# RCA REVIEW

*a technical journal*

RADIO AND ELECTRONICS
RESEARCH • ENGINEERING

VOLUME XVIII DECEMBER 1957 NO. 4

PRINTED IN U.S.A.

RCA REVIEW, published quarterly in March, June, September, and December by RCA Laboratories, Radio Corporation of America, Princeton, New Jersey. Entered as second class matter July 3, 1950 at the Post Office at Princeton, New Jersey, under the act of March 3, 1879. Subscription price in the United States and Canada; one year $2.00, two years $3.50, three years $4.50; in other countries: one year $2.40, two years $4.30, three years $5.70. Single copies in the United States, $.75; in other countries, $.85.

# RCA REVIEW

*a technical journal*

RADIO AND ELECTRONICS
RESEARCH • ENGINEERING

*Published quarterly by*

RCA LABORATORIES

*in cooperation with all subsidiaries and divisions of*

RADIO CORPORATION OF AMERICA

VOLUME XVIII | DECEMBER, 1957 | NUMBER 4

*CONTENTS*





**RCA REVIEW** is regularly abstracted and indexed by *Industrial Arts Index Science Abstracts* (I.E.E.-Brit.), *Electronic Engineering Master Index, Chemical Abstracts, Proc. I.R.E.*, and *Electronic & Radio Engineer.*

# RCA REVIEW

## BOARD OF EDITORS

# NEW APPROACHES TO THE AMPLIFICATION OF MICROWAVES*

By

James P. Wittke

RCA Laboratories,
Princeton, N. J.

***Summary***—*Two types of "molecular" microwave amplifiers—the maser and the parametric amplifier—are described. The maser amplifies by converting internal energy of molecules to radiation energy. Parametric amplification is based on the "negative resistance" introduced into a circuit by a nonlinear reactance when driven by a strong high-frequency signal. The basic principles governing the operation of these amplifiers are discussed. Both types of amplifier have relatively narrow bandwidths, but have excellent noise properties. This low-noise aspect of their behavior will probably determine their application in the near future.*

## Introduction

TWO new types of solid-state microwave amplifiers—masers and parametric amplifiers—are considered in this paper. Both of these operate on principles of amplification basically different from those used in the more conventional devices such as klystrons, magnetrons, and traveling-wave tubes. In conventional *electron* tubes, microwave energy is obtained by the direct conversion of kinetic or electrostatic potential energy of the electrons in an interaction between the microwave field and moving charges. In the new amplifiers, this interaction is absent or only manifests itself indirectly. Instead, the interaction is between uncharged molecular matter and the field. For this reason, these devices are termed molecular, rather than electronic, amplifiers. This paper gives a physical picture of how these devices work, rather than a formal analysis of their operation.

## Masers

The first type of molecular amplifier to be considered is the maser. Certain aspects of the physics of molecules must be understood before its mode of operation can be explained. Consider an isolated molecule of a gas. Even the simplest of these is quite a complex structure

* Presented at the National Electronics Conference, Chicago, Ill., October, 1957.

consisting of various atoms, which in turn consist of negative electrons bound by electrical forces to positively charged atomic nuclei. When put together in a molecule, these tiny fragments do not form a rigid structure, but one that can vibrate, rotate, and perform other relative motions about its center of mass. There is a certain amount of internal energy associated with these motions, and it is on this internal energy that maser action is based. Experiments indicate that these internal molecular motions have the following interesting property: if the internal energy of a molecule is measured at any time, it will *always* be found to be one of an (infinite, perhaps) set of *discrete* energies. This situation is schematically indicated in Figure 1. Thus a molecule may be found with an internal energy $W_1$ or $W_2$, but will never be found with any intermediate energy $W$, where $W_2 > W > W_1$. (This can be understood only on the basis of quantum mechanics.)

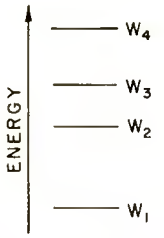



Fig. 1—Schematic representation of energy states (or levels).

An isolated molecule, then, has discrete internal energy levels. If many such molecules are brought together to form a solid, the strong interactions between the molecules in a solid will tend to alter the situation, turning the series of discrete levels into a quasi-continuum of very closely spaced levels. This is illustrated in Figure 2. Despite this, it is possible to find elements which, in certain solids, can still have discrete energy levels associated with them. One example of great practical importance is connected with electronic paramagnetism. An electron is of itself paramagnetic; that is, it acts like a small permanent magnet. This behavior is intimately connected with its *spin,* or angular momentum, which every electron also intrinsically possesses. An electron behaves as if it were a charged ball, continuously spinning about an axis through its center, and acting like a magnet because of the current loops of the rotating charge.

Consider the electrons in a solid. In general, they are "paired." This means that for every electron spinning in one direction there is another spinning in the exactly opposite direction. The magnetic dipole

moments of such a pair of electrons cancel each other out, leaving no net magnetic effect. In special solids, however, this pairing of electron moments is incomplete, and elements in the crystal (ions), fixed in location, can be found with unpaired electrons localized on them. When such a paramagnetic ion is placed in an applied magnetic field, it has an orientational energy depending on the angle between the magnetic dipole of the unpaired spins and the applied field. This is analogous to the energy of a compass needle in an applied field. The lowest energy state which the needle, if unrestrained, will assume is one of alignment with the field. It takes energy to displace the needle from this orientation, and this energy is stored as interaction energy between the field of the compass needle and the applied field. This energy is just $W = -\mu H \cos\beta$, where $\mu$ is the dipole moment of the spins (or compass

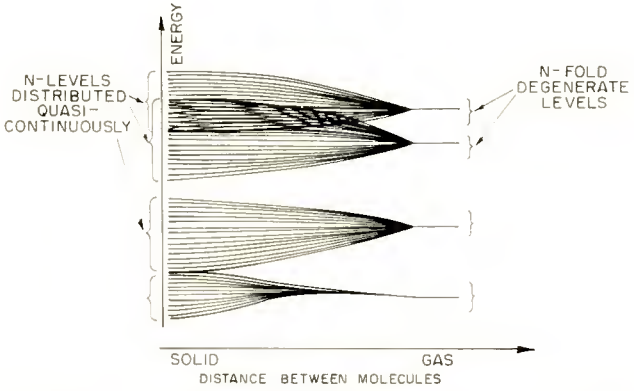



Fig. 2—Conversion of discrete energy levels into energy bands as isolated molecules are brought together to form a solid.

needle), $H$ the applied field and $\beta$ the angle between them. Not all orientational energies are possible for the spins, however; again, only certain discrete (quantized) energies can be found in a measurement. If the ions are sufficiently separated from each other by the nonmagnetic constituents of the crystal, so that their interactions with each other are weak, the result is a high concentration of elements having sharp energy levels, despite the fact that the elements are imbedded in a solid.

The next consideration is how such ions interact with microwave radiation. Not all microwave fields will interact with a given system; the frequency of the field must be proper. The interacting frequencies are related to the energy levels by the Bohr frequency condition, $f = (W_m - W_n)/h$. Here $(W_m - W_n)$ is the difference in energy be-

tween two internal energy states, $f$ is the microwave frequency, and $h$ is Planck's constant. Energy differences corresponding to visible light are a few electron-volts in magnitude. Infrared frequencies range from about one to about $10^{-3}$ electron-volt. The energy associated with a microwave transition is even smaller, typically only $10^{-4}$ to $10^{-5}$ electron-volt.

There are three possible forms which the interaction between ions and field can assume: spontaneous emission, absorption, and induced (stimulated) emission. Spontaneous emission is the radiation emitted (in the absence of an applied stimulating field) by an ion excited to a state of high internal energy. This radiation causes the familiar glow of a gas discharge, where swiftly moving electrons collide with, and thus excite, gas atoms and molecules. These excited molecules spontaneously drop to lower energy states, giving up their excess internal energy as light. This tendency to radiate spontaneously falls off rapidly as the excess internal energy, and hence the frequency of the radiation, is reduced; in the microwave region, spontaneous emission is generally completely negligible.

In absorption, the second form of field–ion interaction, an ion in the lower ($W_n$) of a pair of energy levels is subjected to a microwave field of frequency $f = (W_m - W_n)/h$. The ion absorbs an amount $(W_m - W_n)$ of energy from the field and jumps to the upper energy state $W_m$. Absorption is also a familiar process in the optical frequency range. An example is given by the blue of a copper sulfate solution. Here the copper ion absorbs red light and jumps to an excited electronic state. With red light thus selectively absorbed by the solution, only the blue portion of the spectrum can be transmitted.

Induced emission is just the reverse of absorption. An applied field of frequency $f = (W_m - W_n)/h$, acting on an ion in the upper state ($W_m$), *induces* it to give up its excess internal energy to the field and drop to the lower energy state. The internal energy thus lost by the ion to the field is coherent with the inducing field.

The simple mechanical analog of Figure 3 may make these processes clearer. A quantity of sand represents the solid, each grain representing an individual ion. Two platforms, one above the other, represent two energy levels, the potential energy of the sand standing for the internal energy of the ions. Thus, sand grains on the top platform represent ions in the higher energy state, while sand grains on the lower represent low-energy ions. A conveyer wheel carries sand between the two levels. In this model, spontaneous emission is represented by a tendency of sand to fall off the upper platform to the lower without using the conveyer wheel. This process will be neglected in

the following discussion. The wheel is coupled to a dynamotor (the microwave field). Ions (sand grains) in each energy state (on each platform) are continuously fed onto the conveyor; the ions delivered to one or another energy state immediately start their return journey, as long as the field is applied. The microwave field (the dynamotor) continuously transfers ions from one state to the other. If originally all the ions (sand grains) are in the lower level, absorption of energy from the microwave field (the dynamotor) occurs. If they are instead originally in the upper level, the internal energy of the ions (potential energy of the sand) can be converted to energy of the microwave field. (The sand on the upper platform in coming to the lower runs the dynamotor as a generator.)

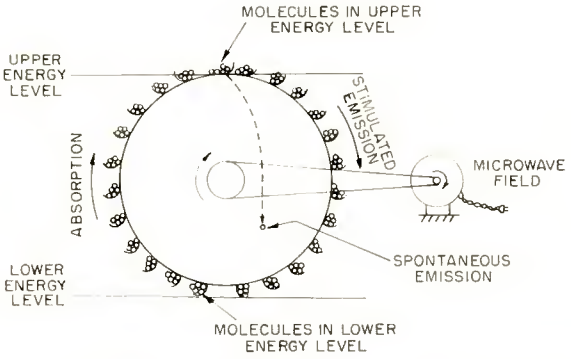



Fig. 3—Mechanical model illustrating the interaction between molecules and a microwave field. The sand grains represent molecules, the platforms, energy levels, and the conveyer wheel driven by the motor, the effect of the microwave field.

This sand-and-conveyer-wheel analog is crude. It does, however, illustrate some important points. One is that if the two levels are equally populated there is no net emission or absorption; no energy is transferred to or from the dynamotor when equal amounts of sand are being raised and lowered.

With this brief discussion of the different interactions between ions and field, the origin and meaning of the term "maser" can be understood. It is made up of the first letters of the phrase "Microwave Amplification by Stimulated Emission of Radiation." The amplification comes from the stimulated (induced) release of internal energy from the ions or molecules.

All molecular systems, when allowed to come into thermal equilibrium with their surroundings, have more molecules in the lower of

each pair of energy states than in the upper, and hence are absorptive. To get a net emission of energy, and hence amplification, more molecules must be in the upper of two levels. This is the basic requirement for maser action. Fulfilling this requirement is the primary problem of building a maser amplifier. The active molecules in a maser must be in a nonequilibrium state where, for the two levels involved, more molecules are in the upper than in the lower state. It is clear that energy must be put into the ionic system by the excitation mechanism to produce such a state. This is the energy source for the amplifier. There are several ways of obtaining a suitable emissive state. Before considering these, however, two other effects should be mentioned. These are the relaxation and the saturation effects.

A system in equilibrium is absorptive; to obtain an amplifier, the ions must somehow be excited to a nonequilibrium state. If the excitation mechanism setting up this emissive state is turned off, the system will return to the absorptive state characterizing equilibrium by several processes. These are known as the relaxation processes. They are generally nonradiative; that is, they result in the conversion of excess internal energy into heat, rather than into electromagnetic radiation. Frequently, the return to an equilibrium state follows an exponential law. The time constant describing the decay of the excess internal excitation is called the *relaxation time.** The relaxation time in a molecular system is analogous to the logarithmic decrement characterizing a tuned circuit. If the drive on a tuned circuit is suddenly removed, the amplitude of oscillations will decay at a rate determined by the Q of the circuit; this decay time corresponds to the molecular relaxation time.

The relaxation time of the solid used imposes a restriction of great practical significance; the excitation process must be such as to repeat the excitation of each ion within a relaxation time. With very short relaxation times, (less than a microsecond, for example), no suitable excitation process has thus far been devised. This severely limits both the number of materials that can be used to build masers, and the operating conditions.

The effect of a microwave field (of proper frequency) on a lower state ion is to raise it to the upper state, during which process it absorbs energy from the field. The effect on an upper state ion is the

* Two distinct relaxation times are recognized: the one discussed above, known as the spin-lattice relaxation time $T_1$, and a time to lose phase-coherence between the interactions of the various molecules and the field, known as the spin-spin relaxation time $T_2$. For the purposes of this paper, this distinction represents a needless complication, and the concepts will be referred to interchangeably as the "relaxation time".

reverse. Why, then, do the ions not continuously oscillate back and forth between the two states, first absorbing, then emitting energy? In our sand-and-conveyer model, will not the sand keep shuttling back and forth between the platforms, with no net energy exchange? The answer is that, except for excitation and relaxation processes, this would indeed be the case. However, if the excitation process puts ions into the upper state while they are still emitting after the previous excitation and before they have become absorptive, a net emission of energy can take place and hence amplification can be obtained. (If sand is lifted off the conveyer wheel and put on the upper platform before it reaches the lower platform, on its downward journey, it will transfer net energy to the dynamotor.)

The rate at which ions tend to oscillate back and forth between the levels depends on the strength of the microwave field. On exact resonance, the oscillation frequency (the speed of the conveyer wheel) is proportional to the amplitude of the microwave field. If the field is strong, the tendency to pass through the emissive part of the cycle and start absorbing is greater than it is for weak fields. It can be shown that the net effect of this, when a large number of ions is considered and excitation and relaxation processes are taken into account, is to limit the total emitted power as the inducing power is raised. In usual terminology, the amplifier *saturates*. While the power level at which a maser saturates depends on the ionic system and on the microwave structure containing it, the saturating power is typically well under a watt, and in extreme cases may run as low as $10^{-10}$ watt. From this it is clear that these devices are low power ones. This is no real limitation on their use as exceedingly low-noise amplifiers, however, for such amplifiers are useful when the signal is not much out of noise, and hence very weak.

As mentioned earlier, several excitation means have been devised. However, only one method of exciting a solid system of paramagnetic ions will be described here.[1] Consider a solid with ions having only three possible (orientational) energies in a magnetic field, as shown in Figure 4. Here the molecular electric fields found in the crystal have distorted the levels so that they are not equally spaced as would be expected for a free ion. (This distortion is essential for the operation of this excitation method.) At thermal equilibrium, $N_1 > N_2 > N_3$, and the system is absorptive at all the three frequencies, $f_{12}$, $f_{23}$, and $f_{13}$. However, this situation is altered if a strong microwave field is applied at the frequency $f_{13}$. With such a strong field, the ions in

[1] N. Bloembergen, "Proposal for a New Type Solid State Maser," *Phys. Rev.*, Vol. 104, p. 324, October, 1956.

states 1 and 3 oscillate back and forth between the two levels faster than the relaxation processes can restore them to the equilibrium condition. With the $f_{13}$ resonance thus saturated, the populations of the two states 1 and 3 are approximately equalized: $N_1' = N_3'$. In general, this process, which pumps considerable energy into the ionic system, will also disturb somewhat the population of the intermediate state; $N_2' \neq N_2$. The saturating field therefore acts to establish a new set of level populations $N_1' = N_3'$, $N_2'$. This is a steady-state, although nonequilibrium, condition.

Two possibilities exist:

$$(1) \quad N_1' = N_3' > N_2'.$$

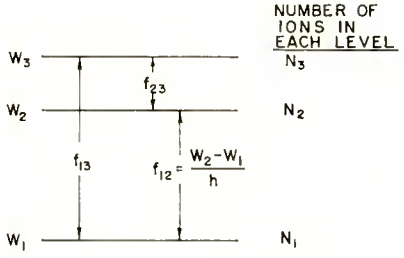



Fig. 4—Energy-level diagram illustrating the three-level method of exciting molecules into an emissive condition.

In this case it is seen that since $N_3' > N_2'$, a net emission of energy will occur upon application of a microwave field at a frequency $f_{23}$. An amplifier can therefore be built operating at this frequency.

$$(2) \quad N_1' = N_3' < N_2'.$$

With this alternative, $N_2' > N_1'$ and amplification can be achieved at the frequency $f_{12}$. Thus, by saturating the resonance at the frequency $f_{13}$, amplification can be obtained at one of the two frequencies $f_{12}$ or $f_{23}$.

The choice of material to use in a maser amplifier depends on many factors. The energy levels of some systems are essentially independent of applied magnetic (or electrostatic) fields. In such a case the operating frequency is fixed by the choice of the material. Certain excitation methods only work with certain types of molecular systems, and this factor may dominate the choice. In the three-level scheme described, it is clear that the levels must be unequally spaced, or the

emission between the intermediate level and the appropriate outer level will be cancelled out by the absorption between the intermediate level and the other outer level. In practice it may also prove impossible to saturate the high-frequency resonance, either because the electromagnetic coupling between the levels is too weak, or because the competing relaxation processes are too strong to be overcome with available microwave power sources. The effect of the crystalline field distortion of the energy-level positions must also be considered in determining the resonant frequencies. Because of the complexity of these requirements on the molecular system in a maser, only a relatively few crystals are suitable for use. Extensive investigations are now under way in several laboratories in an attempt to discover better materials.

One important practical aspect of present solid-state paramagnetic-ion masers should be mentioned. All methods of exciting these materials require relatively long relaxation times—of the order of $10^{-4}$ to $10^{-3}$ seconds or longer. At present, such relaxation times can only be attained by lowering the temperature of the crystal to that of liquid helium. While from a noise point of view this is desirable, there can be no question that the necessity of maintaining the crystal at liquid helium temperatures poses severe practical problems. Studies to find crystals with suitable properties at higher temperatures are also being undertaken.

It is now of interest to consider how a suitable excited solid can be used to construct a practical amplifier. Several methods of building a maser suggest themselves. One is to put the active material into a waveguide, with suitable matching elements. When the signal to be amplified passes down the guide it induces the active material to give up its internal energy, producing a growing wave of much the same type as in a traveling-wave tube. While theoretically possible, there are several practical difficulties with such a traveling-wave maser. With many molecular systems, including all gaseous ones, the gain per unit length along such a guide is very low, perhaps 1.001 per centimeter. Therefore, to get useful gains an amplifier several meters long is required. Other similar practical difficulties tend to make use of this traveling-wave approach equally unfeasible for solid systems.

Another way to use an emissively excited material is in a resonant cavity. By thus concentrating the microwave fields the need for a large structure is eliminated. However, by putting the emissive system in a cavity, positive feedback is automatically introduced, making the amplifier regenerative. As is well known, it is easy to obtain high gains with such an amplifier, but positive feedback has the concurrent disadvantage of reducing the amplification bandwidth. In masers this

can be quite serious. With low available molecular powers, a large amount of feedback is necessary for useful gain. This means that one operates near the oscillation point and consequently runs the risk of some instability. Nevertheless, this seems to offer the best available approach at present, and with improved materials and excitation means, the difficulties will undoubtedly be overcome.

One such device that has been built has been made to work as an oscillator.[2] The crystal used was one of hydrated lanthanum ethyl sulfate. One half of one per cent of the nonmagnetic lanthanum ions

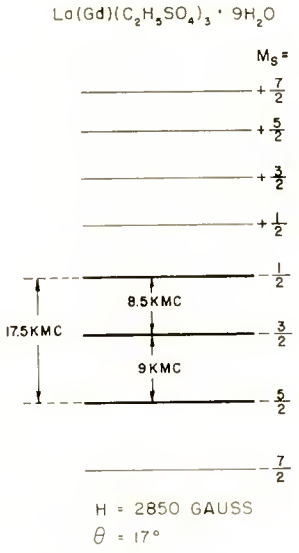



Fig. 5—Energy levels of gadolinium ions in a lanthanum ethyl sulfate crystal.

were replaced by paramagnetic gadolinium ions. These have eight possible orientational energy levels, as shown in Figure 5. Cerium ions were also added to the crystal. These changed the relaxation times in a useful way without altering the energy levels. From Figure 5 it is seen that three of the energy levels are separated by amounts corresponding to $f_{12} = 9 \times 10^9$ and $f_{23} = 8.5 \times 10^9$ cycles. (These separations occur in a field of 2,850 gauss with a suitable crystal orientation.) Thus $f_{13} = 17.5 \times 10^9$ cycles. The crystal was placed in a microwave cavity resonant at both $9 \times 10^9$ and $17.5 \times 10^9$ cycles. The cavity and crystal were immersed in liquid helium and placed between the poles of an

[2] H. E. D. Scovil, G. Feher, and H. Seidel, "Operation of a Solid State Maser," *Phys. Rev.*, Vol. 105, p. 762, January, 1957.

electromagnet that supplied the d-c field. A strong microwave signal at $17.5 \times 10^9$ cycles was coupled into the cavity. Under these conditions, the intermediate level was more highly populated than the lower level, and hence molecular power was available at $9 \times 10^9$ cycles.

By increasing the strength of the saturating signal, the excess population in the intermediate state increased until the power available from the crystal spin system equalled the power losses to the cavity walls and out of the waveguide coupling at $9 \times 10^9$ cycles. At this point the amplifier broke into oscillation, the strength of which increased as the pumping power was raised. Fifteen microwatts of power at $9 \times 10^9$ cycles was observed with 200 milliwatts of $17.5 \times 10^9$ cycles pumping power. At pumping powers between about 60 milliwatts and 95 milliwatts, (the threshold of oscillation), the molecular

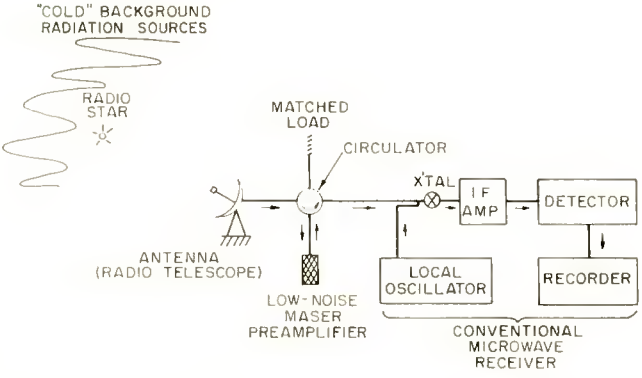


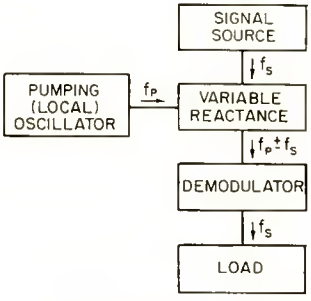

Fig. 6—Schematic representation of a possible use of a maser amplifier in radio astronomy.

power was more than enough to compensate wall losses in the cavity, but not enough to compensate both wall losses and losses to the coupled waveguide. Under such conditions the device amplified signals incident on the cavity. Figure 6 shows the way such a device could be used in an actual microwave system.

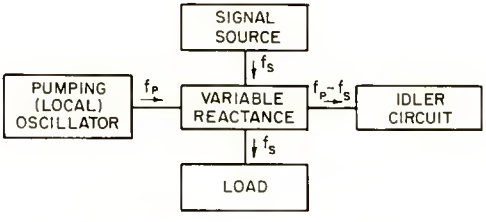
## Parametric Amplifiers

Another class of molecular device is the parametric amplifier. Unfortunately, this class of amplifier does not lend itself as readily to simple mechanical analogy as the maser, and so an analytical rather than a descriptive approach will be used. A parametric amplifier utilizes a variable reactance to obtain gain, and while it thus resembles a magnetic or dielectric amplifier, its principle of operation is somewhat

different. Briefly, a magnetic or dielectric amplifier is a modulator with gain; there is more power in the high-frequency sidebands about the local oscillator frequency than in the signal applied to the variable reactance to generate the sidebands (See Figure 7a). The (low-frequency) signal acting on the variable reactance modulates the strong high-frequency pumping (local oscillator) signal to produce relatively high-power sidebands. However, to get an amplified signal out at the




(a) MAGNETIC OR DIELECTRIC AMPLIFIER




(b) PARAMETRIC AMPLIFIER

Fig. 7—(a) Block diagram of a magnetic or dielectric amplifier, with the frequency spectrum and power flows indicated. (b) Block diagram of a parametric amplifier, showing the frequency spectrum and power flows.

same frequency as the input, a demodulator must also be used. In the parametric amplifier, on the other hand, power supplied by the local oscillator is directly converted to signal power *at the signal frequency.* In both types of devices a high-power high-frequency local oscillator supplies the energy needed for amplification.

To understand the operation of a parametric amplifier, and to see more clearly wherein it differs from a magnetic amplifier, the properties of a variable reactance must be examined. Manley and Rowe[3]

[3] J. M. Manley and H. E. Rowe, "Some General Properties of Nonlinear Elements—Part I. General Energy Relations," *Proc. I.R.E.*, Vol. 44, p. 904, July, 1956.

have discussed the energy relationships between the various frequency components in such an element. (It is assumed in the following that the nonlinear reactance characteristic is single-valued.) Assume that power can flow in the variable reactance only at (at most) four frequencies: a signal frequency $f_0$, a higher local oscillator frequency $f_1$, and the two sideband frequencies $f_+ = f_1 + f_0$ and $f_- = f_1 - f_0$ about the local oscillator frequency. This means that all other frequencies see an open (or short) circuit across the reactor. In this case, Manley and Rowe show that the following expressions hold:

$$W_1 = f_1 \left[ \frac{(-W_+)}{f_+} + \frac{(-W_-)}{f_-} \right],$$

$$W_0 = f_0 \left[ \frac{(-W_+)}{f_+} - \frac{(-W_-)}{f_-} \right].$$

Here $W_1$ and $W_0$ are the powers flowing *into* the variable reactance at the frequencies $f_1$ and $f_0$, respectively. $(-W_1)$ and $(-W_0)$ therefore represent power flow from the reactance into the associated circuit.

First these expressions will be applied to the case of a magnetic (or dielectric) amplifier. It can readily be seen that the local oscillator always supplies power to the reactance $(W_1 > 0)$ when there are passive loads coupled to the sideband frequencies $(W_\pm < 0)$. The signal source, on the other hand, can either supply or absorb power, depending on whether $\frac{|W_+|}{f_+} > \frac{|W_-|}{f_-}$ or the reverse. Generally both local oscillator and signal source supply power which is converted to relatively high-level sideband power. However, the fact that under certain conditions power can be given up to, rather than taken from, the signal source presents the possibility of instability. This explains certain spurious outputs that have been observed from magnetic amplifiers with only local oscillator power supplied. It also explains the basis of parametric amplification.

Parametric amplification makes use of this instability, which effectively adds "negative resistance" to the signal circuit, to develop gain directly in the signal circuit. From the Manley–Rowe expressions given above, it is clear that one can maximize the power into the signal circuit $(-W_0)$ by minimizing the power into the upper sideband frequency circuit $(-W_+)$. If the upper sideband is completely suppressed, the equations become

$$W_1 = \frac{f_1}{f_-}(-W_-)$$

$$W_0 = -\frac{f_0}{f_-}(-W_-).$$

These expressions indicate that the local oscillator supplies power both at the lower sideband frequency $f_-$ and at the signal frequency $f_0$. The size of the negative resistance thus introduced into the signal circuit can be shown to increase with applied local oscillator power. If it is larger than the positive resistance introduced by inevitable losses in the coupling to the (lossless) variable reactance, amplification at the signal frequency results. If the negative resistance is made even greater, so that it exceeds in magnitude the total positive resistance of the signal and reactance circuits, the system will oscillate and become a signal generator.

It is important to note the role of the difference-frequency sideband. If this too is suppressed by open-circuiting, no power can flow into the variable reactance at any frequency. In general, it is impossible to have power at only two incommensurable frequencies coupled through a lossless variable reactance. The circuit at the "idler" frequency must be able to absorb energy from the system if amplification is to be obtained.

Recently, methods have been proposed to secure amplification in the microwave frequency range using this general principle.[4] To understand these, the "magnetostatic modes" of a ferrite must be considered. A ferrite crystal contains a very large number of unpaired electrons that are tightly coupled to each other by a strong exchange interaction. In an applied magnetic field, all of these spins can precess about the field as a unit. However, the coupling between the spins is not so strong as to remove the many degrees of freedom inherent in a system of a large number of elements, and other motions, which can be considered as disturbances on the uniform magnetization, are also possible. These are the magnetostatic modes.[5] These modes are distributed quasi-continuously in frequency. The lowest possible frequency depends upon the shape of the ferrite sample, but is always equal to, or greater than, $\gamma(H - 4\pi M)$. Here $\gamma$ is the gyromagnetic ratio for the

[4] H. Suhl, "Proposal for a Ferromagnetic Amplifier in the Microwave Range," *Phys. Rev.*, Vol. 106, p. 384, April, 1957.

[5] L. R. Walker, "Magnetostatic Modes in Ferromagnetic Resonance," *Phys. Rev.*, Vol. 105, p. 390, January, 1957.

ferrite, $H$ is the applied field, and $M$ is the saturation magnetization of the ferrite. One of the magnetostatic modes is the familiar uniform precession mode at frequency $\omega = \gamma H$. A detailed analysis of the system[6] shows that there is coupling between the uniform precession mode and the other magnetostatic modes. Each mode of the ferrite represents a tuned circuit coupled to the reactance. Instead of couplings at only three frequencies, as assumed above, there are now a huge number of them. If the pumping (local oscillator) frequency is chosen as $\omega_p = \omega$, the uniform precession frequency, then a large number of different pairs of magnetostatic modes will have frequencies adding up to $\omega_p$. These pairs can therefore act as effective signal and idler circuits at various frequencies $\omega_s$ and $\omega_p - \omega_s$. A load can only be coupled at best to a small number of these modes, and so the power supplied at the pumping frequency would be dissipated as heat in the many load-uncoupled circuits, preventing practical use of this "magnetostatic" type of operation in a parametric amplifier.

There are, however, other possible ways to use a ferrite in such an amplifier. One is the "semistatic" mode of operation. In this, the applied magnetic field $H$ is adjusted to be greater than $8\pi M$. (It is assumed that the lowest magnetostatic mode is at a frequency $\gamma(H - 4\pi M)$.) Now no two magnetostatic modes have frequencies whose sum is equal to the precession frequency $\omega_p = \omega$. The ferrite sample is placed in a resonant cavity tuned so that it constitutes either the signal frequency circuit at $\omega_s$ or the idler circuit at $\omega_p - \omega_s$, and one magnetostatic mode is used to supply the other needed circuit at $\omega_p - \omega_s$ or $\omega_s$, respectively.

In the third type of operation, the "electromagnetic," a doubly resonant cavity is made to supply both low-frequency circuits, and so no magnetostatic modes other than the uniform precession one are used. These three cases are represented in Figure 8. In Figure 8a, the magnetostatic type of operation, the cross-hatched region represents the frequency range in which the large number of pairs of low-frequency modes can be found. In Figure 8b, the semistatic case, the region in which the cavity mode must lie, as well as the accompanying range of magnetostatic modes, is shown. The electromagnetic case is shown in Figure 8c, where the cross-hatched region indicates where the two cavity modes must lie.

Operation of this type of ferromagnetic parametric amplifier is also possible under conditions where the pump frequency, $\omega_p$, is different

[6] H. Suhl, "The Nonlinear Behavior of Ferrites at High Microwave Signal Levels," *Proc. I.R.E.*, Vol. 44, p. 1270, October, 1956.

from the uniform precession frequency, $\omega$. However, the pumping power required in such a case is higher than in the resonant case, and hence will in general be less practical. Although the above discussion has been confined to the ferromagnetic case, it is clear that the general amplification principle can be applied to variable capacitances. Reverse-biased semiconductor junctions can provide an element that will operate in this fashion at microwave frequencies.[7]

The ferromagnetic type of parametric amplifier has been operated as an oscillator and amplifier at $4.5 \times 10^9$ cycles.[8] A strip-line resonator was used, with coupling by means of coaxial line probes. To simplify the circuit, the signal and idler frequencies were made degenerate.

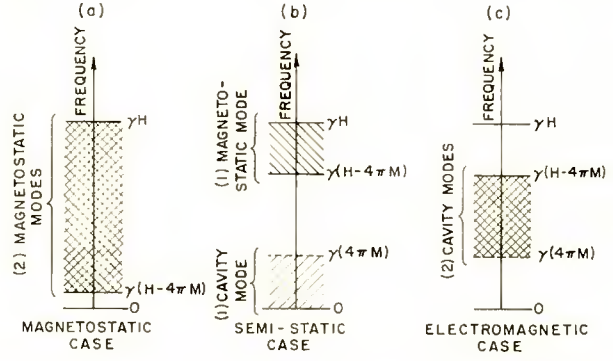



Fig. 8—Schematic representation of the frequency and circuit relations in the three types of ferromagnetic parametric amplifier: (a) magnetostatic mode; (b) semistatic mode; (c) electromagnetic mode.

The local oscillator power at $9 \times 10^9$ cycles was supplied along a waveguide to the ferrite, which was located at the center of the strip-line resonator. To avoid overheating the system, short pulses of local oscillator power were used. With a peak pumping power of 20 kilowatts, the device oscillated at a level corresponding to 100 watts output. By reducing the pumping power about one decibel below the oscillation threshold, an amplifier gain of 8 decibels was observed.

## Discussion

The scope of this paper has been limited to a brief description of the way in which masers and parametric amplifiers work. Important

---

[7] L. J. Giacoletto and J. O'Connell, "A Variable-Capacitance Germanium Junction Diode for UHF," *RCA Review*, Vol. XVII, p. 68, March, 1956.

[8] M. T. Weiss, "A Solid-State Microwave Amplifier and Oscillator Using Ferrites," *Phys. Rev.*, Vol. 107, p. 317, July, 1957.

aspects of these amplifiers as practical devices, such as bandwidths, power handling capabilities, and noise properties, have been ignored or only mentioned briefly. As both types of amplifiers act in a regenerative fashion, gain is no problem in practical embodiments. Bandwidth, on the other hand, is a problem: for these devices, in addition to the loss of bandwidth brought about by regeneration, have a relatively narrow frequency response. Present masers have bandwidths of the order of one megacycle or less (drastically less in some cases), and although there is hope of raising this figure to one hundred megacycles, much work remains to be done before such a figure is actually achieved. Nothing has been published on parametric amplifier bandwidths, but the situation is probably quite similar. It is in regard to noise properties that these new amplifiers are outstanding. Very few measurements have been made, but noise figures as low as 1.0 db[9] have been obtained with masers, and it is expected that parametric amplifiers will also have better noise properties than conventional electronic tubes. Masers have low power handling capabilities, but as mentioned, this is not too serious a matter in certain applications. Parametric amplifiers are much better than masers in this respect.

From the above it is clear that molecular amplifiers will not replace electron tubes in microwave circuits, at least not for a long time. It is their low noise figures that indicate where these new devices will first find application. Wherever weak signals need to be picked out of background noise, where the relatively narrow bandwidths can be tolerated, molecular amplifiers can be of great use. They should be particularly valuable in picking out signals where the background noise is characteristic of very low noise temperatures, such as radar signals with the antenna mainly picking up noise from a very "cold" sky.

---

[9] J. P. Gordon and L. D. White, "Experimental Determination of the Noise Figure of an Ammonia Maser," *Phys. Rev.*, Vol. 107, p. 1728, September, 1957.

# ELECTRON MOBILITY IN THE GERMANIUM-SILICON ALLOYS

By

B. Goldstein

RCA Laboratories,
Princeton, N. J.

***Summary***—*This paper describes a method for measuring the conductivity mobility of a semiconductor, and an application of this method to the measurement of the majority carrier mobility in the germanium–silicon alloy system.*

*The method involves the determination of charge carrier densities by means of the voltage dependence of a p-n junction transition capacitance. This technique is particularly suitable to materials with charge-carrier lifetimes too small for drift experiments.*

*Applying the method to the germanium–silicon alloy system, a similar qualitative behavior has been found for the conductivity and Hall mobilities, as changes are made in alloy composition. However, significant quantitative differences between the measured and calculated values of* $(\mu_H/\mu)$ *indicate that some of the assumptions customarily made in calculating scattering factors may not be sufficiently accurate.*

## Introduction

THE mobility of electrons (or holes) in a semiconductor in which the current is carried predominantly by electrons (or holes) is given by

$$\mu = \sigma/Ne, \tag{1}$$

where $e$ is the electronic charge, $\sigma$ is the conductivity and $N$ is the charge-carrier density. To determine the mobility, it is necessary to know the conductivity and the charge-carrier density. The former is usually easy to determine; however, the latter is not so directly determined. The Hall effect is usually employed to determine the charge-carrier density, but sometimes it is inappropriate because of a very high charge-carrier density or because the use of a magnet may be awkward. In the method described in this paper, the charge-carrier density is determined from the net space-charge density existing in the depletion layer of an abrupt p-n junction.

## Theory

Consider a metal–semiconductor contact of the type shown sche-

matically[1] in Figure 1. Electrons have been displaced from the conduction band of the semiconductor to the metal, warping the energy levels to make the Fermi levels in the metal and the semiconductor coincident, a condition necessary for thermodynamic equilibrium. There then exists in the semiconductor a layer immediately adjacent to the contact which has lost electrons and thus is positively charged. This extends a distance $x_0$ from the contact and is called the depletion layer. Similarly, the region in the metal immediately adjacent to the contact in which the electrons from the semiconductor have accumulated can be called an accretion layer. The extent of this accretion layer in the metal is negligible compared to the extent of the depletion layer in the semiconductor. The potential and space-charge distribu-

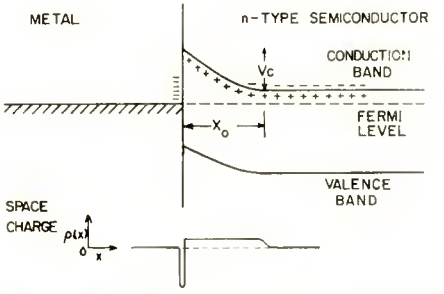



Fig. 1—Energy diagram of metal–semiconductor contact. Shown in the diagram are the Fermi level, the contact potential, $V_c$, and the extent of the depletion layer, $x_0$, when no external voltage is applied across the contact.

tion after equilibrium has been established is shown qualitatively in Figure 1. Poisson's equation for the depletion layer can be written

$$\nabla^2 V(x) = -\frac{4\pi}{k}\rho(x), \qquad (2)$$

where $V$ is the potential, $k$ is the dielectric constant of the semiconductor, and $\rho(x)$ is the space-charge distribution. Since the depletion layer resides within the semiconductor, $\rho(x)$ is simply the density of immobile, positively charged donor atoms, $N$, which have contributed electrons to the conduction band. Assuming a homogeneous distribution of donors, Equation (2) becomes

[1] W. Schottky, "Semiconductor Theory of Blocking-Layer and Point Rectifier," *Z. fur Physik*, Vol. 113, p. 367, 1939.

$$\nabla^2 V(x) = \frac{4\pi}{k} Ne. \tag{3}$$

This is the value of $N$ required for Equation (1) since $N$ is also the number of charge carriers available for conduction, assuming that each donor atom can supply one electron.

The solution to Equation (3), with appropriate boundary conditions, may be written as

$$V_c + V_a = \frac{2\pi Ne\, x_a^2}{k}, \tag{4}$$

where $V_c$ is the contact potential, $V_a$ is any reverse voltage applied across the barrier, and $x_a$ is the extent of the depletion layer due to sum of $V_c$ and $V_a$. If we set up the ratio $\Delta Q/\Delta V_a$ to determine the dynamic junction capacitance, where $\Delta Q$ is the charge swept into (or out of) the depletion layer due to the application of $\Delta V_a$, then

$$C = \frac{\Delta Q}{\Delta V_a} = \frac{kA}{4\pi x_a}, \tag{5}$$

where $A$ is the effective junction area. Substituting this result into Equation (4) and rearranging terms,

$$N = \frac{8\pi}{keA^2} \frac{dV_a}{d(1/C^2)}, \tag{6}$$

in which $dV_a/d(1/C^2)$ and $N$ are evaluated at the edge of the depletion layer inside the semiconductor.* Thus, by plotting the voltage dependence of the junction capacitance in the form $V_a$ versus $1/C^2$, and measuring the slope of the resulting straight line, $N$ can be determined if the junction area and the dielectric constant of the semiconductor material are known.

If one extrapolates the straight line of the $V_a$ versus $1/C^2$ plot to the point where $1/C^2$ is zero (i.e., where $x_a = 0$), then the voltage intercept at this point should theoretically give $V_c$ the contact potential or intrinsic barrier height (see Equation (4)).[2] For the present

* Equations (5) and (6) are general expressions valid for p-n junctions with any arbitrary distribution of doping.

[2] H. Kroemer, "The Apparent Contact Potential of a Pseudo-Abrupt P-N Junction," *RCA Review*, Vol. XVII, p. 515, December, 1956.

purpose, however, only the slope of that part of the curve corresponding to high reverse voltage is of interest. At these high voltages the end of the depletion region is well into the bulk of the semiconductor, far from the junction. It is, of course, in the interior of the semiconductor that we wish to measure the charge carrier density, $N$.

## Application

To apply the method, pellets of indium and aluminum were alloyed onto germanium–silicon crystals to produce abrupt junctions. This method is preferred to electroplating or evaporating a metallic layer onto the semiconductor. With these other methods potential differences may occur due to the formation of surface oxide layers or inversion layers. Such potentials might interfere with the measurements. It has been shown that the Schottky theory holds for the abrupt alloy junction as well as for the abrupt metal–semiconductor junction.[3] The impurity concentration in the impurity saturated alloyed region is so much higher than it is in the semiconductor, that the depletion layer is almost entirely within the semiconductor.

Details of the experimental procedure are as follows. Single-crystal wafers of germanium–silicon alloy were first cleaned in CP4 etch.[4] Fifteen-mil-diameter pellets of pure indium for the high-germanium-content alloy, and of pure aluminum for the high-silicon-content alloy, were then fired onto the crystal wafers at 550°C and 750°C, respectively, for 10 minutes. These pellet materials were chosen for their good wetting and alloying properties. After firing, the wafers were cooled slowly. Contacts to the indium pellets were soldered; contacts to the aluminum pellets were made with silver paste. Base connections were soldered to areas of the crystal wafer which had been etched in hot potassium hydroxide and then copper plated. This technique was used to produce reliable ohmic contacts. The capacitance of the abrupt junction was then measured as a function of reverse bias voltage using a tuned circuit with a sensitivity of about 0.1 micro-microfarad. The junction bias voltage was isolated from the tuned circuit by a blocking condenser.

## Discussion of Errors

The first source of error to be considered is the estimation of the

[3] W. Shockley, "The Theory of p-n Junctions in Semiconductors and p-n Junction Transistors," *Bell. Sys. Tech. Jour.*, Vol. 28, p. 435, July, 1949.

[4] F. L. Vogel, W. G. Pfann, H. E. Corey, and E. E. Thomas, "Observations of Dislocations in Lineage Boundaries in Germanium," *Phys. Rev.*, Vol. 90, p. 489, May, 1953.

effective junction area. Note that in Equation (6) the area appears as $A^2$. The difficulty encountered is that the alloying of the indium or aluminum proceeds to such a depth that the side walls of the alloyed region contribute significantly to the junction area. This error is corrected by using information obtained from measurements of junction capacitance on material of known doping. Since the aluminum alloying procedure gives a similar junction depth, the same correction was applied. The uncertainty in area is estimated to be ±6 per cent. The conductivity, $\sigma$, was measured with an uncertainty of about ±5 per cent. Other, less important, sources of error are in the measure-

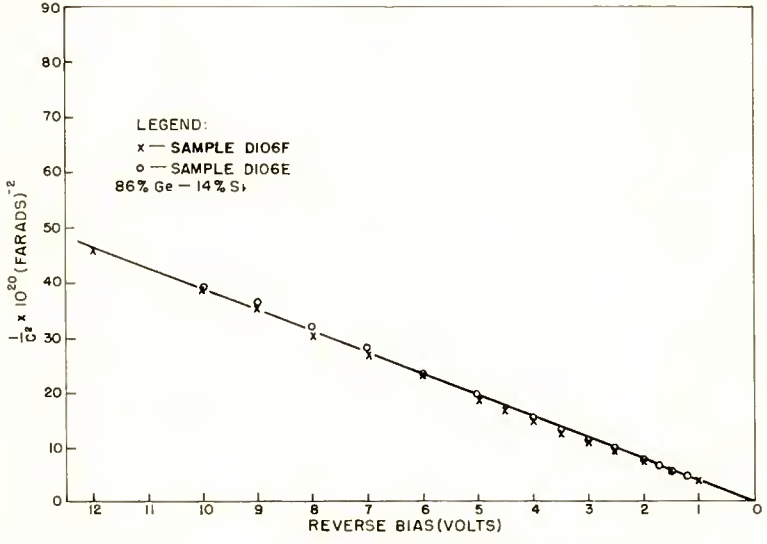



Fig. 2—Voltage dependence of junction capacitance.

ments of the capacitance and voltage. The estimate of the over-all uncertainty in the mobility measurement is about 10 per cent.

## Measurements

Typical $1/C^2$ versus $V_a$ data for abrupt junctions are shown in Figure 2. The measurements were made on adjacent slices of a crystal composed of 86 per cent germanium and 14 per cent silicon. These data demonstrate the reliability of the measuring system. From the slopes of the curves, the area of the junctions, and the dielectric constant,[5] charge carrier densities of 1.9 and $1.8 \times 10^{15}$ per cubic centimeter were

[5] The dielectric constants of the germanium–silicon alloy system were measured by R. Braunstein, RCA Laboratories.

calculated. After measuring the conductivities of these samples, the mobility was calculated with the use of Equation (1).

The mobilities calculated for electrons in n-type germanium–silicon alloys are plotted in Figure 3. The measurements were restricted to the range where single crystals could be grown. This range covers only up to 25 per cent silicon on the germanium side and 2 per cent germanium on the silicon side. The conductivity mobility as a function of alloy composition drops sharply as silicon is added to germanium, passes through a broad minimum, and then rises sharply

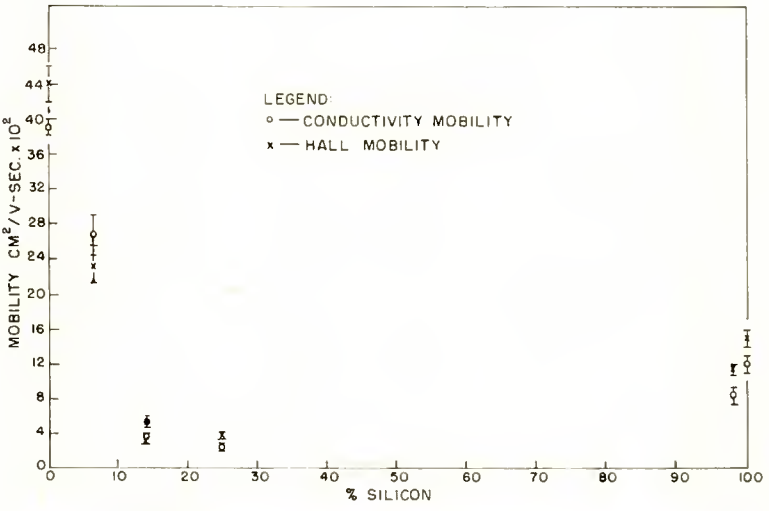



Fig. 3—Electron mobility in the germanium–silicon alloy system.

at the silicon end. The corresponding Hall mobilities, $R\sigma$, for which the same general behavior is found[6] are also plotted in Figure 3. The pertinent information relating to each of the samples measured is listed in Table I.

In Table I the values for pure germanium and pure silicon are taken from the literature.[7–10] As they refer to the germanium–silicon alloys, $\mu_H$ is the experimentally determined Hall mobility; $\mu$ and $N$ are the conductivity mobility and net donor density, respectively, as determined in this work; $(\mu_H/\mu)_{\text{calc.}}$ is the calculated ratio of the Hall mobility to conductivity mobility assuming a specific form (power law) for the energy dependence of the relaxation time;[6] and $(\mu_H/\mu)_{\text{meas.}}$ is

[6] Data supplied by M. Glicksman, RCA Laboratories.

the experimentally determined ratio given in columns 3 and 4 in the table.

Although it is not the purpose in this paper to present any detailed discussion of the mobility or the scattering mechanism, a few observations relevant to the data presented in Figure 3 and Table I are in orders. As seen in Figure 3, there is generally good agreement between the changes in Hall mobility and conductivity mobility due to

*Table I*—Mobility Data for the Germanium–Silicon Alloy System

| Sample No. | % Si | N | Hall Mobility ($\mu_H$) $cm^2/V$-sec. | Conductivity Mobility ($\mu$) $cm^2/V$-sec. | $\left(\frac{\mu_H}{\mu}\right)_{meas.}$ | $\left(\frac{\mu_H}{\mu}\right)_{calc.}$ |
|---|---|---|---|---|---|---|
| Pure Germanium[7,8] | 0 | $10^{15}$ | 4400 | 3900 | 1.13 | .92 |
| D106T | 6-7% | $1.5 \times 10^{15}$ | 2330 | 2660 | .88 | .90 |
| D106E | 14% | $1.9 \times 10^{15}$ | 525 | 3333 | 1.57 | |
| D106F | 14% | $1.8 \times 10^{15}$ | 510 | 375 | 1.36 | |
| D164H | 25% | $2.5 \times 10^{16}$ | 370 | 240 | 1.52 | 1.00 |
| S318D | 98% | $9.7 \times 10^{15}$ | 1150 | 730 | 1.38 | .97 |
| Pure Silicon[8,10] | 100% | $10^{15}$ | 1550 | 1350 | 1.15 | .98 |

changes in alloy composition. The sharp drop in mobility in the alloy is probably due to the presence of "disorder scattering"[11] caused by the random distribution of foreign atoms in the host lattice. However, in comparing the theoretical and the experimentally determined values of $(\mu_H/\mu)$ a discrepancy appears; the measured value is larger than the theoretical one by an amount in excess of experimental error, especially for the alloys with higher silicon content. This may be due to uncertainties in the assumptions underlying the theoretical calculations. For example, a simple power-law dependence of the scattering relaxation time on energy is probably not correct in these cases; or perhaps impurity scattering is present for the more highly doped samples which should effect the Hall mobility and conductivity mobility

[7] P. P. Debye and E. M. Conwell, "Electrical Properties of N-Type Germanium," *Phys. Rev.*, Vol. 93, p. 693, February, 1954.

[8] M. Prince, "Drift Mobilities in Semiconductors I Germanium," *Phys. Rev.*, Vol. 92, p. 681, November, 1953: "Drift Mobilities in Semiconductors II Silicon," *Phys. Rev.*, Vol. 93, p. 1204, March, 1954.

[9] F. J. Morin and J. P. Maita, "Electrical Properties of Silicon Containing Arsenic and Boron," *Phys. Rev.*, Vol. 96, p. 28, October, 1954.

[10] G. W. Ludwig and R. L. Watters, "Drift and Conductivity Mobility in Silicon," *Phys. Rev.*, Vol. 101, p. 1699, March, 1956.

[11] L. Nordheim, "Electron Theory of Metals," *Annalen der Physik*, Vol. 9, p. 607, May 20, 1931 and p. 641, May 28, 1931.

differently. In the region around 15 per cent silicon, where interband scattering is probably present due to the proximity of the silicon and germanium conduction bands, an added difficulty in the calculations of $(\mu_H/\mu)$ is the uncertainty in the contribution to the conductivity by carriers in each band. For this reason the values of $(\mu_H/\mu)$ have not been included for the 14 per cent silicon samples.

## Conclusions

A method has been described for determining the net impurity concentration $|N_D - N_A|$ in a semiconductor. This method is based on measuring the transition capacitance of a p-n junction as a function of reverse bias. The method has been applied to a measurement of the conductivity mobility, $\mu$, of the germanium–silicon alloy system. Although the gross changes of the conductivity mobility with composition measured by this method follow closely those of the Hall mobility, it is found that the experimentally determined values of $(\mu_H/\mu)$ are significantly different from the theoretically calculated values. This indicates that some of the assumptions regarding the various scattering factors, customarily made in calculating $(\mu_H/\mu)$, may not be sufficiently accurate.

# SOME GRAPHICAL APPROACHES TO CODING PROBLEMS

By

Jacques Dutka*

RCA Defense Electronic Products,
New York, N. Y.

***Summary***—*The problem of constructing error-detecting and error-correcting codes for use in communicating information in binary coded form has received considerable attention in recent years. Some graphical methods for constructing such codes are presented, their geometrical interpretations are discussed, and some illustrative examples are worked out.*

## Introduction

A CONVENIENT approach to a study of the construction of binary codes may be had by considering the problem of assigning addresses for a number of locations in a radio network. It is assumed that when a message is broadcast, the body of the message is preceded by the address of the receiving location for which the message is intended. If one or more errors occur in the received address, such as a "1" appearing for a "0" or vice versa, it is possible for a location to receive a message not intended for it, and for the location for which the message was intended not to receive it at all. The problem then arises of finding a systematic way of assigning addresses in a form such that the effect of errors in transmission is minimized.

There are a number of different ways of doing this. In some cases, it is sufficient to assign a set of addresses to the locations in such a fashion that if a single error occurs in the address, the existence of an error will be manifest to the recipients of the message. Such an assignment of addresses is called a single-error-detecting code. The procedure may be extended to double-error-detecting codes, etc. Another possible approach is to assign a set of addresses to the locations in such a fashion that if a single error occurs in the address the existence of the error will be evident and, in addition, the location for which the message was intended will recognize that the message was

* Dr. Dutka is also an Adjunct Associate Professor of Electrical Engineering at Columbia University, New York, N. Y.

intended for it despite the error in the address. An assignment of addresses which accomplishes this is called a single-error-correcting code. Also, double-error-correcting codes, etc. can be constructed.

Numerical methods for constructing error-detecting and error-correcting binary codes have been devised by M. J. E. Golay,[1] R. W. Hamming[2] and others. This paper describes a graphical method for constructing error-correcting and error-detecting codes; the method appears to be novel and to offer some advantages in visualization.

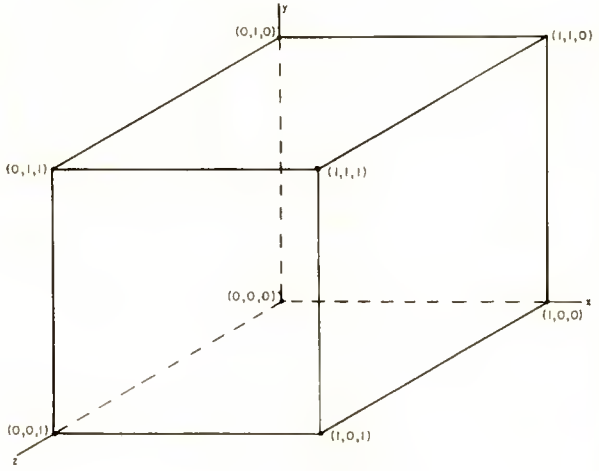



Fig. 1—A unit cube.

## Single-Error-Detecting Codes

Let us assume for illustrative purposes that an address for a location is given by an m bit code. If m = 3, then the possible addresses are 000, 001, 010, 011, 100, 101, 110, 111. To form a single-error-detecting code, it is necessary to choose a set of addresses each member of which differs from any of the others by at least two bits. One way of visualizing the problem is to consider the addresses 000, 001, . . . , 111 as the coordinates of the vertices of a unit cube (see Figure 1). From such a diagram, it is apparent how a code can be constructed. One need merely begin with any vertex, say (0, 0, 0), choose a second

[1] M. J. E. Golay, "Notes on Digital Coding," *Proc. I.R.E.*, Vol. 37, p. 657, June, 1949.

[2] R. W. Hamming, "Error Detecting and Error Correcting Codes," *Bell Sys. Tech. Jour.*, Vol. 29, p. 147, April, 1950.

vertex which is two edges away (a "distance" of 2), such as (1, 1, 0), then a vertex which is a distance of 2 from (1, 1, 0), such as (1, 0, 1), and then finally choose (0, 1, 1), which is a distance of 2 from (1, 0, 1). But a difficulty in this procedure is that it is not easy to keep track of which vertices have been previously enumerated.

A more systematic procedure is to form a Hamiltonian circuit of the cube — a continuous line which consists of unit steps along the edges which passes through each vertex of the cube once and only once. Thus, we get for example, on starting with (0, 0, 0) the vertices (1, 0, 0), (1, 1, 0), (0, 1, 0), (0, 1, 1), (0, 0, 1), (1, 0, 1) and (1, 1, 1) in succession. This circuit of the vertices by which one proceeds from

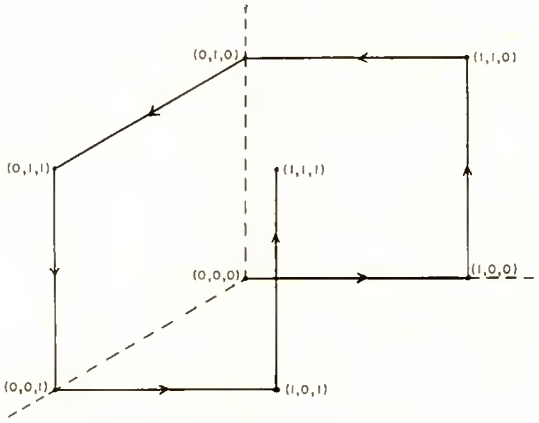



Fig. 2—A Hamiltonian circuit.

one vertex to the next by changing just one coordinate (bit) at a time is equivalent to the construction of a reflected binary, or Gray code (see Figure 2). By choosing alternate vertices along this circuit, we can obtain a set of addresses each member of which differs from any of the others by two bits. Thus, we obtain, for instance, the sequence (0, 0, 0), (1, 1, 0), (0, 1, 1) and (1, 0, 1) and this set of addresses forms a single-error-detecting code. If a nonexistent address, say (1, 1, 1) is received, it is immediately apparent that an error has occurred, but it is not known whether the correct address should have been (1, 1, 0), (0, 1, 1), or (1, 0, 1). (Other possible vertices are excluded since it has been assumed that only a single error could occur.)

The method described above for constructing a single-error-detect-

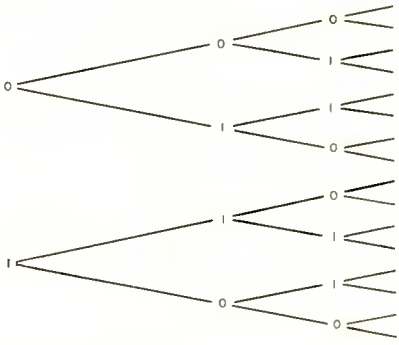


Fig. 3—Construction of a reflected binary code.

ing code by first forming a Hamiltonian circuit can, in principle, be extended to the case where the $2^n$ vertices of an n-dimensional hypercube are considered. But a difficulty arises in that it is hard to draw a usable diagram of a hypercube on a (two-dimensional) sheet of paper. (This problem is considered in greater detail in the next section.) A straightforward method of forming a reflected binary code (Hamiltonian circuit) is to draw a graph of the type shown in Figure 3. In this figure, reading from left to right, we get 000, 001, 011, 010, 110, 111, 101, 100. The method of extension of the branching process consists of repetitions of the bits 0 1 and the reflected pattern 1 0.

## Graphical Representation of a Hypercube

Let us consider a set of $2^4$ addresses which correspond to the vertices of a unit hypercube in four dimensions (tesseract). These vertices can be represented in the pattern of squares shown in Table I. This pattern is constructed as follows: The vertices which are immediately adjacent to a given vertex, either horizontally or vertically,

Table I

| A | | | B |
|---|---|---|---|
| 1100 | 1000 | 1001 | 1101 |
| 0100 | 0000 | 0001 | 0101 |
| 0110 | 0010 | 0011 | 0111 |
| 1110 | 1010 | 1011 | 1111 |
| D | | | C |

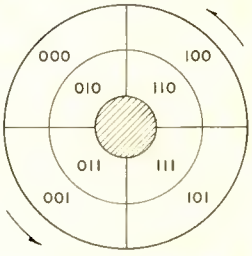



Fig. 4—Plan view of a torus.

differ from that vertex by one bit. Thus, the "singletons" associated with 0000, say, are 0100, 0001, 0010 and 1000. The side denoted by AB is to be thought of as coinciding with the side DC and the side AD is to be thought of as coinciding with the side BC. Thus, this pattern is a two-dimensional representation of the outer surface of a torus or anchor ring. The vertices which are two squares removed either horizontally or vertically from a given vertex, or those which are one square removed from the vertex diagonally, differ from that vertex by two bits. Thus, the "doublets" associated with 0000, say, are 0101, 1010, 0110, 1100, 1001 and 0011. It is evident how by proceeding along rows and columns one square at a time (as a rook would move on a chessboard) we can obtain a Hamiltonian circuit of the vertices and this is equivalent to the generation of a reflected binary code. For example, one such rook's tour which generates a reflected binary code is 1100, 1000, 1001, 1101, 0101, 0001, 0000, 0100, 0110, 0010, 0011, 0111, 1111, 1011, 1010, 1110.

The corresponding pattern for the case of a cube is easily constructed. Figure 4 is a plan view of a torus with a pattern on its surface showing eight cells to which the vertices of a cube are assigned.

The method of constructing patterns which represent Hamiltonian circuits of the vertices of hypercubes in n = 5, 6, . . . dimensions is now apparent. Table II shows the arrangement for n = 5. ABCD

Table II

| A | | | | B | | | E |
|---|---|---|---|---|---|---|---|
| 01100 | 01000 | 01001 | 01101 | 11101 | 11001 | 11000 | 11100 |
| 00100 | 00000 | 00001 | 00101 | 10101 | 10001 | 10000 | 10100 |
| 00110 | 00010 | 00011 | 00111 | 10111 | 10011 | 10010 | 10110 |
| 01110 | 01010 | 01011 | 01111 | 11111 | 11011 | 11010 | 11110 |
| D | | | | C | | | F |

represents one torus pattern and BEFC represents a torus pattern which is the reflected image of ABCD in the line BC. Thus, the image of each element of ABCD is obtained by reflecting the element in BC and replacing the initial "0" by "1".

Similarly, a pattern for a hypercube in six-dimensional space with $2^6 = 64$ vertices can be readily constructed as shown in Table III. Here, the pattern BGHC is a reflection of ABCD in the line BC and CHIF is a reflection of DCFE in the line CF and of BGHC in the line CH. The entire pattern represents the three-dimensional surface of a hypertorus in four-dimensional space. Through each element of the torus, there will pass three mutually perpendicular circles (in contrast with the two perpendicular circles of an ordinary three-dimensional torus). Thus, passing through the element 000000, we

Table III

| A | | | | B | | | | G |
|---|---|---|---|---|---|---|---|---|
| 001100 | 001000 | 001001 | 001101 | 011101 | 011001 | 011000 | 011100 | |
| 000100 | 000000 | 000001 | 000101 | 010101 | 010001 | 010000 | 010100 | |
| 000110 | 000010 | 000011 | 000111 | 010111 | 010011 | 010010 | 010110 | |
| 001110 | 001010 | 001011 | 001111 | 011111 | 011011 | 011010 | 011110 | |
| D | | | | | | | | H |
| 101110 | 101010 | 101011 | 101111 C | 111111 | 111011 | 111010 | 111110 | |
| 100110 | 100010 | 100011 | 100111 | 110111 | 110011 | 110010 | 110110 | |
| 100100 | 100000 | 100001 | 100101 | 110101 | 110001 | 110000 | 110100 | |
| 101100 | 101000 | 101001 | 101101 | 111101 | 111001 | 111000 | 111100 | |
| E | | | | F | | | | I |

have one circle consisting of 000000, 000001, 000101, 000100 moving horizontally; a second circle consisting of 000000, 001000, 001010, 000010 moving vertically; a third circle consisting of 000000, 010000, 110000, 100000 moving in a direction perpendicular to the others.

Graphical patterns for vertices of hypercubes in higher dimensions can be constructed in an analogous manner.

## Single-Error-Correcting Code

It will now be shown how a single-error-correcting code can be constructed, which utilizes the graphical description discussed above. For simplicity, the method will be applied to the case of $2^5 = 32$ vertices, but the method of extending the procedure to the general case will be evident. To obtain a single-error-correcting code, it is necessary to

Table IV

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| x | ⊗ | x | | | | x | |
| ⊗ | x | ⊗ | x | | x | ⊗ | x |
| x | ⊗ | x | | | | x | |
| | x | | | | | | |

isolate a set of vertices each of which is at least a distance of three from every other vertex of the set. Starting with any vertex of the hypercube in five dimensions, say 00000 in Table II,

we first circle all the $\binom{5}{1} = 5$ singletons associated with 00000 and also record all the $\binom{5}{2} = 10$ doublets associated with it. (The symbol $\binom{n}{r} = \frac{n!}{r!(n-r)!}$ denotes the combination of $n$ things taken

Table V

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| xx | xx | x | x | | | x | x |
| xx | x | x | x | | x | x | x |
| xx | xx | x | x | | | x | x |
| x | xx | x | x | x | | x | x |

$r$ at a time.) The resulting configuration is a graphical representation of a Hamming sphere of radius 2 as shown in Table IV.

Next, we choose an element which has not been eliminated, say 01110. In general, if we wish to get as many addresses as possible in the final result, it is advisable to choose each new element as close as possible to the previous, for instance, the equivalent of a knight's move away on a chessboard. If we now record all the singletons and doublets associated with 01110, we get the pattern of Table V. There are now only six vertices left unmarked of the original thirty-two. A selection of any one of them, say 11101, yields a pattern as shown in Table VI, in which all but one of the vertices, here 10011, is marked. Thus, we

Table VI

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| xxx | xx | xx | xx | x | x | xx | xx |
| xx | x | x | xx | x | xx | x | xx |
| xx | xx | x | x | x | | x | x |
| x | xx | x | xx | xx | x | x | x |

Table VII

| 000 100 100 | 000 100 000 | 000 100 001 | 000 100 101 | 000 110 101 | 000 110 001 | 000 110 100 | 000 110 000 | 010 110 100 | 010 110 000 | 010 110 001 | 010 110 101 | 010 100 101 | 010 100 001 | 010 100 000 | 010 100 100 | 010 000 100 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 000 101 100 | 000 101 000 | 000 101 001 | 000 101 101 | 000 111 101 | 000 111 001 | 000 111 100 | 000 111 000 | 010 111 100 | 010 111 000 | 010 111 001 | 010 111 101 | 010 101 101 | 010 101 001 | 010 101 000 | 010 101 100 | 010 001 100 |
| 000 001 100 | 000 001 000 | 000 001 001 | 000 001 101 | 000 011 101 | 000 011 001 | 000 011 100 | 000 011 000 | 010 011 100 | 010 011 000 | 010 011 001 | 010 011 101 | 010 001 101 | 010 001 001 | 010 001 000 | 010 001 100 | 010 000 100 |
| 000 000 100 | 000 000 000 | 000 000 001 | 000 000 101 | 000 010 101 | 000 010 001 | 000 010 100 | 000 010 000 | 010 010 100 | 010 010 000 | 010 010 001 | 010 010 101 | 010 000 101 | 010 000 001 | 010 000 000 | 010 000 100 | 010 000 100 |
| 000 000 110 | 000 000 010 | 000 000 011 | 000 000 111 | 000 010 111 | 000 010 011 | 000 010 110 | 000 010 010 | 010 010 110 | 010 010 010 | 010 010 011 | 010 010 111 | 010 000 111 | 010 000 011 | 010 000 010 | 010 000 110 | 010 000 110 |
| 000 001 110 | 000 001 010 | 000 001 011 | 000 001 111 | 000 011 111 | 000 011 011 | 000 011 110 | 000 011 010 | 010 011 110 | 010 011 010 | 010 011 011 | 010 011 111 | 010 001 111 | 010 001 011 | 010 001 010 | 010 001 110 | 010 001 110 |
| 000 101 110 | 000 101 010 | 000 101 011 | 000 101 111 | 000 111 111 | 000 111 011 | 000 111 110 | 000 111 010 | 010 111 110 | 010 111 010 | 010 111 011 | 010 111 111 | 010 101 111 | 010 101 011 | 010 101 010 | 010 101 110 | 010 101 110 |
| 000 100 110 | 000 100 010 | 000 100 011 | 000 100 111 | 000 110 111 | 000 110 011 | 000 110 110 | 000 110 010 | 010 110 110 | 010 110 010 | 010 110 011 | 010 110 111 | 010 100 111 | 010 100 011 | 010 100 010 | 010 100 110 | 010 100 110 |
| 001 100 110 | 001 100 010 | 001 100 011 | 001 100 111 | 001 110 111 | 001 110 011 | 001 110 110 | 001 110 010 | 011 110 110 | 011 110 010 | 011 110 011 | 011 110 111 | 011 100 111 | 011 100 011 | 011 100 010 | 011 100 110 | 011 000 110 |
| 001 101 110 | 001 101 010 | 001 101 011 | 001 101 111 | 001 111 111 | 001 111 011 | 001 111 110 | 001 111 010 | 011 111 110 | 011 111 010 | 011 111 011 | 011 111 111 | 011 101 111 | 011 101 011 | 011 101 010 | 011 101 110 | 011 001 110 |
| 001 001 110 | 001 001 010 | 001 001 011 | 001 001 111 | 001 011 111 | 001 011 011 | 001 011 110 | 001 011 010 | 011 011 110 | 011 011 010 | 011 011 011 | 011 011 111 | 011 001 111 | 011 001 011 | 011 001 010 | 011 001 110 | 011 000 110 |
| 001 000 110 | 001 000 010 | 001 000 011 | 001 000 111 | 001 010 111 | 001 010 011 | 001 010 110 | 001 010 010 | 011 010 110 | 011 010 010 | 011 010 011 | 011 010 111 | 011 000 111 | 011 000 011 | 011 000 010 | 011 000 110 | 011 000 110 |
| 001 000 100 | 001 000 000 | 001 000 001 | 001 000 101 | 001 010 101 | 001 010 001 | 001 010 100 | 001 010 000 | 011 010 100 | 011 010 000 | 011 010 001 | 011 010 101 | 011 000 101 | 011 000 001 | 011 000 000 | 011 000 100 | 011 000 100 |
| 001 001 100 | 001 001 000 | 001 001 001 | 001 001 101 | 001 011 101 | 001 011 001 | 001 011 100 | 001 011 000 | 011 011 100 | 011 011 000 | 011 011 001 | 011 011 101 | 011 001 101 | 011 001 001 | 011 001 000 | 011 001 100 | 011 001 100 |
| 001 101 100 | 001 101 000 | 001 101 001 | 001 101 101 | 001 111 101 | 001 111 001 | 001 111 100 | 001 111 000 | 011 111 100 | 011 111 000 | 011 111 001 | 011 111 101 | 011 101 101 | 011 101 001 | 011 101 000 | 011 101 100 | 011 101 100 |
| 001 100 100 | 001 100 000 | 001 100 001 | 001 100 101 | 001 110 101 | 001 110 001 | 001 110 100 | 001 110 000 | 011 110 100 | 011 110 000 | 011 110 001 | 011 110 101 | 011 100 101 | 011 100 001 | 011 100 000 | 011 100 100 | 011 100 100 |
| 101 100 100 | 101 100 000 | 101 100 001 | 101 100 101 | 101 110 101 | 101 110 001 | 101 110 100 | 101 110 000 | 111 110 100 | 111 110 000 | 111 110 001 | 111 110 101 | 111 100 101 | 111 100 001 | 111 100 000 | 111 100 100 | 111 100 100 |
| 101 101 100 | 101 101 000 | 101 101 001 | 101 101 101 | 101 111 101 | 101 111 001 | 101 111 100 | 101 111 000 | 111 111 100 | 111 111 000 | 111 111 001 | 111 111 101 | 111 101 101 | 111 101 001 | 111 101 000 | 111 101 100 | 111 101 100 |
| 101 001 100 | 101 001 000 | 101 001 001 | 101 001 101 | 101 011 101 | 101 011 001 | 101 011 100 | 101 011 000 | 111 011 100 | 111 011 000 | 111 011 001 | 111 011 101 | 111 001 101 | 111 001 001 | 111 001 000 | 111 001 100 | 111 001 100 |
| 101 000 100 | 101 000 000 | 101 000 001 | 101 000 101 | 101 010 101 | 101 010 001 | 101 010 100 | 101 010 000 | 111 010 100 | 111 010 000 | 111 010 001 | 111 010 101 | 111 000 101 | 111 000 001 | 111 000 000 | 111 000 100 | 111 000 100 |
| 101 000 110 | 101 000 010 | 101 000 011 | 101 000 111 | 101 010 111 | 101 010 011 | 101 010 110 | 101 010 010 | 111 010 110 | 111 010 010 | 111 010 011 | 111 010 111 | 111 000 111 | 111 000 011 | 111 000 010 | 111 000 110 | 111 000 110 |
| 101 001 110 | 101 001 010 | 101 001 011 | 101 001 111 | 101 011 111 | 101 011 011 | 101 011 110 | 101 011 010 | 111 011 110 | 111 011 010 | 111 011 011 | 111 011 111 | 111 001 111 | 111 001 011 | 111 001 010 | 111 001 110 | 111 001 110 |
| 101 101 110 | 101 101 010 | 101 101 011 | 101 101 111 | 101 111 111 | 101 111 011 | 101 111 110 | 101 111 010 | 111 111 110 | 111 111 010 | 111 111 011 | 111 111 111 | 111 101 111 | 111 101 011 | 111 101 010 | 111 101 110 | 111 101 110 |
| 101 100 110 | 101 100 010 | 101 100 011 | 101 100 111 | 101 110 111 | 101 110 011 | 101 110 110 | 101 110 010 | 111 110 110 | 111 110 010 | 111 110 011 | 111 110 111 | 111 100 111 | 111 100 011 | 111 100 010 | 111 100 110 | 111 100 110 |
| 100 100 110 | 100 100 010 | 100 100 011 | 100 100 111 | 100 110 111 | 100 110 011 | 100 110 110 | 100 110 010 | 110 110 110 | 110 110 010 | 110 110 011 | 110 110 111 | 110 100 111 | 110 100 011 | 110 100 010 | 110 100 110 | 110 100 110 |
| 100 101 110 | 100 101 010 | 100 101 011 | 100 101 111 | 100 111 111 | 100 111 011 | 100 111 110 | 100 111 010 | 110 111 110 | 110 111 010 | 110 111 011 | 110 111 111 | 110 101 111 | 110 101 011 | 110 101 010 | 110 101 110 | 110 101 110 |
| 100 001 110 | 100 001 010 | 100 001 011 | 100 001 111 | 100 011 111 | 100 011 011 | 100 011 110 | 100 011 010 | 110 011 110 | 110 011 010 | 110 011 011 | 110 011 111 | 110 001 111 | 110 001 011 | 110 001 010 | 110 001 110 | 110 001 110 |
| 100 000 110 | 100 000 010 | 100 000 011 | 100 000 111 | 100 010 111 | 100 010 011 | 100 010 110 | 100 010 010 | 110 010 110 | 110 010 010 | 110 010 011 | 110 010 111 | 110 000 111 | 110 000 011 | 110 000 010 | 110 000 110 | 110 000 110 |
| 100 000 100 | 100 000 000 | 100 000 001 | 100 000 101 | 100 010 101 | 100 010 001 | 100 010 100 | 100 010 000 | 110 010 100 | 110 010 000 | 110 010 001 | 110 010 101 | 110 000 101 | 110 000 001 | 110 000 000 | 110 000 100 | 110 000 100 |
| 100 001 100 | 100 001 000 | 100 001 001 | 100 001 101 | 100 011 101 | 100 011 001 | 100 011 100 | 100 011 000 | 110 011 100 | 110 011 000 | 110 011 001 | 110 011 101 | 110 001 101 | 110 001 001 | 110 001 000 | 110 001 100 | 110 001 100 |
| 100 101 100 | 100 101 000 | 100 101 001 | 100 101 101 | 100 111 101 | 100 111 001 | 100 111 100 | 100 111 000 | 110 111 100 | 110 111 000 | 110 111 001 | 110 111 101 | 110 101 101 | 110 101 001 | 110 101 000 | 110 101 100 | 110 101 100 |
| 100 100 100 | 100 100 000 | 100 100 001 | 100 100 101 | 100 110 101 | 100 110 001 | 100 110 100 | 100 110 000 | 110 110 100 | 110 110 000 | 110 110 001 | 110 110 101 | 110 100 101 | 110 100 001 | 110 100 000 | 110 100 100 | 110 100 100 |

obtain the set 00000, 01110, 11101, 10011, each element of which differs from every other element by at least three bits.

By applying the method outlined above, C. N. Merrell has constructed a graphical representation for a nine-bit code as shown in Table VII. From Table VII he has derived the set of 32 addresses shown in Table VIII. This is a single-error-correcting code.

Table VIII

| | DECIMAL | BINARY | | DECIMAL | BINARY |
|---|---|---|---|---|---|
| (1) | 1 | 000, 000, 001 | (17) | 384 | 110, 000, 000 |
| (2) | 6 | 000, 000, 110 | (18) | 391 | 110, 000, 111 |
| (3) | 24 | 000, 011, 000 | (19) | 409 | 110, 011, 001 |
| (4) | 31 | 000, 011, 111 | (20) | 414 | 110, 011, 110 |
| (5) | 42 | 000, 101, 010 | (21) | 427 | 110, 101, 011 |
| (6) | 45 | 000, 101, 101 | (22) | 428 | 110, 101, 100 |
| (7) | 51 | 000, 110, 011 | (23) | 434 | 110, 110, 010 |
| (8) | 52 | 000, 110, 100 | (24) | 437 | 110, 110, 101 |
| (9) | 75 | 001, 001, 011 | (25) | 458 | 111, 001, 010 |
| (10) | 76 | 001, 001, 100 | (26) | 461 | 111, 001, 101 |
| (11) | 82 | 001, 010, 010 | (27) | 467 | 111, 010, 011 |
| (12) | 85 | 001, 010, 101 | (28) | 468 | 111, 010, 100 |
| (13) | 96 | 001, 100, 000 | (29) | 481 | 111, 100, 001 |
| (14) | 103 | 001, 100, 111 | (30) | 486 | 111, 100, 110 |
| (15) | 121 | 001, 111, 001 | (31) | 504 | 111, 111, 000 |
| (16) | 126 | 001, 111, 110 | (32) | 511 | 111, 111, 111 |

## Conclusions

In principle, the graphical methods developed here are sufficient for the construction of binary error-detecting and error-correcting codes. In practice, however, while codes of these types can readily be constructed if words of moderate length are involved, it is found that (as is usual with graphical methods), the procedure is quite cumbersome when applied to the construction of codes containing longer words, and in this case the possibility of developing programs for computers warrants investigation.

# A CARRIER-ENERGIZED BISTABLE CIRCUIT USING VARIABLE-CAPACITANCE DIODES

By

E. O. Keizer

RCA Laboratories,
Princeton, N. J.

***Summary***—*A variable-capacitance junction diode, when used in a simple circuit driven from a high-frequency a-c source, can cause that circuit to have a bistable characteristic suitable for dynamic memory, or to have a sensitive output–input characteristic suitable for control or detection purposes. This paper outlines the principles of operation of the basic bistable circuit and describes several variations, including a transistor version in which the two junctions of a transistor are used in place of two variable-capacitance diodes. A two-diode circuit is described for which the output–input characteristic is bistable for a range of nearly ten to one in input voltage amplitude. Using an energizing carrier level of about one-half volt, a d-c output change of several volts may be obtained. At an energizing frequency of 2 megacycles, transitions require about two microseconds. The input power per circuit is approximately one milliwatt.*

## Introduction

An experimental variable-capacitance diode having relatively high capacitance and low series resistance (forming a low-loss, high-Q capacitive component) has been described in the literature.[1] Experimental diodes of similar type have since been a factor in circuit development. This paper describes a simple, carrier-energized, bistable circuit utilizing both the variable-capacitance and the nonlinear conductance characteristics of such diodes. For purposes of discussion, a specific experimental diode is used to illustrate the application; however, the treatment is sufficiently general to be applicable to other semiconductor diodes with somewhat similar characteristics.

## Basic Bistable Circuit

### *Description of Experimental Diodes*

The low-frequency electrical characteristics of the diodes used in the bistable circuit tests were quite similar to those of the diodes described by Giacoletto and O'Connell.[1] For a typical unit at five volts

[1] L. J. Giacoletto and J. H. O'Connell, "A Variable-Capacitance Germanium Junction Diode for UHF," *Transistors I*, RCA Laboratories, Princeton, N. J., 1956, p. 221.

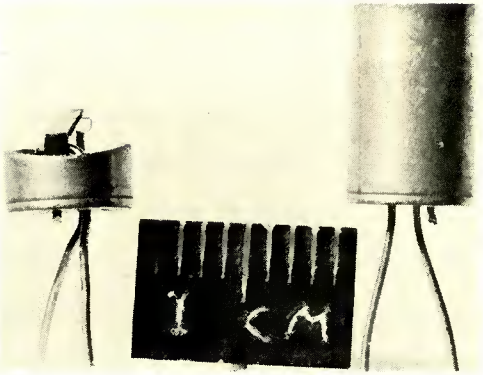



Fig. 1—Experimental variable-capacitance diode.

reverse bias, the capacitance was about 50 micromicrofarads and the leakage current at room temperature was about 0.3 microampere, while the series base resistance was about 0.5 ohm. Peak inverse voltage, limited by conduction in the reverse direction, was about 16 volts. The diodes were mounted in hearing-aid transistor-type cases and had short flexible leads. The size and mounting are illustrated in Figure 1. The capacitance of these germanium alloy-junction units varies inversely as the square root of the instantaneous effective reverse bias voltage.

*Basic Circuit*

The variable-capacitance diode bistable circuit is shown in basic form in Figure 2. A high-frequency a-c voltage source drives a series combination of an inductance and capacitances (resistance, $R$, is at first to be considered large). The effective capacitance of the variable-capacitance diode depends upon the amplitudes of the a-c and d-c voltages across its terminals. Because the diode is a rectifier, the d-c voltage is essentially equal to the peak a-c voltage applied to the diode,

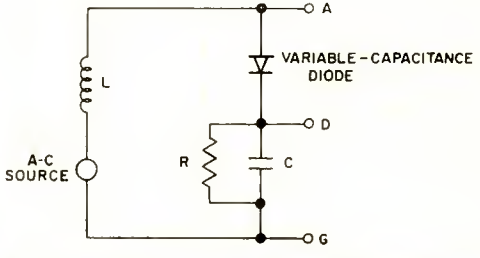



Fig. 2—Variable-capacitance diode bistable circuit.

which in turn depends upon the frequency and amplitude of the voltage source and upon the circuit components. When the frequency and component values are properly chosen, effective capacitance variation is such that the circuit may be tuned through series resonance from one side to the other by changing the amplitude of the source. If the effective $Q$ of the circuit exceeds a minimum value, the change is abrupt, and a complete cycle of raising and lowering the input voltage amplitude traces out a pseudo-hysteresis curve. In Figure 3 the rectified d-c voltage measured between terminals $D$ and $G$ of the circuit of Figure 2 has been plotted against the input amplitude. The segments of the characteristic will be traced out in the order 1-2, 2-3, 3-4, 4-5,

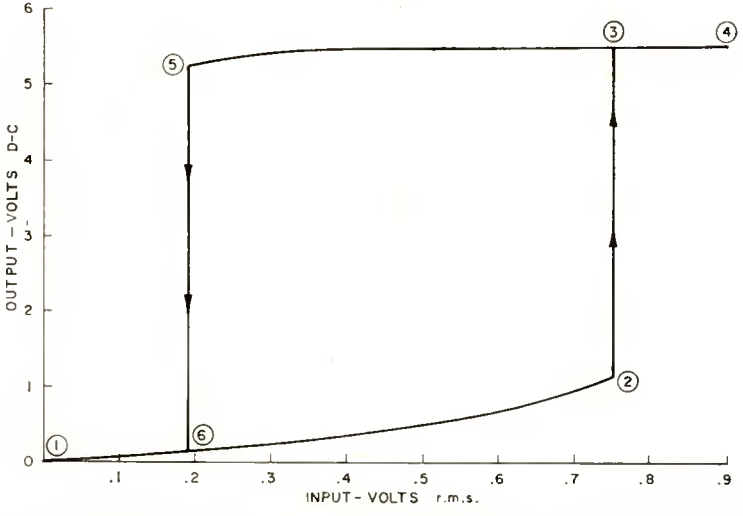



Fig. 3—Pseudo-hysteresis characteristic obtained by varying the amplitude of the input volatge.

5-6, 6-1, 1-2, etc., if the start of the input amplitude variation cycle is begun near zero input amplitude. The a-c voltage appearing between terminals $A$ and $D$ of Figure 2 varies with input amplitude in a similar fashion. These pseudo-hysteresis characteristics pertain to the circuit, not to the diode itself.

As shown by Figure 3, the characteristic is bistable over a considerable range of input voltage amplitude. Within this range, the two states may be easily distinguished by the substantial difference in output amplitudes. The input may be varied within this range (for example, 0.2 to 0.7 volt) and the output will continue to trace along one segment of the characteristic without forming minor loops or affecting the critical voltages which bound the bistable region.

If the frequency of the input voltage rather than its amplitude is varied, the *output voltage versus frequency* characteristic will also have a bistable region. This is illustrated in Figure 4. In this case the peak-to-peak a-c across the diode and the fixed capacitor was taken as the output.

*Elementary Principle of Operation*

If the variable-capacitance diode were replaced in Figure 2 with a hypothetical variable-capacitance element sensitive to bias voltage and the bias were varied by external means, the output versus bias curves for three different inputs might be represented approximately by the three dotted response curves in Figure 5. Now consider the action of a rectifying diode shunted across the bias-sensitive capaci-

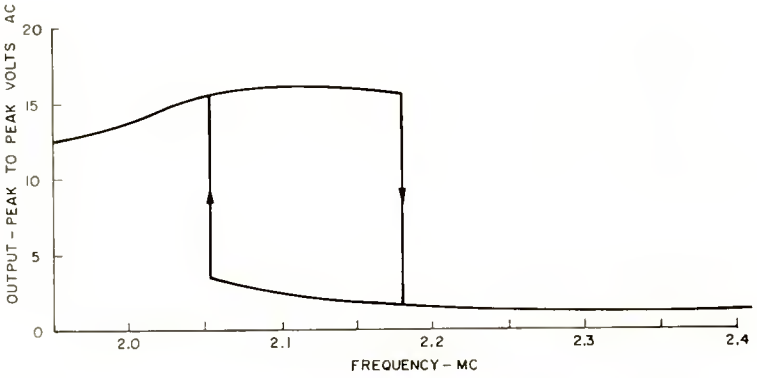



Fig. 4—Pseudo-hysteresis characteristic obtained by varying the frequency of the input voltage.

tance when the external bias is removed. The diode will rectify the a-c voltage appearing across the variable capacitance, resulting in a self-bias appearing across both the variable capacitance and the fixed capacitance, $C$. The magnitude of this bias will equal the peak of the a-c voltage across the variable capacitance, the relation being a linear one as shown by the solid straight line on Figure 5. For all levels of input voltage, the steady-state operating point will remain on this straight line. Since the operating point must also be on the a-c response curve corresponding to the particular level of input being used, it must be where the two curves (the response curve and the self-bias curve) cross. For small amplitudes of input, there is only one crossing. For somewhat larger input amplitudes, however, there are three crossings. For still larger input amplitudes there is again only one crossing.

At any point where the response curve exceeds the self-bias curve, there will be additional self-bias generated, so the operating point will shift to one of higher bias, continuing to shift until a point of coincidence of the two curves occurs, i.e., until the peak a-c is just enough to maintain the self-bias necessary to tune the circuit to maintain that amount of peak a-c. At any point where the response curve falls below the self-bias curve, there will be no diode conduction, and the self-bias will leak away, moving the operating point toward lower bias, until a point of coincidence of the curves is again reached. The center point of crossing, in cases where there are three, is not a stable operating

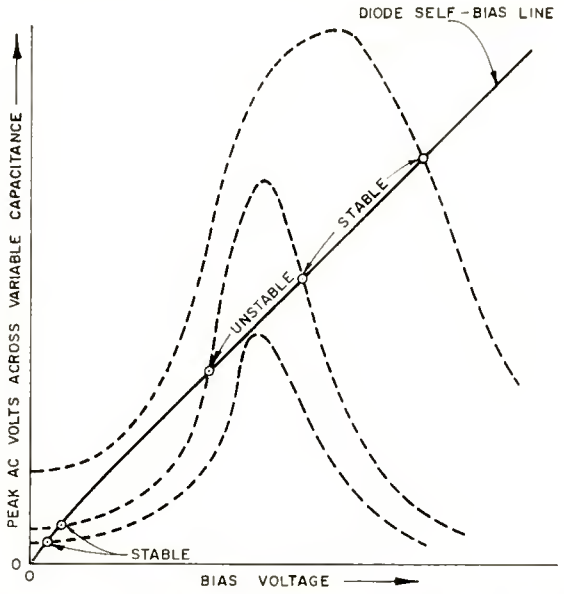



Fig. 5—Curves showing responses, self-bias line, and operating points.

point, since any perturbation away from it will result in a shift to one of the other points of crossing. However, all of the other points are stable, since any perturbation about them is countered by a restoring action.

We may define the largest input voltage for which the response curves still cross the self-bias line in three places as the upper critical input voltage, and similarly the smallest such as the lower critical input voltage. These correspond to the input voltages at point 2 and point 5 of Figure 3. The effective $Q$ of the response curves and their position along the bias axis affect the location and separation of these voltages. If the series inductance or the input frequency is increased,

the response will be moved toward a higher bias condition (assuming higher bias means less capacitance), increased input will be needed, and the back-voltage limitation of the diode may be reached before the bistable region is reached.

If the $Q$ of the circuit is made lower, as for example by decreasing $R$, the separation of the upper and lower critical levels will decrease, and a point may be reached where the circuit is no longer bistable, although a steep S-shape remains in the output–input amplitude characteristic.

*Practical Operation*

In practice there are some further considerations. The capacitance of actual variable-capacitance diodes follows the instantaneous total reverse bias, resulting in a considerable flattening of one half of the a-c voltage waveform and peaking of the other half. It also results in some effective broadening and shifting in the frequency of the response curves as a function of the amplitude of the input. Also, in the case of the experimental diodes used for most of the tests, conduction in the reverse direction limited the peak-to-peak a-c voltage to about 16 volts; this flattened the self-bias line at slightly under six volts. (The peak amplitude of the half cycle of the a-c wave that opposed the average bias was about 6 volts and the peak amplitude of the other half cycle about ten volts.) Reverse-bias conduction contributes to the flatness of the top portions of the curves of Figures 3 and 4. Still another practical effect encountered under certain conditions of operation is a tendency for a spontaneous modulation or subfrequency generation to occur. It occurs over a certain range of inputs when the circuit is in the upper or high-level state and the output is just below the reverse-conduction-limited level. By operating the circuit with sufficient input that the level of the upper bistable state is limited by reverse conduction, or with a small input, the tendency to generate subfrequencies can be avoided.

*Triggering*

In the bistable region, operation may be triggered between states by several means. As already indicated, the input amplitude may be momentarily changed beyond the critical amplitude, or it may be changed in frequency. By changing the d-c bias from an external source stiff enough to override the self-bias, the same effect may be achieved. The tuning of the circuit may be changed by using an additional variable-capacitance diode or a saturable inductance. Under proper conditions, light energy falling on the junction area of the diode can also be used to control or trigger the output.

The output pulse or bias change derived from a transition usually will be more than enough to trigger or switch another similar circuit. Depending on the kind of coupling used, two of these circuits may be interconnected to achieve either flip-flop or multivibrator action. Coupling means include use of a common input impedance, coupling between the outputs, or a combination of the two.

*Transition Characteristic*

The transition characteristics are illustrated in the photographs of Figure 6. A sine-wave amplitude modulated 2.2-megacycle input

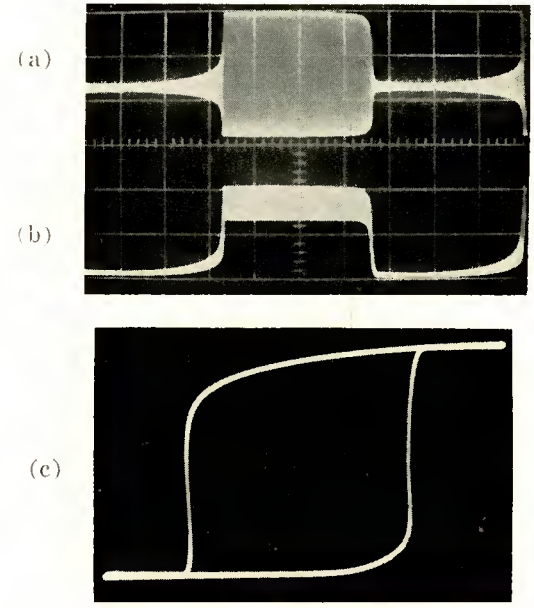



Fig. 6—(a) and (b) Waveforms observed at terminal A and D of Figure 2. (c) Output–input characteristic of circuit of Figure 2.

was applied to the circuit of Figure 2; $C$ was 330 micromicrofarads, $R$ was 50,000 ohms, and $L$ was a 64-105 microhenry tunable inductance tuned to give the bistable responses shown. Oscilloscope probes, each having a capacitance of approximately 15 micromicrofarads to terminal $G$, were connected to terminal $A$ and to terminal $D$. The waveform shown in Figure 6(a) is that seen at terminal $A$, and the waveform shown in Figure 6(b) is that seen at terminal $D$. Each major division of horizontal trace represents 20 microseconds in these figures; the transitions take about 2 microseconds. Figure 6(c) is another photograph taken under the same circuit conditions, but where the

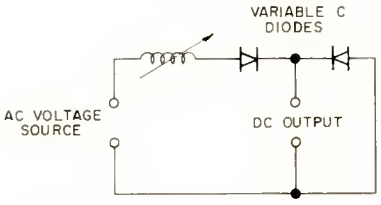



Fig. 7—Series diode version for maximum sensitivity and response speed.

filtered output from $D$ was applied to the vertical deflection input of the oscilloscope and the generator output was rectified, filtered, and applied to the horizontal deflection input.

Decreasing $C$ or increasing the input frequency (smaller $L$ needed) shortened the rise time, but had little effect on the fall time. On the other hand, a very large $R$ results in a short, sharp transition, followed by a long slow decay. The value of $C$ may be decreased, but as it approaches the value of the variable-capacitance diode, the width of the bistable region and the amount of output decrease.

## Other Versions of the Circuit

### *Series Diode Version for Maximum Sensitivity*

Reducing $C$ to a very low value may restrict the operating range of the circuit. This occurs because a small value of fixed $C$ places a relatively large fixed reactance in series with the variable-capacitance

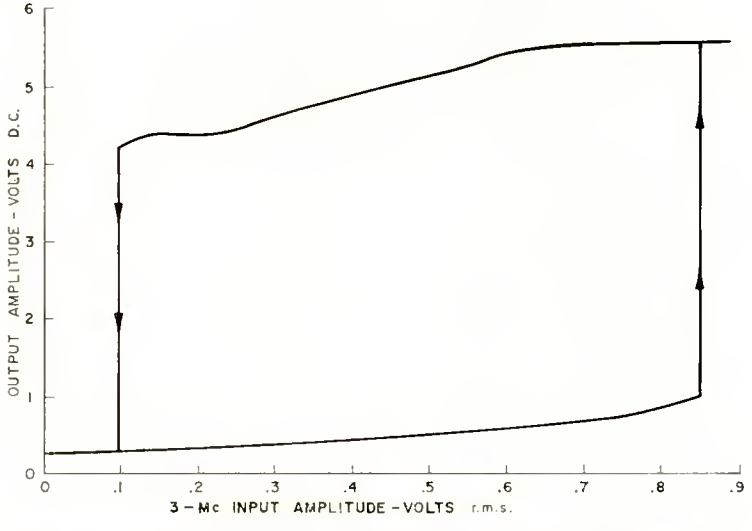



Fig. 8—Typical characteristic obtained with circuit of Figure 7.

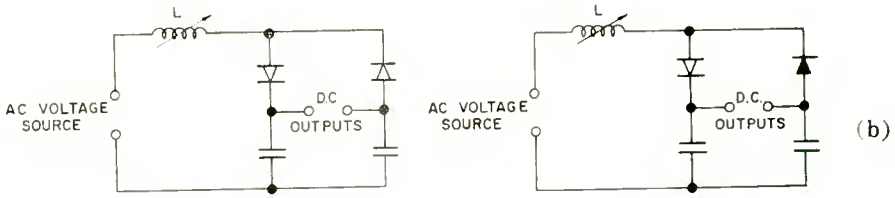



Fig. 9—Circuits showing use of two diodes to provide both polarities of d-c output.

diode. This reduces the range of tuning of the circuit that can be brought about by changes of capacitance of the diode. It also results in division of the a-c voltage so that only a portion of it appears across the diode. The circuit of Figure 7 shows how these effects may be overcome. A second variable-capacitance diode may be used in place of the fixed capacitor to maintain a highly bistable characteristic while at the same time providing a minimum of capacitance across $R$. The capacitances of both diodes, varying together, result in a maximum range of tuning of the circuit and in full utilization of the a-c voltage. A typical output–input characteristic for this circuit is shown in Figure 8. A bistable characteristic extending over a ten-to-one range of input voltages is not uncommon with this circuit if $R$ can be kept large.

*D-C Outputs of Different Polarity*

The circuit shown in Figure 9(a) shows how another variable-capacitance diode may be connected to provide two d-c outputs of

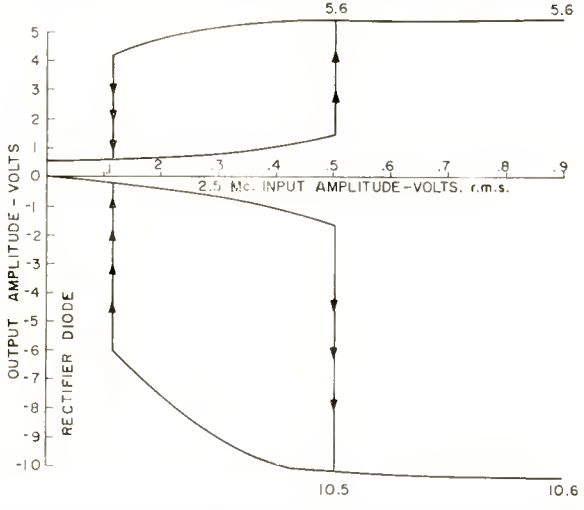



Fig. 10—Output–input characteristic for double-diode circuit of Figure 9(b).

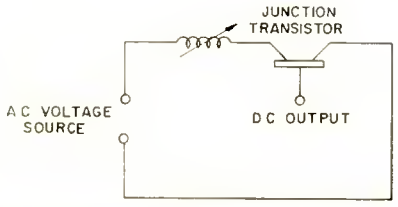



Fig. 11—Transistor version of bistable circuit.

different polarity. Typical output–input characteristics are similar to those shown in Figure 3, except that there will be two, one of each output polarity. A point-contact diode can be substituted for one of the variable-capacitance diodes, as shown in Figure 9(b). Figure 10 shows typical output–input characteristics for this connection. It will be noted that the output from the point-contact diode is greater than that from the variable-capacitance diode. As explained earlier, there is considerable flattening of one side of the a-c waveform, and peaking of the other. The variable-capacitance diode develops bias determined by the flattened half of the cycle, while the point-contact diode develops bias on the peaked half of the cycle. There is about a two-to-one difference in outputs.

*Transistor Version*

A transistor version, one in which a junction transistor is used to

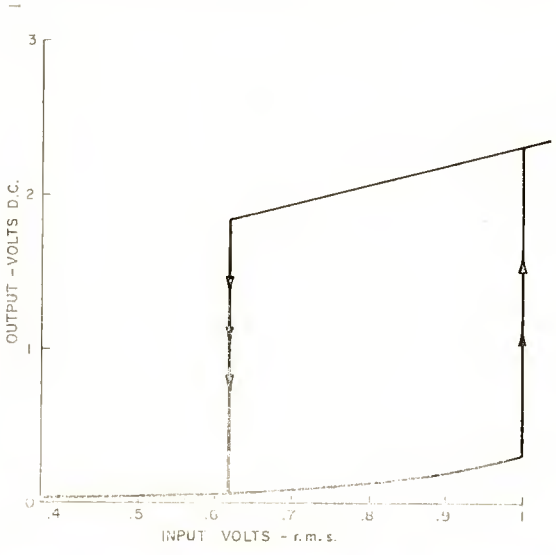



Fig. 12—Bistable characteristic of transistor version.

replace both of the variable-capacitance diodes in the series diode version, is shown schematically in Figure 11. This circuit uses the emitter–base junction capacitance and the base–collector capacitance in series to form a circuit that looks like the series variable-capacitance diode circuit of Figure 7. It is not precisely similar, however, because there is the possibility of transistor action during conduction. When the transistor is used in this manner, the resistance of its base and base lead are not normally part of the r-f circuit, and therefore do not lower the $Q$. The bistable output–input characteristic obtained with a 2N109 transistor is shown in Figure 12.

ACKNOWLEDGMENT

The author is pleased to acknowledge the assistance of Howard D. Helms, who was responsible for much of the experimental portion of the development and to whom the series-diode concept is due.

# THE PHYSICS OF THE CATHODE*

By

L. S. Nergaard

RCA Laboratories,
Princeton, N. J.

***Summary***—*Recent work on thermionic emitters suggests some generalizations bearing on all electron emitters. Four propositions are advanced for consideration and discussion. The propositions are:*

1. *Every cathode is a reducing agent.*
2. *Every cathode lives in equilibrium with its environment.*
3. *Every cathode is a dispenser cathode.*
4. *Monolayer film emitters do not exist.*

*Evidence to support these propositions is adduced. The evidence for the first three propositions is regarded as conclusive. The evidence for the fourth proposition is inconclusive but it suggests that thoriated-tungsten, BaO on tungsten, Ba on tungsten, the L-cathode, the impregnated cathode, and the E-cathode are semiconducting emitters in at least limited temperature ranges.*

## Introduction

SINCE the late war, a great deal of effort has been devoted to the study of thermionic emitters, particularly alkaline-earth-oxide, thoria, and "dispenser" cathodes. These studies have produced a wealth of detailed information about various emitters, have led to a few innovations, and have greatly enhanced the understanding of the physics and chemistry of emitters. This is all to the good. However, one wonders if close attention to detail has not tended to obscure the more general content of the work. Perhaps some generalizations might be discernable if one were to back off and view the scene as a whole; perhaps recent work would cast new light on older work if viewed from another and more remote standpoint. With these thoughts in mind, the writer has scanned the scene and, as a result, offers some generalizations for consideration and discussion. Now, nothing is so apt to provoke discussion as a categoric statement. Accordingly, the generalizations advanced are put in the form of four propositions, to wit:

1. Every cathode is a reducing agent.
2. Every cathode lives in equilibrium with its environment.

* An invited paper presented at the XII General Assembly of the URSI, August 30, 1957, at Boulder, Colorado; URSI reference: 1957/No. 147/Commission VII.

3. Every cathode is a dispenser cathode.
4. Monolayer film cathodes do not exist.

It is unlikely that all of these propositions can be proved, and no attempt is made to do so. However, they are discussed to give them at least an air of reality.

## Every Cathode is a Reducing Agent

That every cathode is a reducing agent is a mere matter of definition. Pauling states "Oxidation is the removal of electrons from an atom or group of atoms. Reduction is the addition of electrons to an atom or group of atoms" and "An atom, molecule, or ion which takes up electrons is called an oxidizing agent and one which liberates electrons is called a reducing agent."[1] One further quotation is of interest, "Every electron reaction involves an oxidizing agent and a reducing agent, which are closely related to one another." According to these definitions, then, the transfer of an electron from the cathode to the anode in a diode is an oxidation–reduction process. The merit of viewing electron emission as an oxidation process is that it emphasizes the chemical aspect of the process and focuses attention on the chemistry of the cathode as well.

The "strength" of a reducing agent is measured by its chemical activity, a quantity which determines, in part, the extent to which the agent will reduce another substance. The chemists have a convenient formalism, the mass-action law, for computing the results of a reaction. Consider the reaction.

$$aA + bB \rightleftarrows cC + dD. \tag{1}$$

When the reaction has reached equilibrium, and provided that no reactant is exhausted before equilibrium is reached, the mass action law states that

$$\frac{[C]^c\,[D]^d}{[A]^a\,[B]^b} = K, \tag{2}$$

in which the bracketed quantities are the activities. K is the equilibrium constant and is given by

$$K = e^{\Delta H/kT} \tag{3}$$

[1] Linus Pauling, *General Chemistry*, W. H. Freeman and Co., San Francisco, California, 1947.

where $\Delta H$ is the heat of reaction. For ideal systems, the activities are proportional to the concentrations or partial pressures of the reactants and products referred to some standard reference state. For our purposes, it will be convenient to use the concentrations relative to the "density of states" in the gas phase as activities, i.e.,

$$[n] = \frac{n}{N_c} \tag{4}$$

where

$$N_c = 2\left(\frac{2\pi m kT}{h^2}\right)^{3/2} \tag{5}$$

This formalism, which is tantamount to making the free energy of the system a minimum, is so convenient that it should see wider use amongst physicists. As an example of its use, consider a set of donors D, of concentration $N_D$, lying at an energy $\epsilon$ below the bottom of the conduction band of an insulator, in equilibrium with electrons in the conduction band. The electrons are obtained by the reaction

$$D \rightleftarrows D^+ + e. \tag{6}$$

Then the mass-action low states that

$$\frac{[D^+]\,[e]}{[D]} = \frac{\left(\frac{N_{D^+}}{N_c}\right)\left(\frac{n}{N_c}\right)}{\left(\frac{N_D}{N_c}\right)} = e^{-\epsilon/kT}. \tag{7}$$

Now the number of ionized donors $D+$ is equal to the number of free electrons $n$, so

$$n = \sqrt{N_D N_c}\, e^{-\epsilon/2kT}, \tag{8}$$

a familiar result which holds when $n \ll N_D$, i.e., when the donors are not exhausted before equilibrium is reached (if they are, $n = N_D$).

It will be noted that in the example above, the electrons were treated on the same basis as the other reactants and products and that their activity was written $n/N_c$, i.e.,

$$a_e'(s) = \frac{n}{N_c} = \sqrt{\frac{N_D}{N_c}}\, e^{-\epsilon/2kT} = e^{-\mu'/kT} \tag{9}$$

where $a_e'(s)$ signifies the activity of the electrons in the solid and $\mu'$ is the Fermi energy measured with respect to the bottom of the conduction band. The absolute activity is the activity referred to the gas phase. Therefore, the absolute activity of the electrons in the solid is

$$a_e(s) = a_e'(s) \exp\left[\frac{-X}{kT}\right] = \exp\left[\frac{-\mu' + X}{kT}\right], \tag{10}$$

where $X$ is the electron affinity of the insulator. But $\mu' + X = \mu$, the Fermi energy measured with respect to the "vacuum level," i.e., the work function of the insulator. Thus, the absolute activity of the electrons in a solid is determined solely by the work function of the solid. The "partial potential" or chemical potential $\mu$ is given by[2]

$$\mu = kT \ln a(g). \tag{11}$$

Thus, the Fermi level and chemical potential are identical for the electrons.

Because the "strength" of a reducing agent is measured by its electronic chemical potential, it is now possible to state the initial proposition in quantitative terms: Every cathode is a reducing agent, and at a given temperature its chemical activity increases in direct proportion to its emission.[3]

This conclusion is of prime importance because it implies that the better a cathode is, the more likely it is to react with oxidizing agents and lose its activity. This consideration leads naturally to the second proposition.

### Every Cathode Lives in Equilibrium with Its Environment

Having arrived at the view that the cathode is a chemically active reducing agent, it is pertinent to examine with what it can react. It can react with the base which supports it and with the gaseous ambient which surrounds it. Because the ambient gases are in part due to the

[2] R. W. Fowler and E. A. Guggenheim, *Statistical Thermodynamics*, Cambridge University Press, Cambridge, Mass., 1939.

[3] R. H. Plumlee, "The Electron Donor Centers in the Oxide Cathode," *RCA Review*, Vol. 17, p. 231, June, 1956.

envelope, the leads and all other electrodes, the cathode is, in effect, in equilibrium with everything within the vacuum envelope and, if the envelope is not impervious to everything, in equilibrium with the rest of the universe. This equilibrium is not a thermodynamic equilibrium because the system is not all at one temperature nor have all possible reactions run to completion. Nevertheless, the system is in a kind of dynamic equilibrium so that the system appears stable on a time scale of hours or thousands of hours. If this were not true, vacuum tubes would have no utility. Of course, it might be argued that because vacuum tubes are stable for long periods, no reactions at all take place. Then tubes would last indefinitely and this seems to be contrary to fact.

The reaction of a cathode with its supports is usually made trivial by using an inert support. In the case of the oxide cathode, the oxide is usually supported on nickel or platinum–iridium, neither of which reacts significantly with the oxide under normal operating conditions. The oxide does react with impurities diffusing out of the support. These impurities are usually reducing "metals" such as silicon and magnesium, and may be present in the support by accident or intent. Their role in the operation of a vacuum tube will be considered in the next section.

Vacuum tubes are usually pumped to a vacuum of $10^{-7}$ mm of Hg or better. It is interesting to note that at a pressure of $10^{-7}$ mm of Hg, about $3 \times 10^{-2}$ monolayers of atoms strike the cathode surface per second. If all of these atoms stuck to the surface and did not re-evaporate subsequently, the cathode thickness would increase about 0.005 centimeter per 1,000 hours. This illustration points up the necessity for considering the cathode in relation to the ambient gases.

The cathode contributes to its own ambient by evaporating its constituents. In the case of metallic cathodes, the evaporated metal condenses on the envelope, leads and other electrodes, and because of its low vapor pressure at the temperature of these surfaces usually plays no further role in the operation of the tube. In the case of cathodes such as the oxide cathode and L cathode, Ba and BaO evaporate. The evaporated Ba probably reacts with ambient $O_2$ to form BaO. Because of its relatively high vapor pressure, BaO may re-evaporate if the electrodes on which it rests run hot during operation. Furthermore, it is known that BaO dissociates under electron bombardment and releases $O_2$ which can oxidize the cathode and reduce its activity. The mechanism by which this occurs is not fully understood. A recent paper advances the hypothesis that electron bombard-

ment produces electron–hole pairs which result in a reduction of the dissociation energy.[4]

The ambient gases which surround an oxide cathode and may react with it are for the most part $O_2$, $H_2O$, $H_2$, CO, and $CO_2$ with traces of $CH_4$ and other hydrocarbon fragments.[5-7] That these ambients do react with the cathode is most strikingly demonstrated in the experiments of Plumlee who used a mass spectrometer to study the partial pressures of the various species in contact with an oxide cathode as a function of temperature, total pressure and the voltage applied to the diode of which the cathode formed a part.[7] A typical behavior is

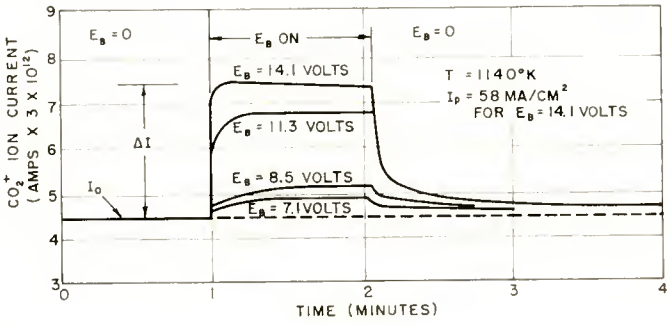



Fig. 1—The evolution of carbon dioxide from an oxide-coated cathode in a diode as a function of time, with anode voltage as a parameter (after Plumlee).

shown in Figure 1 which is taken from Plumlee's paper. The figure shows the variation of the partial pressure of $CO_2$ with time when anode voltage is applied to and removed from the diode. It will be noted that the partial pressure of $CO_2$ increases with the applied anode voltage. This is an example of the increase of chemical activity of a charged reactant with electric field, the most familiar manifestation of which is the Boltzmann law

$$n = n_o e^{-eV/kT}. \tag{12}$$

[4] S. Yoshida, N. Shibata, Y. Igarashi, and H. Arata, "Radioactive Isotope Study of the Dissociation of Barium Oxide under Electron Bombardment," *Jour. Appl. Phys.*, Vol. 27, p. 497, May, 1956.

[5] R. H. Plumlee and L. P. Smith, "Mass Spectrometric Study of Solids I. Preliminary Study of Sublimation Characteristics of Oxide Cathode Materials," *Jour. Appl. Phys.*, Vol. 21, p. 811, August, 1950.

[6] I. Pelchowitch, "A Study of the Evaporation Products of Alkaline Earth Oxides," *Philips Research Reports*, Vol. 9, p. 42, February, 1954.

[7] R. H. Plumlee, "Electrolytic Transport Phenomena in the Oxide Cathode," *RCA Review*, Vol. 17, p 190, June, 1956.

A similar behavior is displayed by the ambient $O_2$. The removal of oxygen from an oxide is a reduction process which should raise the Fermi level (chemical potential) of the oxide and make the cathode more active thermionically and chemically. The role of this reduction mechanism in actual cathodes will be discussed in the next section.

There is also evidence that the direct evaporation of reducing metals by the anode of a tube may effect the activity of a cathode.[8,9] In the cited papers, a spectrographic analysis of aged oxide coatings revealed traces of magnesium in the outer layers. This magnesium could have come from the anode, which contained 0.09 per cent magnesium, during the outgasing of the anode (3 minutes at 1,300°K) or during the life test of the tube if the anode ran at a temperature of 350°K or higher.[3] Whatever the source of the magnesium, and the evidence strongly suggests the anode, it is a strong reducing agent and its presence in the cathode must have contributed considerably to keeping the cathode in a reduced state, i.e., in a thermionically active state.

The observation above that the magnesium could have been evaporated by the anode during outgasing *or* during life points up another aspect of vacuum-tube chemistry, namely, the effect of processing. The chemists have a rule of thumb that states that near room temperature the rate of reaction doubles for every ten-degree increase of temperature. This is tantamount to saying that the activation energy of ordinary reactions is about 0.7 electron volt. If this activation energy is used to estimate the life equivalent of heating an electrode, it turns out that the outgasing at 1,300°K for 3 minutes of an electrode that normally runs at 500°K is the equivalent of about 1,000 hours of normal operating life. However, activation energies may be considerably higher, for example, the heat of sublimation of BaO is 3.82 electron volts.[10] Hence, the evaporation of BaO during a 3-minute "hot-shot" at 1,300°K is also equal to the evaporation during 1,000 hours of life at a cathode temperature of 1,000°K. These considerations emphasize the fact that the performance of a vacuum tube depends critically on the reactions which occur during the brief period devoted to processing, particularly in the commercial production of tubes.

---

[8] L. A. Wooten, G. E. Moore and W. G. Guldner, "Measurement of Excess Barium in Practical Oxide-Coated Cathodes," *Jour. Appl. Phys.*, Vol. 26, p. 937, August, 1955.

[9] G. E. Moore, L. A. Wooten and J. Morrison, "Excess Ba Content of Practical Oxide-Coated Cathodes and Thermionic Emission," *Jour. Appl. Phys.*, Vol. 26, p. 943, August, 1955.

[10] J. P. Blewett, H. A. Liebhafsky and E. F. Henelly, "The Vapor Pressure and Rate of Evaporation of Barium Oxide," *Jour. Chem. Phys.*, Vol. 7, p. 478, July, 1939.

The burden of this section has been that a cathode cannot be considered as an element apart in a vacuum tube. It communicates with every other part of the tube and its life and performance during life depend as much upon the constitution and condition of other electrodes and the envelope of the tube as upon its own initial constitution. It follows that it is pointless to produce a better cathode unless all other electrodes are improved equally because all electrodes including the cathode, will find a common level and the performance of each will be determined by all of the others.

## Every Cathode is a Dispenser Cathode

In view of the oxidizing ambient in contact with the cathode in a tube, some provision must be made to keep the cathode in a reduced state. This is done by providing a source of reducing potential. Because the system is not in thermodynamic equilibrium, but in dynamic equilibrium, reducing potential is lost by the cathode by evaporation to cooler electrodes where it is far less effective in maintaining cathode activity. Hence, to keep the cathode active, reducing potential must be supplied (dispensed) to the cathode continuously.

In the case of pure metallic cathodes, such as tungsten filaments, the metal provides its own reducing potential. At the temperatures required for adequate electron emission from tungsten, the mass-action law acts to keep the tungsten unoxidized in the face of substantial pressures of ambient oxygen. In the case of thoriated tungsten, the reducing potential is provided by the thoria included in the tungsten. The thoria is reduced by the tungsten, the thorium diffuses to the surface at a rate which just matches the loss by evaporation at the normal operating temperature.

Some time ago, Hull invented a "dispenser cathode," so-called because the dispensation is obvious.[11] Recently, this type of dispenser cathode has been greatly refined in the form of the L-cathode. Its dispensation has been thoroughly studied and the results have been published in a series of papers.[12,13]

In the case of the ordinary oxide cathode, the dispensation has

---

[11] A. W. Hull, "The Dispenser Cathode," *Phys. Rev.*, Vol. 56, p. 86, July, 1939.

[12] E. S. Rittner, R. H. Ahlert and W. C. Rutledge, "Studies of the Mechanism of Operation of the L Cathode I," *Jour. Appl. Phys.*, Vol. 28, p. 156, February, 1957.

[13] W. C. Rutledge and E. S. Rittner, "Studies of the Mechanism of Operation of the L Cathode II., *Jour. Appl. Phys.*, Vol. 28, p. 167, February, 1957.

been less obvious but it is now generally recognized that the "impurities" in the base metal constitute an important source of reducing potential.[14-18] Dispensation takes place by diffusion of the active materials out of the base metal. Studies of the diffusion rates of various reducing additives in nickel are in progress and, before long, one or more additives with diffusion rates appropriate to maintaining the oxide cathode at an acceptable level of activity for long periods will in all probability be incorporated in cathode base metals as a matter of course.

While reducing additives dispensed by the base metal of an oxide cathode are useful in affecting the initial reduction of the oxide during activation and in maintaining activity subsequently, they are not necessary, either to activate or to maintain activity. This is borne out by the use of "passive" nickels for base metals and by the achievement of active cathodes supported on ceramics.[19] In these instances, the oxide itself serves as a dispenser and the primary reducing potential is supplied by the anode power supply. It was noted earlier that the partial pressure of $O_2$ in a tube depends on the anode voltage of a diode.[7] This is shown in more detail in Figure 2. With the cathode in a relatively inactive state, the dependence of oxygen evolution on anode voltage is as shown by the upper curve. If the anode voltage is held at a fixed voltage, say 8 volts, the cathode activates and at the same time the rate of evolution of $O_2$ drops to a lower value as indicated by the dotted line. Then the dependence on anode voltage is as shown by the middle curve. Further activation by operation at a fixed anode voltage results in the lower curve. The field produced by the applied voltage increases the chemical activity of the oxygen in the BaO and leads to a partial elimination of $O_2$ by the oxide. This is tantamount to a reduction of the oxide and increases its thermionic activity. As

---

[14] H. E. Kern and R. T. Lynch, "Initial Emission and Life of a Planar-Type Diode as Related to the Effective Reducing Agent Content of the Cathode Nickel," *Phys. Rev.*, Vol. 82, p. 574, May, 1951.

[15] E. S. Rittner, "A Theoretical Study of the Chemistry of the Oxide Cathode," *Philips Research Reports*, Vol. 8, p. 184, June, 1953.

[16] R. W. Peterson, "Activation, Diffusion and Thermionic Emission in Oxide Coated Cathodes," *Phys. Rev.*, Vol. 99, p. 1651, September, 1955.

[17] J. J. Lander, H. E. Kern, and A. L. Blach, "The Solubility and Diffusion Coefficient of Carbon in Nickel," *Phys. Rev.*, Vol. 85, p. 389, January, 1952.

[18] R. W. Peterson, D. E. Anderson and W. G. Shepherd, "Influences of the Cathode Base on the Chemical Activation of Oxide Cathodes," *Jour. Appl. Phys.*, Vol. 28, p. 22, January, 1957.

[19] G. E. Moore and H. W. Allison, "Emission of Oxide Cathodes Supported on a Ceramic," *Jour. Appl. Phys.*, Vol. 27, p. 1316, November, 1956.

the cathode activates, the rate of evolution of $O_2$ decreases for a given voltage. Thus, the activation is self-limiting and the cathode seeks a level of activity appropriate to the demands on it. The variation of activity with current demand is quite conspicuous in transmitting tubes having large cathode areas (10-20 $cm^2$). If such tubes are put on the shelf for a month, the cathodes may become quite inactive. However, operation at a reduced power level will restore the cathodes to normal activity in a few minutes.

Thus, every practical cathode has within it a source of reducing potential. This potential may be brought to bear on cathode activity thermally, as in the case of metallic cathodes and "dispenser" cathodes

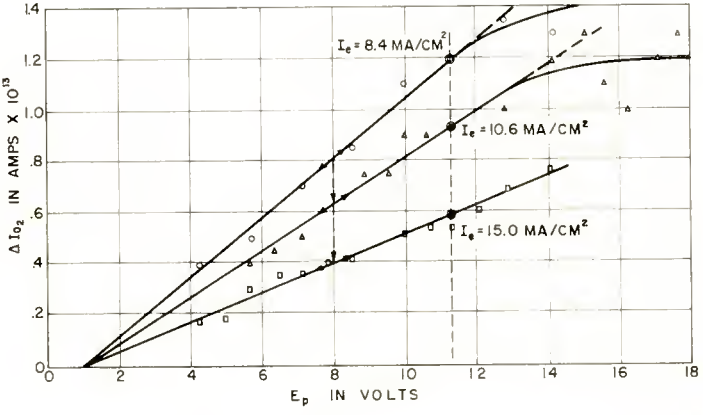



Fig. 2—The evolution of oxygen from an oxide-coated cathode in a diode as a function of anode voltage, with the state of activity of the cathode as a parameter (after Plumlee).

with diffusion sources, or electrically, as in the case of the ordinary oxide cathode.

## Monolayer Emitters Do Not Exist

The proposition that monolayer emitters do not exist is certainly controversial; it is hoped that it is also provocative. The proposition arose from speculation as to how experiments on thoriated-tungsten filaments would have been interpreted if these emitters had been discovered only recently, i.e., after the notions of how semiconducting cathodes operate had become fairly well established.

In the spirit of inquiry, suppose that the thoriated-tungsten cathode had just been discovered and that only pure metallic emitters and oxide-coated emitters, both BaO and $ThO_2$, were known. Then the

first thought might be to describe the new emitter in terms of a semiconductor model. The description might go as follows:

The thoriated cathode is prepared by including thoria (0.5 to 0.75 per cent by weight) in the tungsten. It is activated by flashing at 2,400°K for 1 to 2 minutes, then dropping the temperature to 2,100°-2,200°K for a brief period after which the emitter is ready for operation at 1,800°-2,000°K, with an emission density of about 1.8 amperes per square centimeter at the usual operating temperature of 1,950°K.[20] At the flashing temperature, the thoria probably dissociates, the oxygen and thorium diffuse to the surface and along the surface and recombine to form thoria crystallites at sites where epitaxial growth is possible. The growth of crystallites on hot substrates is well known.[21-34*] Refer-

[20] S. Dushman, "Thermionic Emission," *Rev. Mod. Phys.*, Vol. 2, p. 381, October, 1930.

[21] W. G. Burgers and C. J. Dippel, "On the Crystalline State of Thin Calciumfluoride Films," *Physica*, Vol. 1, p. 549, February, 1934.

[22] W. G. Burgers and J. J. A. Ploos van Amstel, "Oriented Oxidation of Barium," *Physica*, Vol. 3, p. 1057, December, 1936.

[23] H. A. Stahl, "The Structure of Extremely Thin Layers Evaporated in Kinetic Vacuum Systems," *Jour. Appl. Phys.*, Vol. 20, p. 1, January, 1949.

[24] H. A. Stahl, "The Thermal Behavior of Evaporated Layers in Vacuum Devices," *Jour. Appl. Phys.*, Vol. 20, p. 8, January, 1949.

[25] D. M. Evans and H. Wilman, "Crystal Growth and Orientation in Deposits Condensed from the Vapour," *Acta. Cryst.*, Vol. 5, p. 731, October, 1952.

[26] W. Friemel and I. N. Stranski, "On the Formation of Needle-Like Potassium Crystals," *Naturwissenschaften*, Vol. 43, p. 79, May, 1956.

[27] J. A. Becker, "The Use of the Field Emission Electron Microscope in Absorption of W on W and Ba on W," *Bell Sys. Tech. Jour.*, Vol. 30, p. 907, October, Part I, 1951.

[28] P. N. Russel and A. Eisenstein, "Thermionic Emission and Electron Diffraction from Thin Fiims of Barium Oxide," *Jour. Appl. Phys.*, Vol. 25, p. 954, August, 1954.

[29] W. Buckel and R. Hilsch, "Superconductivity and Resistance of Tin With Lattice Faults," *Zeitschrift fur Physik*, No. 3, Vol. 131, p. 420, 1952.

[30] W. Buckel, "Electron-Diffraction Photographs of Thin Metal Layers at Low Temperatures," *Zeitschrift fur Physik*, No. 2, Vol. 138, p. 136, 1954.

[31] W. Buckel and R. Hilsch, "Influence of Low-Temperature Condensation on the Electrical Resistance and Superconductivity of Various Metals," *Zeitschrift fur Physik*, No. 2, Vol. 138, p. 109, 1954.

[32] H. Bülow and W. Buckel, "Electron-Diffraction Pictures of Thin Metallic Layers at Low Temperatures," *Zeitschrift fur Physik*, No. 2, Vol. 145, p. 141, 1956.

[33] W. Buckel and R. Hilsch, "Superconductivity and Electrical Resistance of Novel Tin-Bismuth Alloys," *Zeitschrift fur Physik*, No. 1, Vol. 146, p. 27, 1956.

[34] H. Bülow, "Electron-Diffraction Pictures of Novel Tin-Bismuth Alloys," *Zeitschrift fur Physik*, No. 3, Vol. 148, p. 321, 1957.

* This list is by no means exhaustive.

ences (27) and (28) are of particular interest in the present connection.

Becker has studied the adsorption and migration of Ba on W in the field-emission microscope.[27] He finds that Ba starts migrating at temperatures ranging from 370°K for the (110) plane to 800°K for the (100) plane. The temperature for evaporation depends on the surface coverage and ranges from 1,050° to 1,600°K on the (110) plane. He also finds that crystallites form at about 300°K on all of the crystal faces studied ((110), (211), (100), (111), (611)) and that they disappear at 600°K for the (110) and (211) faces and at 800°K for the (111) and (611) faces. The crystallite sizes range from 200 to 600 Å with a median of about 400 Å.

Russell and Eisenstein have studied the thermionic emission and structure of BaO evaporated onto a nickel surface.[28] The thermionic activity increased with deposition until approximately 20 monolayers (measured by a radioactive tracer technique) had been deposited and showed no increase for the next thirty monolayers. The emission density of these films was comparable to that of ordinary oxide cathodes (0.4 amp $cm^{-2}$ at 1,000°K). BaO deposited with the receiver below 450°K and subsequently heated to 800°K for a short time and BaO deposited at 800°K showed the diffraction patterns characteristic of BaO. Subsequent heating of the crystalline films to 1,070°K or above caused them to become amorphous and the amorphous form persisted as oxide was removed by evaporation and was accompanied by decreasing thermionic emission. The highest emission was obtained from a crystalline film. The work function of this film was about 1.3 electron volts, a value comparable to that of ordinary oxide cathodes.

These two papers are particularly interesting because the first gives direct visual evidence of surface migration and crystal growth and the second correlates thermionic emission with crystal growth. It appears that BaO crystallites on nickel are much more stable than Ba crystallites (melting point = 983°K) on tungsten. This is to be expected because of the high heat of dissociation of BaO. It might be expected that Ba on W might form BaO by reaction with ambient $O_2$. This possibility was precluded by the very low pressures in Becker's tube—"the vacuum conditions were improved to such an extent that the residual gas produced only barely detectable effects after about one week."

Schottky plots of the emission from thoriated tungsten show marked deviations from the Schottky law.[35] In this respect they resemble oxide

[35] K. T. Compton and I. Langmuir, "Electrical Discharges in Gases, Part I," *Rev. Mod. Phys.*, Vol. 2, p. 123, April, 1930.

cathodes. Instead of plotting the current–voltage characteristic on semilogarithmic paper, as for Schottky plots, it is interesting to plot them on log-log paper. Then, in a large number of cases, they are linear with a slope of one-half as shown in Figure 3. The current–voltage relation,

$$i = \alpha V^{1/2}, \tag{13}$$

is suggestive of the conduction mechanism in a semiconductor when the field is high enough so that the electrons become "hot," i.e., so that the drift velocity of the electrons exceeds their mean thermal velocity. Thus the form of the current–voltage characteristic in the high-field region is not inconsistent with a semiconductor model.

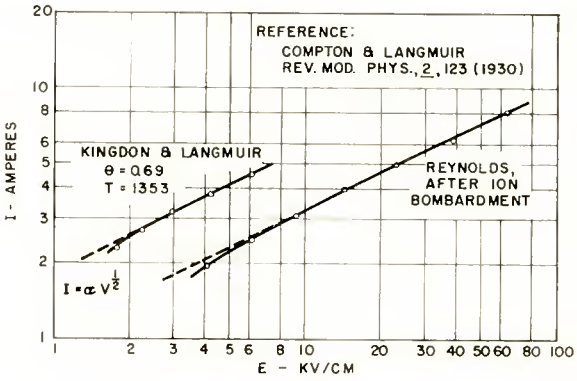



Fig. 3—Schottky-effect data on thoriated-tungsten plotted on logarithmic paper.

The work function of a thoriated tungsten cathode depends on the state of activity of the cathode. Langmuir has defined a parameter $\theta$ to describe the state of activity. The parameter is given by the relation

$$\theta = \frac{\phi_\theta - \phi_w}{\phi_{th} - \phi_w} \tag{14}$$

where $\phi_w$ = the work function of clean tungsten
= 4.53 electron volts,
$\phi_{th}$ = the minimum work function of a thoriated surface
= 2.63 electron volts,
$\phi_\theta$ = the work function of a partially activated surface.

The parameter $\theta$ is known as the "coverage" of the tungsten surface but the relation of $\theta$ to the actual coverage of the surface is obscure.

Kingdon has matched the emission characteristics of thoriated emitters approximately by the empirical expression[36]

$$\frac{i}{T^2} = (a_1 e^{\theta} + a_2 e^{1-\theta} - 1) \exp\left[ -\frac{e}{kT} (\phi_w - \theta [\phi_N - \phi_{th}]) \right]. \tag{15}$$

If $\theta$ represented an actual coverage, one would expect it to appear as a linear multiplier of the exponential. It is interesting to note that the minimum work function $\phi_{th}$ is the work function of fully activated thoria cathodes.[37]

Thoria cathodes have been studied in detail.[37–40] They have a work function of 2.6 electron volts and display two states of activity,[37] one of which is stable and has an A-constant of 2.6 amperes $cm^{-2}$ $deg^{-2}$, and one which is unstable under current flow and has an A-constant of 8-16 amperes $cm^{-2}$ $deg^{-2}$. The conductivity of thoria has activation energies ranging from 0.58-3.2 electron volts.[38,39] Choosing a value of 1.1 electron volts[38] as representative of a stable cathode makes the electron affinity 1.5 electron volts. Using this value it is interesting to compute the electron concentration in the conduction band and the coverage of the surface of a thoriated-tungsten cathode assuming that it is in fact a thoria cathode. The data of Dushman and Ewald serve admirably.[20] They list the emission current, $I_\theta$, the A-constant, $A_\theta$, and the work function, $\phi_\theta$, as a function of $\theta$. Their tabulation is shown in Figure 4. The computation proceeds as follows: The emission equation can be written in the following forms

$$I_\theta = A_\theta T^2 \exp\left[ \frac{-e\phi_\theta}{kT} \right] \tag{16}$$

$$= \sigma n\, e\, \bar{v} \exp\left[ \frac{-eX}{kT} \right] \tag{17}$$

---

[36] K. H. Kingdon, "Electron Emission From Adsorbed Films on Tungsten," *Phys. Rev.*, Vol. 24, p. 510, November, 1924.

[37] W. E. Danforth, "Thorium Oxide and Electronics," *Advances in Electronics*, Vol. V, Academic Press, Inc., New York, New York, 1953, p. 169.

[38] D. A. Wright, "Thermionic Emission From Oxide Coated Cathodes," *Proc. Phys. Soc.*, Vol. 62B, p. 188, March, 1949.

[39] W. E. Danforth and F. H. Morgan, "Electrical Resistance of Thoria," *Phys. Rev.*, Vol. 79, p. 142, July, 1950.

[40] D. A. Wright, "A Survey of Present Knowledge of Thermionic Emitters," *Proc. Inst. Elec. Eng.*, Vol. 100, p. 125, May, 1953.

$$= \sigma e \bar{v} N_c \exp\left[\frac{-e\phi_\theta}{kT}\right] \tag{18}$$

neglecting any reflection coefficient. In these expressions

$\sigma$ = the fractional coverage of the surface,
$n$ = the density of electrons in the conduction band,
$\bar{v}$ = the mean thermal velocity of the electrons normal to the emitting surface.

The density of electrons in the conduction band is given by

$$n = N_c \exp\frac{\epsilon_F}{kT} \tag{19}$$

DATA OF DUSHMAN AND EWALD
REV. MOD. PHYS., 2, 381 (1930)

| $\theta$ | $I_\theta$ | $A_\theta$ | $\phi_\theta$ | $\xi_F$ | n | $\sigma n$ | $\sigma$ |
|---|---|---|---|---|---|---|---|
| 1. | 1.166 | 3.0 | 2.63 | 1.13 | $4.46^{17}$ | $5.13^{15}$ | $1.15^{-2}$ |
| 0.95 | .349 | 1.50 | 2.71 | 1.21 | $2.49^{17}$ | $1.55^{15}$ | $6.23^{-3}$ |
| 0.83 | .118 | 2.08 | 2.94 | 1.44 | $6.13^{16}$ | $5.24^{14}$ | $8.55^{-3}$ |
| 0.72 | .0594 | 3.74 | 3.15 | 1.65 | $1.71^{16}$ | $2.64^{14}$ | $1.54^{-2}$ |
| 0.56 | .0195 | 7.76 | 3.45 | 1.95 | $2.69^{15}$ | $8.67^{13}$ | $3.22^{-2}$ |
| 0.43 | .0064 | 10.86 | 3.69 | 2.19 | $6.33^{14}$ | $2.84^{13}$ | $4.48^{-2}$ |
| 0.25 | .0010 | 15.81 | 4.06 | 2.56 | $6.61^{13}$ | $4.44^{12}$ | $6.73^{-2}$ |
| 0.18 | | (8.2) | 4.17 | 2.67 | $3.34^{13}$ | | |
| AVERAGE | | 6.39 | | | | | $2.66^{-2}$ |

$$\overline{A_\theta} = \frac{\overline{\sigma} e \overline{v} N_c}{T^2} = 6.31$$

Fig. 4—Thoriated-tungsten emission data interpreted in terms of a semiconductor model (see text).

where $\epsilon_F$ is the Fermi energy measured from the bottom of the conduction band. It is given by

$$\epsilon_F = \phi_\theta - X. \tag{20}$$

The product of $n$ and $\sigma$ is obtained from Equation (17)

$$n\sigma = \frac{I_\theta}{e\bar{v}} \exp\frac{eX}{kT}. \tag{21}$$

Having computed $n$ and $n\sigma$, $\sigma$ can be computed. The results of the computation are shown in Figure 4. Equations (16) and (18) may be combined to yield

$$A_\theta = \frac{\sigma e \bar{v} N_c}{T^2}. \tag{22}$$

If the mean value of $\sigma$ as computed in Figure 4 is substituted in this relation, the result is

$$\bar{A}_\theta = 6.31,$$

in reasonable agreement with the value obtained by averaging Dushman and Ewald's $A_\theta$'s ($\bar{A}_\theta = 6.39$). This computation then leads to an average fractional coverage of 2.7 per cent and a dependence of the

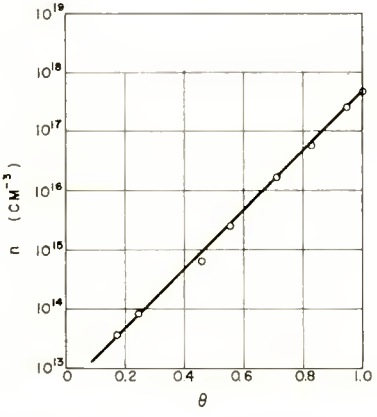



Fig. 5—The density of conduction electrons versus Langmuir's coverage factor ($\theta$) assuming a semiconductor model for thoriated-tungsten emitters.

carrier concentration in the conduction band on $\theta$ as shown in Figure 5. These concentrations are in the range normal for impurity semiconductors.

To examine the relation between the size of crystallites and a monolayer of coverage, assume one monolayer is consumed in forming the crystallites. The lattice constant of $ThO_2$ is 3.3 Å and 1 centimeter of tungsten surface is covered by $7.6 \times 10^{14}$ atoms of thorium.[20] If $n$ crystallites each containing $N$ molecules of thoria form, then

$$n N = 7.6 \times 10^{14}. \tag{23}$$

If the crystallites are cubes, the length of each side will be*

* The crystallites will not necessarily be cubes, but will take on the shape giving the minimum of free energy.

$$l = 3.3 \times 10^{-8} N^{1/3}, \tag{24}$$

and the fractional coverage is

$$\sigma = nl^2, \tag{25}$$

assuming that only the outermost surface contributes to the emission. These equations may be solved simultaneously to yield

$$l = \frac{2.73}{\sigma} \times 10^{-8}. \tag{26}$$

Thus, if $\sigma = 2.7 \times 10^{-2}$, the edge of crystallite has a length of about

$$l = 100 \text{ Å}. \tag{27}$$

This is of the order of magnitude found by Becker for Ba on W.[27] Then each crystallite contains about $3 \times 10^4$ molecules and there are about $3 \times 10^{10}$ crystallites per square centimeter of surface.

In passing, it is interesting to note that Fan's value for the heat of sublimation of $ThO_2$ (184 kilocalories per mole)[41] is not much different from that cited by Langmuir for the evaporation of Th from W (204 kilocalories per mole).[42]

This interpretation of thoriated-tungsten data in terms of a semiconductor model has the following merits:

1. It accounts for the equality of the work functions of "fully covered" thoriated-tungsten surfaces ($\theta = 1$) and thoria cathodes.

2. It provides a rational interpretation of $\theta$ as a measure of the position of the Fermi level in a semiconductor.

3. It leads to carrier concentrations in the semiconducting cathode which are in accord with those found in $ThO_2$ and BaO cathodes.

The thoriated-tungsten cathode is not the only cathode whose performance can be re-examined in the light of modern solid-state models. Others are the Ba on W cathode of Becker,[43] the alkaline earth thin

---

[41] H. Y. Fan, "Thermionic Emission from Sintered Cathode of Thoria and Tungsten Mixture," *Jour. Appl. Phys.*, Vol. 20, p. 682, July, 1949.

[42] I. Langmuir, "The Electron Emission from Thoriated Tungsten Filaments," *Phys. Rev.*, Vol. 22, p. 357, October, 1923.

[43] J. A. Becker, "Phenomena in Oxide-Coated Filaments," *Phys. Rev.*, Vol. 34, p. 1323, November, 1929.

films of Moore and Allison,[14] the L-cathode, and the E-cathode.[15]

In 1939 Becker published his well-known curve of the emission of Ba on a tungsten surface versus fractional surface coverage $\theta$.[13] The curve is reproduced in Figure 6. He found a maximum of emission and defined $\theta$ as unity at the maximum. If this cathode is interpreted as a semiconducting cathode, the semiconductor is probably BaO formed from the deposited Ba and ambient $O_2$. A simple computation shows that with an accommodation constant of unity and an oxygen vapor pressure of $10^{-8}$ mm of Hg, a monolayer of $O_2$ will be deposited on the surface in about 10 seconds. Hence, there should be no difficulty

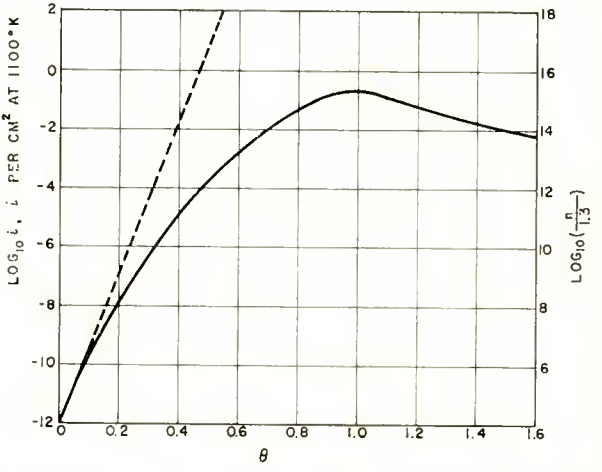



Fig. 6—J. A. Becker's curve of the emission of a barium-on-tungsten emitter versus fractional coverage.

about an adequate supply of oxygen to form the oxide even at lower partial pressures of $O_2$. By a computation similar to that for thoriated tungsten, it is found that

$$n\sigma = 2.11 \times 10^{14} I. \tag{28}$$

If the actual emitter is BaO, the density of electrons in the conduction band is given by

$$n = \sqrt{N_D N_c} \exp\left[\frac{-e\epsilon_D}{kT}\right], \tag{29}$$

[14] G. E. Moore and H. W. Allison, "Thermionic Emission of Thin Films of Alkaline Earth Oxides by Evaporation," *Phys. Rev.*, Vol. 77, p. 246, January, 1950.

[15] G. J. Ekkers, "The E-Cathode," presented at the Fourteenth Annual Conference on Electron Tube Research, Boulder, Colorado, June, 1955.

where $N_D$ = density of donors,
$\epsilon_D$ = energy below the conduction band of donors.

According to DeVore, $\epsilon_D \approx 1.4$ electron volts for BaO.[46] At maximum activation a donor density of about $5 \times 10^{17}$ per cubic centimeter is reasonable. Then, the free electron concentration for $\theta = 1$ is about

$$n \approx 2.6 \times 10^{15},$$

and the coverage is about

$$\sigma \approx 1.6 \times 10^{-2}.$$

The density of conduction electrons varies with $\theta$ as shown by Becker's curve with the scale shown on the right of Figure 6.

If the crystallites are formed from a monolayer of BaO and the crystallites are cubes, there are about $10^9$ crystallites per square centimeter of surface each containing about $10^5$ molecules with a cube edge of about 350 Å (compare with Becker's crystallites of Ba on W with a mean dimension of 400Å).

Moore and Allison have studied the thermiomic emission versus coverage of BaO on tungsten molybdenum, tantalum, and zirconium.[44] They used radioactive tracer techniques to determine the "coverage," so there is no uncertainty in the amount of material deposited. For tungsten, molybdenum and tantalum the behavior was almost identical; deposition of BaO reduced the work function of the surface at a rate of about 40 volts per monolayer down to the normal work function of an oxide coating ($\phi \approx 1.0$ electron volts), i.e., *one tenth of a monolayer displayed the work function of an ordinary oxide coating.* The effect of deposition on the A-constant was somewhat erratic, but in general $A$ decreased from about $10^2$ for a clean surface to $10^{-2}$ for one monolayer of deposit. Retarding-field measurements were made to determine the shift in contact potential with deposition. The work functions obtained in this way agree with those obtained from Richardson plots. The retarding field curves show departures from the Boltzmann law of the kind studied by Ferris.[47] One of these curves is shown in Figure 7. It will be noted that the departure from the Boltzmann law $\Delta V$ is linear with galvanometer deflection up to a deflection of about 40 centimeters. This is in itself suggestive of cathode-coating resistance.

---

[46] H. B. DeVore, "Photoconductivity Study of Activation of Barium Oxide," *RCA Review*, Vol. XIII, p. 453, December, 1952.

[47] W. R. Ferris, "Some Characteristics of Diodes with Oxide-Coated Cathodes," *RCA Review*, Vol. X, p. 134, March, 1949.

The sensitivity of the galvanometer is not given but the internal evidence is that the sensitivity was $10^{-8}$ ampere per centimeter. If this is correct, the resistance is about $10^6$ ohms. The diameter of the filament is not given but the heating–current–temperature curve suggests a diameter of 0.0025 inch and the active length is given as 1.5 inch. These dimensions lead to an emitting area of 0.076 square centimeter. If the coverage of a surface by one monolayer is about 2 per cent, as previously computed, then the actual emitting area in this case is

$$A_e = 0.076 \times 2 \times 10^{-2} \times 0.22 = 3.3 \times 10^{-4} \text{ square centimeter,}$$

and the resistance per square centimeter of emitting area is

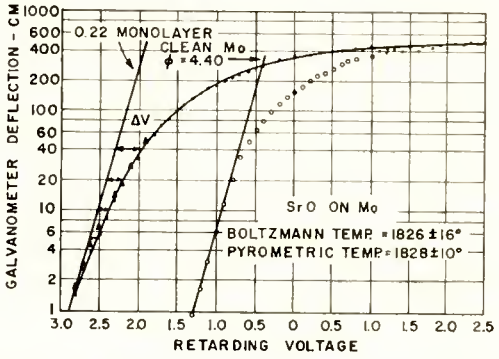



Fig. 7—Retarding-field curves showing the effect of strontium-oxide deposited on clean molybdenum (after Moore and Allison).

$$R = 10^6 \times 3.3 \times 10^{-4} = 300 \text{ ohms.}$$

This is not unusual for oxide-coated cathodes.

The presence of cathode resistance is revealing; a monolayer is easily tunneled by electrons and should display no resistance. In their conclusion Moore and Allison state: "Within the uncertainty of the experiment, the work functions obtained agree quantitatively with those published by others for bulk oxide coatings, which are ordinarily explained in terms of their properties as semiconductors. There is at present no theoretical reason for expecting such agreement." This section of the present paper adopts the attitude that there *are* theoretical reasons for expecting agreement.

The L-cathode is usually construed as a barium-on-oxygen-on-tungsten emitter.[12] For temperatures above 1,250°K, the current–

voltage characteristics of diodes with L-cathodes do indeed satisfy the Child–Langmuir (corrected for initial velocities) within experimental accuracy. At lower temperatures, however, the current–voltage characteristics depart from the 3/2 law in a manner very similar to the departures shown by oxide cathodes and known to be due to oxide resistance. A set of characteristics at lower temperatures taken on a diode with an L-cathode, is shown in Figure 8. As the temperature is reduced, the characteristics depart more and more from a 3/2 law

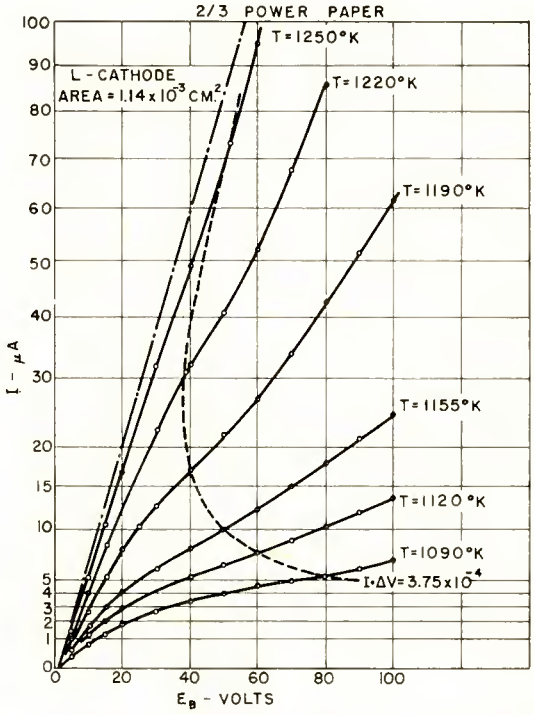



Fig. 8—Current–voltage characteristics of a diode with an L-cathode with cathode temperature as a parameter.

but in no case show the asymptotic behavior tube expected at emission saturation. In fact, the curves show inflection points above which they break upwards. If it is assumed that the true perveance line of the diode is given by the perveance computed from the diode geometry, the dot-dash line of Figure 8, and that departures from this line result from voltage drop in the cathode film, then the inflection points may be connected, approximately, by a line of constant film dissipation as shown by the dotted curve. A possible inference is that the inflection

points correspond to a dissipation (about 0.33 watt per square centimeter) which begins to raise the film temperature above that measured pyrometrically without plate current.

Again assuming the dot-dash line to be the true perveance line and assuming departures from the line to be due to film resistance, the current–voltage characteristics of the film are as shown in Figure 9. Within the accuracy of the data they are linear at low voltages but break upward at about the line of constant dissipation of Figure 8,

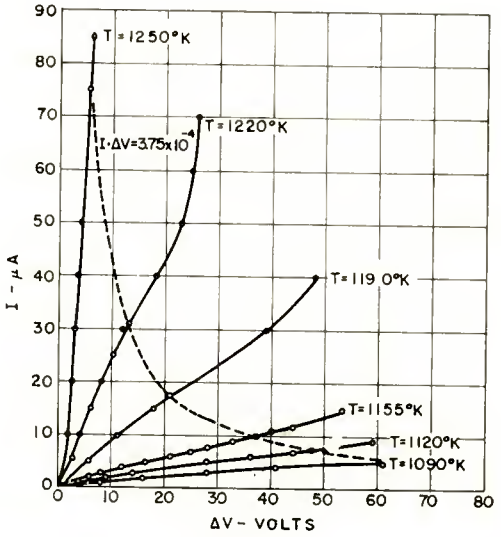



Fig. 9—Current–voltage characteristics of the "coating" of the L-cathode of Figure 8, assuming that departures from the perveance line are due to voltage drop within the "coating."

as expected. Whereas an error in estimation of the perveance line results in minor errors in computing film resistance at low temperatures, it makes a large difference for the higher temperatures where the deviations are small. Furthermore, the perveance changes somewhat with temperature because of the thermal expansion of the cathode support. Hence no reliable estimate of the resistance activation energy is possible.

Impregnated cathodes show similar departures from the Child–Langmuir law before emission saturation sets in.[18]

[18] R. Levi, "Improved 'Impregnated' Cathode," *Jour. Appl. Phys.*, Vol. 26, p. 639, May, 1955.

In June 1956, Ekkers described a new cathode prepared by arcing nickel, tungsten or molybdenum in a rare gas plus a small amount of oxygen.[45] The cathode was reported to be stable and to give an emission of up to 800 amperes per square centimeter at 1,600°C. When the condensable vapors are frozen out in a DE-2 (a Brown–Boveri high-voltage rectifier with an E-cathode) by immersing the tube in liquid nitrogen, we find a less spectacular emission. Then diode characteristics have the form shown by BaO cathodes and the currents are of the order of milliamperes. The activation energy of the apparent resistance turns out to be 0.7-0.8 electron volt, a value found in alkaline-earth oxide cathodes. X-ray analyses of residues from the cathode and surfaces of other electrodes facing the cathode show the presence of barium, probably evaporated from the barium getter used in these tubes. If, in accord with the attitude adopted in this section, the emission is attributed to crystallites of BaO, the fractional surface coverage is about $3 \times 10^{-3}$. E-cathodes removed from rectifiers to vacuum diodes show a behavior similar to rectifiers with the condensable vapors frozen out. The characteristics of such a diode are shown in Figure 10. The inference to be drawn in the present context is obvious. As to the emission mechanism in the presence of mercury vapor, this is believed to have no direct connection with thermionic emission and will be discussed elsewhere.

As a final example of "monolayer films," consider the recent study of the adsorption of strontium and barium on tungsten by Moore and Allison.[49] They studied the thermionic activity and rates of desorption of strontium and barium on tungsten and determined the amount of metal on their emitters by radioactive tracer techniques, so their "coverage" $\theta$ ($\theta = 1$ equals one monolayer) refers to the actual amount of material present. They find a change of work function with $\theta$ as shown in Figure 11. As in the case of alkaline-earth oxides on tungsten, they find that the major change in work function occurs for the first two tenths of a monolayer. They attribute this change in work function to the formation of a dipole layer on the surface. This dipole layer is not ionic but arises from the polarization of the adsorbed atoms by the underlying metal and their neighbors.

To interpret this result in terms of a semiconductor model, it is assumed that the deposited metal reacts with ambient $O_2$ and forms oxide crystallites. Consider the case of barium. In the inactive state the conductivity of BaO exhibits an activation energy of 3.9 electron

[49] G. E. Moore and H. W. Allison, "Adsorption of Strontium and Barium on Tungsten," *Jour. Chem. Phys.*, Vol. 23, p. 1609, September, 1955.

volts for temperatures below about 1,200°K.[50] The band-gap of BaO is about 5 electron volts.[51] Then a density of traps or acceptors of about $1.5 \times 10^{14}$ per cubic centimeter lying 2 volts above the valence band gives an activation energy of 3.9 electron volts. The electron affinity of BaO is about 0.5 electron volt,[51] so the work function in the inactive state is about 4.4 electron volts, i.e., roughly the work

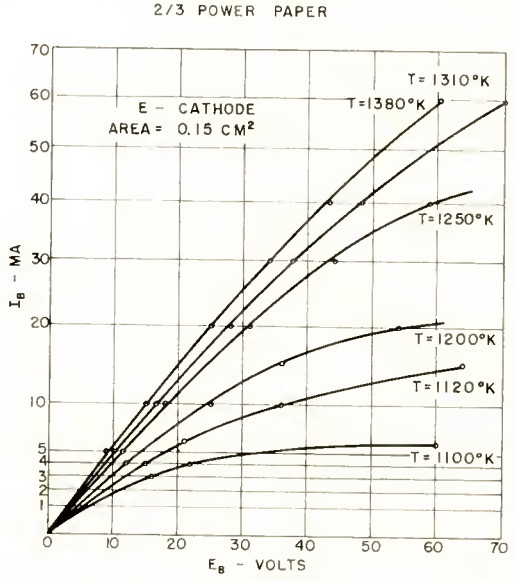



Fig. 10—Current–voltage characteristics of an E-cathode in a vacuum diode.

function of tungsten. Suppose the oxide is now activated by subjecting it to excess Ba. Then blue color centers acting as donors form.[52,53] These lie 2.0 electron volts below the conduction band. As their number is increased the Fermi level $\epsilon_F$ and work function $\phi$ drop rapidly as shown in Figure 12. When the density of color centers $N_D$ exceeds the trap density by a factor of three, the work function has dropped to about 2.4 electron volts. In Moore and Allison's experiment, $\theta = 1$ gives

[50] N. B. Hannay, D. McNair and A. H. White, "Semi-conducting Properties in Oxide Cathodes," *Jour. Appl. Phys.*, Vol. 20, p. 669, July, 1949.

[51] L. S. Nergaard, "Electron and Ion Motion in Oxide Cathodes," Halbleiter Probleme III, F. Vieweg and Sohn, Braunschweig, Germany.

[52] W. C. Dash, "Optical Absorption and Photoconduction in the Visible and Near Infrared in Single Crystals of BaO," *Phys. Rev.*, Vol. 92, p. 68, January, 1953.

[53] R. L. Sproull, R. S. Bever and G. Libowitz, "Oxygen Vacancies in Barium Oxide," *Phys. Rev.*, Vol. 92, p. 77, January, 1953.

$\phi \approx 2.2$ electron volts for Ba. According to the present interpretation, BaO crystallites form on the deposition of Ba and the crystallites are at the same time activated by excess Ba. According to Timmer, the vapor pressure of Ba required to maintain the blue-color-center density which gives a work function of 2.4 electron volts is about $5 \times 10^{-14}$ mm of Hg, certainly not a prohibitively large pressure.[54]

Moore and Allison have compared their desorption experiments with the desorption theory of Langmuir and with an equation from the

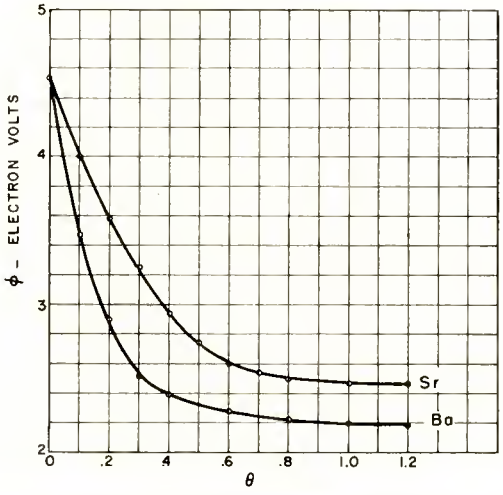



Fig. 11—The work functions of barium- and strontium-covered tungsten emitters as a function of fractional coverage (computed from data of Moore and Allison).

theory of absolute reaction rates.[55] They conclude that requirements have not reached the refinement necessary to choose between theories. The latter theory is particularly interesting because it contains only two disposable constants and it fits the experimental results very well. Moore and Allison use the equation in the form

$$-\frac{d\theta}{dt} = \frac{kT}{h}\,\theta \exp\left(\frac{-\epsilon + 4\theta V}{kT}\right). \tag{30}$$

For Ba, they find for the disposable constants

[54] C. Timmer, "Density of Color Centers in Barium Oxide as a Function of the Vapor Pressure of Barium," *Jour. Appl. Phys.*, Vol. 28, p. 495, April, 1957.

[55] S. Glasstone, K. Laidler and H. Eyring, "The Theory of Rate Processes," McGraw-Hill Book Co., Inc., New York, New York, 1941.

$$\epsilon = 3.68 \text{ electron volts,}$$

$$V = 0.13 \text{ electron volt.}$$

The constant $\epsilon$ is the activation energy for desorption from a clean surface and V is the repulsive energy per pair of adatoms. In the present interpretation, $\epsilon - 4V$ (3.16 electron volts) is the heat of sublimation of BaO on tungsten (3.82 electron volts for bulk BaO[10]) and $4\theta V$ represents the increase in binding energy as the volume of the crystallites approaches zero (the heat of dissociation of BaO vapor is about 6.4 electron volts).[56]

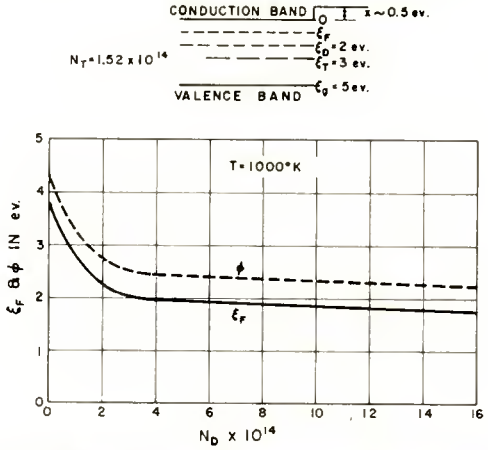



Fig. 12—Top—an energy-level diagram for a barium-oxide emitter. Bottom—the work function and Fermi level of the barium-oxide emitter as a function of donor density.

## Conclusion

This paper has set forth a number of propositions which the writer hopes will elicit discussion and will lead to a broader view of thermionic emitters and their limitations. The proposition that monolayer emitters do not exist has been documented in some detail. In some cases the evidence for semiconductor emission, rather than monolayer emission, is convincing, at least for considerable temperature ranges. In other cases, the evidence is not so convincing. However, the present cursory examination of the problem suggests that further study of aggregation on surfaces and electrical properties of small aggregates on surfaces is in order. Only after such studies have been made will it be possible to describe the emission mechanisms of many existing cathodes unambiguously.

[56] J. P. Blewett, "The Properties of Oxide-Coated Cathodes," *Jour. Appl. Phys.*, Vol. 10, p. 668, October, 1939.

# INFLUENCE OF HYDRATION-DEHYDRATION OF THE GERMANIUM OXIDE LAYER ON THE CHARACTERISTICS OF P-N-P TRANSISTORS

By

J. Torkel Wallmark* and R. R. Johnson†

RCA Laboratories,
Princeton, N. J.

***Summary***—*It has been found that when germanium p-n-p transistors are subjected to a change in temperature, the current-transfer ratio, $\alpha_{fe}$, shows a corresponding change, approaching an asymptotic value in approximately 48 hours. Simultaneously the saturation current of the emitter and collector junctions show a similar change. In this paper, this effect is interpreted in terms of a hydrated oxide layer on the germanium surface.*

*The effect can be reduced either by minimizing the amount of water inside the transistor enclosure, or by using a drying agent (a getter) in the enclosure with a water retention that is considerably stronger than that of the germanium oxide layer.*

## Introduction

THE influence of water vapor on the characteristics of the germanium surface has long been studied to obtain information about the surface layer and its effects on the stability characteristics of transistors. A review of such studies has been published by Kingston.[1] Even minute traces of water produce such large effects that hermetic encapsulation of transistors in a dry surrounding had to be introduced before their characteristics could be made reasonably stable.

As shown by Bardeen and Brattain,[2] and later elaborated by others,[3-5] water vapor introduces positive trapped charge on the ger-

---

* RCA Laboratories, Princeton, N. J.

† RCA Semiconductor Division, Somerville, N. J.

[1] R. H. Kingston, "Review of Germanium Surface Phenomena," *Jour. Appl. Phys.*, Vol. 27, p. 101, February, 1956.

[2] J. Bardeen and W. H. Brattain, "Surface Properties of Germanium," *Bell Sys. Tech. Jour.*, Vol. 32, p. 1, January, 1953.

[3] H. Statz, G. A. deMars, L. Davis and A. Adams, "Surface States on Silicon and Germanium Surfaces," *Phys. Rev.*, Vol. 101, p. 1272, February, 1956.

[4] R. H. Kingston, "Water-Vapor-Induced n-Type Surface Conductivity on p-Type Germanium," *Phys. Rev.*, Vol. 98, p. 1766, June, 1955.

[5] H. C. Montgomery and W. L. Brown, "Field-Induced Conductivity Changes in Germanium," *Phys. Rev.*, Vol. 103, p. 865, August, 1956.

manium surface, driving the surface conductivity towards n-type. This paper shows that at least part of this process has a long time constant (of the order of several hours). Because of the time required for the process to reach equilibrium, it has sometimes been called the "48-hour" effect. The effect is present in varying degrees in most commercial transistors. It shows up as a slow change in surface potential of the germanium surface, and therefore also in the current-transfer ratio and the saturation current, when the transistor is brought from one temperature to another.

Evidence is presented to support the view that the chemical change of the surface responsible for the change in surface potential is a chemical condensation (hydration–dehydration) of the hydrous germanium oxide layer on the surface. In this process the oxide layer, when subjected to a reduction in relative humidity, gives off water slowly. Thus, new bonds of the type Ge-O-Ge are formed leading to a reduction in the solubility of the oxide.

## The 48-Hour Effect

It has been found that when germanium p-n-p transistors are life tested at elevated temperatures (e.g., 85°C), and at intervals are brought back to room temperature for routine measurements, the current-transfer ratio, $\alpha_{fc}$ (formerly called $\alpha_{rb}$), shows a slow decline. A final value is apparently reached after some 48 hours. Simultaneous with the decrease in $\alpha_{fc}$ a corresponding increase is found in the saturation current for the emitter and collector junctions suggesting that the primary factor responsible for the change is the surface potential. Figure 1 shows results from measurements on a group of 18 germanium p-n-p transistors, hermetically encapsulated in dry air (Dew point −10° to −30°F) with a silicon resin covering the transistor surface. The transistors were life tested for 1,500 hours of shelf life at 85°C. At this point they were removed from the oven, cooled to room temperature and measured during storage at room temperature. The values of $\alpha_{fc}$ have been normalized to the initial reading taken 20 minutes after removal (which was chosen to allow temperature equilibrium to be reached). The decrease in $\alpha_{fc}$ after 48 hours amounts to nearly 20 per cent, although the value differs with the method of encapsulation as described later.

Since $\alpha_{fe}$ depends on several different factors, an independent test was made connecting the change to the surface recombination velocity, $s$, and thereby the surface potential. In this test a number of transistors were placed in an oven at 105°C for 25 hours, and then held at room temperature for an equal period. Measurements of $\alpha_{fe}$ at

room temperature were made at intervals. In the high temperature run the units were removed from the oven, cooled to room temperature, measured, and again placed in the oven. The results for a typical unit are shown in Figure 2. These results were analyzed according to the theory of Webster,[6] the details of which are given in Appendix I. The conclusion of this analysis is that the gradual change of the curves after subjecting the transistor to a new temperature can be accounted for by a change in $s$. Furthermore, relative values of $s$ can be obtained with good approximation from the inverse of the measured values of $\alpha_{fe}$ at 1 milliampere emitter current.

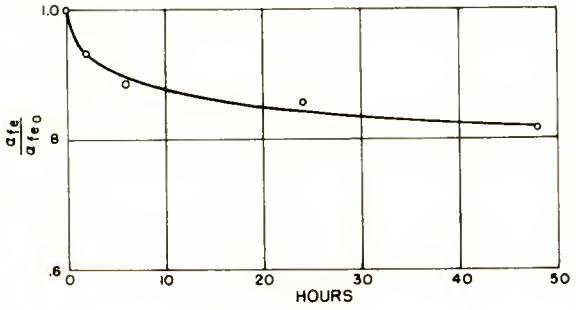



Fig. 1—The 48-hour effect. $\alpha_{fe}/(\alpha_{fe})_0$ versus time, where $(\alpha_{fe})_0$ is the value measured 20 minutes after removal of the units from the oven. Each point an average of 18 units.

That the 48-hour effect is entirely reversible is shown by the experimental results shown in Figure 2. Prior to the experiment the transistors had been aged at 105°C for a sufficient time so that the change in $\alpha_{fe}$ caused by oxidation[7] would be small and not interfere with the measurements. If this is not done the resulting $\alpha_{fe}$ will be a superposition of the two effects, namely the reversible change of the 48-hour effect and the nonreversible decrease caused by oxidation.

If the process responsible for the 48-hour effect is governed by an activation energy it would be expected that $s$ should follow a logarithmic dependence of time. In Figure 3 the results of Figure 1 have been replotted in a semilogarithmic scale. Upper and lower quartile are shown. The straight line in Figure 3 represents the equation

[6] W. M. Webster, "On the Variation of Junction-Transistor Current-Amplification Factor with Emitter Current," *Proc. I.R.E.*, Vol. 42, p. 914, June, 1954.

[7] J. T. Wallmark, "Influence of Surface Oxidation on $\alpha_{ei}$ of Germanium p-n-p Transistors," *RCA Review*, Vol. XVIII, p. 255, June, 1957.

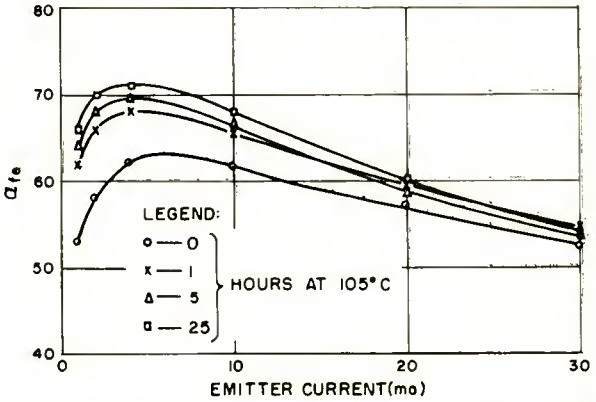



Fig. 2a—$\alpha_{fe}$ versus $I_e$ for a unit stored at 105°C and measured at 25°C.

$$1/\alpha_{fe} = k_1 \log t + k_2, \tag{1}$$

which is the equation of a process governed by an activation energy. From the limited data of Figure 2 it appears that the activation energy is small, probably less than 3,000 calories per mole.

## Influence of Oxide Thickness

Conceivably, the slow change found in the 48-hour effect might be connected to a transport time through the oxide layer on the germanium surface. Recently it has been shown that germanium oxidizes with a measurable rate even at room temperature.[7,8] Therefore, it

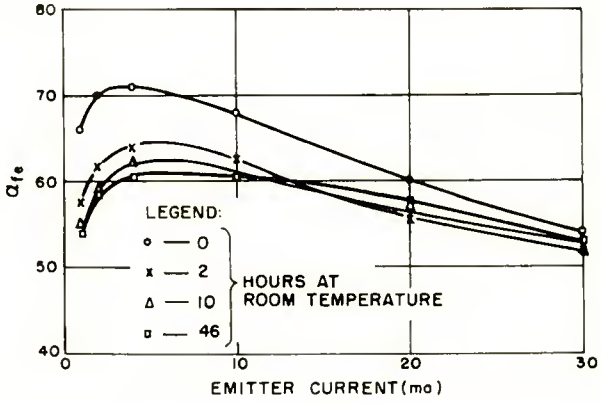



Fig. 2b—$\alpha_{fe}$ versus $I_e$ for the same unit stored and measured at 25°C after the 105°C period.

[8] R. J. Archer, "Optical Measurement of Film Growth on Silicon and Germanium Surfaces in Room Air," *Jour. Electrochem. Soc.*, Vol. 104, p. 619, October, 1957.

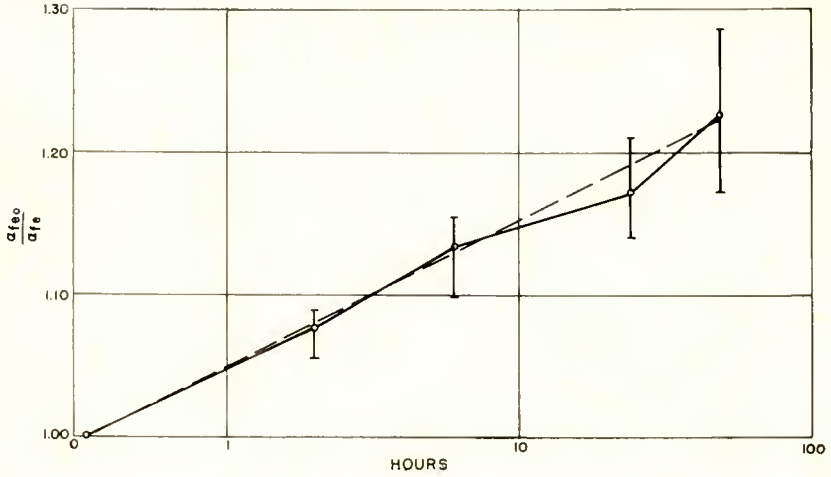



Fig. 3—Replot of the results in Figure 1 on a semilogarithmic scale. Each point an average of 18 units. Upper and lower quartile shown.

should be possible to measure the 48-hour effect on the same transistors early in life when the oxide thickness is small, and again late in life when the oxide layer has grown appreciably. Any difference in the characteristic decay time or magnitude of the effect could then be correlated with the oxide thickness. Figure 4 shows the results of such measurements in which the values of $\alpha_{fe}$ 48 hours after removal from an 85°C oven have been normalized to the values observed 20

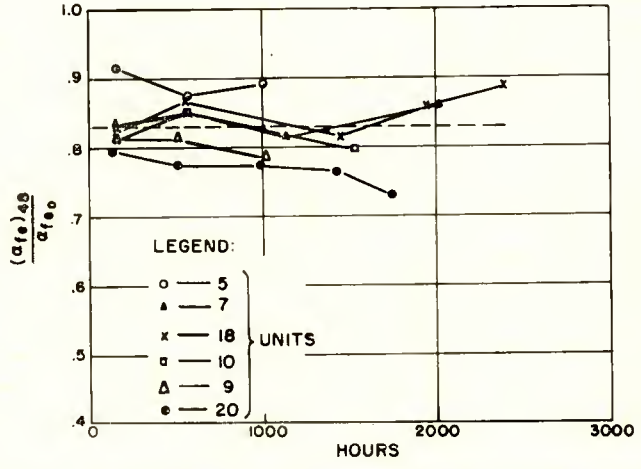



Fig. 4—The 48-hour effect at different stages of life. Each point an average of several units.

minutes after the removal. The measurements were made during several thousand hours of 85°C shelf life. During this time it is estimated that the oxide layer grew by approximately a factor of 4. However, no significant correlation between the amplitude of the 48-hour effect and the oxide thickness is apparent.

Figure 5 shows measurements similar to those of Figure 3 which were made during the life of a group of 7 transistors. No significant correlation can be found for the characteristic decay time derived from these results and the oxide layer thickness.

It is concluded, then, that the characteristic decay time of the 48-hour effect is not dependent on a transport process through the oxide layer.

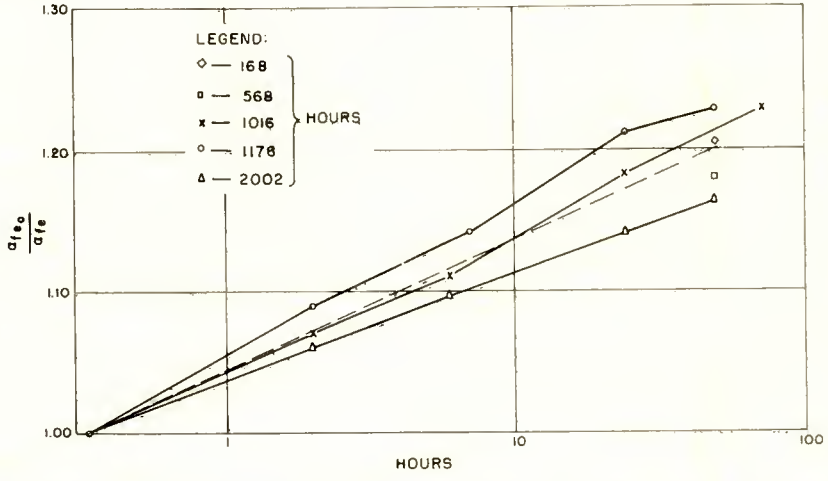



Fig. 5—One of the runs of Figure 4 plotted on a semilogarithmic scale.

Furthermore the independence of oxide thickness is hard to associate with a surface (of the oxide layer) phenomenon which definitely should decrease in amplitude as the oxide grows.

## Influence of the Dryness of the Encapsulant

Experiments with different encapsulants revealed that the water retention capabilities and the dryness of the encapsulant had an influence on the amplitude of the 48-hour effect. Figure 6 shows results for units encapsulated in a vacuum of $10^{-4}$ mm Hg after baking in a vacuum of 0.5 mm Hg at 90°C for 6 hours. In this case the 48-hour effect amounts to only 3 per cent.

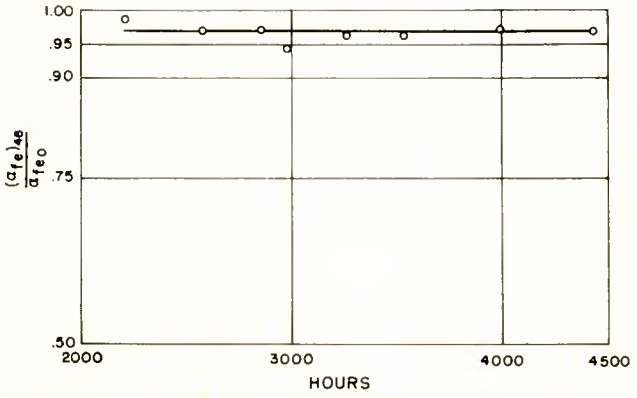



Fig. 6—The 48-hour effect for units encapsulated in a vacuum of $10^{-4}$ mm Hg. Each point an average of 5 units.

Figure 7 shows results for units encapsulated in a metal can containing molecular sieve (sodium and calcium alumino-silicates), prebaked to 220°C for 16 hours. Again the 48-hour effect amounts to only 3 per cent.

It is believed that the first of these tests corresponds to conditions of extreme dryness mentioned previously, while the second test exemplifies an encapsulant with high water retention capabilities.

These experiments suggest that minute amounts of water in the can (50 parts of water per million parts of encapsulant) are responsible for the effect. The results suggest that the germanium oxide layer has a water retention that is less than that of molecular sieve but larger than, or comparable to, that of the silicon resin encapsulant used in the tests shown in Figures 1-5.

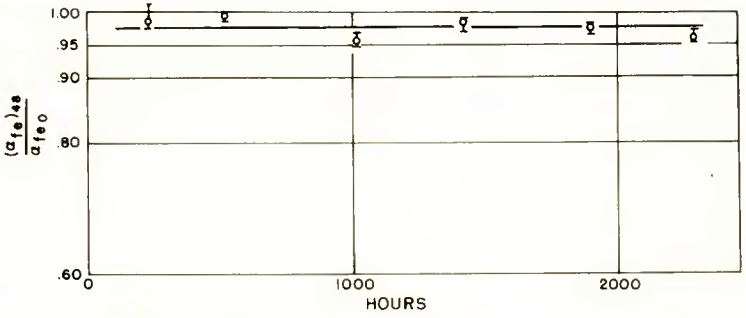



Fig. 7—The 48-hour effect for units encapsulated with dry molecular sieve. Each point an average of 19 units.

### Solubility Measurements

An important clue to the nature of the 48-hour effect is the fact that the solubility of the oxide layer changes with heat treatment. More specifically the oxide layer when heated in a dry atmosphere becomes gradually insoluble in dilute NaOH at a rate that matches the rate of change of $\alpha_{fe}$. Measurements of this effect were made in the following way. Samples of 8 ohm-centimeter n-type germanium were heated to 105°C for various lengths of time in air with a relative humidity of approximately 2 per cent. Thereafter the samples were washed in dilute NaOH to remove the oxide layer. The solution was evaporated to dryness and the residue was analyzed quantitatively in a spectrograph for germanium content. The details of the procedure will be published separately. The results are shown in Figure 8. Results for 100 per cent relative humidity are also shown in Figure 8 as a dashed curve. These results[7] were obtained at 85°C and extrapolated to 105°C.

The interpretation of these results suggests the presence of two competing processes. One is a straight oxidation process as found from the high relative humidity experiments. The other is a gradual decrease in the solubility of the oxide layer as found in the low relative humidity experiment. The solubility decrease, therefore, must be related to the loss of water from the oxide layer at low relative humidities. To develop a quantitative expression for these processes, let us assume that the oxidation rate is not too different at high and low relative humidities. This assumption is supported by the nearly identical initial rise of the two curves. Assume further that a simple energy activated process following the Elovich equation for the loss of water. Then the following differential equation governs the increase in soluble oxide:

$$\frac{dx}{dt} = \frac{Lt^{-\frac{2}{3}}}{3} - \frac{M}{t + t_0}, \tag{2}$$

where $x$ is the soluble oxide weight, and $L$ and $M$ are constants.

Fitting Equation (2) to the experimental results in Figure 8 we obtain quantitative expressions for the two processes taken separately. The expression for the growth of total oxide weight matching the dashed curve in Figure 8 is

$$x_{ox} = 4.9t^{\frac{1}{3}}. \tag{3}$$

Similarly, the growth of insoluble oxide weight derived from matching the solid curve of Figure 8 with Equation (2) is

$$x_{\text{insol}} = 16 \log (t + 4.9) + N. \quad (4)$$

The conclusion derived from this experiment is that the drying-out process of the oxide layer as expressed in Equation (4) has a time dependence that agrees well with the 48-hour effect.

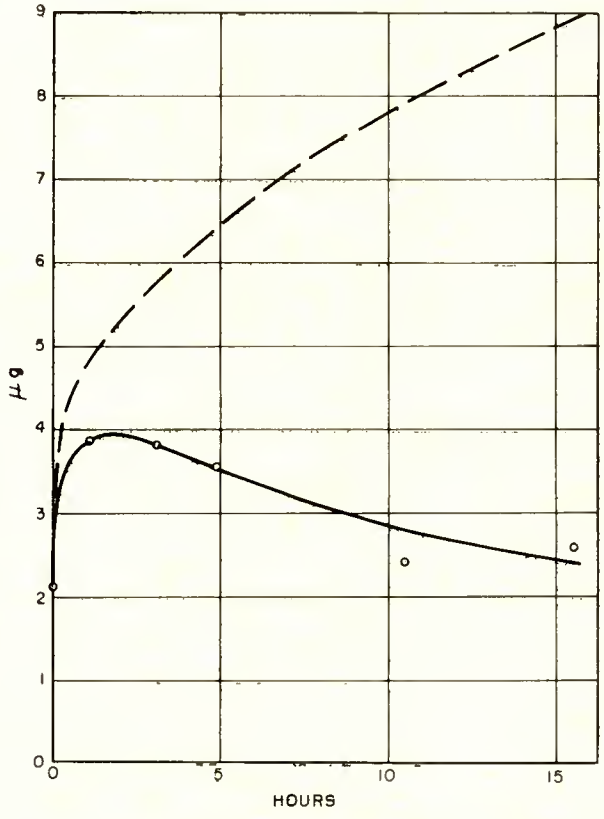



Fig. 8—Amount of oxide soluble in 5 minutes in 0.25 normal NaOH after oxidation in room air at 105° C.

## Phenomenological Model of the Surface

In order to account for the observed behavior of the surface, the following model is presented. It is known that the germanium surface in a transistor is covered by a layer of oxide of a thickness of approximately 10-20 monolayers.[7,8] This oxide is soluble in $H_2O$, NaOH, and HF, and therefore has been assumed to consist of germanium dioxide.

The oxide layer is probably not crystalline as evidenced by X-ray investigations.[9] Initially the oxide layer is formed in an electrolyte such as KOH, and subsequently washed in $H_2O$. Therefore it has been suggested that the oxide contains OH-groups,[10] and possibly other groups, depending upon the chemical history of the unit. In the case of silicon, such a hydrous, noncrystalline silicon oxide (silica gel) is well known. As shown below, the picture of a hydrous germanium oxide in the form of a glassy or polymorphous substance, in analogy with silica gel containing a variable amount of water, fits the experimental data very well.

It is well known that silica gel, which may be thought of as a mixture of different silicic acids, gives off water through chemical

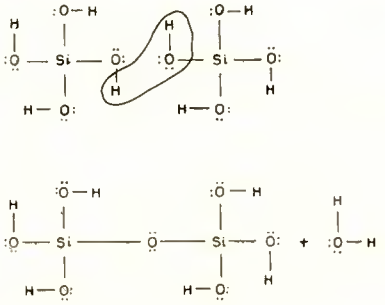


Fig. 9—The process of condensation. Two molecules disilicic acid form one molecule of metasilicic acid giving off one molecule of water.

condensation. In this process simple silicic acids form irregular Si–O–Si bonds, transforming to higher silicic acids as shown in Figure 9 for the transition of disilicic acid to metasilicic acid. The metasilicic acid in turn may condense further by combining with other acid molecules. This process is reversible with relative humidity, as is well known from the practical utilization of silica gel as a drying agent. What has been said here about silicon is true also about germanium (and other elements as well). Such a process would introduce irregular cross-links creating a glass-like structure without crystalline regularity. The solubility of such a structure would decrease markedly with amount of cross-linkage.

The hydration of the oxide layer introduces positive trapped charge, most likely one positive charge per broken cross-link. It is interesting

[9] R. D. Heidenreich, as reported in Reference (2).

[10] D. R. Turner, "The Anode Behavior of Germanium in Aqueous Solutions," *Jour. Electrochem. Soc.*, Vol. 103, p. 252, April, 1956.

in this connection that OH— groups have been suggested as donors in oxide cathode coatings.[11] The positive charge constitutes a slow trap which according to the model is located in the oxide layer, and not only on the surface as has sometimes been assumed. However, the number of slow traps originated by the hydration is small, one or two orders of magnitude smaller than the number of slow traps originated by oxidation. When a germanium transistor is heated to an elevated temperature, water is given off from the interior of the can, from the encapsulants, and from the germanium oxide layer. If the water retention of the germanium oxide is larger than that of its surrounding, the equilibrium will be shifted to a stable state in which more water is found in the germanium oxide layer. This process will proceed at a rate that is governed by the activation energy of the absorption of water in the oxide, assuming that diffusion through the oxide is fast and therefore not rate determining. When the transistor is cooled down again the water equilibrium will shift back to its original state at a slightly slower rate, as the temperature is lower.

## Discussion of Results

A considerable amount of data exists on the surface traps, their number, energy levels, etc. However, before the chemical–physical nature of the traps is known, it is difficult to design surface characteristics for specific device requirements. This paper points out the connection between the chemical nature of an important species of slow traps and the resulting surface potential.

Some practical consequences of the effect described are evident in the measurements of transistor characteristics. When a transistor has been removed from operation at a temperature above the ambient, the characteristics of the transistor will drift for some 48 hours until the change becomes imperceptible. This is especially true in life testing transistors, an operation which is often carried out at close to the rated maximum temperature. If the time between removal of the transistor from the high temperature and the subsequent measurement is not controlled, the measured values of $\alpha_{fe}$ and saturation current will show considerable spread, making rapid and reliable interpretation of trends difficult.

Another easily overlooked consequence is that the initial measurements on the transistor before the life test and the subsequent measurements during life are not comparable unless allowance is made for

[11] R. H. Plumlee, "Electron Donor Centers in the Oxide Cathode," *RCA Review*, Vol. XVII, p. 231, June, 1956.

the 48-hour effect. If this effect is not taken into account, the subsequent measurements of $\alpha_{fe}$ (and saturation current) will represent non-equilibrium values which are too high (or low). The usual drop in $\alpha_{fe}$ after the initial measurement, caused by progressive oxidation, is thereby masked, giving a false impression of stability.

There are two possible ways of reducing the 48-hour effect. One is the meticulous removal of moisture from the encapsulation and the transistor, heating the transistor for several hours at a temperature higher than the rated temperature. Another method is to employ a gettering agent with a water retention which is higher than that of the germanium oxide layer even at elevated temperatures. In a practical process, a combination of the two methods may offer advantages.

As an interesting by-product of the results, the magnitude of the 48-hour effect may offer an extremely sensitive and reliable method of testing the hermetic enclosure of transistors for leaks. In this application the 48-hour effect is considerably more sensitive than the conventional method of waiting for the reverse current–voltage characteristic to deteriorate.

## Acknowledgment

It is a pleasure to acknowledge helpful discussions with E. L. Jordan, R. H. Plumlee, H. M. Pollack, L. R. Shardlow, and W. M. Webster, assistance with the spectrographic method by M. C. Gardels and H. M. Whitaker, and assistance with some of the measurements by R. R. Vannozzi.

## Appendix

As shown by Webster[6]

$$1/\alpha_{fe} = Psg(z) + Q(1+z), \tag{5}$$

$$z = KI_e, \tag{6}$$

where $Q$ and $K$ are factors that depend on transistor dimensions and $g(z)$ is a field factor. $P$ is a factor containing the effective area $A_s$ in which surface recombination takes place. This area is given by

$$A_s = \frac{\pi}{4}(d_c^2 - d_e^2) + 2\frac{\pi}{4}[(d_c + L)^2 - d_c^2], \tag{7}$$

where

$$L = (D_p \tau_{eff})^{1/2}. \tag{8}$$

For $d_e = 0.015$ inch, $d_c = 0.045$ inch, and $\tau_{eff} < 15$ microseconds, Equation (7) becomes

$$A_s \approx \frac{\pi}{4} d_c^2 (1 + 4 L/d_c). \tag{9}$$

For surface recombination dominating over bulk recombination,

$$\tau_{eff} \approx M/s. \tag{10}$$

where $M$ is a proportionality constant. Therefore,

$$A_s \approx \frac{\pi}{4} d_c^2 \left[ 1 + \frac{4}{d_c} \left( \frac{D_p M}{s} \right) \right]^{1/2}. \tag{11}$$

Equation (11) represents a very slowly varying function of $s$. Furthermore, for relatively small changes in $s$, the variation of $A_s$ can be accounted for by a somewhat reduced value of $M$, which can be determined from experiments. It will be assumed therefore that all terms in Equations (5) and (6) are independent of $s$, except $s$ itself.

Fitting Equations (5) and (6) to the initial curve in Figure 2, one obtains

$$Ps = 4.0,$$
$$Q = 0.032,$$
$$K = 1.0 \text{ ma}^{-1}.$$

From these values it is clear that the influence on $\alpha_{fe}$ of the second term in Equation (5) measured at 1 milliampere emitter current is less than 2 per cent. As the total change in $\alpha_{fe}$ at the same current is approximately 20 per cent, the second term can be neglected. Therefore at 1 milliampere

$$1/\alpha_{fe} \propto s.$$

# SIMPLIFIED TREATMENT OF ELECTRIC CHARGE RELATIONS AT A SEMICONDUCTOR SURFACE

By

E. O. Johnson

RCA Laboratories,
Princeton, N. J.

*Summary—Simple, graphical representation is used to describe the electric charge and potential relations at a semiconductor surface. The balance between trapped and mobile charge at the surface is considered for both equilibrium and nonequilibrium situations. The graphical treatment illuminates some of the contemporary surface measuring techniques and shows what happens at metal–semiconductor and gas–semiconductor contacts. P-N junctions, which in a real sense are internal surfaces, are also depicted graphically for both equilibrium and nonequilibrium cases. This treatment serves to emphasize the close similarity between the surface and junction phase-boundary systems, establishing a useful and perhaps novel viewpoint for some aspects of junction behavior.*

## Introduction

THIS paper describes many of the salient electrical features of a semiconductor surface in simple physical terms that are based on the presently accepted model of the surface.[1] It is hoped that the paper will help give a "feel" for surface physics to those who do not have the time or inclination to read through the already voluminous literature on the subject. For those interested in probing deeper, some of the most important references are given. Mathematics is reduced to a minimum. An effort has been made, however, to be rigorous and to point out exceptions, qualifications, and uncertainties. It is assumed that the reader has at least an elementary knowledge of semiconductor physics.

The tool employed in this exposition is a graphical representation that is both simple and rigorous. It describes charge in "slow" and "fast" surface states and also the space charge, which is composed of fixed donors, acceptors, and mobile holes and electrons. It enables one to see at a glance how charge balance ensues and how the various parameters are related for both equilibrium and nonequilibrium situations. This method of presentation illuminates some of the contem-

[1] R. H. Kingston, "Review of Germanium Surface Phenomena," *Jour. Appl. Phys.*, Vol. 27, p. 101, February, 1956.

porary surface measuring techniques and also shows what happens at metal–semiconductor and gas–semiconductor contacts.

Both p-n junctions and surfaces are phase boundaries and so might be expected to have many features in common. The graphical method is shown to be a particularly useful technique for exploring this comparison. A somewhat novel approach to some aspects of junction behavior results.

## Surface States

It is common practice to describe the electrical properties of a bulk semiconductor in terms of an energy-band model which originates from the periodicity of the crystal lattice. Consider an ordinary crystal of germanium, strongly n-type, with all of the donors ionized. The specimen will be electrically neutral with as many electrons in the conduction band as there are positively charged donors. Assume that there are no states in the forbidden gap that can capture and immobilize electrons in the bulk. The situation will then be as shown in Figure 1. Now suppose that a surface is generated along the section x-x by cleaving the crystal. If no net charge is transferred between the halves of the crystal in the act of cleaving, these halves will still be electrically neutral.

In the act of cleaving the crystal something new has been added. First of all, the periodicity of the crystal has been destroyed at section x-x. Secondly, the crystal structure will probably be damaged and strained near the cleavage plane. Thirdly, the surface at this plane will be exposed to the influence of the surrounding atmosphere. All of these disturbances are thought to be capable of introducing localized allowed energy levels, or states, in the normally forbidden energy gap. Tamm[2] and Shockley[3] have shown theoretically that the rupture of periodicity should create localized allowed energy states, some of which may be in the forbidden gap. There should be about one state per surface atom. It is also reasonable to expect states in the forbidden gap if the surface of the germanium is covered with a crystalline germanium oxide. This follows because the transition between the bulk and the oxide is also a disturbance on the periodicity although, perhaps, not as severe as in the bulk–vacuum transition. Experiments with various ambients give good justification for supposing that adsorbed or absorbed molecules or atoms of oxygen, ozone, nitrogen, ammonia, water vapor, etc. generate allowed energy states in the

[2] I. Tamm, "Über Eine Mögliche Art Der Elektronem Bindung An Kristall Oberflächen," *Physik Z. Sowjetunion*, Vol. 1, p. 733, 1932.

[3] W. Shockley, "On the Surface States Associated with a Periodic Potential," *Phys. Rev.*, Vol. 56, p. 317, August, 1939.

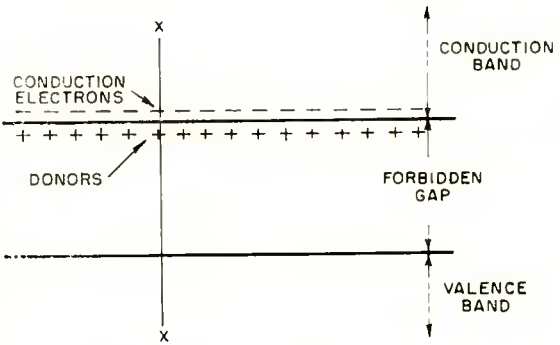



Fig. 1—Energy bands inside semiconductor.

forbidden gap at the surface. Although the details of charge-trapping mechanisms are still obscure, a variety of experiments indicate that trapping does occur.

Figure 2 shows the specimen of Figure 1 as it would look it five electrons from the conduction band had fallen into surface states. Five unneutralized donors are left behind. The crystal still maintains electrical neutrality since no charge has been generated or lost. However, in accordance with perfectly straightforward electrostatic principles,[4,5] there will be a rise in potential across the depletion region containing the locally unneutralized positive donor charges. This

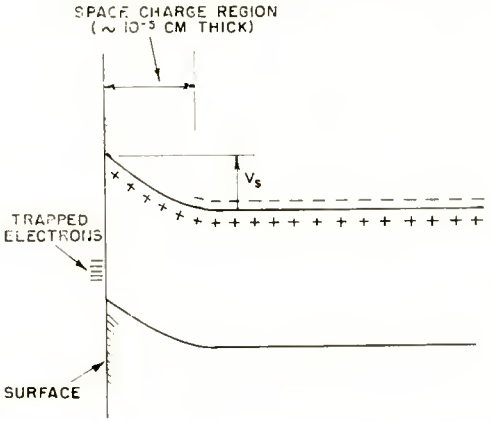



Fig. 2—Effect of trapped electrons on energy bands near surface.

[4] C. G. B. Garrett and W. H. Brattain, "Physical Theory of Semiconductor Surfaces," *Phys. Rev.*, Vol. 99, p. 376, July, 1956.

[5] R. H. Kingston and S. F. Neustadter, "Calculation of the Space Charge, Electric Field, and Free Carrier Concentration at the Surface of a Semiconductor," *Jour. Appl. Phys.*, Vol. 26, p. 718, June, 1955.

potential, denoted in the figure by $V_s$, is often called the surface potential. The region of unneutralized charge across which it occurs is known as the surface space-charge region. In its most general form this region contains, besides the donors shown in the figure, acceptors, mobile holes, and electrons.

The presence of the surface states was first invoked to explain the behavior of a metal–semiconductor contact.[6,7] Their existence was later inferred from the unexpectedly low response of the field-effect transistor.[8] Direct optical evidence for their presence has also been reported.[9] This takes the form of the visible light that is emitted when a semiconductor surface is sandblasted. At least some of this light may be caused by the energy released by electrons falling from the conduction band down into the traps.

Many experiments carried out in the last few years strongly indicate that there are two different classes of surface states. The first class has states of a relatively large density ($\sim 10^{12}$ or more states per square centimeter in germanium). These states are called "slow" because they exchange charge very slowly with the bulk material. The time constants of charge exchange are measured in seconds or tenths of a second. These states are strongly affected by the ambient atmosphere and supposedly exist on or within the oxide[10] layer which covers the semiconductor surface. Atmospheres of $O_2$, $O_3$, and other substances having a strong electron affinity cause the slow states to acquire a negative charge, at least with germanium. Surfaces which are reputed to be atomically clean[11] show a very strong tendency to trap electrons, presumably by the unsaturated bonds of the surface atoms. Wet air, on the other hand, causes the slow states to acquire a positive charge and also increases the speed of interaction with

---

[6] J. Bardeen, "Surface States and Rectification at a Metal Semi-Conductor Contact," *Phys. Rev.*, Vol. 71, p. 717, May, 1947.

[7] W. E. Meyerhof, "Contact Potential Difference in Silicon Crystal Rectifiers," *Phys. Rev.*, Vol. 71, p. 727, May, 1947.

[8] W. Shockley, *Electrons and Holes in Semiconductors*, D. Van Nostrand Co., New York, 1950, pp. 29-33.

[9] P. Aigrain and D. Jenny, private communication.

[10] As is usually done, the term "oxide" is here taken to mean any sort of composite layer or structure that has grown onto the bulk material by virtue of interaction with the surrounding atmosphere. For a more complete discussion, see reference 1.

[11] P. Handler, "Electrical Properties of a Clean Germanium Surface," *Semiconductor Surface Physics*, University of Pennsylvania Press, Philadelphia, Pa., 1957, p. 23.

charge from the bulk material.[12,13] Little has been done to determine the density and energy distribution of the slow states. These states are of great importance in device technology since they largely determine the surface potential, and this seems to be an important parameter in all surface phenomena.

The second type of surface states is called "fast" because these states and the bulk material can exchange charge with time constants of the order of microseconds or less. The radically different time constants of the two types of states afford a very convenient means of separating their behavior. The fast states are supposedly located at or near the interface between the bulk and the oxide; they seem to be relatively unaffected by the atmospheric ambient. These states seem to be relatively less numerous (possibly up to about $10^{11}$ states per square centimeter of surface) than the slow states and probably play a smaller role in determining the surface potential. At least some of these fast states are the centers upon which surface recombination takes place. There is not yet complete agreement on the energy levels of fast states.

States at discrete energy levels have been found in germanium at various energies ranging between about $\pm 0.20$ electron volt from midgap.[13-18] Evidence for a continuous distribution of state energies has also been found.[19,20] Silicon has not been as intensively studied

---

[12] R. H. Kingston and A. L. McWhorter, "Relaxation Time of Surface States on Germanium," *Phys. Rev.*, Vol. 103, p. 534, August, 1956.

[13] G. C. Dousmanis, private communication.

[14] H. Statz, G. A. deMars, L. Davis and A. Adams, "Surface States on Silicon and Germanium Surfaces," *Phys. Rev.*, Vol. 101, p. 1272, February, 1956.

[15] H. Statz, G. A. deMars, L. Davis and A. Adams, "Surface States on Silicon and Germanium Surfaces," *Phys. Rev.*, Vol. 106, p. 455, May, 1957.

[16] J. Bardeen, R. E. Coovert, S. R. Morrison, J. R. Schrieffer and R. Sun, "Surface Conductance and the Field Effect on Germanium," *Phys. Rev.*, Vol. 104, p. 47, October, 1956.

[17] A. Many, E. Harnik and Y. Margoninski, "Surface Recombination Processes in Germanium and Their Investigation by Means of Transverse Electric Fields," *Semiconductor Surface Physics*, University of Pennsylvania Press, Philadelphia, Pa., 1957, p. 85.

[18] S. Wang and G. Wallis, "Field Effect on an Illuminated Ge Surface and Investigation of the Surface Recombination Process," *Phys. Rev.*, Vol. 105, p. 1459, March, 1957.

[19] H. C. Montgomery and W. L. Brown, "Field-Induced Conductivity Changes in Germanium," *Phys. Rev.*, Vol. 103, p. 865, August, 1956.

[20] C. G. B. Garrett and W. H. Brattain, "Distribution and Cross-Sections of Fast States on Germanium Surfaces," *Bell Sys. Tech. Jour.*, Vol. 35, p. 1041, September, 1956.

as germanium. There is, however, a better agreement on the "state" energy levels. So far, discrete levels approximately 0.45 electron volt from midgap are the only ones reported.[13,11,21] It is obvious that more study will be required to settle the question of the fast-state energy levels.

It should be pointed out that the distinction between fast and slow states has usually been a fairly clean-cut one. However, this need not necessarily be the case for all surfaces. Some, can have states which give a wide spread in time constants. For example, there is some experimental evidence to show that there is a small, but significant, fraction of slow states that have time constants as short as $10^{-3}$ second.[12,22]

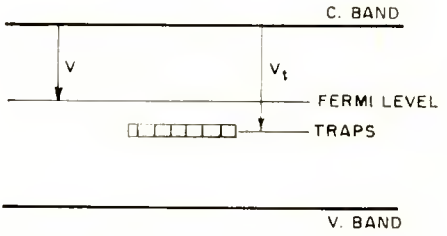



Fig. 3—Energy diagram for bulk traps.

## Surface-State Charge

Bulk traps or states are shown in the energy band diagram of Figure 3. The squares represent the location of these states. These are assumed to be located at a discrete energy level, $V_t$, below the edge of the conduction band. The Fermi level is represented by the line located at an energy $V$ from the edge of the conduction band. Under equilibrium conditions with no excess carriers present, the traps will be filled with electrons when the Fermi level is well above the traps, and essentially empty when it is well below. When it passes through the traps, the traps will be half filled with electrons. This behavior is described by the well known Fermi occupancy function[23] sketched in Figure 4.

---

[21] B. Bernstein and C. S. McCarthy, Jr., "Variation of Surface Recombination Velocity with Surface Potential on Silicon," *Bull. Am. Phys. Soc.*, Vol. 2, p. 130, March, 1957.

[22] G. C. Dousmanis and E. O. Johnson, "Time Constants of Slow States in Germanium Surfaces," *Bull. Am. Phys. Soc.*, Vol. 2, p. 170, April, 1957.

[23] W. Shockley and W. T. Read, Jr., "Statistics of the Recombinations of Holes and Electrons," *Phys. Rev.*, Vol. 87, p. 835, September, 1952.

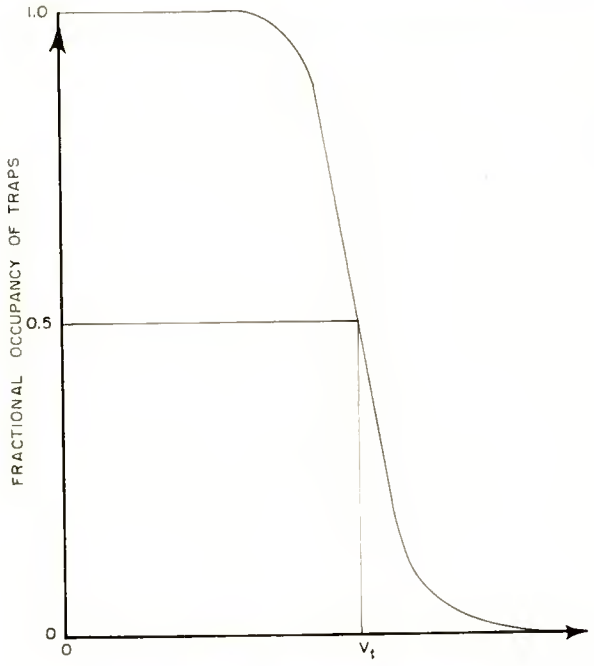



Fig. 4—Trap occupancy function.

The same situation must hold for traps on the surface, except that the surface potential $V_s$ becomes the important parameter as indicated in Figure 5. The positive sense for conventional potential is indicated by the arrow. As $V_s$ increases, the energy bands at or near the surface bend upwards and the Fermi level sweeps downwards along the surface. When the Fermi level passes through the traps, $V_s$ has the algebraic value $(\phi - \nu)$. When $V_s$ is smaller than $(\phi - \nu)$, the Fermi

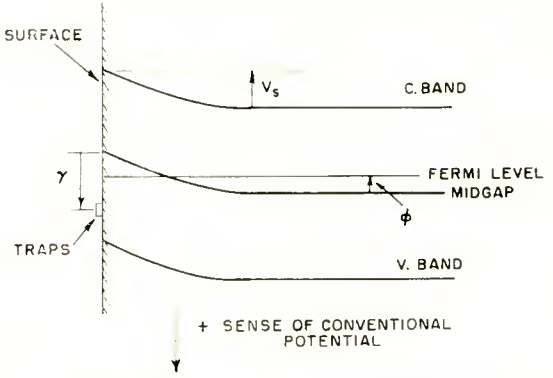



Fig. 5—Energy diagram for surface traps.

level intersects the surface above the trap level and the traps tend to be filled. On the other hand, the traps are essentially empty when $V_s$ is large. The fraction of traps occupied by electrons is represented by the Fermi occupancy function as shown in Figure 6 by the solid curve. Note that $V_s$ is used as the independent variable.

When excess carriers are present the trap occupancy depends upon the excess carrier densities and also upon the relative capture cross sections of the traps for holes and electrons. For example, if the electron capture cross section is relatively large, then the traps will tend to fill up with electrons when excess carriers are present. On the other hand, if the hole capture cross section is relatively large, the traps will tend to fill with holes, i.e., empty of electrons. As might be expected, the trap occupancy depends upon the relative densities

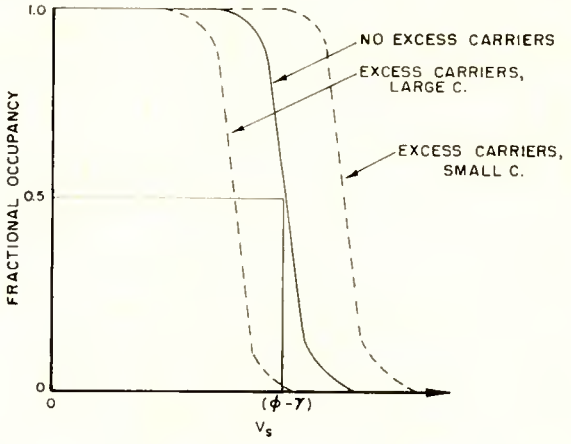



Fig. 6—Surface trap occupancy function—effect of excess carriers.

of the excess carriers as well as upon the hole to electron capture cross section ratio, $C$. The effect of excess carriers upon the occupancy function is typified by the dashed curves in Figure 6. One curve applies to a large value of $C$ and the other to a small value. For simplicity it is assumed that the excess hole and electron densities are equal. These curves are of the same form as the solid one but are shifted horizontally. The exact expression for the occupancy factor is derived from the Shockley–Read theory[23] and is described elsewhere.[17,20]

There are a number of possibilities to consider in connection with the electric charge on the traps. For example, it is conceivable that the empty traps can be neutral, or can be positively or negatively charged. Thus, as the traps fill with electrons, they will become nega-

tive, neutral, or doubly negatively charged, respectively. However, in subsequent arguments about trapped charge, no generality is lost by the implicit assumption that the empty traps are neutral.

## Space Charge

The sketch pertinent to this discussion is shown in Figure 7. Although an n-type specimen is indicated, the following discussion applies to any semiconductor regardless of conductivity type. In addition to the surface potential, $V_s$, previously discussed, another potential, $\phi_s$, is defined in the figure. This, too, is often spoken of as the surface potential.[1] It is frequently more convenient to use in mathematical analyses, and its magnitude and polarity give an immediate indication of the extent to which the surface layer of the bulk material is p or

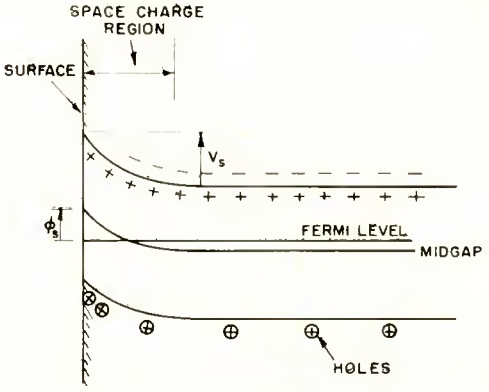



Fig. 7—Energy diagram for surface space-charge region.

n type. As shown in the figure, both $V_s$ and $\phi_s$ are considered as having negative values. A negative value of $\phi_s$ signifies a p-type surface because the Fermi level is below midgap and there are more holes than electrons at the surface. A positive $\phi_s$ signifies an n-type surface because the Fermi level is above midgap and there are more electrons than holes at the surface. The surface is intrinsic when $\phi_s$ is zero. For strongly negative values of $\phi_s$, the surface has a high hole conductivity, for strongly positive values of $\phi_s$ it has a high electron conductivity, and for values of $\phi_s$ near zero the conductivity is a minimum. This behavior can be used to obtain information about the value of $\phi_s$.[1] Thin samples are used and the value of $\phi_s$ is modulated by the ambient atmosphere or by an external electric field. Some assumption about the mobility of carriers near the surface is always necessary, however. There has been a theoretical study of surface

mobility[21] but as yet no definitive experiments. The uncertainty about this important parameter introduces some doubt as to the accuracy of many conductivity-type surface measurements.

In the subsequent discussion, as in the one preceding this section, the term "surface potential" applies only to $V_s$ unless otherwise noted.

The inversion of conductivity type at the surface, which occurs when the Fermi level crosses over the midgap potential, creates what is called an "inversion layer." The layer in the figure would be called a p-type inversion layer, or "channel." One could equally well picture an n-type inversion layer, or channel, on a p-type specimen. The opposite situation, where the bands bend in a direction to make the surface layer higher in conductivity than in bulk, but of the same conductivity type, creates what is known as an "accumulation layer."

Poisson's equation relates the surface potential, $V_s$, to the total charge, $Q_{sc}$, in the space-charge region.[4,5] This charge is composed of fixed donors and acceptors, and also mobile electrons and holes. The densities of free electrons and holes in the space-charge region are exponentially related to their values in the bulk by the familiar Boltzmann relations. Thus, for the case shown in Figure 7, the hole density at the surface will increase exponentially as the bands bend upwards with increasing $V_s$. The electron density, on the other hand, will decrease exponentially. It is relatively easy to find the explicit expression for the total charge in the space-charge region as a function of $V_s$ and the bulk semiconductor properties (conductivity and dielectric constant). Some samples of such solutions are shown in Figure 8. These apply to uniformly doped n-type germanium for the case where energy bands bend upwards making the surface tend towards inversion layer conditions. The charge relations are symmetrical and give the same type of curves with p-type specimens having surfaces tending towards n-type inversion. In these cases $V_s$ is positive, instead of negative as noted in the figure.

The general shape of the curves is not difficult to understand. When the doping is so low that intrinsic conditions prevail, the space charge is dominated by the exponential increase of the minority carriers at the surface. Thus a straight line appears on the semilog plot, except near $V_s = 0$. When the doping is heavier, the donor or acceptor charge causes a gently sloping plateau to appear on the curves. This plateau is described by the Sckottky depletion layer solution. The height of the plateau increases with the doping. The plateaus are washed out at high values of $V_s$ when the exponential increase in the minority

---

[21] J. Robert Schrieffer, "Effective Carrier Mobility in Surface Well," *Phys. Rev.*, Vol. 94, p. 1420, June, 1954.

carrier charge causes this charge to exceed the fixed donor or acceptor charge. The exponential increase in the minority carrier density is, of course, unaffected by the doping so that the upper portions of the curves all must have the same slope. The upper portions of the curves, however, must move out to higher values of $V_s$ as the doping increases. This causes the curves to cross and necessarily follows because the minority carrier density in the bulk decreases as the doping increases.

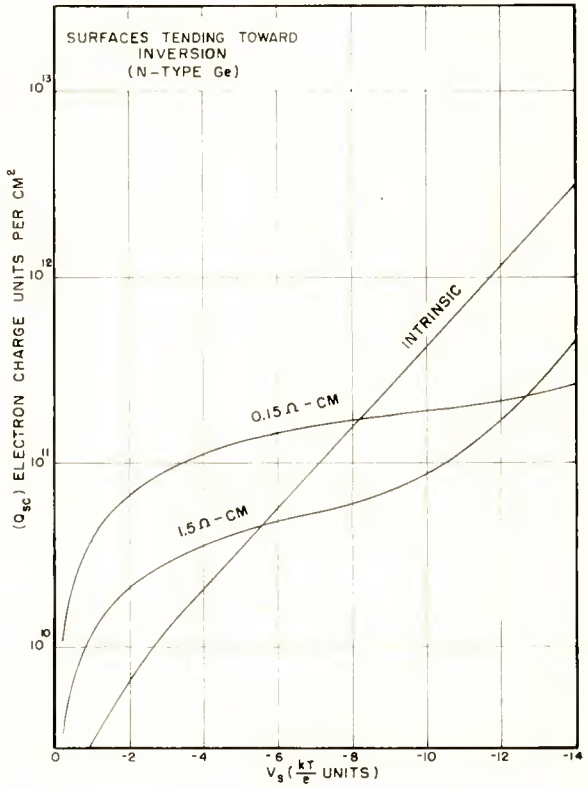



Fig. 8—Space charge for surfaces tending toward inversion.

Hence, larger values of $V_s$ are required to make the exponential increase in minority carrier charge overtake the fixed donor or acceptor charge. The knee of a plateau region always occurs close to one $kT/e$ unit of surface potential. This potential is sufficient to reduce the majority carrier charge, opposite in sign to the fixed donor or acceptor charge, to a small fraction of the fixed charge.

Curves for accumulation layer conditions are shown in Figure 9. The space charge increases very rapidly with surface potential because

of the exponential increase of the majority carrier charge. Thus, for the same magnitude of the surface potential, the space charge is larger than in the inversion layer case. Again, because of the symmetry of the space-charge relations, the same curves are obtained for accumulation layers on p-type material.

The electric field at the surface edge of the space charge layer is obtained from the space charge by Gauss' theorem. This field, in

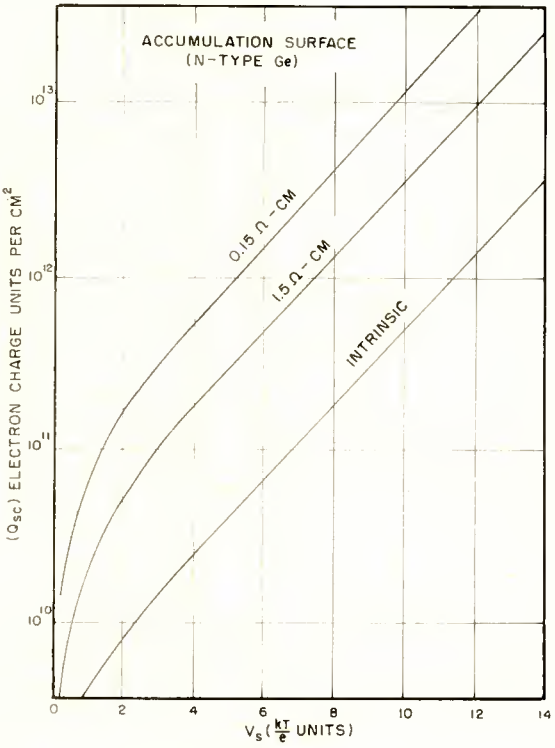



Fig. 9—Space charge for accumulation surface.

volts per centimeter, is roughly $10^{-7}$ times the electron charge units per square centimeter values given in Figures 8 and 9. Thus, the field is $10^4$ volts per square centimeter for intrinsic material at $|V_s| \cong 7kT/e$ units of surface potential.

For reasons of electrical neutrality the charge in the space-charge region must be balanced by the immobile charge in the surface traps plus whatever charge may reside on an electrode outside the surface. Electrically it makes no difference whether the balancing charge is

in the surface traps or somewhere in the region outside the surface. This fact is used to advantage in experiments where charge variation on an external electrode is employed to modulate the charge in the space-charge region, and hence the potential $V_s$.

The surface space-charge region finds its analogue in the sheath which surrounds the electrodes in a gas discharge plasma.[25] In this case, however, the space charge is composed of positive ions or electrons en route to the electrode. There is no fixed space charge, like that from donors and acceptors, and the particles are usually engaged in free-fall motion. Hence the familiar 3/2 power law applies as it does to electrons moving between the cathode and anode in a vacuum tube. In a metal, as with a semiconductor, there must also be a surface space-charge region. This, however, is very thin ($\sim 10^{-7}$ centimeter) and has a very low value of $V_s$ because of the extremely high charge concentrations in the bulk metal.

When excess carriers are present, the space-charge curve for surfaces tending toward inversion takes on the forms shown in Figure 10. These refer to 0.65 ohm-centimeter n-type germanium. The fractional increase in the minority carrier density is represented by the parameter, $\Delta_m$. The increase in carriers may result from electron–hole pair creation by light or from injection at a junction. The plateau due to the fixed donor charge tends to disappear as the minority carrier charge increases with injection. The increase in space charge at low surface potentials is small because here the donor charge outweighs the minority carrier charge.

The behavior with accumulation layer conditions is somewhat different because the space charge can only increase as fast as the majority carrier charge. In percentage this is usually much smaller than the minority carrier increase so that usually the accumulation layer space charge is relatively little affected by injection.

## Surface Capacitance

It is pertinent at this point to say a few words about the electrical capacitance of the surface. Electrical capacitance is associated with the differential change in charge on a system as a differential change is made on the electrical potential of the system. Accordingly, we can define capacitances for the charge in the slow and fast surface states and also for the charge in the space-charge region. These capacitances would be, respectively,

[25] J. D. Cobine, *Gaseous Conductors*, McGraw-Hill Book Co., New York, New York, 1941.

$$C_{ss} = \frac{dQ_{ss}}{dV_s}, \quad C_{fs} = \frac{dQ_{fs}}{dV_s}, \quad C_{sc} = \frac{dQ_{sc}}{dV_s}.$$

The capacitance $C_{ss}$ is simply the slope of the curve in a diagram of $Q_{ss}$ versus $V_s$, provided that we confine our attention to a time scale that is slow compared to the slow-state time constant. An estimate of tens of microfarads per square centimeter of surface has been given.[4] For shorter times $C_{ss}$ necessarily is smaller because $Q_{ss}$ would

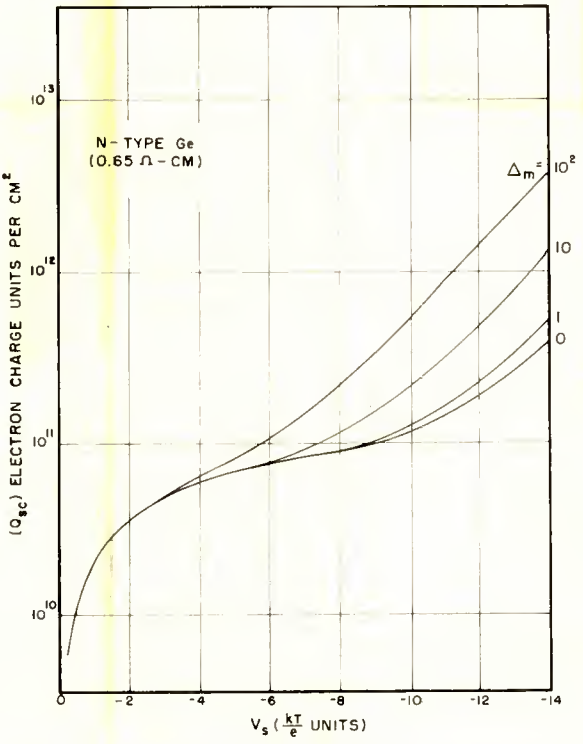



Fig. 10—Effect of excess carriers on space charge.

not have time to change with $V_s$. The same reasoning holds for $C_{fs}$ except that the time constant is much shorter. Likewise, the slope of the space charge curves (Figures 8, 9, 10) is identically $C_{sc}$ provided that the time scale is slow enough so that minority carrier kinetics do not complicate the picture. For such slow time scales the holes and electrons are always effectively in thermal equilibrium as $V_s$ changes (due, for example, to an externally applied field) and the system can be handled with the usual Fermi-level picture. For fast changes the

system becomes complicated and requires a more sophisticated treatment. The capacitance is no longer the slope of the static-space-charge curves.

In principle, the capacitance $C_{fs} + C_{sc}$ can be measured by making observations of the capacitance between the surface layers and an electrode mounted adjacent to them. The measurement would be carried out with frequencies in the approximate range 30 to 1,000 cycles. This is too high for the charge in the slow states to play a role and too slow for the system to become complicated by carrier kinetics and fast-state time constants. Brown[26] has described measurements of this sort and some studies have also been carried out at RCA Laboratories. The technique is difficult because the electrode-to-surface capacitance (~30 micromicrofarads) is much smaller than the space-charge capacitance (~30,000 micromicrofarads). Qualitative agreement with other surface measuring techniques has been obtained.

## Charge Balance and the Equilibrium Value of the Surface Potential

The foregoing information will now be put together graphically to show how electrical charge apportions itself on the surface and also how the equilibrium value of the surface potential is determined.

In Figure 11 the ordinate represents the logarithm of electrical charge, $Q$, and the abscissa the surface potential, $V_s$. A space-charge curve, $Q_{sc}$, is shown on this graph along with a curve for the charge, $Q_{ss}$, in slow states of high density existing at a discrete energy level. The dotted curve, which will be neglected for the moment, refers to the charge $Q_{fs}$ in fast surface states of relatively low density which also exist at a discrete energy level. The space-charge (positive) curve applies to a surface tending toward inversion on n-type, 0.65 ohm-centimeter germanium. To facilitate graphical manipulation, the negative charge in the slow and fast states is plotted in the same quadrant as the positive space charge. If over-all electrical neutrality exists and the charge in the fast states is neglected, it immediately follows that the system will be in equilibrium at the point where the slow state curve intersects the space-charge curve. The surface potential defined by this point of intersection is denoted in the figure by $V_{so}$.

For the case represented in the sketch it is apparent that the value of $V_{so}$ is a stronger function of the energy level of the slow states than it is of the slow-state density. Furthermore, it is apparent that the

[26] W. L. Brown, "Capacity Changes in the Surface Layer of a Semiconductor," *Bull. Am. Phys. Soc.*, Vol. 1, p. 48, January, 1956.

slow-state density could be very high, even as much as one state per surface atom, without resulting in an unreasonable value of $V_{so}$. If the slow-state density is low, so that the plateau of the slow-state curve intersects the space-charge curve, then it is obvious that the surface potential will be most sensitive to the slow-state density. If the slow states are located at a number of different discrete energy levels, then the composite charge curve can be represented by the sum of the curves for each discrete state, assuming that the statistics of each state are

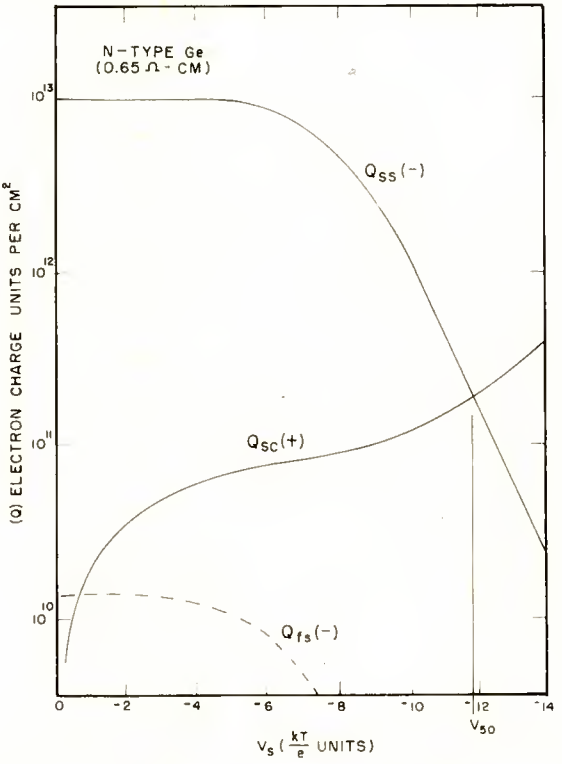



Fig. 11—Charge balance between space and trapped charge.

independent of each other. For a distribution of states, the composite curve for $Q_{ss}$ would tend to vary more slowly with $V_s$ than for the case of states at a discrete energy level.

Intuitively, one would expect the value of $V_{so}$ to be more stable if the curve for the trapped charge cuts across the space-charge curve at a steep angle, rather than at a shallow angle. The latter situation, for example, could occur in the situation sketched in Figure 12 where the space-charge curve intersects a plateau on the slow-state curve.

Here the latter curve represents the composite charge in a number of states that have different densities and energy levels. It is conceivable that some sort of noise or surface instability could arise from situations of this sort where the surface potential is relatively ill-defined. Since the slow states exchange charge with the bulk quite slowly (time constants of the order of seconds, or tenths of seconds), such noise or instability might be expected to have strong low-frequency components.

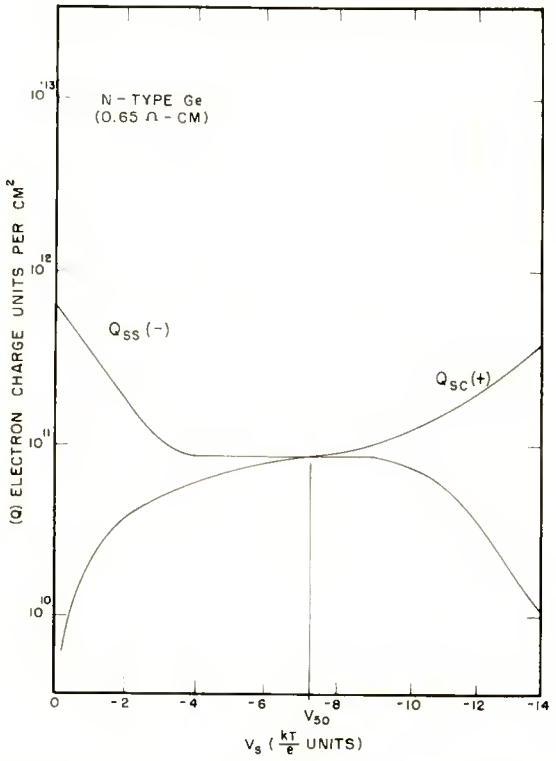



Fig. 12—Charge balance between space and multiple energy level trapped charge.

As pointed out earlier, little has been done to determine the densities and energy levels of the slow states. One would like to know, for example, what the effect of the ambient atmosphere is on the slow-state density and energy distributions. This is obviously an interesting problem to tackle since it should lead to some clarification of the structural details of the surface.

Now consider the role played by the fast surface states. Their charge will add algebraically to that of the slow states if independent

statistics are assumed. The composite curve for the charge in both types of surface states will then be the one that determines $V_{so}$ by the point of intersection with the space-charge curve. If the fast-state density is relatively low, the fast states can never play much of a role in fixing $V_{so}$. If their density is high, but their energy locates them as shown in Figure 11, the effect also will be small. However, if the fast-state density is high and they are strategically located in energy, these states could play a dominating role.

### Charge Balance under Nonequilibrium Conditions

The extension of the preceding graphical techniques to steady-state, nonequilibrium situations is simple and has a number of interesting ramifications.

If the fast states are neglected, the situation for the n-type, 0.65 ohm-centimeter specimen with an inversion layer will be as shown in Figure 13. A high density of slow states at a discrete energy level is again assumed. As before, the fractional increase in the minority carrier density is represented by a parameter $\Delta_m$. The equilibrium value of the surface potential is $V_{so}$; the equilibrium charge on the slow states is $Q_o$.

There are now two distinctly different situations to consider. The first is the case where the slow-state charge is not given time to change, and the second is the case where the slow-state charge is given time to come into equilibrium with the bulk. The first case can be treated using Figure 13. The trapped charge will remain constant as the factor $\Delta_m$ increases so that the system will have to remain on the horizontal line passing through $Q_o$. The surface potentials for each value of $\Delta_m$, denoted by $V_1$, $V_2$, and $V_3$ in the figure, are defined by the intersections of the horizontal line with the appropriate space-charge curve. The potential differences $(V_{so} - V_1)$ and $(V_{so} - V_2)$ would have been termed the surface photovoltages had the excess carriers been generated by light. An analytical description of the surface photovoltage is given elsewhere.[4,27]

A group at Bell Telephone Laboratories has studied the differential surface photovoltage occasioned by differentially small changes in the minority carrier density.[4,28] The case, such as sketched in the figure, where there are large changes in the minority carrier density has been

[27] E. O. Johnson, "Large-Signal Surface Photovoltage Studies with Germanium," *Phys. Rev.*, to be published.

[28] W. H. Brattain and C. G. B. Garrett, "Combined Measurements of Field Effect, Surface Photo-Voltage and Photoconductivity," *Bell Sys. Tech. Jour.*, Vol. 35, p. 1019, September, 1956.

studied at RCA Laboratories. In both studies there has been good experimental agreement with the analytical version of the behavior illustrated in the figure. The surface photovoltage is of practical interest because it enables the experimenter to deduce the value of $V_o$, an important parameter in many surface phenomena.

It is pertinent to discuss the effect of the fast states before going on to the case where the slow states are allowed to achieve equilibrium with the bulk. The fast states certainly have to be considered in

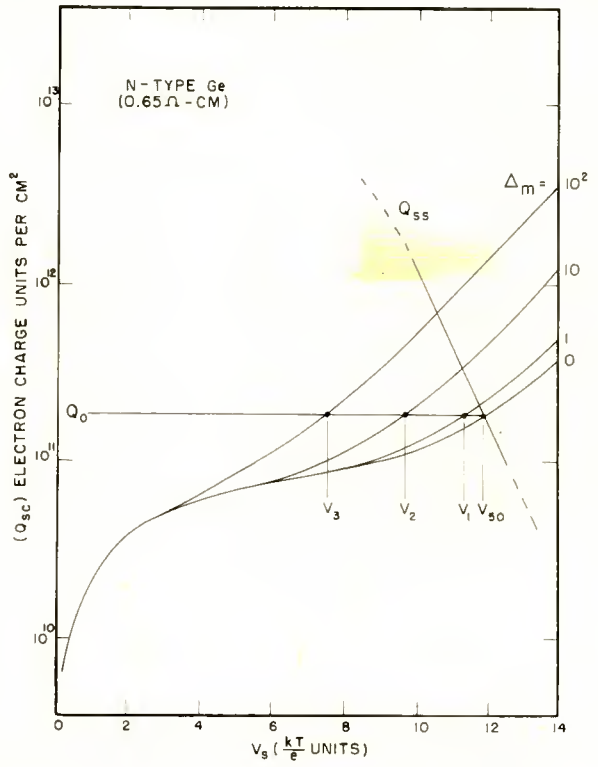



Fig. 13—Charge balance with excess carriers—constant trapped charge.

surface photovoltage measurements because their charge can change at a far faster rate than that ordinarily used to inject the minority carriers. In such experiments the carriers are injected at a rate (~60 cycles) that is fast compared to the slow-state time constants, but slow enough to preclude complications with minority carrier kinetics. The fast states can affect the surface photovoltages only if their charge changes appreciably as the system moves from $V_{so}$ towards zero surface potential. The situation shown in Figure 14 illustrates the case where

the fast states would have a pronounced effect on the surface photovoltage. The effect of injection (see Figure 6) on the fast-state charge curve represented by $Q_{fs}$ has not been introduced. Qualitatively, this will not affect the following discussion. As the injection factor, $\Delta_m$, is increased the system will not move along the horizontal lines, but, instead, will come into charge balance at the surface potentials denoted $V'_1$, $V'_2$, and $V'_3$. This is necessitated by the fact that the charge $Q_{fs}$ adds to the charge $Q_o$ as the injection is increased. It is obvious that

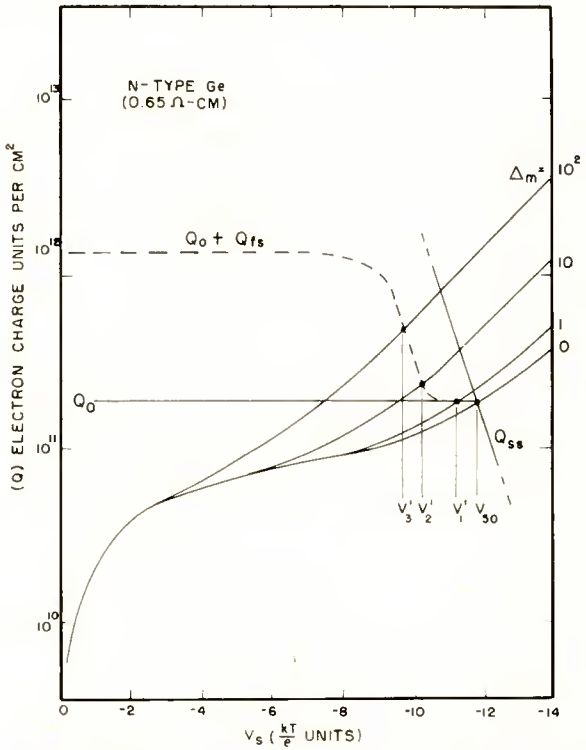



Fig. 14—Charge balance with excess carriers—variable fast-state charge.

the potentials $V'_1$, $V'_2$, and $V'_3$ will not, in general, be the same as $V_1$, $V_2$, and $V_3$ which occur if the fast states are negligible in effect. It is to be noted that the charge change due to the fast states could be either positive or negative depending upon the character of the fast states. Although not treated here, the former case could be treated in the same general manner.

The fast states will not have any effect on the surface photovoltages if they are low in density or if they are located at energies well removed from $V_{so}$. Surface photovoltages measured in these labora-

tories[27] agree well with the picture presented in Figure 13, but not at all with that of Figure 14. From this, conclusions can be made about the density and energy levels of the fast states.[27]

The case where the slow-state charge has time to come into equilibrium with the bulk is sketched in Figure 15. Note that both the space charge and the slow-state charge change with injection. Charge balance will again occur at the intersections of the appropriate curves. These intersections define new surface potentials $V''_1$ and $V''_2$ which

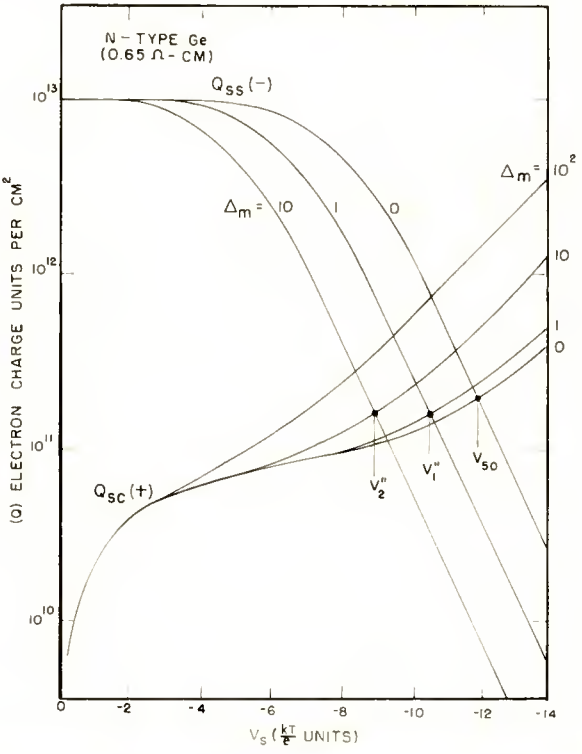



Fig. 15—Charge balance with excess carriers—variable slow-state charge.

are generally different from those indicated in Figure 13 for the case where the slow-state charge remains constant. The qualitative effect of the slow states on the surface photovoltage is seen to be similar to that caused by the fast states. The widely different time constants of these two different types of states can be employed to differentiate the type of state causing deviations in the photovoltage. It is obvious, at least in principle, that the differences between the equilibrium and nonequilibrium behavior of the slow states could be used to give

information on the density and energy distribution of these states.

This completes the qualitative discussion of the charge relations at a free semiconductor surface. In the two following sections the same graphical treatment will be extended to cover metal–semiconductor contacts, gas–semiconductor contact, and some features of a p-n junction. These subjects are very closely related to the preceding treatment because they deal with space-charge regions and in some important cases with trapped charge also.

## Metal–Semiconductor Contacts

Metal–semiconductor contacts have long been of practical interest in the solid-state field. They have attracted theoretical interest because their behavior usually has not been in good accord with simple theory.[6,7] In particular, their rectifying capabilities have been hard to explain. The following graphical treatment is a generalization of the one described by Bardeen in an early paper[6] aimed at explaining the fact that the potential across the semiconductor space-charge region was not modified to the extent expected from the difference in work functions between a typical metal and a semiconductor. It will become clear how Bardeen was led to the conclusion that the results can be explained if there is a large density of surface states on the semiconductor. These states can play the major role in charge exchange with the metal so that there will be little change in the semiconductor surface space charge. Thus the potential change across the space-charge region should indeed be smaller than one would expect if no surface states were present. The system to be described is shown in Figure 16. In its undisturbed state the n-type semiconductor has a work function, $\phi$, an electron affinity, $\chi$, a surface potential, $V_s$ (by virtue of surface states), a half-gap potential, $V_g$, and a displacement, $\phi_b$, of the Fermi level above midgap potential. From the figure, $\phi = \chi + V_s + (V_g - \phi_b)$. The metal has a work function $\phi_m$ and is located at a distance $d$ from the semiconductor. It will be assumed that $\phi_m$ and $\chi$ are independent of $d$. It should be noted, however, that these quantities may change significantly if there is intimate contact ($d \sim 10^{-7}$ centimeter) between the metal and semiconductor.

The external electric charge that the semiconductor surface region will "see" as a result of the proximity of the metal surface is $Q_m = CV$, where $C$ is $\epsilon/4\pi d$, the capacitance between the two surfaces, $\epsilon$ is the effective dielectric constant in the space between the metal and the semiconductor, and $V$ is the potential difference, $\phi_m - \phi = \phi_m - \chi - V_s - (V_g - \phi_b)$. Thus,

$$Q_m = \frac{\epsilon}{4\pi d} [\phi_m - \chi - V_s - (V_g - \phi_b)].$$

It is implicitly assumed in this discussion that a conduction path is somewhere present to enable appropriate charge transfer to take place between the metal and the semiconductor.

Charge balance at the semiconductor surface can now be determined graphically as in the previous cases but with the charge $Q_m$ taken into account. This is shown in Figure 17. The quantities $Q_{sc}$ and $Q_{ss}$ are given as before. With the metal at a great distance, the equilibrium value of the surface potential must be $V_{so}$. The charge $Q_m$ is a straight line (on a linear scale) with a negative slope whose value depends upon the value of the variable parameter $d$. The intercept of this line with the horizontal axis occurs at the value of $V_s$ that makes $\phi_m = \phi$. If there are no slow states the system will come into equilibrium at

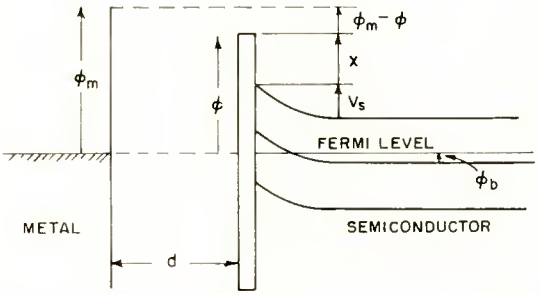



Fig. 16—Energy diagram for metal–semiconductor contact.

the point of intersection of $Q_m$ and $Q_{sc}$. This defines a surface potential $V_m$. For very large values of $d$, $V_m$ approaches zero. At very small values of $d$ the curve for $Q_m$ approaches a vertical line and $V_m$ approaches a value that makes $\phi_m = \phi$. This value is the one to be expected on the grounds of the simple picture, usually shown in textbooks, that has the space-charge region making the adjustment to bring the free-space potentials of the metal and semiconductor surfaces into coincidence. This simple picture hardly ever applies because semiconductor surfaces always seem to have a nonzero value of surface potential, indicating the presence of surface states.

If surface states are present, the charge to be balanced by $Q_{sc}$ is the sum of $Q_m$ and $Q_s$ shown in the figure by the dashed line. It will be assumed in the subsequent discussion that the surface states remain unchanged in density and energy level as the metal is brought up to the semiconductor. This assumption would probably be a bad one for

close contact of the two surfaces. For the situation shown in the figure the surface potential will differ negligibly from $V_{so}$. As $d$ is decreased, however, the surface potential will move over and finally approach the value that makes $\phi_m = \phi$. Many other possible combinations of the parameters can be imagined. In any case, however, the shift in the surface potential caused by the approach of the metal will be equal to the difference between $V_{so}$ and whatever potential is needed to make

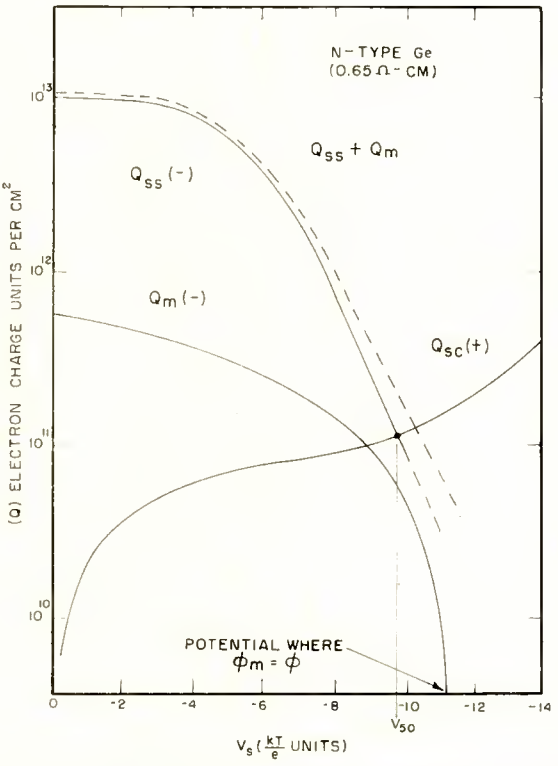



Fig. 17—Charge balance for metal–semiconductor contact.

$\phi_m = \phi$. It is in this sense that the surface states modify the adjustment of the semiconductor surface potential. The absolute density of the surface states themselves, as suggested by Bardeen, need not play a dominant role; the state energy levels may be of greater importance. Although the above treatment is an idealized one it gives a good physical outline of what one can expect to happen at metal–semiconductor contacts.

## Gas–Semiconductor Contact

As pointed out by P. B. Weisz[29] and M. Boudart[30] the adsorption of gas atoms or molecules on a semiconductor surface can be visualized in much the same manner as a metal–conductor contact. One thinks of the ionization potential or electron affinity of an impinging particle as a sort of work function. As with the metal–semiconductor contact the surface space-charge region of the semiconductor has to adjust itself so that the appropriate equilibrium conditions of the gas–semiconductor system are satisfied.

To see how these concepts operate, consider the sample system in Figure 18. The semiconductor work function is $\phi$ and the electron affinity of each impinging particle is $A$. For simplicity, surface states are neglected. At contact, electrons will flow out of the semiconductor

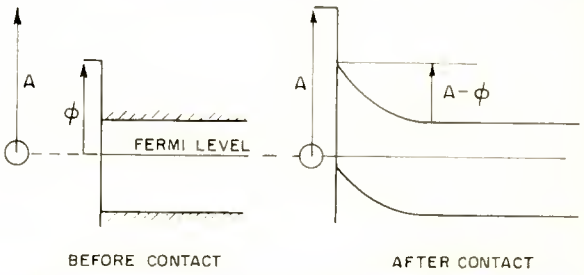



Fig. 18—Energy diagram for gas–semiconductor contact.

to be held by adsorbed atoms at the surface. The electron flow will cease when enough electrons have been transferred to bring the free-space potentials of the gas atoms and semiconductor surface into coincidence as shown on the right in the figure. A potential $A - \phi$ will appear across the space-charge region. This specifies a particular charge $Q_o$ as shown on the charge diagram in Figure 19. This, in turn, fixes the number of particles that can take part in the charge transfer process at the surface.

This picture oversimplifies the situation because the quantities $A$ and $\phi$ may be modified at contact. Other phenomena, such as polarization,[29] can also play a role.

It is easy to follow through the case where surface states are present. These states should, in general, markedly change the number of particles that can be adsorbed.

---

[29] P. B. Weisz, "Effects of Electronic Charge Transfer Between Adsorbate and Solid on Chemisorption and Catalysis," *Jour. Chem. Phys.*, Vol. 21, p. 1531, September, 1953.

[30] M. Boudart, "Electronic Chemical Potential in Chemisorption and Catalysis," *Jour. Am. Chem. Soc.*, Vol. 74, p. 1531, March, 1952.

This idealized picture of adsorption on semiconductors is an interesting one in many respects. It is consistent with a number of phenomena that have been observed. It says that the heat of adsorption should vary from $A - \phi$ for bare surfaces to zero for maximum coverage. Varying heats of adsorption are indeed observed. However, it is well to note that other phenomena could also cause this. The

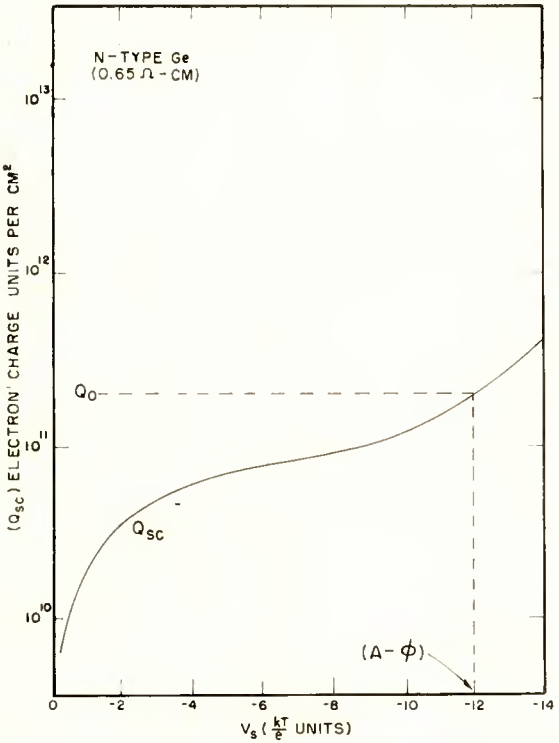



Fig. 19—Charge balance for gas–semiconductor contact.

picture further predicts that the maximum quantity of particles adsorbed should depend upon the specimen doping and also upon the excess carriers present. Both of these predictions have been verified. The picture is also consistent with a number of catalytic phenomena that have been observed in recent years. Hauffe and Schwab of Germany have pioneered in this new field.[31]

[31] K. Hauffe, "Gas Reactions on Semiconducting Surfaces and Space Charge Boundary Layers," pp. 259-282; G. Schwab, "Experiments Connecting Semiconductor Properties and Catalysis," pp. 283-296, *Semiconductor Surface Physics*, University of Pennsylvania Press, Philadelphia, Pa., 1957.

## P-N Junctions

A p-n junction can be considered as two surfaces back-to-back. There are two space-charge regions, one in each type of material. If there is no trapped charge the system will come into balance when the charge in one space-charge region is exactly balanced by charge in the other one. In addition to this, the carrier densities must join smoothly, or match, at the interface. For thermal equilibrium conditions this is equivalent to saying that the Fermi levels must match at the interface. Equivalently, both the electric-charge and carrier-density conditions will be satisfied when the electric field and the carrier densities are continuous across the interface. It can be shown that the carrier-density matching condition determines the total potential across the junction, while the space-charge matching condition determines the

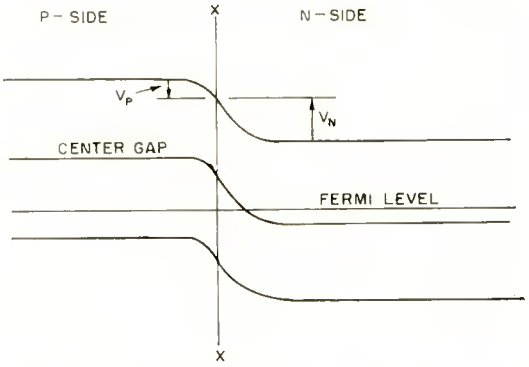



Fig. 20—Energy diagram for p-n junction.

apportionment of the total potential between the two sides of the junction.

First consider the thermal equilibrium case with the usual band picture shown in Figure 20. The p-side is much more heavily doped than the n-side. The interface plane is at x-x. The condition for the match between the Fermi levels is seen to be equivalent to equal $\phi_s$ potentials (see Figure 7) on both sides of the interface. The graphical treatment of the system is sketched in Figure 21. The space charge, along with the electron density at the interface, is plotted vertically on a logarithmic scale; the abscissas have a linear scale and refer to the space-charge potentials. The space-charge curve on the p-side is appropriate for heavy doping (0.013 ohm centimeter) and that on the n-side, for the light doping (6.5 ohm-centimeter). Also displayed on each side are curves for the electron density, n, at the interface. Since the electron density at the interface is an exponential function

of the surface potential, these curves must appear as straight lines with equal slopes on the semilog coordinate system. The vertical intercept is higher on the n-side because the electrons are the majority carriers on this side.

The hole densities could equally well have been used. Because of the way the hole and electron densities are related it is only necessary to specify one or the other in obtaining the match of carrier densities at the interface.

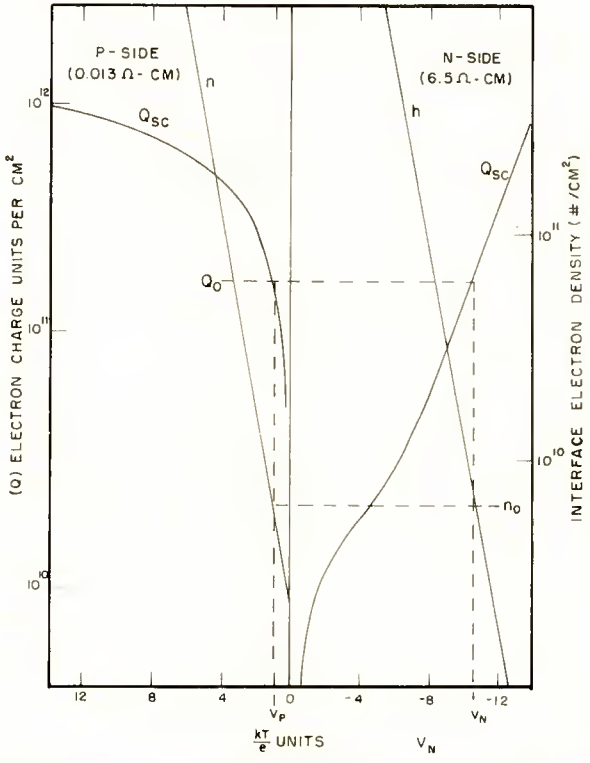



Fig. 21—Charge and interface carrier density balance at p-n junction—no excess carriers.

By trial and error a space charge $Q_0$ is selected to make the potential $V_p$ and $V_n$ give equal values of $n_0$ on each side. The equilibrium solution is indicated by the dashed lines. A few trials are sufficient to show that the solution is unique. A well known result is seen immediately; the largest space-charge potential appears on the less heavily doped side. Regardless of the disparity in doping, however, roughly one $kT/e$ unit of potential will always remain on the heavily doped side. This follows because the knee of the space charge curve is always

located near this potential. The reason for this was pointed out in the discussion of Figure 8. It should also be noted that the graphical treatment will handle the system even if one side is intrinsic, or if it is of the same conductivity type as the other side. The simple Sckottky depletion layer treatment could not handle such cases.

The effect of traps at or near the interface can be introduced by adding the trapped charge curves of the form shown in Figure 6 to either of the space-charge curves. The composite curves are then used in the graphical treatment. Qualitatively, one can think of the trapped charge as locally modifying the donor and acceptor densities.

The case where excess carriers are present and the junction is open circuited is equally easy to treat graphically. The charge balance and carrier density matching conditions must still hold. The only modifications are the change in shape of the space-charge curves (see Figure 10) and the upward displacement of the carrier density curves. For a given increase in the excess carriers, the density curve on the p-side will be displaced upwards more than the one on the n-side, because the electrons are minority carriers on the p-side, but majority carriers on the n-side. The exact relationship between the displacement on each side is somewhat complex. It is chosen in such a way that for all densities of excess carriers the hole densities, and also the electron densities, match on both sides of the interface.

For a given density of excess carriers the system pictured in Figure 21 is modified to that shown in Figure 22. The solution, again, can be obtained by trial and error. It can be seen that (1) both space-charge voltages have decreased, (2) the space charge on each side has decreased, and (3) the electron density at the interface has increased. These changes are accentuated as the excess carrier densities are increased still further. The space-charge curves are further modified, and the line for the interface density on the p-side has an intercept on the ordinate axis that approaches the one on the n-side. The intercepts will be equal only at the intrinsic limit. As this limit is approached at very high densities of injected carriers the space-charge voltages and the space charges approach zero. The largest decrease in space-charge voltage obviously occurs on the high-resistivity side.

The total decrease in the sum of the space-charge voltages for a given level of injected carriers is known as the junction photovoltage. This is indeed the analogue to the surface photovoltage discussed previously. The junction photovoltage is normally analyzed in terms of the diode equation which, as it is ordinarily used, cannot account for the saturation of the photovoltage, nor can it handle the case where trapped charge plays a role. On the other hand, the more basic treat-

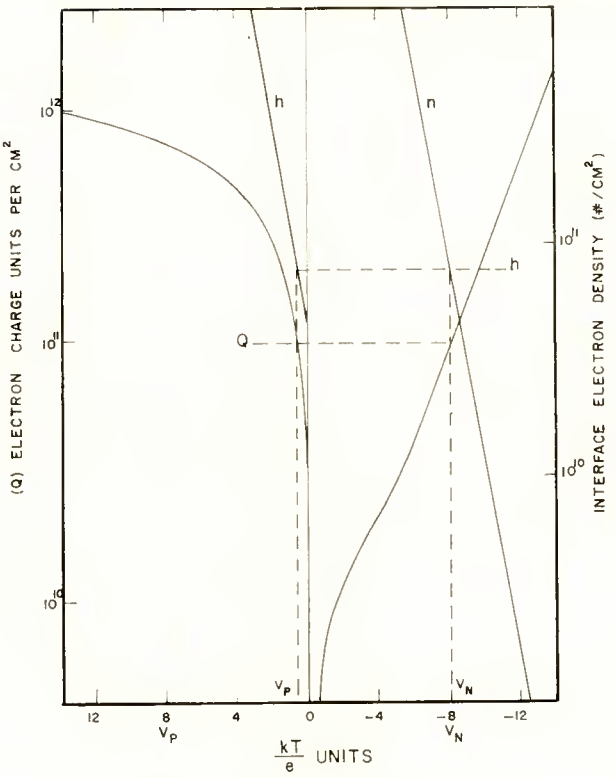



Fig. 22—Charge and interface carrier density balance at p-n junction—excess carriers.

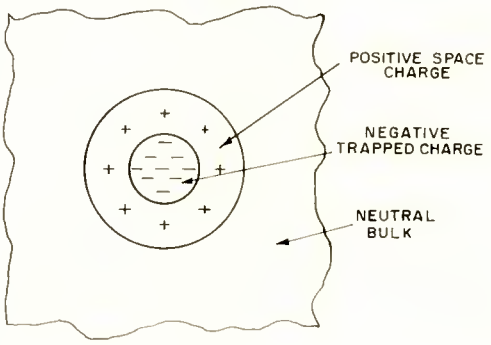



Fig. 23—Sketch of space-charge region surrounding localized bulk trapped charge.

ment portrayed graphically above can handle, in principle at least, most aspects of the junction photovoltage.

## Miscellaneous Systems

The close analogy among a p-n junction, a surface, and a two-electrolyte system has been pointed out by Brattain.[32] In this analogy the two electrolytes have different ion concentrations and are separated by a semipermeable membrane. This electrolyte system will come into equilibrium exactly in the same manner as the p-n junction treated earlier. There must be balance between the space charges in both electrolytes and there also must be a match in the ion densities at the interface. Brattain and Garrett have carried the analogy further with studies of the interface between germanium and an electrolyte.[33]

An interesting extension of the surface picture has been advanced to explain slow recombination phenomena in the bulk of a semiconductor.[34] Here localized trapped charge, presumably due to dislocations, is pictured as being surrounded by a spherical space-charge well as indicated in Figure 23. The trapped charge will always be balanced by the space charge as with the surface. The system can obviously be handled by the same graphical technique as the surface, for both equilibrium and nonequilibrium situations. When there are excess carriers present and the trapping states act as recombination centers, there will be a rather complicated recombination process. In particular, Morrison has shown that this process has a distribution of time constants and leads to long tails on the excess carrier decay curve. Morrison believes that these might be associated with bulk 1/f noise.

---

[32] Physics Colloquium at Princeton University, January 24, 1957.

[33] W. H. Brattain and C. G. B. Garrett, "Experiments on the Interface Between Germanium and an Electrolyte," *Bell Sys. Tech. Jour.*, Vol. 34, p. 129, January, 1955.

[34] S. Roy Morrison, "Recombination of Electrons and Holes at Dislocations," *Phys. Rev.*, Vol. 104, p. 619, November, 1956.

# COMPARISON OF THE SEMICONDUCTOR SURFACE AND JUNCTION PHOTOVOLTAGES

By

E. O. Johnson

RCA Laboratories,
Princeton, N. J.

***Summary***—*It is shown analytically that the surface and junction photovoltages are almost identical phenomena. The functional dependence of the two photovoltages on the light-injected carrier densities is exactly the same, except in the region of saturation. On the other hand, the surface and the junction are basically different with respect to the effect of electric charge changes in traps. These charge changes can have a profound effect on the surface photovoltage but no direct effect at all on the total junction photovoltage.*

*The analysis applies to any type of junction, LH or p-n, and accounts, in distinction to the ordinary diode equation, for saturation effects that stem from changes in the majority carrier concentrations. A possible mechanism for anomalous junction behavior is briefly discussed.*

## Introduction

A PHOTOVOLTAGE can be generated at a semiconductor surface as well as at a junction.[1] The surface photovoltage is a useful parameter in surface studies and has been analyzed in detail.[2,3] In fact, it seems to have been more thoroughly analyzed than the more familiar junction photovoltage. Saturation effects and those due to charge changes in traps have been treated in detail. To the best of the author's knowledge, an equally complete analysis has not been made of the junction photovoltage. The latter photovoltage is usually treated in terms of the diode equation.[4] In its usual form this equation can not account for saturation effects nor does it have much to say about the effect of traps on the photovoltage. These have been of particular interest as a means by which anomalous junction behavior might be explained.[5]

---

[1] W. H. Brattain and J. Bardeen, "Surface Properties of Germanium," *Bell Sys. Tech. Jour.*, Vol. 32, p. 1, January, 1953.

[2] C. G. B. Garrett and W. H. Brattain, "Physical Theory of Semiconductor Surfaces," *Phys. Rev.*, Vol. 99, p. 376, July, 1955.

[3] E. O. Johnson, "Large Signal Surface Photovoltage Studies with Germanium," *Phys. Rev.*, to be published.

[4] C. Kittel, *Introduction to Solid State Physics*, John Wiley & Sons, Inc., New York, 1953, pp. 291, 294.

[5] H. Kleinknecht and K. Seiler, "Single Crystals with p-n Junctions of Silicon," *Zeit. fur Physik*, Vol. 139, p. 599, No. 5, 1954.

A number of questions might now be asked.[6] How are the surface and junction photovoltages related? How can the more complete analysis of the surface photovoltage be applied to the junction? What effects do charge traps in the vicinity of the junction have on the junction photovoltage? How does the junction photovoltage apportion itself on each side of the junction interface? It is the object of this paper to answer such questions.

Although it might appear similar to a "half junction," the surface space-charge region functions very much like a true junction. The surface and junction give exactly the same photovoltage for a given density of light-injected carriers, except in the saturation region where the two systems are not strictly comparable. One major difference between the surface and junction lies in the fact that the surface photovoltage can be profoundly affected by charge changes in traps in the depletion region, while the junction photovoltage can not. Crudely speaking, this is because trapped charge determines the "doping" for the surface; for a junction, the doping is determined during fabrication. Anomalous behavior of the junction, however, may possibly be associated in part with changes in the effective doping due to either bulk or surface effects.

The following analysis was facilitated by the fact that the complete space-charge and trapped-charge relations for a surface had already been worked out in detail.[2,3] To handle junctions it was only necessary to modify boundary conditions. The basic space-charge and carrier-density relations, and their use in describing the surface photovoltage, are briefly reviewed in the first section. The second section treats, both analytically and graphically, the case of a junction where no excess carriers are present. Relations are developed to show the apportionment of the contact potential on each side of the junction for all combinations of doping. The third section deals with the case where excess carriers are present and an open-circuit photovoltage appears across the junction. All doping combinations are taken into account and the junction photovoltage is compared with that generated at a surface. The effect of traps is discussed. In the fourth section brief comments are made about a modified form of the diode equation and also about the extension of the analysis of the third section to the high-injection, closed-circuit case.

## Surface Photovoltage

The presently accepted picture[7] of the energy bands at a semi-

[6] These questions and the suggestion for the analysis were brought to the author's attention by P. Rappaport and J. J. Loferski.

[7] R. H. Kingston, "Review of Germanium Surface Phenomena," *Jour. Appl. Phys.*, Vol. 27, p. 101, February, 1956.

conductor surface under equilibrium conditions is shown in Figure 1. The charge in the space-charge region is composed of fixed donors, acceptors, and mobile holes and electrons. This charge balances the charge immobilized in trapping states that exist on or near the surface proper. This equality of opposite charge forms a dipole layer at the surface. There are two types of trapping states. One type is called "slow" because it exchanges charge slowly (time constants $\approx 10^{-1}$ second) with the bulk material. These states are ambient sensitive and presumably located in or on the surface oxide layer. The other type of state is "fast" because it can exchange charge with the bulk in times of the order of microseconds. This type of state is thought

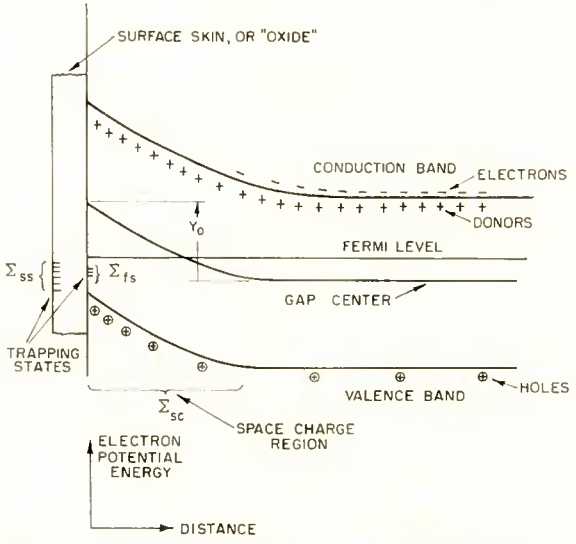



Fig. 1—Energy bands at a semiconductor surface.

to be relatively ambient insensitive and located at the bulk–oxide interface.

The charge in the space-charge region is computed from Poisson's equation and the usual Boltzmann statistics. This is conveniently accomplished by using the symbolism of Garrett and Brattain,[2] which is defined in Figure 1 and Table I. The number of charge units $\Sigma_{sc}$ in the space-charge region and the electric field $d\psi/dx$, in volts per centimeter, at the surface edge of the space-charge region is found to be[2]

$$\Sigma_{sc} = n_i L F = \begin{bmatrix} 3.5 \times 10^9, \text{Ge} \\ 3.94 \times 10^7, \text{Si} \end{bmatrix}, \quad \frac{d\psi}{dx} = \left( \frac{2}{\beta L} \right) F = \begin{bmatrix} 370, \text{Ge} \\ 8.93, \text{Si} \end{bmatrix} F, \tag{1}$$

where $F$, the space-charge factor, is a dimensionless quantity given by

$$F = \mp \sqrt{\Lambda(e^{-Y}-1) + \Lambda^{-1}(e^{Y}-1) + (\Lambda-\Lambda^{-1})\, Y + \Lambda(e^{Y}+e^{-Y}-2)\, \Delta_p}. \tag{2}$$

*Table I*—Symbols

$\Sigma_{sc}$ = number of electron charge units per square centimeter of surface, or interface, area in the space-charge region.

$\Sigma_{ss}$ = number of electron charge units per square centimeter of surface in slow states.

$\Sigma_{fs}$ = number of electron charge units per square centimeter of surface in fast states.

$\psi$ = electrostatic potential.

$Y$ = electrostatic potential, in $kT/e$ units, across a space-charge region. This has a negative value if the energy bands bend upwards in the conventional band picture and a positive value of the bend is downwards. $Y_0$ refers to the equilibrium value of $Y$. For junctions, $Y_0$ will designate the contact potential.

$\Delta Y = Y - Y_0$, the surface or junction photovoltage.

$\Lambda = \left(\dfrac{p_0}{n_0}\right)^{1/2} = \dfrac{p_0}{n_i} = \dfrac{n_i}{n_0}$, the doping factor.

$L = \left[\dfrac{\mathcal{E}}{2\pi e n_i B}\right]^{1/2}$, a characteristic length differing only slightly from the Debye length. At room temperature $L = 1.4 \times 10^{-4}$ for Ge, and $5.8 \times 10^{-3}$ for Si.

$\mathcal{E}$ = dielectric constant of material.

$\beta = e/kT$, where $e$ is the absolute value of the electric charge, and $k$ and $T$ have their usual meanings.

$n_i$ = intrinsic carrier density. At room temperature, $n_i = 2.5 \times 10^{13}$ for germanium, and $6.8 \times 10^{10}$ for silicon.

$n_0$ = free-electron density deep in bulk under equilibrium conditions.

$p_0$ = free-hole density deep in bulk under equilibrium conditions.

$p_P, n_P$ = hole and electron densities in $P$-material at junction interface.

$p_N, n_N$ = hole and electron densities in $N$-material at junction interface.

$\Delta_n, \Delta_p$ = injection factor, fractional increase in electron and hole densities due to injection. When applied to junction, prefixes $P$ and $N$ are affixed to denote side of application.

$\Delta n, \Delta p$ = increase in equilibrium electron and hole densities due to light-generated carriers. Note:

$$\Delta_n = \frac{\Delta n}{n_0}, \qquad \Delta_p = \frac{\Delta p}{p_0}.$$

The sign convention is such that the negative sign is used when $Y > 0$, and the positive sign when $Y < 0$. There are four terms under the radical. The first of these is proportional to the hole charge; the second, to the electron charge; the third, to the fixed donor and acceptor charge; and the fourth, to the mobile carrier charge added by injection. The above equations were derived under the assumption that all the donors and acceptors were ionized and that the carrier diffusion lengths are long compared to the width of the space-charge region ($\leq 10^{-5}$ centimeter).

The fractional increase $\Delta_p$ in the hole density due to injection is related to the fractional increase $\Delta_n$ in the electron by

$$\Lambda^2 \Delta_p = \Delta_n. \tag{3}$$

This equation has its origin in the fact that the light-generated carriers are added in pairs. This can be seen from the definitions and relations listed in Table I.

The value of $F$, $F_0$ given by Equation (2), when $\Delta_p = \Delta_n = 0$, is plotted in Figure 2 for a surface tending toward inversion.[8] In this case, the space charge arises from minority carriers and fixed charge of the same sign. Because of the symmetry of $F_0$, the curves in Figure 2 apply equally well to $P$- or $N$-type material. The same is true for the accumulation layer case whose curves are shown in Figure 3.[8] Here the space charge arises primarily from majority carriers. Figure 4 shows how the space-charge factor, $F$, behaves as a function of carrier injection for a surface tending toward inversion.

The form of the space charge curves can be understood as follows: For intrinsic material the space charge arises from one sign of carrier and increases exponentially with surface potential, except near $Y_0 = 0$. It thus appears as a straight line on the semilogarithmic plot. The gently sloping plateau on the curves in Figure 2 arises from fixed donor or acceptor charge. This region is closely described by the Schottky depletion layer formula. The knee on the plateau occurs at roughly one $kT/e$ unit of surface potential. This potential is sufficient to reduce the majority carrier charge, opposite in sign to the fixed charge, to a relatively small value. The plateau disappears at high surface potentials due to the exponential increase of the minority carrier charge. This exponential increase sets in at higher surface potentials for heavily doped material because the heavy doping reduces

[8] R. H. Kingston and S. F. Neustadter, "Calculation of the Space Charge, Electric Field, and Free Carrier Concentration at the Surface of a Semiconductor," *Jour. Appl. Phys.*, Vol. 26, p. 718, June, 1955.

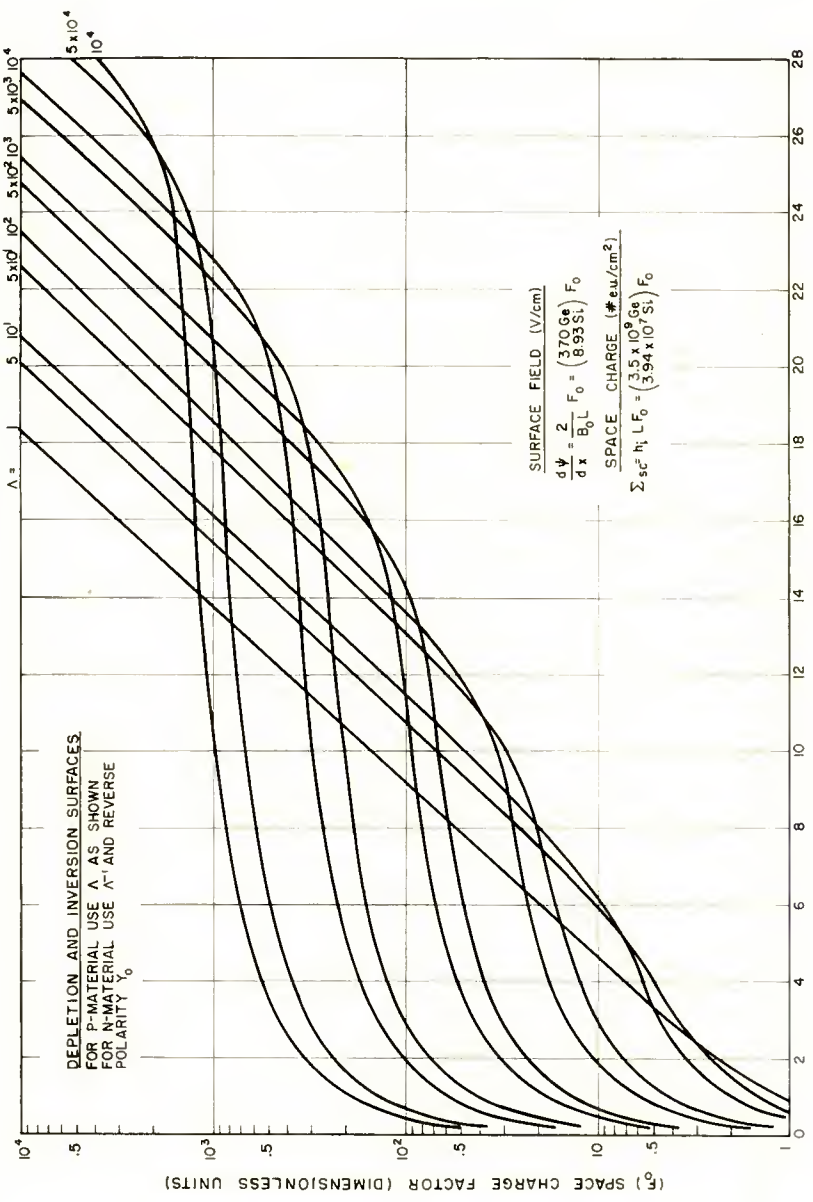



Fig. 2—Space-charge factor for surfaces tending toward inversion no injection.

the minority carrier concentration. On the other hand, injection increases the minority carrier density and causes the exponential increase to set in at lower surface potentials. The accumulation layer curves show more charge, at a given surface potential, than the inversion layer ones because the majority carriers are involved. The knee in these curves also occurs close to one $kT/e$ unit of potential

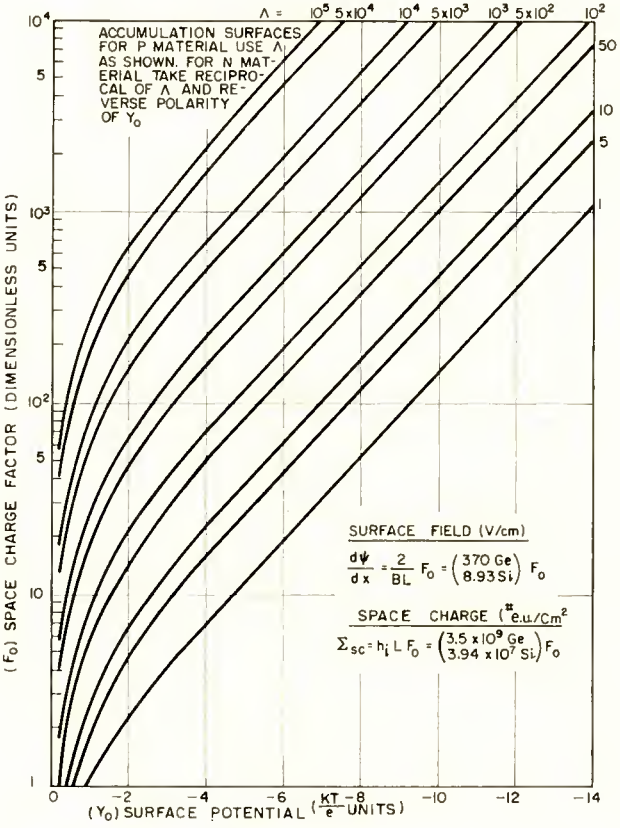



Fig. 3—Space-charge factor for accumulation layer surfaces, no injection.

because this potential is sufficient to make the majority carrier charge increase appreciably above the fixed donor (or acceptor) charge. Injection has little effect on the accumulation layer curves unless it is high enough to make an appreciable fractional increase in the majority carrier density.

The relation between Λ and resistivity for room temperature is plotted in Figure 5 for germanium, and in Figure 6 for silicon.

The surface photovoltage is usually studied for the simple case where the trapped charge in the surface states remains constant as the carrier injection is changed. Because of the electrical neutrality requirement the space charge must also remain constant. The surface

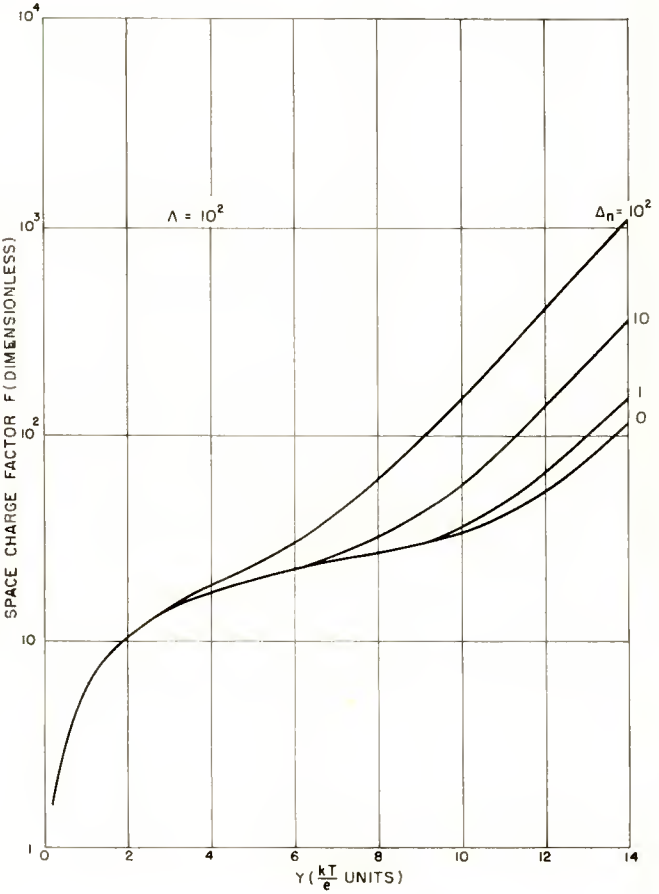



Fig. 4—Effect of injection on the space charge factor for case $\Lambda = 10^2$.

potential will then change with carrier injection in a manner prescribed by Equation (2), when $F$ remains constant. This value of $F$ is, moreover, equal to the value when $\Delta_p = 0$, that is, $F_0$. Using the definition

$$F_Y^2 = \Lambda(e^{-Y} - 1) + \Lambda^{-1}(e^{Y} - 1) + (\Lambda - \Lambda^{-1})Y, \tag{4}$$

it is found that[3]

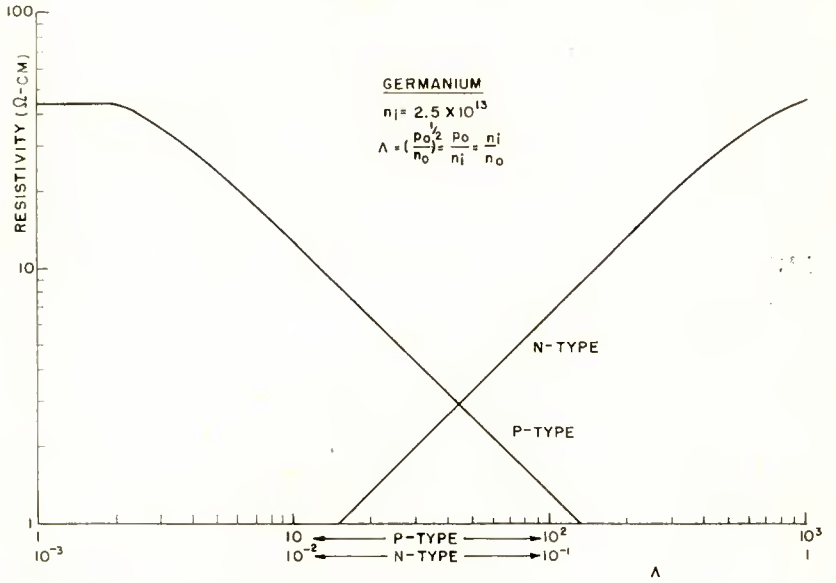



Fig. 5—Relation between doping factor Λ and resistivity for germanium

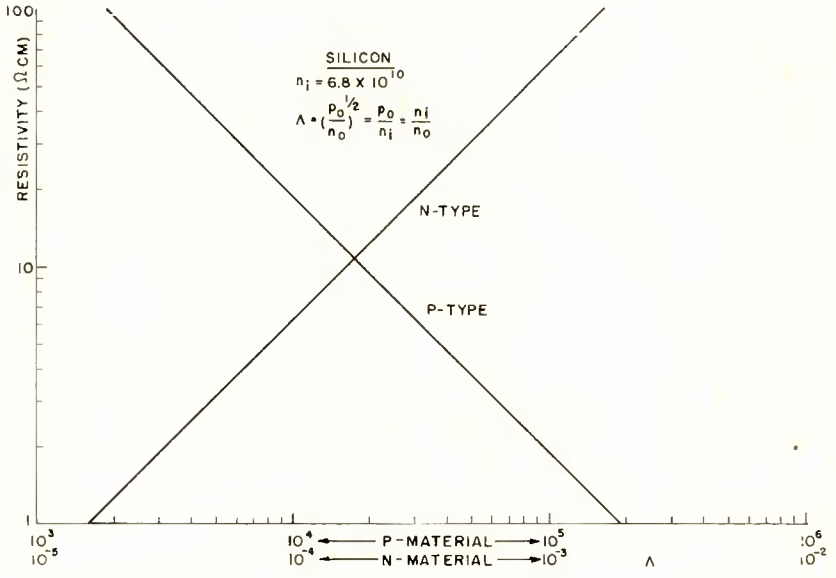



Fig. 6—Relation between doping factor Λ and resistivity for silicon.

$$\Delta_p = \frac{F_0^2 - F_Y^2}{\Lambda(e^Y + e^{-Y} - 2)}. \tag{5}$$

It is much more convenient to use this relation directly rather than to attempt to find an explicit relation for $Y$, and hence for $Y - Y_0$, the surface photovoltage. Some typical curves derived from Equation (5) are shown in Figure 7. These curves could also have been deduced

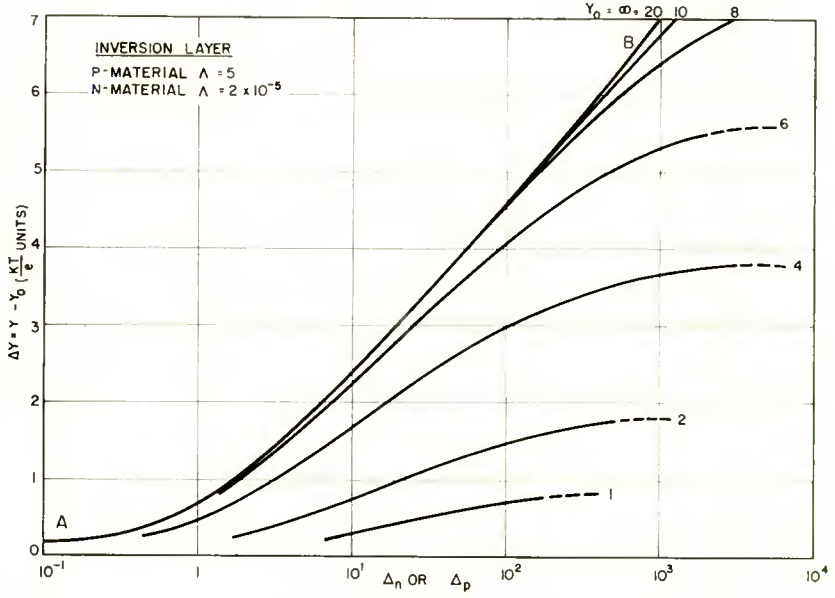



Fig. 7—Surface photovoltage as a function of injection for case $\Lambda = 5$.

graphically.[3,9] Equation (5) neglects the Dember potential generated in the bulk material by the ambipolar field that serves to equalize the flow rates of diffusing holes and electrons.[10]

At low values of the injection parameter the surface photovoltage varies linearly. This fact has been applied to advantage in using the photovoltage to monitor the time behavior of minority carriers.[11] At intermediate values of injection a logarithmic relation exists between

[9] E. O. Johnson, "Simplified Treatment of Electric Charge Relations at a Semiconductor Surface," *RCA Review*, to be published.

[10] W. van Roosbroeck, "Theory of the Photomagnetoelectric Effect in Semiconductors," *Phys. Rev.*, Vol. 101, p. 1713, March, 1956.

[11] E. O. Johnson, "Measurement of Minority Carrier Lifetimes with the Surface Photovoltage," *Jour. Appl. Phys.*, to be published.

injection and the photovoltage. At high values of injection the photovoltage curves saturate, approaching an asymptotic value that is equal in magnitude to the equilibrium potential across the space-charge region. This represents the condition where the energy bands at the surface are completely flattened. The portion AB of the surface photovoltage curves (see Figure 7) is the limiting curve for $Y_0 > 10$. It corresponds to the case where the minority carrier charge dominates so that Equation (5) reduces to

$$\Delta_p = \frac{\Lambda e^{-Y_0} - \Lambda e^{-Y}}{\Lambda e^{-Y}} = e^{Y-Y_0} - 1, \tag{6}$$

or, $Y - Y_0 =$ surface photovoltage $\Delta Y = \ln(1 + \Delta_p)$. It will be shown later that the unsaturated junction photovoltage is described by exactly the same equation.

The surface photovoltage is detected and measured with a light-transparent electrode capacitively coupled to the surface.[1-3]

The more complicated case, where the trapped charge changes with carrier injection, has been handled in detail elsewhere.[3,9] The situation was treated graphically and it was shown that these charge changes could have a very pronounced effect on the surface photovoltage. A simple way of looking at these effects is as follows: The surface trapped charge determines the equilibrium value of surface potential $Y_0$. If this trapped charge remains constant, the system must trace out the appropriate photovoltage curve, for example, one of the curves in Figure 7. On the other hand, if the trapped charge changes with injection, then the effective value of $Y_0$ also changes. Consequently, the system traces out a curve that corresponds to a mixture of $Y_0$'s. If one considers the trapped charge as determining the conductivity type, or "doping," of the surface edge of the space charge region, then one might think of the "doping" as having changed during injection. This contrasts with a junction where the doping is fixed during fabrication. As will be seen, the junction photovoltage is unaffected by charge changes in traps.

The doping factor $\Lambda$ deserves a few words of comment with respect to nonuniformly doped specimens. First of all, let us assume that the light shines on the surface and is effectively absorbed in a short distance ($\sim 10^{-6}$ centimeter for band-gap radiation). The carriers generated at the surface will diffuse away into the bulk material, their density falling off in an exponential manner characterized by the diffusion length. Deep in the bulk, beyond the depth where the last excess carriers have recombined, the local doping of the material can

obviously have no effect upon photovoltages. Only the doping in the region a few diffusion lengths in from the surface can affect the photovoltages. The doping in the space-charge region can have a direct effect upon the surface photovoltage as can be seen from Equation (5). The effect of the doping in the region lying between the bulk edge of the space charge region and the plane where the last excess carriers have recombined is not so obvious. If the doping varies gradually, the Dember potential will be affected. Local "bumps" of doping will contribute no appreciable photovoltage if the bumps are smaller in dimension than a diffusion length. This follows because the photovoltage generated on one side of the bump, considered as a junction, will be cancelled by that generated on the other side. On the other hand, if the two sides of the bump are separated by more than a diffusion length, the local junction photovoltages will not cancel each other because of the difference in the excess carrier density on each side. It is also plausible that changes in the bulk diffusion length, due to injection-induced lifetime changes, could introduce variations of the anomalous components contributing to the surface photovoltage.

## P-N Junction, Equilibrium Case

For this and the nonequilibrium case a p-n junction can be considered as being two surfaces back-to-back as shown in Figure 8.[9] Two boundary conditions must be satisfied. First, for reasons of electrical neutrality, the charge in one space-charge region must be exactly balanced by that in the other. Second, the hole and electron densities must join smoothly at the interface. For the equilibrium case this is the same as specifying that the Fermi levels line up at the interface. An equivalent, more compact, way to specify the boundary conditions is to state that the electric field and carrier densities must be continuous across the interface.

The space-charge equality condition for the equilibrium case is handled by applying Equation (2), with $\Delta_p = 0$, to both the $P$ and $N$ sides and equating. This gives

$$\Lambda_P(e^{-Y_P}-1) + \Lambda_P^{-1}(e^{Y_P}-1) + (\Lambda_P-\Lambda_P^{-1})Y_P$$
$$= \Lambda_N(e^{-Y_N}-1) + \Lambda_N(e^{Y_N}-1) + (\Lambda_N-\Lambda_N^{-1})Y_N. \quad (7)$$

The density matching condition is satisfied if

$$n_P = nie^{Y_P-\ln\Lambda_P} = nie^{Y_N-\ln\Lambda_N},$$

and

$$p_P = nie^{-(Y_P - \ln \Lambda_P)} = nie^{-(Y_N - \ln \Lambda_N)} = p_N .$$

Both equations lead to the relation

$$Y_P - \ln \Lambda_P = Y_N - \ln \Lambda_N. \tag{8}$$

This gives the well-known result

$$Y_0 = Y_P - Y_N = \ln \left( \frac{\Lambda_P}{\Lambda_N} \right) = \ln \left( \frac{{}_P p_0}{{}_N p_0} \right) = \ln \left( \frac{{}_N n_0}{{}_P n_0} \right), \tag{9}$$

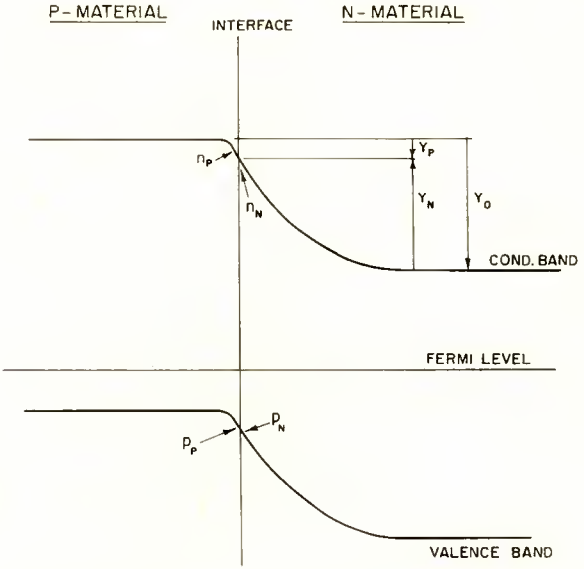



Fig. 8—Energy diagram for a p-n junction. Equilibrium case, no injection.

where $Y_0$ is the so-called contact potential across the junction. In the above equations, and in those to follow, the subscripts or prefixes, $P$ and $N$, refer to the $P$ and $N$ sides.

If Equation (8) is substituted into Equation (7), explicit expressions for $Y_N$ and $Y_P$ are obtained:

$$Y_N = \frac{(\Lambda_P + \Lambda_P^{-1}) - (\Lambda_N + \Lambda_N^{-1}) - (\Lambda_P - \Lambda_P^{-1}) \ln \left( \frac{\Lambda_P}{\Lambda_N} \right)}{(\Lambda_P - \Lambda_P^{-1}) - (\Lambda_N - \Lambda_N^{-1})}, \tag{10}$$

$$Y_P = Y_N + \ln\left(\frac{\Lambda_P}{\Lambda_N}\right). \tag{11}$$

The case where both sides of the junction are equally, but oppositely, doped makes $\Lambda_P = \Lambda_N^{-1}$ and leads to the expected equal division of potential across each side of the interface. That is, $Y_P = -Y_N = \ln \Lambda_P = -\ln \Lambda_N$.

To examine the case where one side is more heavily doped than the other, let

$$\Lambda_P = x\Lambda_N^{-1}, \tag{12}$$

where $x$ is to be some arbitrarily chosen factor describing the extent that one side is more heavily doped than the other. If $x > 1$, the $P$ side is more heavily doped than the $N$ side. The opposite is true for $x < 1$. Using the definition (12) in Equations (10) and (11), there results

$$Y_N = \frac{(\Lambda_P + \Lambda_P^{-1}) - (x\Lambda_P^{-1} + x^{-1}\Lambda_P) - (\Lambda_P - \Lambda_P^{-1}) \ln x^{-1}\Lambda_P^2}{(\Lambda_P - \Lambda_P^{-1}) - (x\Lambda_P^{-1} - x^{-1}\Lambda_P)}, \tag{13}$$

$$Y_P = Y_N + \ln x^{-1}\Lambda_P^2. \tag{14}$$

From these it is seen that if $x > 1$, $Y_N > Y_P$, and vice versa. With increasing disparity in doping, the potential on the heavily doped side asymptotically approaches a minimum value of $kT/e$. This corresponds to the knee in the space-charge curves shown in Figure 2.

As pointed out elsewhere,[9] a graphical representation of the junction relations is very helpful in showing how the system fits together. A typical case is shown in Figure 9. On the left is the space-charge factor curve for $P$-type, 0.013 ohm-centimeter germanium and, on the right, the curve for the space-charge factor in 6.5 ohm-centimeter $N$ material. The horizontal dashed line $F_0$ gives equal charge intercepts on both curves, thus satisfying the electrical neutrality condition. The lines $n_P$ and $n_N$ represent the electron densities in each material at the interface. Both lines have the same slope but different intercepts on the axis of ordinates. The intercept on the left side is lower than the one on the right because the electrons are minority carriers in the $P$ material, and majority carriers in the $N$ material. The system is in equilibrium, corresponding to the previous analytical solution, when the space charge factor $F_0$ is chosen to give $Y_P$ and $Y_N$ values that make $n_P = n_N$. The equilibrium solution is shown in the figure with

$Y_P = 1$, $Y_N = -10.5$, $F_0 = 62$, $n_P = n_N = 6.4 \times 10^9$ per cubic centimeter. The space charge $\Sigma_{sc}$ on each side is found from Equation (1) to be $3.5 \times 10^9 \times 62 = 2.17 \times 10^{11}$ electron charges per square centimeter of interface. It is of interest to note that the contact potential $Y_0 = Y_P - Y_N$ can be obtained directly from the horizontal displacement between the lines $n_P$ and $n_N$. As can be seen from Equation (9), no recourse to the space-charge relations is necessary. It is only necessary


Fig. 9—Graphical solution for p-n junction in equilibrium, no injection.

to use the space-charge curves when the division of potential between the two sides is desired. This result has to do with the fact that the contact potential is the potential which balances the tendency for carrier diffusion caused by carrier density gradients. Equation (9) simply states that the density of carriers at one point is related to the density at some other point by the Boltzmann factor $e^{-Y_0}$ where $Y_0$ is the electric potential between the points. As for nonuniform

doping, the local variations in doping between these points are immaterial: only the doping at the end points is of consequence when these points are outside depletion regions. The requirement that the end points be outside depletion regions is equivalent to the analytical specification that the end points be greater than about a Debye length from regions where the doping changes rapidly over a Debye length. This requirement is not satisfied for the surface case so that it is not surprising that the surface photovoltage is dependent upon charge relations in the space-charge region.

### P-N Junction, Open-Circuit Nonequilibrium Case

When excess carriers are present the situation is somewhat more complicated. The same conditions for equality of space charge and carrier density continuity at the interface must still hold, however. For the space-charge balance,

$$\begin{aligned}\Lambda_P(e^{-Y_P}-1) + \Lambda_P^{-1}(e^{Y_P}-1) + (\Lambda_P-\Lambda_P^{-1})\,Y_P + \Lambda_P(e^{Y_P}+e^{-Y_P}-2)\,{}_P\Delta_p \\ = \Lambda_N(e^{-Y_N}-1) + \Lambda_N^{-1}(e^{Y_N}-1) + (\Lambda_N-\Lambda_N^{-1})\,Y_N \\ + \Lambda_N\,(e^{Y_N}+e^{-Y_N}+2)\,{}_N\Delta_p\,.\end{aligned} \tag{15}$$

For the electron match at the interface,

$$\begin{gathered}n_P = n_N \\ n_i(1+{}_P\Delta_n)\,e^{Y_P-\ln\Lambda_P} = n_i(1+{}_N\Delta_n)\,e^{Y_N-\ln\Lambda_N}\,,\end{gathered} \tag{16}$$

hence

$$e^{Y_P-Y_N} = \frac{\Lambda_P}{\Lambda_N}\,\frac{1+{}_N\Delta_n}{1+{}_P\Delta_n}\,.$$

For the hole match

$$\begin{gathered}p_P = p_N \\ n_i(1+{}_P\Delta_p)\,e^{-Y_P+\ln\Lambda_P} = n_i(1+{}_N\Delta_p)\,e^{-Y_N+\ln\Lambda_N}\end{gathered} \tag{17}$$

and

$$e^{Y_P-Y_N} = \frac{\Lambda_P}{\Lambda_N}\,\frac{1+{}_P\Delta_p}{1+{}_N\Delta_p}\,.$$

Equations (16) and (17) can be equated to give

$$\frac{1+{}_N\Delta_n}{1+{}_P\Delta_n} = \frac{1+{}_P\Delta_p}{1+{}_N\Delta_p}\,. \tag{18}$$

Using the relations

$$ {}_P\Delta_p = {}_P\Delta_n \Lambda_P^{\ 2}, \tag{19} $$

$$ {}_N\Delta_n = {}_N\Delta_p \Lambda_N^{\ 2}, \tag{20} $$

which follow from Equation (3) applied to the $N$ and $P$ sides, Equation (18) is solved to give

$$ {}_P\Delta_p = \frac{1}{2}\left(1 + \frac{1}{\Lambda_P^{\ 2}}\right)\left[-1 + \sqrt{1 + \frac{4\,{}_N\Delta_p}{\Lambda_P^{\ 2}\left(1 + \frac{1}{\Lambda_P^{\ 2}}\right)^2}(1 + \Lambda_N^{\ 2} + {}_N\Delta_p\Lambda_N^{\ 2})}\right] \tag{21} $$

In this equation the factor ${}_N\Delta_p$ is the independent and arbitrarily chosen injection parameter. The expected result, ${}_P\Delta_n = {}_N\Delta_p$, is obtained from Equation (21) for the symmetrical case $\Lambda_P = \Lambda_N^{-1}$.

From Equations (19) and (20) and the equations for the electron interface densities it is found that

$$ n_P = n_i(1 + {}_P\Delta_p\Lambda_P^{\ 2})\, e^{Y_P - \ln\Lambda_P}, \tag{22} $$

$$ n_N = n_i(1 + {}_N\Delta_p\Lambda_N^{\ 2})\, e^{Y_N - \ln\Lambda_N}. \tag{23} $$

Corresponding expressions can be obtained for the holes. The total voltage across the junction is found either from Equation (16) or (17) and is

$$ Y_P - Y_N = \ln\left(\frac{\Lambda_P}{\Lambda_N}\right)\frac{1 + {}_N\Delta_n}{1 + {}_P\Delta_n} = \ln\left(\frac{\Lambda_P}{\Lambda_N}\right)\frac{1 + {}_N\Delta_p\Lambda_N^{\ 2}}{1 + {}_P\Delta_p\Lambda_P^{\ 2}}, \tag{24} $$

where the appropriate substitutions have been made from Equations (19) and (20). The junction photovoltage $\Delta Y$ is obtained by taking the difference between Equations (24) and (9). Thus

$$ \Delta Y = \ln\left(\frac{\Lambda_P}{\Lambda_N}\right) - \ln\left(\frac{\Lambda_P}{\Lambda_N}\right)\frac{1 + {}_N\Delta_p\Lambda_N^{\ 2}}{1 + {}_P\Delta_p\Lambda_P^{\ 2}} = \ln\frac{1 + {}_P\Delta_p\Lambda_P^{\ 2}}{1 + {}_N\Delta_p\Lambda_N^{\ 2}}. \tag{25} $$

This is the voltage (in units of $kT/e$) that would be measured with a high-resistance voltmeter. Equation (25), like Equation (9) for the contact potential, makes no use of the charge balance relation; only if the apportionment of potential is desired need the charge-balance relation be used. This behavior is basically the same as in the equilibrium case; it is only a matter of having an electrostatic potential to balance the tendency for diffusion flow caused by carrier density gradients. As can be seen by considering the limiting case of infinitely high injection, the effect of the injection is to make the material become more intrinsic. Density gradients caused by the doping are diluted so that the potential across the junction decreases, approaching zero in value for very high injection. As for nonuniform doping, the nonequilibrium junction case is more complicated than the equilibrium one. Outside the depletion layer region the situation is closely similar to the surface case, except that there are two different pieces of bulk material to consider; the algebraic sum of potentials developed in each piece of bulk material appears to the external circuit. Also, account must be taken of the geometrical relation between the region of illumination and the junction area. With respect to local doping variations in the depletion region, the situation is basically the same as in the equilibrium case. Injection-induced charge changes in traps in the depletion region should, therefore, have no effect on the photovoltage because these changes can be considered as changes in the local doping. However, trap charge changes can effectively change the space-charge curves and so can have an effect on the division of potential between the two sides of the junction. In summary, the reason why the junction behaves differently from the surface with respect to charge changes in traps is roughly the following: With a junction one deals with the diffusion-balancing electrostatic potential existing between two regions (where the electrodes are effectively attached) that actually have a different doping and are both located outside the space-charge region. The system is one wherein particle densities are related solely by the net intervening potential difference. In the surface case the two regions normally have exactly the same doping; the potential difference arises from trapped charge outside the regions being considered. If this trapped charge changes, then the potential between the two regions changes. Furthermore, in the surface case one of the regions is in the space charge, or depletion, layer. This directly violates the requirement mentioned in the preceding section and makes the surface photovoltage dependent upon charge relationships in the space charge layer.

If the traps introduce strong recombination in the junction region,

the junction will act as though it were shunted with a resistance.[12] This could have an effect on the open-circuit photovoltage.

Returning to Equation (25), consider the simple case where there is equal doping on the $N$ and $P$ sides so that $\Lambda_P = \Lambda_N^{-1}$. Then, as pointed out previously, ${}_P\Delta_p = {}_N\Delta_n$, and ${}_N\Delta_n$ can be replaced by ${}_N\Delta_p\Lambda_N^2$ to make

$$\Delta Y = \ln \frac{1 + {}_N\Delta_p}{1 + {}_N\Delta_p\Lambda_N^2}. \tag{26}$$

For small injection, where $\frac{1}{2}({}_N\Delta_p)^2 \ll {}_N\Delta_p$, the linear approximation,

$$\Delta Y = \ln (1 + {}_N\Delta_p) \cong {}_N\Delta_p \tag{27}$$

applies. For intermediate injection, where ${}_N\Delta_p \geq 1$ and $\Lambda_N^2\,{}_N\Delta_p \ll 1$,

$$\Delta Y = \ln {}_N\Delta_p . \tag{28}$$

For large injection, where ${}_N\Delta_p\Lambda_N^2 \gg 1$,

$$\Delta Y = \ln \frac{1}{\Lambda_N^2} = \ln \frac{\Lambda_P}{\Lambda_N}. \tag{29}$$

The results obtained in Equations (27) and (28) correspond with those obtained from the familiar diode equation. However, the result in Equation (29) can only be extracted from the diode equation if it is modified in the manner shown in the next section. Equation (29) represents the saturation limit when the energy bands are completely flat across the junction so that the magnitude of the junction photovoltage must be exactly equal to that of the original contact potential. The saturation region for this case will start when ${}_N\Delta_p\Lambda_N^2 \simeq 10^{-1}$, or when ${}_N\Delta_p \simeq 10^{-1}/\Lambda_N^2$. For 4-ohm-centimeter germanium, $\Lambda_N^{-1} \simeq 50$ and ${}_N\Delta_p \simeq 250$, a value that can be attained with a focussed microscope light if the effective carrier lifetime is about 50 microseconds. For almost complete saturation, $\Lambda_N^2\,{}_N\Delta_p \simeq 10$ and ${}_N\Delta_p \simeq 2.5 \times 10^4$, a value too large to be obtained with any ordinary light source.

The junction photovoltage from Equation (26) is plotted in Figure 10 for several different values of doping. Up to the saturation region, these curves coincide perfectly with the surface photovoltage curves shown in Figure 7. This is because Equation (26), below saturation, is identical with Equation (6). Below the region of saturation the junction photovoltage curves for nonsymmetrical cases of doping are

[12] W. Shockley, *Electrons and Holes in Semiconductors*, D. Van Nostrand, Inc., New York, 1950.

identical with the curves in Figure 10. The doping in the nonsymmetrical case, as in the symmetrical case, only affects the point at which saturation sets in.

Strictly speaking, account must be taken of potentials developed in the bulk material on each side of the interface. These potentials are associated with the Dember,[10] or ambipolar, field which acts to equalize the flow rates of holes and electrons away from the region of illumination. This field always acts in a direction to slow down the

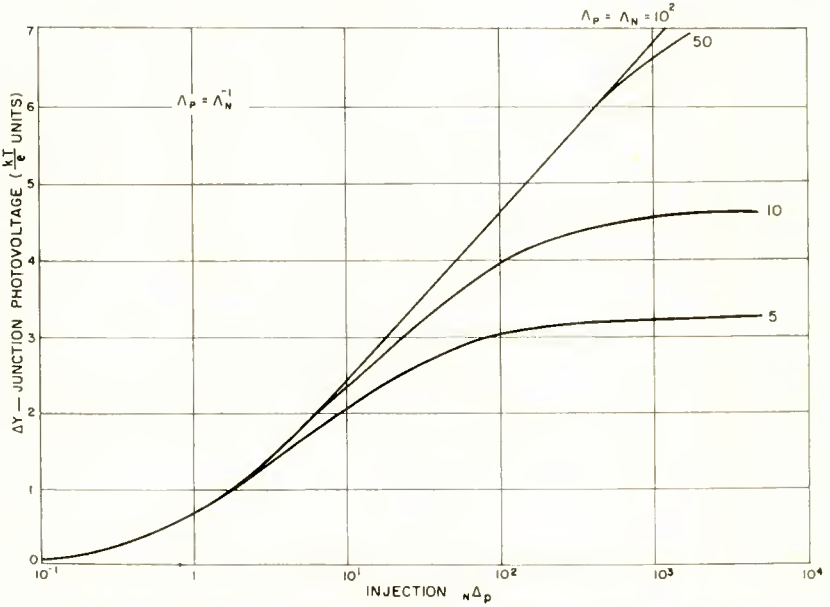



Fig. 10—Junction photovoltage for symmetrical case where $\Lambda_P = \Lambda_N^{-1}$.

diffusion of the most mobile carrier. The value of this potential seen by the external circuit is the algebraic sum of the Dember potentials in the bulk material on each side of the junction. The total effective potential is zero for the symmetrical doping case. For nonsymmetrical doping the Dember potential is close to zero on the heavily doped side and, at most a few $kT/e$ units on the lightly doped side. Thus, the Dember potential can usually be disregarded.

It is usually more practical to use graphical means to determine the division of junction potential for the nonequilibrium case; the analytical approach involves a simultaneous solution of Equations (15) and (21). As pointed out elsewhere,[9] the graphical solution for the nonequilibrium and equilibrium cases are carried out similarly. In the former case, however, the injection-modified space charge and interface

density functions, Equations (2), (22), and (23), are used with each value of injection. In brief, the following changes accompany injection: First, the voltage across each space-charge region decreases, the major change taking place across the lightly doped side; second, the space charge on each side decreases; third, the carrier density at the interface increases; fourth, charge changes in traps in the space-charge region can alter the space-charge curves and effect the division of potential across the two sides of the junction, but there is no effect on the total photovoltage.

Anomalous behavior of the junction photovoltage is experimentally evidenced by photovoltage curves which lie below the theoretical ones and have a shallower slope. Such curves can be fitted to the simple diode equation only if the exponent $e/kT$ in the exponential factor is multiplied by an empirical factor[5] which commonly lies in the range 0.5 to 1.0. Silicon is a far worse offender in this respect than germanium. The preceding analysis can only account for anomalous behavior if injection is considered as changing the effective values of $\Lambda_P$ and $\Lambda_N$ in Equations (21) and (25). Such changes are effectively equivalent to what would happen if there were nonuniform doping and injection-induced changes in the bulk lifetime. The latter would cause changes in the diffusion length in the region where the carrier density is falling away due to bulk recombination. The reasoning leading to this conjecture was discussed earlier. It is also conceivable that the surface near the junction could give rise to similar effects. Surface recombination, for example, can be greatly affected by injection and so could be a potent source of variations in the effective diffusion length of carriers near the surface. These conjectures are mentioned here only as possibilities that should be considered in any detailed treatment of anomalous effects at junctions, particularly departures from the simple diode equation.

## Additional Comments

If account is taken of the variation of the majority carrier density in the usual derivation[4] of the diode equation, an expression comparable to Equation (25) is obtained. For comparison with the ordinary diode equation, this modified expression can be written as

$$I = I_0 [(1 + M) e^{\lambda v} - 1], \tag{30}$$

where the symbols have their usual meanings and $M$ is roughly equal to the fractional increase in the majority carrier density on either side of the junction. At low injection levels $M = 0$ and Equation (30) is identical with the ordinary diode equation.

To a first approximation, the analysis for the nonequilibrium case can be applied to take into account current flow across a junction under high injection level conditions. A junction does not "care," except for the net current flow, whether the net voltage across it stems from a photoeffect or from an externally applied potential. In the open-circuit photovoltage case, hole–electron pairs flow away together from the illuminated junction area and recombine deep in the bulk material. There is no net current flow because the hole and electron currents are equal. The only potential developed in the bulk material is a relatively insignificant Dember potential. Corresponding to a given value of photovoltage at the junction there is a definite excess carrier density. Now, if the light is removed and an external potential, equal in magnitude to the photovoltage, is applied to the junction, the same values of excess carrier densities will be present. This follows because the nonequilibrium Fermi levels will have to have very closely the same relationships with each other and with the applied potential as they did when the photovoltage was present.[12] There is, however, a notable difference. The carriers do not flow away from the junction in pairs; the holes, as minority carriers, flow away from the junction in the $n$ material, and the electrons flow away, as minority carriers, in the $p$ material. The holes in the $n$ material combine with electrons coming from this side of the circuit; the electrons in the $p$ material recombine with holes coming from this side of the circuit.[12] The junction voltage was previously related to the excess carrier density, and now the excess carrier density can be related to the total junction current by using the usual diffusion flow relations. The relation,

$$I = I_0 \left[ (1 + {}_N\Delta_p \Lambda_N{}^2)\ e^{\Delta Y} - 1 \right], \tag{31}$$

results for the most simple case, a $p$-$n$ junction with symmetrical doping. This relation becomes quite complicated when the doping is not symmetrical but, in principle, any type of junction, *LH* included, can be handled by the analysis. Equations (30) and (31) are obviously of the same form. As pointed out by Webster,[13] high injection levels in a transistor cause the diffusion coefficients, which appear here in $I_0$, to increase by a factor ranging between unity and two. This factor takes into account the electric field induced drift of the carriers in the bulk material. Thus, to a first approximation, the effect of high injection on bulk conditions can be taken into account in Equation (31).

[13] W. M. Webster, "On the Variation of Junction-Transistor Current-Amplification Factor with Emitter Current," *Proc. I.R.E.*, Vol. 42, p. 914, June, 1954.

# THEORY OF PARAMETRIC AMPLIFICATION USING NONLINEAR REACTANCES

By

S. Bloom and K. K. N. Chang

RCA Laboratories,
Princeton, N. J.

*Summary—The parametric amplifier is analyzed phenomenologically in terms of an equivalent-circuit model. The model consists of a signal circuit resonant at $\omega_1$, an idling circuit at $\omega_2$, and a pumping circuit at $\omega_3 = \omega_1 + \omega_2$, these three circuits being coupled across a nonlinear inductance. The analysis is general enough to delineate the conditions on the signal level and circuit parameters which lead to distortionless amplification. Expressions are derived for power gain, bandwidth, and noise factor for the case in which the signal and idling frequencies are well separated and for the degenerate case in which these two frequencies are equal.*

## Introduction

THE general power relations for nonlinear reactances have been discussed by Manley and Rowe.[1] In particular, they have shown, under very general conditions, that if a nonlinear reactance is driven by a low-frequency signal source and a single higher-frequency pumping source, then the flow of power at the difference frequency will introduce a negative resistance into the signal circuit. This negative resistance increases with increasing pumping power. With the production of negative resistance in the signal circuit, the possibility exists of signal amplification, or of free oscillation, if this resistance is made large enough.

It has been known for some time that a negative resistance could be produced by such a modulation process.[2-4] Recently, however, Suhl[5] has pointed out the possibility of realizing these effects in the *microwave* region by making use of the nonlinearities associated with ferromagnetic resonance at high power levels. An amplifier and oscillator

---

[1] J. M. Manley and H. E. Rowe, "Some General Properties of Non-Linear Elements," *Proc. I.R.E.*, Vol. 44, p. 904, July, 1956.

[2] R. V. L. Hartley, "Oscillations in Systems with Non-Linear Reactance," *Bell Sys. Tech. Jour.*, Vol. 15, p. 424, July, 1936.

[3] J. M. Manley and E. Peterson, "Negative Resistance Effects in Saturable Reactor Circuits," *Trans. A.I.E.E.*, Vol. 65, p. 870, December, 1946.

[4] V. D. Landon, "The Use of Ferrite-Cored Coils as Converters, Amplifiers and Oscillators," *RCA Review*, Vol. X, p. 387, September, 1949.

[5] H. Suhl, "Proposal for a Ferromagnetic Amplifier in the Microwave Range," *Phys. Rev.*, Vol. 106, p. 384, April, 1957.

based upon Suhl's proposal has been built by Weiss.[6] In the Suhl–Weiss case the nonlinear reactance is essentially an inductance; however, nonlinear capacitances can also be used.

In common with the three-level maser molecular amplifier[7]—to which it bears a superficial resemblance—the parametric amplifier has a rather limited bandwidth. This is due to the use of sharply tuned signal and idling (i.e., difference-frequency) circuits. The sharper the tuning the larger the negative resistance and also the less the wasted power being drained from the pump power supply into unwanted sidebands. However, it is in its noise behavior that the parametric amplifier appears most promising in comparison with electron-beam devices. For, again as with the maser, the ultimate source of noise is the Johnson noise "spontaneously emitted" by the equivalent circuit resistances. Thus, because the parametric amplifier can operate at room temperature or below, the ultimate noise factor may be in the order of 3 decibels or less. Although this is higher than the noise factor of a solid-state maser which *must* operate at very low temperatures, the parametric amplifier surpasses the maser in power-handling capacity.

Because there are many nonlinear systems which might lead to realizations of parametric microwave amplifiers, it is of interest to have a general analysis of this type of amplifier. In this paper the parametric amplifier is analyzed in terms of an equivalent circuit using a nonlinear inductance. (The capacitative case would follow a quite similar analysis.) The method used is general enough to describe not only linear but also nonlinear amplification.

In brief, the analysis proceeds as follows. A signal generator at frequency $\omega_1$ and a pumping power source at frequency $\omega_3$ are impressed across a nonlinear inductance, or "reactor," whose flux density is taken to be a quadratic function of the current. A quadratic function is used rather than a cubic, not only because it is simpler to analyze but also because the cubic (or higher order) case cannot give distortionless amplification.* The modulation process produces many upper and lower

---

[6] M. T. Weiss, "A Solid-State Microwave Amplifier and Oscillator Using Ferrites," *Phys. Rev.*, Vol. 107, p. 317, July, 1957.

[7] N. Bloembergen, "Proposal for a New Type Solid State Maser," *Phys. Rev.*, Vol. 104, p. 324, October, 1956.

* The present analysis is of the conventional parametric amplifier in which only one pumping source is used, the pumping power being at a frequency higher than that of the signal to be amplified. In this case a quadratic nonlinearity must be used in order to have a region of linear amplification. However, two-pump parametric amplifiers are possible and these have the advantage of allowing the signal frequency to exceed the pumping frequencies. In such cases cubic nonlinearities may also be used. A particularly simple version of such a system may have the two pumps degenerate into one pump with a frequency lower than that of the signal.

sideband frequencies, but it is assumed that only the first lower sideband, $\omega_2 = \omega_3 - \omega_1$, can flow, all other sidebands being filtered out. From the nonlinearity relation between flux and current, one can find the voltages at each frequency produced across the reactor by the flow of the total three-frequency current. These modulation voltages, when combined with the internal potential drops and the externally applied voltages, lead to three simultaneous equations relating the impressed voltages to the currents at $\omega_1$, $\omega_2$, and $\omega_3$. From these equations one obtains for the signal circuit at $\omega_1$ the voltage and current relation which describes the amplifier. The nondegenerate case, in which the signal and idling frequencies are well separated, is treated in the next

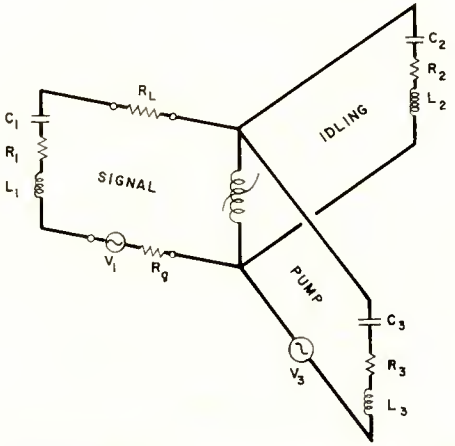



Fig. 1—Equivalent circuit of a parametric amplifier. The signal circuit (resonant at $\omega_1$), idling circuit ($\omega_2$), and pumping circuit ($\omega_3 = \omega_1 + \omega_2$), are coupled across a nonlinear inductance.

section; the degenerate case, in which $\omega_1 = \omega_2$, is treated in the last section.

## Analysis of Nondegenerate Case

Consider the three resonant circuits shown in Figure 1. Each circuit contains a tank resistance $R_i$, inductance $L_i$, and capacitance $C_i$ $(i = 1,2,3)$. In addition, the signal circuit contains a voltage generator $V_1 \cos \omega_1 t$, with internal resistance $R_g$, and a load resistance $R_L$. The energy supply for the system comes from the pumping voltage generator $V_3 \cos \omega_3 t$. These three tank circuits are coupled to a nonlinear inductance, or reactor, the flux linkage through this reactor being taken as a quadratic function of the current, e.g.,

$$\phi = L_0 i(t) - \mathcal{L} i^2(t), \tag{1}$$

where $L_0$ is the linear inductance, $\mathcal{L}$ is the nonlinearity coefficient, and $\mathcal{L}i$ is presumed small compared to $L_0$. The resistance of the reactor is considered to be lumped in with the tank resistances. The resonant frequencies of the three circuits are

$$\omega_i = 1/\sqrt{C_i(L_0 + L_i)} \tag{2}$$

with

$$\omega_3 = \omega_1 + \omega_2. \tag{3}$$

The current through the reactor consists of three frequency components,

$$i(t) = I'_1 \cos(\omega_1 t + \theta_1) + I'_2 \cos(\omega_2 t + \theta_2) + I'_3 \cos(\omega_3 t + \theta_3), \tag{4}$$

where $\theta_1$ and $\theta_3$ are the phase angles between the currents and applied voltages at signal and pumping frequencies, respectively, and where $\theta_2 + \theta_1 - \theta_3$ is the phase angle between the current through and the voltage across the reactor at the idling frequency. The voltage across the reactor is

$$v(t) = d\phi/dt = v_1(t) + v_2(t) + v_3(t),$$

which by Equations (1) and (4) has the three relevant components

$$\begin{aligned} v_1 &= -\omega_1 L_0 I'_1 \sin(\omega_1 t + \theta_1) + \omega_1 \mathcal{L} I'_2 I'_3 \sin(\omega_1 t - \theta_2 + \theta_3), \\ v_2 &= -\omega_2 L_0 I'_2 \sin(\omega_2 t + \theta_2) + \omega_2 \mathcal{L} I'_1 I'_3 \sin(\omega_2 t - \theta_1 + \theta_3), \\ v_3 &= -\omega_3 L_0 I'_3 \sin(\omega_3 t + \theta_3) + \omega_3 \mathcal{L} I'_1 I'_2 \sin(\omega_3 t + \theta_1 + \theta_2), \end{aligned} \tag{5}$$

other frequency components being ignored. Going now to complex notation, the current components are written as

$$I_i = I'_i \exp(j\theta_i).$$

Then the complex impedances due to the reactor, obtained by dividing the voltages of Equation (5) by the corresponding currents of Equation (4), are

$$z_1 = j\omega_1 L_0 - j\omega_1 \mathcal{L} I_3 I_2^* / I_1 ,$$
$$z_2 = j\omega_2 L_0 - j\omega_2 \mathcal{L} I_3 I_1^* / I_2 , \tag{6}$$
$$z_3 = j\omega_3 L_0 - j\omega_3 \mathcal{L} I_1 I_2 / I_3 .$$

Now, the signal frequency in general is $\omega$ and may be detuned from the resonant frequency, $\omega_1$, of the signal circuit. If the signal is detuned then the idling current is also detuned, being at $\omega_3 - \omega$ rather than at $\omega_2$. The pump frequency is assumed to be always at $\omega_3$. Thus the general voltage–current equations for the three circuits are

$$V_1 = I_1 (R_T + jX_1) - j\omega \mathcal{L} I_3 I_2^* ,$$
$$0 = I_2 (R_2 + jX_2) - j(\omega_3 - \omega) \mathcal{L} I_3 I_1^* , \tag{7}$$
$$V_3 = I_3 (R_3 + jX_3) - j\omega_3 \mathcal{L} I_1 I_2 .$$

Equations (7) are correct provided each circuit sees an infinite impedance looking into the other two circuits. That is, it is assumed that the three resonant frequencies are sufficiently separated and the circuit $Q$'s are sufficiently large that the only current flowing in any given circuit is that having a frequency near that circuit's resonance. The degenerate case, in which $\omega_1 = \omega_2$, is treated separately in the last section.

The total positive resistance in the signal circuit is

$$R_T = R_1 + R_g + R_L , \tag{8}$$

and the circuit reactances in the absence of nonlinearity are

$$X_1 = \omega (L_0 + L_1) - \frac{1}{\omega C_1} = R_1 Q_1^{\circ} \left( \frac{\omega}{\omega_1} - \frac{\omega_1}{\omega} \right),$$
$$X_2 = (\omega_3 - \omega) (L_0 + L_2) - \frac{1}{(\omega_3 - \omega) C_2} = R_2 Q_2^{\circ} \left( \frac{\omega_3 - \omega}{\omega_2} - \frac{\omega_2}{\omega_3 - \omega} \right), \tag{9}$$
$$X_3 = \omega_3 (L_0 + L_3) - \frac{1}{\omega_3 C_3} = 0,$$

the $Q^{\circ}$ being the unloaded quality factors.

The last terms on the right of Equations (6) or (7) give the Manley–Rowe[1] power relations. The powers entering the reactor at the three frequencies are

$$p_1 = -\omega_1 |I_1 I_2 I_3| \mathcal{L},$$

$$p_2 = -\omega_2 |I_1 I_2 I_3| \mathcal{L},$$

$$p_3 = \omega_3 |I_1 I_2 I_3| \mathcal{L},$$

or

$$p_3/\omega_3 = -p_1/\omega_1 = -p_2/\omega_2 . \tag{10}$$

Thus, if power enters the reactor at the pump frequency, power leaves the reactor at the signal and idling frequencies.

Upon eliminating the idling current, $I_2$, and the pump current, $I_3$, from the first of Equations (7), one obtains, for the signal circuit,

$$\frac{V_1}{I_1} = R_T - R + j\left(X_1 - \frac{RX_2}{R_2}\right), \tag{11}$$

where the *negative resistance*, $R$, is given by

$$R = \alpha R_T \frac{\dfrac{\omega(\omega_3 - \omega)}{\omega_1 \omega_2}}{\left| 1 + j\dfrac{X_2}{R_2} + \left| \dfrac{I_1}{i_0} \right|^2 \right|^2}, \tag{12}$$

$R$ being a positive number. Here

$$\alpha \equiv \left( \sqrt{\frac{\omega_1 \omega_2}{R_T R_2}} \frac{\mathcal{L} V_3}{R_3} \right)^2 \tag{13}$$

is dimensionless, and

$$i_0 \equiv \sqrt{\frac{R_2 R_3}{\mathcal{L}^2 \omega_3 (\omega_3 - \omega)}} \tag{14}$$

has the dimensions of current. It is to be noted from Equation (13) that if the pump frequency is detuned from $\omega_3$ then $V_3/R_3$ is replaced by $V_3/(R_3 + jX_3)$ and the negative resistance is reduced.

From Equation (12) it is seen that the negative resistance is a decreasing function of the signal current, $I_1$. Thus distortionless amplification requires that

$$\left| \frac{I_1}{i_0} \right|^2 \ll 1, \tag{15}$$

which, by Equation (14), means that the pump and idling resistances should be large, the nonlinearity small, and the signal current small. If in place of Equation (1) a cubic nonlinearity is used together with an idling frequency which is now $\omega_3 - 2\omega_1$, then a similar analysis shows that the negative resistance is always a function of the signal current and thus results in distortion.

An alternative method of analyzing the parametric amplifier is to take the nonlinear inductance to vary at the pumping frequency, e.g.,

$$L(t) = L_0 + \tilde{L} \sin \omega_3 t,$$

and to use the circuit of Figure 1 in which the signal and idling circuits are coupled across $L(t)$, and in which the pumping circuit no longer enters explicitly. The quantity $\tilde{L}$ is then found to be related to the $\mathcal{L}$ of Equation (1) by

$$\tilde{L} = 2\mathcal{L}I_3 .$$

Hence $\tilde{L}$ is not a constant, but depends upon the pump current. Thus the "variable parameter" method using $\tilde{L}$ is of limited utility; it can give no information regarding nonlinear amplification in which the dependence of the pump current upon the signal current, $I_1$, causes distortion.

*Power Gain*

The power gain definition to be used is

$$G = \frac{P_{\text{out}}}{\hat{P}_{\text{in}}},$$

where the power output across the load resistance is

$$P_{\text{out}} = R_L |I_1|^2 = |V_1|^2 \frac{R_L}{(R_T - R)^2 + (X_1 - RX_2/R_2)^2} .$$

Here $\hat{P}_{\text{in}}$ is the power which would flow into the amplifier if the input were matched;

$$\hat{P}_{\text{in}} = |V_1|^2/4R_g ,$$

which is the maximum available input power. Thus the available power gain is

$$G = \frac{4R_g R_L}{(R_T - R)^2 + (X_1 - RX_2/R_2)^2}. \tag{16}$$

On resonance ($\omega_1 = \omega$), the gain has its maximum value

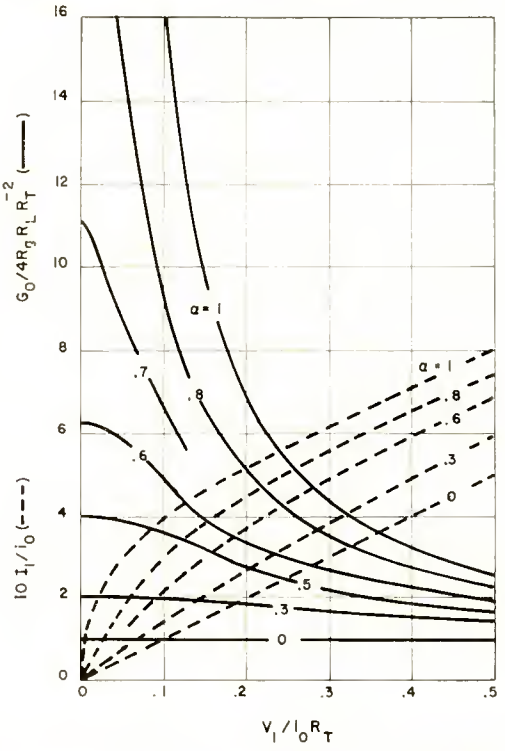



Fig. 2—The normalized power gain on resonance, $G_0/(4R_gR_LR_T^{-2})$, and the normalized current in the signal circuit, $I_1/i_0$, as functions of the normalized signal voltage, $V_1/i_0R_T$, for various values of the parameter $\alpha \equiv (\sqrt{\omega_1\omega_2/R_TR_2}\,\mathcal{L}\,V_3/R_3)^2$. The amplifier is distortionless in the region of small $V_1/i_0R_T$, where the gain curves are flat and the current curves are linear.

$$G_0 = \frac{4R_gR_LR_T^{-2}}{[1 - \alpha/(1 + |I_1/i_0|^2)]^2}. \tag{17}$$

The normalized power gain on resonance, $G_0/(4R_gR_LR_T^{-2})$, and the normalized signal current, $I_1/i_0$, are shown in Figure 2 as functions

of the normalized signal voltage, $V_1/(i_0R_T)$, for various values of the parameter $\alpha$. When this parameter is zero there is no negative resistance, and when it is unity the gain is infinite in the absence of $V_1$, i.e., the signal (and idling) circuit freely oscillates. The dashed curves in Figure 2 show the signal current, the full curves show the gain. Linear amplification occurs in the region of small $V_1/(i_0R_T)$ where the gain curves are essentially flat.

*Bandwidth*

The frequency bandwidth for linear amplification is found by considering Equations (12) and (16) with the distorting term $I_1^2$ neglected. Since $\omega(\omega_3 - \omega)/(\omega_1\omega_2)$ is very nearly unity within the bandwidth, Equation 12 reduces to

$$R = \frac{\alpha R_T}{1 + (X_2/R_2)^2}.$$

The half-power frequencies are the roots of $G(\omega) = G_0/2$, or

$$2(1-\alpha)^2 = \left[1 - \frac{\alpha}{1 + (X_2/R_2)^2}\right]^2 + \left[\frac{X_1}{R_T} - \frac{\alpha}{1 + (X_2/R_2)^2}\frac{X_2}{R_2}\right]^2. \tag{18}$$

Since the half-power frequencies are near the resonant frequency, $\omega_1$, the reactances of Equation (9) may be approximated by

$$\begin{aligned} X_1/R_1 &\simeq 2Q^\circ{}_1(\omega - \omega_1)/\omega_1 \equiv (R_T/R_1)\, s \\ X_2/R_2 &\simeq 2Q^\circ{}_2(\omega_1 - \omega)/\omega_2 = -sc, \end{aligned} \tag{19}$$

where

$$c \equiv \omega_1 Q_2/(\omega_2 Q_1). \tag{20}$$

In Equation (20) the loaded $Q$'s are $Q_1 = R_1Q^\circ{}_1/R_T$ and $Q_2 = Q^\circ{}_2$. Equation (18) thus becomes, in terms of $s \equiv X_1/R_T$,

$$\begin{aligned} s^6c^4 + s^4c^2\,[2 + c^2 + 2\alpha c - 2c^2(1-\alpha)^2] + s^2\,[1 + 2c^2(1-\alpha)(2\alpha - 1) \\ + 2\alpha c + \alpha^2c^2] - (1-\alpha)^2 = 0. \end{aligned} \tag{21}$$

The two real roots of Equation (21) are of equal magnitudes and opposite signs. Thus, to within the accuracy of Equations (19), the

gain-versus-frequency curves are symmetrical about the peak value at $\omega = \omega_1$. The roots of Equation (21) are shown in Figure 3 as a function of $\alpha$, for various values of $c$. Figure 3 also shows the normalized gain at resonance with distortion neglected:

$$G_0/(4R_gR_LR_T^{-2}) = (1-\alpha)^{-2}. \tag{22}$$

In the region of normalized-gain values of 10 or greater ($\alpha \gtrsim .7$), the bandwidth curves can be approximated by $s \approx (1-\alpha)/(1+c)$. This gives a full percentage bandwidth of $B \equiv 2\,|\omega - \omega_1|/\omega_1 \approx (1-\alpha)/(Q_1 + Q_2\omega_1/\omega_2)$. Since the voltage gain on resonance is

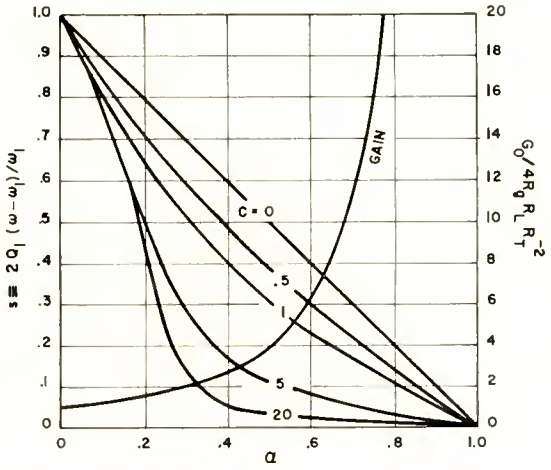



Fig. 3—Nondegenerate case, with distortion neglected; the normalized power gain on resonance, and the bandwidth function, as functions of the parameter $\alpha$ for various values of $c \equiv \omega_1 Q_2/\omega_2 Q_1$.

$$G_v = R_L(I_1/V_1)_{\omega_1} = (R_L/R_T)/(1-\alpha),$$

then the product of bandwidth and voltage gain is

$$G_vB \approx R_L\omega_2/(R_T\omega_1Q_2),$$

where $Q_1$ is taken to be small compared to $Q_2\omega_1/\omega_2$ because of the loading of the signal circuit.

*Noise Factor*

The noise factor is defined as the input signal-to-noise ratio divided by the output signal-to-noise ratio,

$$F = (\hat{P}_{in}/N_{in})/(P_{out}/N_{out}), \tag{23}$$

where the input noise power is taken to be the maximum available thermal noise power delivered from a resistance at room temperature, e.g.,

$$N_{in} = k\,T_0\,\Delta f, \qquad (T_0 = 290^\circ\ \mathrm{K}). \tag{24}$$

Also, $\hat{P}_{in}$ is taken to be the signal input power under matched conditions; hence the available power gain is $G = P_{out}/\hat{P}_{in}$ and therefore

$$F = N_{out}/(k\,T_0\,\Delta f\,G). \tag{25}$$

Now, the *ultimate* noise in the parametric amplifier is the Nyquist–Johnson thermal noise in the circuit resistances.* To compute the noise output power, $N_{out}$, due to the resistances, one returns to the general circuit equations (7), replacing signal voltages by noise voltages. Thus Equations (7) become

$$\begin{aligned} \mathcal{E}_1 &= I_1\,(R_T + jX_1) - j\omega\mathcal{L}\,I_3\,I_2^* \\ \mathcal{E}_2 &= I_2\,(R_2 + jX_2) - j\,(\omega_3 - \omega)\,\mathcal{L}\,I_3\,I_1^* \\ V_3 &= I_3\,(R_3 + jX_3) - j\omega_3\mathcal{L}\,I_1\,I_2\,, \end{aligned} \tag{26}$$

where the noise voltages in the signal and idling circuits are, respectively,

$$\begin{aligned} |\mathcal{E}_1|^2 &= 4k\Delta f\,(R_g T_0 + R_1 T), \\ |\mathcal{E}_2|^2 &= 4k\Delta f R_2 T. \end{aligned} \tag{27}$$

The internal resistance of the signal generator, $R_g$, is at the reference temperature, $T_0$, and the amplifier resistances $R_1$ and $R_2$ are at the ambient temperature $T$. In the third of Equations (26), $V_3$ should actually be written as $V_3 + \mathcal{E}_3$, but because the pump voltage is always much larger than any noise voltage arising from $R_3$, this correction may be neglected. Furthermore, the use of $\mathcal{E}_1$ and $\mathcal{E}_2$ in Equations (26) is legitimate, even though these are noise quantities, because these two noise voltages are uncorrelated. Finally, the bandwidths appearing in Equations (27) are actually $\Delta f_1$ and $\Delta f_2$, respectively; however

* A more general noise theory has been worked out by H. Heffner and G. Wade of Stanford University, in which account is also taken of fluctuations in the pumping source.

$$\Delta f_1 = |\omega_1 - \omega| = |\omega_3 - \omega - \omega_2| = \Delta f_2 .$$

Since one is interested in the noise behavior when the amplifier is linear, the distorting term in $I_1 I_2$ in the third of Equations (26) may be dropped. Solving the second and third equations for $I_2$ and $I_3$ and substituting in the first of Equations (26), one obtains

$$\mathcal{E}_1 - \frac{\omega \mathcal{L} V_3}{R_3 (R_2 - jX_2)} \mathcal{E}_2^* = I_1 \left[ R_T + jX_1 - \frac{\omega(\omega_3 - \omega)\ \mathcal{L}^2 V_3^2}{(R_2 - jX_2)\ R_3^2} \right],$$

where $X_3$ is zero since the pump is presumed to be always at $\omega_3$. Squaring the above equation, taking $\mathcal{E}_1 \mathcal{E}_2^*$ to be zero since $\mathcal{E}_1$ and $\mathcal{E}_2$ are uncorrelated, one gets

$$|\mathcal{E}_1|^2 + \frac{\omega}{\omega_3 - \omega} \frac{R}{R_2} |\mathcal{E}_2|^2 = |I_1|^2 [(R_T - R)^2 + (X_1 - RX_2/R_2)^2], \quad (28)$$

where $R$ is the negative resistance of Equation (12) in the absence of distortion.

The noise power output of the amplifier is

$$N_{\text{out}} = R_L |I_1|^2.$$

Making use of the power gain expression of Equation (16), one thus obtains from Equation (28)

$$N_{\text{out}} = \frac{G}{4R_g} \left[ |\mathcal{E}_1|^2 + \frac{\omega}{\omega_3 - \omega} \frac{R}{R_2} |\mathcal{E}_2|^2 \right] \quad (29)$$

Hence, by Equations (25) and (27), the ultimate noise factor is

$$F = 1 + \frac{T}{T_0} \left( \frac{R_1}{R_g} + \frac{\omega}{\omega_3 - \omega} \frac{R}{R_g} \right) \quad (30)$$

which, at resonance, becomes

$$F = 1 + \frac{T}{T_0} \left( \frac{R_1}{R_g} + \frac{\omega_1}{\omega_2} \frac{\alpha R_T}{R_g} \right) \quad (31)$$

It is seen from Equation (29) that the negative resistance contributes

noise to the signal circuit because of noise current flowing in the idling circuit. Typically, the term in parentheses in Equation (31) is of order unity, thus giving a noise factor in the order of 3 decibels at room temperature. Refrigeration, and the use of small ratios $\omega_1/\omega_2$, will of course lower this value.

## Analysis of Degenerate Case

In the preceding section it was assumed that the signal and idling frequencies were well separated. This is no longer true in the degenerate case; then $\omega_1 = \omega_2 = \omega_3/2$ and a new analysis is necessary.

In contrast to Equation (4), the current through the reactor now has only two frequency components

$$i(t) = I'_1 \cos(\omega_1 t + \theta_1) + I'_3 \cos(\omega_3 t + \theta_3). \tag{32}$$

The two components of voltage across the reactor thus become

$$v_1 = -\omega_1 L_0 I'_1 \sin(\omega_1 t + \theta_1) + \omega_1 \mathcal{L} I'_1 I'_3 \sin(\omega_1 t + \theta_3 - \theta_1).$$

$$v_3 = -\omega_3 L_0 I'_3 \sin(\omega_3 t + \theta_3) + (1/2)\,\omega_3 \mathcal{L} I'^2_1 \sin(\omega_3 t + 2\theta_1).$$

The complete voltage–current equations in complex notation are then

$$V_1 = I_1(R_T + jX_1) - j\omega \mathcal{L} I_3 I_1^*$$

$$V_3 = I_3(R_3 + jX_3) - j\omega_3 \mathcal{L} I_1^2/2. \tag{33}$$

The $I_1^2$ term in Equations (33) causes distortion and will be neglected in this small signal analysis. Because $V_1$ and $V_3$ are real, Equations (33) result, after some algebra, in the following voltage–current relation for the signal circuit:

$$V_1/I_1 = (R_T - R)\,(1 + j\tan\theta_1), \tag{34}$$

where

$$-\tan\theta_1 = \frac{X_1/R_T - \sqrt{\beta}/(1+\tan^2\theta_3)}{1 - (\sqrt{\beta}\sin 2\theta_3)/2}$$

$$-\tan\beta_3 = X_3/R_3.$$

Here the negative resistance is

$$R = (\sqrt{\beta}\,R_T/2)\,[\sin 2\theta_1 + \sin(2\theta_1 - 2\theta_3)],$$

where
$$\beta \equiv \left[\frac{\omega \mathcal{L} V_3}{R_T R_3}\right]^2 \doteq \left[\frac{\omega_1 \mathcal{L} V_3}{R_T R_3}\right]^2 \tag{35}$$

The parameter $\beta$, which is a measure of the magnitude of the negative resistance, is the counterpart of the $\alpha$ of Equation (13).

In practice, the pump frequency will be at resonance ($\omega_3$); thus $X_3 = 0 = \theta_3$, so that the negative resistance and the signal phase angle become

$$R = \sqrt{\beta}\, R_T \sin 2\theta_1$$
$$\tan \theta_1 = \sqrt{\beta} - X_1/R_T. \tag{36}$$

*Power Gain*

The power gain is

$$G = \frac{R_L |I_1|^2}{V_1^2/4R_g}$$

or, by Equation (34),

$$G = \frac{4R_g R_L R_T^{-2}}{(1 - \sqrt{\beta} \sin 2\theta_1)^2 (1 + \tan^2 \theta_1)}.$$

The phase angle, $\theta_1$, can be expressed in terms of the signal-circuit reactance, $X_1$, by Equation (36); thus the gain expression becomes

$$G = 4R_g R_L R_T^{-2} \frac{1 + (\sqrt{\beta} - X_1/R_T)^2}{[1 + (\sqrt{\beta} - X_1/R_T)^2 - 2\sqrt{\beta}\,(\sqrt{\beta} - X_1/R_T)]^2}, \tag{37}$$

where the reactance is given in terms of the loaded $Q$ by

$$X_1/R_T = Q_1(\omega/\omega_1 - \omega_1/\omega).$$

For signal frequencies not too far removed from the resonant frequency, $\omega_1$, the reactance can be approximated by

$$X_1/R_T \simeq 2Q_1(\omega - \omega_1)/\omega_1 \equiv s. \tag{38}$$

When the signal frequency, $\omega$, is at signal-circuit resonance, $\omega_1$, the reactance $X_1$ is zero. This however is not quite the condition for maximum gain, as can be shown by a differentiation of Equation (37)

with respect to $X_1$. Rather, the signal frequency giving maximum gain lies slightly below $\omega_1$. The frequency of maximum gain is shown in Figure 4 as a function of $\beta$.

However, the true maximum gain exceed the gain at resonance by only a few per cent. The maximum gain will therefore be taken as the gain at resonance; by Equation (37) with $X_1 = 0$, it is

$$G_0 = 4R_gR_LR_T^{-2}\,(1+\beta)/(1-\beta)^2, \tag{39}$$

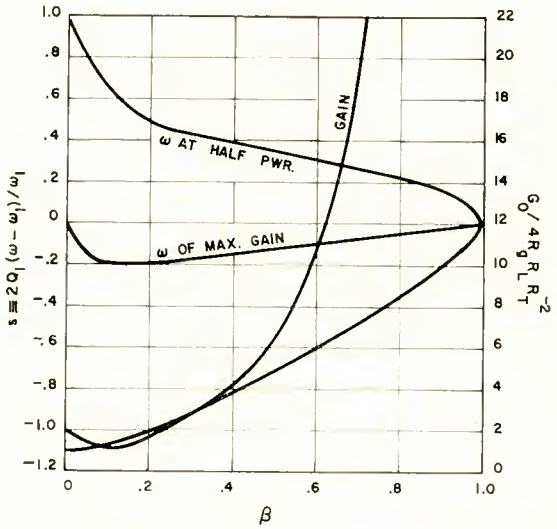



Fig. 4—Degenerate case ($\omega_1 = \omega_2$), with distortion neglected. The frequency of maximum gain, the normalized gain on resonance, and the bandwidth function, as functions of the parameter $\beta \equiv (\omega_1 \mathcal{L} V_3/R_TR_3)^2$.

which is the counterpart of Equation 22. The normalized gain on resonance is shown in Figure 4.

*Bandwidth*

The half-power bandwidth is found from the solution of $G(\omega) = G_0/2$, or by Equations (38) and (39);

$$\frac{1+\beta}{2(1-\beta)^2} = \frac{1+(\sqrt{\beta}-s)^2}{[1+(\sqrt{\beta}-s)^2+2\sqrt{\beta}\,(\sqrt{\beta}-s)\,]^2}.$$

This can be expanded to read

$$s^4 + s^2 [4\beta(1-\beta)/(1+\beta)] + s [4\sqrt{\beta}\,(1-\beta)^2/(1+\beta)] - (1-\beta)^2 = 0. \tag{40}$$

The roots of Equation (40) give the half-power frequencies; they are shown in Figure 4. Whereas the nondegenerate case gave symmetrical gain curves, it is seen from Figure 4 that the gain curves in the degenerate case are asymmetric, the lower half-power frequencies lying farther away from $\omega = \omega_1$ than do the upper half-power frequencies.

*Noise Factor*

The ultimate noise factor in the degenerate case can be written down at once; it is simply

$$F = 1 + \frac{T R_1}{T_0 R_g}. \tag{41}$$

Because there is no separate idling circuit to contribute noise through the negative resistance into the signal circuit, the negative resistance is "noiseless"; $R$ does not enter Equation (41) as it does Equation (30).

The degenerate case, then, is the more attractive mode of operation. It has a lower noise factor than does the nondegenerate case, it has a bandwidth greater by about a factor of 2 for the same gain, and it requires only two resonators rather than three.

# RCA TECHNICAL PAPERS†

## Third Quarter, 1957

**Any request for copies of papers listed herein should be addressed to the publication to which credited.**

"Activation of an Oxide Cathode by Deposition of Alkaline Earth Metal Ions via a Mass Spectrometer," R. M. Matheson, L. S. Nergaard and R. H. Plumlee, *RCA Review* (September) .... 1957

"Alignment of TV Chassis Using Marker-Adder Techniques," R. Samuel, *Service* (July) ..... 1957

"Analysis of Nuclear Radiation Effects on Transistors," D. B. Kret, *Electronic Design* (July 15) ..... 1957

"Antenna-Multiplex System Design," H. K. Schlegelmilch, O. K. Nilssen, and W. Y. Pan, *Electronics* (July) ..... 1957

"Application of the Ising-Bethe Theory to the Susceptibilities of the Polymorphic Forms of MnS," P. J. Wojtowicz, *Phys. Rev.* (July 15) ..... 1957

"Bizmac," A. L. Malcarney, *Signal* (September) ..... 1957

"A Circularly-Polarized Corner Reflector Antenna," O. M. Woodward, Jr., *Trans. I.R.E. PGAP* (July) ..... 1957

"Class A Transistor Power Amplifier Design," R. Minton, *Electronic Design* (July 15) ..... 1957

"Class B Transistor Power Amplifier Design," R. Minton, *Electronic Design* (September 15) ..... 1957

"Corner Reflector Antennas with Arbitrary Dipole Orientation and Apex Angle," R. W. Klopfenstein, *Trans. I.R.E. PGAP* (July) 1957

"Effect on Edge Dislocations on the Alloying of Indium to Germanium," J. I. Pankove, *Jour. Appl. Phys.* (September) .... 1957

"Effects of Radiation Damping on Spin Dynamics," S. Bloom, *Jour. Appl. Phys.* (July) ..... 1957

"Effects of Voltage-Divider Characteristics on Multiplier Phototube Response," R. W. Engstrom and E. Fischer, *Rev. Sci. Instr.* (July) ..... 1957

"Electron Beam Focusing in Three-Anode Guns for Traveling-Wave Tubes," T. S. Chen and L. Kovack, *Journal of Electronics and Control* (September) ..... 1957

"The Equivalent Circuit of the Drift Transistor," J. Almond and R. J. McIntyre, *RCA Review* (September) ..... 1957

"Exponential Transmission Lines as Resonators and Transformers," R. N. Ghose, *Trans. I.R.E. PGMTT* (July) ..... 1957

"Ferrite Transducers for Electromechanical Filters," G. S. Hipskind, *Electronic Design* (September 1) ..... 1957

"Get More from Your Volt Ohmyst," R. Samuel, *RCA Rad. and Tele. Serv. News* (August) ..... 1957

"High-Voltage Regulator Tubes for Color Television Receivers," R. E. Byram, *Trans. I.R.E. PGED* (July) ..... 1957

"How to Get High Quality Performance from the TK-26 3-Vidicon Film Chain," T. J. Shipferling and H. N. Kozanowski, *Broadcast News* (August) ..... 1957

"An Image-Converter Tube for High-Speed Photographic Shutter Service," R. G. Stoudenheimer and J. C. Moor, *RCA Review* (September) ..... 1957

† Report all corrections or additions to *RCA Review*, RCA Laboratories, Princeton, N. J.

"Junction Capacitance & Related Characteristics Using Graded Impurity Semiconductors," L. J. Giacoletto, *Trans. I.R.E. PGED* (July) ..... 1957

"Magnetic Field Dependence of the Seebeck Effect in Germanium," M. C. Steele, *Phys. Rev.* (July 1) ..... 1957

"On the Measurement of Component Reliability," I. K. Munson, *Trans. I.R.E. PGRQC* (July) ..... 1957

"Modern Theater Sound-Service Procedures," E. Stanko, *Jour. S.M.P.T.E.* (September) ..... 1957

"Molecular Amplification and Generation of Microwaves," J. P. Wittke, *Proc. I.R.E.* (Letter to the Editor) (July) ..... 1957

"Napierian Logarithms," M. S. Corrington, *Proc. I.R.E.* (September) (Letter to the Editor) ..... 1957

"N-Halogen Organic Compounds as Cathode Materials for Primary Batteries," C. K. Morehouse and R. Glicksman, *Jour. Electrochem. Soc.* (August) ..... 1957

"Noise-Level Reductions of Barriers," M. Rettinger, *Jour. S.M.P.T.E.* (July) ..... 1957

"Noise Wave Excitation at the Cathode of a Microwave Beam Amplifier," W. R. Beam, *Trans. I.R.E. PGED* (July) ..... 1957

"Operation and Performance of the 6866 Display Storage Tube," E. M. Smith, *RCA Review* (September) ..... 1957

"Oxygen Sorption Phenomena on Cadmium Selenide Crystals," R. H. Bube, *Jour. Chem. Phys.* (August) ..... 1957

"Phonetic Typewriter," H. F. Olson and H. Belar, *Trans. I.R.E. PGA* (July-August) ..... 1957

"Photocell Measures Light Direction," J. T. Wallmark, *Electronics* (July) ..... 1957

"Photoleitfähigkeit von Zinkoxyd bei Ohmschen und sperrenden Kontakten," H. J. Gerritsen, W. Ruppel, and A. Rose, *Sonderabdruck Aus Helvetica Physica Acta* (August) ..... 1957

"Proposal for a Nuclear Quadrupole Maser," R. Braunstein, *Phys. Rev.* (Letter to the Editor) (August 15) ..... 1957

"Quasi-Electric and Quasi-Magnetic Fields in Nonuniform Semiconductors," H. Kroemer, *RCA Review* (September) ..... 1957

"Schwärzung von ZnS und CdS-Einkristallen durch Licht," W. J. Merz, *Sonderabdruck Aus Helvetica Physica Acta* (August). 1957

"Single Sideband Receivers," H. F. Comfort, *Signal* (July) ..... 1957

"A Thin-Window Cathode-Ray Tube for High-Speed Printing with 'Electrofax'," R. Olden, *RCA Review* (September) ..... 1957

"A Transistor Marker Beacon Receiver," R. G. Erdman, *Trans. I.R.E. PGANE* (September) ..... 1957

"Transistor Receiver Video Amplifiers," M. C. Kidd, *RCA Review* (September) ..... 1957

"A Transistorized Horizontal-Deflection System," H. C. Goodrich, *RCA Review* (September) ..... 1957

"Troubleshooting Color Circuits," J. R. Meagher, *RCA Rad. and Tele. Serv. News* (July) ..... 1957

"TV Horizontal Retrace: Problems and Sclutions," S. Wlasuk, *Service* (September) ..... 1957

"Ultraviolet TV Provides Data on Cancer Cells," *Broadcast News* (August) ..... 1957

"Use of Lithium in Photoemissive Cathodes," A. H. Sommer, *Rev. Sci. Instr.* (Notes) (August) ..... 1957

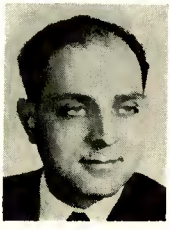
"Versuch einer chemischen Deutung der Energieniveaux in Festkorpern," W. Ruppel, A. Rose, and H. J. Gerritsen, *Sonderabdruck Aus Helvetica Physica Acta* (August) ..... 1957

"Voltage Measurements with an Oscilloscope," R. Samuel, *Electronic Technician* (September) ..... 1957

# AUTHORS

STANLEY BLOOM received the B.S. degree in Physics and Mathematics from Rutgers University in 1948. In 1949 he received the M.S. degree in Physics and in 1952, the Ph.D. degree in Physics from Yale University. In 1952 he joined the RCA Laboratories where he has engaged in research on noise problems in traveling-wave tubes, magnetic resonance, and parametric amplifiers. Dr. Bloom is a Member of the American Physical Society, Sigma Xi, and Phi Beta Kappa.

K. K. N. CHANG received the B.S. degree from the National Central University, Nanking, China, in 1940; the M.S. degree in Electrical Engineering from the University of Michigan in 1948, and the D.E.E. degree in 1954 from the Polytechnic Institute of Brooklyn. From 1940 to 1945, he was associated with the Central Radio Manufacturing Works, Kunming, China, working on radio receivers and from 1945 to 1947, he was a radio instructor in the Office of Strategic Service, U. S. Army, China Theatre. Since 1948, he has been at RCA Laboratories, Princeton, N. J., where he has engaged in research on magnetrons, traveling-wave tubes, beam-focusing devices, and parametric amplifiers. Dr. Chang is a member of Sigma Xi.



JACQUES DUTKA received the B.S. degree from the City College of New York, and the A.M. and Ph.D. degrees from Columbia University. From 1946 to 1953 he taught mathematics at Rutgers and Princeton Universities, and from 1953 to 1956 he was a mathematician with the Norden Laboratories. Since 1956 he has been a Senior Engineer with the Surface Communications Laboratory of RCA Defense Electronic Products in New York City. In addition he has, since 1954, been an Adjunct Associate Professor in the Electrical Engineering Department of Columbia University. Dr. Dutka is a member of Sigma Xi, the American Mathematical Society, and the Institute of Mathematical Statistics.

B. GOLDSTEIN (see *RCA Review*, Vol. XVIII, No. 2, June 1957, p. 288.)

EDWARD O. JOHNSON received the B.S. degree in Electrical Engineering from Pratt Institute of Brooklyn in 1948. From 1941-1945 he served as an electronic technician in the U. S. Navy. He has subsequently studied at Princeton University (Electrical Engineering) and at the Swiss Federal Institute of Technology (Physics). From 1948 to 1954 he was engaged in research at RCA Laboratories, Princeton, in the gaseous conduction field. After his return from a year's leave of absence at Zurich he pursued research in the field of solid-state surfaces. Mr. Johnson is a member of the Institute of Radio Engineers, Sigma Xi, and Tau Beta Pi. He received honorable mention in the Eta Kappa Nu recognition of 1952 and was co-recipient of the 1953 I.R.E. Editor's Award.

RONALD R. JOHNSON received the B.S. degree in Electrical Engineering from West Virginia University in 1952. Prior to graduation he worked for two years as a chemical analyst engaged in quantitative and qualitative analysis for the DuPont Company. He joined the RCA Tube Division in June, 1952, and was assigned to the semiconductor design activity where he has worked on various transistor projects including transistor surface treatments and reliability. Mr. Johnson is a member of Eta Kappa Nu.

E. O. KEIZER received the B.S. degree in Electrical Engineering from Iowa State College, Ames, Iowa in 1940. He then joined the RCA Victor Division, Camden, N.J. In 1942 he was transferred to the RCA Laboratories, Princeton, N. J., where he is now supervisor of the receiver and circuit applications group. Since 1942 he has been engaged in work on receiver circuits, television receiver circuits and on related problems. Mr. Keizer is a Senior Member of the I.R.E. and a Member of Tau Beta Pi, Phi Kappa Phi, Eta Kappa Nu, and Sigma Xi.

L. S. NERGAARD (see *RCA Review*, Vol. XVIII, No. 3, September 1957, p. 436.)

J. T. WALLMARK (see *RCA Review*, Vol. XVIII, No. 2, June 1957, p. 290.)

JAMES P. WITTKE received the degree of Mechanical Engineer from Stevens Institute of Technology in 1949. After a year and a half spent studying hydrodynamic forms as a Fellow at the Experimental Towing Tank in Hoboken, New Jersey, he entered the graduate school of Princeton University. Here he studied in the physics department, receiving the M.A. degree in 1952, and the Ph.D. degree in Physics in 1955. He taught for one year as instructor in physics at Princeton before joining the Technical Staff of the RCA Laboratories in Princeton. Since coming to RCA Laboratories he has been engaged in an experimental and theoretical study of ways to achieve low-noise microwave amplification using molecular processes. Dr. Wittke is a member of the American Physical Society, Sigma Xi, and Tau Beta Pi.

# RCA REVIEW

*a technical journal*

RADIO AND ELECTRONICS
RESEARCH • ENGINEERING

# INDEX

VOLUME XVIII

TABLE OF CONTENTS

**March**



**June**

## AUTHORS, VOLUME XVIII