WILLIAM SHOCKLEY

Transistor technology evokes new physics

Nobel Lecture, December 11, 1956

Introduction

The objective of producing useful devices has strongly influenced the choice of the research projects with which I have been associated. It is frequently said that having a more-or-less specific practical goal in mind will degrade the quality of research. I do not believe that this is necessarily the case and to make my point in this lecture I have chosen my examples of the new physics of semiconductors from research projects which were very definitely motivated by practical considerations.

An important fraction of United States industry adheres to the idea that research of a fundamental character is worthwhile from a practical point of view. This is outstandingly the case at Bell Telephone Laboratories where my co-prizewinners and I, together with our many colleagues, carried out the work described in these lectures. The attitude of Bell Telephone Laboratories has undoubtedly resulted to a substantial degree from the viewpoints of the four men who have been its research directors and subsequently its presidents. Each of these men, H. D. Arnold, F. B. Jewett, O. E. Buckley, and M. J. Kelly, has also been active and effective in governmental or civic affairs. All had obtained a thorough indoctrination in the research viewpoint in the course of their doctorate training in physics. My personal contact with two of these men had a significant influence on my planning of semiconductor research programs as I shall mention below.

My decision to come to Bell Telephone Laboratories immediately after obtaining my Ph.D. in 1936 was strongly influenced by the fact that my supervisor would be C. J. Davisson. Upon my arrival I was assigned by Dr. M. J. Kelly to an indoctrination program in vacuum tubes. In the course of this program Dr. Kelly spoke to me of his ideal of doing all telephone switching electronically instead of with metal contacts. Although I did not choose to continue work on vacuum tubes and was given freedom to pursue basic research problems in solid-state physics, Dr. Kelly's discussion left me continually alert for possible applications of solid-state effects in telephone

switching problems. Insofar as my contribution to transistor electronics has hastened the day of a fully electronic telephone exchange, it was strongly stimulated by the experiences given me during my early years at the Laboratories.

Before leaving the subject of research in industry, I would like to express some viewpoints about words often used to classify types of research in physics; for example, pure, applied, unrestricted, fundamental, basic, academic, industrial, practical, etc. It seems to me that all too frequently some of these words are used in a derogatory sense, on the one hand to belittle the practical objectives of producing something useful and, on the other hand, to brush off the possible long-range value of explorations into new areas where a useful outcome cannot be foreseen. Frequently, I have been asked if an experiment I have planned is pure or applied research; to me it is more important to know if the experiment will yield new and probably enduring knowledge about nature. If it is likely to yield such knowledge, it is, in my opinion, good fundamental research; and this is much more important than whether the motivation is purely esthetic satisfaction on the part of the experimenter on the one hand or the improvement of the stability of a high-power transistor on the other. It will take both types to « confer the greatest benefit on mankind » sought for in Nobel's will.

The five basic imperfections

Before discussing the selected group of experiments in transistor physics, I shall extend somewhat farther Dr. Bardeen's remarks[1] about the characteristics and interactions of electrons, holes, donors and acceptors. For this purpose I shall use a reference to water solutions of acids, bases and salts as an aid to exposition.

The dissociation of pure water to positive hydrogen ions and negative hydroxyl ions satisfies the mass-action law

$$[H^+]\ [OH^-] = \text{function of } T$$

where the concentration of H_2O is so nearly constant for dilute solutions that it may be taken as constant and not shown explicitly. This equation has as its analogue

$$n \cdot p = f(T)$$

for a semiconductor where n and p are the electron and hole concentrations. The equation is accurate, provided neither n nor p is so large that its statistics become degenerate. The normal electron-pair bond here plays the role of an undissociated water molecule. In pure deionized water and in a pure semiconductor, electrical neutrality requires that the concentration of the positive charges equals that of the negative charges:

$$[H^+] = [OH^-] \text{ and } p = n$$

A semiconductor which is so pure that impurities do not affect the property being considered is called *intrinsic* in regard to such a property. Using the subscript i to designate this case, we have

$$p_i = n_i \text{ and } n_i^2 = f(T)$$

The concentration n_i is referred to as the intrinsic concentration.

The chemical analogue to an n-type semiconductor is a base and the charge neutrality equations are

$$[H^+] + [Na^+] = [OH^-]$$

$$p + N_d = n$$

where N_d is the concentration of donors and it is assumed that all donors are fully ionized.

Similarly a p-type semiconductor is analogous to an acid:

$$[H+] = [OH-] + [Cl-]$$

$$p = n + N_a,$$

where N_a is the acceptor concentration.

A neutral salt also has its analogue which is called a *compensated* semiconductor. For this case the donor and acceptor concentrations are equal so that the equation for electrical neutrality

$$p + N_d = n + N_a$$

reduces to equality of n and p so that each is equal to n_i. The electrical conductivity of a perfectly compensated silicon or germanium crystal is almost equal to that of an intrinsic crystal; there may be a small difference due to the reduction of mobility by scattering by the charged ions. The difference between the low conductivity of the perfectly compensated semiconductor and the high conductivity of a neutral salt solution arises from the fact that the donors and acceptors are tied in place in a semiconductor while the cations and anions of a salt have mobilities comparable to that of an OH-ion.

Compensation, or rather overcompensation, plays a vital role in the manufacture of semiconductor devices. For example, it permits the conversion of n-type material to p-type by the addition of acceptors without the prior removal of donors. A crystal grown from an n-type melt may be changed to p-type by adding an excess of acceptors.

The words *majority* and *minority* are frequently useful in discussing semiconductors. In an n-type specimen, the majority carriers are electrons and the minority carriers are holes.

Holes, electrons, donors and acceptors represent four of the five classes of imperfections that must be considered in semiconductor crystals in order to understand semiconductor effects. The fifth imperfection has been given the name *deathnium.* The chemical analogue of deathnium is a catalyst. In the case of water as the analogue to a crystal, there is, so far as I know, no important corresponding catalyst. What deathnium does is to hasten the establishment of equilibrium between holes and electrons. If, due to the various possible disturbances important in transistor electronics, the concentration of minority carriers is, for example, substantially raised, then the minority carriers will combine with the majority carriers to produce normal electron-pair bonds, by this means restoring equilibrium. Deathnium catalyzes this recombination process. The symbols for the five imperfections are shown in Table 1.

Table 1.

1. - (excess) electron
2. + hole
3 deathnium
4. ⊕ donor
5. ⊖ acceptor

The role of deathnium can be illustrated in terms of the phenomenon of photoconductivity. If light shines on a germanium crystal, then the pairs of

electrons and holes that are formed will impart a conductivity to the crystal. This conductivity is known as photoconductivity. If the source of light is removed, the photoconductivity will die away, due to the recombination of the holes and the electrons. Thus, if an electron falls into an incomplete bond, one hole-electron pair will be eliminated.

The photoconductivity dies away with a characteristic time known as the lifetime. Thus, after the light is turned off, the photoconductivity will drop to approximately one-half its value in one lifetime. This process continues with a reduction of approximately one-half in each subsequent period of one lifetime.

If the process of recombination of holes and electrons were a direct one, the lifetime would be the same in all germanium crystals. It is found experimentally, however, that two otherwise very similar germanium crystals will have lifetimes that differ by as much as 1,000 fold. In one crystal, the lifetime may be a millisecond, whereas in another it may be a microsecond. This variation in lifetime requires the presence of some sort of imperfection which catalyzes the recombination of the holes and the electrons.

Actually, there are several forms of deathnium. For example, if electrons having an energy of several million electron volts fall upon a germanium crystal, the lifetime is subsequently reduced[2]. From the investigation at Purdue University, it is known that such bombardment produces disorder of the germanium atoms[3]. A high-energy electron can eject a germanium atom bodily from its normal position in the crystal structure, thus leaving a vacancy behind, where there should be an atom, and causing the ejected atom to become either an extra atom or an interstitial atom fitting into a place in the structure which would normally be empty. It has been found at Bell Telephone Laboratories that these disordering effects function as deathnium. It has also been found that copper and nickel chemical impurities in the germanium produce marked reductions in lifetime[4].

The way in which deathnium catalyzes the recombination process is in-

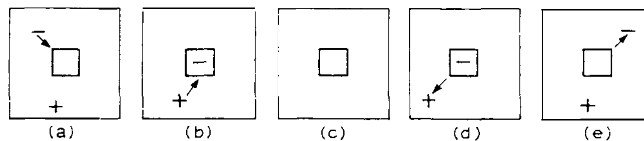

Fig. 1. A recombination center (deathnium) captures alternately an electron and a hole and thus catalyzes their recombination, as shown in parts (a), (b), and (c). The thermally activated generation process is shown in (d) and (e).

dicated in Fig. 1. In part (*b*) of this figure, an electron is captured by a deathnium center. The deathnium center thus becomes a baited trap which is ready to capture a hole. If a hole comes near to the deathnium center, the electron can drop into it, thus forming a normal covalent bond, and the deathnium center is then uncharged and ready to repeat the process.

It is characteristic of all microscopic processes that they may go backwards as well as forwards. Thus, the deathnium center may generate hole-electron pairs as well as eliminate them. The generation process is indicated in parts (*d*) and (*e*) of Fig. 1. In part (*d*) deathnium center captures an electron from an adjoining normal electron-pair bond. This produces a hole which wanders off. Somewhat later, the deathnium center ejects the electron and thus reverts to its empty state in which it is ready either to recombine or to generate another hole-electron pair.

Under conditions of thermal equilibrium, both the recombination process and the generation process proceed continuously. The energy required to generate the hole-electron pair is furnished by the thermal energy of vibration of the atoms in the germanium crystal. The condition of thermal equilibrium is achieved when the two processes balance. For germanium at room temperature, this leads to a conductivity of about 0.02 $ohm^{-1} cm^{-1}$.

Since the concentration of holes and electrons under equilibrium conditions is governed by a mass-action law, the product np is independent of the concentration of deathnium. For example, if the concentration of deathnium is doubled, both the rate of generation and the rate of recombination are doubled, but the equilibrium concentrations of holes and electrons are unaltered.

Evidence that the deathnium mechanism shown in Fig. 1 is correct has been obtained by studying the dependence of the rate of recombination upon hole and electron densities[5]. These studies are found to be in general agreement with the predictions based on the statistical model of Fig. 1.

The field effect

The experiment which played the largest role in stimulating the transistor electronics program at Bell Telephone Laboratories was the so-called field-effect experiment. I devised this experiment as a direct result of trying to invent a semiconductor amplifier having separate input and output circuits. From the then known properties of semiconductors, I concluded that a thin

film of silicon or germanium would have a substantial change in its conductivity if it were made into one of a pair of condenser plates and the condenser strongly charged. The surface charge, if represented as a change in mobile carriers, could appreciably increase or decrease the conductance of the film.

A number of experiments were initially tried using evaporated layers and other thin layers. These all gave negligible effects and progress was at a standstill until Bardeen proposed his theory of surface states to explain the lack of an observable effect.

Bardeen's model also explained a number of other previously mysterious phenomena[6] and led to the suggestion of doing the field-effect experiment at liquid-air temperature to immobilize the surface states. This gave the first positive result. How this led to a sequence of experiments culminating in the point-contact transistor is discussed in Dr. Bardeen's lecture[1]. The first publication of a positive effect was made by G. L. Pearson and the writer in 1948[7].

At the present time the field-effect experiment is playing a very important role in measuring the properties of semiconductor surfaces[8].

On the practical side the field-effect has been utilized to make transistor amplifiers having interesting properties quite different from those of junction transistors.

Injection and drift

At the time of the discovery of the point-contact transistor by Bardeen and Brattain there were a number of unresolved questions regarding its mode of operation. The original transistor showed evidence that the coupling between the input or emitter point and the output or collector point took place through conduction in a thin surface layer of conductivity type opposite to the underlying or base material. Somewhat later the idea that the emitter point might actually be injecting minority carriers into the body of the semiconductor developed. The development of this idea came as a result of two independent events: the invention of the junction-transistor structure by the present writer (as discussed below, injection plays an essential role in the junction transistor) and the observation by J. N. Shive[9] that transistor action could occur with the points on opposite sides of a thin slab of semiconductor.

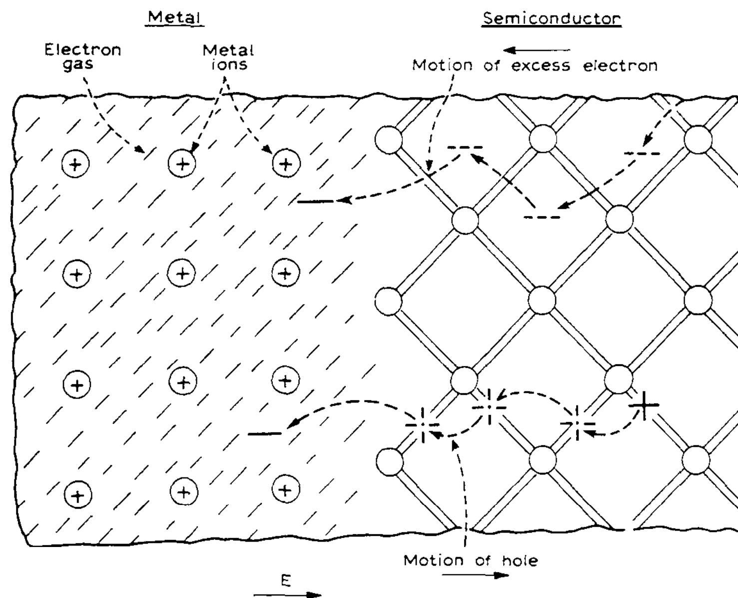

Fig. 2. Two possible mechanisms for current flow near an emitter point as described in text.

In order to test the model of carrier injection, J. R. Haynes and the author collaborated in the drift-mobility experiment or « Haynes' experiment » on germanium specimens[10]. In order to understand the significance of the experiment in elucidating transistor action the mechanism of current flow at the metal-semiconductor contact must be considered.

In Fig. 2, the metal is represented in a highly pictorial fashion. The valence electrons in a metal are thought of as forming an electron gas, which permeates the entire structure. Thus, the electrons are not held in position in valence bonds as they are in an insulator. The electron gas can flow freely through the structure of the metal, and this fact accounts for the high conductivity of metals. In the upper part of Fig. 2 one of the processes for removing electrons from the semiconductor is represented. Since the semiconductor is n-type, it contains excess electrons; these excess electrons may be drawn to the metal by its positive charge and thus enter the metal to produce a current of electrons flowing out of the emitter point through the connecting lead.

Another possible mechanism for electron transfer from semiconductor to metal is shown in the lower part of Fig. 2. In this case, an electron is withdrawn from one of the valence bonds adjacent to the metal. This process

also transfers an electron from the semiconductor to the metal, but when the transfer occurs, a hole is left behind. The hole is repelled by the positive charge on the emitter contact and moves deeper into the semiconductor.

Both of the processes discussed above have the same effect so far as the metal emitter point and conducting leads to the emitter point are concerned. Both produce a net flow of electrons from semiconductor to the emitter point and through the leads to the emitter terminal. It is thus evident that some more subtle experiment than simply measuring the current to the emitter point is necessary to show that both processes of electron removal from the semiconductor occur. Suitable experiments have been planned and performed, with the result that it is possible to show that both of the processes of Fig. 2 occur and also to determine the fraction of current carried by each. In fact, in a good emitter point it can be shown that more than 90 per cent of the current is carried by the process which injects holes into the semiconductor, and less than 10 per cent by the process which removes electrons.

In an ideal emitter point, all of the current would be carried by the hole-injection process. The reason for this result is that the electron-removal process does not disturb the state of affairs within the semiconductor. If electrons are removed from the semiconductor in the neighborhood of the emitter point, they are promptly replaced by electrons flowing from more distant parts of the semiconductor, and these electrons in turn are replaced by other electrons flowing in from whatever contact to the semiconductor completes the electrical-current path or circuit. In the hole-injection process the situation is quite different. Normally, the number of holes in the n-type semiconductor is negligible so far as a contribution to the conductivity is concerned. However, when electrons are removed from the valence bonds and holes are injected, relatively large numbers of holes will be introduced. The conductivity of the semiconductor will be increased in the neighborhood of the emitter point in much the same fashion that it would be if light were to shine on the semiconductor and produce hole-electron pairs. This disturbance in the electronic structure can be used to produce amplifying action in the transistor.

Instead of discussing the quantitative experiment which is used to distinguish between the two processes shown in Fig. 2, I shall describe a qualitative experiment which shows that hole injection does occur at an emitter point. This experiment permits quantitative studies to be made of the behavior of holes and provides a method for the direct measurement of diffusion and drift.

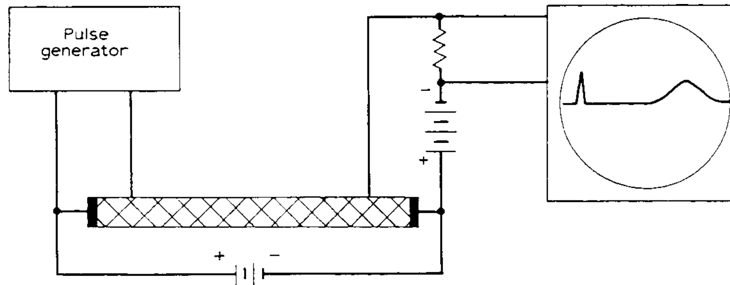

Fig. 3. Schematic representation of experiment to observe the drift and diffusion of injected holes in n-type germanium.

The experimental arrangement, which was first carried out in this form by J. R. Haynes, is illustrated diagrammatically in Fig. 3. The germanium specimen is in the form of an elongated point-contact transistor. There is, however, an extra contact on the base. The germanium is present as a rod, about 1/32 of an inch in cross section and approximately 1 inch long. A « sweeping field » is applied from end to end of the rod by a battery. This field acts in such a direction as to draw electrons from right to left through the rod. If any holes were introduced in the rod, they would drift from left to right.

When the pulse generator at the left-hand point contact, or emitter point, operates, the emitter point is biased positive and thus in the forward direction. According to the ideas presented in Fig. 2, this condition causes holes to be injected into the rod. These holes are then drawn down the rod by the sweeping field. After a time they arrive in the neighborhood of the collector point, which, as the figure shows, is biased negative. It thus tends to attract holes, and some of the holes flow to the collector point and thus contribute to the current flowing in the collector circuit. This current flows through a resistor, and the voltage across the resistor is applied to the vertical plates of a cathode-ray oscilloscope.

Under operating conditions, the operation of the pulse generator is accomplished electronically and is synchronized with the functioning of the oscilloscope, so that just before the switch is closed, the electron beam in the oscilloscope starts to move across the tube face from left to right. At time t_1 the switch to the emitter point is closed for a brief moment; the time of closing is indicated by a « pick up » signal on the face of the oscilloscope. After this nothing happens until time t_2 when some of the holes arrive at the collector point; the concentration of holes builds up for a moment and

then decays as the group of holes injected at time t_2 passes by the collector point. The arrival pulse at the collector point is not so sharp as the « pick up » pulse because the holes, which were injected approximately at one point and at the same time, spread out by diffusion so that by the time the group of holes reaches the collector point it is relatively large in extent along the rod.

It is evident that this experiment permits observation and measurement of both diffusion and drift. It is possible to measure the distance between the points and the electric field between the points; by calibrating the oscilloscope, the time of travel may be measured. Thus the drift velocity may be measured directly, verifying the fact that the disturbance occurring at the emitter point behaves precisely as would be expected if the emitter point injected small numbers of positive carriers into the rod. For example, if the distance between the points is doubled, the time lag between pick-up at t_1 and the arrival of the pulse is also doubled. This result shows that the carriers drift at a constant drift speed in the rod. But if the sweeping field is doubled, the time lag is cut in half. This fact shows that the speed of the carriers is proportional to the electric field. If the polarity of the sweeping field is reversed, we would expect the injected carriers to be drawn to the left in the filament so that none arrive at the collector point, and it is found experimentally that this is true.

As was indicated above, the spread of the time of arrival of holes is a measure of the diffusion constant. From studies of the dependence of this spread upon the transit time from emitter to collector, it can be verified that the holes spread out in accordance with the laws expected for diffusion phenomena. The value of the diffusion constant D can also be measured.

J. R. Haynes and his colleagues have performed various experiments of this sort. They have also experimented with the case of electron injection into p-type germanium and have dealt with the two corresponding cases for silicon. The values of mobility and diffusion constant which they obtain in this way[11] are tabulated in Table 2.

Table 2. Mobilities in cm^2/volt sec and diffusion constants in cm^2/sec.

	Electrons		*Holes*	
	μ	D	μ	D
Silicon	1.200	30	250	6.5
Germanium	3.600	93	1.700	43

It should be noted from Table 2 that in each case the ratio of diffusion constant to mobility is approximately 1/40, and the dimensions of this quantity are in volts. In other words, the ratio of D to μ is 25 millivolts. This value has a fundamental significance, and the relationship between D and μ is commonly known as the Einstein relationship. This relationship has recently been investigated in detail, by the means described above, for germanium[12]. The significance of this value of 25 millivolts is that an electron moving with random thermal energy will, on the average, be just capable of surmounting a potential hill of 25 millivolts. In other words, 25 millivolts is the electrostatic potential corresponding to thermal energy for one electron. Put in another way, it can be stated that if any electron was set in motion with thermal energy in free space against any electric field, the electron would be slowed down by the electric field and by the time it had moved 25 millivolts against the field its velocity would be brought to zero and it would start to move in the opposite direction. The fact that a value of 25 millivolts is obtained shows that the charge of the carriers which are drifting and diffusing in the Haynes experiment is the electronic charge. If it were half or twice this value, for example, the ratio of D to μ would be 50 or 12.5 millivolts, respectively.

When the Haynes experiment was first carried out, the procedure was varied by placing the two points on opposite sides of the filament and by using filaments of much greater width than thickness. All these experiments indicated that the interaction between the points took place through the bulk rather than over the surface.

Hot electrons and Ohm's law

Another fundamental experiment on the behavior of electrons and holes in semiconductors had its origin in a practical aim. Sometime during 1948 in an attempt to see how semiconductors could be made to amplify at very high frequencies, I came upon an idea which was independently discovered and first published as the *Staueffekt* by H. A. Krömer[13]. Krömer does not appear to have considered the effect from the point of view of amplification, and its utilization for this purpose was first published in 1954[14].

The basic principle of the staueffect may be understood as follows : a hole in the valence band will lose energy to phonons at a maximum rate P (max)

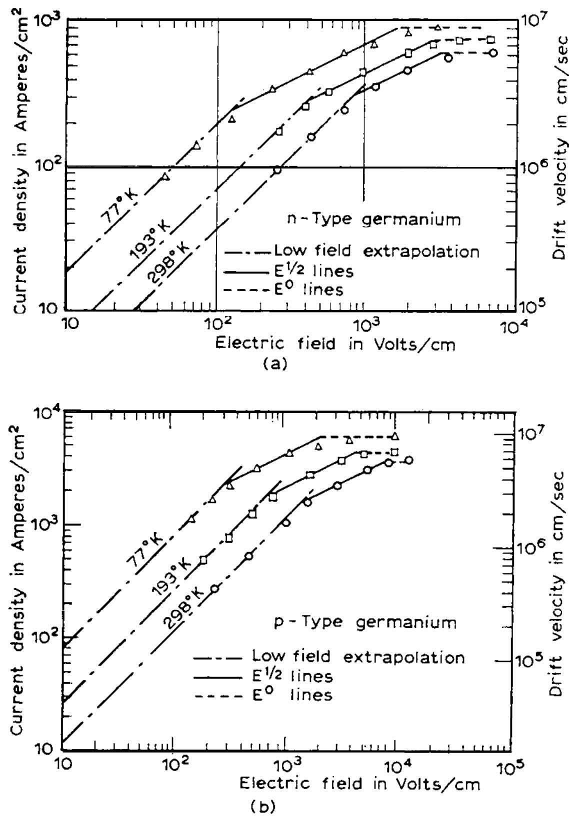

Fig. 4. *(a, b)*. Current density as a function of electric field. *(a)* n-type germanium; *(b)* p-type germanium.

when its energy is somewhere near the middle of the band. If subjected to an electric field E, its drift velocity v_d must be limited by

$$qEv_d < P(\max)$$

since it cannot gain energy indefinitely. What happens in more detail is this: if the hole gains too much energy, it moves to such low states in the valence band that its effective mass changes sign and it is decelerated instead of accelerated by the field.

The negative resistance effects were not observed. However, experimental situations were produced by E. J. Ryder and W. Shockley[15] in which the random energies of holes and electrons were raised to equivalent temperatures of many thousands of degrees.

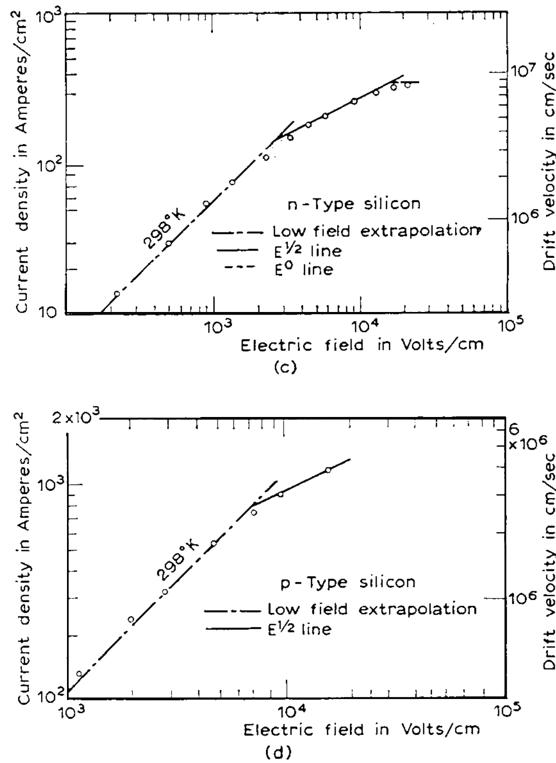

Fig. 4. *(c, d)*. Current density as a function of electric field. (*c*) n-type silicon; (*d*) p-type silicon.

These effects represent a fundamental deviation from Ohm's law in the sense that they result solely from the magnitude of the electric field and not from side effects such as changing numbers of holes and electrons due to changing temperature.

Fig. 4 shows several curves representing how the drift velocity of holes and electrons varies with electric field as reported by Ryder[16]. Although the general features of the curves are well understood, they are not yet adequately explained quantitatively[17]. The discrepancy between simple theory and experiment probably results from the complex structure of the energy surfaces for holes and electrons, that is, from deviations from the single-effective-mass model of semiconductors.

From a practical point of view, non-linearities in the drift velocity are important in calculating the characteristics of semiconductor devices.

The p-n junction and single crystals

The rectifying contacts of chief importance in the early days of transistor physics were made by pressing pointed wires on the semiconductor and subsequently conditioning them (« forming ») by passing heavy current. It was difficult at the time to interpret these procedures from an atomic point of view and this is still true at present.

An important decision in the year following the invention of the transistor was to place emphasis upon p-n junctions in the interior of the semiconductor rather than upon point contacts. From earlier work it was known that such junctions were rectifying and photosensitive. However, their characteristics were not entirely understood in terms of atomic models. Since a p-n junction can occur in a single crystal simply by a change in impurity content from donor dominated in the n-region to acceptor dominated in the p-region, it is a far simpler structure than a metal semiconductor contact and its behavior highly predictable on theoretical grounds. It seemed logical, therefore, to attempt to understand p-n junctions first and point contacts later.

Another reason for choosing to emphasize p-n junctions was the possibility of producing junction transistors, a possibility published in 1949[18].

A p-n junction is the simplest so-called *compositional* structure in semiconductor electronics. By a compositional structure is meant a single crystal of semiconductor in which the composition in terms of the donor and acceptor content varies in a systematic and controlled way. Before describing the theory of the functioning of a p-n junction, I would like to say something about the way in which p-n junctions whose behavior was satisfactory from a theoretical point of view were first achieved at Bell Telephone Laboratories. This history also is an example of the interaction of practical needs on a research program.

To begin, with attempts were made, particularly by M. Sparks, to produce p-n junctions by allowing molten drops of germanium of one conductivity type to fall upon a heated block of germanium of the other conductivity type. Although p-n junctions were obtained by these means, their characteristics failed to live up to the predictions of theory. (The problems were almost certainly those of cleanliness - the importance of copper was not known at the time.)

As a result of the needs of the Development Department for uniform material for the purpose of making experimental transistors, a program was

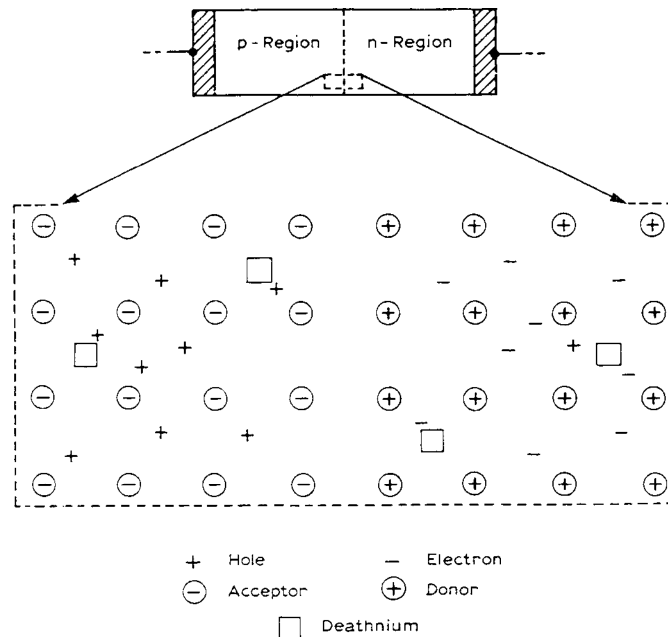

Fig. 5. A p-n junction and the distribution of imperfections in it. (For simplicity, compensating pairs of donors and acceptors are not shown.)

set up to grow large single crystals of germanium. It is interesting to note that the actual decision and emphasis on this program was due principally to J. A. Morton, who headed the transistor-development group, rather than by my own group or other research groups. I felt at the time that scientific studies could adequately be carried out by cutting small specimens from polycrystalline masses of material. As a result of Morton's urging, G. K. Teal, with the collaboration of J. B. Little, made an experimental apparatus for « pulling » single crystals of germanium from a melt of germanium heated in a graphite crucible into which was dipped a small single-crystal seed[19]. The recent advances of transistor science and technology are founded on these crystals. The addition of acceptors to an n-type melt during pulling changed the solidifying material from n- to p-type and gave the first good p-n junctions.

Another extremely important development in the preparation of materials should be mentioned. This is the method of *zone refining* invented by W. G. Pfann, also of Bell Telephone Laboratories[20]. Noting that impurities tend to be more soluble in molten germanium than in solid germanium.

Pfann devised a system of repetitive purification by crystallization. By making an elongated graphite boat and providing means to heat locally a small length or zone of it, he was able to pass molten zones successively from one end to the other of the germanium and by this means to sweep impurities clear of the major portion of the crystal and to concentrate them near one end. By this means germanium crystals having one impurity atom in about 10^{10} germanium atoms have been produced. The density of impurities in these crystals is thus comparable to the density of molecules in a gas at a pressure of 10^{-5} millimeters of mercury. It is appropriate to call zone refining the vacuum pump of transistor electronics.

Fig. 5 represents a p-n junction. In discussing its electrical properties, we will be concerned with the five kinds of imperfections shown in the lower part of the figure. From a mechanical point of view the crystal is practically homogeneous and perfect. A typical concentration for impurities in the crystal might be 10^{15} cm^{-3}. This density of imperfections is so low that if one were to traverse a row of atoms from end to end in the crystal one would, on the average, strike only about ten imperfections. Thus the crystal structure itself is only slightly altered by the presence of the imperfections. From the electrical point of view, on the other hand, the imperfections have profound effects.

As is shown in Fig. 5, the electrons are found chiefly in the n-region where they neutralize the chemical space charge of the donors, and the holes are similarly found in the p-region. In order for this situation to persist, as it does under conditions of thermal equilibrium, there must be an electric field present in the crystal. The idea that an electrical field is present in a conductor under conditions of thermal equilibrium is at first surprising. However, the necessity for such an electric field can readily be seen in terms of Fig. 5. Let us first of all suppose that no electric field is present across the junction; then as a result of diffusion, holes will cross the junction from left to right and electrons will similarly diffuse across the junction from right to left. As a result, there will be a net transfer to the right of positive charge across the junction. This will result in an electric field which will build up to just such a point that further current flow is prevented.

The electric field in the p-n junction is represented in Fig. 6 *(a)* in terms of an electrostatic potential ψ. The remaining parts of the figure show how this electric field and electrostatic potential arise from the charge densities involved. The chemical charge densities are shown in parts (*b*) and (*c*) of the figure. In this example it has been assumed, as represented by the N_d

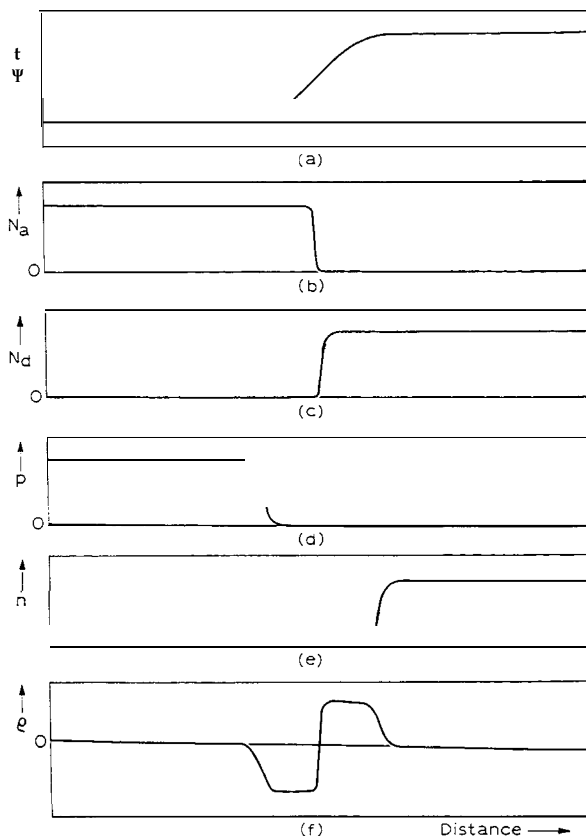

Fig. 6. Potential and charge distribution in a p–n junction.

and N_a curves, that the transition from n-type to p-type occurs abruptly at the junction and that compensation of one impurity type by another is not involved. In the presence of the electrostatic potential shown in part *(a)* of the figure the holes tend to be pushed to the left. As a result, the hole density drops to a small value before the junction itself is reached. Electrons having a negative charge tend to move to the points of highest electrostatic potential, and thus they also are not found near the center of the junction. As a consequence, the chemical charge density is not compensated by holes or electrons in the immediate neighborhood of the junction. Thus an electrostatic dipole layer is formed at the junction, the charge density being as shown in part *(f)* of the figure. This dipole layer is seen to be of just the nature necessary to produce the electrostatic potential shown in part *(a)*.

Mathematically, what is done in order to determine the shape of the elec-

trostatic potential in Fig. 6 *(a)* is to solve a differential equation. If the dependence of electrostatic potential upon distance is regarded as the unknown, then from it and certain principles of statistical mechanics it is possible to write an expression for the charge density due to the holes and the electrons. These charge densities can be combined with those due to the chemical imperfections in order to obtain a differential equation for the electrostatic potential. This differential equation is Poisson's equation, which relates derivatives of the electrostatic potential to the charge density. When this equation is solved, it is found that the situation in the p-n junction under thermal equilibrium conditions is as represented in Fig. 6.

Under conditions of thermal equilibrium no net current of either holes or electrons will flow across the junction. It is advantageous, however, to consider this equilibrium situation as arising from compensating currents. We shall illustrate this by considering the flow of holes back and forth across the junction. Although the density of holes is small in the n-region, it is still appreciable and plays a vital role in the behavior of the p-n junction. Let us consider the course of a hole that arrives in the n-region by climbing the potential hill as illustrated in Fig. 7. Having climbed the hill and arrived at the plateau of constant electrostatic potential in the n-type region, it will then proceed to move by a random diffusive motion. The most probable outcome of this motion will be that it will diffuse to the steep part of the

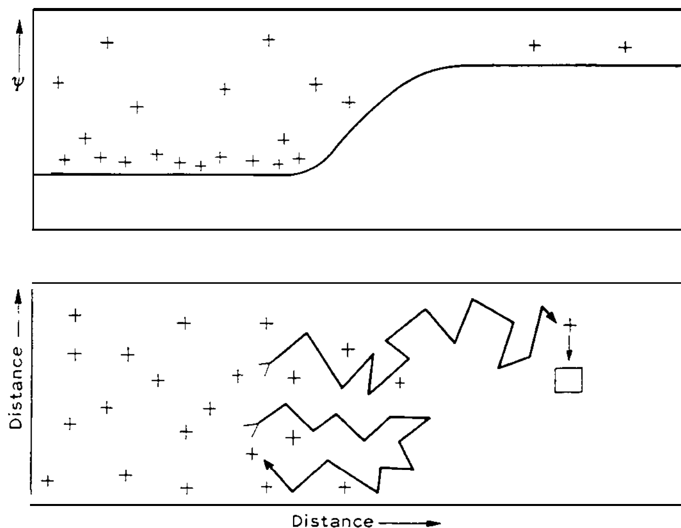

Fig. 7. Hole current from p-region to n-region in a p-n junction.

hill and slide back down into the p-type region. We shall not be concerned with holes which follow this particular course. On the other hand, it may, by chance, diffuse more deeply into the n-type region. In this event, it will on the average diffuse for lifetime τ, and subsequently it will be captured by a deathnium center in which it recombines with an electron.

The average depth to which holes diffuse in the n-type region depends upon the lifetime. The holes spread out in the region by diffusion. When the suitable differential equation describing this process is solved, it is found that the average depth to which they penetrate is given by the equation.

$$L = \sqrt{D\tau}$$

where L is known as the diffusion length, D is the diffusion constant for holes, and τ is the lifetime for holes in the n-region. Thus under equilibrium conditions a current of holes flows from the p-region into the n-region and penetrates on the average one diffusion length L before recombining with electrons.

Under equilibrium conditions a *principle of detailed balance* holds. This principle of statistical mechanics says that each process and its opposite occur equally frequently. Hence we must conclude that the flow of holes from the p-region into the n-region, followed by recombination, must be exactly balanced by a reverse process. The reverse process is thermal generation of holes through deathnium centers, followed by diffusion to barrier where they slide down into p-type region.

The application of voltage to the terminals of the device shown in Fig. 5 destroys the exact balance of the two currents just discussed. In considering the application of voltage we shall neglect any voltage drops at the contacts between the metal electrodes of Fig. 5 and the semiconductors. At the end of this section we will return briefly to the reasons why such voltage drops may be neglected. The effect of the application of voltages upon the currents is represented in Fig. 8. In part *(a)* of this figure we show the thermal equilibrium condition. The two currents previously discussed are represented by I_f and I_g, these currents standing, respectively, for the current of holes entering the n-region and recombining and the current generated in the n-region and diffusing to the barrier[21]. For the condition of thermal equilibrium these two currents are equal and opposite. In part *(b)* of the figure the situation for a large « reverse » bias is shown. For reverse bias, negative voltage is applied to the p-region and positive to the n-region so that the electro-

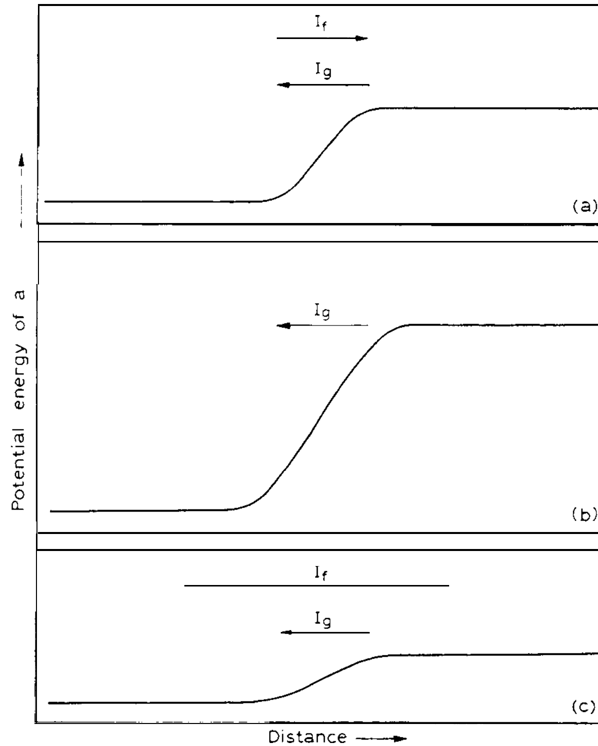

Fig. 8. Dependence of recombination and generation currents upon bias. (*a*) Thermal equilibrium; (*b*) reverse bias; (*c*) forward bias.

static potential difference between the two regions is increased. If the electrostatic potential is sufficiently high, corresponding to the situation shown in part *(b),* then practically no holes can climb the potential hill and I_f drops nearly to zero. This situation is represented by showing I_f as a vector of negligible length, whereas I_g has practically the same value as it has for the case of thermal equilibrium. In general, the diffusion length L is large compared to the width of the dipole or space-charge region. Hence the region where I_g, arises is practically unaffected by reverse bias and I_g is thus independent of reverse bias. This independence of current upon bias is referred to as *saturation.*

When forward bias is applied, the situation shown in Fig. 8 (*c*) occurs and I_f increases. This increase is described in terms of the energy difference for a hole in the n-region and p-region. This energy difference is equal to the charge of the electron times the electrostatic potential differences be-

tween the two sides. We can apply a general theorem from statistical mechanics to a consideration of the number of holes which, by chance, acquire sufficient energy to climb the potential hill. This theorem states that each time the potential hill is increased by one thermal unit of energy, *kT,* then the number of holes capable of climbing the higher hill is reduced by a factor of I/e. Since the potential barrier is already present under conditions of thermal equilibrium, it follows also that each lowering of the barrier by an amount *kT* will increase the current by a factor of e. The change in height of the barrier caused by the applied voltage V is $-qV$, where the polarity is so chosen that positive values correspond to plus potentials applied to the p-region and q is the absolute value of the charge of the electron. $V = 0$ is the equilibrium case, and for this case I_f is equal to I_g. Hence, in general, the recombination current is

$$I_f = I_g \exp qV/kT$$

This gives rise to a total current of holes from p-region to n-region, given by the difference

$$I_f - I_g = I_g [\exp (qV/kT) - I]$$

This current is zero when $V = 0$, increases exponentially to large values for positive V, and decreases to a negative saturation value of I_g when V is negative and substantially larger than kT/q.

Similar reasoning can be applied to the electron current flowing across the junction. The applied potential which lowers the potential barrier for holes, evidently lowers it also for electrons; consequently, large electron currents flow under the same voltage conditions that produce large hole currents. In both cases these large currents correspond to flows of minority carriers into each region. In both cases the current is in the sense of transferring positive charge from the p-region to the n-region. In one case this is carried in the form of positive imperfections, namely holes, moving from p to n, and in the other case it is due to negative imperfections, electrons, moving in the opposite direction. For the case of reverse biases the potential rise is larger and the holes tend to be retained in the p-region and the electrons in the n-region. A saturation current due to generation in both regions flows. If the total saturation current is called I_s, then the total current for any applied voltage V is given by the formula

$$I = [\exp (qV/kT) - I] \quad I_s$$

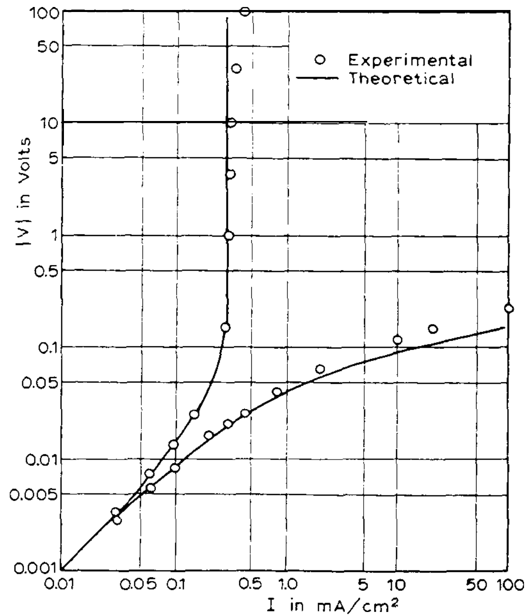

Fig. 9. Rectification characteristic for a p-n junction.

Evidently, I_s, is the sum of the two generation currents. This equation is found to be well-satisfied for p-n junctions in germanium, and a comparison of the rectification curve[22] as measured with the theoretical formula is given in Fig. 9. It should be noted that the separation between the forward and reverse branches of the curves corresponds to a factor of e when the voltage is kT/q = 25 mV. This is exactly the factor predicted by the preceding equation. This agreement between theory and experiment is evidence that the imperfections which carry the current in a p-n junction have the charge of the electron. If they had twice this charge, a value of 12.5 mV should be obtained; if half the charge, the value should be 50 mV.

For large forward biases, the potential barrier between the n- and p-type is nearly eliminated. Under these conditions large concentrations of minority carriers flow across the barrier and density of the majority carriers may be substantially disturbed. Under these conditions it is no longer valid to consider that the minority carriers diffuse in a field-free region, as is the case when nearly equilibrium carrier densities exist. Although these large signal conditions are of general interest, we shall not consider them further here.

There are a number of ways in which the diffusion theory for rectification in the p-n junction can be tested experimentally. We shall consider a test

based upon a photoelectric effect. Photons of sufficient energy can generate hole-electron pairs when they are absorbed in germanium. This generation adds to the thermal generation produced in the deathnium centers. If the light is focused on the specimen in the form of a small spot, then it is possible to generate minority carriers in either region at a precise distance from the junction. The current flowing across the junction as a result of generation by the light should then decrease with increasing distance of the light from the junction. It can be shown that the probability that a minority carrier generated at a distance x from the junction diffuses to junction before recombining is simply

$$\exp\,(-x/L)$$

where L is the diffusion length of the minority carrier in the region where it is generated. This exponential dependence of response upon distance from the junction has been verified directly by Goucher and his colleagues[22]. They have found also that the value of L determined by studies of this sort is consistent with that necessary to explain value of I_s in the rectification formula.

For purposes of illustration we shall consider the value of the diffusion length for a typical example. A typical lifetime for a minority carrier is 10^{-4} sec and the diffusion constant for electrons in germanium is 93 cm^2/sec. These lead to a diffusion length of about 0.1 cm or 1 mm.

Since the light acts as a current generator, its effect may be readily included in an equivalent circuit for a p-n junction. This is illustrated in Fig. 10. Here I_l is the current of minority carriers generated by the light. The equivalent circuit shown corresponds to a case in which the light is focused at a distance x from the junction. If the light is distributed, the appropriate average of the exponential probability factor should be used. If the light falls on both the n-type and the p-type region, this average should take into account the fact that the diffusion length for holes in the n-region is probably different from that for electrons in the p-region. Fig. 10 emphasizes the importance of considering the light as acting as a current generator. If the equivalent circuit of Fig. 10 is operated in an open-circuit condition, then a photovoltage will be developed. This photovoltage will in general be nonlinear in the light intensity because of the rectification characteristics of the rectifier. In the equivalent circuit of Fig. 10, however, the current generator is linear in the light intensity. This independence of the current generator of biases applied to the p-n junction has been verified over a wide range of ex-

perimental conditions by Goucher[22] and his colleagues, and by Pietenpol[23].

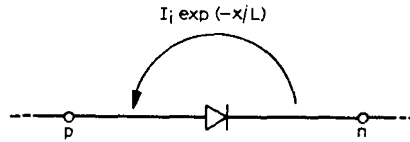

Fig. 10. Equivalent circuit for a p-n junction photocell.

The junction transistor[18,24,25]

The junction transistor is a compositional structure composed of two p-n junctions placed back to back. In its general behaviour as an amplifying device, a junction transistor shows great similarities to a vacuum-tube triode, or thermionic valve. Fig. 11 shows an n-p-n junction transistor in an amplifying circuit, the transistor being in the form of a sandwich with a layer of p-type germanium interposed between two layers of n-type germanium. Non-rectifying electrical contacts are made to the three layers. Under operating conditions the n-type region on the right, known as the collector, is biased positive so as to become attractive to electrons. As a result, a reverse bias appears between it and the middle region, known as the base. The current flowing across this reverse-biased junction can be controlled by an input signal applied between the base layer and the n-type region to the left, known as the emitter. As I shall describe below in more detail, the bias across the emitter junction controls the electron flow into the base region. In effect, the emitter junction acts like the region between the cathode and the grid in the vacuum tube. Electrons which enter the base region have a high probability of diffusing to the collector junction, and thus the flow of electrons from emitter to collector can be varied by varying the potential across the emitter base junction. The action is very similar to that controlling the flow of electrons from the cathode to the anode in the thermionic triode.

Junction transistors can be fabricated in a variety of ways. The compositional structure can be produced in a crystal-growing machine by techniques like those used for making simple p-n junctions. As the crystal is grown from a melt containing antimony, a pellet containing indium is dropped into the melt, and a second pellet containing antimony is dropped in a few seconds later. The portion of the crystal which grows between the dropping of the two pellets is rich in indium and is consequently p-type. The second pellet

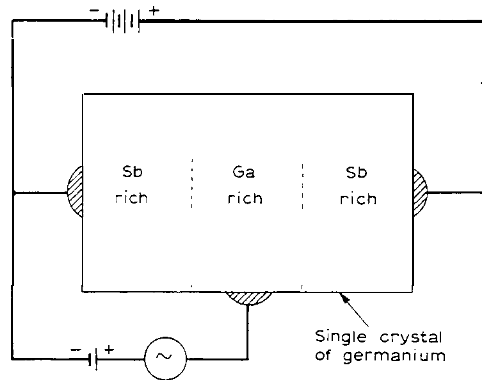

Fig. 11. The structure of a junction transistor and the bias supply for its operation in an amplifying circuit.

over-compensates the effect of the added indium and the subsequent material is again n-type. From such a single crystal small rods may be cut and contacts made. There are a number of technical processes involved in proceeding from the point of growing the crystal to the production of a packaged, stable transistor. I shall not endeavor to discuss these in the lecture.

An alternative technique for producing the compositional structure[25a] starts with a thin plate of germanium which subsequently plays the role of the p-type region. A pellet of metal containing a donor is placed on this plate. The plate and pellet are then raised to such a temperature that the metal melts and dissolves a small amount of germanium. When the metal and germanium are subsequently cooled, the germanium precipitates from the metal and grows back onto the crystal structure of the base material. This regrown germanium carries with it some of the donors contained in the molten metal and thus grows an n-type region. In the fabrication of a transistor, pellets are placed on both sides of a thin plate and allowances are made for the degree to which they dissolve germanium on the two sides. This process has been used in the production of a large fraction, if not the majority, of transistors made to date.

From the point of view of an electron the situation in an operating transistor is as represented in Fig. 12. This diagram shows the variation in potential energy for an electron along a line going from emitter to collector in a transistor biased like that shown in Fig. 11. The reverse bias at the collector junction produces a large drop in potential on the right-hand side. The varying bias across the emitter junction changes the height of the hill and thus

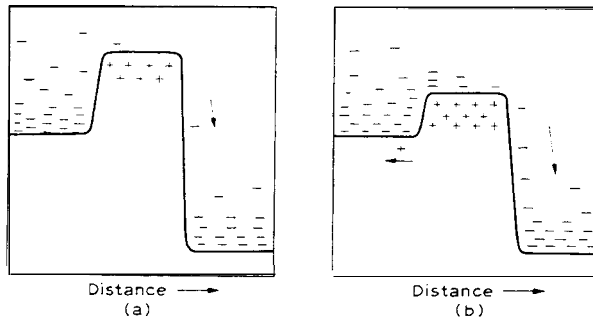

Fig. 12. The potential energy of an electron in an n-p-n junction transistor for two values of bias across the emitter junction. In *(b)* the forward bias is greater than in *(a).*

varies the diffusion current of electrons into the base. The base layer is very thin and contains very little deathnium. Consequently, the probability is much higher for an electron to diffuse through the base layer and arrive at the collector junction than for it to combine with a hole through a deathnium center in the base layer. In a well-built junction transistor, in fact, the electron flow through the base region proceeds so efficiently that the electron current flowing through the base layer to the collector may exceed the current combining in the base layer by a factor of 100 or more. This means that the input currents flowing to the base layer through the base lead may control the output currents 100 times larger flowing to the collector region. Although this situation is not nearly as ideal as in a vacuum triode, in which the grid current is still smaller compared with the anode current, it permits the junction transistor to operate as a highly efficient amplifier.

So far as voltage requirements are concerned, the junction transistor is far superior to any vacuum tube, because a junction transistor can be brought fully into its operational range with voltages as small as 50-100 mV. Such voltages make the potential hill at the collector junction shown on Fig. 12 several times the thermal voltage (kT/q = 25 mV). As a consequence, any electron reaching the collector junction is sure to slide down the hill and has a negligible chance of returning.

The junction transistor has almost ideal pentode characteristics in terms of collector saturation. This also follows from the same considerations which make it operate well at low collector voltages. In fact, in some junction transistors there is a negligible variation in current over a voltage range of a hundredfold or more, say from 100 mV to 10 volts on the collector.

The junction transistor described in Figs. 11 and 12 is only one of a

large family of transistors. The p-n-i-p and n-p-i-n transistors are variations of this form especially designed for low capacitance between the collector and the base[25b]. The so-called junction tetrode is a special form of junction transistor in which the current flow is controlled so as to occur only over a small region of the base[25c]. Junction transistors of these types have been used in oscillating circuits at frequencies as high as 1000 Mc/s.

In closing this section, I should like to point out that the junction transistor operates more flexibly in terms of power than do vacuum tubes. This is of great importance, since it means that junction transistors may be used efficiently in some cases where the power level to be amplified is very much smaller than the heater current required in the most efficient of vacuum tubes. The possibility of operating at very low power is due to the low voltages necessary to operate a junction transistor, and because of its small size and the high degree of purity possible, the current due to generation by deathnium centers can be made smaller than a microampere. It is thus possible to make a junction transistor which can be put into an amplifying condition with a power input substantially less than a microwatt. The same transistor, however, can operate at voltages as high as 10 volts and currents of the order of 10 mA. Thus, it can cover a power range effectively of 100,000. This fact indicates that there will probably be a much smaller diversity of transistor types than there are of vacuum tube types.

Conclusion

In a lecture of reasonable scope, it is no longer possible to do justice to the many branches of semiconductor physics that have developed since the discovery of the transistor effect. In the previous sections I have tried to illustrate how the stimulation of aiming for practical goals led to an experimental program which brought out many new and important aspects of electronic phenomena in semiconductors.

Not all of the important work has, of course, been motivated by highly practical ends which could be so immediately achieved once the fundamental science was understood. Much of the work on surface states[8,26] should now be regarded as being in the pure-science class inasmuch as it is not possible to see at all directly how the scientific results will find utility. However, it seems highly probable that once the phenomena of surface states are thoroughly understood from a scientific point of view, many useful sug-

gestions will arise as to how this knowledge may be employed to make better devices.

Also in the class of purely scientific work are the very beautiful determinations of the energy band structure in silicon and germanium that have resulted from *cyclotron resonance* in silicon and germanium[27,28,29]. This work has led to a more detailed knowledge of energy band surfaces than has ever existed before, except possibly for some of the most simple and almost trivial cases.

One of the large new areas that has begun to grow as a result of the large amount of transistor technology may be called « The Theory of Solid Solutions in Silicon and Germanium ». Recently, interactions between donors and acceptors of various types have been given active consideration[30]. It is evident that interesting phenomena are associated with the incorporation of any foreign element from the Periodic Table in silicon and germanium. Important distinctions are being drawn between the same atoms in a substitutional position and an interstitial position. It is also being found that some atoms may exist in a semiconductor in not merely two states of charge as in the case of a simple donor and acceptor but in as many as four different conditions of charge[31]. It seems highly probable that this area of work will become a field of active interest in physical chemistry in the years to come.

Among the phenomena motivated in considerable measure by practical aims we should mention the phenomena of avalanche breakdown which is useful in voltage limiting diodes and protection devices. A strong motivation for following up early leads which suggested that such effects might occur, came from the conversation I once had with Dr. O. E. Buckley, whom I have mentioned before in this lecture. Dr. Buckley pointed out the need for a device for protecting telephones from damage due to voltages induced by lightning near telephone lines. It was in large measure my knowledge of this need that gave emphasis to this work. The original interpretations[32] which appeared to be consistent with the dielectric breakdown mechanism once proposed by Zener[33] (direct rupture of electron-pair bonds by strong electric fields with the resultant production of hole-electron pairs) seems now to have been in error. The investigation of the same phenomena by McKay[34] and his collaborators has led to a new branch of semiconductor electronics dealing with impact ionization of holes in electrons moving in intense electric fields in semiconductors. Both the effect and its theory are playing an active role in a certain class of switching transistors.

In closing this lecture, I would like to refer to a paragraph written in my

book in 1950[35]. I am pleased to see that the predictions of the paragraph appear to have been borne out to a considerable extent and I feel that it is now as appicable as it was then:

« It may be appropriate to speculate at this point about the future of transistor electronics. Those who have worked intensively in the field share the author's feeling of great optimism regarding the ultimate potentialities. It appears to most of the workers that an area has been opened up comparable to the entire area of vacuum and gas-discharge electronics. Already several transistor structures have been developed and many others have been explored to the extent of demonstrating their ultimate practicality, and still other ideas have been produced which have yet to be subjected to adequate experimental tests. It seems likely that many inventions unforeseen at present will be made based on the principles of carrier injection, the field effect, the Suhl effect, and the properties of rectifying junctions. It is quite probable that other new physical principles will also be utilized to practical ends as the art develops. »

It is my hope to contribute to the fulfilment of the predictions of this paragraph through my new organization in California.

1. J. Bardeen, *Nobel Lecture,* this volume, p. 318.

1a. J. H. Scaff, H. C. Theurer, and E. E. Schumacher, Trans. *AIME,* 185 (1949) 383.

2. W. Shockley, *Electrons and Holes in Semiconductors,* D. Van Nostrand Co., Inc., New York, 1950, Ch. 12.

3. K. Lark-Horovitz, *Semi-Conducting Materials,* Butterworth Scientific Publications, London, 1951, pp. 47-48.

4. J. A. Burton, G. W. Hull, F. J. Morin, and J. C. Severiens, *J. Phys. Chem.,* 57 (1953) 853-859.

5. The theory of this process has been developed by W. Shockley and W. T. Read, Jr., in *Phys. Rev.,* 87 (1952) 835-842. Experimental findings in agreement with the theory have been obtained by Burton, Hull, Morin, and Severiens (Ref. 4) and by R. N. Hall, *Phys. Rev.,* 83 (1951) 228; *Ibid.,* 87 (1952) 387.

6. J. Bardeen, *Phys. Rev.,* 71 (1947) 717-727.

7. W. Shockley and G. L. Pearson, *Phys. Rev., 74 (1948) 232-233.*

8. W. H. Brattain, *Nobel Lecture,* this volume, p. 377.

9. J. N. Shive, *Phys. Rev., 75 (1949) 689-690.*

10. J. R. Haynes and W. Shockley, *Phys. Rev., 75 (1949) 691.*

11. J. R. Haynes and W. Shockley, *Phys. Rev.,* 81 (1951) 835-843.
J. R. Haynes and W. Westphal, *Phys. Rev., 85 (1952) 680.*

12. Transistor Teacher's Summer School, *Phys. Rev., 88 (1952) 1368-1369.*

13. H. Krömer, *Z. Physik, 134 (1953) 435-450.*
14. W. Shockley, Bell *System Tech. J., 33 (1954) 799-526.*
15. E. J. Ryder and W. Shockley, *Phys. Rev.,* 81 (1951) 139-140.
16. E. J. Ryder, *Phys. Rev., 90 (1953) 767-769.*
17. W. Shockley, *Bell System Tech. J., 30 (1951) 990-1034.*
18. W. Shockley, *Bell System Tech. J., 28 (1949) 435.*
19. G. K. Teal and J. B. Little, *Phys. Rev., 78 (1950) 647.*
20. W. G. Pfann, *Trans. AIME, 194 (1952) 747;* W. G. Pfann and K. M. Olsen, *Phys. Rev., 89 (1953) 322; Bell Lab. Record, 33 (1955)* 201.
21. The subscripts *f* and *g* may be thought of as *forward* and *generation.* This mixed choice avoids subscript *r,* which might be either *reverse* or *recombination.* In this decision, *forward* is equivalent to *recombination* and *generation* to *reverse.*
22. F. S. Goucher *et al., Phys. Rev.,* 81 (1951) 637.
23. W. J. Pietenpol, *Phys. Rev., 82 (1951) 120.*
24. W. Shockley, M. Sparks, and G. K. Teal, *Phys. Rev., 83 (1951)* 151.
25. W. J. Pietenpol and R. L. Wallace, *Proc. I.R.E., 39 (1951) 753.*
25a. R. N. Hall and W. C. Dunlap, *Phys.* Rev., 80 (1950) 467.
25b. J. M. Early, *Bell System Tech. J., 33 (1954)* 517.
25c. R. L. Wallace, L. G. Schimpf, and E. Dickten, *Proc. I.R.E., 40* (1952) 1395.
26. For a recent review and references, see R. H. Kingston (Ed.), *Semiconductor Surface Physics,* University of Pennsylvania Press, Philadelphia, 1957.
27. W. Shockley, *Phys. Rev., 90 (1953) 491.*
28. G. Dresselhaus, A. F. Kip, and C. Kittel, *Phys. Rev., 92* (1953) 827. R. N. Dexter, B. Lax, A. F. Kip, and G. Dresselhaus, *Ibid., 96 (1954)* 222.
29. R. N. Dexter, H. J. Zeiger, and B. Lax, *Phys. Rev.,* 104 (1956) 637-644 (references) .
30. H. Reiss, C. S. Fuller, and F. J. Morin, *Bell System Tech. J.,* 35 (1956) 535-636.
31. H. H. Woodbury and W. W. Tyler, *Phys. Rev., 105 (1957) 84.*
32. K. B. McAfee, E. J. Ryder, W. Shockley, and M. Sparks, *Phys. Rev., 83* (1951) 650.
33. C. Zener, *Proc. Roy. Soc. London,* A 145 (1934) 523.
34. K. G. McKay, *Phys. Rev., 94* (1954) 877.
35. W. Shockley, *Electrons and Holes in Semiconductors,* D. Van Nostrand C O ., Inc., New York, 1950, p. 349.