JOHN BARDEEN

# Semiconductor research leading to the point contact transistor

*Nobel Lecture, December 11, 1956*

## *Introduction*

In this lecture we shall attempt to describe the ideas and experiments which led to the discovery of the transistor effect as embodied in the point-contact transistor. Some of the important research done subsequent to the discovery will be described in the following lectures by Shockley and Brattain. As we shall see, the discovery was but a step along the road of semiconductor research to which a great many people in different countries have contributed. It was dependent both on the sound theoretical foundation largely built up during the thirties and on improvement and purification of materials, particularly of germanium and silicon, in the forties. About half of the lecture will be devoted to an outline of concepts concerning electrical conduction in semiconductors and rectification at metal-semiconductor contacts as they were known at the start of our research program.

The discovery of the transistor effect occurred in the course of a fundamental research program on semiconductors initiated at the Bell Telephone Laboratories in early 1946. Semiconductors was one of several areas selected under a broad program of solid-state research, of which S. O. Morgan and W. Shockley were co-heads. In the initial semiconductor group, under the general direction of Shockley, were W. H. Brattain, concerned mainly with surface properties and rectification, G. L. Pearson, concerned with bulk properties, and the writer, interested in theoretical aspects of both. Later a physical chemist, R. B. Gibney, and a circuit expert, H. R. Moore, joined the group and made important contributions, particularly to chemical and instrumentation problems, respectively.

It is interesting to note that although Brattain and Pearson had had considerable experience in the field prior to the war, none of us had worked on semiconductors during the war years. We were able to take advantage of the important advances made in that period in connection with the development of silicon and germanium detectors and at the same time have a

fresh look at the problems. Considerable help was obtained from other groups in the Laboratories which were concerned more directly with wartime developments. Particular mention should be made of J. H. Scaff, H. C. Theuerer and R. S. Ohl.

The general aim of the program was to obtain as complete an understanding as possible of semiconductor phenomena, not in empirical terms, but on the basis of atomic theory. A sound theoretical foundation was available from work done during the thirties:

(1) Wilson's quantum mechanical theory[1], based on the energy band model, and describing conduction in terms of excess electrons and holes. It is fundamental to all subsequent developments. The theory shows how the concentration of carriers depends on the temperature and on impurities.

(2) Frenkel's theories of certain photoconductive phenomena[2] (change of contact potential with illumination and the photomagneto electric effect) in which general equations were introduced which describe current flow when non-equilibrium concentrations of both holes and conduction electrons are present. He recognized that flow may occur by diffusion in a concentration gradient as well as by an electric field.

(3) Independent and parallel developments of theories of contact rectification by Mott[3], Schottky[4] and Davydov[5]. The most complete mathematical theories were worked out by Schottky and his co-worker, Spenke.

Of great importance for our research program was the development during and since the war of methods of purification and control of the electrical properties of germanium and silicon. These materials were chosen for most of our work because they are well-suited to fundamental investigations with the desired close coordination of theory and experiment. Depending on the nature of the chemical impurities present, they can be made to conduct by either excess electrons or holes.

Largely because of commercial importance in rectifiers, most experimental work in the thirties was done on copper oxide ($Cu_2O$) and selenium. Both have complex structures and conductivities which are difficult to control. While the theories provided a good qualitative understanding of many semiconductor phenomena, they had not been subjected to really convincing quantitative checks. In some cases, particularly in rectification, discrepancies between experiment and theory were quite large. It was not certain whether the difficulties were caused by something missing in the theories or by the fact that the materials used to check the theories were far from ideal.

In the U.S.A., research on germanium and silicon was carried out during

the war by a number of university, government and industrial laboratories in connection with the development of point-contact or « cat's whisker » detectors for radar. Particular mention should be made of the study of germanium by a group at Purdue University working under the direction of K. Lark-Horovitz and of silicon by a group at the Bell Telephone Laboratories. The latter study was initiated by R. S. Ohl before the war and carried out subsequently by him and by a group under J. H. Scaff. By 1946 it was possible to produce relatively pure polycrystalline materials and to control the electrical properties by introducing appropriate amounts of donor and acceptor impurities. Some of the earliest work (1915) on the electrical properties of germanium and silicon was done in Sweden by Prof. C. Benedicks.

Aside from intrinsic scientific interest, an important reason for choosing semiconductors as a promising field in which to work, was the many and increasing applications in electronic devices, which, in 1945, included diodes, varistors and thermistors. There had long been the hope of making a triode, or an amplifying device with a semiconductor. Two possibilities had been suggested. One followed from the analogy between a metal semiconductor rectifying contact and a vacuum-tube diode. If one could somehow insert a grid in the space-charge layer at the contact, one should be able to control the flow of electrons across the contact. A major practical difficulty is that the width of the space-charge layer is typically only about $10^{-4}$ cm. That the principle is a sound one was demonstrated by Hilsch and Pohl[6], who built a triode in an alkali-halide crystal in which the width of the space-charge layer was of the order of one centimeter. Because amplification was limited to frequencies of less than one cycle per second, it was not practical for electronic applications.

The second suggestion was to control the conductance of a thin film or slab of semiconductor by application of a transverse electric field (called the *field effect).* In a simple form, the slab forms one plate of a parallel plate condenser, the control electrode being the other plate. When a voltage is applied across the condenser, charges are induced in the slab. If the induced charges are mobile carriers, the conductance should change with changes of voltage on the control electrode. This form was suggested by Shockley; his calculations indicated that, with suitable geometry and materials, the effect should be large enough to produce amplification of an a.c. signal[7].

Point-contact and junction transistors operate on a different principle than either of these two suggestions, one not anticipated at the start of the program. The transistor principle, in which both electrons and holes play a role,

was discovered in the course of a basic research program on surface properties.

Shockley's field-effect proposal, although initially unsuccessful, had an important bearing on directing the research program toward a study of surface phenomena and surface states. Several tests which Shockley carried out at various times with J. R. Haynes, H. J. McSkimin, W. A. Yager and R. S. Ohl, using evaporated films of germanium and silicon, all gave negative results. In analyzing the reasons for this failure, it was suggested[8] that there were states for electrons localized at the surface, and that a large fraction of the induced charge was immobilized in these states. Surface states also accounted for a number of hitherto puzzling features of germanium and silicon point-contact diodes.

In addition to the possibility of practical applications, research on surface properties appeared quite promising from the viewpoint of fundamental science. Although surface states had been predicted as a theoretical possibility, little was known about them from experiment. The decision was made, therefore, to stress research in this area. The study of surfaces initiated at that time (1946) has been continued at the Bell Laboratories and is now being carried out by many other groups as well[9].

It is interesting to note that the field effect, originally suggested for possible value for a device, has been an extremely fruitful tool for the fundamental investigation of surface states. Further, with improvements in semiconductor technology, it is now possible to make electronic amplifiers with high gain which operate on the field-effect principle.

Before discussing the research program, we shall give first some general background material on conduction in semiconductors and metal-semiconductor rectifying contacts.

### *Nature of conduction in semiconductors*

An electronic semiconductor is typically a valence crystal whose conductivity depends markedly on temperature and on the presence of minute amounts of foreign impurities. The ideal crystal at the absolute zero is an insulator. When the valence bonds are completely occupied and there are no extra electrons in the crystal, there is no possibility for current to flow. Charges can be transferred only when imperfections are present in the electronic structure, and these can be of two types: *excess electrons* which do not

fit into the valence bonds and can move through the crystal, and *holes,* places from which electrons are missing in the bonds, which also behave as mobile carriers. While the excess electrons have the normal negative electronic charge *-e,* holes have a positive charge, *+e.* It is a case of two negatives making a positive ; a missing negative charge is a positive defect in the electron structure.

The bulk of a semiconductor is electrically neutral; there are as many positive charges as negative. In an intrinsic semiconductor, in which current carriers are created by thermal excitation, there are approximately equal numbers of excess electrons and holes. Conductivity in an *extrinsic* semiconductor results from impurity ions in the lattice. In n-type material, the negative charge of the excess electrons is balanced by a net positive space charge of impurity ions. In p-type, the *positive* charge of the holes is balanced by negatively charged impurities. Foreign atoms which can become positively charged on introduction to the lattice are called *donors;* atoms which become negatively ionized are called *acceptors.* Thus donors make a semiconductor n-type, acceptors p-type. When both donors and acceptors are present, the conductivity type depends on which is in excess. Mobile carriers then balance the *net* space charge of the impurity ions. Terminology used is listed in the table below:

**Table I.**

| *Designation of conductivity type* | | *Majority carrier* | *Dominant impurity ion* |
|---|---|---|---|
| n-type | excess | electron ($n/cm^3$) | donor |
| p-type | defect | hole ($p/cm^3$) | acceptor |

These ideas can be illustrated quite simply for silicon and germanium, which, like carbon, have a valence of four and lie below carbon in the Periodic Table. Both crystallize in the diamond structure in which each atom is surrounded tetrahedrally by four others with which it forms bonds. Carbon in the form of a diamond is normally an insulator; the bond structure is complete and there are no excess electrons. If ultraviolet light falls on diamond, electrons can be ejected from the bond positions by the photoelectric effect. Excess electrons and holes so formed can conduct electricity; the crystal becomes photoconductive.

The energy required to free an electron from a bond position so that it and the hole left behind can move the crystal, is much less in silicon and germanium than for diamond. Appreciable numbers are released by thermal excitations at high temperatures; this gives intrinsic conductivity.

Impurity atoms in germanium and silicon with more than four valence electrons are usually donors, those with less than four acceptors. For example, Group V elements are donors, Group III elements acceptors. When an arsenic atom, a Group V element, substitutes for germanium in the crystal, only four of its valence electrons are required to form the bonds. The fifth is only weakly held by the forces of Coulomb attraction, greatly reduced by the high dielectric constant of the crystal. The energy required to free the extra electron is so small that the arsenic atoms are completely ionized at room temperature. Gallium, a typical Group III acceptor, has only three valence electrons. In order to fill the four bonds, Ga picks up another electron and enters the crystal in the form of a negative ion, $Ga^-$. The charge is balanced by a free hole.

While some of the general notions of excess and defect conductivity, donors and acceptors, go back earlier, Wilson[1] was the first to formalize an adequate mathematical theory in terms of the band picture of solids. The band picture itself, first applied to metals, is a consequence of an application of quantum mechanics to the motion of electrons in the periodic potential field of a crystal lattice. Energy levels of electrons in atoms are discrete. When the atoms are combined to form a crystal, the allowed levels form continuous bands. When a band is completely occupied, the net current of all of the electrons in the band is zero. Metals have incompletely filled bands. In insulators and semiconductors, there is an energy gap between the highest filled band and the next higher allowed band of levels, normally unoccupied.

The relations are most simply illustrated in terms of an energy-level diagram of the crystal. In Fig. 1 is shown a schematic energy-level diagram of an intrinsic semiconductor. Electrons taking part in the chemical bonds form a continuous band of levels called the valence band. Above these is an energy gap in which there are no allowed levels in the ideal crystal, and then another continuous band of levels called the conduction band. The energy gap, $E_G$, is the energy required to free an electron from the valence bonds. Excess, or conduction, electrons have energies in the lower part of the conduction band. The very lowest state in this band, $E_C$, corresponds to an electron at rest, the higher states to electrons moving through the crystal with

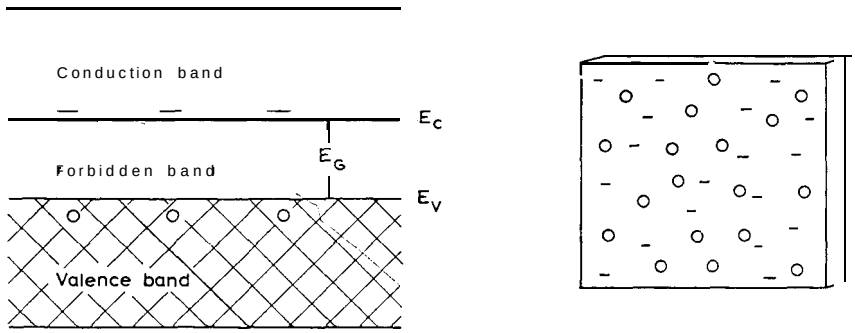


Fig. 1. Energy-level diagram of an intrinsic semiconductor. There is a random distribution of electrons and holes in equal numbers.

additional energy of motion. Holes correspond to states near the top of the valence band, $E_V$, from which electrons are missing. In an intrinsic semiconductor, electrons and holes are created in equal numbers by thermal excitation of electrons from the valence to the conduction band, and they are distributed at random through the crystal.

In an n-type semiconductor, as illustrated in Fig. 2a, there is a large number of electrons in the conduction band and very few holes in the valence band. Energy levels corresponding to electrons localized around Group V donor impurity atoms are typically in the forbidden gap and a little below the conduction band. This means that only a small energy is required to ionize the donor and place the electron removed in the conduction band. The charge of the electrons in the conduction band is compensated by the positive space charge of the donor ions. Levels of Group III acceptors (Fig. 2b) are a little above the valence band. When occupied by thermal excitation of electrons from the valence band, they become negatively charged. The space charge of the holes so created is compensated by that of the negative acceptor ions.

Occupancy of the levels is given by the position of the Fermi level, $E_F$. The probability, $f$, that a level of energy $E$ is occupied by an electron is given by the Fermi-Dirac function:

$$f = \frac{1}{1 + \exp (E - E_F)/kT}$$

The energy gap in a semiconductor is usually large compared with thermal energy, $kT$ (~ 0.025 eV at room temperature), so that for levels well above $E_F$ one can use the approximation

$$f \simeq \exp\,[-\,(E - E_F)/kT]$$

For levels below $E_F$, it is often more convenient to give the probability

$$f_p = 1 - f = \frac{1}{1 + \exp\,[(E_F - E)/kT]}$$

that the level is unoccupied, or « occupied by a hole ». Again, for levels well below $E_F$,

$$f_p \simeq \exp\,[-\,(E_F - E)/kT]$$

The expressions for the total electron and hole concentrations (number per unit volume), designated by the symbols $n$ and $p$ respectively, are of the form

$$n = N_C \exp\,[-\,(E_C - E_F)/kT]$$

$$p = N_V \exp\,[-\,(E_F - E_V)/kT]$$

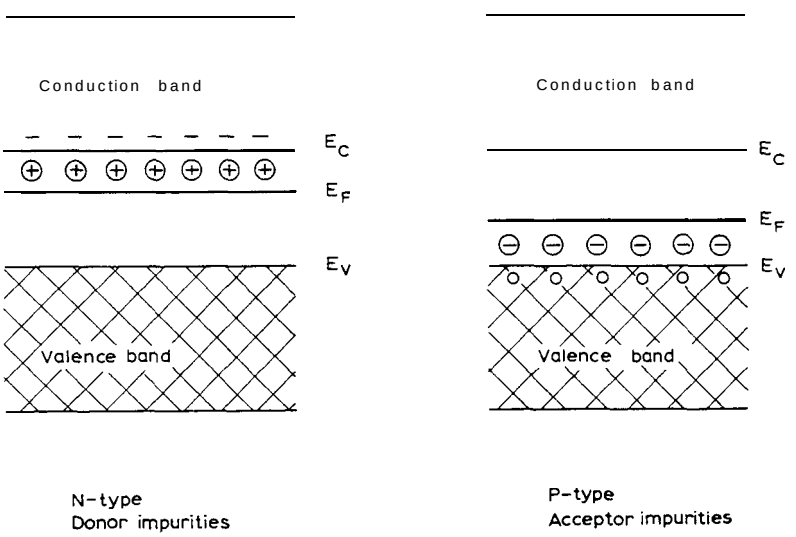


Fig. 2. Energy-level diagrams for n- and p-type semiconductors.

where $N_C$ and $N_V$ vary slowly with temperature compared with the exponential factors. Note that the product $np$ is independent of the position of the Fermi level and depends only on the temperature:

$$np = n_i^2 = N_C N_V \exp\left[-(E_C - E_V)/kT\right] = N_C N_V \exp\left[-E_G/kT\right]$$

Here $n$ is the concentration in an intrinsic semiconductor for which $n = p$.

In an n-type semiconductor, the Fermi level is above the middle of the gap, so that $n \gg p$. The value of $n$ is fixed by the concentration of donor ions, $N_d^+$, so that there is electrical neutrality:

$$n - p = N_d^+$$

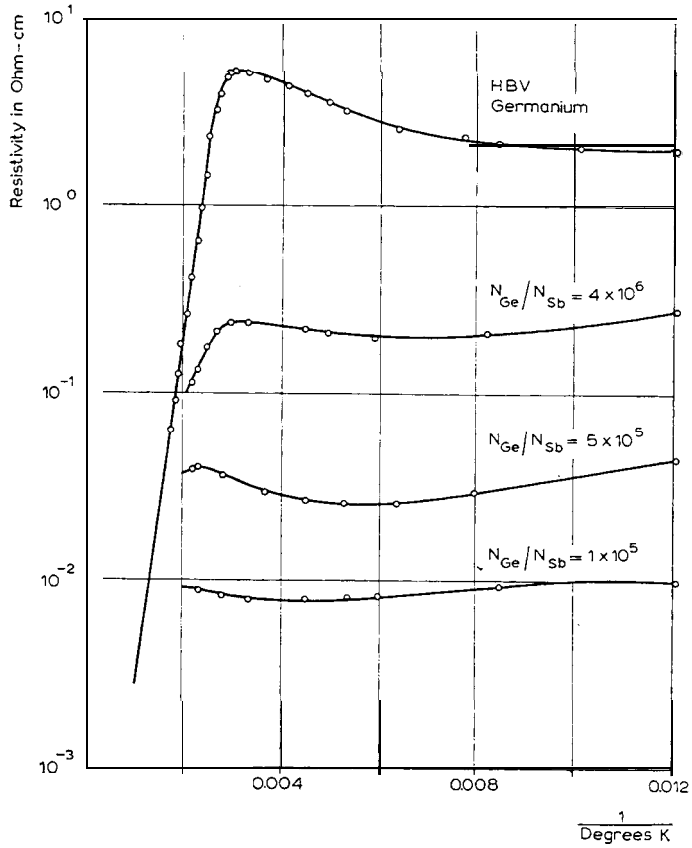


Fig. 3. Conductivity *vs.* $1/T$ for germanium with antimony added as a donor impurity.

The minority carrier concentration, $p$, increases rapidly with temperature and eventually a temperature will be reached above which $n$ and $p$ are both large compared with $Nd+$ and the conduction is essentially intrinsic. Correspondingly in a p-type semiconductor, in which there are acceptor ions, $p \gg n$, and the Fermi level is below the center of the gap.

The Fermi level is equivalent to the chemical potential of the electrons. If two conductors are electrically connected together so that electrons can be transferred, the relative electrostatic potentials will be adjusted so that the Fermi levels of the two are the same. If the n- and p-type materials of Fig. 2 are connected, a small number of electrons will be transferred from the n-type to the p-type. This will charge the p-type negatively with respect to the n-type and raise the electrostatic potential energy of the electrons accordingly. Electron transfer will take place until the energy levels of the p-type material are raised relative to those of the n-type by the amount required to make the Fermi levels coincide.

The amount of impurity required to make significant changes in the conductivity of germanium or silicon is very small. There is given in Fig. 3 a plot, on a log scale, of the resistivity *vs.* 1/T for specimens of germanium with varying amounts of antimony, a donor impurity. This plot is based on some measurements made by Pearson several years ago[11]. The purest specimens available at that time had a room temperature resistivity of about 10-20 ohm cm, corresponding to about one donor atom in $10^8$ germanium atoms. This material (H.B.V.) is of the sort which was used to make germanium diodes which withstand a high voltage in the reverse direction (High Back Voltage) and also used in the first transistors. The purest material available now corresponds to about one donor or acceptor in $10^{10}$. The resistivity drops, as illustrated, with increasing antimony concentration; as little as one part in $10^7$ makes a big difference. All specimens approach the intrinsic line corresponding to pure germanium at high temperatures.

Conduction electrons and holes are highly mobile, and may move through the crystal for distances of hundreds or thousands of the interatomic distance, before being scattered by thermal motion or by impurities or other imperfections. This is to be understood in terms of the wave property of the electron; a wave can travel through a perfect periodic structure without attenuation. In treating acceleration in electric or magnetic fields, the wave aspect can often be disregarded, and electrons and holes thought of as classical particles with an effective mass of the same order, but differing from the ordinary electron mass. The effective mass is often anisotropic, and dif-

ferent for different directions of motion in the crystal. This same effective mass picture can be used to estimate the thermal motion of the gas of electrons and holes. Average thermal velocities at room temperature are of the order of $10^7$ cm/sec.

Scattering can be described in terms of a mean free path for the electrons and holes. In relatively pure crystals at ordinary temperatures, scattering occurs mainly by interaction with the thermal vibrations of the atoms of the crystal. In less pure crystals, or in relatively pure crystals at low temperatures, the mean free path may be determined by scattering by impurity atoms. Because of the scattering, the carriers are not uniformly accelerated by an electric field, but attain an average drift velocity proportional to the field. Ordinarily the drift velocity is much smaller than the average thermal velocity. Drift velocities may be expressed in terms of the mobilities, $\mu_n$ and $\mu_p$ of the electrons and holes respectively*.

In an electric field $E$,

$$(V_d)_n = -\mu_n E$$
$$(V_d)_p = \mu_p E$$

Because of their negative charge, conduction electrons drift oppositely to the field. Values for pure germanium at room temperature are $\mu_n$ = 3,800 $cm^2/volt\ sec$; $\mu_p$ = **1,800** $cm^2/volt$ sec. This means that holes attain a drift velocity of 1,800 cm/sec in a field of one volt/cm.

Expressions for the conductivity are:

$$\text{n-type}: \quad \sigma_n = ne\mu_n$$
$$\text{p-type}: \quad \sigma_p = pe\mu_p$$
$$\text{intrinsic}: \quad \sigma = ne\mu_n + pe\mu_p$$

It is not possible to determine $n$ and $\mu_n$ separately from measurements of the conductivity alone. There are several methods to determine the mobility; one which has been widely used is to measure the Hall coefficient in addition to the conductivity. As part of the research program at the Bell Laboratories, Pearson and Hall made resistivity measurements over a wide range of temperatures of silicon containing varying amounts of boron (a Group III ac-

* A subscript $n$ (referring to negative charge) is used for conduction electrons, $p$ (positive) for holes.

ceptor) and of phosphorus (a Group V donor). Analysis of the data[10] gave additional confirmation of the theory we have outlined. Similar measurements on germanium were made about the same time by Lark-Horovitz and co-workers, and more recently more complete measurements on both materials have been made by other groups. The result of a large amount of experimental and theoretic work has been to confirm the Wilson model in quantitative detail.

Carriers move not only under the influence of an electric field, but also by diffusion; the diffusion current is proportional to the concentration gradient. Expressions for the particle current densities of holes and electrons, respectively, are

$$j_p = p\mu_p E - D_p \operatorname{grad} p$$
$$j_n = n\mu_n E - D_n \operatorname{grad} n$$

Einstein has shown that mobilities and diffusion coefficients are related:

$$\mu = \frac{e}{kT} D$$

where $k$ is Boltzmann's constant. Diffusion and conduction currents both play an important role in the transistor.

The diffusion term was first considered by Wagner in his theory of oxidation of metals. The equations were worked out more completely by Frenkel[2] in an analysis of the diffusive flow which occurs when light is absorbed near one face of a slab, as shown schematically in Fig. 4. The light quanta raise

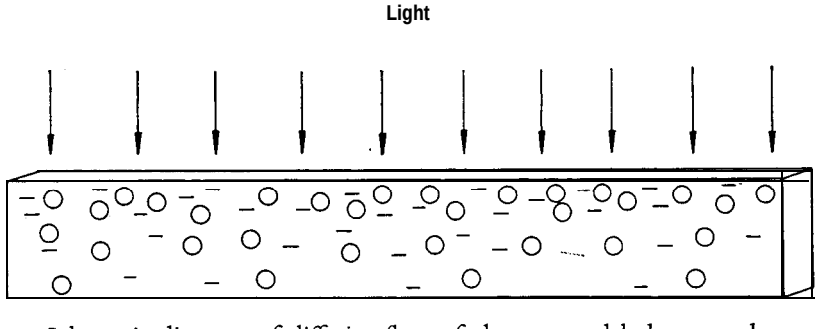


Fig. 4. Schematic diagram of diffusive flow of electrons and holes created near the surface by absorption of light.

electrons from the valence to the conduction bands, creating conduction electrons and holes in equal numbers. These diffuse toward the interior of the slab. Because of recombination of conduction electron and hole pairs, the concentration drops as the diffusion occurs. Frenkel gave the general equations of flow when electrons and holes are present in attempting to account for the Dember effect (change in contact potential with light) and the photomagnetoelectric (PME) effect. The latter is a voltage analogous to a Hall voltage observed between the ends of a slab in a transverse magnetic field (perpendicular to the paper in the diagram). The Dember voltage was presumed to result from a difference of mobility, and thus of diffusion coefficient, between electrons and holes. Electrical neutrality requires that the concentrations and thus the concentration gradients be the same. Further, under steady-state conditions the flow of electrons to the interior must equal the flow of holes, so that there is no net electrical current. However, if $D_n$ is greater than $D_p$, the diffusive flow of electrons would be greater than that of holes. What happens is that an electric field, $E$, is introduced which aids holes and retards the electrons so as to equalize the flows. The integral of $E$ gives a voltage difference between the surface and the interior, and thus a change in contact potential. As we will mention later, much larger changes in contact potential with light may come from surface barrier effects.

## *Contact rectifiers*

In order to understand how a point-contact transistor operates, it is necessary to know some of the features of a rectifying contact between a metal and semiconductor. Common examples are copper-oxide and selenium rectifiers and germanium and silicon point-contact diodes which pass current much more readily for one direction of applied voltage than the opposite. We shall follow Schottky's picture[4], and use as an illustration a contact to an n-type semiconductor. Similar arguments apply to p-type rectifiers with appropriate changes of sign of the potentials and charges. It is most convenient to make use of an energy-level diagram in which the changes in energy bands resulting from changes in electrostatic potential are plotted along a line perpendicular to the contact, as in Fig. 5. Rectification results from the potential energy barrier at the interface which impedes the flow of electrons across the contact.

The Fermi level of the metal is close to the highest of the normally oc-

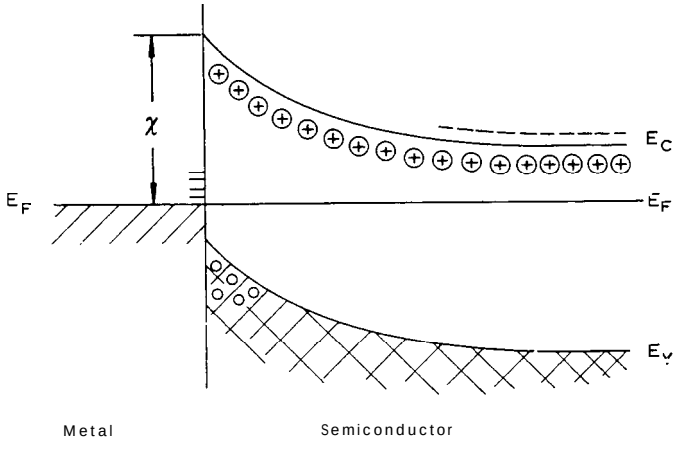


Fig. 5. Equilibrium energy-level diagram for a metal-semiconductor rectifying contact along a line perpendicular to the interface. Variations in the energy bands of the semiconductor result from changes in electrostatic potential due to the layer of uncompensated space-charge. The overall change in potential from the surface to the interior is such as to bring the Fermi level in the interior of the semiconductor into coincidence with that of the metal. In this example, there is an inversion from n-type conductance in the bulk to p-type at the surface.

cupied levels of the conduction band. Because of the nature of the metal-semiconductor interface layers, a relatively large energy, $\chi$, perhaps of the order of 0.5 eV, is required to take an electron from the Fermi level of the metal and place it in the conduction band in the semiconductor. In the interior of the semiconductor, which is electrically neutral, the position of the Fermi level relative to the energy bands is determined by the concentration of conduction electrons, and thus of donors. In equilibrium, with no voltage applied, the Fermi levels of the metal and semiconductor must be the same. This is accomplished by a region of space charge adjacent to the metal in which there is a variation of electrostatic potential, and thus of potential energy of the electron, as illustrated.

In the bulk of the semiconductor there is a balance between conduction electrons and positive donors. In the barrier region which is one of high potential energy for electrons, there are few electrons in the conduction band. The uncompensated space charge of the donors is balanced by a negative charge at the immediate interface. It is these charges, in turn, which produce the potential barrier. The width of the space-charge region is typically of the order of $10^{-5}$ to $10^{-4}$ cm.

When a voltage is applied, most of the drop occurs across the barrier layer. The direction of easy flow is that in which the semiconductor is negative relative to the metal. The bands are raised, the barrier becomes narrower, and electrons can flow more easily from the semiconductor to the metal. In the high resistance direction, the semiconductor is positive, the bands are lowered relative to the metal, and the barrier is broadened. The current of electrons flowing from the metal is limited by the energy barrier, $\chi$, which must be surmounted by thermal excitation.

If $\chi$ is sufficiently large, the Fermi level at the interface may be close to the valence band, implying an inversion from n-type conductivity in the bulk to p-type near the contact. The region of hole conduction is called, following Schottky, an inversion layer. An appreciable part of the current flow to the contact may then consist of minority carriers, in this case holes. An important result of the research program at the Bell Laboratories after the war was to point out the significance of minority carrier flow.

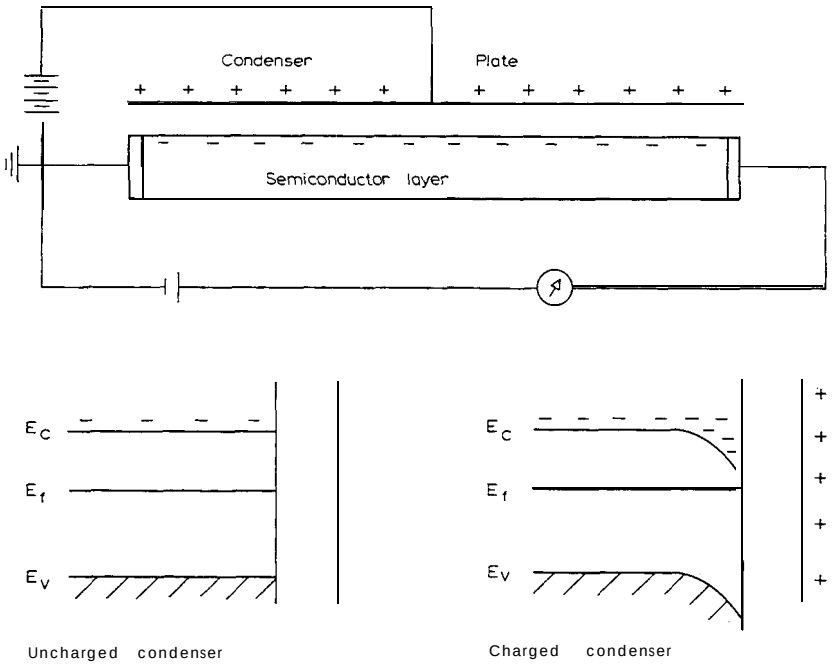


Fig. **6.** Schematic diagram of a field-effect experiment for an n-type semiconductor with no surface states.

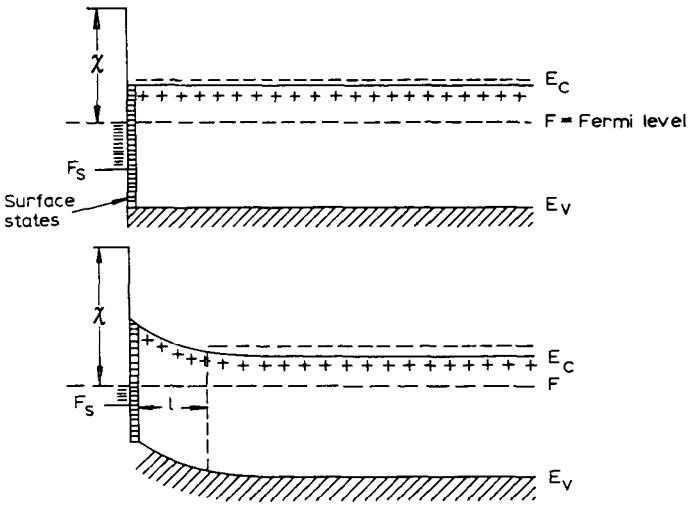


Fig. 7. Formation of a space-charge barrier layer at the free surface of a semiconductor

*Experiments on surface states*

We have mentioned in the introduction that the negative result of the field-effect experiment was an important factor in suggesting the existence of surface states on germanium and silicon, and directing the research program toward a study of surface properties. As is shown in Fig. 6, the experiment consists of making a thin film or slab one plate of a parallel plate condenser and then measuring the change in conductance of the slab with changes in voltage applied across the condenser. The hypothetical case illustrated is an n-type semiconductor with no surface states. When the field plate is positive, the negative charge induced on the semiconductor consists of added electrons in the conduction band. The amount of induced charge can be determined from the applied voltage and measured capacity of the system. If the mobility is known, the expected change in conductance can be calculated readily.

When experiments were performed on evaporated films of germanium and silicon, negative results were obtained; in some cases the predicted effect was more than one thousand times the experimental limit of detection. Analysis indicated that a large part of the discrepancy, perhaps a factor of 50 to 100, came from the very low mobility of electrons in the films as compared with bulk material. The remaining was attributed to shielding by surface states.

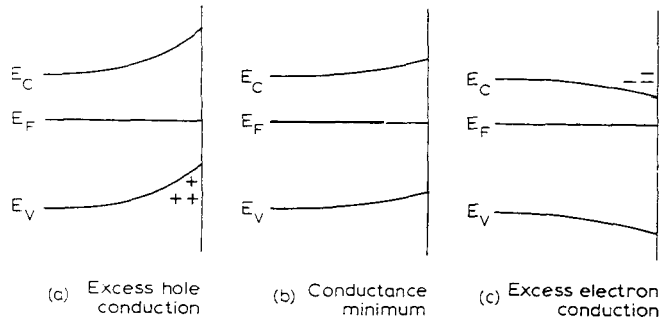


Fig. 8. Types of barrier layers which may exist at the free surface of an n-type semiconductor: *(a)* excess conductance from an inversion layer of p-type conductivity; (*b*) near the minimum surface conductance; (c) excess conductance from an accumulation layer of electrons.

It was predicted that if surface states exist, a barrier layer of the type found at a metal contact might be found at the free surface of a semiconductor. The formation of such a layer is illustrated schematically in Fig. 7. Occupancy of the surface levels is determined by the position of the Fermi level at the surface. In the illustration, it is presumed that the distribution of surface states is such that the states themselves would be electrically neutral if the Fermi level crossed at the position *Fs* relative to the bands. If there is no surface barrier, so that the Fermi level crosses the surface above *Fs,* there are excess electrons and a net negative charge in the surface states. When the surface as whole is neutral, a barrier layer is formed such that the positive charge in the layer is compensated by the negative surface states charge. If the density of surface states is reasonably high, sufficient negative charge is obtained with the Fermi level crossing only slightly above *Fs.*

Types of barriers which may exist at the surface of an n-type semiconductor are illustrated in Fig. 8. On the left (a) the energy bands are raised at the surface so as to bring the valence band close to the Fermi level. An inversion layer of opposite conductivity type is formed, and there is excess conductance from mobile holes in the layer. Negative charge on the surface proper is balanced by the charge of holes and of fixed donor ions in the barrier region. In *(b)* the bands are raised at the surface, but not enough to form a barrier layer. The surface resistance is near a maximum. In (c), the bands bend down so as to form an *accumulation* layer of excess electron conductance near the surface. The charge on the surface proper is now positive, and is balanced by the negative charge of the excess electrons in the layer.

The postulated existence of surface states and surface barrier layers on the free surface of germanium and silicon accounted for several properties of germanium and silicon which had hitherto been puzzling[8]. These included (1) lack of dependence of rectifier characteristics on the work function of the metal contact, (2) current voltage characteristics of a contact made with two pieces of germanium, and (3) the fact that there was found little or no contact potential difference between n- and p-type germanium and between n- and p-type silicon.

While available evidence for surface states was fairly convincing, it was all of an indirect nature. Further, none of the effects gave any evidence about the height of the surface barrier and of the distribution of surface states. A number of further experiments which might yield more concrete evidence about the surface barrier was suggested by Shockley, Brattain and myself. Shockley predicted that a difference in contact potential would be found between n- and p-type specimens with large impurity concentration. A systematic study of Brattain and Shockley[12] using silicon specimens with varying amounts of donor and acceptor impurities showed that this was true, and an estimate was obtained for the density of surface states. Another experiment which indicated the presence of a surface barrier was a measurement of the change in contact potential with illumination of the surface. This is just the Dember effect, which Frenkel had attempted to account for by the difference in mobilities of the electrons and holes generated by the light and

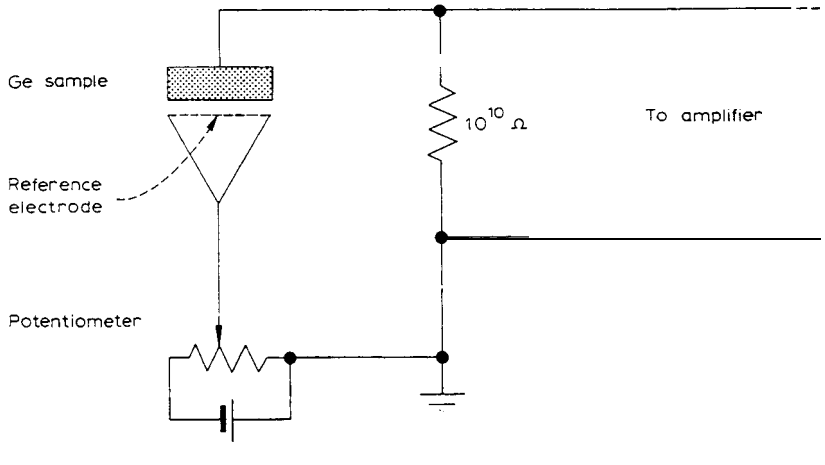


Fig. 9. Schematic diagram of apparatus used by Brattain to measure contact potential and change of contact potential with light.

diffusing to the interior. It was found[13] that the change is usually much larger and often of the opposite sign than predicted by Frenkel's theory, which did not take into account a surface barrier.

Some rather difficult experiments which at the time gave negative results have been carried out successfully much later by improved techniques, as will be described by Dr. Brattain in his talk.

Apparatus used by Brattain to measure contact potential and change in contact potential with illumination is shown in Fig. 9. The reference electrode, generally platinum, is in the form of a screen so that light can pass through it. By vibrating the electrode, the contact potential itself can be measured by the Kelvin method. If light chopped at an appropriate frequency falls on the surface and the electrode is held fixed, the change with illumination can be measured from the alternating voltage developed across the condenser. In the course of the study, Brattain tried several ambient atmospheres and different temperatures. He observed a large effect when a liquid dielectric filled the space between the electrode and semiconductor surface. He and Gibney then introduced electrolytes, and observed effects attributed to large changes in the surface barrier with voltage applied across the electrolyte. Evidently ions piling up at the surface created a very large field which penetrated through the surface states.

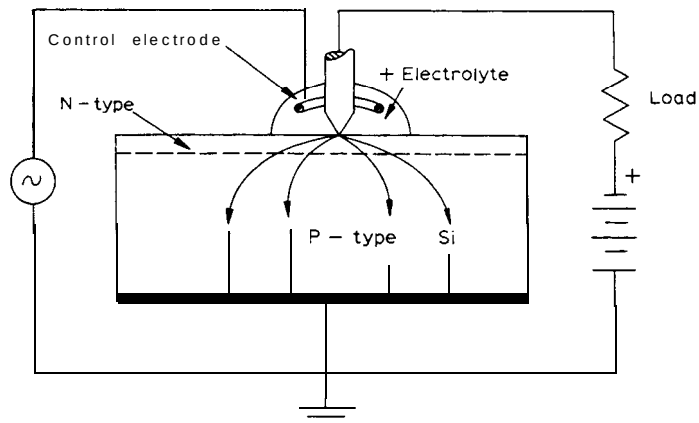


Fig. 10. Diagram of experiment used to observe effect of the field produced by an electrolyte on an inversion layer of n-type conductance at the surface of a p-type silicon block. Negative potential applied to the probe in the electrolyte decreases the number of electrons in the inversion layer and thus the current of electrons flowing to the point contact biased in the reverse direction. Arrows indicate the conventional direction of current flow; electrons move in the opposite direction.

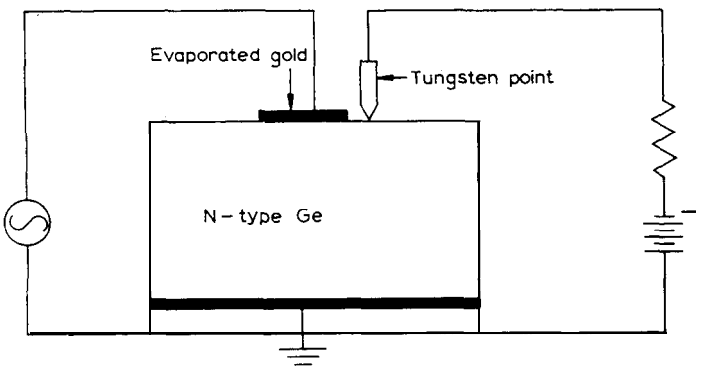


Fig. 11. Diagram of experiment in which the transistor effect was first observed. Positive voltage applied to the gold spot introduced holes into the n-type germanium block which flowed to the point contact biased in the reverse direction. It was found that an increase in positive voltage increased the reverse current. When connected across a high impedance, the change in voltage of the point contact was larger than the change at the gold spot, both measured relative to the base electrode.

### *Experiments on inversion layers*

Use of an electrolyte provided a method for changing the surface barrier, so that it should be possible to observe a field effect in a suitable arrangement. We did not want to use an evaporated film because of the poor structure and low mobility. With the techniques available at the time, it would have been difficult to prepare a slab from bulk material sufficiently thin to observe a sizable effect. It was suggested that one could get the effect of a thin film in bulk material by observing directly the flow in an inversion layer of opposite conductivity type near the surface. Earlier work of Ohl and Scaff indicated that one could get an inversion layer of n-type conductivity on p-type silicon by suitably oxidizing the surface. If a point contact is made which rectifies to the p-type base, it would be expected to make low resistance contact to the inversion layer.

The arrangement which Brattain and I used in the initial tests is shown in Fig. 10. The point contact was surrounded by, but insulated from, a drop of electrolyte. An electrode in the electrolyte could be used to apply a strong field at the semiconductor surface in the vicinity of the contact. The reverse, or high resistance direction is that in which point is positive relative to the block. Part of the reverse current consists of electrons flowing through the n-type inversion layer to the contact. It was found that the magnitude of

this current could be changed by applying a voltage on the electrolyte probe, and thus, by the field effect, changing the conductance of the inversion layer. Since under static conditions only a very small current flowed through the electrolyte, the set-up could be used as an amplifier. In the initial tests, current and power amplification, but not voltage amplification, was observed. As predicted from the expected decrease in number of electrons in the inversion layer, a negative voltage applied to the probe was found to decrease the current flowing in the reverse direction to the contact.

It was next decided to try a similar arrangement with a block of n-type germanium. Although we had no prior knowledge of a p-type inversion layer on the surface, the experiments showed definitely that a large part of the reverse current consisted of holes flowing in an inversion layer near the surface. A positive change in voltage on the probe decreased the reverse current. Considerable voltage as well as current and power amplification was observed.

Because of the long time constants of the electrolyte used, amplification was obtained only at very low frequencies. We next tried to replace the electrolyte by a metal control electrode insulated from the surface by either a thin oxide layer or by a rectifying contact. A surface was prepared by Gibney by anodizing the surface and then evaporating several gold spots on it. Although none made the desired high resistance contact to the block, we decided to see what effects would be obtained. A point contact was placed very close to one of the spots and biased in the reverse direction (see Fig. 11). A small effect on the reverse current was observed when the spot was biased positively, but of *opposite* direction to that observed with the electrolyte. An increase in positive bias *increased* rather than decreased the reverse current to the point contact. The effect was large enough to give some voltage, but no power amplification. This experiment suggested that holes were flowing into the germanium surface from the gold spot, and that the holes introduced in this way flowed into the point contact to enhance the reverse current. This was the first indication of the transistor effect.

It was estimated that power amplification could be obtained if the metal contacts were spaced at distances of the order of 0.005 cm. In the first attempt, which was successful, contacts were made by evaporating gold on a wedge, and then separating the gold at the point of the wedge with a razor blade to make two closely spaced contacts. After further experimentation, it appeared that the easiest way to achieve the desired close separation was to use two appropriately shaped point contacts placed very close together.

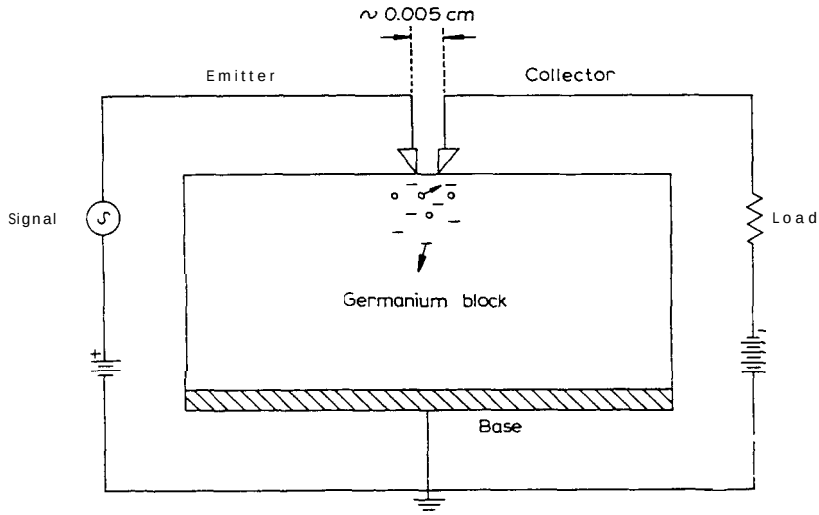


Fig. 12. Schematic diagram of point-contact transistor.

Success was achieved in the first trials; the point-contact transistor was born[14].

It was evident from the experiments that a large part of both the forward and reverse currents from a germanium point contact is carried by minority carriers, in this case holes. If this fact had been recognized earlier, the transistor might have come sooner.

Operation of a point-contact transistor is illustrated in Fig. 12. When operated as an amplifier, one contact, the emitter, is biased with a d.c. voltage in the forward direction, the second, the collector, in the negative or high resistance direction. A third contact, the base electrode, makes a low resistance contact to the block. A large part of the forward current consists of holes flowing into the block. Current from the collector consists in part of electrons flowing from the contact and in part of holes flowing toward the contact. The collector current produces an electric field in the block which is in such a direction as to attract holes introduced at the emitter. A large part of the emitter current, introduced at low impedance, flows in the collector circuit. Biased in the reverse direction, the collector has high impedance and can be matched to a high impedance load. There is thus a large voltage amplification of an input signal. It is found[14] that there is some current amplification as well, giving an overall power gain of 20 db. or more. An increase in hole current at the collector affects the barrier there in such a way as to enhance the current of electrons flowing from the contact.

The collector current must be sufficiently large to provide an electric field to attract the holes from the emitter. The optimum impendance of the collector is considerably less than that of a good germanium diode in the reverse direction. In the first experiments, it was attempted to achieve this by treating the surface so as to produce a large inversion layer of p-type conductivity on the surface. In this case, a large fraction of the hole current may flow in the inversion layer. Later, it was found that better results could be obtained by electrically forming the collector by passing large current pulses through it. In this case the surface treatment is less critical, and most of the emitter current flows through the bulk.

Studies of the nature of the forward and reverse currents to a point contact to germanium were made by making probe measurements of the variation of potential in the vicinity of the contact[15]. These measurements showed a large increase in conductivity when the contact was biased in the forward direction and in some cases evidence for a conducting inversion layer near the surface when biased in the reverse direction.

Before it was established whether the useful emitter current was confined to an inversion layer or could flow through the bulk, Shockley[16] proposed a radically different design for a transistor based on the latter possibility. This is the junction transistor design in which added minority carriers from the emitter diffuse through a thin base layer to the collector. Independently of this suggestion, Shive[17] made a point-contact transistor in which the emitter and collector were on opposite faces of a thin slab of germanium. This showed definitely that injected minority carriers could flow for small distances through bulk material. While transistors can be made to operate either way, designs which make use of flow through bulk material have been most successful. Junction transistors have superseded point-contact transistors for most applications.

Following the discovery of the transistor effect, a large part of research at the Bell Laboratories was devoted to a study of flow on injected minority carriers in bulk material. Much of this research was instigated by Shockley, and will be described by him in the following talk.

Research on surface properties of germanium and silicon, suspended for some time after 1948 because of the pressure of other work, was resumed later on by Brattain and others, and is now a flourishing field of activity with implications to a number of scientific fields other than semiconductors such as adsorption, catalysis, and photoconductivity. This research program will be described by Dr. Brattain in his talk.

It is evident that many years of research by a great many people, both before and after the discovery of the transistor effect, has been required to bring our knowledge of semiconductors to its present development. We were fortunate enough to be involved at a particularly opportune time and to add another small step in the control of Nature for the benefit ofmankind. In addition to my colleagues and to others mentioned in the lecture, I would like to express my deep gratitude to Drs. M. J. Kelly and Ralph Bown for the inspired leadership of the Laboratories when this work was done.

1. A. H. Wilson, *Proc. Roy. Soc. London,* A 133 (1931) 458; A 134 (1932) 277; A 136 (1932) 487.
2. J. Frenkel, *Physik. Z. Sowjetunion, 8* (1935) 185.
3. N. F. Mott, *Proc. Roy. Soc. London,* A 171 (1939) 27.
4. W. Schottky, Z. *Physik,* 113 (1939) 367; 118 (1942) 539.
5. B. Davydov, *J. Tech. Phys. U.S.S.R., 5* (1938) 87.
6. R. Hilsch and R. W. Pohl, Z. *Physik,* III (1938) 399.
7. Amplifiers based on the field-effect principle had been suggested earlier in the patent literature (R. Lillienfeld and others), but apparently were not successful. Shockley's contribution was to show that it should be possible according to existing semiconductor theory to make such a device. An early successful experiment is that of W. Shockley and G. L. Pearson, *Phys. Rev.,* 74 (1948) 232.
8. J. Bardeen, *Phys. Rev.,* 71 (1947) 717.
9. A review is given in the lecture of Dr. Brattain, this volume, p. 337.
10. G. L. Pearson and J. Bardeen, *Phys. Rev.,* 75 (1949) *865.*
11. *See* K. Lark-Horovitz, *Elec. Eng.,* 68 (1949) 1047.
12. W. H. Brattain and W. Shockley, *Phys. Rev.,* 72 (1947) 345.
13. W. H. Brattain, *Phys. Rev.,* 71 (1947) 345.
14. J. Bardeen and W. H. Brattain, *Phys. Rev.,* 74 (1948) 230; 75 (1949) 1208.
15. W. H. Brattain and J. Bardeen, *Phys. Rev.,* 74 (1948) 231.
16. W. Shockley, *Electrons and Holes in Semiconductors,* D. Van Nostrand Co., Inc., New York, 1950, p. 86.
17. J. N. Shive, *Phys. Rev.,* 75 (1949) *689.*