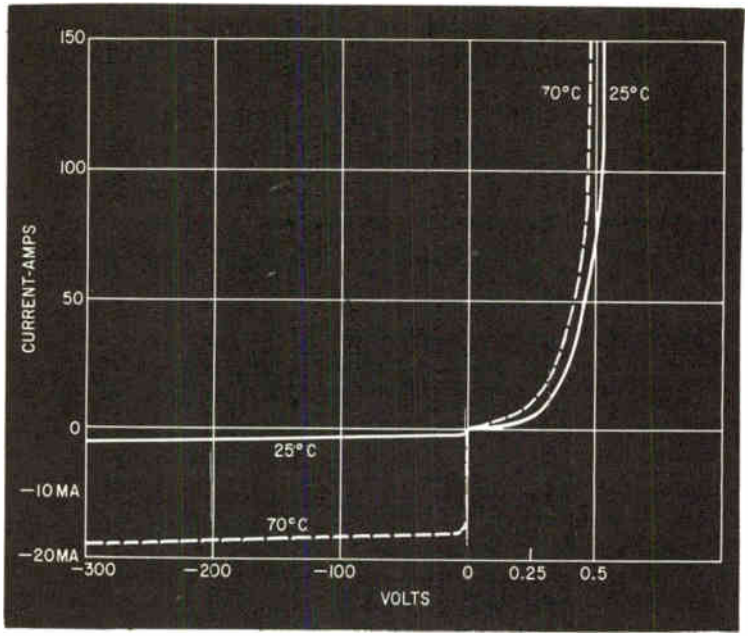


Fig. 607. Characteristic curves for a high-power germanium rectifier at two different temperatures. Note the changes of both current and voltage scales on going negative (reverse), and the extremely favorable rectifying properties.

to alloying less easily. There are a number of other methods of producing p-n junctions, but one of them will be enough for now — the diffusion process. Again, the germanium is first cut up into wafers, one for each junction. These are heated nearly to melting point in an atmosphere of the opposite kind of impurity as a vapor. The impurity atoms diffuse slowly into the solid crystal, converting the surface layer into p-type (if the crystal is n-type).

Although it takes many hours to convert a layer only 0.001 inch thick, the slowness allows the depth to be controlled accurately within millionths of an inch, and many junctions can be made at one time. This process is especially convenient when junctions of large area have to be made for handling heavy currents. For example, it is used to make silicon rectifiers rated at 100 amperes.

Ohmic junctions

As we have seen, it is necessary to have nonrectifying connections between each type of semiconductor and the metal connecting wires or terminals. One technique is to use solder containing

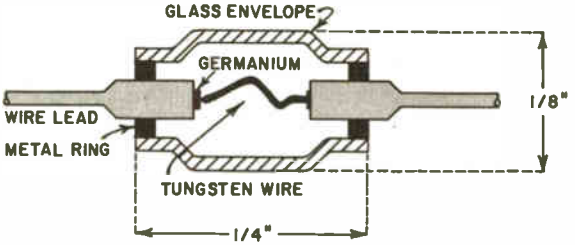


Fig. 608. *Enlarged section of a typical point-contact germanium diode, as used for rf detection.*

a relatively large proportion — say 1% — of the same kind of impurity as the crystal to which connection is being made. These soldering materials are called p+ type or n+ type, as the case

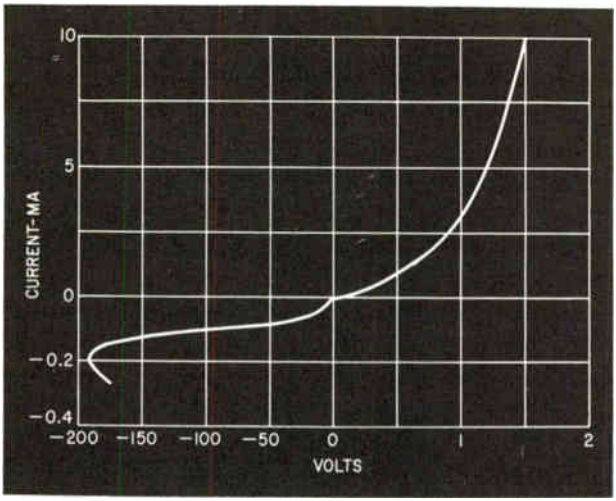


Fig. 609. *Characteristics curve for the type of diode shown in Fig. 608.*

may be, so the metal-to-crystal joint is a pp+ or nn+ junction.

An alternative process, suitable for very small devices such as high-frequency signal diodes, is bonding. A gold wire is made to adhere to a thin metal coating on the crystal by a combination of heat and high pressure.

Final processes

After the junctions have been formed, the external surfaces of the crystal must again be etched and then washed with very pure water, to remove any contamination which might short-circuit the semiconductor junction or cause deterioration and short life. Water vapor is particularly harmful, so not only must the device be dried thoroughly in a dust-free atmosphere, but it must be protected from moisture by hermetic sealing, or "encapsulation", in its container. Fig. 606 shows a section of a high-power germanium rectifier, and Fig. 607 characteristic curves for two temperatures. Note the low forward voltage required even for very heavy currents.

As may be imagined in the light of the impurity requirements of the crystals, all the processes leading to the complete products must be done with cleanliness greatly exceeding that in a surgical operating theater. The manufacturing plant must be specially air-conditioned throughout, and all the staff wear special clothing. Some operations are carried out in "glove boxes" filled with inert gas, often with the help of microscopes. Automation is being adopted to reduce costs and improve uniformity of products.

Point-contact rectifiers

For high-frequency signal circuits, the capacitance of even a small junction diode is too much, and point-contact types are preferred. Fig. 608 shows a typical construction. A piece of n-type germanium, about 0.05 inch square by 0.02 inch thick, is soldered to a metal base by rf heating. It has a pointed, springy tungsten wire pressing on its face. Forming, done by passing a current of about 1 ampere for 1 second with the point positive, produces a very small p-type zone around the point. Fig. 609 shows a typical characteristic curve for this kind of diode, many millions of which have been produced since about 1946. They are used mainly in television receivers and computers. Cheaper to make than junction types, they are preferred wherever acceptable, especially for computers, which may each contain as many as 10,000 of them.

The greater the impurity, the lower the resistance and the better the performance at very high frequencies, but the lower the back voltage that can be withstood. For example, some silicon point-contact diodes are suitable for frequencies up to 10,000 mc, but they can stand only a few volts and pass appreciable reverse current. Others, suitable only for comparatively low fre-

quencies, can stand over 1,000 volts in reverse and have a reverse resistance of many megohms. Using what might be described as subsubminiature or microscopic construction, point-contact diodes built into tiny waveguides have been successful at 70,000 mc (just over 4 mm wavelength).

Silicon rectifiers

A major advantage of silicon over germanium is its greater energy gap, which enables it to work at a much higher temperature — say 250°C instead of 100°C — without excessive back

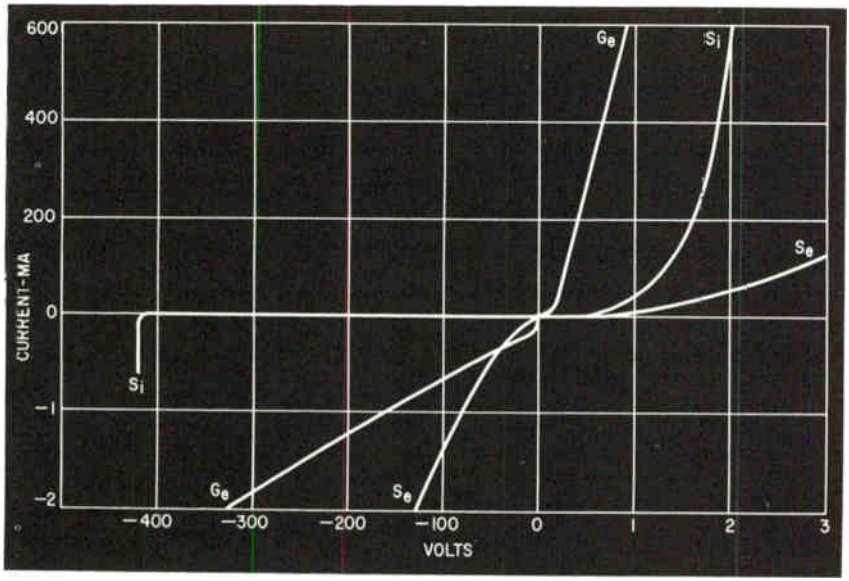


Fig. 610. *Here the rectifying characteristics of germanium (Ge), silicon (Si) and selenium (Se) are plotted together for comparison. Temperature, 55°C.*

current. But it requires about double the voltage to drive the forward current, which means that its power loss is greater and efficiency lower. Its higher melting point — 1,420°C instead of 940°C — and its chemical affinities, make it more difficult to prepare.

The processes are similar to those described for germanium, with extra precautions. One difficulty is to find a material for the crucible in the crystal-pulling apparatus (Fig. 604) which doesn't contribute impurity at the high working temperature. A recent idea is to use the silicon itself, in the solid. To enable silicon to

be melted in it without melting the whole lot, the heat is generated by bombardment from electron guns as in a cathode-ray tube. Of course, the whole process has to be carried out in a vacuum.

Copper-oxide rectifiers

The performance of copper-oxide rectifiers depends largely on the precise nature of the materials and processes used. In fact, success or failure is said to hang on whether or not the copper came from Chile! In view of the known influence of minute amounts of impurity, that is not really incredible.

In broad outline, manufacture consists in stamping out discs of this very pure copper and heating them in air at a temperature of about 1,000°C for a few minutes (or 500°C for several hours), giving them a coating of red (cuprous) oxide on one side. There is no difficulty in making contact with the clean copper side, of course, but the sort of contact made with the oxide is of vital importance. Various materials and methods have been used, including graphite, lead, nickel and even gold, pressed, sprayed, plated or evaporated on.

This kind of rectifier resembles germanium in not liking high temperatures, so except for low-current uses such as in instruments it is usual for the rectifying discs to be interleaved with large brass discs to serve as cooling fins. They make this kind of rectifier rather bulky for power applications.

Selenium rectifiers

Selenium, like silicon, can stand higher temperatures than copper oxide, and passes less back current, but has a higher voltage drop. Manufacture differs because of the nature of the material. Each "cell" consists of a base plate, usually of steel or aluminum, which is given a thin coating of nickel or bismuth. This is to permit a good contact to be made with the selenium, which is next applied. Having a low melting point — 217°C — it can be put on in liquid or at least plastic form. After other heat treatment, an electrode of some such material as cadmium sulfide is applied. The reason for the choice is that in contact with the selenium it forms a thin layer of a compound of both, which is the essential "barrier layer". In some types, an intermediate material is applied to form this layer.

Electrical forming is done to increase the backward resistance, by passing pulses of current somewhat as for point-contact diodes.

Choice of rectifier

One feature common to all four types when properly constructed is that apart from electrical overload or penetration by moisture, etc., their life is so long that no limit has yet been experienced. If any has an advantage in this respect it is selenium,

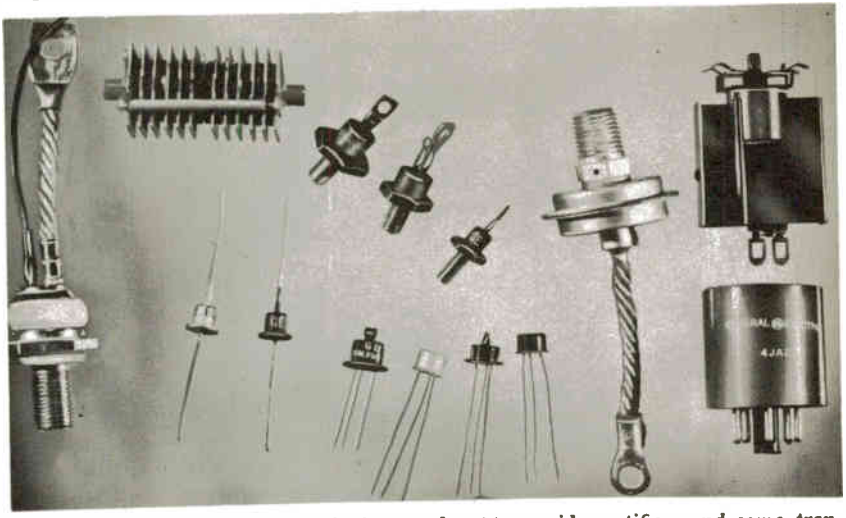


Fig. 611. *Silicon, germanium, selenium and copper-oxide rectifiers and some transistors.*

because it can recover from momentary "punctures" caused by pulses of excessive back voltage. (This feature is shared to a less extent by copper oxide.) Moreover, it is second only to silicon in its ability to stand high temperatures.

Silicon is smaller and lighter than selenium for the same power rating, has less reverse current, can work at a higher temperature and at higher frequencies. So selenium may, in time, be superseded by silicon.

Except for its temperature limitation and higher reverse current, germanium is even better than silicon, because it requires less than half the voltage to drive current through it, and so less power is lost in the rectifier. That partly makes up for its lower temperature limit. The efficiency (which means the proportion of the ac power put in that comes out as dc) is as high as 98.5%, — a remarkable figure. Germanium has had a head start over silicon on account of its being easier to process, and will probably hold its place except where temperature would be too high or a lower reverse current is required. It is particularly indicated for

heavy-current low-voltage rectification. In Fig. 610, which compares selenium, silicon and germanium, silicon shows up strikingly to advantage in its reverse characteristic, germanium in its forward characteristic, and selenium is inferior to either at both ends.

Copper oxide has an advantage over selenium and silicon in its low forward voltage drop and (when aged) stable characteristics. So it has been much used for instruments and control purposes. But it seems likely to be displaced by germanium, which among other things can be used at higher frequencies.

It is never very safe to prophesy, but it does seem that germanium and silicon will soon displace copper oxide, selenium, vacuum and gas tubes, and perhaps even the mercury-arc rectifiers used for high power.

Fig. 611 shows silicon, germanium, selenium and copper-oxide rectifiers. Compare these with the four transistors shown in the lower center of the photo.

chapter 7

transistors

WHEN we remember how beneficial Lee de Forest found putting a third electrode into a vacuum-tube diode, as far back as 1906, it is surprising that few people seemed to have thought of doing the same thing with a crystal diode and not until 1948 did anyone have any success with it. That was when Bardeen and Brattain of the Bell Telephone Laboratories did so, earning for themselves a Nobel prize.

These first transistors were point-contact types, which are now practically obsolete, and moreover difficult to understand, so we shall begin with the junction transistor, which dates from 1950.

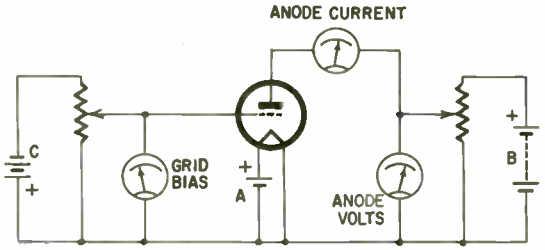


Fig. 701. *This is a usual arrangement for plotting triode tube characteristic curves.*

Assuming we know how vacuum triodes work, the easiest way to understand semiconductor triodes is by comparison, noting that in a general way they correspond to tubes but have some important differences.

Comparison with tubes

Fig. 701 shows a familiar tube setup, with an A-battery to heat

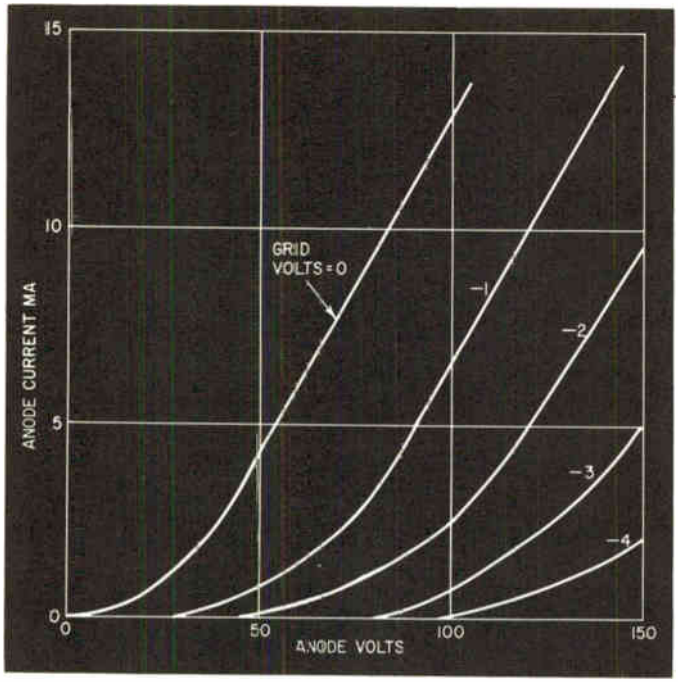


Fig. 702. *Anode current/voltage curves for various grid bias voltages plotted as in Fig. 701.*

the filament or cathode, providing its electrons with so much energy that many of them escape from the surface like bubbles of steam from boiling water. The B-battery provides a positive attraction for these electrons, pulling them across the vacuum to the anode, where they continue around the circuit, making the milliammeter indicate current. The C-battery provides negative voltage bias to the grid. This is the wrong polarity to attract electrons, so no current flows to the grid; in fact, the negative grid more or less cancels the attraction of the positive anode. Being nearer to the cathode, a few volts there can offset many volts at the anode; hence, the ability of the tube to amplify. By varying anode and grid voltages and noting the current readings, one can plot a set of characteristic curves, as in Fig. 702.

Now compare the corresponding transistor circuit, Fig. 703. The transistor itself is, in effect, a double junction diode — a thin layer of p-type material sandwiched between two n-types. So it is called an n-p-n transistor. There are also p-n-p transistors, but the n-p-n is easier to compare with a tube.

Transistor electrodes

One difference we soon see: there is no A-battery. For the reasons given in Chapter 4, n-type material contains vast numbers of electrons ready for action without any heating. So that is a big saving already, to say nothing of the convenience of not having to wait for a cathode to heat. This n-piece is called the *emitter* — a name that would at least equally well suit the tube cathode.

The other n-piece, corresponding to the anode, is called the *collector* — again, a very suitable alternative to "anode" or "plate".

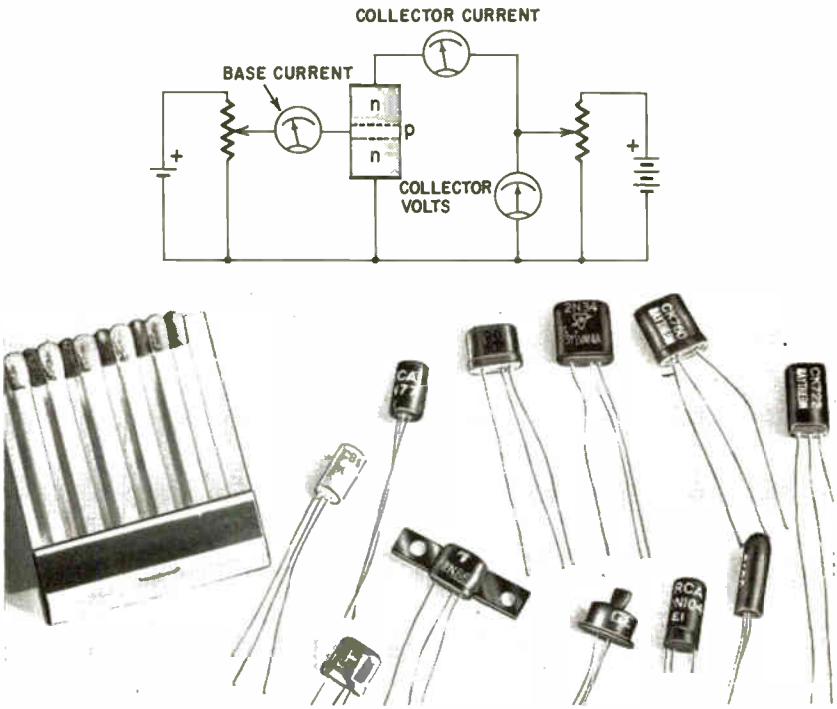
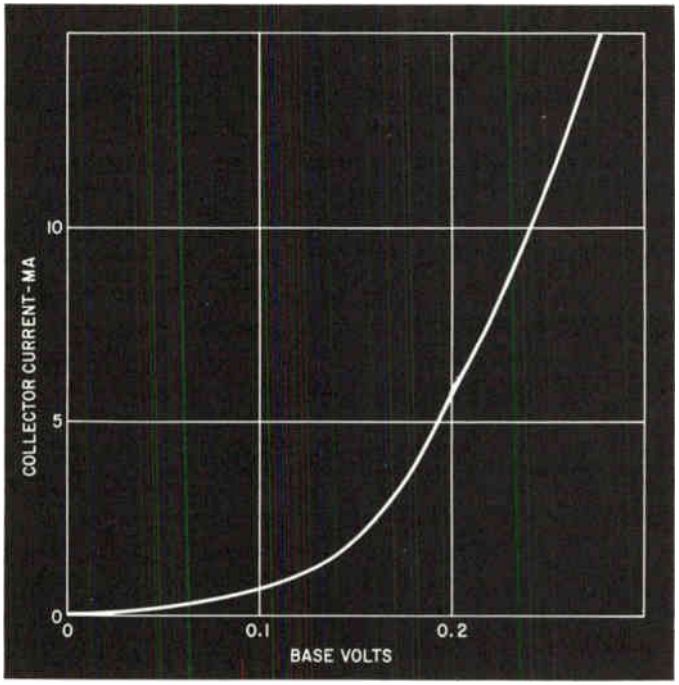


Fig. 703. *The corresponding set-up for plotting transistor curves. The photo shows a few of the many transistor shapes.*

Ignoring the connection to the left, we see that with whichever polarity a voltage is applied between emitter and collector, it will cause hardly any current to flow, because there are two junction diodes back-to-back so one or the other of them must be in the reverse direction. With the collector positive as shown, it is the upper diode that holds back current. The stoppage is because there are no free electrons in the middle layer, which is p-type material, nor holes in the top layer, which is n-type.

The middle layer, for historical rather than functional reasons, is called the *base*. If it is made positive with respect to the emitter, as in the diagram, current can flow around that circuit because the emitter and base form a diode in the right direction. The emitter provides the electrons which stream across the frontier into the base. These electrons tend to keep on flowing in the same direction and soon find themselves crossing the other frontier into the collector, where they are welcomed by the positive collector voltage. To insure that most of the electrons go that way, rather than into the base circuit, the base is made extremely thin — about 1/1,000 inch or less. So the electrons have hardly crossed into the base before they are through it and into the collector. And, whereas the emitter is usually heavily "doped" with impurity to insure that plenty of electrons are available, the base is lightly doped, so that few holes are there for the electrons to drop into and be lost as far as the collector is concerned. On the other hand, low base impurity means high base resistance, which

Fig. 704. *The mutual conductance curve of a transistor is far from straight, so it is usual to regard collector current as controlled by base current.*



is undesirable; so a compromise must be made between these conflicting requirements.

Transistor characteristics

In practice, usually 95–99% of the emitter current is snapped up by the collector. This figure, slightly less than 1, is usually denoted by the Greek letter alpha (α). Looking at the same fact in another way, we can say that the collector current varies 20

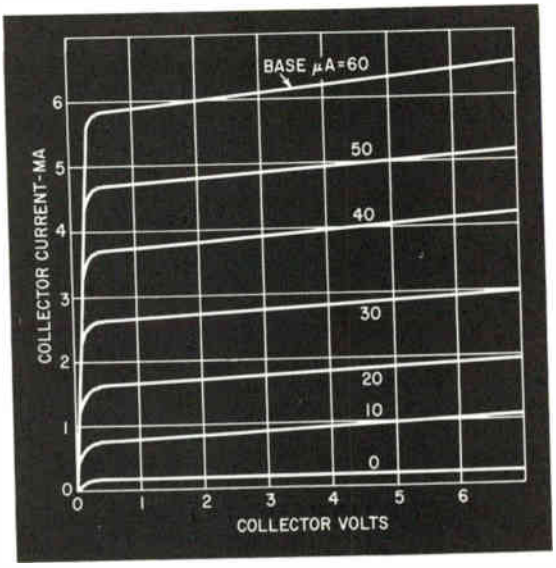


Fig. 705. *These transistor curves correspond to the tube curves in Fig. 702. The squarer shapes indicate much higher efficiency.*

to 100 times as much as the base current. This figure, which represents the current amplification factor of the transistor in such a circuit, is often denoted by the Greek letter beta (β).

We could, of course, plot collector current against base voltage as in Fig. 704, just as we do for the corresponding tube quantities to calculate the mutual conductance. We see, however, that the two things are not at all in direct proportion, so that the transistor mutual conductance is not even approximately constant. On the other hand, as Fig. 705 shows, collector current is very closely proportional to base current, so β is very constant.

Here, then, are some more contrasts to note. In place of negative grid voltage to control anode current downward, we use positive base current to control collector current upward. And,

although the transistor whose characteristics are plotted in Fig. 705 is only a triode, its curves are like those of a pentode tube. In fact, in several ways they are much better. Whereas a pentode needs 50–100 volts on the anode before the curve straightens out into the useful horizontal part (and then isn't very straight), the transistor begins to be useful with no more than a fraction of one volt at the collector. Also, instead of the curves for equal intervals of grid voltage getting closer and closer together as anode current is reduced, so that there is distortion if one works down to anywhere near zero, the base-current lines are evenly spaced right down to zero. All this makes a transistor much more efficient than a tube. Because there is no vacuum space-charge to overcome, a collector battery can have a much lower voltage than a tube B-battery for a given output. And the base voltage is much less than the usual grid bias.

More advantages and disadvantages

All the foregoing information about n-p-n transistors can be applied to p-n-p types by reversing all the polarities and substituting "holes" for "electrons." The fact that these opposite types can be produced is one of the most important advantages of transistors over tubes, because it simplifies symmetrical circuits — such as push-pull — and enables circuits to be devised that would be impossible with tubes.

Other advantages of transistors are their small size, rugged construction and indefinitely long life.

On the other hand, from a tube user's point of view, the fact that the transistor's input not only takes current but even has a much lower impedance than the output looks unfavorable. For some purposes, this certainly is a problem, but most often

Fig. 706. *This circuit symbol for a transistor is appropriate for the point-contact type, but less so for junction types.*

it can be largely overcome by circuit design. Another complication, as we shall see, is that the input and output circuits are not so free from interaction as with tubes. Transistors are more difficult to manufacture in quantity between close limits. And germanium types, at least, are more affected by high temperature.

Symbols

In the original point-contact transistor, there were two wire contacts (emitter and collector) impinging close together on a

block of semiconductor (the base). The symbol shown in Fig. 706, in which the emitter is distinguished by an arrow head, was therefore quite appropriate. It should certainly be used whenever one wants to indicate a point-contact transistor in a circuit diagram. Unfortunately, it continues to be used also as the symbol for a junction transistor, which it does not resemble.

Note that this symbol indicates p-n-p transistors. To show an n-p-n type (Fig. 707) reverse the arrow head.

The three configurations

There are three basic ways in which tubes can be connected — see Fig. 708. (This applies to triodes, and also tetrodes and pentodes when the extra electrodes are not used for signals but

Fig. 707. *Symbol for n-p-n transistors.*

are connected to fixed-potential points.) The most usual is (a), known as the common (or grounded) cathode configuration. It has the advantages of giving greatest amplification, and of the

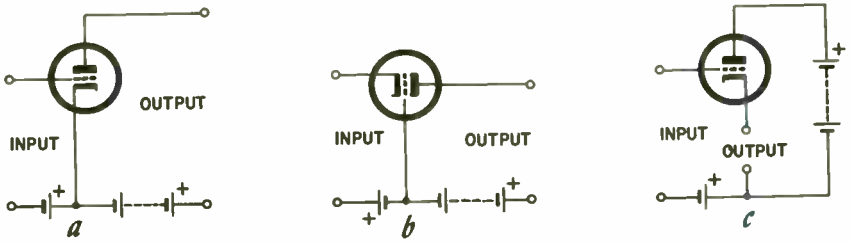


Fig. 708. *The three basic tube configurations: (a) common cathode; (b) common grid; and (c) common anode (or cathode follower).*

input (when properly biased) not passing dc. For special purposes, such as amplification at very high frequencies, the common grid configuration (b) is sometimes used, in spite of its comparatively low input impedance. The common anode arrange-

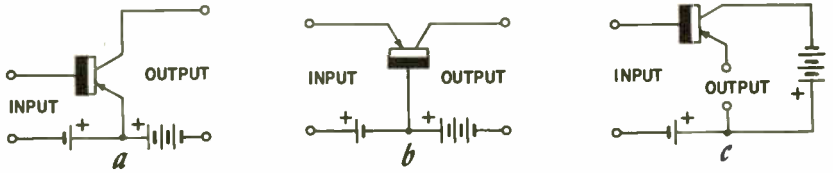


Fig. 709. *The three corresponding transistor configurations: (a) common emitter; (b) common base; (c) common collector.*

ment (c), better known as the cathode follower, gives no voltage amplification at all, but is much used for its very high input and low output impedances.

Transistors have three corresponding configurations: common emitter, common base and common collector (Fig. 709). As the transistors shown are p-n-p, the batteries have opposite polarities to those in Fig. 703. The circuit connections are not the only things that correspond to the tube configurations; their comparative behavior is much the same too.

Unfortunately the original point-contact transistors were first used in common-base circuits (b), with the result that their parameters were defined for that configuration. For example, the input/output current ratio in this case is the emitter-current/collector-current ratio, denoted by α. But quite soon transistor

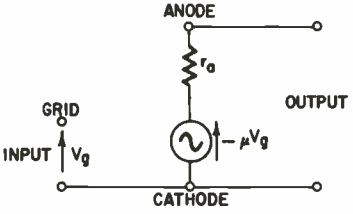


Fig. 710. *The most usual equivalent circuit for a triode tube.*

circuitry followed tube usage, so that the common-emitter configuration (a) is now by far the most used. It is a pity that the opportunity was not taken of making the primary parameters apply to it just as tube parameters such as μ and g_m apply to common-cathode circuits, those for the other configurations being expressed in terms of them.

Equivalent circuits

In studying tube behavior, much use is made of the "equivalent circuit" or "equivalent generator." There are only two basic ones; the more common of them is shown in Fig. 710, in which v_g stands for the input signal voltage. (In the other equivalent there is a current generator in parallel with r_a and the output.)

When we turn to this aspect of transistors, we are somewhat discouraged to find that there are almost as many varieties as Heinz! In fact there may be more, and some of them are very complicated, notably those devised to represent transistor behavior at high frequencies. The subject is too involved to go

into fully here, but one sample (Fig. 711) will serve to bring out an important distinction between tubes and transistors.

This sample can be described as the series-voltage T-circuit equivalent of the common-emitter configuration, and is about the simplest possible. Incidentally, it illustrates what a pity it was that the parameters were not originated for this configuration, so that the collector resistance could be plain r_c, instead of our having to borrow r_c from common-base and multiply it by $(1 - \alpha)$. Similarly for the generator voltage. It also shows, of course, the conductive input path, via r_b and r_e, which are both quite low (typical values: $r_e = 20$ ohms; $r_b = 600$ ohms). But the main object of gazing at this diagram is to take in the fact that the input current not only causes an amplified output current by way of the

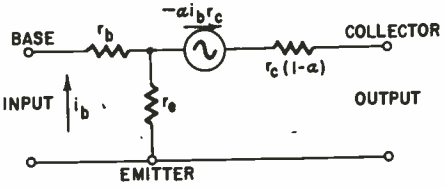


Fig. 711. *One of many transistor equivalent circuits — a T type for the common-emitter configuration.*

imaginary generator in the output circuit (as does the input voltage in Fig. 710), but it also affects the output directly by its voltage drop across r_e. More significant still, the output current, likewise flowing through r_e, directly affects the input. Except in certain circumstances where these effects are very small, they greatly complicate calculation of transistor performance and circuit design. This contrasts with tubes in which, at low frequencies at least, there is almost complete separation between input and output.

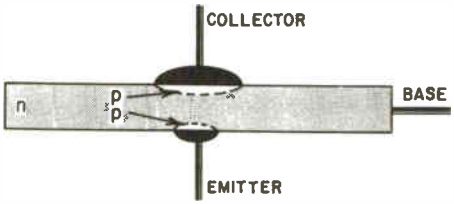
Correspondingly, in place of the simple tube parameters, μ, g_m and r_a (only two of which need be specified, because the third follows) at least four are needed, and different people specify different sets of four, and even the same sets are not everywhere denoted by the same symbols. It is necessary to study this aspect of transistors at some length, and preferably in more than one book, to be sure of understanding the data sheets of all makers.

Manufacture

The techniques used for diodes apply to a large extent to transistors, too. The production of low-impurity single-crystal semiconductor material involves doping it with precise, small pro-

portions of impurity, cutting it into small sections, making non-rectifying contacts for terminal wires, removing contamination and sealing in capsules. The main additional requirements are to insure that the material between the two junctions — the base — is sufficiently thin and yet hasn't too high a resistance between its working part and the base terminal. The processes for achieving this on such a microscopic scale must be capable of turning out large quantities reasonably similar in performance.

The grown-junction technique has been used for producing considerable quantities of n-p-n transistors, but if the base layer is to be thin enough for good results there is obviously a difficult



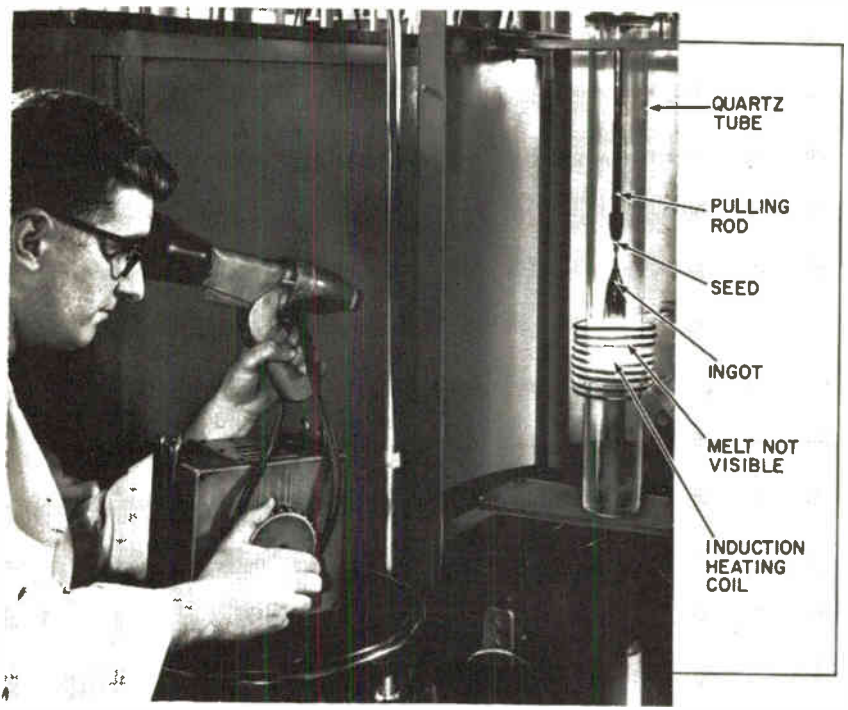


Fig. 712. *Section of an alloyed or fused-junction transistor. The photo shows a method for growing a single germanium crystal.* (Raytheon Mfg. Co.)

problem in attaching a connection to it. In form, a grown-junction transistor is not unlike the one shown in Fig. 703.

More favored is the alloy-junction construction, used for both n-p-n and p-n-p types, shown in Fig. 712. A thin wafer (say) of n-type germanium has small pellets of the opposite kind of impurity material (such as indium) applied to both sides. The temperature is raised sufficiently for some of the indium to alloy with the germanium, forming p-type regions as shown. By carefully proportioning the dimensions, temperature, and time of application, the unaffected layer between — which is the working part of the base — can be made to the desired thickness. The inactive part is much thicker and therefore puts less resistance into the connecting lead.

These are only the two most elementary methods; new ones are continually being developed.

Transistors for high frequencies

The speed of electrons through solid material is much slower than across the vacuum of a tube, and holes are about twice as slow as electrons. So even when the base of a transistor is made as thin as 1/1,000 inch, the time taken to cross it prevents the flow from being varied at very high frequency. Even if there were no other limitations, this would set an upper limit of a few megacycles for germanium and only about 1 megacycle for silicon.

There are, moreover, capacitances between the various electrodes, and these vary considerably with working conditions. And the storage effect, explained in the previous chapter, puts another restriction on high-frequency operation.

The obvious ways of increasing the possible operating frequency are first, to make the base thinner and, second, to reduce the emitter and collector junction areas. The second method is

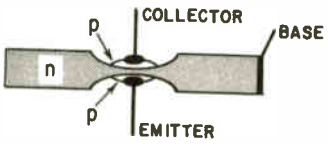


Fig. 713. *Section of surface-barrier transistor, in which the base is hollowed out from each side so that the current carriers have a shorter and therefore quicker journey across.*

not difficult, but of course it reduces the power rating. The first requires ingenious methods of manufacture if it is to be carried far enough to make transistors workable at vhf and beyond. Fortunately, halving the thickness increases the frequency fourfold.

Surface-barrier transistor

One of the first methods resulted in what is known as the surface-barrier transistor. It is somewhat similar to the form shown in Fig. 712, except that before the emitter and collector are applied the center part of the base is tunneled away from both sides by squirting tiny jets of etching liquid at it, until they are perhaps as near as 0.0001 inch to joining up. The collector and emitter are then deposited on this thin web; Fig. 713. If they are alloyed on, the base is thinned still more; this is called micro-alloying.

One advantage is that although the base is so thin where the carriers have to cross it, it is thick enough elsewhere to ensure a

Fig. 714. The double-diffused technique enables transistors to be made with extremely thin bases.

reasonably low-resistance connection to the base terminal. This technique is commercially successful, and a cut-off frequency of about 70 mc is obtainable. The power rating is low, because the collector voltage has to be kept down to perhaps 6 volts to prevent it from "punching through" the thin base.

Diffused transistors

In the previous chapter we saw how a very thin layer of a crystal can be converted to the opposite type by gaseous impurity diffusion, forming a p-n junction. The same process can be modified to produce the two junctions needed for a transistor. Let us suppose that it is to be an n-p-n type. A wafer of n-type material is used, and one face of it is exposed to vaporized impurities of both donor and acceptor elements. They are so chosen that the acceptor diffuses faster than the donor and goes ahead to make a p-type layer. The donor element follows behind, but in sufficiently greater concentration to "compensate" the acceptor and restore the material to n-type.

At this stage the result is as shown in Fig. 714. The main body of the original wafer, unaffected by the diffusion, forms the collector. An ohmic contact is made with it in one of the usual ways. The single diffused p layer, which may be only 0.00005 inch thick, is the base; and the double diffused n layer is the emitter.

There is the delicate problem here of making a connection with the base, which is about 1/30 the thickness of a sheet of India

paper. One solution is to use a material which penetrates through the emitter layer but makes a reversed-diode junction with it, so that it is virtually unconnected to it. Its junction with the base, on the other hand, is of a highly conductive kind.

Mesa transistor

In a later variety of diffused transistor, commercially available, only the base layer is diffused into the collector, and a gold contact is made to it. The emitter is then alloyed on by its side, and the active base area is reduced by etching away the material,

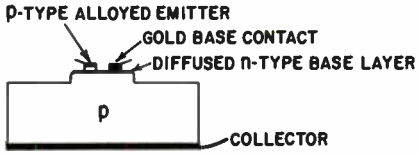


Fig. 715. *Section of mesa transistor, in which another arrangement is adopted to speed transit across the base and thereby raise the possible signal frequency.*

leaving a sort of plateau, as in Fig. 715. Because of its resemblance to the geographical formation of that name, it is called the mesa transistor.

For high-frequency transistors, the diffusion process has important advantages besides enabling the base to be made as thin as one pleases. Instead of the acceptor impurity being uniformly distributed through the thickness, it is most concentrated at the boundary with the emitter and least at the collector, which was its farthest line of advance. This helps in two ways. It results in less collector-to-base capacitance, and it causes a potential gradient through the base, of the right polarity to speed the electrons on their way to the collector. This contrasts with the situation in a uniform base, where there is consequently almost no electric field, so the electrons (or holes) can only diffuse across it slowly, like a drop of ink in a glass of water.

Drift types

Transistors having this current-accelerating feature, whether obtained by diffusion or alloying, are called (rather inappropriately, one might think) drift types. If it is obtained by diffusion, they are sometimes known as graded-diffused. The result of such techniques is the possibility of operation at frequencies up to or even beyond 1,000 mc. Obviously such thin layers are very limited

in the voltages they can stand, and a typical maximum rating is 6 volts.

Intrinsic transistors

Another way of reducing collector-to-base capacitance is to interpose a thin layer of "intrinsic" (negligible impurity) material between them. This kind of transistor is referred to as n-p-i-n (or p-n-i-p) type.

Tetrode transistors

Still another way of improving high-frequency performance is to use a tetrode transistor. How this helps is quite different from the principle of tetrode tubes, in which the fourth electrode acts as an electric screen between anode and grid. In the tetrode transistor, the fourth connection is a second base contact,

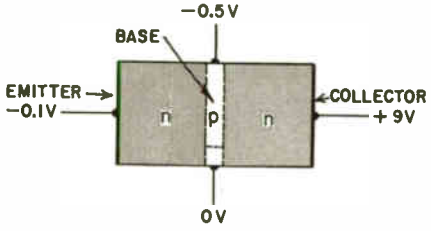


Fig. 716. *In the tetrode transistor there are two contacts to the base, enabling the working junction area to be restricted.*

opposite the first. As shown in Fig. 716, a potential difference is maintained between the two base connections such that most of the base is "biased off", preventing the flow of current through it. Only the remainder, below the dotted line, is positive with respect to the emitter, so is the only effective part. The reduced effective base area reduces the interelectrode capacitance, and the shorter distance from the working parts of the base to the lower terminal reduces r_b. Both these effects are beneficial at high frequencies.

Unipolar or field-effect transistor

The next device to be described resembles the last in general form and in having a sort of throttle action, but is really very different indeed. In fact, although it is known as the unipolar or field-effect transistor (sometimes abbreviated to "fieldistor") there might well be some doubt about whether it is a transistor at

all, since it has only one p-n junction and works on an entirely different principle, more akin to that of the vacuum triode.

As Fig. 717 shows, it consists of a chunk of n-type material (the body) which has an ohmic contact at each end, one called the source and the other the drain. When an emf is applied between these contacts, electrons flow through the body from negative to positive. At each side of the body, or all around it, there is p-type material, called the gate. When this is biased negatively, it repels the flowing electrons and forces them into a narrower channel, increasing the effective resistance of the body. This action reminds one of the grid in a tube. If now the drain voltage is increased, the result is not altogether what one might expect. It is true that it accelerates the electrons faster and so tends to increase the current. But, by making all parts of the body more positive with respect to the source, it increases the gate bias and tends to throttle the current. Beyond a certain drain voltage the acceleration of the electrons is almost exactly offset by the constriction of the gap, so the drain current curve flattens out, like the anode current/voltage curve of a pentode. This means that the device has a high output resistance. And because (unlike a transistor) the input is a *reversed* p-n junction, the input resistance is also high — of the order of megohms.

One condition for high-frequency operation is clearly present: an electric field to accelerate the current across. To increase the

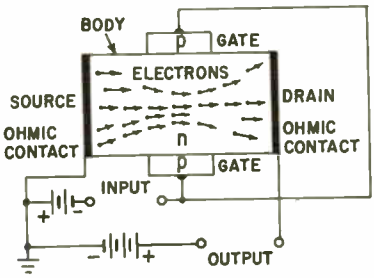
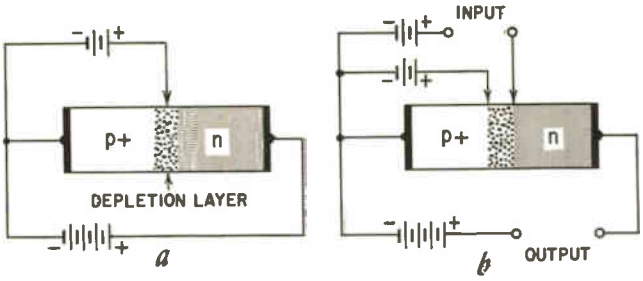


Fig. 717. *Section of field-effect "transistor," which has only one p-n junction, the end connections being non-rectifying. The "gate" acts rather like the grid of a tube, controlling the flow of current through the device.*

field from a given drain voltage, and to reduce the transit time still more, the dimensions must be small. The word "chunk" was perhaps misleading, for length and width are of the order of



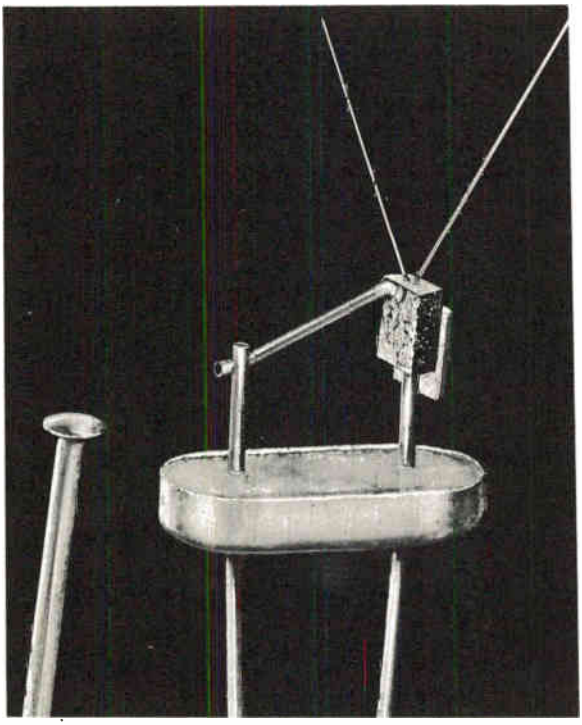


Fig. 718. *Triode and tetrode varieties of the spacistor, another device that works on a different principle from the ordinary transistor. The photo shows the construction of a spacistor.*

0.00025 inch! About 1,000 mc is the theoretical cut-off, but the maximum practical attainment seems to be more like 50 mc. However, 500 mc is claimed for a French variety called the technetron.

Spacistor

Lastly (for the present) there is the spacistor. This consists of a p-n junction with one or more extra contacts. Fig. 506-b showed

that when a reverse bias is applied to a junction it clears the main lot of electrons and holes — the majority carriers — out of a layer between the n and p regions, which is therefore called the depletion layer. The fixed charges left high and dry by these retreating carriers charge the p side of the layer negative and the n side positive, so the depletion layer is sometimes called the space-charge layer. The diagrams in Chapter 5 showed equal numbers of charges on each side, but suppose now that the p region is much more heavily doped with impurity than the n region. Since there must always be equal total charges on each side, the depletion layer must extend a comparatively long way into the n region to uncover enough positive charges to balance the dense crowd of negative charges standing at the boundary on the p side.

It is to this extended depletion layer that the extra contacts are made. Fig. 718-a shows a spacistor triode, which is one with a single side contact. The positive space charge makes it inject electrons into the depletion layer. Here they find themselves in a strong electric field due to the opposing charges, so they are accelerated towards the positive side — the n region. The reverse voltage connected across the junction is sufficient to bring it near the point of breakdown, and the fast-moving electrons knock others out of the atoms they hit, causing an avalanche effect, multiplying the number originally injected. So there is current amplification.

Better results are obtained with a tetrode structure, in which there are two side contacts: one to inject electrons and the other (called the modulator) to control them, as in Fig. 718-b. There is

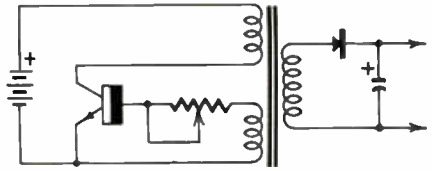


Fig. 719. *Basic circuit of the transistor square-wave oscillator for stepping up the voltage of dc.*

an obvious manufacturing difficulty in locating the contacts at the effective points in a layer thinner than a piece of paper, but when properly made the result is similar to a pentode without its bulk and power consumption, and very high frequencies have

been attained experimentally. Input and output resistances are of the order of 30 megohms, and the tetrode has low coupling between input and output.

Power limitations

When a tube or transistor is delivering its maximum ac signal power, it is alternately subjected to peak voltage with current cut off and peak current with voltage at its lowest. So one way in which its power-handling is limited is by the maximum voltage and current it can stand. Another way is by temperature rise.

In general, transistors are more limited by voltage than are tubes. If the collector voltage is increased beyond a certain point, even though collector current may be almost entirely cut off by absence of base bias, the collector-to-base junction will break down, just like any other diode. The maximum allowable collector voltage depends on the type of transistor, but is usually less than 50, and often much less.

With regard to current, however, transistors have an advantage over tubes. Fig. 705 shows that only a small fraction of a volt is needed to make a lot of current flow. This applies equally to types in which the current is measured in amperes rather than milliamperes. By contrast, tubes require a considerable voltage to make much current flow across the vacuum, because of the space charge, and this voltage generates heat at the anode, perhaps enough to make it red hot. A transistor junction of the same area carrying the same current receives comparatively little heat.

This does not, however, mean that the size of the junction for that amount of current can be greatly reduced or the current greatly increased. Although some tubes enjoy a full normal life with red-hot anodes, germanium transistors break down below the boiling point. The trouble is that even a moderate rise in temperature enormously increases the number of hole-and-electron pairs in both n- and p- materials, so current begins to flow freely in whichever direction the emf beckons. Instead of the zero-base-current curve in Fig. 705 almost coinciding with the zero-collector-current horizontal, it rises steeply, upsetting proper transistor action. Worse still, the more the current rises the hotter the junction becomes, and the transistor is likely to destroy itself quickly by this thermal runaway.

Precautions must be taken against this in the design of power transistor circuits, and by arranging for heat to be conducted rapidly away from the junction. Power transistors are made to

be mounted in close contact with a mass of cool metal, such as a suitable part of the equipment chassis, called a heat sink.

Uses of transistors

However little signal power a tube may be required to handle, it has to be fed with power to heat its cathode, and with voltage to attract electrons across the vacuum. Transistors require neither, so they are the obvious choice for low-power equipment such as hearing aids and personal radios. In these, they have entirely superseded tubes. Their advantages are less where power would in any case be drawn from a power line at low cost, but the ability to reduce size and weight to a fraction of what would be necessary with tubes and, in many cases, to render the equipment independent of power lines, is leading to a takeover by transistors in radio and television and many other directions. They are especially convenient for portable measuring instruments — meters, signal generators, oscilloscopes, etc.

Characteristic curves such as those in Fig. 705 show that a transistor working with a load in its collector circuit can, at signal peaks, be in either of two states, both of which cause very little power to be wasted in the transistor:

1. Collector voltage relatively large and current near zero, and
2. Collector current large and voltage near zero.

This means that in systems such as computers and controllers, where they are needed as electronic switches, working only in these two conditions, small transistors can handle relatively large amounts of power. This enables computers to be made smaller,

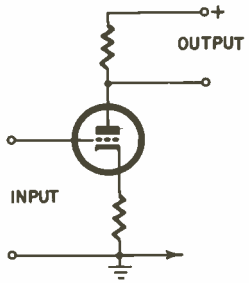


Fig. 720. *A cathode resistor is a common device for providing grid bias for a tube.*

give off less heat, consume less power, and work longer and more reliably without replacements than tube-operated equivalents. A typical computer, incorporating 5,000 transistors, needs only 20 watts dc: with tubes the consumption would run into kilowatts.

By back-coupling, a transistor can be made to switch itself alternately between these two states of maximum current and no current. This can be regarded as an extreme case of oscillation. A third winding on the coupling transformer can be used to step this square-wave ac up to any desired voltage, which can then be rectified and smoothed in the usual way to provide high-voltage dc from a low-voltage source, such as a car battery. This is useful for portable tube circuits, oscilloscopes, etc. Because of the square-cut transistor characteristics, the power efficiency of these units can be as high as 90% and, as Fig. 719 shows, the basic circuit is very simple.

These are only a few of the more important uses of transistors. At present, it is still not possible to combine very high frequency and high power in transistors, so tubes are likely to be used in transmitters for some time. Tubes also have advantages in very high-resistance circuits. But it does seem that semiconductor devices will, in time, largely displace vacuum and gas-filled devices.

Base biasing

One of the commonest methods of providing grid bias for a tube is by means of a resistor in the cathode circuit, as in Fig. 720. Anode current passing through this resistor makes the cathode positive with respect to the grid, which is the same thing as making the grid negative with respect to the cathode. An inci-

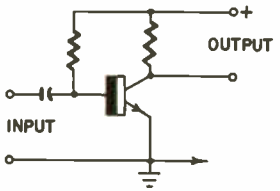


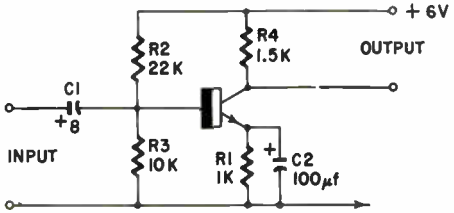
Fig. 721. *Base bias for a transistor has to be the same polarity as the collector, so is usually provided from the same supply through a resistor as here.*

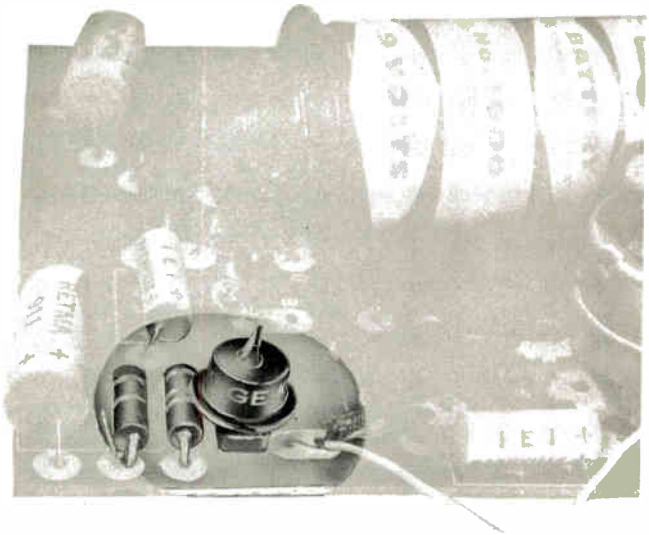
dental advantage is that if the tube tends to pass too much current, it automatically increases the bias voltage, which reduces the current.

An n-p-n transistor, which has the same collector-circuit polarity as a tube, needs positive base current and voltage. So if the same method were adopted as in Fig. 720, the bias would be of the wrong polarity. A possible method is to supply base bias current through a resistor from the collector supply, as in Fig. 721. Suppose, for example, the required bias is 0.1 milliampere and the supply is 6 volts. The difference in voltage between base and emitter is usually much less than one volt, so for this calculation it can

be neglected. The resistance needed to pass 0.1 milliampere when 6 volts is applied is 60,000 ohms.

However sound it may look on paper, this simple scheme is no good in practice. Since base current is not sufficiently related to collector current, there is no automatic compensation for the inevitable differences between individual transistors bearing the same type number. Because of manufacturing difficulties, these differences are apt to be rather large, and, in any case, any one transistor varies with temperature. So the usual method is to combine the previous two, using one resistor as in Fig. 721 to provide too much base current, so that another resistor as in Fig. 720 is needed to bias in the opposite polarity, cutting it





(Allied Radio Corp.)

Fig. 722. *To stabilize the collector current against changes of transistor and temperature, the simple arrangement for base bias current in Fig. 721 is usually supplemented by (1) a potential divider R2—R3 instead of R2 alone, and (2) an emitter resistor R1. The photo shows a resistance-coupled af stage.*

down to the required amount. Suppose now that the emitter current (which is nearly the same as the collector current) tends to be high; the voltage drop across the emitter resistor will tend to be large, making the emitter more positive and thereby acting to reduce the base current. But the voltage change is usually so small in comparison with the voltage available to the upper resistor that the base current supplied through it is not very much affected and the compensating action is small. To improve this, the base voltage is held more nearly constant by adding a third resistor as in Fig. 722, making a potential divider. The lower the values of R2 and R3, the better it works, but of course it would not do to make them very low or there would be too much current wasted in them and the input resistance would be too low.

Practical transistor circuits

The component values shown in Fig. 722 are typical for a stage of resistance-coupled af amplification. Apart from R2 and R3, the circuit is the same as for a tube.

Most other transistor amplifier and oscillator circuits look remarkably like their tube counterparts, but remember that due to the relatively very low input impedance of a transistor the numerical design of circuits is quite different. For the same reason, much greater efficiency can be obtained with transformer coupling, than with resistor coupling.

Besides the biasing arrangements, the other most notable difference in transistor radio circuits is the provision for counteracting the internal feedback, usually by neutralizing capacitors. Details are given in books on transistor circuits.

Switching transistors

Like tubes, any transistor can be used as a switch or relay, a small base current being used to turn on a much larger collector current. The transistor in Fig. 719 can be regarded as a switch automatically turning itself on and off at short intervals. Transistors used in computers are operated as switches, and to achieve the high speeds of modern computers they must be capable of working in the megacycle ranges. The charge storage effect is particularly undesirable when the object is to change over as quickly as possible from full to no current. And low resistance is obviously advantageous if much current is to be controlled. Transistors designed for high radio frequencies can be used for switching, but the tendency is to design special types for the purpose.

Among these is the symmetrical transistor, in which emitter and

collector are identical, so that either can be used as the other. This is sometimes convenient in switching circuits.

Another is the avalanche transistor, so named because it makes use of the avalanche effect, which is the very rapid current multiplication that follows when the collector junction is broken down by excess reverse voltage. Reverse current builds up in a small fraction of a microsecond and would, of course, burn out the transistor if the circuit were not provided with enough resistance to limit it to a safe value. Using this device, pulses of current can be produced with extremely steep rises.

When all that one wants is either full current or no current, there is no point in having the stable and linear characteristics that are essential for distortionless amplification. In fact, it is better if a single little signal pulse switches the transistor right over from one extreme state to the other as quickly as possible. Two ordinary transistors or tubes are normally needed to do this. But special "bi-stable" devices, resembling thyratrons rather than vacuum tubes in their behavior, have been produced. Let us now look at some of these.

Triple-junction transistors

Suppose we add one extra layer to a transistor, making three p-n junctions in all, and connect it through ohmic junctions at each end to a battery as in Fig. 723. We might expect the reverse junction in the middle would prevent any current flow other than

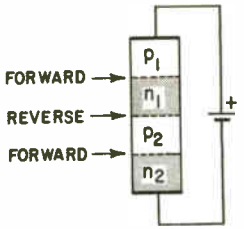


Fig. 723. *The four-layer triple-junction diode transistor; a bi-stable device.*

the usual very small leakage current, until the voltage was raised sufficiently to break the junction down. In other words, the whole thing should behave as an ordinary reverse-connected rectifier, the two end layers p_1 and n_2 doing nothing significant.

And that is just what happens. So far as impurity conduction is concerned, there are only electrons in n_1 and they are attracted towards p_1 by the positive pole of the battery. Similarly the holes in p_2 are attracted towards n_2. So the only current carriers that can cross the middle junction are holes in n_1 and electrons in p_2 caused by such disturbances as heat.

However, when the voltage is raised to a certain point the whole device flips suddenly over to a wide-open state, in which protective resistance would be needed to prevent a burn out. At the same time, the voltage across it drops from perhaps 100 to 0.5. Fig. 724 shows a typical charactertistic curve.

This second state will be easier to understand if we see that the device is equivalent to two ordinary transistors connected as in Fig. 725. Each of the two inner layers in Fig. 723 acts as a base for one transistor and a collector for the other. We must remember that the current from B_2 to C_1, plus the current from C_2 to B_1 must be equal to the current from E_1 to E_2 via the battery. If the two transistors were similar, we would expect each of the two parallel currents to be about half the battery current. But we know that under normal working conditions the base current is usually less than one-tenth of the collector current. That is obviously impossible here. If we examine transistor characteristic

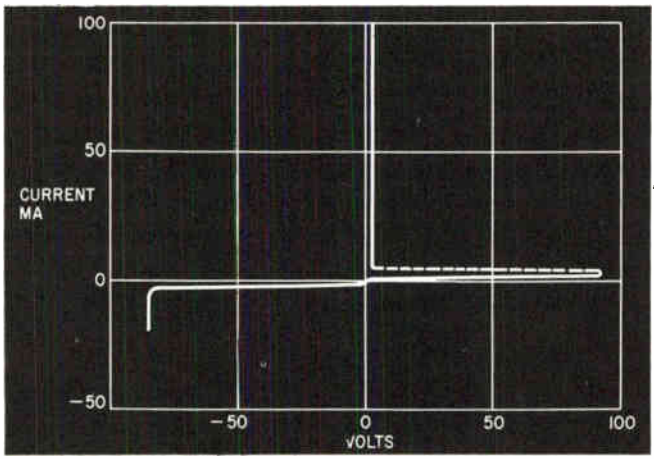


Fig. 724. *Characteristic curve of the transistor in Fig. 723. The dotted line represents the sudden switch-over from "open" to "closed" circuit.*

curves, we shall find that the base current can be only as much as half the collector current if the collector voltage is very low indeed — only a few millivolts. So nearly all the battery voltage is used in making the base of the n-p-n transistor positive to its emitter, and the base of the p-n-p transistor negative to its emitter, both having the effect of opening the transistors wide to current, like vacuum tubes with the grids connected positive.

If the current is reduced to a certain minimum, say a milliampere or two, the device springs back to its nonconducting state.

Compared with gas-discharge tubes, which behave in a somewhat similar way, these semiconductor switches differ in their very low resistance and voltage drop in the "on" state. So for computors and other applications they are more suitable. Fig. 726 shows a simple use for one as a sawtooth waveform generator. Capacitor C charges through R1 and R2 until the voltage across the transistor reaches the breakdown point, when it short-circuits and discharges C rapidly through R2. When the discharge current drops below the amount needed to hold the switch "on", it opens and the cycle begins again.

These four-layer units can be produced in various ways: by double diffusion at one end for three of the layers, with the fourth added at the same end by alloying; by diffusion at opposite ends, plus alloy at one or by single diffusion plus alloy at both ends.

Multi-layer triodes

The usefulness of the foregoing bi-stable devices can be increased by making a connection to one of the two inner layers. If it is made to n_1 in Fig. 723, the device is called a *thyristor;* if to p_2, a controlled rectifier or a *trinistor.* In all of these, the main end-to-end current can be switched on by pulses applied to the third or control electrode. The thyristor is notable for its very rapid switching — "on" in something like one-tenth of a microsecond. The amount of power controlled is only a few watts, whereas controlled rectifiers are made to handle up to 20 amperes

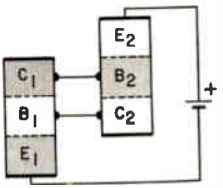


Fig. 725. *The transistor in Fig. 723 can be regarded as two ordinary transistors (one n-p-n and one p-n-p) connected like this.*

at hundreds of volts. Switching times are slightly more leisurely — as much as one microsecond for "on" and 10 microseconds for "off". Another difference is that, like the four-layer diodes, they block current in the reverse direction, whereas the thyristor conducts.

Devices in this class fall into two main groups: small low-power types for circuit switching, computers, etc., and larger types for power control. Compared to gas thyratrons, they have many advantages, such as low voltage drop, high efficiency, fast switching off, no filament and (with silicon) higher allowable temperature.

Hook transistor

A four-layer triode connected the same way as a trinistor, but smaller and with different characteristics, is known as the hook transistor. Like the point-contact transistor, described at the end of this chapter, the hook transistor gives both voltage and current

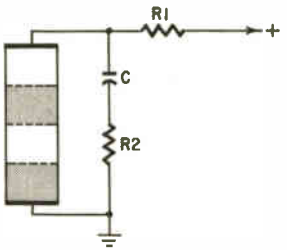
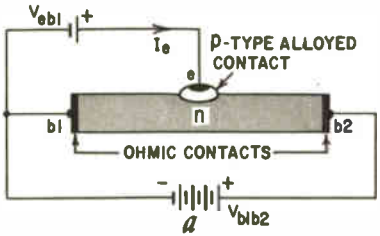


Fig. 726. *Simple sawtooth oscillator circuit using a four-layer diode. R2 is much less than R1 — just enough resistance to prevent too much current going through the diode.*

amplification, and an output in phase with the input. (The ordinary transistor, common-emitter connected, like the vacuum tube gives its voltage output in inverse phase to the input.)

Negative-resistance devices

Another feature of the point-contact transistor is that it can have a negative-resistance input; that is to say, over a certain range



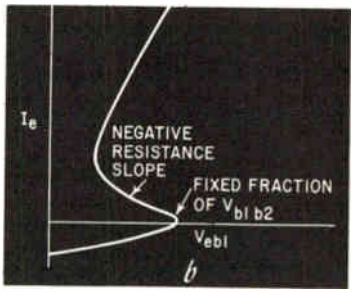


Fig. 727. *Unijunction "transistor," which looks rather like the fieldistor but works quite differently. Its characteristic curve doubles back, providing a range of negative resistance.*

of voltage the current increases with decreasing voltage. It is therefore capable of setting up oscillation in the input circuit without the aid of feedback. Without an oscillatory circuit, a negative-resistance device flips over from one current-voltage state to another, as in the switching devices we have already considered. The point-contact transistor is not an ideal component, for reasons to be mentioned. However, other negative-resistance semiconductor devices have appeared. We have come across some of them already under the heading of Switching Transistors.

Unijunction transistor

Another of them, the unijunction transistor, in some respects resembles the unipolar transistor, and like it has a doubtful claim to the title "transistor". Its alternative name — double-base diode — is perhaps more accurate. Comparing Fig. 727-a with Fig. 717 shows that the main difference is that the side contact — in this case called the emitter — is worked in a positive range of voltage. The body, of n-type germanium or silicon, acts as a potential divider for the voltage V_{b1b2} applied to it, so the part of it in contact with the emitter receives a certain fraction of that voltage. When the emitter voltage (V_{eb1}) is equal to it, zero emitter current (I_e) flows; and from this point, as Fig. 727-b shows, growth of I_e is accompanied by fall of V_{eb1}, the reason being that when current is flowing in the emitter circuit b_1 and e together form a forward-biased junction diode, and b_2 and e a reverse-biased diode. When the change-over has been completed, the remainder of the curve is typical of a forward-biased diode.

Tunnel diode

A more recent negative-resistance device is quite avowedly a diode, being known in fact as the tunnel diode, but is included here on account of its undiodelike characteristics, which are likely to be found very useful. At the time of writing, tunnel diodes have been made to oscillate at some thousands of megacycles per second and to switch over in less than 1 mμsec. (=1/1,000,000,000 sec.). In contrast with some of the semiconductor devices described, the impedance is extremely low, and a major problem in utilization is to keep the power-supply and circuit impedances

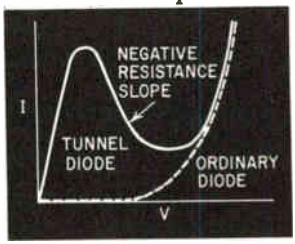


Fig. 728. *This negative-resistance curve of a tunnel diode should be compared with the forward characteristic of an ordinary diode (dotted) and with the negative-resistance curve shown in Fig. 727.*

low enough. Compared with the ordinary kind, tunnel diodes are very heavily doped, and the forward current rises steeply at a much lower voltage, as shown in Fig. 728. This rise is followed by a fall, representing negative resistance, beyond which the curve is similar to that of an ordinary diode.

The name "tunnel" has nothing to do with its shape; it refers to a peculiarity of atomic behavior whereby in certain circum-

stances a charged particle can, as it were, tunnel through a potential barrier which according to simple theory would stop it. Hence the current flow indicated by the left-hand part of the curve.

A valuable feature of the tunnel diode is that it will work over an exceptionally wide range of temperature: from about —269°C to +200°C for germanium and +400°C for silicon.

Although both Fig. 727-b and Fig. 728 show negative-resistance slopes, there is a fundamental difference between the two. One has three different levels of current at a single voltage and is said to be current-controlled; the other has the same current at three different voltages and is said to be voltage-controlled.

Point-contact transistors

Although point-contact types are now only of historical interest, this brief review of transistors may fittingly conclude with a glance back at them. They are similar to point-contact diodes (Fig. 608) except that *two* springy wires touch the small slab of germanium at points only a few thousandths of an inch apart. Fig. 729 shows that in this original type there was some justification for the term "base," especially since it was usually grounded in circuits. It consisted of n-type germanium, and the transistor was formed similarly to point-contact diodes by passing heavy currents through the contacts for brief periods of time. This

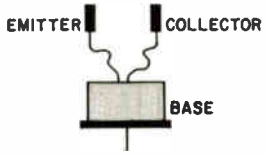


Fig. 729. *Point-contact transistor.*

changed very small regions around the points into p-type material, so the result was a p-n-p transistor.

But there is more to it than this; so much more that in spite of extensive research there is still doubt about exactly how the thing works. In particular, any theory has to account for a remarkable difference as compared with junction transistors—the value of α is greater than 1, being usually about 2 or a little more. This means that increasing the emitter current by a small amount increases the collector current by perhaps twice that amount. One might think this a good thing, but it made these transistors tricky to use; in certain circumstances they were unstable, so that once the current started to increase it went on increasing until the transistor was burned out — an event that could take place in a moment.

chapter 8

photocells

THIS chapter is about semiconductor devices designed to respond to light. They are often called photocells, but it must be understood that not all photocells are semiconductor devices. In fact, the kind with the most familiar results (because used for broadcasting television programs) comes into what is called the photoemissive group, which is a branch of electron tubes rather than of semiconductors and is outside the scope of this book. But, just as we have found it helpful to compare transistors with tubes, we will compare semiconductor photocells with the emissive type.

Light

Before looking at types of photocells, it would be well to consider light itself. It consists of waves of the same nature as those used in radio communication, but much higher in frequency. The highest used so far for radio are in the region of 100,000 megacycles, corresponding to wavelengths of a few millimeters. The frequency of the waves has to be about 4,000 times higher still before the human eye can detect them. Such waves, having a length of less than 1/1,000 millimeter, produce the kind of response we call red. Doubling this frequency (halving its wavelength) brings us to the highest we can see, which is violet. The frequencies in between are seen as colors that come in the order of those in the rainbow. Although in the strict sense "light" is confined to this quite narrow visible band, the word is often used — and will be in what follows — to include frequencies lower than visible (infra-red) and higher than visible (ultra-violet).

Photoemission

We saw in Chapter 3 that the energy of radiated waves increases in direct proportion to their frequency. When ordinary radio waves strike a piece of metal, such as an antenna, they have enough energy to make electrons in the metal vibrate, but not nearly enough to make them jump right out into the open air. Even light waves, with their thousands of times greater energy, aren't energetic enough to knock electrons out of ordinary metals. But specially prepared metal surfaces yield electrons much more easily. For instance, the rare element cesium on antimony emits electrons when illuminated by light waves of the more energetic kinds (violet and blue), and cesium on oxidized silver is sensitive to the less energetic red light and even to some of the invisible (infrared) frequencies.

Fig. 801 shows how this response is utilized in photoemissive cells. The sensitive surface is mounted in an evacuated glass tube

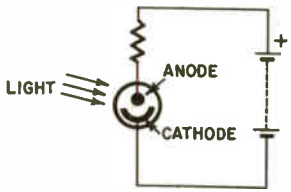


Fig. 801. *The emissive type of photocell is like a tube in which light takes the place of heat at the cathode.*

in such a way that light can reach it. Here it forms a cathode; but instead of being heated, as in the usual vacuum tube, it depends entirely on light to release its electrons. These electrons are collected by a positively-charged anode in series with a suitable load resistance. When the tube is in the dark, there is negligible emission and therefore hardly any voltage across the load resistance. The brighter the light, the more the emission and the greater the voltage. At most, the emission is only a few microamperes, so the load resistance is usually in megohms. In some tube types, a small amount of inert gas is admitted; this considerably increases the current by ionization.

Characteristics of photocells

One of the important characteristics of a photocell is its sensitivity – which means its sensitiveness measured in a definite way, for comparison with other cells. There are several different ways of defining sensitivity, and the subject is too complicated to go into here, but it can be said that the sensitivity of vacuum photocells is comparatively small. Gas-filled cells are typically 5 to 10 times better in this respect.

Another characteristic is speed of response. Vacuum cells respond almost immediately light shines on the cathode, the only delay being due to the very small time — a fraction of a microsecond — taken by the electrons to cross over to the anode. The ionization process in the gas-filled kind is much slower, so that, if the light varies faster than about 1,000 cycles, the response cannot keep up with it. So for fast changes, as in television, vacuum cells must be used in spite of their low sensitivity.

Then, there is what is called the spectral response. This is not, as might be supposed, the sensitivity of photocells to ghosts, but

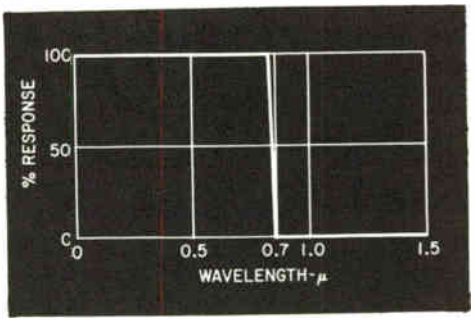


Fig. 802. *On the basis of the Planck quantum law, one would expect the response of a photoemissive tube to vary with wavelength like this, but complications make it like Fig. 803 (1μ = 0.001 mm).*

the way in which it varies with frequency or wavelength. For this purpose it is usual to work in wavelengths, and the most convenient unit of length is the micron, denoted by μ, which is 1/1,000 millimeter. On this scale, violet light begins at about 0.38μ, red light ends at 0.76μ, and from there the infra-red band extends until it meets the shortest waves generated by radio engineers, which are about 1,000μ or 1 millimeter. For practical purposes, it goes up to only about 8μ.

Spectral response

A certain amount of energy is needed to dislodge an electron from a solid surface. The amount depends on the kind of surface. Most surfaces need more than 4 electron-volts (eV). The energy units of visible light (photons) range from 1.65 eV at the red end to 3.3 at the violet end, so are insufficient. The special feature of the surfaces used on the cathodes of photoemissive cells is their low eV figure. Cesium, for example, is only about 1.8 eV,

corresponding to 0.7μ wavelength, so light of almost all colors is enough to extract electrons from it.

According to this theory, we would expect a spectral diagram for this type of cell to be something like Fig. 802, with no response at wavelengths longer than 0.7μ and full response at all shorter. For several reasons, it doesn't work out like that. For instance, practical surface conditions are more complicated than in simple theory, and continuation of response in the ultraviolet direction is cut off by the glass of the tube, which is opaque to such rays. So most spectral diagrams look more like flat re-

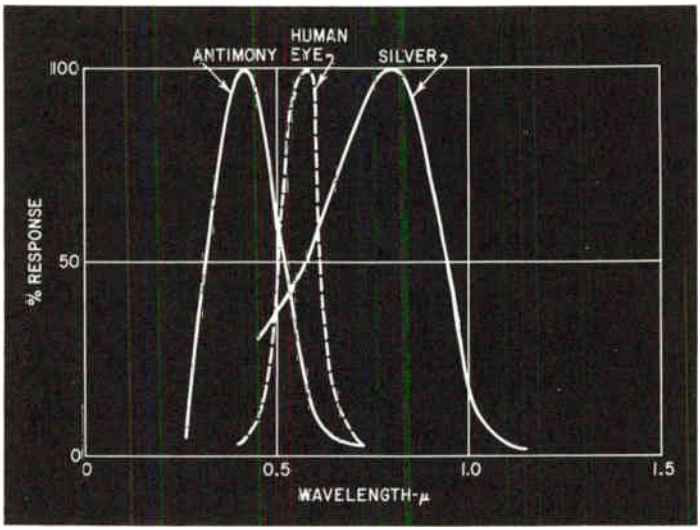


Fig. 803. *Response curves of photoemissive cells having cathodes with cesium on two different kinds of base, and the human eye's response for comparison.*

sonance curves. Fig. 803 shows examples for cesium on silver and cesium on antimony, with the human eye's response for comparison. The first kind of surface would be an obvious choice for light at the yellow and red end and for near infra-red, and the second for light at the blue and violet end and for near ultraviolet.

Photoconductive cells (or photoresistors)

Coming now to semiconductors, we may remember that in Chapter 4 we saw that the pure materials don't conduct unless they are exposed to some kind of energy. This is because there is a gap between the energy states, which are completely occupied

by the valence electrons, and those higher states which they have to reach if they are to be available as current carriers. The energy gap in germanium is about 0.7 eV, so light of all colors is sufficient to hoist electrons across it. But because germanium is opaque, only electrons on the surface are affected, and the conducting path is so thin that even if many electrons were released the resistance would not be greatly reduced. On the other hand, the whole material is affected by heat, and, therefore, changes of temperature affect the resistance far more. So although, in principle, a strip of germanium or silicon could be used to detect or measure light by its change of conductance (or resistance), it is not very suitable in practice. For one thing, it would have to be cooled nearly to absolute zero to prevent the light indications from being drowned by "noise." Moreover, even the purest material is considerably affected by impurities.

The photoconductive response of selenium was discovered as long ago as 1873, and the first selenium cell patented (by C. E. Fritts) in 1884. Selenium is still used, although many other materials have become available. In one recent development, thin layers of selenium, cadmium and gold are deposited on a flexible metal base to make a photocell which can be bent into cylindrical and other curved forms for special applications. One advantage of selenium is that its spectral characteristic is close to that of the eye. But its photoconductive sensitivity and speed of response are not very good. Owing to its high resistance, it is usually spread

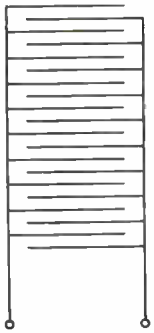


Fig. 804. *Electrode formation commonly used in selenium photoconductive cells, to bring the resistance down to a workable level.*

across comb-like electrodes as in Fig. 804, and even then the resistance is of the megohm order and needs anything up to several hundred volts. We will see later that selenium is more often used in a different (photovoltaic) manner.

One of the best semiconductors for visible-light photoconductive cells is cadmium sulfide. It responds to practically the whole

visible range and near infra-red, passes very little current in the dark (less than 1 μa), and can be made up into rugged cells which pass several milliamperes when light shines on them — sufficient to operate a simple relay without amplification. So it is much used for such purposes as detecting flame extinction in oil-fired furnaces and for street light controls. But its speed of response is too slow for high-frequency light variations. Another material is thallium sulfide, which is made up into "thallofide" cells.

Other commercial types of cell use compounds of lead: lead sulfide, lead selenide and lead telluride. All of these cover a wide part of the infra-red band; the sulfide includes the whole visible range as well, is extremely sensitive, and has a fairly low cell resistance (kilohms rather than megohms), so is a most useful photoconductive material. Its applications for military purposes ("seeing in the dark", and aircraft detection) were developed by the Germans in World War II.

It has, however, a time lag of about 75 microseconds, so where high-speed response to infra-red waves is desired, the choice is

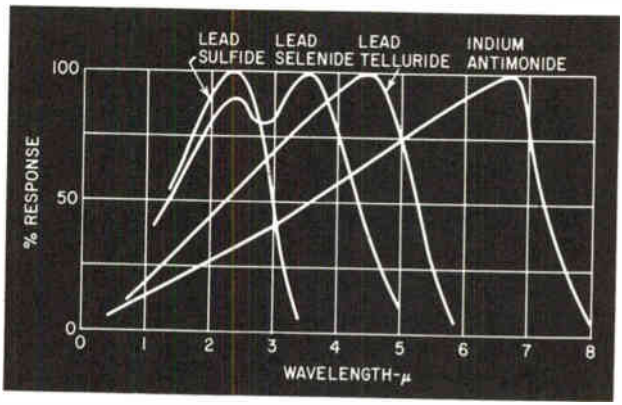


Fig. 805. *Spectral characteristic curves of four semiconductors used for infra-red detection and measurement.*

lead selenide, in spite of its much lower sensitivity. Its spectral response is confined mainly to the infra-red, as shown in Fig. 805.

Lead telluride goes still farther into the infra-red — up to 6μ — and its useful sensitivity and time lag are about halfway between the other two lead compounds, but it has the rather considerable practical disadvantage of having to be cooled to about the temperature of liquid oxygen (—183°C) to obtain a satisfactory signal-to-noise ratio.

Last in this group is another semiconductor compound, indium antimonide, which covers an even wider waveband, right out to 8μ. It is moreover very fast and its resistance is low — typically 75 ohms. The main use of these last three materials is in infra-red spectrometers — scientific instruments for analyzing infra-red radiation according to wavelength. Being so sensitive to infra-red radiation (more familiar to us as heat), the current they pass depends very much on their temperature. To exclude this effect, the radiation signals they are used to detect are "chopped" by some sort of shutter, converting them into alternating signals of a frequency to which the amplifier following the cell is sharply tuned (Fig. 806).

The Vidicon

Most television camera tubes operate on the photoemissive principle. This is limited in sensitivity because the output energy is derived directly from the light. So rather complicated systems have to be used to amplify it. Photoconductivity is not limited in this way, as the energy is derived from the local source, which is controlled by the light, much the same way as a vacuum tube acts

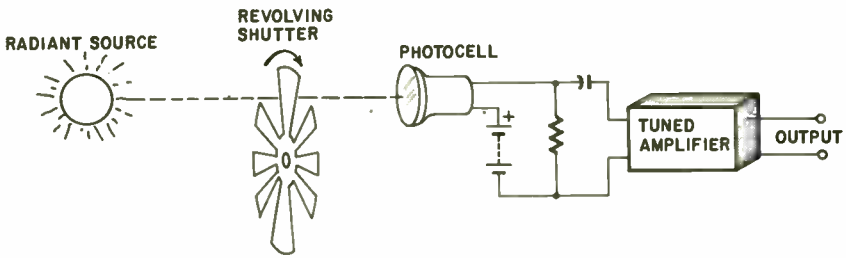


Fig. 806. *To distinguish the desired response from others when using photocells, the light being detected is "chopped" by a moving shutter and the resulting ac signal amplified in a selective amplifier.*

as a grid-controlled relay. It is used in a type of television camera tube known as the Vidicon. Because of its simplicity and small size, the Vidicon is particularly suitable for portable cameras and for industrial use. Tubes have been made as little as half an inch in diameter, though one inch is more usual. The sensitivity is good, and only a tendency to a time lag in poor light prevents it being used even more widely, say in studio cameras.

Fig. 807 omits the details of scanning and focusing coils, gun electrodes, etc., to show more clearly how a semiconductor enables a video signal to be obtained. The semiconductor, one that

has a resistance of many megohms in the dark, is in the form of a very thin disc, the back of which is scanned by the electron beam in the usual way. An image of the scene to be televised is focused on the front of the disc, through a semitransparent conductive layer by which electrical contact is made for maintaining it at about 20 volts positive. Those parts of the disc on which the light shines are rendered more conductive, so that when the beam (which is near ground potential) strikes them, a current flows around the circuit and through a resistance across which signal voltages are developed.

Among semiconductors which can be used are amorphous selenium, antimony sulfide, cadmium sulfide and cadmium selenide.

Photodiodes

Chapter 5 told us how a p-n junction works, and that when it is biased by a voltage in the "reverse" direction, it passes no

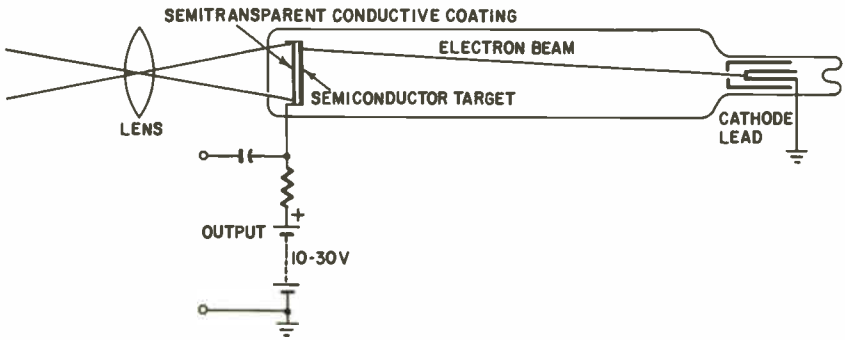


Fig. 807. *The Vidicon television camera tube uses a semiconductor photoconductive screen which is scanned in the usual way to give the video signal.*

current except an amount which is small at ordinary temperatures but increases rapidly if the temperature is raised. Any reverse current is a nuisance when the junction is used in the usual ways as a rectifier. But in a photodiode — which under working conditions is kept permanently biased in reverse to prevent any of the impurity current from flowing — the unwanted reverse current becomes the working current. For, as we saw, it is stimulated not only by heat but also by light. That is why ordinary diodes and transistors must be shielded from light by opaque containers. Photodiodes, on the other hand, are provided with a little window through which light can fall on the junction in such a way as to allow it to reach as many as possible of the ger-

manium atoms, knocking their valence electrons into the "conduction" energy band and leaving behind "holes" in the valence band. These electrons and holes are driven in opposite directions by the reverse bias, creating a current through the diode. The brighter the light, the more the current.

Their spectral range covers most of the visible band, and quite a lot of the infra-red, up to about 2μ. So they are particularly suitable for the light from filament lamps. The speed of response is moderate, and sensitivity is good. But perhaps their biggest virtue is that they are are smallest. They can be made about the size of a large pinhead, yet they will pass several milli-

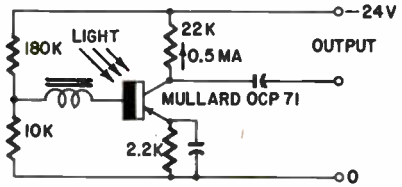


Fig. 808. *Typical circuit diagram for a phototransistor.*

amperes. So they are ideal for "reading" punched cards or tapes in computers and data processing equipment, and for industrial control devices of many kinds.

Phototransistors

For some purposes, it may be necessary to amplify the output from photodiodes. Where the utmost advantage must be taken of their small size, the amplifiers can be located separately. But if there is room for a slightly larger device — still far smaller than other types of photocell — the best idea is to use a phototransistor, in which the photodiode is combined with its own amplifier.

A phototransistor looks like any other transistor, except that it has a small window to admit the light; and it is used in the same kind of circuit, such for example as Fig. 808. Use of a temperature compensating bias system is especially necessary here, because temperature variations compete directly with light variations in changing collector current. This bias system also tends to iron out *slow* changes in light, but the light signals to be detected are assumed to take place rapidly enough to vary the collector current before the emitter-bias capacitor has time to charge or discharge. Temperature cannot change fast enough to take the emitter bias by surprise, as it were.

Fig. 809 is a set of characteristic curves for the same phototransistor. They are the same shape as for an ordinary transistor, the only difference being that they are drawn for illumination (instead of base current) as parameter.

The four-layer bi-stable transistors can also be arranged as phototransistors. The effect of light is to switch the device over from no current to full current, which may be many milliamperes — or even amperes — capable of directly operating a motor, etc.

Photovoltaic cells

All the cells described so far have to be provided with voltage from a battery or other source. Some semiconductor junctions have the useful property of generating an emf themselves under

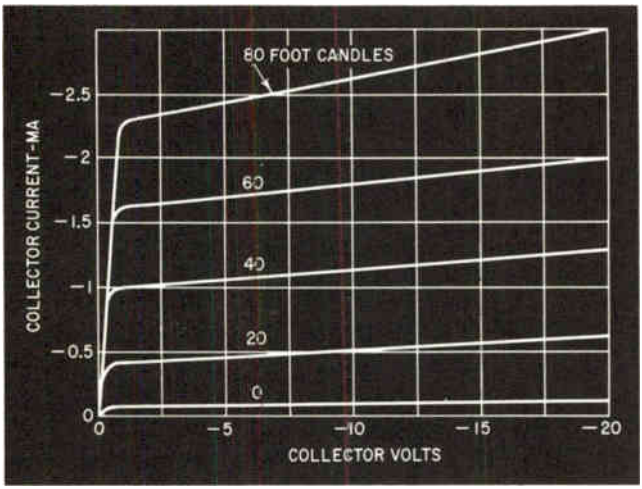
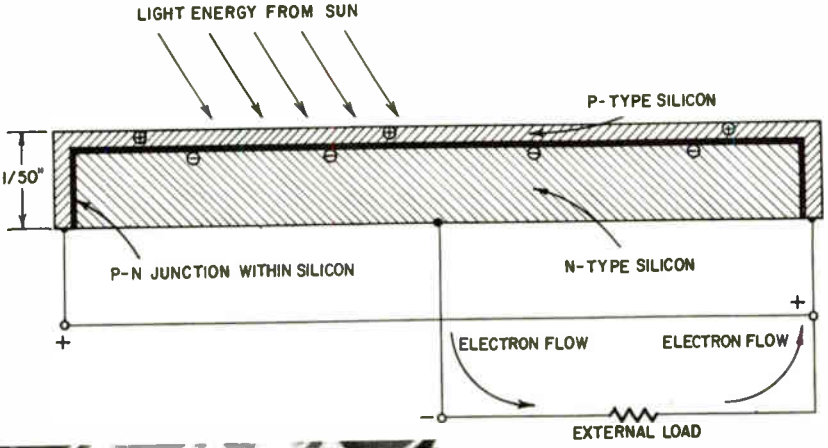


Fig. 809. *Phototransistor characteristic curves, which should be compared with Fig. 705.*

the stimulus of light, so all that is needed is a meter for showing the current it drives. This combination is familiar to most of us in the form of photographic exposure meters. Photocells in this class are called *photovoltaic.*

For such cells, the "historical" photosensitive semiconductor, selenium, is still largely used; but silicon cells are now commercially available in many forms.

Where does their emf come from? Chapter 5 showed how a difference of potential builds up at a p-n junction as a result of the tendency of electrons in the n-zone to stray into the electronless p-zone, and for holes in the p-zone to stray into the n.

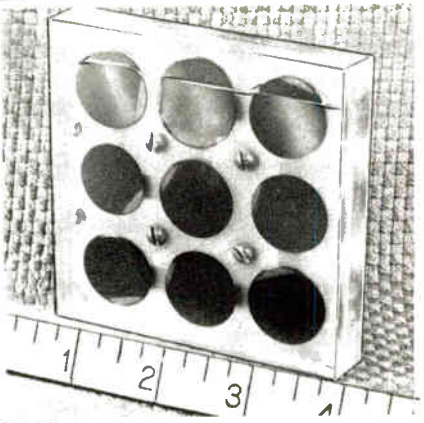


Cross-section of a silicon solar cell.

International Rectifier Corp.

(Above) Single silicon solar cell and plastic holder.

(Below) A solar power converter composed of a modular cluster of nine silicon cells connected in series.



Hoffman Electronics Corp.

Bell Telephone Labs.

(Left) A 400-cell solar-power converter featuring an automatic sun-following device.

Both these movements charge the n zone positively with respect to the p. If we try to detect this difference (about 0.25 volt with germanium) using even a sensitive vtvm, we shall be disappointed, for the reason already explained – that the junctions between the n and p semiconductor zones and their metal terminals also develop potential differences which exactly cancel the p-n potential difference.

This assumes all three junctions are at the same temperature and are equally (if, at all) affected by light. But if the p-n junction is given a different temperature from the others, or is brightly lit while the others are kept dark, this balance is upset by the creation of hole-and-electron pairs, and the device has a small net voltage between its terminals.

To be much good photocells working on this principle obviously have to be arranged so that light can get at the whole of the

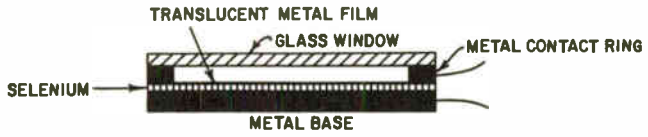


Fig. 810. *Section of a selenium photovoltaic cell.*

junction area. Fig. 810 shows the type of construction. Selenium is deposited on a metal base, with an extremely thin film of a metal (such as gold) on top, through which light is able to shine. The cell is provided with a glass window for protection.

Solar generators

The power efficiency – the proportion of light power converted into usable electrical power – of photovoltaic cells is low; usually only 1 or 2%. But silicon cells have been developed to the point where an efficiency of 10% or more is yielded. That is enough to bring the dream of direct power from the sun into the realm of practical engineering. Already, silicon photovoltaic cells have been used on telephone poles to provide power for repeater amplifiers, and in artificial satellites to drive their radio transmitters and other equipment. Their value rests on the fact that the power never runs down and their life is indefinitely long. But to take care of nights and dull days on terrestrial installations, and eclipses by the earth or other bodies on space vehicles, some surplus of power is needed to charge reserve batteries.

It is estimated that in two days the earth receives more energy from sunlight than could be provided by all known reserves of

fuel, so you may wonder why anyone bothers about nuclear power stations or indeed any other kind? The answer is that some hun-

Fig. 811. *Section of a solar silicon photovoltaic cell for deriving electric power from the sun.*

dreds of cells are needed to generate even 10 watts, and their cost is at present enormous compared with other generators — far too much for any except rather special or experimental uses.

Fig. 811 shows the construction of a typical cell. It consists mainly of a disc of n-type silicon, on the exposed surface of which an extremely thin p-type layer is formed by diffusion of boron. The junction is only 1/10,000 inch below the surface, so can be reached by light. In bright sunlight it generates about 0.5 volt, the p-layer being positive. Current is limited, so many cells have to be connected in series-parallel for battery charging, etc.

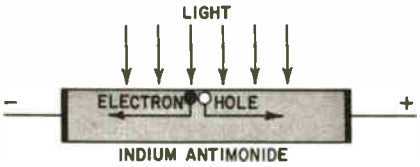


Fig. 812. *Showing the principle of the photoelectromagnetic type of cell, in which a magnetic field must be provided to act at right angles to the light and to the electric current resulting. In this case the direction of the field is into the paper.*

If cost can be drastically reduced, there is a great future for this semiconductor application.

Photoelectromagnetic cells

Mention has already been made of indium antimonide as a semiconductor compound having a very wide spectral range in the photoconductive role, where, of course, a power source is needed to pass current through it. Alternatively, the material can be mounted between the poles of a permanent magnet, which causes a current to be generated when light strikes the cell at right angles to the direction of the magnetic field (Fig. 812).

One of the fundamental electrical effects, used in every power-station generator, is that a conductor moved across a magnetic field generates an emf at right angles to both directions. One difference here is that the conductor (or rather semiconductor)

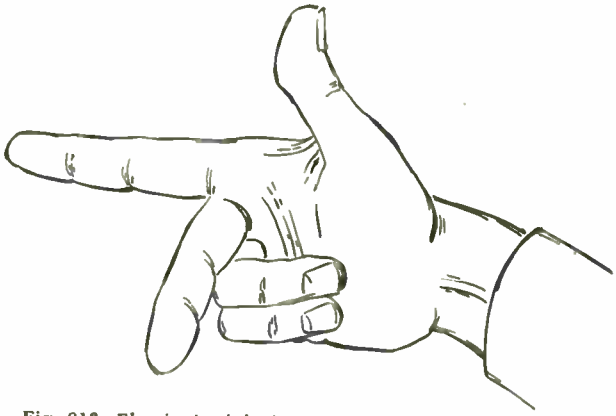


Fig. 813. *Fleming's right-hand rule, from which the direction of the electric current in Fig. 812 is indicated by the second finger if the first finger is pointed in the direction of the magnetic field and the thumb in the original direction of motion of the charges.*

is fixed with respect to the field. But the only reason for moving the conductor is to make the electrons move so that they can be pushed aside by the magnetic field. In this case, light releases electrons at the surface and they drift inward into the material. It is this movement that brings them under the influence of the field, just as if the material containing them were moving in the direction of the light. The direction in which they are pushed is indicated by the second finger in Fleming's right-hand rule when the first finger is pointed in the direction of the magnetic field and the thumb in the direction in which the charges are originally moving (Fig. 813).

Another difference is that in a semiconductor the release of electrons creates holes, and they contribute to the current by turning aside in the opposite direction. Because both holes and electrons are comparatively scarce, the current is much smaller than in a metal.

chapter 9

other semiconductor devices

NEW uses for semiconductors are announced so frequently that it would be impractical to describe in detail all those already known, still less predict what other developments may have appeared by the time these pages are read. Those included in this

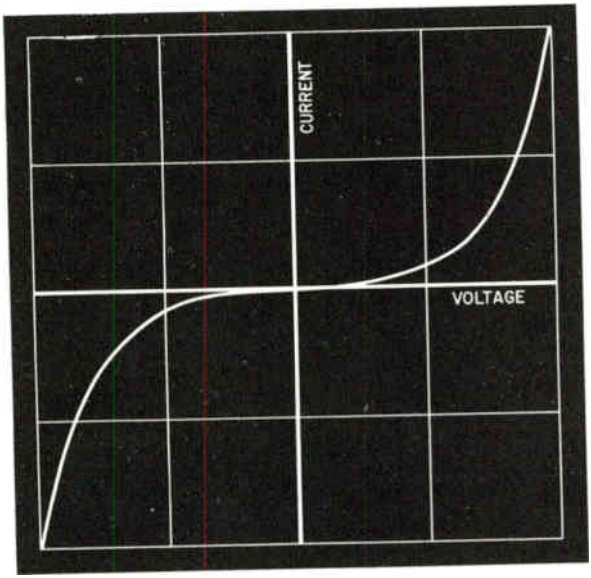
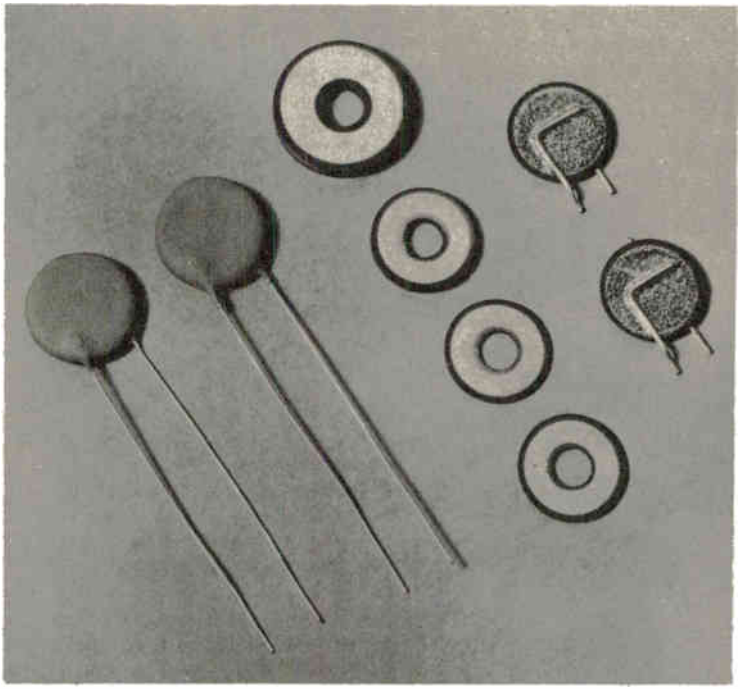


Fig. 901-a. *Typical shape of varistor current/voltage characteristic curve.*

chapter should be regarded as no more than typical of a technology that is expanding perhaps more rapidly than any other.

They can be divided into two groups: first, those in which semiconductors are used simply as materials, in single chunks, layers or what not; and, second, those that make use of junctions between two or more kinds of semiconductors, or semiconductors and metals. Many in this second group have already been men-



(Victory Engineering Corp.)

Fig. 901-b. *Photo shows varistors in different stages of production.*

tioned in the previous chapters; those still to come happen to be essentially diodes, but used for purposes other than rectification or light detection.

Varistors

Certain semiconductor materials conspicuously defy Ohm's law; i.e., the current passing through them is far from directly proportional to the applied voltage, even when the temperature is kept constant. Unlike rectifiers, however, they offer the same resistance to a voltage of either polarity. A notable example is silicon carbide, whose resistance drops very rapidly as the voltage

is increased. Fig. 901-a shows a typical current/voltage curve, in contrast to the straight line of Ohm's law.

This material is extensively used in power supply electrical engineering to provide a safe path when excess voltages build up, as in thunderstorms. This protects the equipment it is shunted

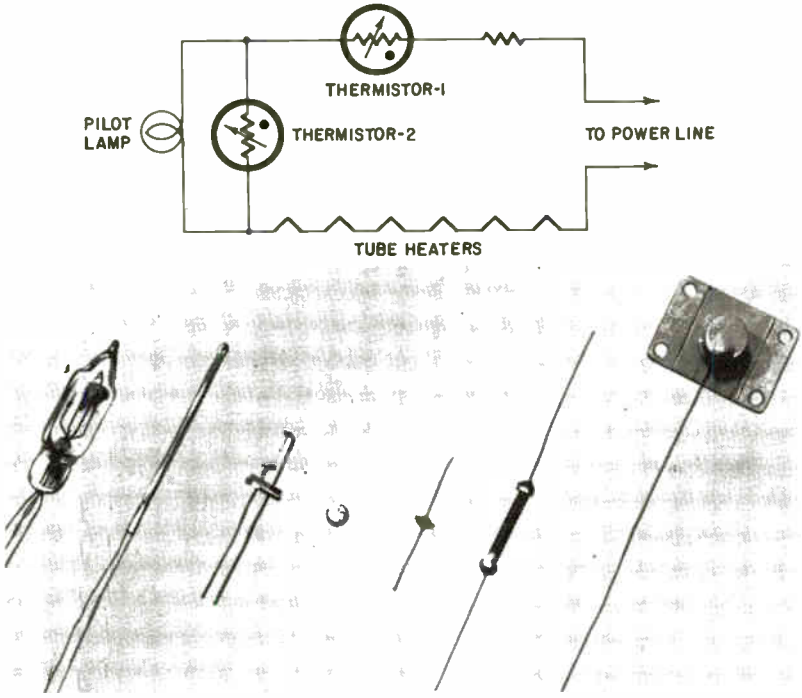


Fig.-902. *Two uses of thermistors in a tube heater circuit. The photo shows various types.*

across and, when the surge has died away, the varistor (as it is called) regains its normal resistance, which is too high to affect the working of that equipment. For high voltage and small current, varistors take the form of thin rods; for the opposite requirements, they are fabricated into thin wide discs (Fig. 901-b).

Thermistors

It is characteristic of all semiconductors that their resistance falls steeply as the temperature is raised. Resistors designed to take advantage of this are called thermistors. Among the semiconductor materials found particularly suitable are oxides of titanium, manganese, cobalt and nickel, either singly or mixed. Like varistors, thermistors come in a great variety of sizes and

shapes, according to the current and voltage ranges for which they are intended. Those most used in radio and electronics are small beads, discs or rods, fused on to their connecting leads.

Thermistors are sometimes used in electronic tube circuits in which the cathode heaters are wired in series. When cold, these heaters have a much lower resistance than when hot, and any pilot lamps in the circuit would be burned out by the surge of current at switching on. The thermistor, having the opposite kind of characteristic, prevents this surge by offering at first a high resistance; then, as it heats, it loses most of this and allows the full working voltage to reach the heaters. Unfortunately, it makes one wait longer for the equipment to come into action.

If a pilot lamp does burn out in such series circuits (and its life is usually less than that of the heaters, run at a much lower temperature), it puts the whole equipment out of action. To prevent this, another thermistor (thermistor–2) is sometimes connected in parallel with it, in Fig. 902. Under normal working conditions, the resistance of thermistor–2 is too high to divert much current from the lamp, but, if the lamp burns out, all the current has to go through the thermistor, heating it and reducing its resistance so much that the tubes receive practically full working current.

Thermistors are also used: in series with the anodes of a directly-heated rectifier tube, to reduce the output for a short time after switching on until the other tubes have heated enough to draw current; to compensate for temperature effects in electronic circuits; to control feedback to keep constant the output of oscillators; and to measure microwave power by the change in temperature of a thermistor in the waveguide.

Hall effect

At the end of the last chapter, we came across an example of how electric charges in motion are turned aside by a magnetic field. The motion described was an exceptional one — diffusion of charges following their release by the energy of light. A much more familiar kind of charge motion is an electric current. Whenever a current flows through a magnetic field which is at least partly at right angles to the direction of the current, the charges are pushed to one side in the same way.

Fleming's hand rule for finding the direction of this push applies to positive current. Negative charges (electrons) flowing in the *same* direction are pushed to the opposite side, as we saw in Fig.

812. But negative charges moving in the same direction as positive charges are a positive current in the *opposite* direction. To be equivalent to a positive current, they must flow in the opposite direction to positive charges, so they are pushed to the same side as positive charges (if any) taking part in the same current. If that sounds a bit confusing, consider the same thing in an alternative way. As compared with positive charges in a current, electrons are subject to two reversals — once because of their opposite direction of flow, and again because of their opposite polarity — and two reversals cancel one another.

Current flowing through a metal consists entirely of electrons. A magnetic field directed across the conductor pushes them a little to one side, making that side more negative than the other. In Fig. 903, where the magnetic field is supposed to be directed away from you into the paper, the current (positive) is moving upward, meaning that electrons are flowing downward. They are, therefore, pushed to the left, so that side is found to be slightly negative. This peculiarity is called the Hall effect, after a young man at Johns Hopkins University who discovered it in 1879.

Because there are such enormous numbers of electrons available for electric currents in metals, even when the current is quite strong they need move only very slowly — say an inch or two per minute. Because the strength of the sideways push is proportional to their speed, it is very small. So small that Hall's professor had failed to detect it, and Hall did so only by using a very thin piece of metal (actually gold leaf) and a sensitive galvanometer.

In semiconductors, there are far fewer current carriers than in metals, so for each unit of current they must move faster and consequently get a far stronger sideways push from the magnetic field. Hall effect is, therefore, much more noticeable in germanium, say, than in copper. It is much more noticeable still in indium antimonide, which is remarkable for the high speed at which electrons and holes can move through it; their *mobility*, as it is called. That is why indium antimonide is so often chosen for devices making use of Hall effect.

Distinguishing holes from electrons

Another way in which semiconductors differ from metals is in having both positive and negative current carriers. To make a current in the same direction, holes must flow in the opposite direction to electrons, so they are pushed to the same side. If

the current were made up equally of both, the sideways tendency would also apply equally to both, but being toward the same side, their charges would cancel out one another (Fig. 904). So the net Hall effect in that material would be nil. If the semiconductor were p-type, with holes predominating, the polarity of the Hall voltage would be reversed compared with n-type.

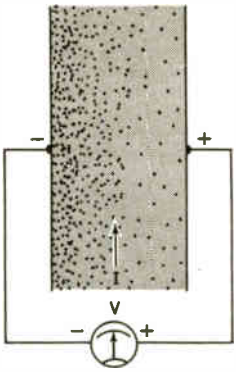


Fig. 903. *Hall effect. Electric charges forming a current are pushed to one side when a magnetic field acts at right angles to the paper; the result is a difference of potential between the two sides.*

Hall not only discovered the effect itself, for which alone he would deserve great credit; but he also carried out a very fine

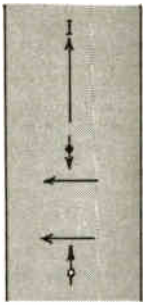
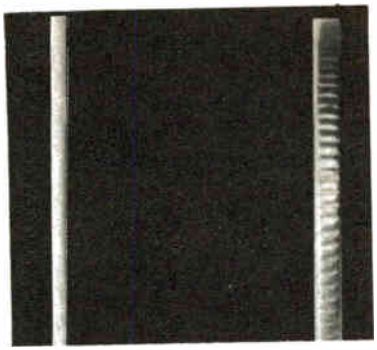
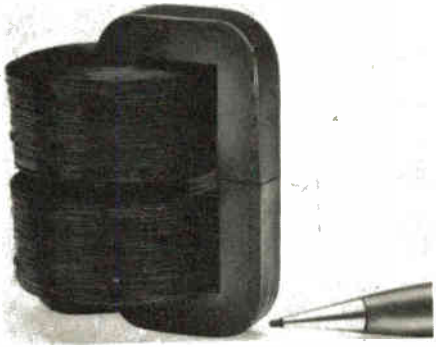


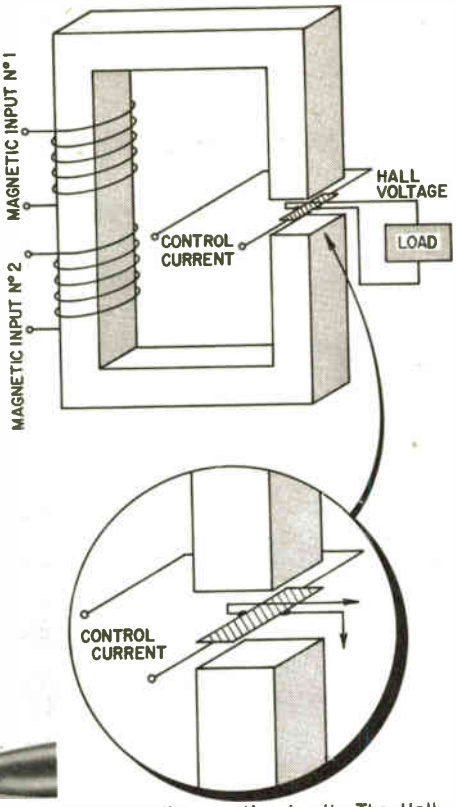
Fig. 904. *Both electrons and holes in a current are pushed to the same side in Hall effect, but being opposite in polarity cause opposite results.*

program of research on it, during which he noted that the sideways voltage depended on the material in which the current was flowing (other things being equal) and, with some of them, was reversed in polarity. He tried hard to explain these differences, and can scarcely be blamed for failing, for the number of mobile current carriers in a material depend in a very complicated way on its atomic structure, knowledge of which was still far in the future. The discovery of electrons even didn't come about until 18 years later, let alone holes. (As a matter of fact, although Hall was careful not to be too cocksure about it, he gave very good rea-

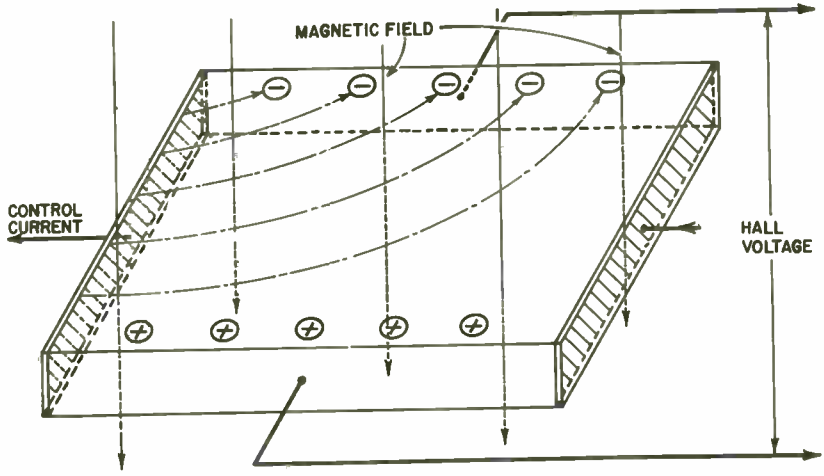




The Hall-effect device (upper photo) is very thin. Typical magnetic circuit (lower photo) uses two center-tapped coils.



The drawing at the upper right shows a typical 2-coil magnetic circuit. The Hall-effect device is placed in the magnetic field between the poles of the electromagnet. The drawing at the bottom shows the relationships of the magnetic field, control current and Hall voltage.



sons for concluding that an electric current through metal consists of moving *negative* charges.)

This is where the simple explanations of hole movement as being really a series of electron movements in the opposite direction break down. For if that were all, then Hall effect would be unable to distinguish between them. Both would give the polarity shown in Fig. 903. What is really needed is an understanding of the nature of the electron on a basis of wave mechanics, which is beyond the scope of this book. It is, however, an undoubted experimental fact that holes behave as if they were mobile positive charges.

Uses of Hall effect

This ability to tell whether current carriers are negative or positive (electrons or holes) is most valuable in research on semiconductors. It is the usual way of finding out their mobility and density in different materials.

The fact that for a given current the Hall voltage is proportional to the flux density of the magnetic field makes it also a convenient way of measuring flux density. A small strip of material — preferably indium antimonide, because that gives the strongest Hall effect — is mounted at the end of a probe so that it can be inserted in the air gap or other place where the field is to be measured. It is arranged so that a standard amount of current is passed through the strip from end to end, and side connections allow the Hall voltage to be read on a meter which is scaled in flux density (Fig. 905). This device can be made so sensitive that a large deflection is given by a fraction of the earth's magnetic field.

Another use for the same effect is in an alternative to the usual "magnetic amplifier" type of clip-on probe for measuring direct current without breaking into the circuit. The magnetic field around the wire carrying the current affects a Hall strip mounted in the probe and gives a reading on a sensitive meter.

Hall effect can also be used for measuring the power of radio waves. Such waves consist of varying electric and magnetic fields, in phase regarding time but at right angles to one another in space. So they supply both the end-to-end electric field and the through magnetic field, and all one has to do is measure the sideways Hall voltage. Of course, it is a little less simple in practice than it sounds, but all experimenters are familiar with that sort of thing!

Still another use is as a modulator for special requirements.

The valuable feature is that the Hall voltage is exactly proportional to both current and flux density, in contrast to tubes and rectifiers which are never quite linear; by varying either, it is modulated. Another feature is that modulation is effective right down to zero frequency. There are many other applications of Hall effect — too many to describe here.

Electron emitters

Although the triode tube was invented in 1906, progress in a big way had to wait until semiconductors came to the rescue

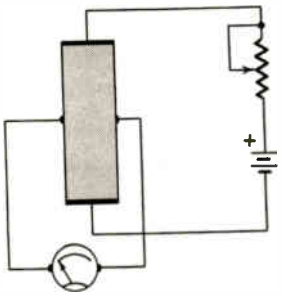


Fig. 905. *A small slip of semiconductor, mounted in a probe and supplied with end-to-end current, can be used to measure magnetic flux density.*

in the 1920's. Until then, simple metal filaments (normally tungsten) were used, and these had to be heated as bright as electric lamps in order to emit a useful number of electrons. The reason, as we already know, is the large amount of energy — equivalent to about 4½ volts per electron — needed to fling the electrons clear of their atoms. For a time the situation was improved by a process which resulted in tungsten filaments being coated with a thin layer of thorium, which emitted at a considerably lower temperature. But that soon gave place to filaments or indirectly heated cathodes, usually consisting of nickel coated with oxides of barium and strontium, which are semiconductors.

Oddly enough, discovery of the good emitting properties of these materials was made before the tube itself, by Wehnelt in 1903. Even now, the reasons for their low emission energy — only about 1 eV, allowing filaments to be run at about 700°C instead of 2500°C for bare tungsten — are not fully understood. But the advantages are enormous. There is not only the saving in power; oxide coatings make indirect heating practical, and develop much better tube characteristics by enabling grids to be placed closer to cathodes.

Phosphors

Everything in this book comes down basically to this: that just as the face of a boxer is knocked into various shapes by the kinds of punches it receives, so atoms are knocked into various shapes by the packets of energy they receive in a number of different ways. (A major distinction is the ease and speed with which an atom can revert to its original condition.) The different shapes are due to one or more of the atom's electrons being knocked into higher energy states. Sometimes they are knocked clean out of the atom, as we have just seen.

The sources of energy we have considered so far have been heat and light. These are needed, for example, to make electrons (or holes) available as current carriers inside semiconductors. Fortunately, the heat of the surroundings, even in the coldest weather, is sufficient.

Another way energy can come is more closely in line with the boxer's punches — bombardment by electrons or other particles. This often knocks other electrons right out of the bombarded material, the effect being well known in tubes as secondary emission. But, as far as semiconductors are concerned, we are more interested in the less violent disturbances of bombarded atoms — those that just lift their electrons into more distant orbits — because these can return to normal, giving back the packets of energy in the form of radiation.

If the energy differences between the higher and lower orbits happen to come within the range 1.65 — 3.3 eV, the radiation is within the visible band, and the end product of the electron bombardment is light. This is so with the semiconductor materials known as phosphors, with which the ends of our television and oscilloscope tubes are lined. Typical phosphors are the oxide, sulfide and silicate of zinc; cadmium sulfide, and magnesium tungstate. The choice depends on the color of light which the electron bombardment is required to excite.

Just as transistor semiconductors need small quantities of impurity to make them useful, phosphors need "activators" to increase the amount of light. The colors are also affected. Typical activators are copper, silver and manganese. Lead and iron are anti-activators or poisoners. Most good phosphors are n-type semiconductors.

In some phosphors, the electrons return within ten billionths of a second of having been knocked into the higher orbits, so the light appears only while the material is actually being bombarded.

This phenomenon is called fluorescence. In others, the electrons are trapped, sometimes for long periods of minutes, before returning and shining forth. This delayed effect is called phosphorescence. It too is affected by the activator.

Scintillation counters

The phosphors in cathode-ray tubes are bombarded by beams consisting of innumerable electrons. But for scientific purposes connected with radioactivity it is most desirable to be able to

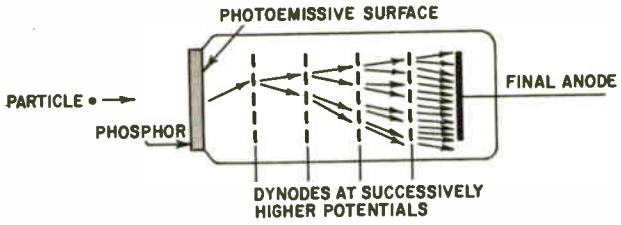


Fig. 906. *Diagram of a scintillation counter using an electron multiplier tube, or photomultiplier. In practice the dynodes have to be specially shaped and arranged so as to guide electrons from one to another correctly.*

observe and count hits made by individual particles. For electrons, crystals of anthracene have been successfully used, and, for the heavier alpha particles, zinc sulfide.

In the earliest work of this kind, scientists had to wait in total darkness for hours to allow their eyes to reach maximum sensitivity to detect the faint splashes of light, or scintillations. But now the work is eased by the use of a photomultiplier to amplify them. This is an electron-multiplier tube modified by having a light-sensitive target as the source of the input signal; it is shown very diagrammatically in Fig. 906. The particle to be detected hits the semiconductor crystal (phosphor) and makes its faint flash of light. The light strikes the target, which is like the cathode of a photoemissive cell, and makes it emit one or more electrons. These are accelerated by the positive potential of the first anode or dynode, and hit it hard enough to knock out several times as many electrons from it; these, in turn, are accelerated to hit a second dynode where the number is multiplied again, and so on. A single tube can in this way be made to give a total amplification of 10,000. From that point, it is a simple matter to rig up an electronic counter to total the number of particles detected, thus releasing a scientist for more rewarding work. Counting-speeds of up to one million per second have been achieved.

Electroluminescence

Yet another way of disturbing semiconductor atoms sufficiently for them to light up is by means of an alternating electric field. So far, this method has not come anywhere near competing economically with the filament and gas-discharge methods of generating light electrically, but for special purposes such as illuminated notices, signs in aircraft, movies, etc., it has advantages. One of these is the flatness of the light source; another is the absence of heat. It also offers intriguing possibilities in television, not yet realized.

To excite the phosphor, it is sandwiched between what are, in effect, the plates of a capacitor. Obviously, ordinary metal plates would be no good, as the light wouldn't be seen. But glass

Fig. 907. *Section of an electroluminescent panel, which glows under the stimulation of an alternating electric field.*

can be given a conducting surface by heating it to nearly its melting point in contact with the vapor of a tin compound, causing atoms of tin to diffuse into the glass. Both the electrodes are of this kind if the light is to be seen on both sides; otherwise, one of them can be a metal plate. Fig. 907 shows the latter arrangement.

The brightness increases very steeply with the electric field strength, which is the voltage per unit distance between the plates. So, for a given supply voltage, this distance should be as small as possible. The limit of thinness is reached when there is a risk of breakdown. The choice of phosphor is much the same as for cathode-ray tubes; a typical one is zinc sulfide activated by copper. To increase its electrical strength, the phosphor powder is usually suspended in a plastic or resin sheet. The active material, or more easily the electrodes, can be formed into letters or other signs.

The brightness increases with frequency, but while there is a definite advantage in using the aircraft 400 cycles rather than the usual 60, a much higher frequency would increase losses excessively.

By connecting an electroluminescent panel in series with a photoconductive cell, light can be amplified. The original light

shining on the cell, as in Fig. 908, reduces its resistance and so applies a greater voltage to the panel, making it light up more brightly and reinforce the original light.

Fig. 909 shows a method of amplifying a light image or picture. The original light shines through a conductive glass sheet which forms one electrode as before, and it illuminates a layer of photoconductor. Where it is brightest, it reduces the photoconductor resistance most and applies greatest voltage to the phosphor, which accordingly glows brightest and shines through the second conductive glass sheet.

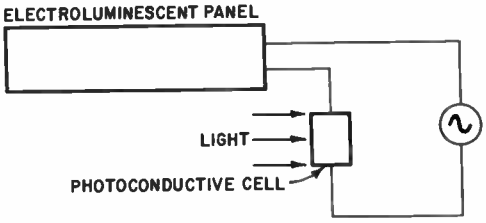


Fig. 908. *In this circuit the strength of field applied to the electroluminescent panel — and hence its brightness — is controlled by a photocell in series. If conditions are right, the light changes produced are greater than those applied to the cell.*

Very small panels — only about 1/4 inch square — have been used as relays and storage devices in electronic computers. Switching one of these panels to an ac supply lights it up. The light shines on a photoconductive detector, which can be used to switch on another panel or indicating device, or operate a hold circuit to maintain the first after the original ac pulse has ended. There are many possibilities, but a restrictive factor is the maximum speed of operation, which by present standards is low.

Masers

Here, we will have to recall once again the basic principles set forth in Chapter 3, especially the direct relationship (discovered by Planck) between the frequency of radiation and the size of the packets or "atoms" of energy (photons) carried by that radiation. For instance, if the radiation frequency happens to be 530 megamegacycles, visible as yellow light, its energy can come only in multiples of 2.2 eV. Fractions are impossible. So if electrons, excited by some form of energy intake into high-energy orbits of their atoms, drop back into orbits 2.2 eV lower in energy,

yellow light must be given out. Similarly for other frequencies.

In certain circumstances too involved to explain here, the atoms of some semiconductor materials have orbital states with much smaller energy differences, so that the radiation has pro-

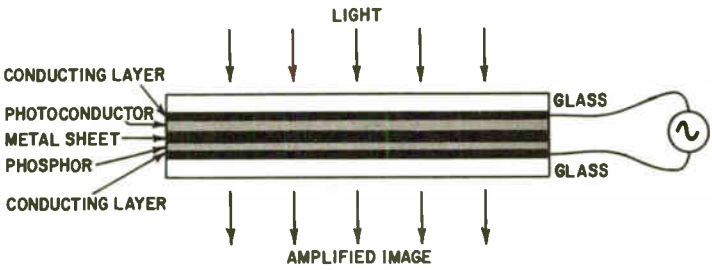


Fig. 909. *Section of panel combining the photoconductive cell in Fig. 908 with the electroluminescent part.*

portionately lower frequencies; low enough in fact, to be classified as radio frequency. The electrons can be excited into the higher-energy states by exposing the material to radiation of that frequency. But no useful purpose is served if the electrons just

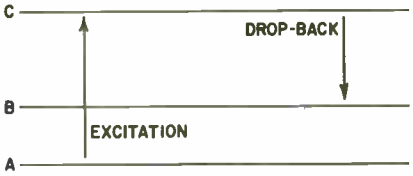


Fig. 910. *Three energy levels employed in a solid-state "maser."*

shuttle back and forth between those two states, alternately taking in and returning photons. However, it is possible to excite them between states A and C (Fig. 910), which might be compared to the first floor and third floor of a house, in circumstances which allow them to drop back from C to B (the second floor).

At these comparatively low frequencies, the electrons are not nearly in such a hurry to drop back as they are at higher frequencies, such as those of light. But they can be hustled by even a very weak radio signal if it has exactly the frequency corresponding to the energy drop between C and B. The energy the electrons give up in their drop boosts the radio signal. So the device is an amplifier.

Now you can see why it was important to excite the electrons

from some state other than B. It would be no use trying to amplify a very weak signal if there was a much stronger exciting source right in the same apparatus, swamping it.

This seems to be a simple type of amplifier. But in practice it is not really at all simple; for one thing, to get conditions right, the semiconductor has to be located between the poles of a very powerful and precisely adjusted magnet. For another thing, it has to be maintained at a temperature of anything up to 270°C below zero. This being so, you may be wondering why anyone would go to such trouble. The point is that ordinary amplifiers generate so much electrical "noise" in themselves that extremely weak microwave signals, such as those received by radio telescopes, are lost in it. Because the device just described is electrically quiet, it increases the range of those very expensive instruments. It is called a *maser,* from the initial letters of "Microwave Amplification by Stimulated Emission of Radiation."

Strain gauges

Engineers often want to know how much their structures are pulled, pushed or bent out of shape by loads and working stresses. They find out by using strain gauges, which are usually folded lengths of thin wire mounted on pieces of paper firmly stuck to the structure being investigated. If the structure is stretched by a load, the wire is stretched too, and its resistance increases. The amount of increase — and hence the mechanical strain — can be measured by a suitable Wheatstone bridge.

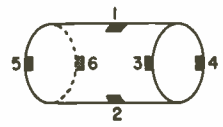


Fig. 911. *Example of a semiconductor strain gauge for measuring torsion. The numbered points are electrical contacts.*

These wire gauges may be superseded by semiconductors, which are about 50 times more sensitive and can be arranged to measure tension, compression, sheer and torsion. Our old friends germanium and silicon are well in front, but compounds such as gallium arsenide have also been tried successfully.

Fig. 911 shows a semiconductor strain gauge made for torsional measurements. Bias is applied between contacts 3 and 5, and 6 and 4. When the gauge is unstressed, contacts 1 and 2 are at the same potential, but torsion between the ends of the semiconductor gives rise to a proportionate voltage.

Measuring submicroscopic distances

We have already noted indium antimonide as an outstanding semiconductor because of its having the largest Hall effect. It also

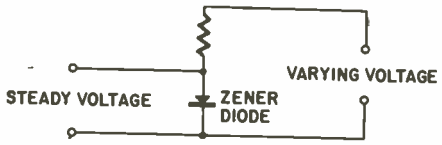


Fig. 912. *Simple voltage-stabilizing circuit using a Zener diode.*

tops the list in what is called the magneto resistance effect. This is an increase in resistance most semiconductors show when they are placed in a magnetic field.

It could of course be used for measuring field strength, but a less obvious application is measuring very small movements by putting it in the part of a field where the strength changes very steeply with distance. In certain experiments, the center part of a small rectangular piece of indium antimonide was cut out and the remainder fitted with metal contacts to make it into a miniature Wheatstone bridge. The part forming two "arms" of the bridge was put between the poles of a powerful magnet so that each arm was equally affected and the bridge was balanced. A very slight movement either way unbalanced the bridge, causing a deflection on a transistor-amplified meter.

By this means, movements of one hundred-millionth of an inch (equal to about one thousandth of one ultra-violet wavelength!) were detected. The theoretical limit is reckoned to be about 250 times smaller still! As it was, however, there was great difficulty in keeping the temperature constant enough to avoid comparatively large shifts due to expansion or contraction of the parts.

Because the equipment can be quite rugged, it is suitable for industrial measurement and control of very small variations in thickness, etc.

Zener diodes

So much for semiconductive materials used singly; now for junctions between different kinds, or with metals.

In using the diodes and transistors described in Chapters 6 and 7, one must take care not to apply so many volts in the reverse direction that the device is broken down, for it would then offer hardly any resistance and would probably be ruined instantly by

the uncontrolled current flow. Zener diodes, though they are specially made for breaking down, are no more immune from sudden death than any other kind; survival depends on their being used in the correct way, i.e., with sufficient resistance in series to limit the current to a safe amount.

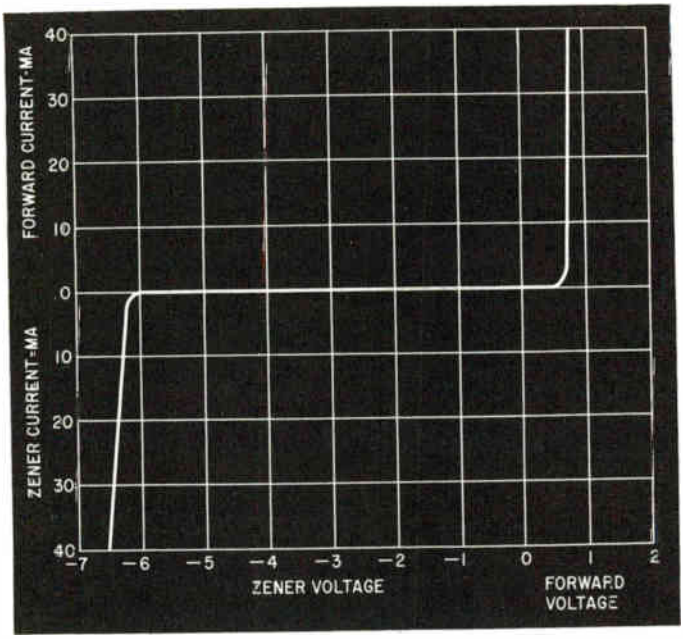


Fig. 913. *Current/voltage characteristic curve of typical Zener diode.*

Their most important characteristic is the voltage at which they break down. Below that their resistance is very high; above, it is very low. So if a Zener diode is connected in series with a fairly high resistance to a supply giving more than its breakdown voltage (Fig. 912), that voltage will exist across the diode almost regardless of variations in the supply voltage. In this way a fluctuating supply can be made to provide a steady voltage.

This arrangement is exactly the same as for gas-filled stabilizer tubes, but whereas such tubes are available only for a limited number of voltages, none below about 60 volts, Zener diodes can be made for any voltage from about 2 volts to hundreds of volts.

Silicon is the most suitable semiconductor, and the construction is similar to that of rectifiers, except that the silicon is doped

with an impurity to give the required low breakdown voltage. For any particular diode, this voltage can be relied upon always to be nearly the same, especially if it is in the region of 6 volts (which happens to be right for many transistor circuits). Lower breakdown voltages tend to fall a little with rise in temperature, and higher ones tend to rise. Fig. 913 shows a typical Zener characteristic curve.

Typical maximum current ratings range from 1 to 500 ma, but there seems to be no limit to what can be made. Obviously, the resistance of the rest of the circuit must be high enough to prevent more than the rated current passing when the supply voltage is at its highest. For example, suppose the diode is rated at 6 volts 50 ma, and the highest supply voltage is 15. That leaves

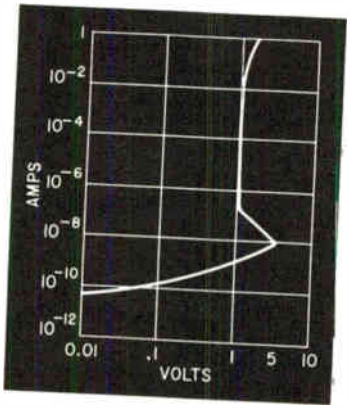


Fig. 914. *This characteristic curve of a cryosar has a negative-resistance section which is useful for computor switching elements. Note the logarithmic scale, used to include a wide range of current.*

9 volts to be absorbed in the resistance, which should therefore be $9/0.05 = 180$ ohms. Allowance must be made for component tolerances. To find how much voltage-regulated current could be drawn off from across the diode, the *lowest* supply voltage has to be regarded. Suppose it was 11 volts; then there would be 5 volts across the resistance, which would pass $5/180 = 0.028$, or 28 ma. If more current were drawn, the voltage across the diode would drop below its breakdown point and the diode would cease to function; the voltage across it would fall just as if the diode were not there.

Cryosars

Not long ago any idea of operating part of a regular plant equipment at a temperature of —269°C would have been ridiculed. But nowadays a supply of the coldest liquid in the world — helium — can be laid on almost as easily as milk. Its importance is that its liquefying temperature, —269°C, is only about 4° above absolute zero, so is usually stated as 4°K, the "K" standing for Kelvin, the famous Scottish scientist. At such a temperature, many things behave very strangely; for instance, lead has no resistance at all, and a current set up in it will continue flowing for months or even years without any emf.

In all the semiconductor devices considered so far, we have taken for granted that the temperature was high enough for practically all the impurity atoms to be ionized, so that the number of carriers (electrons and holes) available to conduct

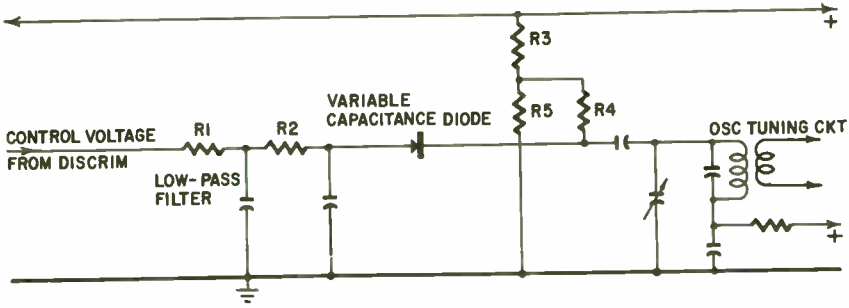


Fig. 915. Afc circuit, using a semiconductor diode as the variable reactance element.

electricity through the material was practically the same as the number of impurity atoms. In other words, that we were working on the flat part of the curve in Fig. 410. But we saw in Chapter 4 that these carriers are not inevitably parted from their atoms; it is heat that does it. Near absolute zero the heat has insufficient energy even for that, hence the steep fall to the left of Fig. 410.

A very cold semiconductor therefore has a high resistance. But only to low voltages. At first the emf can move only the very few carriers that are free. If it is raised, a point is reached where these carriers are moved so fast that they shake loose some of the bound carriers, and a chain reaction ensues — the avalanche again — with the result that the current suddenly becomes perhaps millions of times greater.

This looks very like a low-voltage Zener diode, but with some materials there is a negative-resistance slope between the high

and low resistance parts, as in Fig. 914. Here, then, we have another oscillating or switching device to add to those described in Chapter 7. It has been named the cryosar, from "cryo" (very cold), and switching by avalanche and recombination. Even though liquid helium is not such a rare supply as it used to be, you might well wonder why cryosars should be worth the trouble of maintaining at 4°K. The answer is that they can be unusually cheap, reliable and compact – to the extent that 200,000 could be put into one cubic inch, opening up the way for still more "intelligent" computors.

Diodes as variable capacitors

We have already noted that a semiconductor diode, like a vacuum tube, has interelectrode capacitance, but unlike a tube its capacitance varies quite a lot with the reverse voltage applied – short of breakdown voltage, of course. Fig. 511 in Chapter 5 is an example of the relationship in a typical junction diode.

One obvious use for this is in automatic frequency correction. The purpose of afc is to pull a radio receiver back into tune if it drifts. The general method is to use the dc component of the FM discriminator output voltage (or an added discriminator if the receiver is for AM) to control reactance, either capacitive or inductive. The usual device for doing this is a tube circuit so arranged that changing the grid voltage varies the reactance between two points of the circuit. These two points are connected in parallel with the oscillator tuning circuit and, whenever its frequency tends to drift away from that needed for exact tuning, the resulting positive or negative discriminator voltage

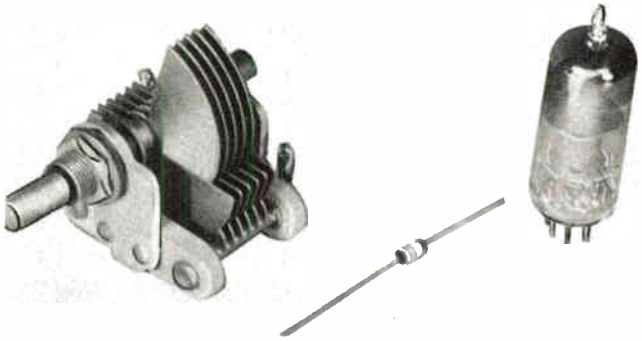


Fig. 916. *The capacitance of this little semiconductor, the size of a ¼-watt resistor, varies with the voltage applied to it.*

varies the tube reactance in the right direction to bring it back.

A semiconductor diode is much smaller and more convenient than the tube used for reactance variation. Fig. 915 shows a typical circuit, which is self-explanatory, except perhaps for the potential divider R3-R5 used to supply a few volts positive bias to the diode to bring it to the center of its capacitance/voltage characteristic curve.

Ordinary germanium and silicon junction diodes are suitable in this role for frequencies up to about 100 mc. Silicon is particularly so, because of its high resistance to reverse voltage, so as a capacitor it has little leakage and a reasonably high Q. A silicon diode for the purpose has been marketed under the name Varicap, and typical Q figures at 50 mc are in the region of 15. This may seem low compared with tank circuits, but it must be remembered that the diode capacitance is usually only a part of the whole tuning capacitance and the overall Q may be much higher. The fact that a diode of this kind is no larger that a 1/4-watt resistor is an attractive feature. Besides afc, applications include voltage-operated tuning, frequency modulation, tunable filters and sensitive remote-control systems. (See Fig. 916.)

Mavars

Most frequency-changers (or mixers) in microwave receivers are diodes, working on the principle that the resistance of the diode is made alternately low and high by the voltage of the local oscillator. This high-frequency variation of resistance modulates the amplitude of the signal being received, giving it a frequency which is usually lower. The disadvantage is that the mixer, being a resistive device, contributes electrical noise which may drown very weak signals.

We have already come across a largely noise-free method (the maser) of amplifying such signals before they come to the mixer, thus insuring that they are able to compete successfully with its noise. Freedom from noise in that system depends on the essential part of it — a semiconductor crystal — being cooled to a very low temperature by liquid oxygen or even helium. That is too expensive and inconvenient for any except very special purposes.

A more practical alternative is to use a nonresistive and, therefore, quiet device. If the incoming signal is fed into a *reactance* which is varied at twice the signal frequency, it is amplified — if necessary, to the point of oscillation. If required, the signal can have its frequency changed by a circuit modification at the same

time. The system is called a parametric amplifier or mavar ("Mixer Amplification by Variable Reactance").

Although the mavar could be used at relatively low frequencies, the need for minimizing noise is greatest at microwave frequencies. The problem then is to vary reactance at thousands of megacycles per second. Of several methods that have been used, one of the best and certainly the simplest is the semiconductor diode. Ordinary types are not effective at these frequencies, but special diffusion types have been used successfully.

Electricity direct from heat

It has been known ever since 1821 (when it was discovered by Seebeck) that heating a junction of two different metals generates an emf. This effect is still commonly used for measuring temperature differences, by means of what are called thermocouples.

It is less generally known that any two different metals in contact give rise to a difference of potential between them. The reason why it is not usually noticed is the same as for p-n junctions: if the circuit is completed, there is bound to be at least one other junction, and, if it is at the same temperature,

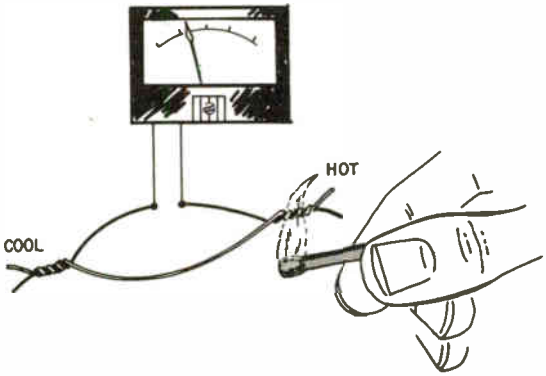


Fig. 917. *An electric current is generated in a circuit comprising two different metals or semiconductors if the junctions are at different temperatures.*

its potential difference exactly cancels that of the first junction. But the amount of the potential difference varies with the temperature. So, if the two junctions are at different temperatures, one of them has a larger potential difference than the other, there will be a net emf around the circuit, and current will flow (Fig. 917).

This emf amounts to only a few microvolts per degree difference in temperature; so a number of junctions are often connected in series to make what is called a thermopile, in which every alternate junction is heated and the others are kept cool. The difficulty about this is that if the wires between are made short and thick to keep the resistance down (so that a lot of current will flow), heat travels along them, reducing the temperature difference and emf. If they are made long and thin to prevent this, their resistance reduces the current. So the power efficiency is low; only about 1%, which compares very unfavorably with other methods of generating electricity.

Semiconductor junctions offer higher efficiency, and at least 7% has been attained in practical units. This can be quite useful where there is no supply of electricity but only a source of heat such as gas or kerosene. Fig. 918 shows the polarity for n- and p- types of semiconductor.

Refrigeration by semiconductors

Thirteen years after Seebeck's discovery, Peltier found that the effect was reversible; i.e., if an electric current is passed around a circuit comprising more than one metal, so that there are at least two junctions, one junction is warmed and the other

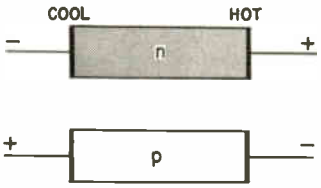
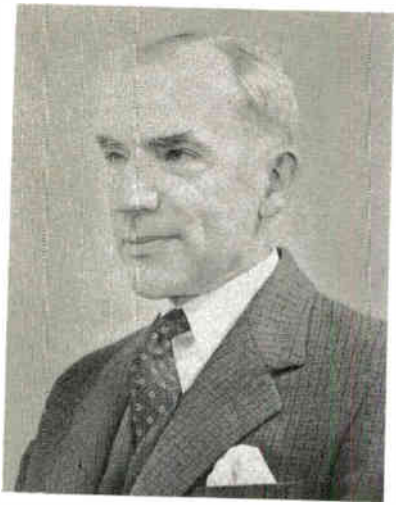


Fig. 918. *Polarities of thermoelectric effect with* n *and* p *types of semiconductor.*

cooled. The junction that gets cooler is the one which, if heated, would generate a Seebeck current in the same direction.

The Peltier effect is seldom noticed, because it is usually quite small and the cooling effect is likely to be more than offset by the heating due to resistance. Attempts to make a refrigerator on this principle, using metal conductors, have been unsuccessful, again because of the difficulty of making the wires short and thick and at the same time long and thin.[1] And again semiconductors offer the answer. It is possible for them to have low heat conductivity and high electrical conductivity. Bismuth telluride is a promising example, and temperature differences of as much as 85°C have been generated electrically.

[1]Electronic refrigerators have reached the patent stage with applications by Philco and Westinghouse for patents on electronic units using the Peltier effect. RADIO-ELECTRONICS Magazine, Oct. 1959, p. 18.



About The Author

MARCUS G. SCROGGIE, *well-known in England for bringing to technical writing a smooth and facile prose style, has written more than 700 technical articles.*

Educated at George Watson's College in Edinburgh, Scotland, he received his B.Sc. in Electrical Engineering from Edinburgh University. Subsequently he obtained his practical training at Bruce, Peebles Ltd., Edinburgh, and Creed Telegraph Co., Croydon. Not content with full days and evenings of working and writing, he made amateur radio his hobby and, using his call, 5JX, was the second amateur to communicate from Scotland to North America. In 1925 he joined Burndept Wireless Ltd. and in 1928 was appointed their chief engineer.

To be able to spend more time in technical writing and engineering consultation, he resigned his full-time position in 1932.

At the outbreak of World War II he joined the Royal Air Force as a radar officer. Throughout the Battle of Britain he was in command of Pevensey, a group of radar stations that were one of the half-dozen around the southeast coast of England that played such a vital role in that battle in 1940. It was at this station where operational experiments were evolved for ground control of interception by night fighters.

In 1945 Marcus Scroggie resumed his civilian life as consultant and technical writer. He is the author of the *Radio Laboratory Handbook* which by now has appeared in six editions. His *Foundations of Wireless* has been translated into French and Spanish. He is also the author of *Television* and *Second Thoughts on Radio Theory*.

index

Printed in the United States of America