U.S. Patent 2,524,035

John Bardeen et al.

Three-Electrode Circuit Element Utilizing Semiconductive Materials

Application filed in 1948

Patent issued to *Bell Labs* in 1950

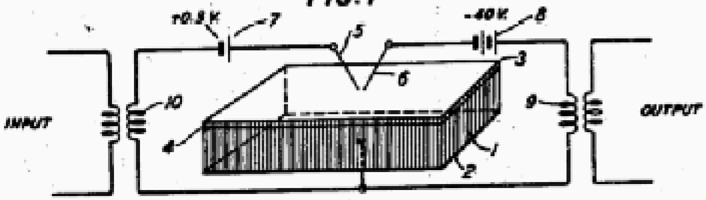
One of the most significant electrical engineering patents of all time begins by listing the following objectives:

> . . . to amplify or otherwise translate electric signals or variations by use of compact, simple, and rugged apparatus of novel type. Another object is to provide a circuit element for use as an amplifier or the like which does not require a heated thermionic cathode for its operation, and which therefore is immediately operative when turned on. A related object is to provide such a circuit element which requires no evacuated or gas-filled envelope. Col. 1, lines 8-18.

Half a century after the invention of the transistor, the objectives listed above seem modest compared to the revolutionary changes made by this “three-electrode circuit element”:

> The present invention in one form utilizes a block of semiconductor material on which three electrodes are placed. One of these, termed the collector, makes rectifier contact with the body of the block. The other, termed the emitter, preferably makes rectifier contact with the body of the block also. The third electrode, which may be designated the base electrode, preferably makes a low resistance contact with the body of the block. Col. 2, lines 22-30.

FIG. 1



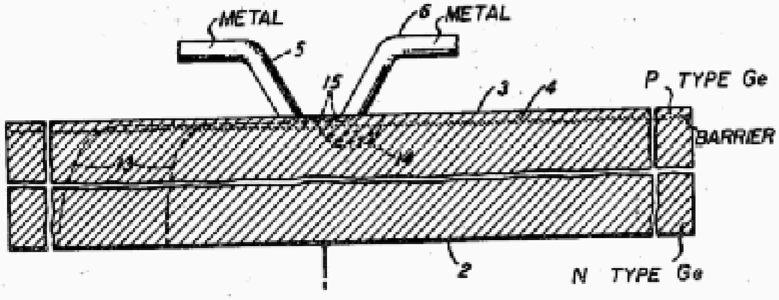


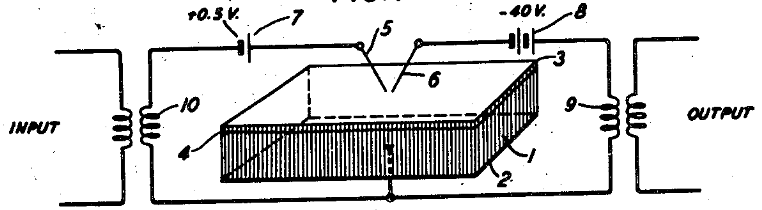
FIG. 1A

Oct. 3, 1950 J. BARDEEN ET AL 2,524,035

THREE-ELECTRODE CIRCUIT ELEMENT UTILIZING
SEMICONDUCTIVE MATERIALS

Filed June 17, 1948 3 Sheets-Sheet 1

FIG. 1



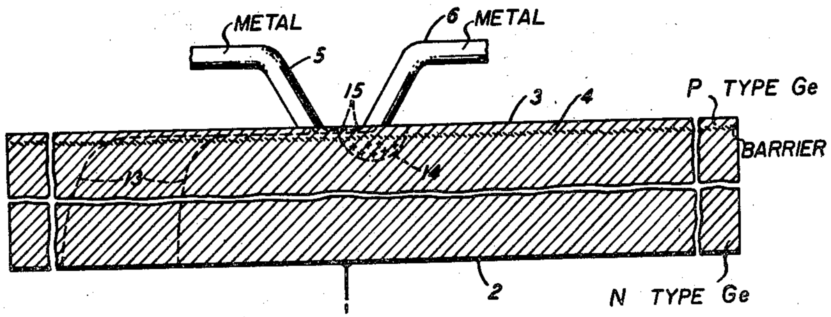


FIG. 1A

FIG. 2

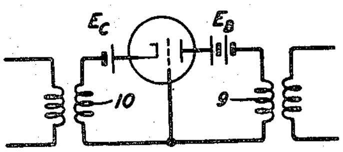


FIG. 10

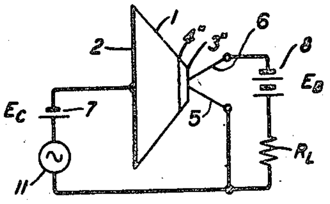


FIG. 11

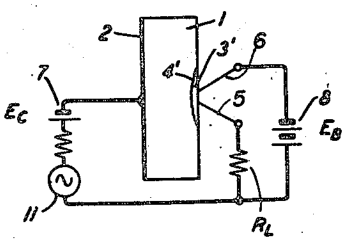
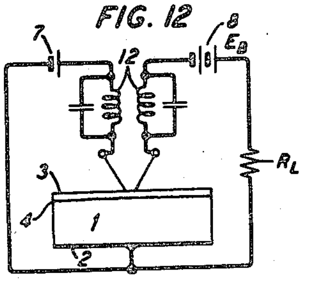


FIG. 12



INVENTORS: J. BARDEEN
W. H. BRATTAIN

BY Harry C. Hart

ATTORNEY

Oct. 3, 1950 J. BARDEEN ET AL 2,524,035

THREE-ELECTRODE CIRCUIT ELEMENT UTILIZING
SEMICONDUCTIVE MATERIALS

Filed June 17, 1948 3 Sheets-Sheet 2

FIG. 3

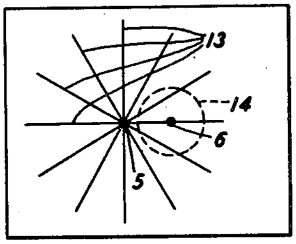


FIG. 3A

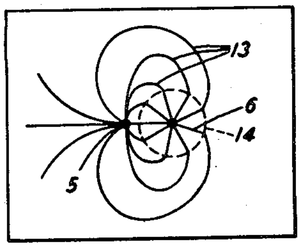


FIG. 4

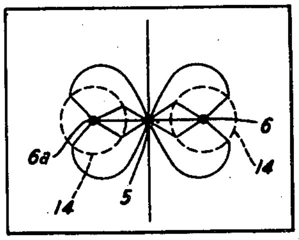


FIG. 5

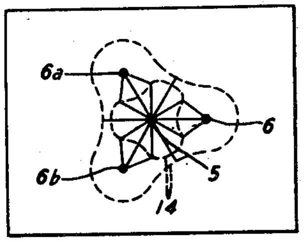


FIG. 6

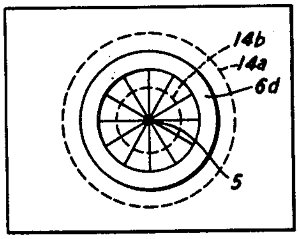


FIG. 7

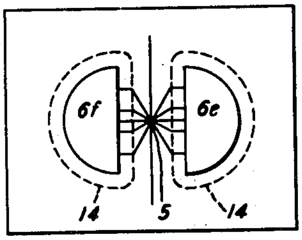


FIG. 8

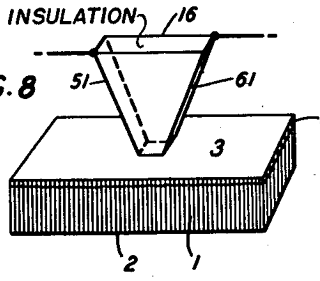
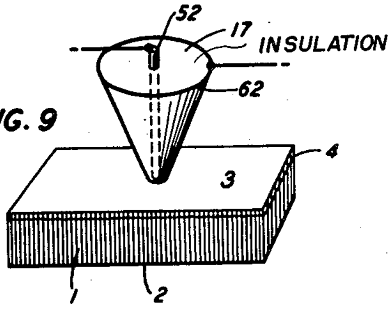


FIG. 9



INVENTORS: J. BARDEEN
W. H. BRATTAIN
BY Harry C. Hart
ATTORNEY

Oct. 3, 1950 J. BARDEEN ET AL 2,524,035

THREE-ELECTRODE CIRCUIT ELEMENT UTILIZING SEMICONDUCTIVE MATERIALS

Filed June 17, 1948 3 Sheets-Sheet 3

FIG. 13

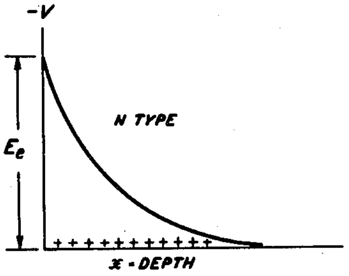


FIG. 14

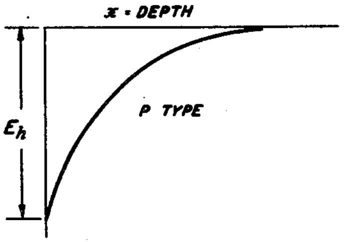


FIG. 15

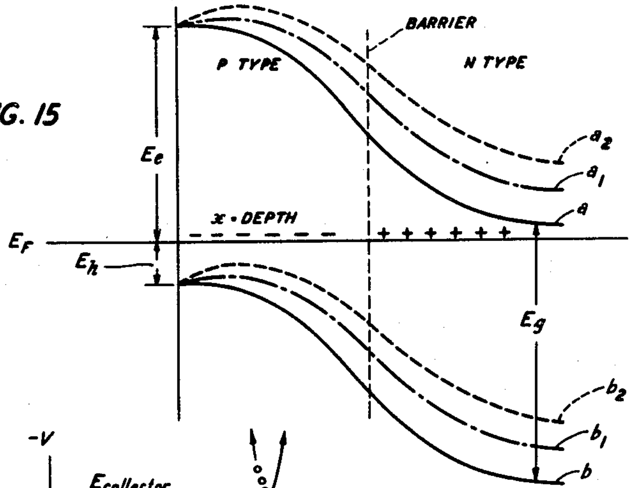
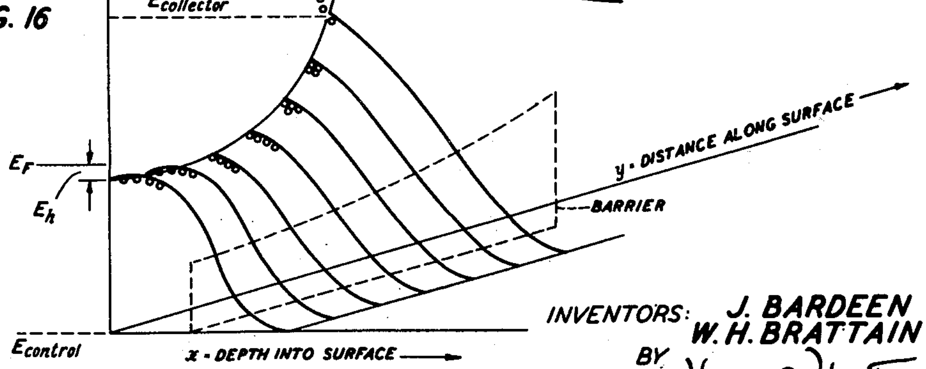


FIG. 16



INVENTORS: J. BARDEEN
W. H. BRATTAIN

BY Harry C. Hart

ATTORNEY

Patented Oct. 3, 1950 **2,524,035**

UNITED STATES PATENT OFFICE

2,524,035

THREE-ELECTRODE CIRCUIT ELEMENT UTILIZING SEMICONDUCTIVE MATERIALS

John Bardeen, Summit, and Walter H. Brattain, Morristown, N. J., assignors to Bell Telephone Laboratories, Incorporated, New York, N. Y., a corporation of New York

Application June 17, 1948, Serial No. 33,466

40 Claims. (Cl. 179—171)

This application is a continuation-in-part of application Serial No. 11,165, filed February 26, 1948, and thereafter abandoned.

This invention relates to a novel method of and means for translating electrical variations for such purposes as amplification, wave generation, and the like.

The principal object of the invention is to amplify or otherwise translate electric signals or variations by use of compact, simple, and rugged apparatus of novel type.

Another object is to provide a circuit element for use as an amplifier or the like which does not require a heated thermionic cathode for its operation, and which therefore is immediately operative when turned on. A related object is to provide such a circuit element which requires no evacuated or gas-filled envelope.

Attempts have been made in the past to convert solid rectifiers utilizing selenium, copper sulfide, or other semi-conductive materials into amplifiers by the direct expedient of embedding a grid-like electrode in a dielectric layer disposed between the cathode and the anode of the rectifier. The grid is supposed, by exerting an electric force at the surface of the cathode, to modify its emission and so alter the cathode-anode current. As a practical matter it is impossible to embed a grid in a layer which is so thick as to insulate the grid from the other electrodes and yet so thin as to permit current to flow between them. It has also been proposed to pass a current from end to end of a strip of homogeneous isotropic semiconductive material and, by the application of a strong transverse electrostatic field, to control the resistance of the strip, and hence the current through it.

So far as is known, all of such past devices are beyond human skill to fabricate with the fineness necessary to produce amplification. In any event they do not appear to have been commercially successful.

It is well known that in semiconductors there are two types of carriers of electricity which differ in the signs of the effective mobile charges. The negative carriers are excess electrons which are free to move, and are denoted by the term conduction electrons or simply electrons. The positive carriers are missing or defect "electrons," and are denoted by the term "holes." The conductivity of a semiconductor is called excess or defect, or N or P type, depending on whether the mobile charges normally present in excess in the material under equilibrium conditions are electrons (Negative carriers) or holes (Positive carriers).

When a metal electrode is placed in contact with a semiconductor and a potential difference is applied across the junction, the magnitude of the current which flows often depends on the sign as well as on the magnitude of the potential. A junction of this sort is called a rectifying contact. If the contact is made to an N-type semiconductor, the direction of easy current flow is that in which the semiconductor is negative with respect to the electrode. With a P-type semiconductor, the direction of easy flow is that in which the semiconductor is positive. A similar rectifying contact exists at the boundary between two semiconductors of opposite conductivity types.

This boundary may separate two semiconductor materials of different constitutions, or it may separate zones or regions, within a body of semiconductor material which is chemically and stoichiometrically uniform, which exhibit different conductivity characteristics.

The present invention in one form utilizes a block of semiconductor material on which three electrodes are placed. One of these, termed the collector, makes rectifier contact with the body of the block. The other, termed the emitter, preferably makes rectifier contact with the body of the block also. The third electrode, which may be designated the base electrode, preferably makes a low resistance contact with the body of the block. When operated as an amplifier, the emitter is normally biased in the direction of easy current flow with respect to the body of the semiconductor block. The nature of the emitter electrode and of that portion of the semiconductor which is in the immediate neighborhood of the electrode contact is such that a substantial fraction of the current from this electrode is carried by charges whose signs are opposite to the signs of the mobile charges normally in excess in the body of the semiconductor. The collector is biased in the reverse, or high resistance direction relative to the body of the semiconductor. In the absence of the emitter, the current to the collector flows exclusively from the base electrode and is impeded by the high resistance of this collector contact. The sign of the collector bias potential is such as to attract the carriers of opposite sign which come from the emitter. The collector is so disposed in relation to the emitter that a large fraction of the emitter current enters the collector. The fraction depends in part on the geometrical disposition of the electrodes and in part on the bias potentials applied. As the emitter is biased in the direction of easy flow, the emitter current is sensitive to small changes in potential between the emitter and the body of the semiconductor, or between the emitter and the base electrode. Application of a small voltage variation between the base electrode and emitter causes a relatively large change in the current entering the semiconductor from the emitter, and a correspondingly large change in the current to the collector. One effect of the change in emitter current is to modify the total current flowing to the collector, so that the overall change in collector current may be greater than the change in the emitter current. The collector circuit may contain a load of high impedance matched to the internal impedance of the collector, which, because of the high resistance rectifier contact of the collector, is high. As a result, voltage amplification, current amplification, and power amplification of the input signal are obtained.

In one form, the device utilizes a block of semiconductor material of which the main body is of one conductivity type while a very thin surface layer or film is of opposite conductivity type. The surface layer is separated from the body by a high resistance rectifying barrier. The emitter and collector electrodes make contact with this surface layer sufficiently close together for mutual influence in the manner described above. The base electrode makes a low resistance contact with the body of the semiconductor. When suitable bias potentials are applied to the various electrodes, a current flows from the emitter into the thin layer. Owing to the conductivity of the layer and to the nature of the barrier, this current tends to flow laterally in the thin layer, rather than following the most direct path across the barrier to the base electrode. This current is composed of carriers whose signs are opposite to the signs of the mobile charges normally in excess in the body of the semiconductor. In other words, when there is a thin layer of opposite conductivity type immediately under the emitter electrode, the current flowing into the block in the direction of easy flow consists largely of carriers of opposite sign to those of the mobile charges normally present in excess in the body of the block; and the presence of these carriers increases the conductivity of the block. The bias voltage on the collector which, as stated above, is biased in the reverse or high resistance direction relative to the block, produces a strong electrostatic field in a region surrounding the collector so that the current from the emitter which enters this region is drawn in to the collector. Thus, the collector current, and hence the conductance of the unit as a whole, are increased. The size of the region in which this strong field exists is comparatively insensitive to variations in the collector potential so that the impedance of the collector circuit is high. On the other hand, the current from the emitter to the layer is extremely sensitive to variations of the emitter potential, so that the impedance of the emitter circuit is low.

It is a feature of the invention that the input and output impedances of the device are controlled by choice and treatment of the semiconductor material body and of its surface, as well as by choice of the bias potentials of the electrodes.

From the standpoint of its external behavior and uses, the device of the invention resembles a vacuum tube triode; and while the electrodes are designated emitter, collector and base electrode, respectively, they may be externally interconnected in the various ways which have become recognized as appropriate for triodes, such as the conventional, the "grounded grid," the "grounded plate" or cathode follower, and the like. Indeed, the discovery on which the invention is based was first made with circuit connections which are extremely similar to the so-called "grounded grid" vacuum tube connections. However, the analogies among the circuits is, of course, no better than the analogy between emitter and cathode, base electrode and grid, collector and anode.

By feeding back a portion of the output voltage in proper phase to the input terminals, the device may be caused to oscillate at a frequency determined by its external circuit elements, and, among other tests, power amplification was confirmed by a feedback connection which caused it to oscillate.

It has been found that the performance of the device is expressed, to a good approximation, by the following functional relations:

$$I_e=f(V_e+R_F I_c) \qquad (1)$$

$$I_c=I_c^0(V_c)+\alpha I_e \qquad (1a)$$

where

I_e=emitter current

I_c=collector current

$I_c^0(V_c)$=collector current with emitter disconnected

V_e=voltage of emitter electrode measured with respect to the base electrode

V_c=voltage of collector electrode measured with respect to the base electrode

R_F=an equivalent resistance independent of bias

α=a numerical factor which depends on the bias voltages

$f(V_e)$ gives the relation between emitter current and emitter voltage with the collector circuit open.

The interpretation of the foregoing Equation 1 is that the collector current lowers the potential of the surface of the block in the vicinity of the emitter relative to the base electrode by an amount $R_F I_c$, and thus increases the effective bias voltage on the emitter by the same amount. The term $R_F I_c$ thus represents positive feedback.

The invention will be fully apprehended from the following detailed description of one embodiment thereof, taken in connection with the appended drawings, in which:

Fig. 1 is a schematic diagram, partly in perspective, showing a preferred embodiment of the invention;

Fig. 1*a* is a cross-section of a part of Fig. 1 to a greatly enlarged scale;

Fig. 2 is the equivalent vacuum tube schematic circuit of Fig. 1;

Fig. 3 is a plan view of the block of Fig. 1, showing the disposition of the electrodes;

Fig. 3*a* is like Fig. 3 but shows the influence of the collector in modifying the emitter current;

Figs. 4, 5, 6 and 7 show electrode dispositions alternative to those of Fig. 1;

Figs. 8 and 9 show electrode structures alternative to those of Fig. 1;

Fig. 10 shows a modified unit of the invention connected for operation in the circuit of a conventional triode;

Fig. 11 shows another modified unit of the invention connected for operation in a "grounded plate" or cathode follower circuit;

Fig. 12 shows the unit of the invention connected for self-sustained oscillation;

is preferably employed as the emitter and the surrounding plate film 62 as collector. The cone and the wedge serve to hold the interelectrode capacities to a minimum while keeping the contacts close together where they bear against the surface of the semiconductor.

Further understanding of the considerations which govern the thickness of the P-type surface layer may be had from the following considerations, which apply specifically to a chemical layer. Fig. 13 is a plot of the electrostatic potential within the body of an N-type semiconductor in contact with a metal. As above stated, the N-type material of the semiconductor contains fixed or bound positive charges. They are believed to be distributed with fair uniformity in depth to a certain distance, beyond which the material is electrically neutral, because the bound positive charges are balanced by equal negative (electron) charges. In accordance with Poisson's equation:

$$\frac{d^2V}{dx^2}=-\frac{4\pi\rho}{\epsilon} \qquad (2)$$

where

V is the potential
x is the distance, measured from the metal into the semiconductor
ρ is the charge density, and
ϵ is the dielectric constant of the material.

Assuming the charge density ρ to be uniform, two integrations give the potential as a function of depth. When plotted, it is a parabola. In the figure, negative potential has been plotted upward. The vertical rise E_e from the Fermi level to the terminus of the curve, i. e., to its intercept with the potential axis, represents the energy which must be imparted to an electron to cause it to move from the metal to the semiconductor. These matters are fully explained in the literature, for example, in "Schottky's Theories of Dry Solid Rectifiers," by J. Joffe, published in "Electrical Communication," vol. 22 (1944–1945) at page 217.

Similarly Fig. 14 shows the potential distribution, for positive holes, within a P-type semiconductor in contact with a metal. In this case the height E_h of the terminus of the curve from the Fermi level represents the energy which must be given to a positive hole to cause it to leave the metal and enter the semiconductor.

Fig. 15 is a composite diagram showing, in the upper curves, the electron energy and in the lower curves the hole energy, within a semiconductor which comprises a thin layer of P-type material separated from a body of N-type material by a barrier. The fixed charges are negative in the P-type material and positive in the N-type, and for simplicity are assumed to be distributed uniformly in each zone. Integration of the charge density, twice, in accordance with Poisson's equation gives the lowermost curves, a, b of the two groups, which represent equilibrium conditions and which, but for an additive constant E_g, are alike. The constant E_g represents the energy difference between the filled band and the conduction band for the particular material.

The middle curves a_1, b_1, of each group represent the conditions when a small negative bias is applied to the semiconductor block 1 with respect to the emitter 5, and the upper curves a_2, b_2, of each group represent the conditions when a signal applied between the emitter and the control electrode further reduces the potential of the block. Evidently the alteration of the block potential with respect to the emitter operates in each case to increase the effective thickness of the P-type layer and so the density of holes and the layer conductivity. Such an increase in conductivity with increase in the forward bias has been observed in connection with the potential probe measurements referred to above.

The rounded peak of the hole potential curve lies below the Fermi level. The greater the thickness of the P-type layer, the more the terminus of this curve falls below the Fermi level, i. e., the greater the magnitude of E_h, and the greater the difficulty for holes to leave the metal of the emitter and enter the semiconductor. Similarly, the thinner the P-type layer, the less is the magnitude of E_h, and the greater the ease with which holes move from the metal of the emitter to the semiconductor and enter it. On the other hand, if the P-type layer is too thin, the conductivity of the layer, which is related to the width of the approximately flat portion of the upper part of the curve b_1 of Fig. 15 will be small. In the vicinity of the collector electrode, the thickness of the P-type layer should be sufficiently small so that the rectification characteristic of the collector is determined primarily by the body of the semiconductor and not by the layer. If, now, the collector is biased in the reverse direction relative to the body, most of the drop from the high voltage on the electrode occurs in the immediate vicinity of the collector, so that the impedance of the collector circuit is high.

The P-type layer is preferably adjusted to an optimum thickness lying between these extremes. Best results are believed to be obtained when its thickness is such that the terminus of the curve falls slightly below the rounded peak. Holes can enter the semiconductor without great difficulty, and tend to collect in the region of greatest negative potential as a cloud of mobile positive charges. They then diffuse away from the emitter—laterally in Fig. 1, perpendicular to the paper in Fig. 15—some of them entering the field 14 of the collector 6.

Because the right-hand part of the lower curve falls well below the left-hand part, positive holes can cross the barrier only with difficulty. Because the P-type layer is thin, the energy E_h, required to cause holes to enter the layer, is small. Therefore holes enter easily under the influence of the positive bias on the emitter 5 and collect in the layer, like air bubbles as it were, at the top of a liquid in a closed vessel. They may easily travel in the layer and parallel with it.

The sense in which, and the reason why the barrier exists, separating a region of P-type conductivity from a region of N-type conductivity, despite the fact that the semiconductor material itself may be chemically and stoichiometrically uniform, may be explained as follows:

From the elementary considerations, the conductivity is given by

$$C=n_1e_1\mu_1+n_2e_2\mu_2 \qquad (3)$$

where

n_1, e_1, μ_1 are the electron density, the electronic charge, and the electron's mobility, respectively, and
n_2, e_2, μ_2 are the corresponding quantities for positive holes.

It is known that

$$n_1=A_1\epsilon^{-\frac{eV_e}{KT}} \qquad (4a)$$

$$n_2=A_2\epsilon^{-\frac{eV_h}{KT}} \qquad (4b)$$

where V_e is the height of the electron space potential curve (*a* of Fig. 15) above the Fermi level, and V_h is, correspondingly, the height of the Fermi level above the hole space potential curve (*b* of Fig. 15) and A_1, A_2, K, and T are constants for a given temperature. Inasmuch as the potential difference between the two kinds of space potential curves is a constant E_g, the conductivity may be written

$$C=A_1\mu_1 e_1\epsilon^{-\frac{eV_e}{KT}}+A_2\mu_2 e_2\epsilon^{-\frac{e(Eg-V_e)}{KT}} \qquad (5)$$

Since the factor $A_1\mu_1 e_1$ does not differ greatly in magnitude from the factor $A_2\mu_2 e_2$, it is a simple matter of calculation to show that this expression is a minimum when

$$V_e=V_h=\frac{E_g}{2} \qquad (6)$$

i. e., that the restivity of the material is greatest at the depth at which the *a* curves and the *b* curves of Fig. 15 lie at equal distances above and below the Fermi level, respectively; and that, furthermore, the resistivity departs rapidly from this maximum value as the space potentials V_e and V_h depart from equality. If

$$V_e<\frac{E_g}{2}$$

the electron conductivity is greater than the hole conductivity, and the conductivity is N-type. If

$$V_e>\frac{E_g}{2}, \text{ or } V_h<\frac{E_g}{2}$$

the hole conductivity is greater than the electron conductivity, and the conductivity is P-type.

Fig. 16 is a three dimensional representation of the conditions which the holes encounter in the course of their travel in the layer from the emitter to the collector—in the figure, parallel with the Y axis. As in Fig. 15, the X axis represents depth measured into the semi-conductor and the V axis which is drawn to an approximately logarithmic scale, represents negative potential. As the holes approach the collector the peak of the potential curve becomes less and less pronounced until finally, at the collector, the region of lowest potential, to which the holes flow, is the collector itself, where they are withdrawn.

Of that part of the emitter current which crosses the barrier, a certain fraction crosses it again in the vicinity of the collector and is collected, thus forming a part of the collector current. The foregoing hypothesis as to the mechanism by which amplification is obtained applies to this fraction of the current as well as to the fraction which proceeds entirely within the layer.

The collector current contains still another component, which consists of a flow of electrons from the collector to the base electrode, crossing the barrier once on its way. A hypothesis as to the manner in which this current component takes part in the amplification process is as follows:

There is a potential hill at the contact point between the collector electrode and the surface layer which offers an impedance to the flow of electrons from the electrode into the semiconductor. In the absence of bias, the height of this hill, indicated by E_e in Figs. 13 and 15, is the energy required to take an electron from the metal and place it in the conduction band of the semiconductor. When the collector is biased in the reverse direction, the effective height of the hill, and so the impedance of the contact point, are reduced by the electric field across the layer and barrier which acts in such a direction as to pull electrons from the electrode. The effect is to increase the flow of electrons into the semiconductor in a way which is similar to the enhancement of current from a thermionic cathode by field-induced emission. When the emitter is connected, and a current of holes flows to the collector, the accumulation of the positive charges of these holes in the vicinity of the collector tends to make the potential fall more rapidly with depth into the material, and so results in an increase in field and a decrease in the effective height of the hill, i. e., in the impedance of the contact point. Thus any increase in that component of the collector current which originates at the emitter is accompanied by a corresponding increase in the other component of the collector current, namely, in the flow of electrons to the base electrode. Hence the total change in collector current may be greater than the change in the emitter current.

From the foregoing description it will be clear that if it is desired to employ a semiconductor block of which the main body is of P-type so that the conductivity of the thin surface layer, whether due to impurities or to space charge effects, is of N-type, the polarities of all the bias sources of Figs. 1, 10, 11 and 12 are to be reversed. It is also to be understood that the magnitudes of the biases for best operation will depend on the semiconductor material employed and on its heat treatment and processing. Furthermore, it is possible to use a P-type layer of one semiconductor material on an N-type body of some other semiconductor material or vice versa. All such variations are contemplated as being within the spirit of the invention.

The invention is not to be construed as limited to the particular forms disclosed herein, since these are to be regarded as illustrative rather than restrictive.

What is claimed is:

1. A circuit element which comprises a block of semiconductive material of which the body is of one conductivity type and a thin surface layer is of the opposite conductivity type, an emitter electrode making contact with said layer, a collector electrode making contact with said layer disposed to collect current spreading from said emitter electrode, and a base electrode making contact with the body of the block.

2. Apparatus as defined in claim 1 wherein the surface layer is of the same chemical material as the block.

3. Apparatus as defined in claim 1 wherein the block is of germanium.

4. Apparatus as defined in claim 1 wherein the block is of N-type germanium and the surface layer is of P-type germanium.

5. Apparatus as defined in claim 1 wherein the block is of high back voltage germanium and to at least a part of one surface of which an anodic oxidation treatment has been applied.

6. A circuit element which comprises a semiconductive supporting body, a thin surface layer of semiconductor material supported by and in electrical contact with said body and differing in conductivity type therefrom, an emitter elec-