

The Bell System Technical Journal

Vol. XXVIII

July, 1949

No. 3

Editorial Note regarding Semiconductors

ALL but one of the papers that comprise this issue discuss practical applications of semiconductors and touch upon their properties as employed in rectifying devices, detectors, and in a new amplifying unit—the so-called transistor. These semiconductor papers all relate to one another and present, as a whole, a current but well developed account of the behavior and uses of these very promising additions to today's vast array of electrical appliances.

Because semiconductors are relative newcomers, few engineers have as yet had occasion to become familiar with their characteristics and the reasons for their somewhat unexpected performance. Accordingly, it seems appropriate to preface the present group of papers with a brief introductory note devoted to the nature of the physical phenomena encountered.

The semiconductors of interest in the communications art are electronic rather than ionic conductors, and include copper oxide, various other oxides, selenium, germanium and silicon. Being electronic conductors, the constituent atoms remain in fixed positions. They may lose or gain electrons during the conduction process but the structure of the conductor as a whole and its chemical composition are not affected.

Basic to the theory of these semiconductors is the idea that electrons can carry current in two distinguishable and distinctly different ways: one being called "excess conduction," "conduction by excess electrons," or simply "conduction by electrons;" and the other being called "deficit conduction" or "conduction by holes." The possibility that these two processes may be simultaneously and separably active in a semiconductor affords a basis for explaining transistor action.

We shall confine our attention to the behavior of electrons within the silicon and germanium type of crystal lattice, and especially as it is modified by minute amounts of suitably chosen impurities.¹

¹ There has been very marked development in the understanding of semiconductors since 1940. This understanding is an outgrowth of the research and development program on crystal rectifiers undertaken in connection with the radar program during the war and continued in several laboratories thereafter. Some of the wartime work was carried out in the Radiation Laboratory of M.I.T., which operated under the supervision of the National Defense Research Committee. The Radiation Laboratories Series volume "Crystal Rectifiers" by H. C. Torrey and C. A. Whitmer reports this program and mentions in particular as chief contributors to crystal research and development in England: the General Electric

Silicon and germanium form what are called "covalent crystals," the atoms being held together by "electron-pair bonds" formed by the valence electrons. The covalent bond in the hydrogen molecule is the simplest electron-pair bond. Figure 1 represents two hydrogen atoms and a hydrogen molecule.² Each atom consists of a proton and one electron. The proton weighs approximately 2,000 times as much as the electron and is a relatively immobile particle about which the electron moves in its orbit or quantum mechanical wave function. In an isolated atom, this wave function has spherical symmetry and the electronic charge is distributed on the average as a diffuse sphere centered about the proton. When the two atoms are brought close together, interaction between the wave functions of the two electrons takes place and the electronic cloud of each becomes modified, as suggested in the figure. The result is to produce an extra accumulation of charge between the two protons which acts to bind them together. According to quantum mechanical laws associated with the "Pauli exclusion principle," the bond is especially stable when it contains precisely two electrons. It is weakened considerably by removal of one electron and is not greatly strengthened by the addition of a third electron. This special stability of the electron-pair bond or covalent bond is a fundamental fact of chemistry which is now quite well understood on the basis of wave mechanics.

The elements carbon, silicon and germanium have the common feature of being tetravalent. Although they possess respectively 6, 14 and 32 electrons all together, in each case only four of these are able to enter into chemical reactions. The remaining electrons are closely bound to the nucleus producing a stable "ionic core" having a net charge of +4 units. This core can be regarded as completely inactive so far as electronic processes in chemical reactions and in semiconductors are concerned.

Each of these atoms tends to form covalent or electron-pair bonds with four other atoms. This tendency is completely satisfied in the diamond lattice which is the crystalline form of all three elements. The lattice, shown in Fig. 2, is a cubic arrangement and may be regarded as made up of four interpenetrating simple cubic lattices like the one formed by the atoms on the four corners of the cube shown. In this structure each typical atom is surrounded by four neighbors regularly placed about it, with which it forms four

Company, British Thompson-Houston Ltd., Telecommunications Research Establishment and Oxford University; and in the United States: the Bell Telephone Laboratories, Westinghouse Research Laboratory, General Electric Company, Sylvania Electric Products, Inc., and the E.I. duPont deNemours and Company. It is also pointed out that the crystal groups at the University of Pennsylvania and Purdue University, who operated under N.D.R.C. contracts, were responsible for much fundamental work.

²The figures in this introduction and the text associated with them, like the following paper on "Hole Injection in Germanium", follow closely the presentation in a book entitled "Holes and Electrons, an Introduction to the Physics of Transistors" now under preparation by W. Shockley.

covalent bonds. These neighbors are arranged on the corners of a regular tetrahedron in conformity with the known chemical behavior of the tetrahedral carbon atom.³ For purposes of discussion of conductivity in these crystals, we shall represent the three-dimensional array in two dimensions as is shown in Fig. 3, indicating that each carbon atom forms an electron-pair bond with four neighbors.

On the basis of this valence bond structure we can intuitively see why diamond should be an insulator. Although it contains a large number of

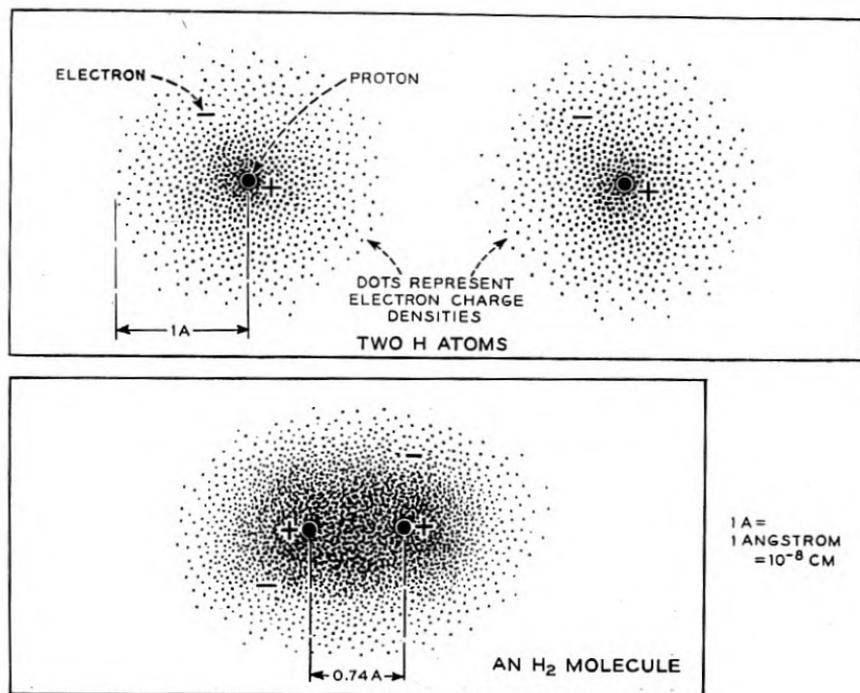


FIG. 1.

electrons, as does a metal, the covalent bond is a quite different structure from the metallic bond. In an ideally perfect diamond crystal, each valence bond would contain its two electrons; therefore, every electron would be tightly bound and thus unable to enter into the conduction process.

Conductivity can be produced in diamond, however, in a number of ways, all of which involve destroying the perfection of the valence bond structure.

³Long before the arrangement of atoms in the diamond crystal was established by X-rays, the organic chemists had concluded that carbon formed four bonds at the tetrahedral angles—a truly remarkable result of inductive reasoning based on observations of the optical properties of solutions of organic compounds.

Thus, if high-energy particles or quanta of radiation fall upon the crystal, they can break the bonds. Conductivity in diamond induced by bombardment in this way has recently received considerable prominence in connection with "crystal counters" which have been used to detect nuclear particles and in experiments on conductivity induced by electron bombardment. An electron ejected from a bond constitutes a localized negative charge in the crystal and, since initially the bond structure was electrically neutral, the electron as it departs from its point of liberation leaves behind an equal, localized positive charge. Such a migratory electron, because it represents

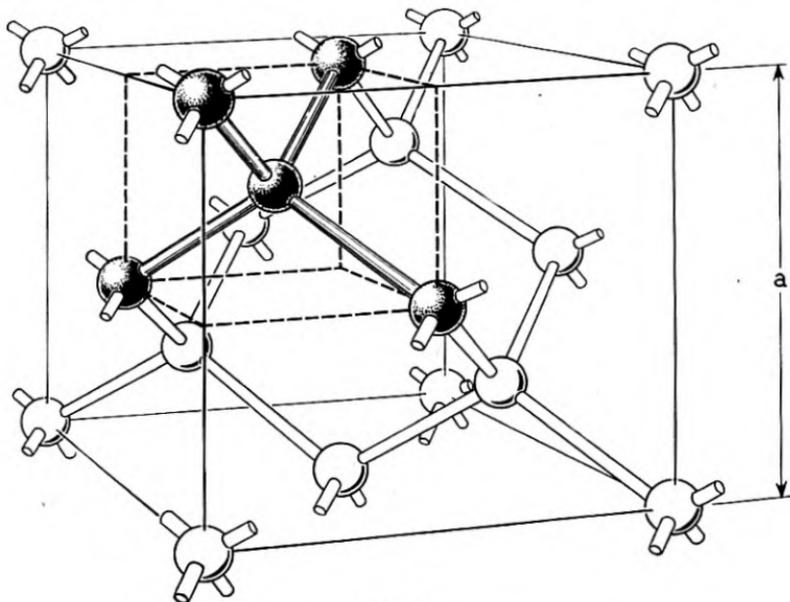


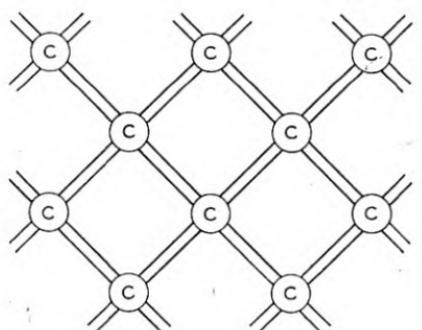
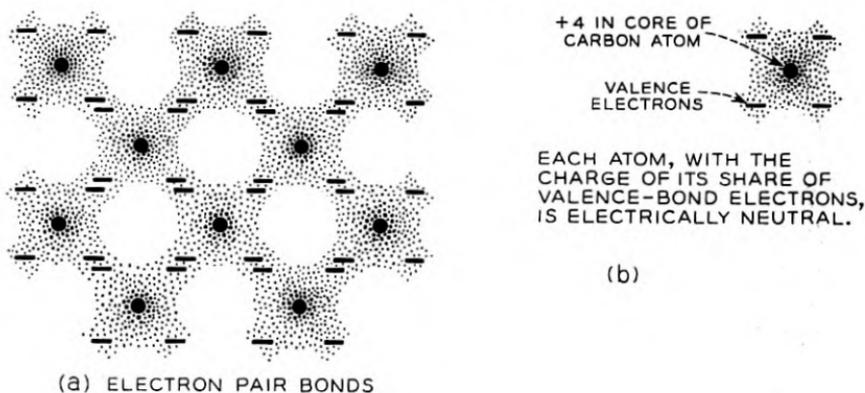
FIG. 2.

an excess over and above that required to complete the bond structure in its neighborhood, is called an "excess electron." Since it cannot enter any of the completed bonds in the lattice, it moves about in the crystal in a random manner under the influence of thermal agitation. If an electric field is applied, it tends to drift in the direction of the applied force and to carry a current. This illustrates conduction by excess electrons (referred to simply as conduction by electrons) and, as we shall see, is to be distinguished from the other process whereby an electron deficit enables conduction to be effected by "holes."

Such a hole, constituting a net, localized, positive charge in the crystal, moves from place to place by a reciprocal motion of electrons in the valence bonds; and, under the influence of an electric field, its random motion ac-

quires a systematic drift. Therefore it also can contribute to the current; in other words, current can flow as well by virtue of a deficit of electrons as by an excess of them.

In an illuminated and bombarded diamond crystal the electrons and holes, produced in pairs by the excitation, will of course drift in opposite directions under the influence of a field; the electron, being negative, drifts in the op-



(c) PLANE DIAGRAM OF DIAMOND LATTICE WITH BONDS REPRESENTED BY LINES



EACH ATOM, WITH THE CHARGE OF ITS SHARE OF VALENCE-BOND ELECTRONS, IS ELECTRICALLY NEUTRAL.

(b)

FIG. 3.

posite direction from the applied field, but its current is in the direction of the field. In the case of the hole, the reciprocal electron motions are once more opposite to the direction of the field (on the average). As a consequence, the net result is that the hole drifts in a direction to increase the current represented by the electrons. If the source of bombardment or illumination is removed, the conductivity dies away and the crystal will return to its normal state. This can occur by the recombination of holes and electrons. Whenever an electron drops into a hole, both the hole and the electron dis-

appear and the bond structure becomes complete, the excess energy being given up to the atoms in the form of thermal vibrations.⁴

If the temperature is sufficiently elevated, spontaneous breaking of some fraction of the covalent bonds by agitation will occur producing electrons and holes in equal numbers. In a diamond this effect would occur at such high temperatures that it would not be observed. However, it plays a major role in silicon and germanium at temperatures well within the range of investigation in the laboratory.

On the basis of quantum mechanical theory, it is found that a very high degree of symmetry exists between the behavior of electrons and the behavior

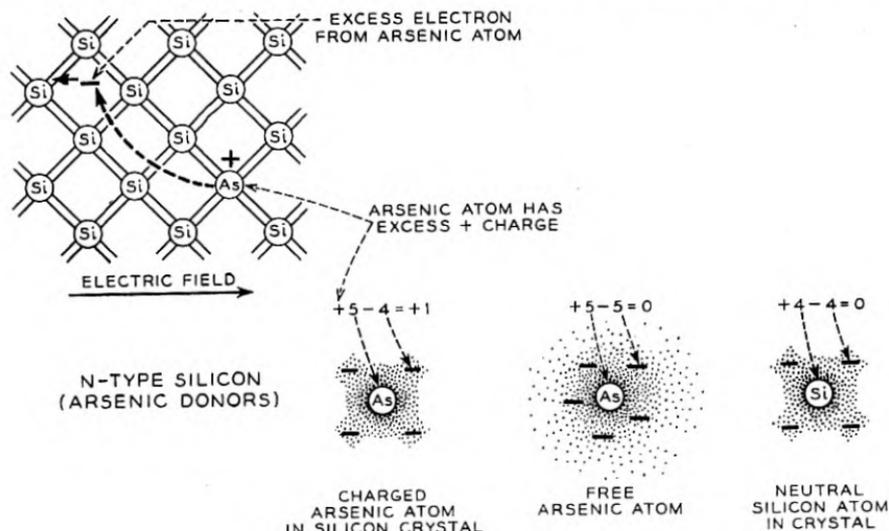


FIG. 4.

of holes. One may think of the hole as moving through the crystal as a positively charged particle with much the same attributes as a free electron except for the sign of its charge.

IMPURITY SEMICONDUCTORS: DONORS AND ACCEPTORS

If the only cases of conductivity open to investigation were like those discussed above, for which electrons and holes are present in equal numbers, the problem of interpreting the data would be very difficult. Fortunately, in the semiconductors silicon and germanium, there are cases in which conductivity is due to excess electrons only or to holes only.⁵

⁴ The process of recombination may actually be much more complicated and may involve intermediate stages in which the hole or the electron is trapped.

⁵ The behavior of silicon with impurities of the sorts discussed here was investigated by Scaff, Theurer, and Schumacher. Their work was stimulated by the development of silicon detectors for microwave use by R. S. Ohl, also of Bell Telephone Laboratories, in the prewar years.

If the conductivity of the sample is due to excess electrons it is called *n-type*, since the current carriers act like negative charges; if due to holes, it is called *p-type*, since the carriers act like positive charges.

Either type of conduction can be produced at will by admixture of a suitable "impurity," a donor such as arsenic yielding an excess of free electrons, while an acceptor like boron causes an electron deficit or a surplus of positive holes. The reason why arsenic and boron serve in these opposite capacities comes readily to hand.

The arsenic atom has five valence electrons surrounding a core having a net charge of +5 units and, when introduced (e.g. in silicon) as a low-fraction impurity, it is believed that each arsenic atom displaces one of the silicon atoms from its regular site and forms four covalent bonds with the neighboring silicon atoms, thus using four of its five valence electrons (see Fig. 4). The extra electron cannot fit into these four bonds and is free to move about the crystal. This excess electron therefore constitutes a mobile localized negative charge. The arsenic atom, on the other hand, is an *immobile* localized positive charge, since its core, with a charge of +5 units, is not neutralized by its share (-4) of the charge in the valence bonds. Its net charge, therefore, just balances that of the excess electron it sets free in the crystal. Thus arsenic impurity atoms add excess electrons but do not disturb the over-all electrical neutrality of the crystal. The negative electrons, however, are somewhat attracted by the positive arsenic atoms and at low temperatures become weakly bound to them. At room temperature, thermal agitation shakes them off so that they become free.

To produce a p-type semiconductor we choose an added impurity, such as boron, having three valence electrons and therefore not enough to complete the valence bond structure surrounding it. The hole in one of the bonds to the boron atom can be filled by an electron from an adjacent bond, and when this occurs the hole migrates away to the bond which just gave up one of its electrons. The boron atom thus becomes an *immobile* localized negative charge. Because of the symmetry between the behavior of holes and electrons, we can describe the situation by saying that the negative boron atom attracts the positively charged hole but that thermal agitation shakes the latter off at room temperature so that it is free to wander about and contribute to the conductivity.

Because of their valencies, phosphorous and antimony, as well as arsenic are in the donor class while aluminum, gallium and indium are additional examples of the acceptor class.

It is beyond the scope of this prefatory note to describe how, by measurements of conductivity and the Hall effect as influenced by the amount of added donor or acceptor, it has been possible to determine the concentration of electrons and holes, as well as to fix the energies needed to remove an electron from a donor, a hole from an acceptor, and to break a covalent bond

between lattice atoms. In samples of germanium of such purity that the amount of added donor or acceptor was too small to determine by conventional chemical methods, the conductivity was still controlled by the processes outlined above. And it is interesting to note that a portion of this investigation was carried out with the aid of radioactive antimony alloyed with the germanium, the radioactive property making possible an accurate count of antimony atoms, though present only in extremely attenuated amounts.

The semiconductor papers in this issue of the Journal will explain how these simple facts of electron exchange give rise to rectifying and amplifying properties.

SEMICONDUCTOR RECTIFIERS AND AMPLIFIERS

A contact between a metal and semiconductor may act as a rectifier, the contact resistance being high for one direction of current flow and low for the opposite. Rectification results from the presence in the semiconductor adjacent to the interface of a potential barrier or hill which the current carriers, electrons or holes, must surmount in order to flow across the junction. The direction of easy flow is that in which the carriers flow from the semiconductor to the metal. An applied voltage which produces a current flow in this direction reduces the height of the potential hill and allows the carriers to flow more easily to the metal. When the voltage is applied in the opposite direction the height of the barrier which the carriers must surmount in going from the metal into the semiconductor is unchanged, to a first approximation, and the resistance of the contact remains high. A *p*-type semiconductor is positive, an *n*-type negative, relative to the metal, in the direction of easy flow.

Rectifying contacts can also be made between two semiconductors of opposite conductivity types. The direction of easy flow is again that for which the *p*-type is positive, the *n*-type negative. The rectifying boundary may separate two regions with different conductivity characteristics within the same crystal.

In some contact rectifiers it is necessary to consider the flow of both types of carriers, electrons and holes, even though one type is overwhelmingly in excess under equilibrium conditions. An example is the germanium point contact rectifier such as the 400 type varistor. The germanium used is *n*-type and the normal concentration of holes is small compared to the concentration of conduction electrons. Nevertheless, a large part of the current in the forward direction consists of holes flowing away from the contact rather than electrons flowing in. The flow increases the concentration of holes in the vicinity of the contact and there is a corresponding increase in the concentration of electrons to compensate for the space charge of the holes. This increase in concentration of carriers increases the conductivity

of the germanium. The holes introduced in this way gradually combine with electrons and disappear so that at large distances the current consists largely of electrons. Similar effects occur at n - p boundaries in germanium; the current in the forward direction consists in part of holes flowing from the p -type region into the n -type region and electrons flowing from the n -type region into the p -type region.

The alteration of concentration of carriers and conductivity by current flow may be used to produce amplification in a number of ways. In the type-A transistor two point contacts are placed in close proximity on the upper face of a small block of n -type germanium. A large area low resistance contact on the base is the third element of the triode. Each point, when connected separately with the base electrode, has characteristics similar to those of the rectifier. When operated as an amplifier, one point, called the emitter, is biased in the forward direction so that a large part of the current consists of holes flowing away from the contact. The second point, called the collector, is biased in the reverse direction. In the absence of the emitter, the current consists largely of electrons flowing from the collector point to the base electrode. When the two points are in close proximity there is a mutual influence which makes amplification possible. The collector current produces a field which attracts the positively charged holes flowing from the emitter, so that a large part of the emitter current flows to the collector and into the collector circuit.

It has been found that rectifying boundaries between n - and p -type germanium may be used both as emitters and collectors, so that it is possible to make transistors without point contacts.

The following five papers are concerned with the behaviors of holes and electrons in semiconductors, with particular emphasis upon rectifying junctions and transistors. The first paper "Hole Injection in Germanium" describes new experiments on the behavior of holes and shows how their numbers and velocities may be measured and how they may be used to modulate the conductivity in the "filamentary transistor." The second paper "Some Circuit Aspects of the Transistor" describes the characteristics and equivalent circuits for the transistor. "Theory of Transient Phenomena in the Transport of Holes in an Excess Semiconductor" describes in mathematical terms a number of the processes encountered in the first paper and brings out interesting features of the nature of an advancing wave front of holes. "The Theory of Rectifier Impedances at High Frequencies" analyzes the behavior of metal-semiconductor rectifiers for high frequencies for the case in which the current is carried by one type of carrier only. As mentioned above, in rectifiers formed from p - n junctions, currents of both holes and of electrons must be considered. Such rectifiers and related subjects are dealt with in "The Theory of p - n Junctions in Semiconductors and p - n Junction Transistors."

Hole Injection in Germanium—Quantitative Studies and Filamentary Transistors*

By

W. SHOCKLEY, G. L. PEARSON and J. R. HAYNES

Holes injected by an emitter point into thin single-crystal filaments of germanium can be detected by collector points. From studies of transient phenomena the drift velocity and lifetimes (as long as 140 microseconds) can be directly observed and the mobility measured. Hole concentrations and hole currents are measured in terms of the modulation of the conductivity produced by their presence. Filamentary transistors utilizing this modulation of conductivity are described.

1. INTRODUCTION

THE invention of the transistor by J. Bardeen and W. H. Brattain^{1, 2, 3} has given great stimulus to research on the interaction of holes and electrons in semiconductors. The techniques discussed in this paper for investigating the behavior of holes in *n*-type germanium were devised in part to aid in analyzing the emitter current in transistors. The early experiments suggested that the hole flow from the emitter to the collector took place in a surface layer.^{1, 2} The possibility that transistors could also be produced by hole flow directly through *n*-type material was proposed in connection with the *p-n-p* transistor.⁴ Quite independently, J. N. Shive⁵ obtained evidence for hole flow through the body of *n*-type germanium by making a transistor with points on opposite sides of a thin germanium specimen. Such hole flow is also involved in the coaxial transistor of W. E. Kock and R. L. Wallace.⁶ Further evidence for hole injection into the body of *n*-type germanium under conditions of high fields was obtained by E. J. Ryder.⁷

In keeping with these facts it is concluded³ that with two points close together on a plane surface, as in the type-A transistor⁸, holes may flow either in a surface layer or through the body of the germanium. For surface flow to be large, special surface treatments appear to be necessary; such treatments were not employed in the experiments described in this paper and the results are consistent with the interpretation that the hole current from the emitter point flows in the interior.

The experiments described in this paper, in addition to any practical implications, serve to put the action of emitter points on a quantitative basis and to open up a new area of research on conduction processes in semicon-

* It is planned to incorporate this material in a book entitled "Holes and Electrons, an Introduction to the Physics of Transistors" currently being written by W. Shockley. This book is to cover much of the material planned for the "Quantum Physics of Solids" series which was discontinued because of the war.

ductors. It is worth while at the outset to contrast some of the new aspects of these experiments with the earlier experimental status of the bulk properties of semiconductors. Prior to the invention of the transistor, inferences about the behaviors of holes and electrons were made from measurements of conductivity and Hall effect. For both of these effects, under essentially steady state conditions, measurements were made of such quantities as lengths, currents, voltages and magnetic fields. The measurement of time was not involved, except indirectly in the calibration of the instruments. Nevertheless, on the basis of these data, definite mental pictures were formed of the motions of holes and electrons describing in particular their drift velocity in electric fields and the transverse thrust exerted upon them by magnetic fields. The new experiments show that something actually does drift in the semiconductor with the predicted drift velocity and does behave as though it had a plus or minus charge, just as expected for holes and electrons. In addition, experiments described elsewhere⁹ show that the effect of sidewise thrust by a magnetic field actually is observed in terms of the concentration of holes and electrons near one side of a filament of germanium.

We shall discuss here evidence that holes are actually introduced into *n*-type germanium by the forward current of an emitter point and show how the numbers and lifetimes of the holes can be inferred from the data. We shall refer to this important process as "hole injection." Discussions of the reasons why an emitter should emit holes are given for point contacts by J. Bardeen and W. H. Brattain^{1, 2, 3} and for *p-n* junctions elsewhere in this journal.⁴ There are other possible ways in which semiconductor amplifiers can be made without the use of hole injection into *n*-type material or electron injection into *p*-type material.* In this paper, however, our remarks will be restricted to semiconductors which have only one type of carrier present in appreciable proportions under conditions of thermal equilibrium; for such cases the theoretical considerations are simplified and are apparently in good agreement with the experiments.

2. THE MEASUREMENT OF DENSITY AND CURRENT OF INJECTED HOLES

The experiment in its semiquantitative form is relatively simple and is shown in Fig. 1.¹⁰ A rod of *n*-type germanium is subjected to a longitudinal electric field E applied by a battery B_1 . Collector and emitter point contacts are made to the germanium with the aid of a micromanipulator. The collector point is biased like a collector in a type-A transistor by the battery B_2 and the signal obtained across the load resistor R is applied to the input of an oscilloscope. At time t_1 the switch in the emitter circuit is closed so that a forward current, produced by the battery B_3 , flows through the emitter point. At t_3 the switch is opened. The voltage wave at the collector, as

* For example see references 1 and 11.

observed on the oscilloscope, has the wave form shown in part (b) of the figure.

These data are interpreted as follows: When the emitter circuit is closed, the electrons in the emitter wire start to flow away from the germanium (i.e. positive current flows into the germanium). These electrons are furnished by an electron flow in the germanium towards the point of contact. The flow in the germanium may be either by the excess electron process or by the hole process. In Fig. 2 we illustrate these two possibilities. At first

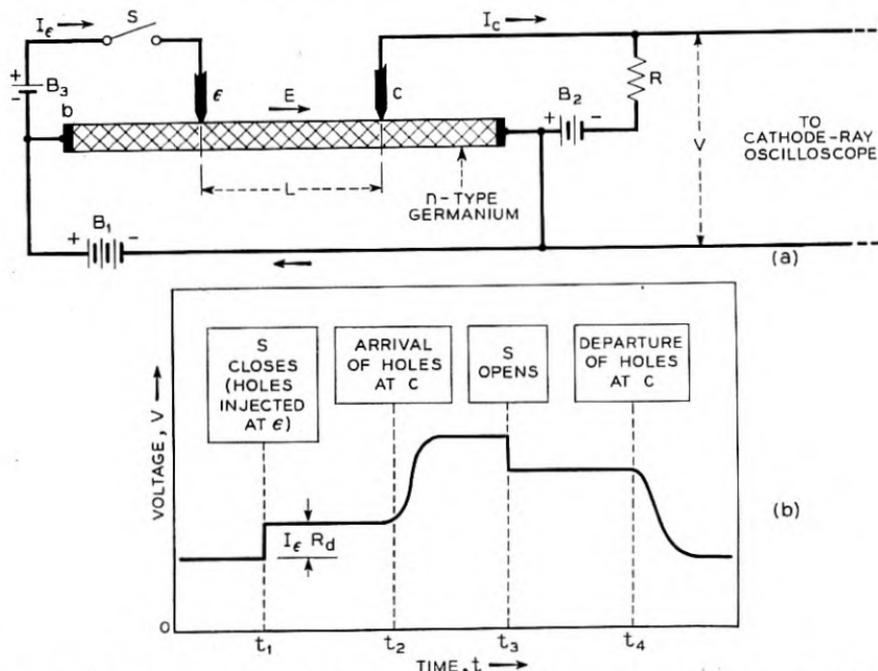


Fig. 1—Experiment to investigate the behavior of holes injected into *n*-type germanium
(a) Experimental arrangement.

(b) Signal on oscilloscope showing delay in hole arrival at t_2 in respect to closing S at t_1 and delay in hole departure at t_4 in respect to opening S at t_3 .

glance it might appear that the difference between these two processes is unimportant since the net result in both cases is a transfer of electrons from the germanium to the emitter point. There is, however, an important difference, one which makes several forms of transistor action possible. In the case of the hole process an electron is transferred from the valence band structure to the metal. After this the hole moves deeper into the germanium. As a result the electronic structure of the germanium is modified in the neighborhood of the emitter point by the presence of the injected holes.

Under the influence of the electric field E , the injected holes drift toward

the collector point with velocity $\mu_p E$, where μ_p is the mobility of a hole, and thus traverse the distance L to the collector point in a time $L/\mu_p E$. When they arrive at the collector point, they increase its reverse current and produce the signal shown at t_2 .

There are two important differences between the signal produced at t_2 and that produced at t_1 . The signal at t_1 , which is in a sense a pickup signal, would be produced even if no hole injection occurred. We shall illustrate this by considering the case of a piece of ohmic material substituted for the

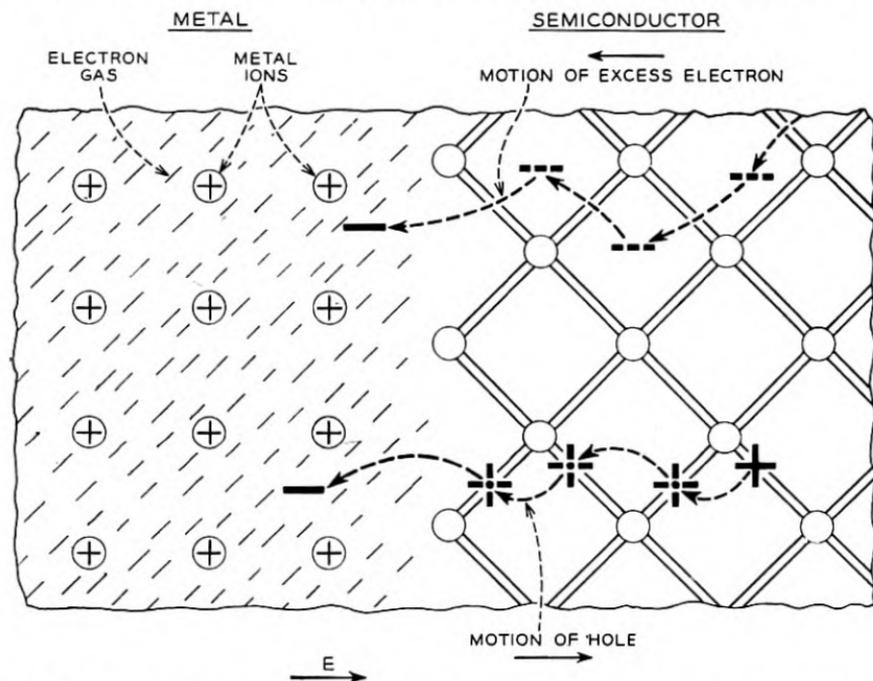


Fig. 2—Electron flow to the metal may be produced by an excess electron moving toward the metal or by bonding electrons jumping (dashed arrows) successively into a hole thus displacing the hole deeper into the semiconductor.

germanium. Conventional circuit theory applies to such a case; however, in order to contrast this purely ohmic case with that of hole injection, we shall also give a description of the conventional theory of signal transmission in terms of the motion of the carriers. According to conventional circuit theory, the addition of the current I_e would simply produce an added IR drop due to current flow in the segment of the specimen to the right of the collector. This voltage drop is denoted as $I_e R_d$ in part (b), R_d representing the proper combination of resistances to take into account the way in which I_e divides in the two branches. This signal will be transmitted from the emitter to the collector with practically the speed of light—the ordinary theory of signal

transmission along a conductor being applicable to it. This high speed of transmission does not, of course, imply a correspondingly high velocity of motion of the current carriers. In fact the rapidity of signal transmission has nothing to do with the speed of the carriers and comes about as follows: If the ohmic material is an electronic conductor, then the withdrawal of a few electrons by the emitter current produces a local positive charge. This positive charge produces an electric field which progresses with the speed of light and exerts a force on adjoining electrons so that they move in to neutralize the space charge. The net result is that electrons in all parts of the specimen start to drift practically instantaneously. They flow into the specimen from the end terminals to replace the electrons flowing out at the emitter point and no appreciable change in density of electrons occurs anywhere within the specimen.*

The distinction between the process just described and that occurring when holes are injected into germanium is of great importance in understanding many effects connected with transistor action. One way of summarizing the situation is as follows: In a sample having carriers of one type only, electrons for example, it is impossible to alter the density of carriers by trying to inject or extract carriers of the same type. The reason is, as described above (or proved in the footnote), that such changes would be accompanied by an unbalanced space charge in the sample and such an unbalance is self-annihilating and does not occur.†

When holes are injected into n -type germanium, they also tend to set up a space charge. Once more this space charge is quickly neutralized by an electron flow. In this case, however, the neutralized state is not the normal thermal equilibrium state. Instead the number of current carriers present has been increased by the injected holes and by an equal number of electrons drawn in to neutralize the holes. The total number of electrons in the specimen will thus be increased, the extra electrons coming in from the metal terminals which complete the circuit with the emitter point. The presence of the holes and the neutralizing electrons near the emitter point modify the conductivity. As we shall show below, this modification of conductivity may be so great that it can be used to measure hole densities and also to give power gain in modified forms of the transistor. We shall summarize this situation as follows: *In a semiconductor containing substantially only one type of current carrier, it is impossible to increase the total carrier concentration by*

* This is a description in words of the result ordinarily expressed in terms of the dielectric relaxation time obtained as follows: $\nabla \cdot I = -\dot{\rho}$, $I = \sigma E = -\sigma \nabla \Psi$, $\nabla^2 \Psi = -4\pi\rho/\kappa = \dot{\rho}/\sigma$ so that $\rho = \rho_0 \exp [-(4\pi\sigma/\kappa)t]$, where I = current density, ρ = charge density, σ = conductivity, E = electric field, Ψ = electrostatic potential, κ = dielectric constant.

† In the case of modulation of conductivity by surface charges,¹¹ a net charge is produced by the field from the condenser plate. The changed charge density extends slightly into the specimen but should not be confused with the true volume effect of hole injection. Such space charge layers are discussed in other articles in this issue.^{4, 12}

injecting carriers of the same type; however, such increases can be produced by injecting the opposite type since the space charge of the latter can be neutralized by an increased concentration of the type normally present.

Thus we conclude that the existence of two processes of electronic conduction in semiconductors, corresponding respectively to positive and negative mobile charges, is a major feature in several forms of transistor action.

In terms of the description given above, the experiment of Figure 1 is readily interpreted. The instantaneous rise at t_1 is simply the ohmic contribution due to the changing total currents in the right branch when the emitter current starts to flow. After this, there is a time lag until the holes injected into the germanium drift down the specimen and arrive at the collector. When the current is turned off at t_3 , a similar sequence of events occurs.

The measured values of the time lag of $t_1 - t_2$, the field E and the distance L can be used to determine the mobility of the holes. The fact that holes, rather than electrons, are involved is at once evident from the polarity of the effect; the disturbance produced by the emitter point flows in the direction of E , as if it were due to positive charges; if the electric field is reversed, the signal produced at t_2 is entirely lacking. The values obtained by this means are found to be in good agreement with those predicted from the Hall effect and conductivity data. The Hall mobility values obtained on single crystal filaments of n - and p -type germanium¹³ are

$$\mu_p = 1700 \text{ cm/sec per volt/cm}$$

$$\mu_n = 2600 \text{ cm/sec per volt/cm}$$

The agreement between Hall effect mobility and drift mobility, as was pointed out at the beginning of this section, is a very gratifying confirmation of the general theoretical picture of holes drifting in the direction of the electric field.

We shall next consider a more quantitative embodiment of the experiment just considered. In Fig. 3, we show the experimental arrangement. In this case it is essential in order to obtain large effects that the cross-section of the germanium filament be small. A thin piece of germanium is cemented to a glass backing plate and is then ground to the desired thickness. After this the undesired portions are removed by sandblasting while the desired portions are protected by suitable jigs consisting of wires, scotch tape, metal plates, etc. After the sandblasting, the surface of the germanium is etched. In this way specimens smaller than 0.01×0.01 cm in cross-section have been produced. The ends of the filament are usually made very wide so as to simplify the problem of making contacts.

Under experimental conditions, a battery like B_1 , of Figure 1 applies a "sweeping" field in the filament so that any holes injected by the emitter

current are swept along the filament from left to right. In the small filaments used for these experiments, the resulting concentration of holes is so high that large changes in conductivity are produced to the right of the emitter point and, as we shall describe below, these changes can be measured and the results used to determine the hole current at the emitter point. In order to treat this situation quantitatively, we introduce a quantity γ defined as follows:

$$\gamma = \text{the fraction of the emitter current carried by holes.}$$

Accordingly, a current γI_ϵ of holes flows to the right from ϵ and produces a hole density, denoted by p , which is neutralized by an equal added electron density. A fraction $(1 - \gamma)I_\epsilon$ of electrons flows to the left; these electrons do not, however, produce any increased electron density to the left of the emitter since they are of the sign normally present in the n -type material. The presence of the holes to the right in the filament increases the conductivity σ (as shown in Fig. 3c) both because of their own presence and the presence of the added electrons drawn in to neutralize the space charge of the holes. The mobility of electrons is greater than the mobility of holes, the ratio being¹³

$$b = \mu_n/\mu_p = 1.5 \text{ for germanium}$$

and the electrons are always more numerous than the holes*

$$n = n_0 + p, \quad (2.1)$$

where n_0 is the concentration of electrons which would be present to neutralize the donors if p were equal to zero; consequently, the current carried by electrons is greater than the current carried by holes. The concentration of holes diminishes to the right due to the fact that holes may recombine with electrons as they flow along the filament.

From this experiment the value of γ and the lifetime of a hole in the filament can be determined. The measurements are made with the aid of the two probe points P_1 and P_2 . The conductance of the filament between these points is obtained by measuring the voltage difference ΔV and dividing it into the current $I_b + I_\epsilon$, no current being drawn by the probes themselves. The necessary formulae for calculating hole density and hole current,

* The notation used in the equations is as follows: n, p, n_0 = respectively density of electrons, of holes, of electrons when no holes are injected. N_d and N_a are the densities of donors and acceptors, assumed ionized so that $n_0 = N_d - N_a$. I_ϵ, I_b, I_c are as shown on Figs. 3 and 9. (I_c used for the probe collector in Figures 1 and 8 does not enter the equations.) $q = |q|$ is the charge on the electron, used to be consistent with Ref. 4, where e is used for 2.718 ...

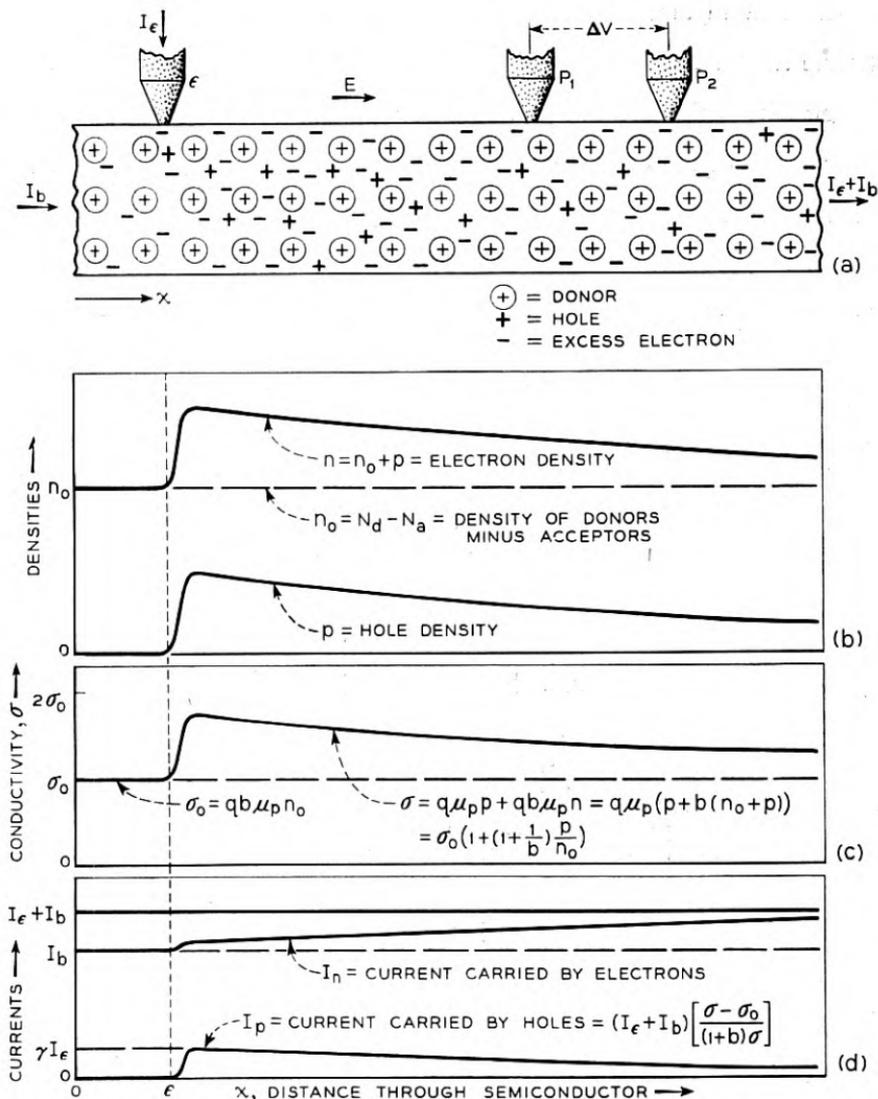


Fig. 3—Method of measuring hole densities and hole currents.

(a) Distribution of holes, electrons and donors. Acceptors, which may be present, the excess of donor density N_d over acceptor density N_a being n_0 .

(b) To the right of the emitter the added hole density p is compensated by an equal increase in electron concentration.

(c) The conductivity is the sum of hole and electron conductivities.

(d) The total current $I_b + I_\epsilon$ to the right of the emitter is carried by I_p and I_n in the ratio of the hole to the electron conductivity.

shown on the Figure, are derived as follows:

$$\text{Normal conductivity } \sigma_0 = q\mu_n n_0, \quad (2.2)$$

$$\begin{aligned} \text{conductivity with holes present } \sigma &= q\mu_n n + q\mu_p p \\ &= q\mu_n(n_0 + p) + q\mu_p p = \sigma_0[1 + (1 + b^{-1})(p/n_0)]. \end{aligned} \quad (2.3)$$

The conductance,

$$G = (I_\epsilon + I_b)/\Delta V,$$

between P_1 and P_2 is proportional to the local conductivity, and hence to

$$1 + (1 + b^{-1})(p/n_0),$$

so that a measurement of the conductance gives a measurement of p/n_0 . Letting G and G_0 be the conductances between the points with and without hole injection, we have

$$\frac{G}{G_0} = \frac{\sigma}{\sigma_0} = 1 + (1 + b^{-1})(p/n_0) \quad (2.4)$$

or

$$\frac{p}{n_0} = \frac{\sigma - \sigma_0}{\sigma_0(1 + b^{-1})} = \frac{(G/G_0) - 1}{1 + b^{-1}}. \quad (2.5)$$

The ratio of hole current to electron current is $q\mu_p p/q\mu_n n$ and the fraction of the current carried by holes is thus

$$\begin{aligned} \frac{I_p}{I_n + I_p} &= \frac{q\mu_p p}{q\mu_n n + q\mu_p p} = \frac{p}{bn_0 + (1 + b)p} \\ &= \frac{p/n_0}{b[1 + (1 + b^{-1})(p/n_0)]} = \frac{1 - (G_0/G)}{1 + b} \end{aligned} \quad (2.6)$$

Hence from the measured values of G , it is possible to obtain the fraction of the current carried by holes. Multiplying this fraction by $I_\epsilon + I_b$ then gives the actual hole current flowing past the probe points.* If there were no decay, the current past the probe points would be γI_ϵ and since I_ϵ is known, γ could be easily determined. Actually, however, there may be quite an appreciable decay. However, if the current I_b is increased, the holes will be swept more rapidly from the emitter to the probes and less decay will result. Thus by increasing I_b , the effect of recombination can be minimized and the value of hole current can be extrapolated to the value it would have in the absence of decay. This value is, of course, γI_ϵ .

* In these calculations the formulae $n = p + n_0$, corresponding to completely ionized donors and acceptors, has been used. In germanium this is a good approximation. For silicon, however, modifications will be necessary.

In Fig. 4 we show some plots of this sort. The ordinate is I_p/I_ϵ which should approach γ as the value of I_b becomes larger. The theory indicates that a logarithmic plot should be used and that the abscissa should be made proportional to transit time so that the case of no decay or zero transit time

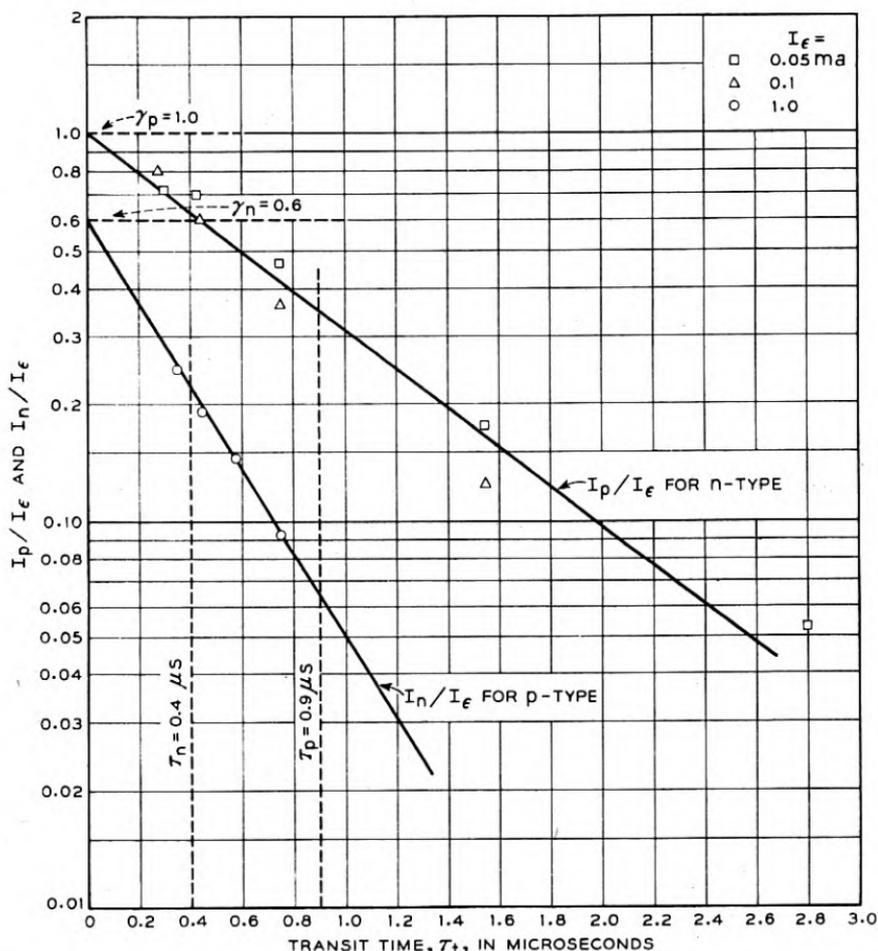


Fig. 4—Extrapolation of measured hole and electron currents to zero transit time in order to determine γ .

comes at the left edge.† The conclusion reached from this plot is that for the case of the *n*-type sample, the value of γ is substantially unity,—all the

† If the lifetime of a hole is τ , then the hole current at the points is $I_p = \gamma I_\epsilon \exp(-t/\tau)$ where t is the transit time to a point midway between the points, say a distance L from the emitter. If the electric field is $E = \Delta V/\Delta L$, then the transit time $t = L\Delta L/\mu_p\Delta V$. Hence if $\ln I_p$, as determined from the ratio of conductivities is plotted against $t = L\Delta L/\mu_p\Delta V$ a straight line with intercept $\ln \gamma I_\epsilon$ and slope $-1/\tau$ should be obtained.

emitter current is holes. For the opposite case in which electrons are injected into *p*-type material,¹⁴ the corresponding value of I_n/I_e extrapolates to 0.6 indicating that for this case 60% of the current is carried by electrons and 40% by holes. For these particular specimens the lifetimes are found to be 0.9 and 0.41 microseconds respectively. There is a body of evidence, some of which we discuss below, that holes combine with electrons chiefly on the surface of the filament.

3. THE INFLUENCE OF HOLE DENSITY ON POINT CONTACTS

The presence of holes near a collector point causes an increase in its reverse current; in fact the amplification in a type-A transistor is due to the modulation of the collector current by the holes in the emitter current. The influence of hole density upon collector current has been studied in connection with experiments similar to those of Fig. 3. After the hole current and the hole density are measured, a reverse bias of 20 to 40 volts is applied. The reverse current is found to be a linear function of the hole density. Figure 5 shows typical plots of such data. Different collector points, as shown, have quite different resistances. However, once data like that of Fig. 5 have been obtained for a given point, the currents can then be used as a measure of hole density. This experimental procedure for determining hole density is simpler than that involved in using the two points and much better adapted to studies of transient phenomena. It is necessary in employing this technique to keep the current drawn by the collector point somewhat smaller than $I_b + I_e$; otherwise the disturbance in the current flow due to the collector current is too great and the sample of the hole current is not representative. Experiments have shown, however, that this condition is readily achieved and that the collector current may be satisfactorily used as a measure of hole density.

The hole density also affects the resistance of a point at low voltage. Studies of this effect have also been made in connection with the experiment of Fig. 3. After the hole density has been determined from measurements of ΔV and $I_b + I_e$, a small additional voltage (0.015 volts) was applied between P_1 and P_2 and the current flowing externally between P_1 and P_2 was measured. From these data a differential conductance, for small currents, is obtained for the two points P_1 and P_2 in series. As is shown in Fig. 6, this conductance is seen to be a linear function of the hole concentration. The conductance of a point contact arises in part from electron flow and in part from hole flow. From experiments using magnetic fields⁹, it has been estimated that under equilibrium conditions the two contributions to the conductance may be comparable. In connection with Fig. 6 it should be noted that the hole concentration on the abscissa is the average hole

concentration throughout the entire cross section; the hole concentration may be much less near the surface due to recombination on the surface.

Techniques of the sort described above can be used to measure the properties of collector points. If a collector point is placed between the emitter and P_1 in Fig. 3, then the hole current extracted by the collector can be determined in terms of the hole current past P_1 and P_2 . By these means an "intrinsic α " for the collector point can be determined. The intrinsic α is

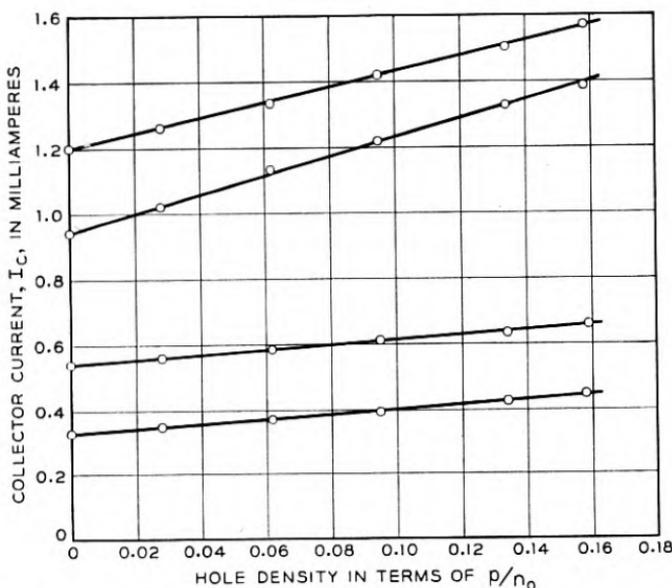


Fig. 5—Dependence of collector current I_c upon average hole density being swept by collector point. Collector biased 20 volts reverse.

defined as the ratio of change in collector current per unit change in hole current actually arriving at the collector.

4. STUDIES OF TRANSIENT PHENOMENA

The technique of using a collector point to measure hole concentrations has been employed in a number of experiments similar to those described in connection with Fig. 1. These experiments give information concerning hole lifetimes, hole mobilities, diffusion and conductivity modulation.

One of the methods employed to measure hole lifetime involves the measurement of the increase in collector current, produced by the arrival of the leading edge of the hole pulse, as a function of the transit time of the holes from emitter to collector. This time is varied by varying the distance between the emitter and the collector points.

In Fig. 7 we show a plot, obtained in this way, from a sample of germanium having dimensions $1.0 \times .05 \times .08$ cm. It is seen that the increase in collector current due to hole arrival decays exponentially with a time constant of 18 microseconds. This time constant increases as the dimensions of the germanium sample are increased so that a time constant of 140 microseconds was measured, using a sample having dimensions $2.5 \times .35 \times .30$

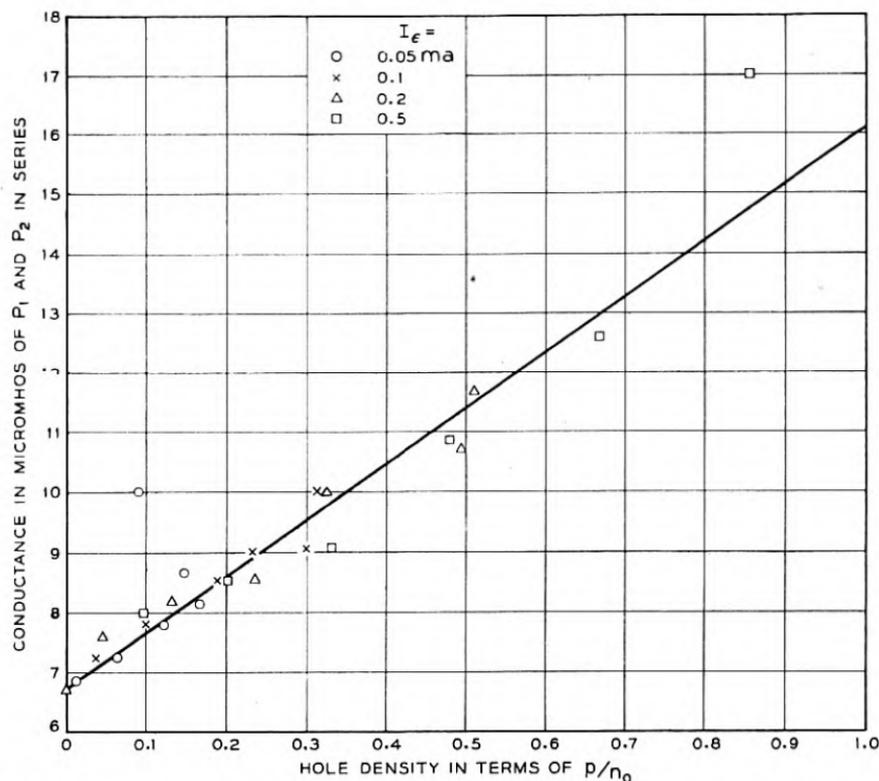


Fig. 6—Conductance of P_1 and P_2 of Fig. 3 in series as a function of p/n_0 , showing that conductance depends on hole concentration but not on currents in filament. For each value of I_ϵ the hole density was varied by varying $I_b + I_\epsilon$ from .038 to 0.78 ma.

cm. Since the holes injected into the interior of this sample can diffuse to the surface and recombine in about 100 microseconds, the process may still be largely one of surface recombination. In any event, it may be concluded that the lifetime in the bulk material used must be at least 140 microseconds. Making use of the electron density determined from other measurements, we conclude that the recombination cross section must be less than 10^{-18} cm². This cross section, which is less than 1/400 the area of a germanium

atom, may be so small because a hole-electron pair has difficulty in satisfying in the crystal the conditions somewhat analogous to conservation of energy and momentum which hinder recombination of electrons and positive ions in a gas discharge. Thus it has been pointed out that a hole-electron pair will have a lowest energy state in which the two current carriers behave something like the proton and electron of a hydrogen atom.¹⁵ Such a bound pair are called an exciton and the energy given up by their recombination is the "exciton energy." In order to recombine they must radiate this energy in

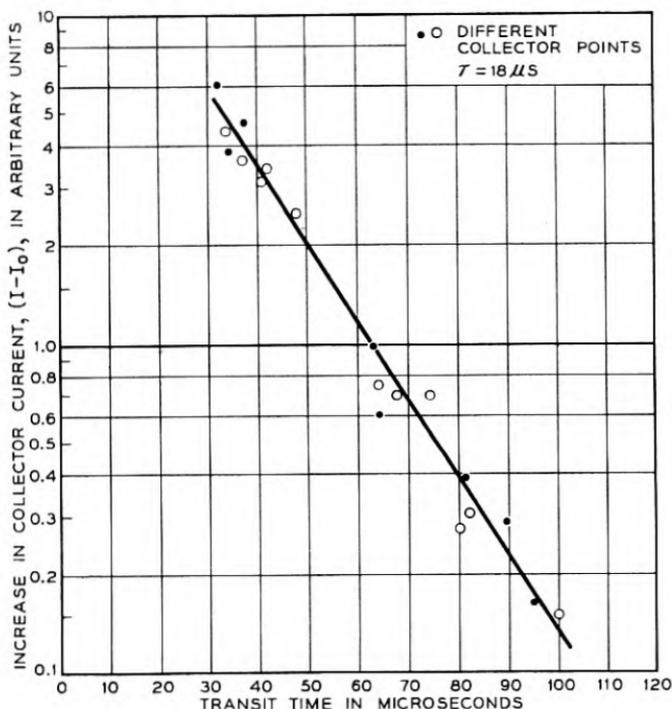


Fig. 7—The decay of injected holes in a sample of *n*-type germanium.

the form of a light quanta (photon) or a quantum of thermal vibration of the crystal lattice (phonon). The recombination time for the photon recombination process can be estimated from the optical constants for germanium and the theory of radiation density using the principle of detailed balancing, which states that under equilibrium conditions the production of hole electron pairs by photon absorption equals the rate of recombination with photon emission; the lifetime obtained in this way is about 1 second at room temperature indicating that the photon process is unimportant.¹⁰ As has been pointed out by A. W. Lawson,¹⁶ the highest energy phonon will have

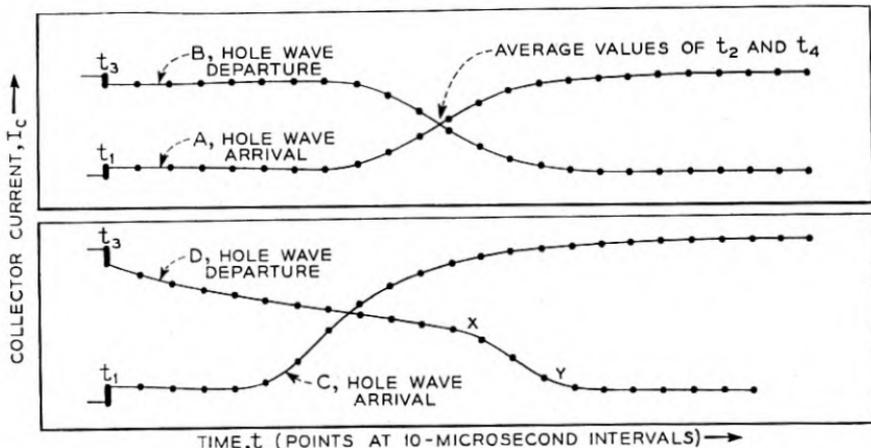
insufficient energy to carry away the "exciton energy" of a hole-electron pair and, therefore, the release of energy will require the cooperation of several phonons with a correspondingly small transition probability.

When a square pulse of holes is injected in an experiment like that of Fig. 1, the leading and trailing edges of the current at the collector point are deformed for several reasons. Due to the high local fields at the emitter point, some of the holes actually start their paths in the wrong direction—i.e. away from the collector; these lines of flow later bend forward so that those holes also pass by the collector point but with a longer transit time than holes which initially started towards the collector. A spread in transit times of this sort is probably largely responsible for the loss of gain at high frequencies in transistors. For the experiments described below, however, this effect is negligible compared to two others which we shall now describe.

On top of the systematic drift of holes in the electric field, there is superimposed a random spreading as a result of their thermal motion. This would cause a sharp pulse of holes to become spread so that after drifting for a time t_d the hole concentration would extend over a distance proportional to $\sqrt{Dt_d}$ where D , the diffusion constant for holes, $= kT\mu_p/q = 45 \text{ cm}^2/\text{sec}$. As a result of this effect, the leading and trailing edges of the square wave of emission current become spread out when they arrive at the collector. This is shown in Fig. 8, curve *A* for the leading edge and *B* for the trailing edge. The points are 10 microsecond marker intervals traced from an oscilloscope, the time being measured from the instant at which the emitter current starts. For *A* and *B* the emitter current was so small compared to the current I_b that the holes produced a negligible modulation of conductivity and each hole moved in essentially the same electric field. It is to be observed that the wave shapes are nearly symmetrical in time about the half rise point and that the *A* and *B* waves are identical except for sign. This is just the result to be expected from diffusion. Furthermore, analysis shows that the spread in arrival time is in good quantitative agreement with the theoretical wave shape using the diffusion constant appropriate for holes. For this case the mid-point of the rise, corresponding to the crossing point of the curves, gives the average arrival time and has been used to obtain an accurate measure of the mobility.

Curves *C* and *D* correspond to conditions in which the emitter current was relatively large—two thirds of the base current. High impedance sources are used so that I_b is constant and I_e is a good flat topped wave. For the currents used in this experiment, the conductivity is appreciably modulated by the presence of holes. This accounts for the shape of curve *C*, corresponding to the arrival of holes at the collector. It is seen that this curve is not symmetrical but is much more gradual towards later times. The reason for

this is that the first holes to arrive are those which have diffused somewhat ahead of the rest and move in material of low conductivity. The later holes travel in an environment of relatively high conductivity and, consequently, in a lower electric field. (Since the current is the same at all points between emitter and collector, the field is inversely proportional to the conductivity.) The transit time for the later holes is, therefore, longer and the hole density builds up more slowly for the latter part of the incoming pulse of holes. The wave form obtained from the trailing edge of the emitter pulse, curve *D*, is in striking contrast with the leading edge, up to



A & B EMITTER CURRENT SMALL, ABOUT 4% OF I_b , SO THAT ALL HOLES MOVE IN THE SAME FIELD.

C LEADING EDGE OF PULSE FOR $I_e = 2/3 I_b$.

D TRAILING EDGE OF HOLE PULSE FOR $I_e = 2/3 I_b$, SHOWING SHARPENING FROM X TO Y DUE TO TENDENCY OF LAGGING HOLES TO CATCH UP.

Fig. 8—Collector current characteristics for the circuit shown in Fig. 1.

point *X*, is due to recombination of holes and electrons; at t_3 the emitter current becomes zero; consequently, the electric field is reduced and the holes arriving at *X* have taken a longer transit time than the holes arriving at t_3 and a larger fraction of them have recombined with electrons. The true trailing edge, running from *X* to *Y*, is appreciably sharper than the leading edge. The reason for this is that holes lagging behind the main body of holes are in a region of relatively low conductivity and high electric field and tend to catch up with the main body. Thus the same effect which lengthens wave *C* acts to shorten wave *D*.

C. Herring has been able to obtain mathematical solutions for the appropriate equations bearing on the matters just discussed. His theory is presented elsewhere in this issue.¹⁷

The delay feature discussed in connection with Figs. 1 and 8 indicates interesting possibilities of using germanium filaments as delay or storage elements.

5. THE THEORY OF THE FILAMENTARY TRANSISTOR

In Fig. 9 we show a transistor with a filamentary structure.¹⁸ Modulation is achieved in this case by injecting holes at the emitter point which flow to the right and modulate the resistance in the output branch between emitter and collector. Structures of this sort can be produced by the sandblasting technique discussed in Section 2. The enlarged ends, which give the

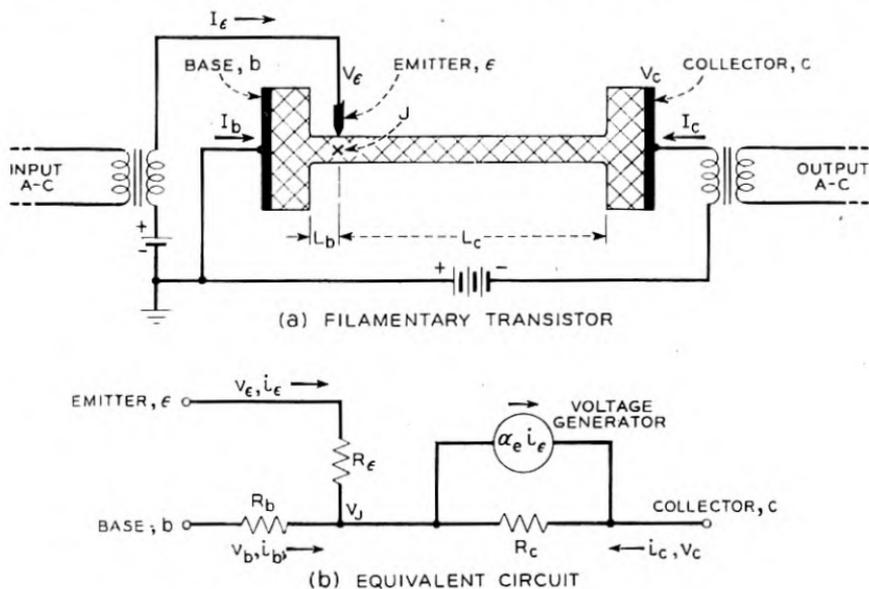


Fig. 9—Filamentary transistor and equivalent circuit.

unit a dumbbell appearance, decrease the problem of making contact to the unit. The large area at the left side serves the additional purpose of reducing unwanted hole emission from the metal electrode and affords an opportunity for any emitted holes to recombine before they enter the narrow part of the unit.

The theory of this transistor is relatively simple and most of the features we shall discuss in connection with it have counterparts in the theory of the type-A transistor. We shall discuss the case for which the injected current is a small fraction of the total current in the filament. Under these conditions we can use a simple linear theory. We shall show that the behavior of the transistor can be given for small a-c signals by the equivalent circuit in

Figure 9(b), which shows the current and voltage relationships in a form equivalent to those used in connection with the type-A transistor.

The point J in Fig. 9 represents a point in the filament near the emitter point. The current from the emitter point will be determined by the difference between its voltage V_ϵ and that of the surrounding semiconductor, namely the voltage at J . Thus we can write

$$I_\epsilon = f_\epsilon(V_\epsilon - V_J). \quad (5.1)$$

For small a-c variations, i_ϵ , v_ϵ and v_J , this equation leads to the relationship

$$i_\epsilon = (v_\epsilon - v_J)f'_\epsilon, \quad (5.2)$$

where f'_ϵ is the derivative of f_ϵ in respect to its argument. Letting $f'_\epsilon = 1/R_\epsilon$ this equation becomes

$$v_\epsilon - v_J = R_\epsilon i_\epsilon. \quad (5.3)$$

This relationship is correctly represented by the R_ϵ branch of the equivalent circuit. The voltage at J , under the assumed operating conditions with I_ϵ positive and much less than I_c , will be $-I_b R_b$ where R_b is the resistance from the base to an imaginary equipotential surface passing through J and $v_b = 0$, corresponding to *grounded base* operation. This leads to

$$v_J = -R_b i_b = +R_b i_\epsilon + R_b i_c, \quad (5.4)$$

since $i_b + i_\epsilon + i_c = 0$. This relationship is obviously satisfied by the R_b branch of the equivalent circuit.

We now come to the collector branch which we have represented as a resistance R_c and a parallel current generator* $\alpha_\epsilon i_\epsilon$. (This circuit is equivalent to another in which the parallel current generator is replaced by a series voltage generator $\alpha_\epsilon R_c i_\epsilon$.) We must show that this part of the equivalent circuit represents correctly the effect of injecting holes into the right arm of the filament. We shall suppose that there is negligible recombination so that the hole current injected at the emitter point flows through the entire filament. (We consider recombination in the next section.) The current I_c in the collector branch thus contains a component $-\gamma I_\epsilon = I_p$ of hole current (minus because of the algebraic convention that positive $I_c (= -I_b - I_\epsilon)$ flows to the left). The added hole and electron concentrations lower the resistance and R_c changes to $R_c + \delta R_c$, where δR_c is negative. The current voltage relationship for this branch of the filament then becomes

$$V_c - V_J = (R_c + \delta R_c) I_c. \quad (5.5)$$

* α_ϵ in the equivalent circuit differs from $\alpha = -(\partial I_c / \partial I_\epsilon) v_\epsilon$ by the relationship $\alpha_\epsilon = \alpha + (\alpha - 1)(R_b / R_c)$, equivalent to equation (6.8).

Our problem is to reexpress this relationship in terms of the small a-c components and show that it reduces to the relationship

$$v_c - v_J = R_c(i_c + \alpha_e i_e) \quad (5.6)$$

corresponding to the equivalent circuit. For small emitter current the analysis is carried out conveniently as follows: The ratio of hole current to the total current is $-\gamma I_e/I_c$. The ratio $(R_c + \delta R_c)/R_c$ corresponds to G_0/G discussed in connection with Fig. 3. The ratio of hole current to total current is given in (2.6) in terms of G_0/G and may be rewritten as

$$-\frac{\gamma I_e}{I_c} = \frac{1 - (G_0/G)}{1 + b} = \frac{-\delta R_c}{(1 + b)R_c}, \quad (5.6)$$

giving

$$\delta R_c = R_c(1 + b)\gamma I_e/I_c. \quad (5.7)$$

(Since I_c is negative and I_e is positive this equation shows that δR_c is negative, i.e., the conductivity has been increased by the hole current.) Putting this value of $R_c + \delta R_c$ into the equation for $V_c - V_J$ gives

$$\begin{aligned} V_c - V_J &= (R_c + \delta R_c)I_c \\ &= R_c[I_c + (1 + b)\gamma I_e]. \end{aligned} \quad (5.8)$$

If we consider small a-c variations in the currents and voltages, this reduces to the equation given by the equivalent circuit with

$$\alpha_e = (1 + b)\gamma. \quad (5.9)$$

The data of Section 2 indicate that for holes injected into *n*-type germanium $\gamma = 1$, and since $b = 1.5$ we obtain $\alpha_e = 2.5$.

The quantity v_J can be eliminated by using $v_J = R_b(i_e + i_c)$ in equation (5.3) for v_e and the small signal form of (5.8) for v_c leading to the pair of equations

$$v_e = (R_e + R_b)i_e + R_b i_c \quad (5.10)$$

$$v_e = (R_b + \alpha_e R_c)i_e + (R_c + R_b)i_c. \quad (5.11)$$

These equations are formally identical with those for the equivalent circuits of the type-A transistor.

It should be emphasized that although hole injection into *n*-type germanium plays a role in both the type-A and the particular form of filamentary transistor shown in Fig. 9, there are differences in the principles of operation. One important feature of the type-A is the high impedance of the rectifying collector contact which, however, does not impede hole flow and another important feature is the current amplification occurring at the collector contact. Neither of these features is present in the filamentary type shown. Instead, the high impedance at the collector terminal arises from the small

cross-section of the filament. The modulation of the output current takes place through the change in body conductivity due to the presence of the added holes, a change which appears to be unimportant in the type-A transistor. In the filamentary type, current amplification is produced by the extra electrons whose presence is required to neutralize the space charge of the holes. Current amplification in the type-A transistor is, probably, also produced by the space charge of the holes³ but the details of the mechanism are not as easily understood.

6. EFFECTS ASSOCIATED WITH TRANSIT TIME

Two important effects arise from the fact that a finite transit time is required for holes to traverse the R_c side of the filament: during this time the holes recombine with electrons and the modulation effect is attenuated for this reason; also the modulation of the conductivity of the filament at any instant is the result of the emitter current over a previous interval and for this reason there will be a loss of modulation when the period of the a-c signal is comparable with the transit time or less.

For the small signal theory, the effect of transit time is readily worked out in analytic terms. We shall give a derivation based on the assumption that the lifetime of a hole before it combines with an electron is τ_p . According to this assumption, the fraction of the holes injected at instant t_1 which are still uncombined at time t_2 is $\exp[-(t_2 - t_1)/\tau_p]$. This means that the effect in the filament at any instant t_2 is the average, weighted by this factor, of all the contributions prior to t_2 back to time $t_2 - \tau_t$ where τ_t is the transit time; holes injected prior to $t_2 - \tau_t$ have passed out of the filament by time t_2 . If the emitter current is represented by $i_{e0}e^{i\omega t}$, the effective average emitter current is

$$i_{e\text{ eff}}(t_2) = i_{e0} \int_{t_2 - \tau_t}^{t_2} e^{i\omega t_1 - (t_2 - t_1)/\tau_t} dt_1 / \tau_t. \quad (6.1)$$

The term dt_1/τ_t is chosen so that a true average is obtained since the sum of all the dt_1 intervals add up to τ_t . The integral is readily evaluated and gives

$$i_{e\text{ eff}}(t_2) = i_{e0} e^{i\omega t_2} \frac{1 - \exp[-i\omega\tau_t - (\tau_t/\tau_p)]}{i\omega\tau_t + (\tau_t/\tau_p)}. \quad (6.2)$$

The result so far as the equivalent circuit is concerned is that obtained by taking α_e as*

$$\alpha_e = \gamma(1 + b)\beta, \quad (6.3)$$

* The derivation of equations (5.10) and (5.11), describing the equivalent circuit, shows that hole injection enters only through the term $\delta R_c I_e$ in (5.8). This term leads only to $\alpha_e R_c i_e = (1 + b)\gamma R_c i_e$ in (5.11) and should be replaced by $(1 + b)\gamma R_c i_{e\text{ eff}} = (1 + b)\gamma\beta R_c i_e$ leading to (6.3).

where

$$\beta = \frac{1 - \exp[-i\omega\tau_t - (\tau_t/\tau_p)]}{i\omega\tau_t + (\tau_t/\tau_p)}. \quad (6.4)$$

β represents the effect of recombination and transit angle, $\omega\tau_t$, in reducing the gain.

We shall consider two limiting cases of this expression. First if $\omega\tau_t$ is very small, the new factor becomes

$$\beta = (\tau_p/\tau_t)(1 - e^{-\tau_t/\tau_p}). \quad (6.5)$$

If τ_t is much larger than τ_p , so that the holes recombine before traversing the filament, then the exponential is negligible and β becomes simply τ_p/τ_t . This means that the effectiveness of the holes is reduced by the ratio of their effective distance of travel to the entire length of the filament, i.e., τ_p/τ_t is the ratio of distance travelled in one lifetime to the entire length of the filament. Essentially the holes modulate only the fraction of the filament which they penetrate. The transit time depends on the field in the filament which is $|V_c - V_J|/L_c$, the absolute value being used since V_c is negative. The transit time is thus

$$\tau_t = L_c/[\mu_p |V_c - V_J|/L_c] = L_c^2/\mu_p |V_c - V_J|. \quad (6.6)$$

For very small emitter currents $V_c - V_J = R_c V_c/(R_c + R_b)$ so that

$$\tau_t = L_c^2(R_c + R_b)/\mu_p R_c |V_c| \quad (6.7)$$

and τ_t is inversely proportional to V_c . For large values of V_c , τ_t approaches zero and β approaches unity. The dependence of β upon V_c has been investigated by measuring α and plotting it as a function of $|1/V_c|$, as shown in Fig. 10. The value of

$$\alpha = -(\partial I_c/\partial I_e)_{V_c} \quad (6.8)$$

is readily found from the equivalent circuit, using equation (5.11), to be

$$\alpha = \frac{R_b}{R_b + R_c} + \frac{\alpha_e R_c}{R_b + R_c}. \quad (6.9)$$

For the particular structure investigated, the values of R_b and R_c , obtained at $I_e = 0$, were in the ratio 1:4. The value of α obtained by extrapolating the data to $|V_c| = \infty$ is 2.2; the value given by the formula for this case with $\beta = 1$, is

$$\alpha = 0.2 + 0.8 \times 2.5 \times \gamma, \quad (6.10)$$

from which we find $\gamma = 1.0$, in agreement with the result of Fig. 4 that

substantially all of the emitter current is carried by holes. The theoretical curve shown on the Figure is

$$\alpha = 0.2 + 0.8 \times 2.5 \times |V_c/10|(1 - e^{10/|V_c|}). \quad (6.11)$$

This corresponds to

$$\frac{\tau_t}{\tau_p} = \frac{10}{|V_c|} = \frac{L_c^2(R_c + R_b)}{\tau_p \mu_p R_c |V_c|}, \quad (6.12)$$

from which it was concluded that for the particular bridge studied τ_p was 0.2 microseconds.

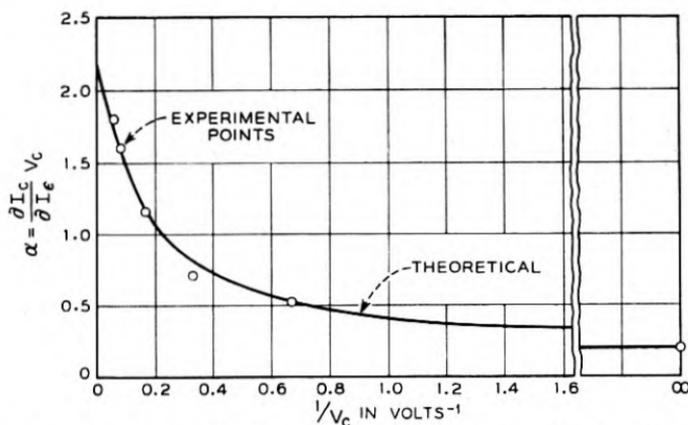


Fig. 10— α versus $1/|V_c|$ showing agreement with the theory for the value of β .

If τ_t is much shorter than τ_p , then the holes penetrate the whole filament and β becomes

$$\beta = \frac{1 - \exp(-i\omega\tau_t)}{i\omega\tau_t} = \frac{e^{-i\omega\tau_t/2} \sin(\omega\tau_t/2)}{(\omega\tau_t/2)}. \quad (6.13)$$

For small values of $\omega\tau_t$, β approaches unity since $(\sin x)/x$ approaches unity as x approaches zero. For $\omega\tau_t/2 = \pi$, the response is zero. This is the condition that $\tau_t = 2\pi/\omega = 1/f$. For this case the filament is just so long that the modulation is averaged over the time of one cycle of the input signal and since this average includes all phases, the modulation vanishes.

Preliminary experiments with filamentary transistors, made in accordance with the principles discussed above, appear to confirm the general aspects of the theory. Power gains of 15 db have been obtained and frequency responses showing a drop of 3 db in α at 10^6 cycles/sec. have been observed. Noise measurements indicate an improvement of 10 to 15 db over the average type-A transistor for comparable conditions of preparation.

ACKNOWLEDGMENTS

We have been aided and encouraged in these experiments by many of our colleagues in the Research and Apparatus Development Departments. We are particularly indebted to W. H. Brattain and H. R. Moore for help with experimental problems, to J. Bardeen and W. van Roosbroeck for assistance with the theory and to P. W. Foy and W. C. Westphal for their many contributions in connection with fabricating and measuring the filaments.

REFERENCES

1. J. Bardeen and W. H. Brattain, *Phys. Rev.*, **74**, 230 (1948).
2. W. H. Brattain and J. Bardeen, *Phys. Rev.*, **74**, 231 (1948).
3. J. Bardeen and W. H. Brattain, *Phys. Rev.*, **75**, 1208 (1949).
4. W. Shockley, *Bell Syst. Tech. J.*, July (1949).
5. J. N. Shive, *Phys. Rev.*, **75**, 689 (1949).
6. W. E. Kock and R. L. Wallace, *Electrical Engineering*, **68**, 222 (1949).
7. E. J. Ryder and W. Shockley, *Phys. Rev.*, **75**, 310 (1949).
8. For reviews of the type-A transistor see reference 3, R. M. Ryder, *Bell Laboratories Record*, Mar. 1949, J. A. Becker and J. N. Shive, *Electrical Engineering*, **68**, 215 (1949) and R. M. Ryder, *Bell Syst. Tech. J.*, July (1949).
9. H. Suhl and W. Shockley, *Phys. Rev.*, **75**, 1617 (1949), **76**, 180 (1949).
10. Experiments of this sort were first reported by J. R. Haynes and W. Shockley, *Phys. Rev.*, **75**, 691 (1949).
11. W. Shockley and G. L. Pearson, *Phys. Rev.*, **74**, 232 (1948).
12. John Bardeen, *Bell Syst. Tech. J.*, July (1949).
13. G. L. Pearson, *Phys. Rev.*, **76**, 179 (1949).
14. Transistors using p -type germanium have been described by W. G. Pfann and J. H. Scaff, *Phys. Rev.*, **76**, 459 (1949). Electron injection in p -type germanium has also been observed by R. Bray, *Phys. Rev.*, **76**, 458 (1949) and *Phys. Rev.*, **76**, 152 (1949).
15. G. Wannier, *Phys. Rev.*, **52**, 191 (1937).
16. Personal communication; a somewhat similar case is treated by B. Goodman, A. W. Lawson and L. I. Schiff, *Phys. Rev.*, **71**, 191 (1947).
17. C. Herring, *Bell Syst. Tech. J.*, July (1949).
18. Transistors of this type, employing p - n junctions as well as point contacts as emitters, have been discussed by W. Shockley, G. L. Pearson, M. Sparks, and W. H. Brattain, *Phys. Rev.*, **76**, 459 (1949).
19. Optical constants for germanium have been published by W. H. Brattain and H. B. Briggs, *Phys. Rev.*, **75**, 1705 (1949). The integration over the radiation distribution was carried out by W. van Roosbroeck.

Some Circuit Aspects of the Transistor

By R. M. RYDER and R. J. KIRCHER

INTRODUCTION

THE purpose of this note is to discuss in a general way some circuit aspects of the transistor. It is rather interesting that in order to discuss its circuit aspects, little direct reference to the transistor is necessary. One needs only certain properties of the transistor which are empirically obtainable by measurement; these properties then determine behavior in the manner prescribed by the methods of general network theory. In principle, one needs no knowledge of the physics of the transistor in order to treat it circuitwise; any "black box" with the same electrical behavior at its terminals would act the same way.

It is rather fortunate for our purposes that the problem does separate nicely in this way. The operation of the transistor is reasonably well understood; but, for calculations of performance from physical properties, the numerical parameters needed are somewhat inaccessible, numerous and complicated. The paper by Shockley¹ gives some calculations of this kind which are illuminating for theoretical understanding. However, just as with electron tubes, practical engineering calculations often do not need to go back to the ultimate physics. Starting from the electrical properties of the transistor as empirically determined by measurements on its terminals, we need go only to the literature of electrical engineering to find much practically useful information on properties of circuits which could be built around the unit.

This method of characterizing the electrical performance of a device more or less independently of its physical construction has come into wide use in recent years. A considerable amount of work has been done with applications to both electron tubes and transistors at the Bell Telephone Laboratories by L. C. Peterson.² The purpose of the present note, however, is not to go deeply into the subject but rather to review it in a general way, indicating applications to some of the simpler transistor circuits and comparisons with electron tubes. For more profound analyses one may refer to Peterson's work.

¹"The Theory of p-n Junctions in Semiconductors and p-n Junction Transistors," W. Shockley, this issue of *The Bell System Technical Journal*.

²"Equivalent Circuits of Linear Active Four-Terminal Networks," L. C. Peterson, *Bell System Technical Journal*, Oct. 1948, pp. 593-622.

The method used for circuit analysis may be grouped under the following headings:

1. Linear problems, like low-level amplifiers or the question of onset of oscillations. Such problems visualize the transistor as making only small excursions from an assumed operating point and are best treated by the method of small-signal analysis. The unit is assigned an equivalent circuit or, in mathematical terms, is dealt with by means of linear equations.
2. Slightly non-linear problems, like Class A power amplifiers. Here the excursions about the operating point are large enough to bring in higher-order effects like harmonic generation or intermodulation, but still small enough so that these effects can be treated by adding to the equivalent circuit certain distortion generators. Mathematically, some terms need to be added to the linear equations but these terms are of the nature of corrections, not big changes.
3. Highly non-linear problems, such as Class B or C amplifiers, oscillators, switches, harmonic generators. Here the excursions about the characteristic are so large as to reduce the linear approximation to the status of a qualitative guide or perhaps to invalidate it entirely; mathematically, the small signal series either require many terms for accuracy or else do not converge at all. These large-signal problems usually have to be treated by methods special for each problem. Frequently one uses graphical constructions from the static characteristics, or analytical methods starting from reasonable approximations to the static characteristics.
4. Finally, in certain highly non-linear problems the non-linear features are in a sense subsidiary; one is really interested in the behavior of a superposed small signal subject to a linear analysis. The non-linear part of the problem may appear in the form of circuit parameters or frequency shifts which may be left for empirical determination. Such problems are exemplified by mixers, modulators, or switches.

The subsequent discussion will emphasize mainly the linear problems where the methods of circuit analysis are most effective, but will touch on some of the other fields occasionally.

THE TYPE A TRANSISTOR

Perhaps at this point is the place to pay our respects to the physics of the transistor. A view of the Type A transistor³, currently being made in small quantities, is shown in Fig. 1. It is about $\frac{1}{2}$ inch long and $\frac{3}{16}$ inch in diameter. Two small phosphor bronze "cat-whiskers" make point contacts close together to a block of germanium. A large area ohmic contact to the

³ "Type A Transistor," R. M. Ryder, *Bell Laboratories Record*, March 1949, pp. 89-93.

germanium constitutes the third electrode, called the base. How it works is shown in a purely descriptive way in Fig. 2. One point, called the collector,

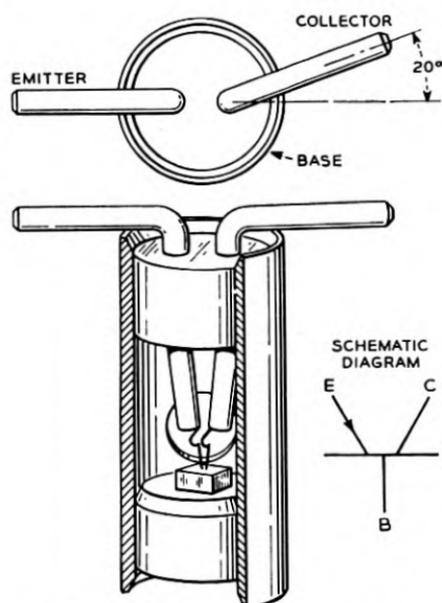


Fig. 1—Cutaway view of transistor.

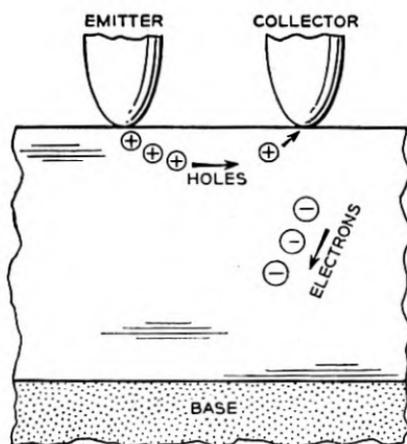


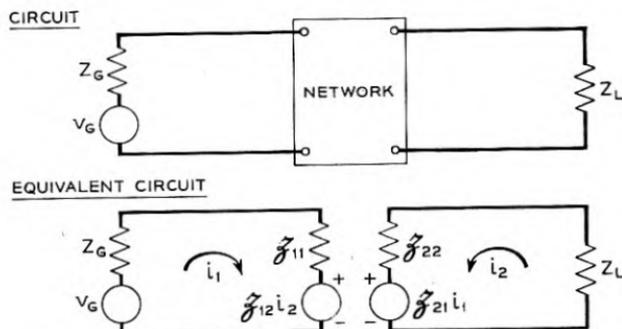
Fig. 2—Transistor mechanism.

is a rectifier biased strongly in the low-conducting direction. It therefore has a rectifying barrier in the germanium near it, which causes the collector impedance to be high. However, the collector can be influenced by the

emitter if the latter is arranged to emit anomalous charge carriers, that is, carriers of the sign not normally present in the interior of the material.

EQUIVALENT CIRCUITS

As has been explained by Bardeen, Brattain, and Shockley, many features of the transistor are nicely explained by this picture of its action; but, for present purposes of circuit analysis, we shall now take the purely empirical



$$\begin{aligned} \text{Equations} \quad & i_1(Z_G + \mathfrak{J}_{11}) + i_2\mathfrak{J}_{12} = v_G \\ & i_1\mathfrak{J}_{21} + i_2(\mathfrak{J}_{22} + Z_L) = 0 \end{aligned}$$

$$\text{Circuit determinant} \quad \Delta = (\mathfrak{J}_{11} + Z_G)(\mathfrak{J}_{22} + Z_L) - \mathfrak{J}_{12}\mathfrak{J}_{21}$$

$$\text{Input impedance} \quad Z_{11} = \mathfrak{J}_{11} - \frac{\mathfrak{J}_{12}\mathfrak{J}_{21}}{\mathfrak{J}_{22} + Z_L}$$

$$\text{Output impedance} \quad Z_{22} = \mathfrak{J}_{22} - \frac{\mathfrak{J}_{12}\mathfrak{J}_{21}}{\mathfrak{J}_{11} + Z_G}$$

$$\text{Operating power gain} \quad G_0 = 4R_G R_L \left| \frac{-\mathfrak{J}_{21}}{\Delta} \right|^2$$

$$\text{Insertion power gain} \quad G_1 = \left| \frac{(Z_G + Z_L)\mathfrak{J}_{21}}{\Delta} \right|^2$$

Fig. 3—Synopsis of general four-pole—impedance analysis.

view and regard the transistor as a *black box* whose performance is to be determined by electrical measurements on its terminals.

A picture of a black box is shown in Fig. 3 along with the equations describing it. The performance is completely characterized if one knows the voltage and current at each of the two pairs of terminals. Now, of these four variables, only two are independent since, if any two are fixed, the other two are determined. One can therefore describe the network in terms of any two variables and, since there are six possible ways to choose a pair of variables from a set of four, there are six ways of describing the network.

To recall what is done for electron tubes is helpful. In the case of a triode

the voltages on grid and plate are usually taken as independent variables; the grid and plate currents are taken as functions of the voltages. It becomes natural, then, to measure tubes with regulated power supplies having low impedances to keep the voltages constant, and one is then naturally led to describe tubes in terms of admittances. Now the trouble with this scheme for transistors is that many of them oscillate when connected to low impedances, that is, many transistors are short-circuit unstable. To avoid this difficulty it is convenient to measure with high impedances in the leads; the analytical counterpart is to regard the currents as independent variables, leading naturally to a description of the transistor in terms of impedances, as shown in the figure.

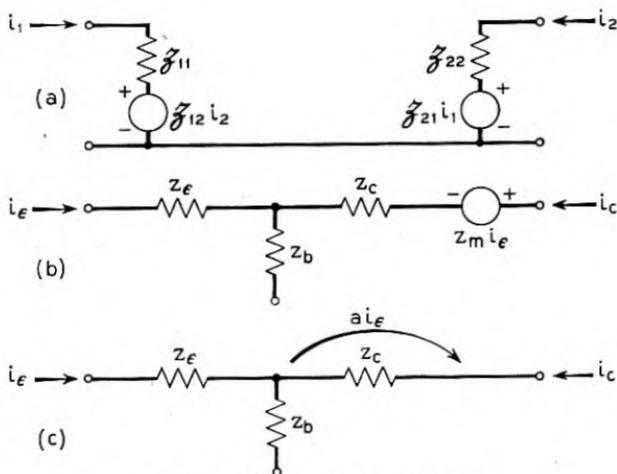
This description by open-circuit impedances happens to be a good one for many purposes, but there is nothing final or unique about it. In fact at high frequencies one of the other descriptions becomes more convenient.

By interpreting the \mathcal{Z} equations as circuit equations, one is led directly to the first equivalent circuit of Fig. 4. A little consideration shows why the \mathcal{Z} 's are called open-circuit impedances. For example, if the second mesh is open-circuited, then the equation says that \mathcal{Z}_{11} is the ratio of input voltage to input current, that is, the input open-circuit impedance; while \mathcal{Z}_{21} is the ratio of output voltage to input current, that is, the open-circuit forward transimpedance. Similarly \mathcal{Z}_{12} is the open-circuit feedback transimpedance and \mathcal{Z}_{22} is the open-circuit output impedance. Most of the subsequent discussion is concerned with low frequencies, where the impedances reduce to resistances.

This equivalent circuit for small signals is only one of many possibilities. Another, which is in fact more frequently used, is shown on Fig. 4. It consists of a T of resistors, each of which is associated with one of the transistor leads, and a voltage generator in series with the collector lead whose ratio to the emitter current is also of the dimensions of a resistance. The elements of this equivalent circuit are related to the former one by a simple subtraction. The other equivalent circuit on Fig. 4 is obtained by converting the series voltage generator to the equivalent shunt current generator, whose ratio to the emitter current is now a dimensionless constant which we shall call a .

These circuits, as well as all the other numerous possibilities, are equivalent in the sense that they all give exactly the same performance for any external connection of the unit. These three, however, are particularly well-behaved in that usually none of the circuit elements is negative; they are readily accessible to measurement; the association of the various circuit elements with corresponding regions within the transistor appears to have some physical significance; and, finally, the parameters are not too dreadfully dependent on the exact operating point used.

In the choice among various equivalent circuits, it appears that the optimum of convenience is also the one which most closely approaches the underlying physical situation. In agreeing to use the *black box* approach we have resolutely ignored the physical details, but here they are presenting themselves in a new way, having sneaked in the back door after we barred the front. Now, however, having chosen an equivalent circuit, we shall continue pursuing the circuit analysis in resolute ignorance of the physics. In what follows various equivalent circuits may be used, depending on the convenience of the moment.



Figs. 4—Some equivalent circuits.

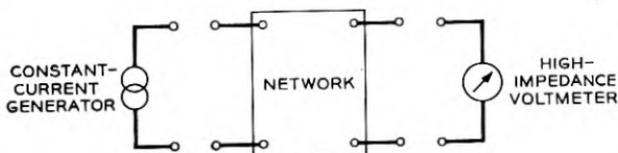


Fig. 5—Principle of measurement method.

The principle of a method used for rapid measurement of the transistor impedances is shown in Fig. 5, illustrating the measurement of forward transimpedance. A pair of terminals of the transistor is driven by a small alternating current of a few thousand cycles from a high impedance generator; the voltage developed is read by a high-impedance voltmeter. By calibrating the meter directly in ohms, one can read off the open circuit resistances of the unit as rapidly as one can switch and read meters.

Average values found by this method for the Type A transistor are shown on Fig. 6, together with data on the direct-current operating point. Since

development is still at an early stage, there are considerable variations between units.

SINGLE STAGE AMPLIFIERS. STABILITY, ELECTRON TUBE ANALOGY

An amplifier can be built in a straightforward manner by using the emitter as input electrode and collector as output electrode, the base being common to the two circuits. This amplifier is therefore called the grounded base amplifier. Figure 7 shows a schematic circuit using the average parameters just mentioned, working between 500 ohms and 20,000 ohms. The amplifier has an operating power gain of 17 db, power output Class A 10 milliwatts, noise figure at 1000 cycles 60 db with a variation inversely with frequency, and frequency response down 3 db at 5 megacycles.

Type A Transistor			
D.C. Operating Point:	$I_e = 0.6 \text{ ma}$	$V_e = 0.7V$	
	$I_c = -2 \text{ ma}$	$V_c = -40V$	
Circuit Parameters:	$r_e = 240 \text{ ohms}$	$r_b = 290 \text{ ohms}$	
	$r_c = 19000 \text{ ohms}$	$r_m = 34000 \text{ ohms}$	
	$\mathcal{Z}_{11} = 530 \text{ ohms}$	$\mathcal{Z}_{12} = 290 \text{ ohms}$	
	$\mathcal{Z}_{21} = 34000 \text{ ohms}$	$\mathcal{Z}_{22} = 19000 \text{ ohms}$	

Fig. 6—Equivalent circuit parameter values.

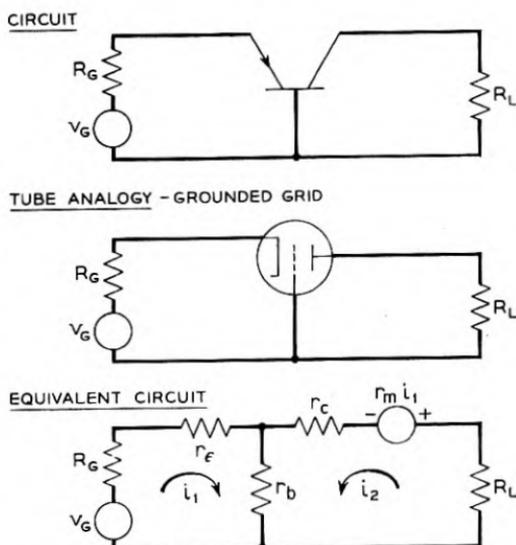
Some comments are in order on how this amplifier compares with an electron tube amplifier. First of all, the amplifying function and the manner of analyzing it from the circuit point of view are very similar, even though the internal mechanisms are markedly different. Secondly, there are qualitative differences in circuit behavior, which are set forth on Fig. 8. The base resistance r_b acts as a positive feedback element which, under adverse conditions, can cause the circuit to oscillate. A necessary condition for stability is that the circuit determinant shall be positive, and this can be written as follows:

$$\frac{r_m}{R_c} < 1 + \frac{R_E}{R_B} + \frac{R_E}{R_C} \quad (1)$$

Here the quantity r_m is the net mutual resistance of the transistor, and the capital R's are the total resistances in the corresponding leads, internal and external. One can see several features, as follows:

1. If $R_B = 0$, the circuit can be stable.
2. If $R_B > 0$, as usual, the circuit can be stable if the emitter and collector lead resistances are large enough or if r_m is not too large. In other words, resistance in the base lead tends toward instability if r_m is large; resistance in emitter or collector leads tends toward stability.

In the grounded base circuit the property of low base resistance is important, since the backward transmission depends directly on this property. In circuit terms, the base impedance is the feedback impedance in the grounded base circuit, and its value helps to set a limit on the stable gain which can be realized.



Equations:

$$i_1(R_G + r_e + r_b) + i_2 r_b = v_G$$

$$i_1(r_b + r_m) + i_2(r_b + r_c + R_L) = 0$$

Circuit determinant $\Delta = (R_G + r_e + r_b)(R_L + r_c + r_b) - r_b(r_b + r_m)$
 > 0 for stability

Input impedance $R_{11} = r_e + r_b - \frac{r_b(r_b + r_m)}{R_L + r_c + r_b}$

Output impedance $R_{22} = r_c + r_b - \frac{r_b(r_b + r_m)}{R_G + r_e + r_b}$

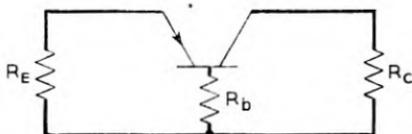
Operating power gain $G_0 = 4R_G R_L \left(\frac{-(r_b + r_m)}{\Delta} \right)^2$

Typical values: For $R_G = 500\omega$, $R_L = 20,000\omega$
 Then $R_{11} = 280\omega$, $R_{22} = 9600\omega$
 $G_0 = 17^{db}$

Fig. 7—Synopsis of grounded base amplifier.

The grounded base circuit has properties which are strongly reminiscent of the grounded grid electron triode amplifier in that both have low input impedance, high output impedance, and no change of signal polarity in transmission. The analogy was pointed out by Shockley. That this similarity is no coincidence can be seen by comparing the third equivalent circuit

above with the triode equivalent circuit of F. B. Llewellyn and L. C. Peterson⁴ in Fig. 9. Both circuits have the same topological form, and have similar impedance levels if the triode is considered to be operating in the frequency range of some tens of megacycles. The most important difference concerns the quantity a , a current amplification factor which, for the transistor, may be considerably greater than unity; while the analogous quantity



Can be stable if:

$$\frac{r_m}{R_c} < 1 + \frac{R_E}{R_b} + \frac{R_E}{R_c}$$

R 's include resistive elements both internal and external to the transistor.

Fig. 8—Stability

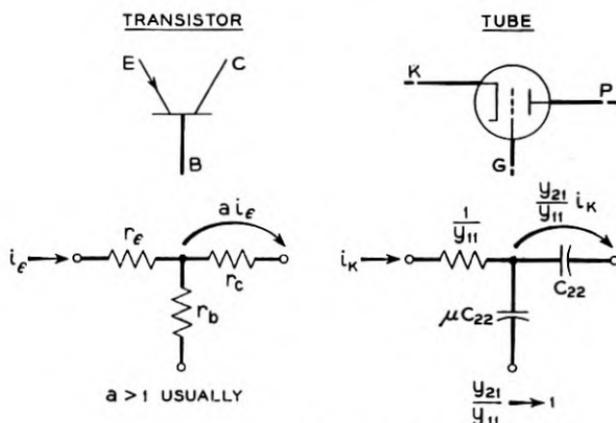


Fig. 9—Transistor-electron tube analogy.

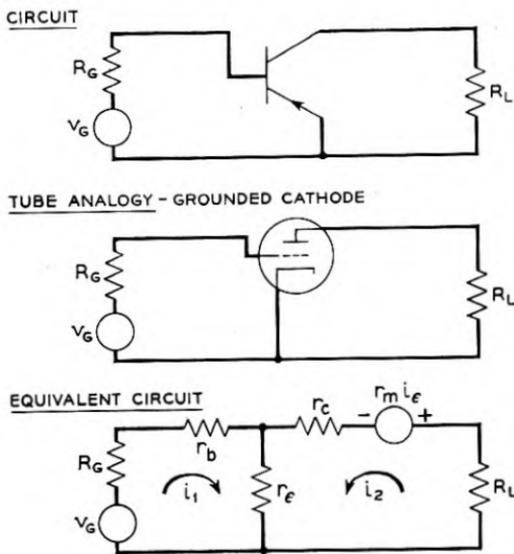
for the triode is close to unity for usual conditions. Another difference, of less importance, is the fact that the tube quantities analogous to r_c and r_b are capacitive reactances; their ratio, however, is like the ratio of r_c to r_b in magnitude.

One of the first consequences of this transistor-tube analogy is the suggestion that different transistor connections analogous to the different electron triode connections may be interesting.³ The analogy makes emitter analogous

⁴ "Vacuum Tube Networks," F. B. Llewellyn and L. C. Peterson, *Proc. I.R.E.*, March 1944, page 159, Fig. 13.

³ Loc. cit.

to cathode, base to grid, and collector to plate; the conventional or grounded cathode tube connection is therefore analogous to the grounded emitter connection of a transistor, shown on Fig. 10. It is found that when $a = 1$ the analogy is fairly close, in that the transistor has comparatively high-



Equations:

$$i_1(R_G + r_b + r_e) + i_2 r_e = v_G$$

$$i_1(r_e - r_m) + i_2(R_L + r_e + r_c - r_m) = 0$$

Circuit determinant:

$$\Delta = (R_G + r_b + r_e)(R_L + r_e + r_c - r_m) + r_e(r_m - r_e) > 0 \text{ for stability}$$

Input impedance

$$R_{11} = r_b + r_e + \frac{r_e(r_m - r_e)}{R_L + r_e + r_c - r_m}$$

Output impedance

$$R_{22} = r_c + r_e - r_m + \frac{r_e(r_m - r_e)}{R_G + r_b + r_e}$$

Operating Gain

$$G_F = 4 R_G R_L \left(\frac{r_m - r_e}{\Delta} \right)^2$$

Backward Operating Gain $G_R = 4 R_G R_L \left(\frac{r_e}{\Delta} \right)^2$

Typical values: For $R_G = 500^\omega$, $R_L = 20000^\omega$. Then $R_{11} = 2100^\omega$, $R_{22} = -6900^\omega$, $G_F = 24^{db}$, $G_R = -19^{db}$

Fig. 10—Synopsis of grounded emitter amplifier.

input impedance, high-output impedance, and changes signal polarity in transmission. When $a > 1$, as is usual, the analogy becomes less close, and feedback effects tend to become large and obnoxious; the open-circuit output impedance is usually negative. This behavior is readily under-

standable from stability considerations, since the base lead is now one of the signal terminals and, as before mentioned, putting resistance in the base lead tends toward instability if a is enough greater than unity. The effect is so severe that often it is worth while to add resistance in the collector lead, thereby reducing a to the neighborhood of unity, and simultaneously reducing the amplifier to a state of greater tractability.

Another feature of the grounded emitter amplifier is that the base resistance r_b is usually negligible, in contrast to its pronounced effect on the reverse transmission of the grounded base amplifier. The role of feedback element is taken over here by the emitter resistance r_e . These considerations have important effects on the properties of cascaded amplifiers and will be reverted to later.

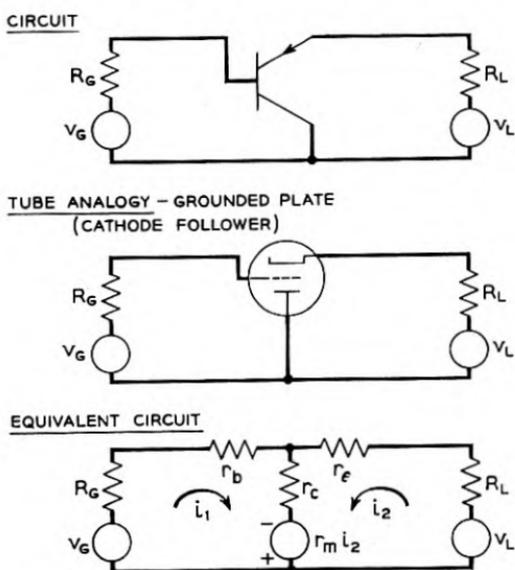
For numerical comparison we might work the grounded emitter amplifier between the same two terminations as the grounded base amplifier above, namely from 500 into 20,000 ohms. It would then have a gain of about 24 db, an improvement of 7 db over the grounded base, with about the same power output and noise figure. This improvement is obtained at greater risk of oscillation; in fact the output impedance of this amplifier is negative.

The remaining tube connection — the cathode follower or grounded plate — is analogous to the grounded collector connection (Fig. 11); again, when $a = 1$ the analogy is fairly close, in that the transistor has high-input impedance, low-output impedance, and no change of polarity in transmission. In fact when $a = 1$ the device is usable in very much the same manner as the cathode follower. The power output is lower than the other connections because the output electrode (the emitter) does not carry much direct current.

However, when we make a greater than 1 the effect is even more pronounced than it was in the grounded emitter case. As a increases from 1, the grounded collector amplifier rapidly loses its resemblance to the cathode follower and begins to transmit in both directions as a bilateral element. When $a = 2$, the operating gains in the two directions are the same; and for $a > 2$ the transmission is actually greater in the "backward" direction. Another curious feature is that, while the "forward" transmission is still without change in signal polarity, the "reverse" transmission inverts the signal polarity.

In any device which is supposed to give gain in both directions, naturally stability must be a controlling consideration. This amplifier is of course still subject to the aforementioned stability condition (1) and it is found that with care one can actually get power gains in both directions of transmission without instability, i.e. a simple bilateral amplifier is present. One numerical example may suffice. Assume a transistor having the properties

$r_e = 250$ ohms, $r_b = 250$ ohms, $r_c = 20,000$ ohms, $r_m = 40,000$ ohms, so that $a = 2$ and both base and emitter resistances r_e and r_b are negligible.



Equations:

$$i_1(R_G + r_b + r_e) + i_2(r_c - r_m) = v_G$$

$$i_1 r_c + i_2(R_L + r_e + r_c - r_m) = v_L$$

Circuit determinant

$$\Delta = (R_G + r_b + r_e)(R_L + r_e + r_c - r_m) + r_c(r_m - r_e) > 0 \text{ for stability}$$

Input impedance

$$R_{11} = r_b + r_e + \frac{r_c(r_m - r_e)}{R_L + r_e + r_c - r_m}$$

Output impedance

$$R_{22} = r_e + r_c - r_m + \frac{r_c(r_m - r_e)}{R_G + r_b + r_e}$$

Operating Gain

$$G_F = 4 R_G R_L \left(\frac{-r_c}{\Delta} \right)^2$$

Backward Operating Gain

$$G_R = 4 R_G R_L \left(\frac{-r_c + r_m}{\Delta} \right)^2 = (1 - a)^2 G_F$$

Typical values:

$$\text{For } R_G = 20000^\omega, R_L = 10000^\omega$$

$$\text{Then } R_{11} = -4100^\omega$$

$$R_{22} = -7600^\omega$$

$$G_F = 15^{db}$$

$$G_R = 13^{db}$$

Fig. 11—Synopsis of grounded collector amplifier

Working between 20,000-ohm terminations, such an amplifier should have 6 db power gain in both directions and should still be stable even if one of its terminations changes 50% in the unfavorable direction.

The grounded emitter connection can also exhibit bilateral properties.

Recapitulating these three single-stage amplifiers, we see that when $a = 1$ their properties are close enough to the analogous electron tube arrangements to be easily remembered; but that, when a is different from 1, their properties begin to diverge from their tube counterparts. Some of these circuits will perform in a simple manner functions which are impossible to the analogous tube connections, although of course the functions could be accomplished by using more tubes or more complicated circuits.

FREQUENCY RESPONSE

So far the analysis of transistors has been given only for the resistive case, appropriate at low frequencies. When the frequency is raised, reactive components appear and the situation becomes more complicated, although of course still subject to the same general methods of analysis.

One might expect that since semiconducting diodes work at microwave frequencies, so also would semiconducting triodes. For the Type A transistor, this hope is blasted because of the essentially different nature of the mechanism, involving as it does the physical transport of charge carriers over appreciable distances. For certain features of the transistor, however, the analogy does hold. For example, the emitter by itself is a diode; and, in keeping with this fact, its open-circuit impedance does not change much with frequency in the range in which we shall be interested. For most engineering purposes the open-circuit input impedance of a Type A transistor may be regarded as a resistance independent of frequency. Such deviations as occur are small and entirely similar to what take place in an analogous diode.

The same situation holds with respect to the base resistance r_b and the collector resistance r_c , that is, they act as one might expect of a diode. The base resistance is substantially constant with frequency; the collector resistance has associated with it a slight amount of capacitance, mostly due to the case, leads, and wiring external to the unit, which gives a variation of properties with frequency in high-impedance circuits. The analogous capacitance on the emitter side is negligible because of the lower value of emitter impedance. One has, therefore, the T of resistors in the equivalent circuit substantially constant with frequency.

The dominant factor governing frequency response of the transistor is therefore largely expressed as a variation of the net mutual impedance r_m or, one may say as well, in the factor a which is the ratio of r_m to r_c .

Measurements of r_m as a function of frequency encounter the practical difficulty that it is impossible to present to the transistor over a wide frequency range an impedance high compared to the collector impedance. It is, however, quite easy to present to the collector a relatively low impedance

(75 ohms), which is constant over the frequency range of interest. Concurrently it is relatively simple to present to the emitter a high impedance,

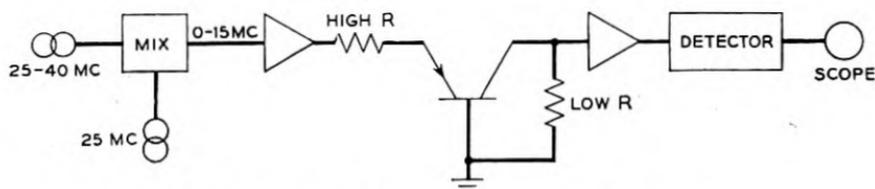


Fig. 12—Sweeper for measuring frequency response.

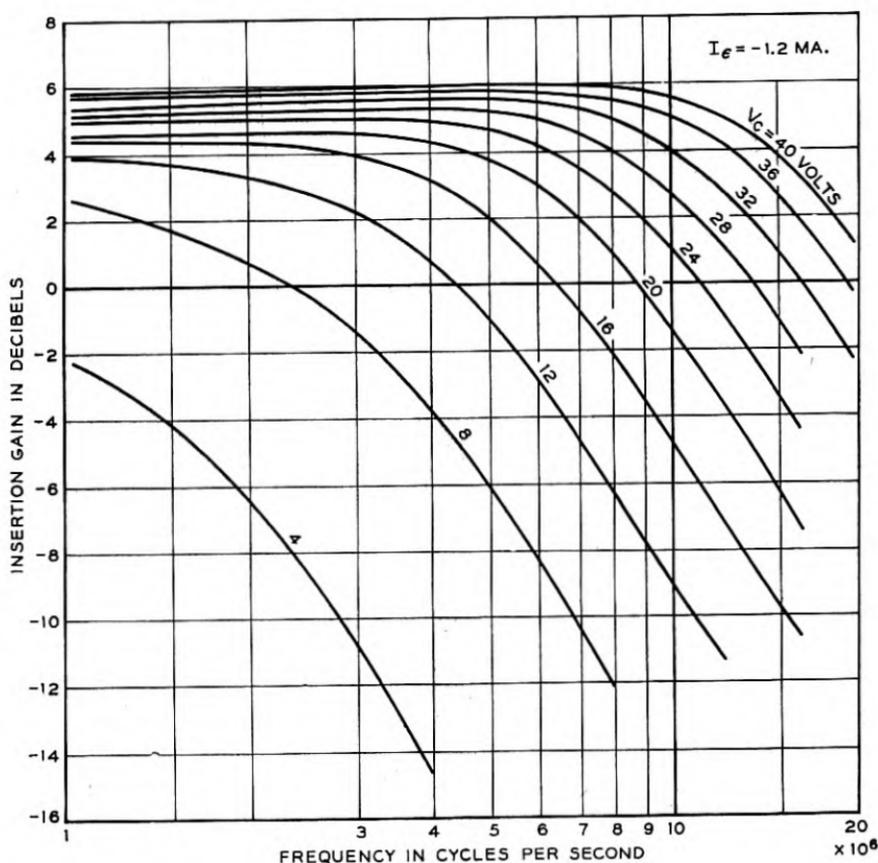


Fig. 13—Alpha versus frequency.

that is, to drive it with a constant current generator. Under these conditions the insertion power gain of the transistor is approximately α^2 , where the current amplification factor α is the ratio of increment in collector current to

increment in emitter current at constant collector voltage.⁵ The quantities α and a are usually nearly the same.

An oscilloscopic presentation of α versus frequency is possible and is a great convenience since many units can be measured quickly and variation with operating point observed directly. The sweep frequency generator built for this purpose is diagrammed in Fig. 12. It presents on an oscilloscope the magnitude of α as a function of frequency from 0 to 15 megacycles. Means are also available for making point-by-point plots which are more accurate, though much slower.

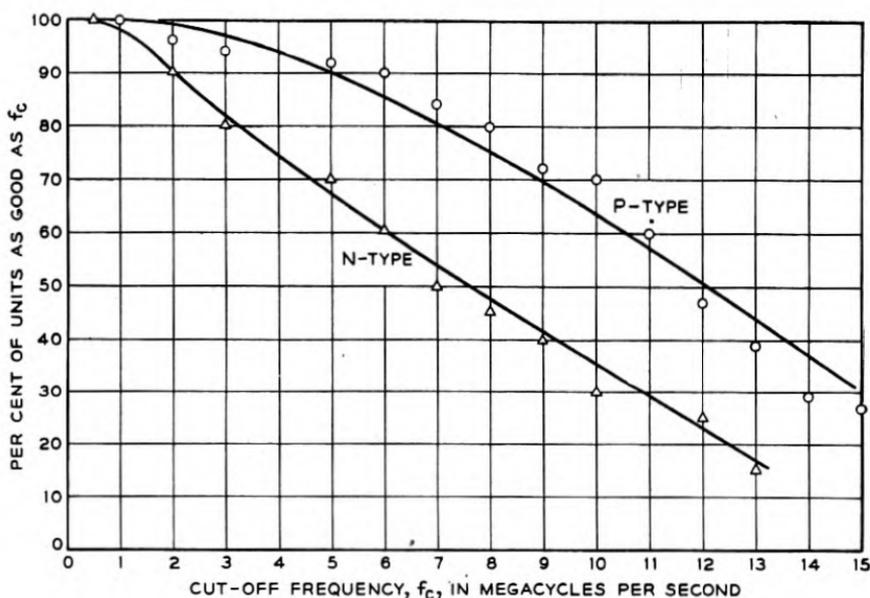


Fig. 14—Cut-off frequency statistics.

A set of curves of current amplification factor α versus frequency, as obtained with this apparatus, is shown in Fig. 13. The cutoff shape is a little sharper than that of a single R-C circuit but less so than that of a pair, one of which is shunt-peaked enough to make the combination flat. The apparent high-frequency asymptote varies in different units from 7 to 11 db per octave.

The phase shift associated with this curve has been found to be related to the amplitude in the same way as if the characteristic were that of a "mini-

⁵ Actually, $\alpha = (\partial I_e / \partial I_b)_{V_c}$ is only one of a set of four circuit parameters h_{ij} whose relationship to I_e and V_c is the same as that of the Z's to I_e and I_c , and which furnish an alternative circuit representation of the transistor. The other three h's can be measured in a similar manner but are of less interest.

imum phase" passive circuit.⁸ Accordingly the phase shift, like the amplitude variation, is also intermediate between a single R-C interstage and the flat compensated pair of interstages.

When variations between curve shapes are not too large, the shape can be characterized by a single parameter which we take as the cutoff frequency f_c . Cutoff is defined as the frequency where the magnitude of α^2 is halved. Some statistical data on cutoff frequency of different units made of N-type and P-type germanium are plotted in Fig. 14. The P-material is somewhat

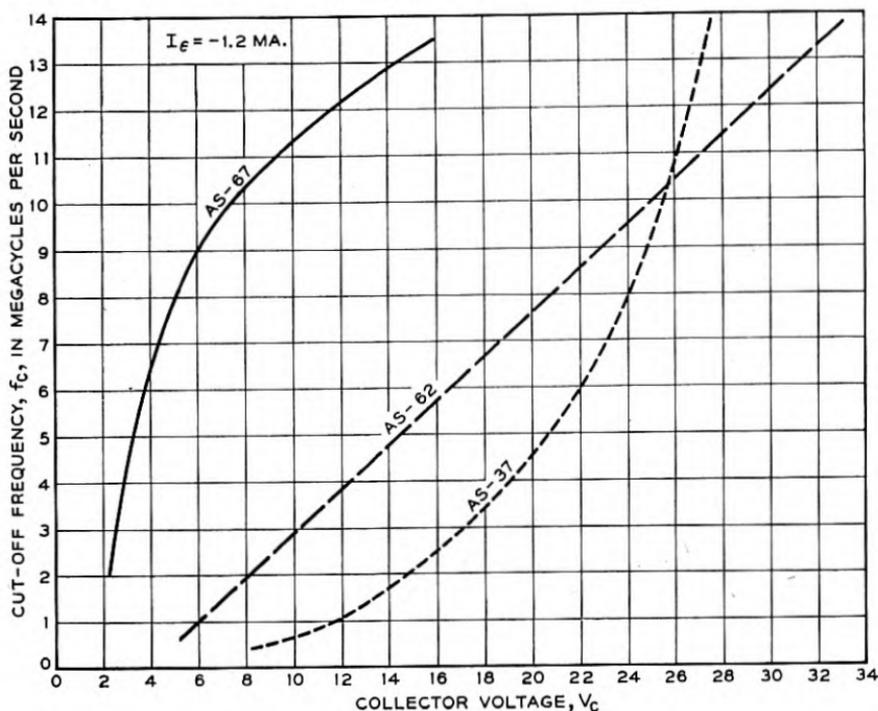


Fig. 15—Cut-off frequency versus collector voltage.

better, in keeping with the fact that the active charge carriers producing the transistor effect in it are electrons having greater mobility than the holes which are active in N-type germanium.

As one changes the operating point of the transistor the frequency response curve changes in such a way that the shape remains sensibly constant on a logarithmic frequency scale, but the scale changes. The cutoff frequency is usually roughly proportional to the collector voltage, with only minor dependence on the other operating parameter, as shown in Fig. 15 unit AS62.

⁸ "Network Analysis and Feedback Amplifier Design," H. W. Bode, D. Van Nostrand Publishing Co., 1945.

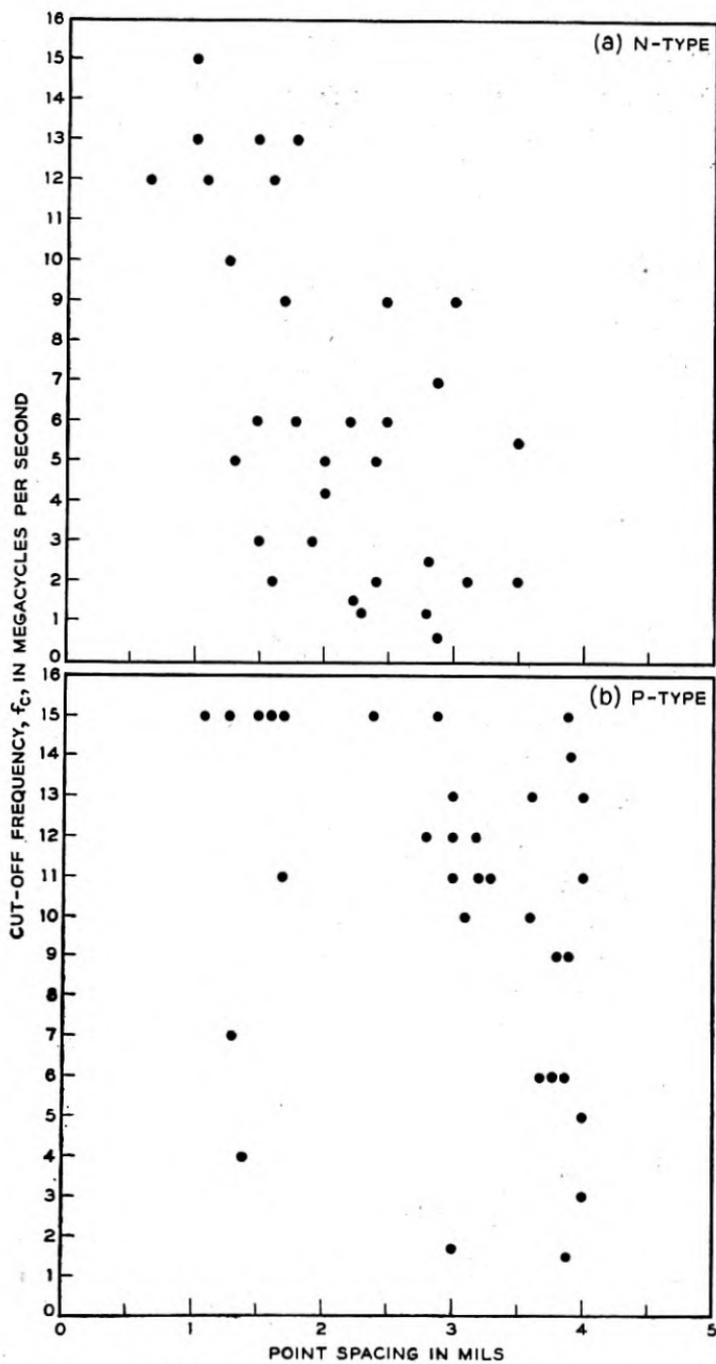


Fig. 16—Cut-off frequency versus point spacing.

Other types of variations of cutoff frequency with collector voltage are exhibited by some transistors.

That frequency cutoff is affected by the spacing between points of the transistor is shown in Fig. 16, which gives some support to the idea that the cutoff frequency might vary inversely as point spacing, other things being equal. However, one has only to look at the graph to see that other things are not equal for, at any given point spacing, the cutoff frequencies of different units vary by almost an order of magnitude. It is, however, clear that point spacing is one of the important factors.

In recapitulation of the measurements of frequency behavior, it appears possible to build Type A transistors with frequency cutoffs well above 10 megacycles. At the present time, the factors determining the frequency behavior are not yet under good control.

CASCADE AMPLIFIERS

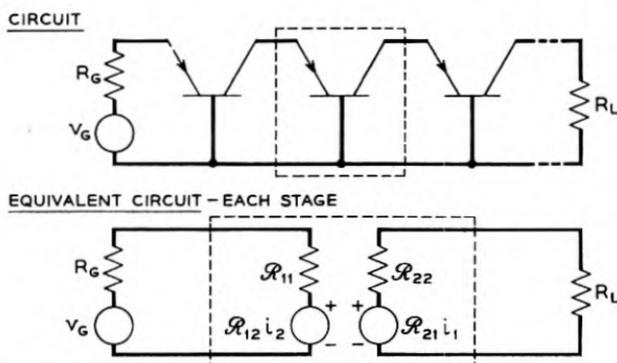
Many cascading possibilities exist, since any connection of the transistor might be used in combination with other connections, as well as involving all the parameter variations which might be made on each single stage. Some of the more elementary possibilities will be mentioned. Since feedback in each unit greatly complicates the situation, the essential features of the amplifiers may become clearer by discussing an idealized case where feedback is absent or greatly reduced. For similar reasons, the preliminary discussion is confined to frequencies low enough so that the equivalent circuits are purely resistive.

Perhaps the most straightforward cascade amplifier is the iterated grounded-base cascade, outlined in Fig. 17. Neglecting feedback, the insertion power gain is nearly equal to the current amplification factor α squared. For the Type A transistor this amounts to some 5 db per stage. For most uses this could be regarded as impractically low, but it might be pointed out that the tube analog (grounded grid cascade) is even worse; for when $\alpha = 1$ the maximum insertion gain is 0 db per stage. Both amplifiers of course can be made practical by interstage transformers (Fig. 18). For the Type A transistor, the matched gain without feedback rises to about 15 db per stage, which still compares favorably in magnitude with most grounded-grid tubes.

When feedback is considered by allowing r_b to return to its usual value of a few hundred ohms, the question of stability becomes important. The nominal Type A transistor is still stable when the cascade interstages are matched, the gain rising to about 21 db per stage. For many units having more than the usual amount of feedback, the interstages cannot be matched without violating the stability condition and therefore encountering os-

cillations; but one can normally count on stable gains of 15 to 20 db per stage, the transformers being perhaps somewhat mismatched.

Interesting possibilities for a good cascade amplifier with more gain than the grounded base cascade are offered by the grounded emitter connection. Incidentally, this gain advantage is also enjoyed by the grounded cathode or conventional tube connection, so that one would expect it to apply here from the electron tube analogy; but in transistors the feature that α may be greater than 1 brings in complications having no simple analogy for tubes.



Without feed back ($\mathcal{R}_{12} = 0$):

Iterative impedance $R_G = \mathcal{R}_{22}$, $R_L = \mathcal{R}_{11}$

Circuit determinant $\Delta = (\mathcal{R}_{11} + \mathcal{R}_{22})^2$

$$\text{Insertion Power Gain } G_I = \left| \frac{-R_{21}}{\mathcal{R}_{11} + \mathcal{R}_{22}} \right|^2$$

$$= \left(\frac{\alpha}{1 + \mathcal{R}_{11}/\mathcal{R}_{22}} \right)^2$$

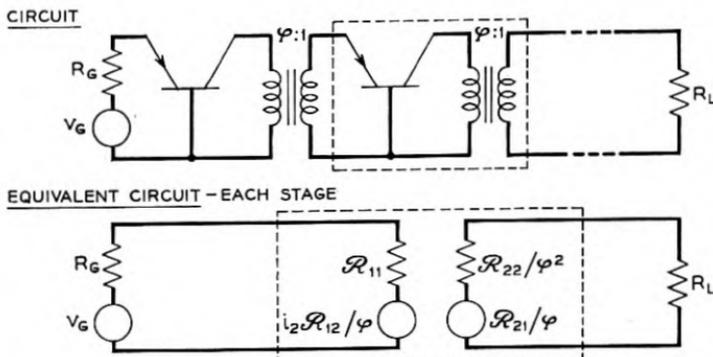
Nominal Type A Gain = 5^{db}

Fig. 17—Synopsis of grounded base cascade.

The iterated grounded emitter cascade without feedback (that is, emitter resistance $r_e = 0$) is unstable for the nominal Type A transistor, but can be stabilized in many ways of which we shall mention only one. The equivalent circuit of Fig. 19 shows an added resistor which may be thought of as adjusting the value of the collector resistance, and tends to make the unit more stable. When this resistor is adjusted to make the total collector resistance R_c about equal to the net mutual resistance r_m , thus reducing the effective value of α to the neighborhood of unity, then the cascade amplifier becomes stable, its gain being sensitive to the exact value chosen for the adjusting resistor. A numerical calculation for the grounded emitter

amplifier using the nominal Type A transistor adjusted in this way gives the following results:

Assuming an adjusted value of collector resistance of 36000 ohms to be satisfactory for stability, then the iterative input impedance is 2300 ohms, output impedance 4000 ohms, and insertion gain about 21 db per stage without transformers. Three-stage stable amplifiers having power gains of about 55 db have been operated.



Without feed back ($R_{12} = 0$):

Iterative impedance $R_G = R_{22}/\varphi^2$, $R_L = R_{11}$

Circuit determinant $\Delta = (R_{11} + R_{22}/\varphi^2)^2$

Insertion Power Gain $G_I = \left(\frac{-R_{21}/\varphi}{R_{11} + R_{22}/\varphi^2} \right)^2$

Maximum when $R_{11} = R_{22}/\varphi^2$

$$G_{I \text{ max.}} = \frac{R_{21}}{4 R_{11} R_{22}} = \frac{1}{4} \alpha^2 \frac{R_{22}}{R_{11}}$$

Nominal Type A Gain: without feed back = 15^{db}
with R_{12} normal = 17^{db}

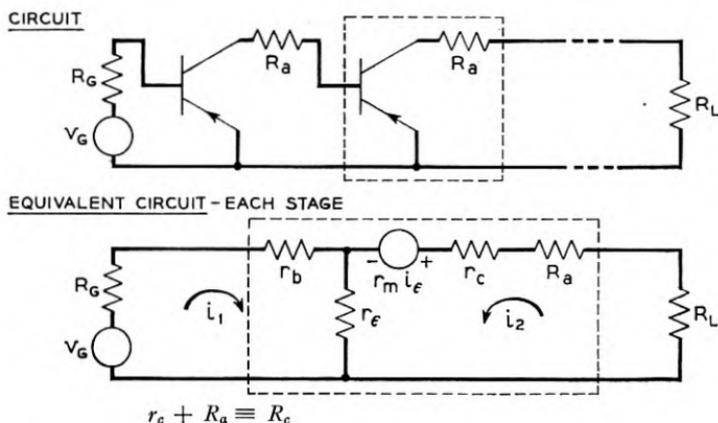
Fig. 18—Synopsis of grounded base cascade with transformers.

Another interesting feature of the grounded emitter amplifier is the ease with which negative feedback may be applied to it. A resistor inserted in the emitter lead gives local negative feedback analogous to the cathode feedback of tubes, while feedback involving several stages is also obtainable by common-lead methods analogous to common-cathode resistors familiar in the tube art. By such means, as is well known, distortion instability or gain variation may be reduced, or power output increased.

Theoretical study of these and other iterative amplifiers, particularly at higher frequencies, is conveniently carried on with the aid of the formulas

of Figs. 20 and 21 which give some of the iterative properties of a general fourpole and the effect thereon of an interstage matching transformer.

The iterative method of course does not exhaust the possibilities of cascade amplifiers. They can also be designed stage by stage. Even when feedback is large they can be cascaded together in the manner used for filter sections. A particular design of this sort is shown in Fig. 22. It is a grounded



Equations:

$$\begin{aligned} i_1(R_G + r_b + r_e) + i_2 r_e &= v_G \\ i_1(r_e - r_m) + i_2(R_L + r_e + R_c - r_m) &= 0 \end{aligned}$$

Circuit determinant $\Delta = \begin{vmatrix} R_G + r_b + r_e & r_e \\ -r_e(r_e - r_m) & R_L + r_e + R_c - r_m \end{vmatrix}$
 > 0 for stability

Without feed back $(r_e = 0)$

Iterative impedance $R_G = R_c - r_m, R_L = r_b$

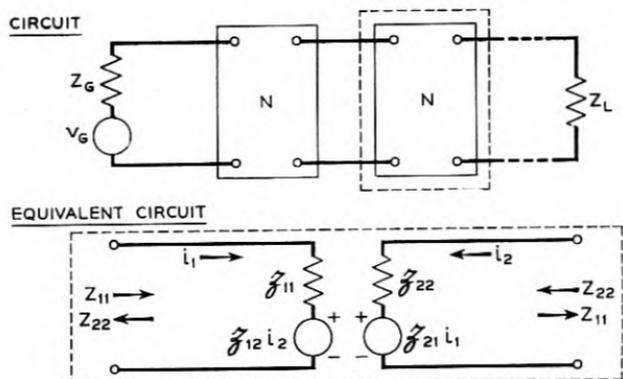
Circuit determinant $\Delta = (r_b + R_c - r_m)^2$

Insertion Power Gain $G_I = \left(\frac{r_m}{r_b + R_c - r_m} \right)^2$

Nominal Type A Gain with $R_c = 36000^\omega$
 without feed back 23^{db}
 with r_e normal 21^{db}

Fig. 19—Synopsis of grounded emitter cascade.

base stage followed by a grounded collector and accordingly has the tube analog grounded-grid, cathode follower, from which one would expect that the terminating impedances would be low and the interstage impedance high. This amplifier matched a 600-ohm line to better than 10% and had 16 db insertion gain, with a bandwidth of about a megacycle. An adaptation for video purposes was made to obtain over a band from 100 cycles to 3.5 megacycles, an insertion gain of 20 db in a 75-ohm coaxial line.



Equations:

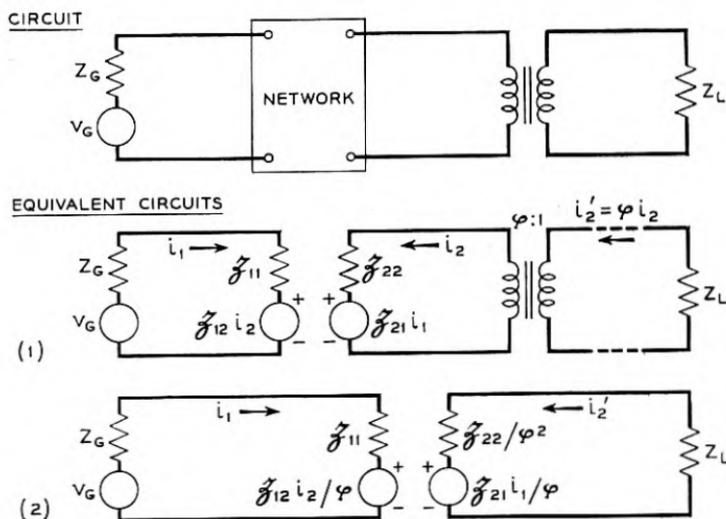
$$\begin{aligned} i_1(\mathcal{D}_{11} + Z_{22}) + i_2\mathcal{D}_{12} &= v_G \\ i_1\mathcal{D}_{21} + i_2(\mathcal{D}_{22} + Z_{11}) &= 0 \end{aligned}$$

Terminations:

$$\begin{aligned} Z_{11} = Z_L &= -\mathcal{D}_{22} + \frac{1}{2}(\mathcal{D}_{11} + \mathcal{D}_{22})(1 + \sqrt{1-y}) \\ Z_{22} = Z_G &= -\mathcal{D}_{11} + \frac{1}{2}(\mathcal{D}_{11} + \mathcal{D}_{22})(1 + \sqrt{1-y}) \\ y &= 4\mathcal{D}_{12}\mathcal{D}_{21}/(\mathcal{D}_{11} + \mathcal{D}_{22})^2 \end{aligned}$$

Circuit determinant $\Delta = \frac{1}{2}(\mathcal{D}_{11} + \mathcal{D}_{22})^2(1 - y + \sqrt{1 - y})$ Insertion Power Gain $G_I = \left| \frac{\mathcal{D}_{21}}{\mathcal{D}_{11} + \mathcal{D}_{22}} \frac{2}{1 + \sqrt{1 - y}} \right|^2$

Fig. 20—Synopsis of iterated cascade of four-poles.

Equations: $i_1(\mathcal{D}_{11} + Z_G) + i_2'\mathcal{D}_{12}/\varphi = v_G$

$$i_1\mathcal{D}_{21}/\varphi + i_2'\left(\frac{\mathcal{D}_{22}}{\varphi^2} + Z_L\right) = 0$$

Fig. 21—Four-pole with ideal transformer.

The foregoing amplifiers both have rather low output powers because of the fact that the emitter, a low-current electrode, is the output electrode. A way of improving this situation has been suggested in the second amplifier schematic shown in Fig. 22. The first stage is a grounded emitter and the second a grounded collector transistor, the latter operating in what we have called the "backward" direction so that the output electrode is the base and the power level is improved. This amplifier can be stabilized by negative feedback obtainable by inserting a resistor in the first stage emitter lead.

These examples emphasize that one can cascade unlike stages and that feedback can be used to stabilize performance, just as with electron tubes. These amplifiers can be further cascaded to obtain more gain. Other possibilities worthy of mention include modifying the design of the first stage

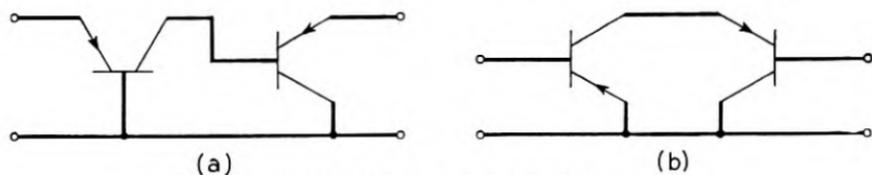


Fig. 22—Non-iterative cascade amplifiers.

of an iterative amplifier to obtain good noise figure, or of the last stage for greater power output.

BAND PASS AMPLIFIERS

Bandpass amplifiers require a few remarks before concluding the small-signal discussion. The design within the band may be carried out by the methods previously discussed; but frequently attention must also be paid to properties outside the band, to an extent unusual with tubes. The reason, of course, is connected with that Dr. Jekyll and Mr. Hyde of transistors, α (or a) greater than 1. When a transistor may be short-circuit unstable, then oscillations may result from the practise usual with electron tube amplifiers of letting the impedances outside the band fall to low values. For the same reason design of power leads requires more care than usual. The problems encountered are somewhat similar to those of tube amplifiers with feedback in that one must pay attention to characteristics far outside the useful band. In the case of transistors, one may have to exercise design care to avoid oscillations even when the gain of the amplifier is less than unity.

LARGE SIGNAL ANALYSIS

Large signals are those which involve considerable excursions over the electrical characteristics of the device and cannot be regarded as small

changes near an assumed operating point. For their general study a most convenient tool is provided by the set of static characteristics of the unit.

Since most analyses begin with the static characteristics, perhaps some excuse is needed for the unorthodox approach which has delayed them to this point. Two reasons may be cited: First, the small-signal behavior is in a sense simpler, being capable of discussion by the familiar linear methods of circuit theory. Second, the small-signal behavior has brought out some features, notably short-circuit instability, which have a bearing on certain features of the static characteristics, on the methods of measuring them, and on the particular manner of expressing them.

A set of characteristics representative of Type A transistor performance is shown in Fig. 23, consisting of four plots, one of each of the electrode voltages against each of the currents with the other current as parameter. Contrary to electron tube practise, rather than the voltages we take the currents as the independent variables. This choice avoids the experimental difficulty that the short-circuit unstable transistors might oscillate if we were to attempt to hold the electrode voltages constant, as well as the concomitant analytical trouble that in that case the voltage-dependent characteristics become double-valued.

The relationship of these characteristics to the open-circuit impedances is direct and quickly shown. Suppose the voltages are expressed formally as functions of the currents:

$$\begin{aligned} V_e &= f_1 (I_e, I_c) \\ V_c &= f_2 (I_e, I_c) \end{aligned} \quad (2)$$

Differentiating, and identifying the differentials as small-signal variables, we get immediately the equations for the open-circuit resistances:

$$\begin{aligned} v_e &= i_e \frac{\partial f_1}{\partial I_e} + i_c \frac{\partial f_1}{\partial I_c} \\ v_c &= i_e \frac{\partial f_2}{\partial I_e} + i_c \frac{\partial f_2}{\partial I_c} \end{aligned} \quad (3)$$

Accordingly, the open-circuit resistances are the slopes of these static characteristics. The reactive components do not appear because our assumptions (2) were not sufficiently general to take them into account or, in other words, the reactive information is not contained in the static characteristics.

Just as there are five other pairs of small signal parameters which could have been chosen, so there are five other ways in which the static characteristics could have been expressed. Often these other ways are convenient for special purposes or are closely connected with particular large signal circuits.

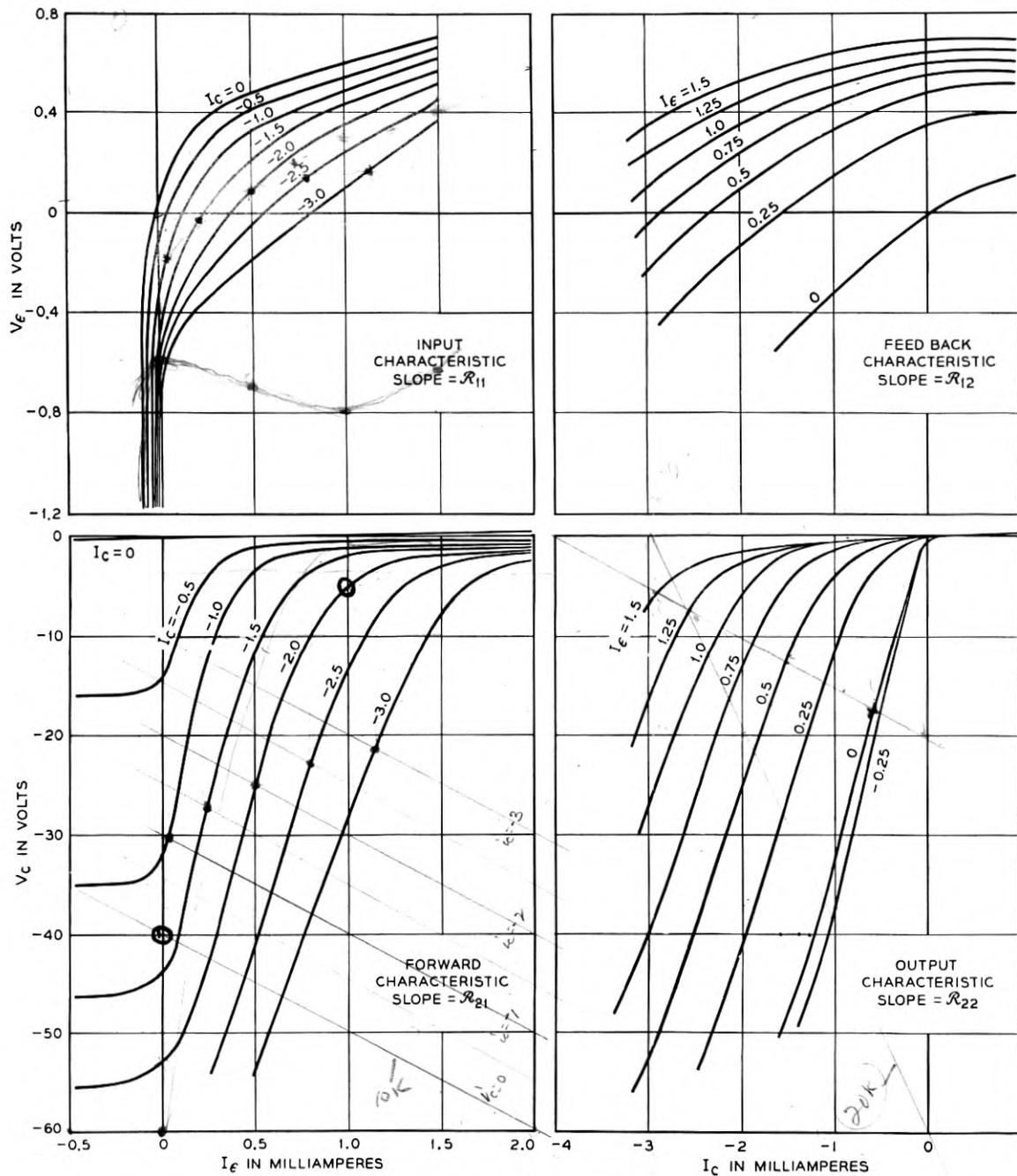


Fig. 23—Static characteristics.

Measurement of the characteristics can be by conventional point-by-point plots or by oscilloscope presentation. An oscilloscopic curve tracer has been built which can show any of the four characteristics for any of the six pairs of independent parameters of the Type A transistor, as well as any two-pole characteristic which might be of interest (such as a negative resistance characteristic).

Occasionally the static characteristics are affected by effects of a thermal nature such that an oscilloscope trace does not give the same results as a slow point-by-point plot. These thermal effects are small in the usual region of operation of the Type A transistor but may become appreciable if the unit is heated by excessive power dissipation in it.

POWER OUTPUT AND DISTORTION

The problem of obtaining good "undistorted" power output from a transistor at low frequencies is one which is conveniently discussed by means of the static characteristics. Analytically this question belongs to the class of slightly non-linear problems but, for descriptive purposes, it is illustrated by the curves of Fig. 24. The family of collector characteristics of a Type A transistor is shown. The region of linear operation is substantially that part of the plot where the curves are uniformly spaced, have constant slope, and lie within the permitted power dissipation of the unit.

In driving a Type A transistor harder and harder in an attempt to get greater power output, one may encounter four types of overload distortion, analogous to the types found in tubes.

1. One may drive the emitter negative into the cutoff region where the collector current fails to respond to changes in emitter potential, corresponding to grid cut-off in a tube.

2. One may drive the emitter positive into an emitter overload region where non-linear distortion may be encountered because the emitter impedance changes with its voltage. The corresponding tube phenomenon is positive grid distortion. For both tubes and transistors this effect is a minor one which may be actually beneficial in practical cases.

3. The collector may be driven down to low potential where it can no longer draw the current required to follow the impressed emitter current variations. This distortion corresponds to plate "bottoming" in electron tubes.

4. The collector may be driven up to high currents where it overloads because of the non-linear voltage response in that region arising from heating effects. This effect has practical consequences something like the overloading of electron tubes which may arise from insufficient cathode emission.

In other words, either emitter or collector may be driven into overload or cut-off and the problem of getting good power output reduces to choosing

an operating point and load impedance such as to avoid these non-linear effects as long as possible. Reverting to Fig. 24, since one wants as large a product of $\Delta V \cdot \Delta I$ as possible, the problem may be thought of in geometrical terms as approximately that of constructing the largest possible rectangle such that a load line extending diagonally across the corners of this rectangle

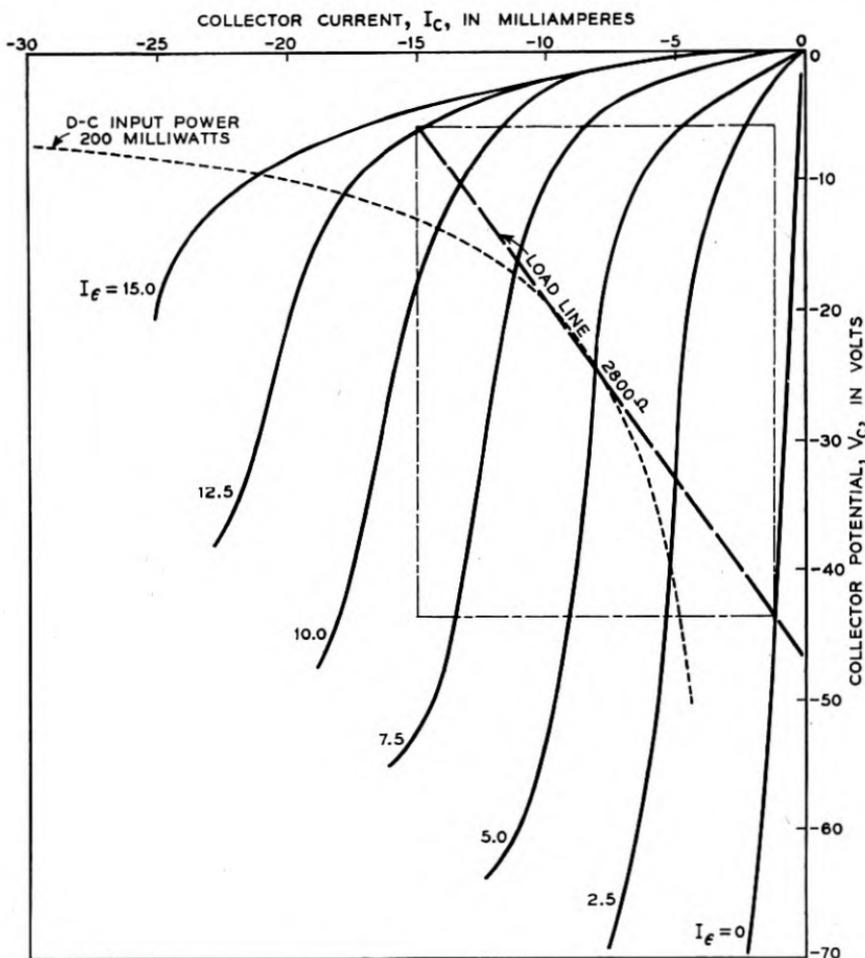


Fig. 24—Collector power output plot.

lies within the "linear" region of operation. The slope of this line gives the load impedance required, its intercept the collector supply voltage (for resistance coupling), and the sides of the rectangle give the extreme values of voltage and current. The center of the rectangle is approximately the quiescent or small-signal operating point.

Under optimum conditions of load impedance and operating point,

one obtains power efficiencies comparable to Class A electron tube operation, that is, 20 to 35% efficiency with a few percent harmonic distortion. As contrasted to recommendations for good low-level gain for the Type A transistor, the optimum conditions for power output have usually involved lower load impedances and higher currents. Representative values may be: load impedance, 5000 ohms; collector current, -8 milliamperes at -35 volts bias; emitter current, 3 milliamperes; power output, 60 milliwatts, with distortion less than ten percent.

One complication of the power transistor is that, when the optimum load impedance is low, the operating point gets nearer to the region where the transistor may tend to oscillate if it happens to be one of the kind which is short-circuit unstable. A saving circumstance here is available in that

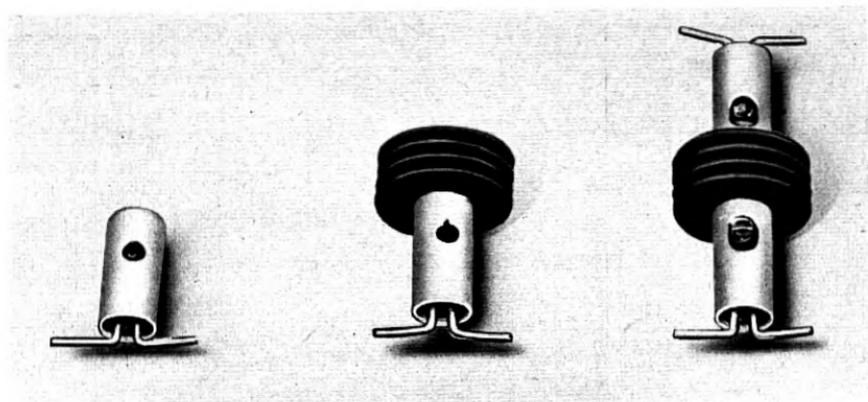


Fig. 25—Some power transistors.

added resistance in the emitter lead tends to promote stability, so that the transistor may be stabilized by operating out of a higher generator impedance, possibly at some cost in reduced gain. A corollary aspect of the same phenomenon is that the input impedance of a high-power transistor may become very low or even negative.

Higher power output from the transistor can also be obtained by increasing the permissible collector dissipation. This has been accomplished by using a thin wafer of germanium directly soldered to a copper base equipped with suitable fins to facilitate the removal of heat generated in the vicinity of the collector point. An increase in allowable dissipation from 200 to 600 milliwatts has been thereby obtained. Output powers of approximately 200 milliwatts at a conversion efficiency of 33% have been realized.

The photograph of Fig. 25 shows on the left the type A transistor, in the center the power version of this unit, and on the right is shown a double

ended type of power transistor using two germanium wafers with a common radiator for push-pull applications.

OTHER LARGE-SIGNAL APPLICATIONS

The static characteristics can be used for calculations of many large-signal circuits of which only a few examples can be given here. The first is a tickler feedback oscillator of Fig. 26, which uses the grounded-base circuit with a resonant circuit in the collector lead, transformer-coupled back to the emitter.

Other circuits making use of the special possibilities of the transistor include an oscillator with anti-resonant circuit in the base lead, or with a

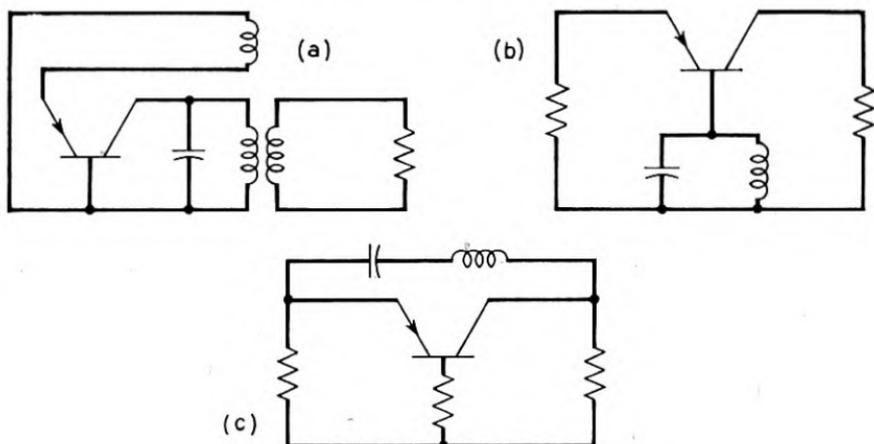


Fig. 26—Transistor oscillators.

series resonant circuit from collector to emitter. Some of these circuits make use of the short-circuit instability peculiar to the transistor and accordingly would not work with electron tubes.

NOISE

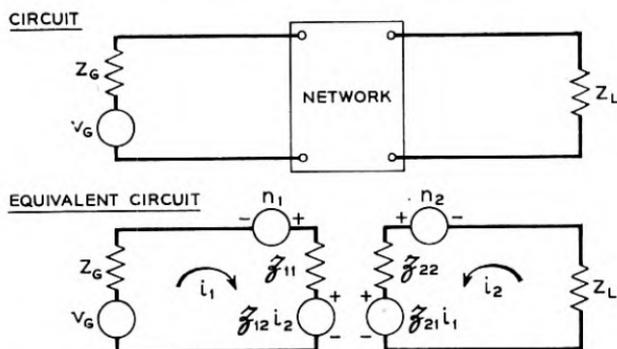
A discussion of small-signal amplifiers would be incomplete without some mention of the limiting factor of noise. The noise has been left to the last, however, because its discussion complicates the circuits slightly, and perhaps because it is not well to present too early an aspect of performance which is at the moment so much inferior to electron tubes.

On the circuit representation of noise as well as signal much work has been done by L. C. Peterson.⁷ It turns out that in the general four-terminal network in which we are interested, a complete noise representation for

⁷ "Signal and Noise in Microwave Tetrode," *Proc. I.R.E.*, Nov. 1947, pp. 1264-1272.

circuit purposes may be obtained by adding two noise generators to the equivalent circuit of four signal parameters, as shown in Fig. 27.

These noise representations are on an entirely similar basis to the signal representations. Just as four elements in any independent configuration suffice for signal description, so two noise generators in either series or shunt in any convenient independent locations can be added to account for the noise. All these representations give the same signal and noise behavior for any external connections. Still, some may be better than others in corresponding to the actual physics of the transistor; presumably the



$$\begin{aligned} \text{Equations: } i_1(Z_G + \mathfrak{Z}_{11}) + i_2\mathfrak{Z}_{12} &= v_G \oplus N_1 \\ i_1\mathfrak{Z}_{21} + i_2(\mathfrak{Z}_{22} + Z_L) &= \oplus N_2 \end{aligned}$$

Circled \oplus signs indicate addition with attention to any correlations which may exist between noise generators or mean square additions if no correlation exists.

$$\text{Noise Figure } F = 1 + \frac{1}{4 kTBR_G} \left\{ \overline{N_1^2} \oplus \overline{N_2^2} \left(\frac{\mathfrak{Z}_{11} + Z_G}{\mathfrak{Z}_{21}} \right)^2 \right\}$$

Fig. 27—Synopsis of general four-pole, including noise.

better representations will show particularly simple behavior, for example, in their dependence upon the d-c operating point of the transistor. The usual choice puts noise voltage generators in series with the emitter and collector leads, as shown.

If the two noise generators were truly independent physical sources of noise, their outputs would be expected to show no correlation and their noise power contributions would be simply additive. This independence is not usually the case for the Type A transistor. By adding the noise outputs and comparing the power in the sum to that in the separate components, correlation coefficients ranging from $-.8$ to $+.4$ have been found. From this the conclusion can be drawn that the physical sources of noise in the network do not act in series with the leads but at least to some extent arise elsewhere

in the transistor and contribute correlated noise output to both the generators of the circuit representation.

The transistor noise is of two types. One is a rushing sound somewhat similar qualitatively to thermal resistance noise; the other is a frying or rough sound which occurs erratically, usually in the noisier units. The noise

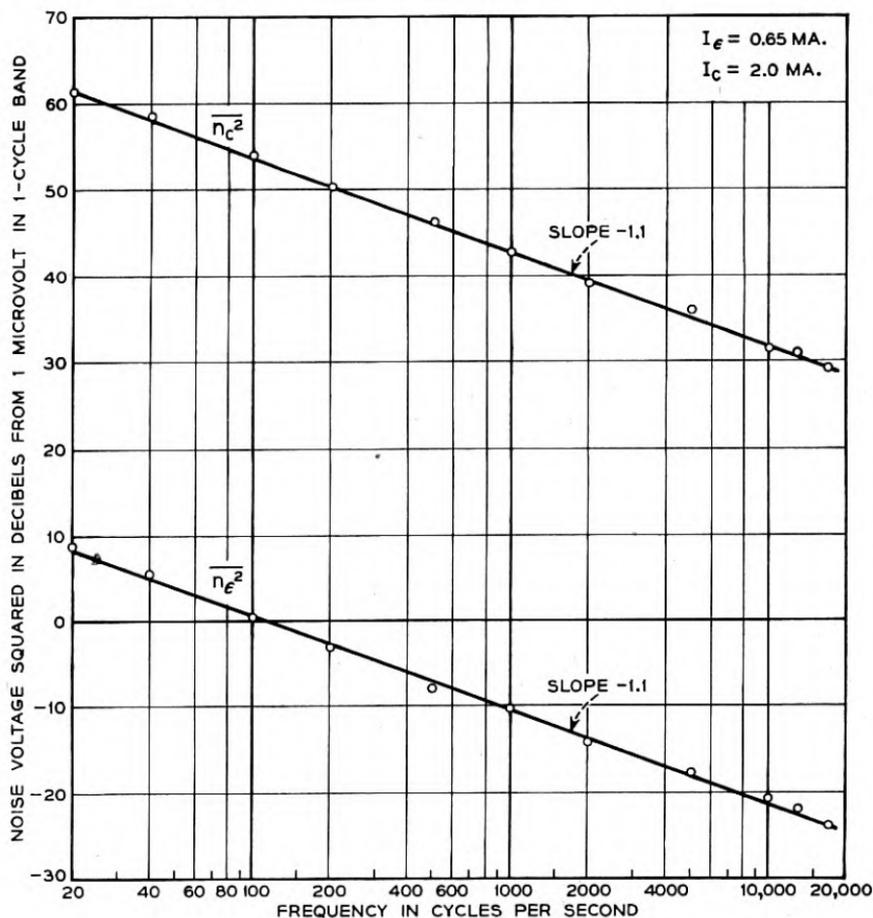


Fig. 28—Transistor noise versus frequency.

power per unit bandwidth varies almost exactly inversely with frequency as shown in Fig. 28, being in this respect reminiscent of contact noise.

Since the noise dependence on frequency is known, its level may be given as noise voltage per unit bandwidth at a reference frequency (1000 cycles). The collector noise usually dominates as far as practical effects on the output are concerned. Representative values are about 100 microvolts per cycle at 1000 cycles for the collector, and one or two microvolts for the emitter.

The noise voltages depend mainly on the collector direct voltage as shown in Fig. 29. While they do vary with the other operating parameter at constant collector voltage, such variations rarely exceed 10 db, which is much less than the variations with collector voltage.

More important than the actual level of the noise is its relation to thermal resistance noise, which is the ultimate limit to amplification. This relationship is conveniently expressed by means of the noise figure, or number of times noisier than amplified thermal noise in the output of the amplifier.

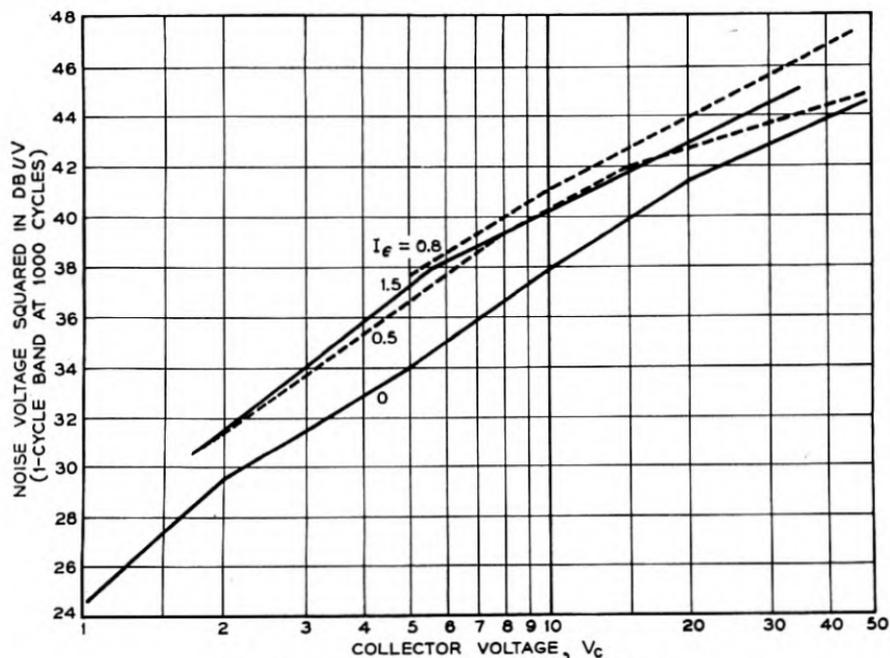


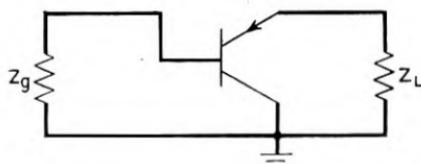
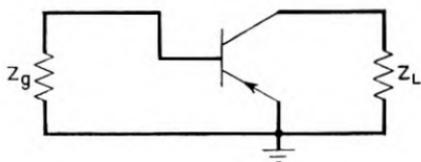
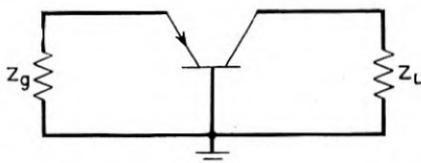
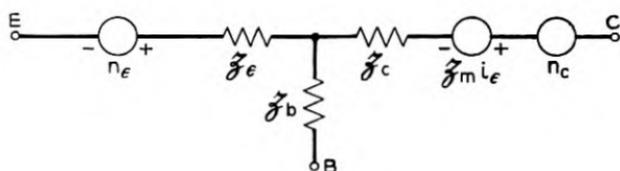
Fig. 29—Transistor noise versus operating point.

A representative noise figure for the Type A transistor at 1000 cycles is 60 db, with individual units ranging from 50 to 70 db.

Noise figure formulas for the three single-stage connections are given in Fig. 30. The noise performance of the three connections would usually not be very different if it were not for stability considerations, which may render unusable the generator impedance which would give optimum performance. Mainly, on account of stability, the grounded base connection may be said to give the best noise performance, with the grounded emitter running a close second.

The noise figure of any device depends upon the generator impedance out of which it works but does not depend upon the load. Accordingly, there exists an optimum generator impedance which gives the best noise

Equivalent Circuit



Grounded Base

$$F = 1 + \frac{1}{4kTBR_G} \left[\overline{N_\epsilon^2} \oplus \overline{N_c^2} \left(\frac{Z_G + z_\epsilon + z_b}{z_m + z_b} \right)^2 \right]$$

Grounded Emitter

$$F = 1 + \frac{1}{4kTBR_G} \left[\overline{N_\epsilon^2} \left(\frac{Z_G + z_m + z_b}{z_m - z_\epsilon} \right)^2 \oplus \overline{N_c^2} \left(\frac{Z_G + z_b + z_\epsilon}{z_m - z_\epsilon} \right)^2 \right]$$

Grounded Collector

$$\text{Forward } F = 1 + \frac{1}{4kTBR_G} \left[\overline{N_\epsilon^2} \left(\frac{Z_G + z_\epsilon + z_b}{z_c} \right)^2 \oplus \overline{N_c^2} \left(\frac{Z_G + z_b}{z_c} \right)^2 \right]$$

$$\text{Backward } F = 1 + \frac{1}{4kTBR_L} \left[\overline{N_\epsilon^2} \oplus \overline{N_c^2} \left(\frac{Z_L + z_\epsilon}{z_m - z_c} \right)^2 \right]$$

Fig. 30—Noise figure formulas.

figure of which the unit is capable. This optimum source impedance is best for signal-to-noise performance, not for signal performance alone; hence, as is well known for vacuum tubes, it is usually not a match for the unit, and in general both the resistive and reactive components of impedance may be mismatched to the unit.

For the transistor at low frequencies in the grounded-base connection, reactive effects are negligible and the emitter noise generator may usually be neglected. Under these conditions the optimum noise figure is obtained from a generator of impedance equal to the open-circuit input resistance of the transistor (not the actual working input resistance, which may be quite different).

The best operating point for low noise is usually obtained at a moderate collector voltage (20 volts) and a small emitter current (0.5 ma.).

SUMMARY

A tentative evaluation of the Type A transistor may be made on the basis of presently available information. Before making it, we should say that a comparison with the field of electron tubes is obviously unfair — there are many against one, and a little one at that. Furthermore the little one is a baby not only in size but in length of time under development. It is only natural that the full possibilities are not yet apparent. With these reservations, we can make the following statements about the present Type A transistor:

Gain: the transistor figure of about 17 db per stage is somewhat low compared to 30 or 40 db obtainable from audio tubes. When the bandwidth is taken into consideration the gain-band product of the transistor is good but, since the excess bandwidth cannot be exchanged for gain, this number is in this case illusory for narrow-band amplifiers. For video amplifiers the comparison is more favorable.

Stability considerations differ from the electron tube in such a way as to be likely to give more trouble at low frequencies. At video frequencies this difference is less marked if we play fair by comparing with a triode tube instead of a pentode. The latter is of course better shielded than the transistor.

Frequency response appears to be practical up to 10 megacycles or more.

Power output efficiency of around 30%, Class A, seems fully comparable to an electron tube, so that a comparison between the two can be based on input d-c power.

Noise figure of 60 db at 1000 cycles is much worse than that of a good electron tube, which can come close to 0 db. In view of the frequency dependence which brings the transistor noise figure down to 30 db at a megacycle, the comparison at video frequencies is less unfavorable, particularly if some developmental improvement can be made.

So far on most counts the comparison is not too favorable but, as we said before, it isn't fair to the baby. In addition there are a number of other considerations which are secondary from the point of view of pure technique but may be dominant from other points of view. Among favorable factors here are: small size; low power drain; no standby power, but instant response when needed; low heating effect when used in large numbers; and ruggedness.

The life of transistors should be fairly long on the basis of diode performance, but the device is too new to permit definite statement. The mechanical simplicity might well lead one to hope for low cost, but no production figures are as yet available.

In fine, even if Type A transistor performance does not excel all electron tubes, it is still good enough for many applications and will be considerably better in the future.

ACKNOWLEDGEMENT

This survey is based on the work of many people, only a few of whom have been mentioned in the text. The examples of circuits have not been numerous or exhaustive, but rather have been used to illustrate the methods adopted; these are general enough to be adapted to the solution of many particular problems.

Theory of Transient Phenomena in the Transport of Holes in an Excess Semiconductor

By CONYERS HERRING

An analysis is given of the transient behavior of the density of holes n_h in an excess semiconductor as a function of time t and of position x with respect to the electrode from which they are being injected. When the geometry is one-dimensional, an exact solution for the function $n_h(x, t)$ can be constructed, provided certain simplifying assumptions are fulfilled, of which the most important are that there be no appreciable trapping of holes or electrons and that diffusion be negligible. An attempt is made to estimate the range of conditions over which the neglect of diffusion will be justified. A few applications of the theory to possible experiments are discussed.

A variety of experiments have been performed, and others are planned, which involve measurement of transient or steady-state phenomena due to the drift of positive holes along a specimen of n -type semiconductor after they have been introduced at an *injection electrode* or *emitter*.¹ These phenomena are presumably a result of the interplay of drift, space-charge, recombination, and diffusion effects. This paper seeks to relate these effects to the phenomena, and its principal contribution is an explicit calculation of the transient phenomena outside the range of small-signal theory, for cases where the geometry is one-dimensional and where certain simplifying assumptions, notably the neglect of diffusion, are justified. Removal of some of these simplifying assumptions and a more careful development of the theory will be necessary in certain applications.

Section 1 discusses the physical assumptions and boundary conditions involved in setting the problem up. Section 2 contains calculations of the distribution of holes along the length of the semiconductor at various times, for the mathematically simplest case where recombination and diffusion are ignored and all currents are held constant after the start of the injection. This simple case illustrates the method of attack to be used in the more general calculations of Section 4, and it is hoped that this sketching of basic ideas will enable the hasty reader to pass on to Section 6 without going

¹ Experiments of this sort have been undertaken with the objective of testing and extending the theoretical interpretation of transistor action proposed by J. Bardeen and W. H. Brattain, *Phys. Rev.*, 75, 1208 (1949), especially as regards the role of volume transport of holes, a role first suggested by J. N. Shive, *Phys. Rev.*, 75, 689 (1949). Examples of the type of experiment discussed in the present paper have been described by: J. R. Haynes and W. Shockley, *Phys. Rev.*, 75, 691 (1949) (transient effects); W. Shockley, G. L. Pearson, M. Sparks and W. H. Brattain, in a paper presented at the Cambridge Meeting of the American Physical Society, June 16-18, 1949 (steady-state transport); W. Shockley, G. L. Pearson, and J. R. Haynes, *Bell Sys. Tech. Jour.*, this issue (steady-state and transient effects).

through the mathematical details of Sections 3, 4, and 5. Section 3 contains the complete differential equations of the problem, including diffusion and recombination, and Section 4 gives the solution when only the diffusion terms are neglected. Section 5 contains some order-of-magnitude estimates regarding diffusion effects. Section 6 summarizes the capabilities of the theory so far developed, presents some obvious generalizations, and discusses an interesting *shock wave* phenomenon which occurs whenever the injected hole current is quickly decreased.

1. BASIC ASSUMPTIONS AND BOUNDARY CONDITIONS

Consider the n -type semiconducting specimen shown in Fig. 1, having electrodes at its two ends, $x = -a$ and $x = b$, respectively, and an injection electrode system at $x = 0$ somewhere in between. Let a current of density j_a per unit area enter at the left-hand end, and let a current of density j_e be injected at $x = 0$. To make the problem strictly one-dimensional, it will be

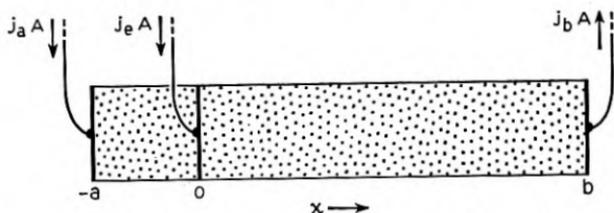


Fig. 1—Idealized experiment on hole transport in one dimension.

supposed that this injection takes place uniformly over the plane cross-section of the specimen at $x = 0$, instead of taking place at isolated points of the surface, as is usually the case in experiments. This idealization will presumably be justified if the thickness of the specimen is small compared with lengths in the x -direction which are significant in the experiment and if the injected positive holes are able to spread themselves uniformly over the cross-section before appreciable recombination has taken place.

Unless otherwise stated, it will be supposed that j_e consists entirely of positive holes, i.e., that the number of electrons withdrawn from the specimen by the electrode at $x = 0$ is negligible compared with the number of holes injected. The currents j_a and j_e need not be constant in time, although most of the analysis to be given below will assume them constant after the time of initiation of j_e .

One can set up differential equations for the variation with x and time of the electron density, n_e , and the hole density, n_h . These equations will in the general case involve migration due to electrostatic fields, diffusion, recombination, trapping, and thermal release of electrons and holes from traps. It will be assumed, however, that trapping and thermal release from traps can be neglected, or, more precisely stated, that creation of mobile

holes and electrons occurs only at the electrodes, and that the disappearance of mobile holes and electrons is caused only by mechanisms which cause holes and electrons to disappear in equal numbers at essentially the same time and place. If this assumption is valid, the charge density due to impurity centers will never differ from its equilibrium value by an amount comparable with the density due to free electrons. This assumption can be expected to be reasonably good for an n -type impurity semiconductor in which the number of donor levels is very much greater than the number of acceptor levels and for which, at the operating temperature, practically all the donor levels have been thermally ionized, while thermal excitation of electrons from the normally full band has not yet become appreciable.

As has just been mentioned, the differential equations for the behavior of the electron and hole densities involve migration under the influence of the local electric field $E(x, t)$. This field is in turn influenced by the space charge due to any inequality between the hole density n_h and the electron excess $(n_e - n_0)$, where n_0 is the normal electron density. If the difference $(n_h - n_e + n_0)$ were comparable with n_h or n_e , the problem would be very complicated. Fortunately, however, this difference cannot have an appreciable value over an appreciable range of x , on the scale of typical experiments. For example, if $(n_h - n_e + n_0)$ were 10^{-2} of n_0 for a range Δx of 1μ , and if n_0 is 10^{15} cm^{-3} , then the difference in field strength on the two sides of Δx would be about 2000v/cm , a field which would outweigh all other fields in the problem and rapidly neutralize the space charge. Moreover, the time required for the evening out of any such abnormally high space charge would be very short, of the order of magnitude of the resistivity of the specimen expressed in absolute electrostatic units ($1 \text{ sec.} = 9 \times 10^{11} \Omega \text{ cm}$). Thus it will be quite legitimate to assume $(n_h - n_e + n_0) = 0$ in all equations of the problem except Poisson's equation which determines the field E , and so n_e can be eliminated from the conduction-diffusion equations for holes and electrons. These two equations can then be used, as is shown below, to determine the two unknown functions n_h and E , Poisson's equation being discarded as unnecessary.

The boundary conditions for these differential equations consist of two parts, the conditions at $t = 0$ and those at and to the left of $x = 0$. In most of the applications to be considered, the injection current j_e will be assumed to commence at $t = 0$. Thus, initially, the specimen will be free of holes and, at $t = 0^+$, will have a field $E_a = j_a/\sigma_0$ in the region $-a < x < 0$, and a field $E_0 = j_b/\sigma_0$ in the region $0 < x < b$, where σ_0 is the normal conductivity of the specimen and $j_b = j_a + j_e$ is the total current density to the right of $x = 0$. The boundary condition at $x = 0$ is determined by the magnitudes of the electronic and hole contributions to the injection current j_e . If no electrons are withdrawn by the electrode at $x = 0$, then the electron currents just to the left and just to the right of $x = 0$ must be equal, and the

hole current densities on the two sides must differ by j_e ; if a part of j_e is due to withdrawal of electrons, then the electronic current will have a corresponding discontinuity. If j_a is positive, i.e., flows from left to right in the specimen, the current can be assumed to be practically entirely electronic over most of the range from $-a$ to 0; i.e., as x becomes negative the hole current must rapidly approach zero and the electron current must rapidly approach j_a . In fact, if diffusion is ignored the electron and hole currents must have these limiting values for any negative x .

The preceding discussion and the mathematics to follow have been couched in purely one-dimensional language, i.e., have been formulated as if the electron and hole densities were functions of x alone, independent of y and z , and as if the semiconductor extended to infinity in the y - and z -directions. However, it is easy to see at each stage that practically the same equations can be written for transport of holes along a narrow filament whose thickness is small compared with the linear scale of the phenomena along its length, even when the density of holes is not uniform over the cross-section of the filament. If the density of holes is uniform over the cross-section, all the equations will of course hold as written. However, recent work² has suggested that holes recombine with electrons so rapidly at the surface that the density of holes may be much smaller near the surface than in the center of the cross-section. In such case all the equations of this memorandum must be interpreted as applying to the mean value, $\bar{n}_h(x)$, of the density of holes, $n_h(x, y, z)$, averaged over the cross-section of the filament; also, the rate of recombination of holes and electrons must be set equal to some function of \bar{n}_h , as yet not reliably known, instead of to a constant times the product of electron and hole densities. This will of course alter most of the quantitative predictions of Section 4, but will not require any change in the method of calculation.

2. FORMULATION AND SOLUTION OF THE PROBLEM WITH NEGLECT OF DIFFUSION AND RECOMBINATION

For this case the electron and hole currents can each be equated to the product of field strength E by particle density n by mobility μ , and the continuity equations are

$$\frac{\partial n_h}{\partial t} = -\frac{\partial}{\partial x} (E\mu_h n_h) \quad (1)$$

$$\frac{\partial n_e}{\partial t} = \frac{\partial}{\partial x} (E\mu_e n_e). \quad (2)$$

² H. Suhl and W. Shockley, paper Q11 presented at the Washington Meeting of the American Physical Society, April 29, 1949; see also Shockley, Pearson, Sparks and Bratman, reference 1.

Since the neutrality condition requires $\frac{\partial n_h}{\partial t} = \frac{\partial n_e}{\partial t}$, subtracting (1) and (2) and integrating gives the equation of conservation of total current:

$$E(\mu_e n_e + \mu_h n_h) = j(t)/e = \text{const. indep. of } x$$

where of course $j = j_b = (j_a + j_e)$ when $0 < x < b$ and when conditions are such that all currents flow from left to right. Putting the neutrality condition $n_e = n_h + n_0$, into the equation gives the following relation between E and n_h :

$$E[(\mu_e + \mu_h)n_h + \mu_e n_0] = j/e \quad (3)$$

This can be used to eliminate either E or n_h from (1). If E is eliminated we have

$$\frac{\partial n_h}{\partial t} = -\frac{\mu_e \mu_h n_0 j}{e[(\mu_e + \mu_h)n_h + \mu_e n_0]^2} \frac{\partial n_h}{\partial x} = -V(n_h) \frac{\partial n_h}{\partial x} \quad (4)$$

where $V(n_h)$ is an abbreviation for the coefficient shown. If, instead, n_h is eliminated from (1) a similar equation results:

$$\frac{\partial E}{\partial t} = \frac{E}{j} \frac{dj}{dt} - V(E) \frac{\partial E}{\partial x} \quad (5)$$

where

$$V(E) = eE^2 \mu_h \mu_e n_0 / j = E \mu_h (E/E_0) \quad (6)$$

where

$$E_0 = j/\sigma_0 \quad (7)$$

i.e., the field necessary to maintain the total current by electronic conduction in the normal state of the specimen. The velocity $V(E)$ is of course numerically the same as the $V(n_h)$ occurring in (4) when E and n_h are related by (3).

The solution can be based on either (4) or (5). We shall use (4), as n_h is the most interesting quantity for direct measurement, and as the differential equation to be given below for the case where diffusion terms are included is simpler when n_h is chosen as the dependent variable.

Equation (4) (or (5)) describes a wave propagated with the variable velocity V . If $j_e \ll j_a$, so that E is never greatly different from E_0 , (4) (or (5)) and (6) indicate that n_h (or E) is propagated with the constant velocity $E_0 \mu_h$, as is of course to be expected. More interesting is the case where j_e and j_a are comparable, so that V departs significantly from constancy. It is tempting to suppose that, for this case also, the curve of n_h against x at any time t can be constructed by taking the graph of n_h against x at $t = 0$ and moving each point of the curve horizontally to the right a

distance $V(n_h)t$. One can, in fact, easily verify that this construction gives a solution of (4), by writing (4) in the form

$$\left(\frac{\partial \lambda}{\partial t}\right)_{n_h} = -\left(\frac{\partial n_h}{\partial t}\right)_x = V(n_h) \left(\frac{\partial n_h}{\partial x}\right)_t$$

whence it is obvious that the function $n_h(x, t)$ defined implicitly by

$$x(n_h, t) = x(n_h, 0) + V(n_h)t$$

satisfies (4) for any form of the arbitrary function $x(n_h, 0)$, and that, conversely, any solution of (4) must be of this form.

Assuming, as in the preceding, that all currents flow from left to right, the boundary conditions at $t = 0^+$ are:

$$n_h = 0 \text{ for } x < 0 \text{ and } x > 0 \quad (8)$$

or, equivalently,

$$\left. \begin{aligned} E &= E_a = j_a/\sigma_0 && \text{for } x < 0 \\ E &= E_0 = (j_a + j_e)/\sigma_0 && \text{for } x > 0 \end{aligned} \right\} \quad (9)$$

The boundary conditions at $x = 0$ are, for $t > 0$,

$$n_h = 0 \text{ or, equivalently, } E = E_a \text{ for } x = 0^- \quad (10)$$

and

$$n_h = n_{h1} \text{ or, equivalently, } E = E_1, \text{ for } x = 0^+ \quad (11)$$

where E_1 and n_{h1} are given by the requirement of continuity of electronic current, i.e.,

$$E_a n_0 \mu_e = E_1 (n_0 + n_{h1}) \mu_e$$

whence, using the relation (3) between E_1 and n_{h1} and expressing E_a as $j_a/n_0 e \mu_e$

$$n_{h1} = \frac{n_0}{\frac{j_a \mu_h}{j_e \mu_e} - 1} \quad (12)$$

or, alternatively,

$$E_1 = E_0 \left[1 - \frac{(\mu_e + \mu_h)}{\mu_h} \frac{j_e}{(j_a + j_e)} \right] \quad (13)$$

According to (12), n_{h1} is small when j_e is small; and, by (13), E_1 is only slightly below E_0 for this case. As j_e increases, n_{h1} increases and E_1 decreases,

and (12) and (13) would make n_{h1} infinite and E_1 zero when $j_c/j_a = \mu_h/\mu_e$. This merely means that the assumptions made in this section, in particular the neglect of diffusion and recombination or the assumption that no electrons are taken out by the injection electrode, must fail to be valid before j_c gets as large as $\mu_h j_a/\mu_e$. It will, in fact, be shown in Section 5 how the presence of enormous concentration gradients makes it essential to consider the effects of diffusion near $x = 0$ when j_c becomes large.

Putting the boundary conditions (8), (9), (10), and (11) into the wave-

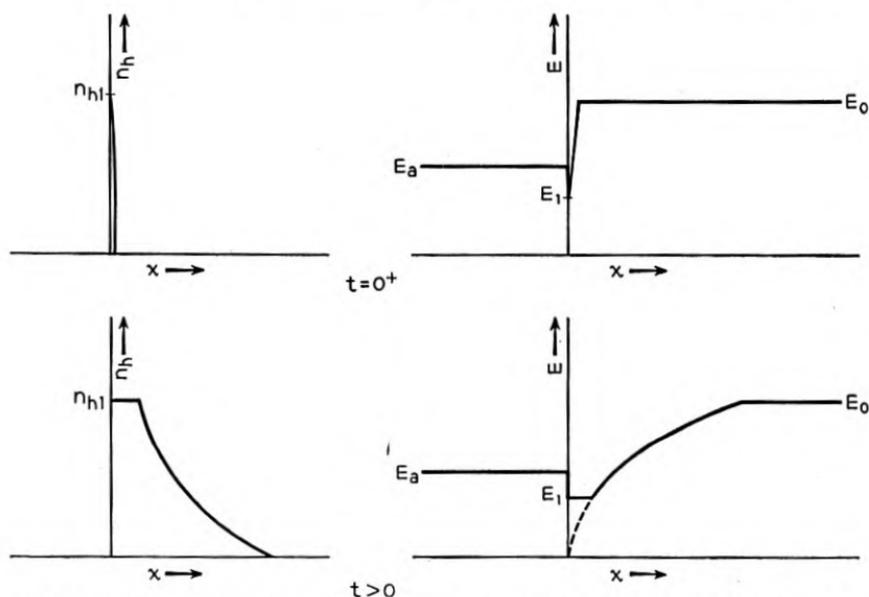


Fig. 2—Schematic variation of hole density n_h and electric field E with distance x from injection electrode and time t after the start of the injected current, in the approximation neglecting diffusion and recombination.

propagation construction described above gives the solution shown schematically in Fig. 2. An infinitesimal instant after $t = 0$, n_h is zero everywhere except in an infinitesimal interval at $x = 0$, where it rises to a maximum value n_{h1} given by (12). This is shown schematically in the upper left diagram of Fig. 2. The corresponding plot of E , shown in the upper right, dips down to E_1 , which is less than either E_a or E_0 , in this infinitesimal interval. After a finite time has elapsed, the curves of n_{h1} and E against x are simply those obtained by moving each point of the right-hand portions of these $t = 0^+$ curves a distance Vt horizontally to the right, as shown in the bottom two sketches. Here V depends on the ordinate in each diagram, taking on its maximum value $E_0\mu_h$ when $n_{h1} = 0$ or $E = E_0$. Since V is proportional to E^2 , the curve in the lower right diagram is a parabola in the range be-

tween the front and the rear of the transient disturbance; this parabola, if continued, would have its vertex at the origin. After a sufficiently long time a steady state will be reached in which the field for positive x has the uniform value E_1 and the density of holes the uniform value n_{h1} .

It is possible to measure n_h as a function of t for fixed x by using a closely spaced pair of probes to measure the potential gradient E , and converting E to n_h by (3); alternatively, the current to a single negatively biased probe can be used as a measure of n_h , if calibrated by the two-probe method. The

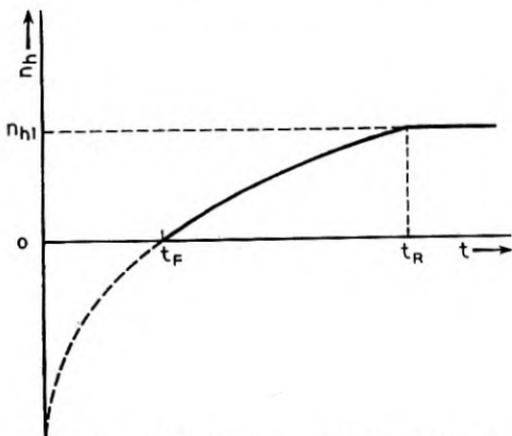


Fig. 3—Schematic variation of hole density n_h with time t after the start of the injected current, at some given distance downstream from the injection electrode, in the approximation neglecting diffusion and recombination.

portion of this curve of n_h against t for which $0 < n_h < n_{h1}$ is given, in the present approximation, by

$$t = x/V(n_h) = \frac{x[(\mu_e + \mu_h)n_h + \mu_e n_0]^2 e}{\mu_e \mu_h n_0 (j_a + j_e)} \quad (14)$$

$$= t_F [1 + (1 + \mu_h/\mu_e)n_h/n_0]^2$$

where

$$t_F = x/E_0 \mu_h \quad (15)$$

is the time of arrival of the front of the disturbance. This curve is a parabola, as shown in Fig. 3; if continued, the parabola would have its vertex on the negative n_h axis, as shown. The rear of the disturbance, at which n_h becomes constant and equal to n_{h1} , arrives at a time t_R given by inserting n_{h1} from (12) into (14):

$$t_R = t_F / [1 - (1 + \mu_e/\mu_h)j_e/(j_a + j_e)]^2 \quad (16)$$

Note that the velocity of advance of the rear of the disturbance is less than that with which the holes drift in the steady-state field E_1 . In other words, wave velocity and particle velocity must be distinguished in phenomena of this sort, although they happen to coincide at the front of the disturbance.

The discussion just given has been based on the assumption that j_a and j_e are independent of time, and that they both flow from left to right in Fig. 1. Time changes in the currents are easily taken into account in the n_h construction of Fig. 2: according to (4), it is merely necessary to move the various points of the curve of n_h against x to the right with the variable velocity $V(n_h, t)$ instead of the constant velocity $V(n_h)$; in addition, n_{h1} will in general not be a constant, so that the part of the curve for small x will no longer be a horizontal line. As for the restriction that the currents all flow from left to right, only a change of notation is needed to make all formulas apply to the case where all currents flow from right to left; and the case where part of j_e flows to the right and part to the left can, obviously, occur only under conditions where the assumptions of this section are not fulfilled, i.e., can occur only if electrons are removed at $x = 0$ or if both diffusion and recombination are important. For, if diffusion is negligible, the existence of a potential maximum at $x = 0$ implies a convergence of electrons from both sides onto the plane $x = 0$, and recombination alone cannot annihilate electrons at a finite rate in an infinitesimal volume.

Mention has already been made of the fact that equations such as (12) and (13) give an infinite density of holes when $j_e/j_a = \mu_h/\mu_e$, and are nonsensical for larger values of j_e/j_a . It is easy to see why any theory which neglects diffusion must break down for values of j_e/j_a of this size and larger if no electrons are removed by the injection electrode. If j_e/j_a is too large, any positive field just to the right of the injection plane $x = 0$ will cause more electrons to flow in the negative x -direction than can be carried off by the current j_a which flows in the region of negative x . This difficulty cannot be eliminated by making the field smaller in the region of small positive x , since making the field smaller requires a higher density of holes to carry the hole current j_e ; and this in turn requires a higher density of electrons to preserve electrical neutrality. Thus, though it may be possible to realize experimental conditions under which the approximations of this section are valid for moderate values of j_e/j_a , increase of j_e/j_a above the critical value will always result in the building up of an enormously high density of holes and electrons near $x = 0$, and one must then consider diffusive transport and possibly other phenomena such as breakdown of the assumption that no electrons are removed by the injection electrode.

It will be shown below that the effect of recombination on the curves of n_h against x at various times t can be taken into account by using a geometrical construction similar to that of Fig. 2 except that, instead of moving the

various points of the curve horizontally to the right with increasing time, one must move them along a family of decreasing curves (cf. Figs. 4, 5, and 6). The effect of diffusion can be described roughly as a migration of each point from one of these curves to another.

3. COMPLETE DIFFERENTIAL EQUATIONS OF THE PROBLEM

As was mentioned in Section 1, the transport of electrons and holes along a narrow filament can be described by one-dimensional equations even if recombination at the surface of the filament causes the distribution of electrons and holes to be non-uniform over its cross-section. In the equations to follow, n_h and n_e will be understood to refer to averages, over the cross-section, of the hole and electron densities, respectively; the electrostatic field E can always be assumed uniform over the cross-section of the filament, if the latter is thin. The as yet uncertain influence of the surface on the rate of recombination of electrons and holes can be allowed for by writing the recombination rate as $n_0 R(n_h/n_0)/\tau$ particles per unit volume per unit time, where R is a function which is asymptotically n_h/n_0 as its argument $\rightarrow 0$, and where τ is the recombination time for small hole densities. For pure volume recombination, $R = n_h n_e / n_0^2 = (n_h/n_0)(1 + n_h/n_0)$, while a conceivable extreme of surface recombination would be $R = n_h/n_0$.

Using this function, the continuity equations for electrons and holes can then be written, with inclusion of recombination and diffusion terms

$$\frac{\partial n_h}{\partial t} = -\frac{\partial}{\partial x} (E\mu_h n_h) - \frac{n_0}{\tau} R\left(\frac{n_h}{n_0}\right) + \frac{\partial}{\partial x} \left(D_h \frac{\partial n_h}{\partial x} \right) \quad (17)$$

$$\frac{\partial n_e}{\partial t} = \frac{\partial}{\partial x} (E\mu_e n_e) - \frac{n_0}{\tau} R\left(\frac{n_h}{n_0}\right) + \frac{\partial}{\partial x} \left(D_e \frac{\partial n_e}{\partial x} \right) \quad (18)$$

where the D 's are the diffusion constants, related to the mobilities μ by the Einstein relation

$$D/\mu = kT/e \quad (19)$$

Using the neutrality condition $n_e = n_0 + n_h$, subtracting (17) from (18) and integrating gives the equation of constancy of current, the generalization of (3):

$$E[(\mu_e + \mu_h)n_h + \mu_e n_0] + \frac{kT}{e} (\mu_e - \mu_h) \frac{\partial n_h}{\partial x} = j(t)/e. \quad (20)$$

Solving for E gives

$$E = \frac{j - kT(\mu_e - \mu_h) \frac{\partial n_h}{\partial x}}{e[(\mu_e + \mu_h)n_h + \mu_e n_0]} \quad (21)$$

which can be substituted into (17) to give a differential equation for n_h alone:

$$\frac{\partial n_h}{\partial t} = -\frac{j}{e} \frac{\partial}{\partial x} \left[\frac{\mu_h n_h}{(\mu_e + \mu_h)n_h + \mu_e n_0} \right] - \frac{n_0}{\tau} R \left(\frac{n_h}{n_0} \right) + \frac{kT}{e} \mu_h \mu_e \frac{\partial}{\partial x} \left[\frac{(n_0 + 2n_h) \frac{\partial n_h}{\partial x}}{(\mu_e + \mu_h)n_h + \mu_e n_0} \right]. \quad (22)$$

The first term on the right represents drift, the second recombination, and the third diffusion. This holds whether j is constant in time or not. However, as the remainder of this memorandum will be devoted to the case where the currents involved are held constant after their initiation, it will be convenient to simplify the notation by introducing a current-dependent scale for x and writing the equation in terms of the dimensionless variables

$$\nu = n_h/n_0, s = t/\tau, \xi = x/E_0\mu_h\tau = xen_0\mu_e/j\mu_h\tau \quad (23)$$

In terms of these (22) becomes simply

$$\frac{\partial \nu}{\partial s} = -\frac{\partial}{\partial \xi} \left[\frac{\nu}{1 + (1 + \mu_h/\mu_e)\nu} \right] - R(\nu) + \left(\frac{j}{j} \right)^2 \frac{\partial}{\partial \xi} \left[\frac{(1 + 2\nu) \frac{\partial \nu}{\partial \xi}}{1 + (1 + \mu_h/\mu_e)\nu} \right] \quad (24)$$

where $R(\nu) = \nu(1 + \nu)$ for pure volume recombination, or $= \nu$ for a surface recombination uninfluenced by the electron density, and where

$$J = (kTe \mu_e^2 n_0^2 / \mu_h \tau)^{1/2} = \sigma_0 (kT/e \mu_h \tau)^{1/2} \quad (25)$$

Numerically the characteristic field is, at 300°K, with $\mu_h = 1700 \text{ cm}^2/\text{v sec}$,³

$$(kT/e\mu_h\tau)^{1/2} = 3.90 (\tau/1\mu\text{s})^{-1/2} \text{ volts/cm} \quad (26)$$

Note that the importance of the diffusion term in (24) goes down inversely as the *square* of the current density used and inversely as the *square* of the recombination time; this is because an increase in the distance the holes travel decreases the distance they diffuse by decreasing the concentration gradient, and also makes a given diffusion distance less serious by comparison with the total distance traveled. Note also that, if $\mu_e = \mu_h$, the last term of (24) reduces simply to $\left(\frac{j}{j} \right)^2 \frac{\partial^2 \nu}{\partial \xi^2}$, but that, if $\mu_e \neq \mu_h$, the diffusion term is not a simple second derivative.

³ G. L. Pearson, paper Q9 presented at the Washington Meeting of the American Physical Society, April 29, 1949.

4. SOLUTION INCLUDING RECOMBINATION BUT NEGLECTING DIFFUSION

It is plausible to expect by analogy with Fig. 2 that (24) can be solved, neglecting the last term, by a similar construction in which the curve of n_h against x at time t is derived from that at time 0 by moving each point to the right along a descending curve, instead of along a horizontal line as before. To show that this is indeed the case, and at the same time to show that the diffusion term cannot so easily be taken into account, let (24) be written, omitting its last term, as

$$\frac{\partial v}{\partial s} = -\Phi(v) \frac{\partial v}{\partial \xi} - R(v)$$

where Φ is just the translation into dimensionless variables of the velocity V encountered in (4). This can be converted into a differential equation for ξ by writing

$$\left(\frac{\partial v}{\partial s}\right)_\xi = -\left(\frac{\partial \xi}{\partial s}\right)_v \left(\frac{\partial \xi}{\partial v}\right)_s$$

and multiplying through by $\left(\frac{\partial \xi}{\partial v}\right)_s$:

$$\left(\frac{\partial \xi}{\partial s}\right)_v = R(v) \left(\frac{\partial \xi}{\partial v}\right)_s + \Phi(v) \quad (27)$$

or with $w = \int \frac{dv}{R(v)}$,

$$\frac{\partial \xi}{\partial s} - \frac{\partial \xi}{\partial w} = \Phi(w)$$

$$\frac{\partial \left(\xi + \int \Phi dw \right)}{\partial s} - \frac{\partial \left(\xi + \int \Phi dw \right)}{\partial w} = 0$$

whence the general solution is

$$\xi = -\int \Phi dw + f(s + w) \quad (28)$$

where f is an arbitrary function. If the same transformation is tried on (24) with the diffusion term retained, the equation corresponding to (27) has an additional term on the right containing a quotient of second and first derivatives of ξ with respect to v , and the simple explicit solution fails.

To apply (28) to explicit calculation, or even to visualize it physically, it is necessary to determine the proper form of the arbitrary function f to fit

the boundary conditions of the problem. This is most conveniently done by introducing a family of curves as suggested by the analogy with Fig. 2. The analogy suggests that we should try to find curves in the ν, ξ plane (the full curves of Fig. 4) such that a point can move along any one of them with velocity components

$$\frac{d\xi}{ds} = \Phi, \quad \frac{d\nu}{ds} = -R.$$

The equation of any such curve is

$$\frac{d\xi}{d\nu} = -\Phi/R$$

or

$$\xi(\nu, \nu_0) = \int_{\nu}^{\nu_0} \frac{\Phi}{R} d\nu \quad (29)$$

where ν_0 , the intercept of the curve on the ν -axis, is taken as a parameter distinguishing the curve in question from others of the family. A point which starts at height ν_0 on the ν -axis at time $s = 0$ will reach height ν at time

$$s(\nu, \nu_0) = \int_{\nu}^{\nu_0} \frac{d\nu}{R}. \quad (30)$$

Thus, after time s , the locus of all points which start at all the various heights ν_0 will be the curve obtained by eliminating ν_0 between (29) and (30) (shown dotted in Fig. 4). That this curve is, in fact, of the form (28) and therefore a solution of the differential equation is easily seen by writing (29) and (30) in terms of integrals taken from some arbitrary but fixed lower limit:

$$\xi(\nu, \nu_0) = -\int^{\nu} \frac{\Phi}{R} d\nu + \int^{\nu_0} \frac{\Phi}{R} d\nu$$

$$s(\nu, \nu_0) = -\int^{\nu} \frac{d\nu}{R} + \int^{\nu_0} \frac{d\nu}{R}.$$

As ν_0 is varied both the integrals with upper limit ν_0 will vary, and either can be expressed as a function of the other:

$$\int^{\nu_0} \frac{\Phi d\nu}{R} = f\left(\int^{\nu_0} \frac{d\nu}{R}\right)$$

whence

$$\xi = -\int^{\nu} \frac{\Phi}{R} d\nu + f\left(s + \int^{\nu} \frac{d\nu}{R}\right)$$

which is identical with (28).

The equations (29) and (30) of course apply only to the portion of the curve of ν against ξ which is derived from starting points ν_0 on the ν axis which are less than the maximum value ν_1 corresponding to the value n_{h1} given by (12): The points for $\nu_0 < \nu_1$ are merely initiated at time $s = 0$ and propagated by the differential equation from then on; the point $\nu = \nu_1$, $\xi = 0$, on the other hand, remains a source at all times from the initiation of the injection onward. Thus the complete curve of ν against ξ for any positive s follows the dotted construction of Fig. 4 from the ξ axis up to where it intersects the full curve corresponding to $\nu_0 = \nu_1$, after which it

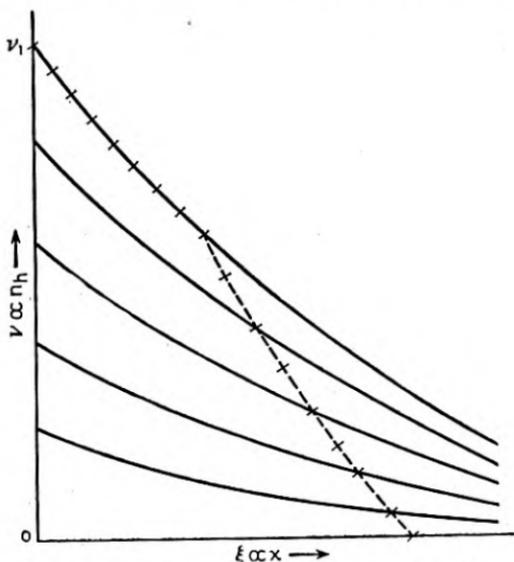


Fig. 4—Schematic illustration of the method of constructing the curve of hole density n_h against distance x from the injection electrode at some given time, taking account of recombination but neglecting diffusion.

follows the latter curve, as indicated by the crosses in the figure. The steady-state distribution is thus simply the full curve for $\nu_0 = \nu_1$.

For explicit calculation for the case of pure volume recombination one must insert $\Phi = 1/[1 + (1 + \mu_h/\mu_e)\nu]^2$, $R = \nu(1 + \nu)$ into (29) and (30). The integrations are easily carried out and give

$$\xi = \left[\frac{1 + \mu_e/\mu_h}{1 + (1 + \mu_h/\mu_e)\nu_0} + \ln \frac{\nu_0}{1 + (1 + \mu_h/\mu_e)\nu_0} + \frac{\mu_e^2}{\mu_h^2} \ln \frac{1 + (1 + \mu_h/\mu_e)\nu_0}{1 + \nu_0} \right] - [\text{same with } \nu \text{ instead of } \nu_0] \quad (31a)$$

$$s = \ln \frac{\nu_0}{1 + \nu_0} - \ln \frac{\nu}{1 + \nu} \quad (32a)$$

For the case of a surface recombination uninfluenced by electron concentration one obtains similarly, with $R = \nu$:

$$\xi = \left[\frac{1}{1 + (1 + \mu_h/\mu_e)\nu_0} - \ln \frac{\nu_0}{1 + (1 + \mu_h/\mu_e)\nu_0} \right] - [\text{same with } \nu \text{ instead of } \nu_0] \quad (31b)$$

$$s = \ln \frac{\nu_0}{\nu} \quad (32b)$$

When $\mu_e = 3\mu_h/2$, as for germanium, (31a) and (31b) become respectively

$$\xi = \left[\frac{5/2}{1 + 5\nu_0/3} + \ln \frac{\nu_0}{1 + 5\nu_0/3} + \frac{9}{4} \ln \frac{1 + 5\nu_0/3}{1 + \nu_0} \right] - [\text{same with } \nu \text{ instead of } \nu_0] \quad (33a)$$

and

$$\xi = \left[\frac{1}{1 + 5\nu_0/3} + \ln \frac{\nu_0}{1 + 5\nu_0/3} \right] - [\text{same with } \nu \text{ instead of } \nu_0] \quad (33b)$$

These can also be written, using (32a) and (32b),

$$\xi = s + \frac{5/2}{1 + 5\nu_0/3} - \frac{5/2}{1 + 5\nu/3} + \frac{5}{4} \ln \left[\frac{(1 + 5\nu_0/3)(1 + \nu)}{(1 + 5\nu/3)(1 + \nu_0)} \right] \quad (34a)$$

and

$$\xi = s + \ln \left(\frac{\nu + 3/5}{\nu_0 + 3/5} \right) - \frac{1}{1 + 5\nu/3} + \frac{1}{1 + 5\nu_0/3} \quad (34b)$$

Figures 5a and 5b show as a full curve the plot of eq. (33a) for the case $\nu_0 = \infty$, and the full curve in Fig. 6 shows in the same way the plot of (33b) for $\nu_0 = \infty$. Changing ν_0 of course merely shifts either curve horizontally. Note the very sharp increase of ν for small ξ , which shows up in pronounced manner on the expanded scale of Fig. 5b. The corresponding values of s , computed from (32a) or (34a), are marked on the curve of Fig. 5; the corresponding marks on the curve of Fig. 6 also represent values of s at intervals of 0.2, but are not labeled with absolute values because (32b) is infinite for $\nu_0 = \infty$.

For large ξ , ν becomes very small and it becomes legitimate to expand the logarithms. The first few terms of the resulting asymptotic expression for ξ are, for $\nu_0 = \infty$ and the recombination function leading to (31a),

$$\xi \sim \left(\frac{\mu_e^2}{\mu_h^2} - 1 \right) \ln (1 + \mu_h/\mu_e) - (1 + \mu_e/\mu_h) - \ln \nu + (3 + 2\mu_h/\mu_e)\nu \quad (35a)$$

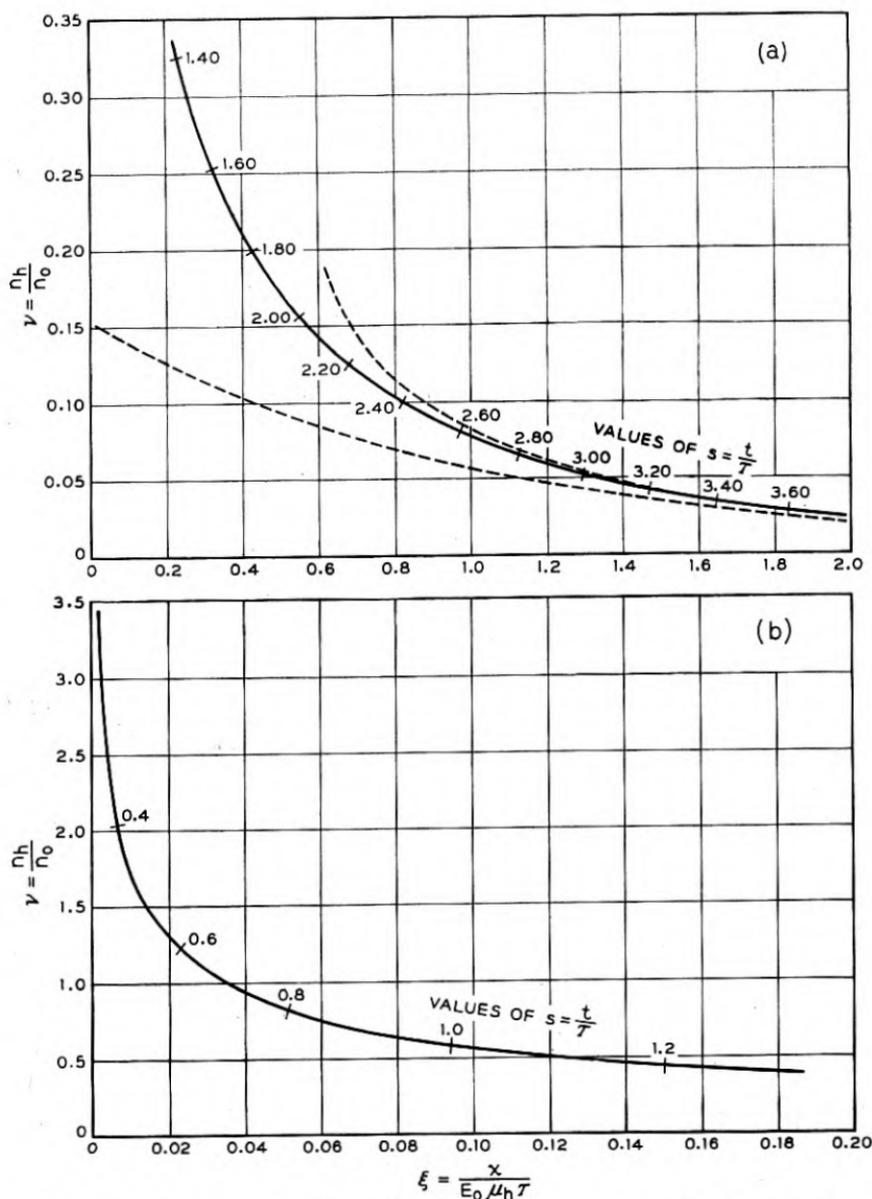


Fig. 5—Steady-state curve of hole density n_h against distance x , for the case of ideal volume recombination (recombination rate = $n_h n_e / \tau n_0$), and asymptotic approximations to this curve.

while, for the recombination function leading to (31b),

$$\xi \sim -\ln(1 + \mu_h/\mu_e) - 1 - \ln\nu + 2(1 + \mu_e/\mu_h)\nu \quad (35b)$$

In Figs. 5a and 6 the lower dotted curve represents the sum of the terms of (35a) or (35b) respectively as far as the term in $\ln\nu$: in this approximation the dependence of ν on ξ is exponential. An exponential behavior of this sort is assumed in the small-signal theory of the modulation of the resistance of a filament of semiconductor by hole injection.⁴ The upper dotted curve

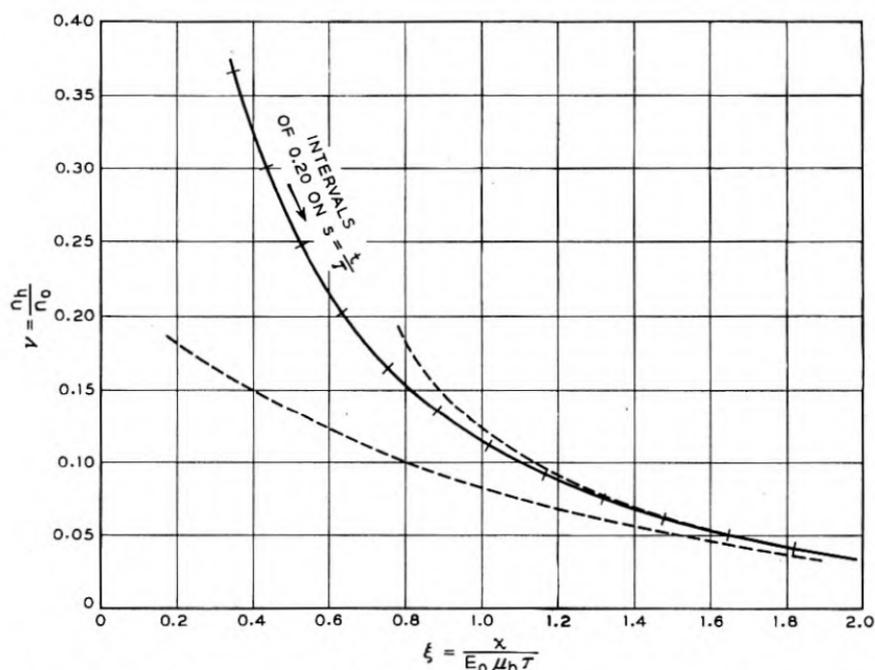


Fig. 6—Steady-state curve of hole density n_h against distance x , for the case of ideal surface recombination (recombination rate = n_h/τ), and asymptotic approximations to this curve.

in Figs. 5a and 6 is a plot of (35a) or (35b), respectively, with the linear term included. It will be seen that in both figures the simple exponential approximation is already quite far off when $\nu = n_h/n_0 = 0.1$, though it improves rapidly for smaller ν .

Figure 7 shows a sample plot of ν against ξ for the case of ideal volume recombination (eqs. (31a) etc.), for the numerical conditions $s = 1$, $\nu_1 = 0.3$ (cf Fig. 4). According to (12), whose validity at $\xi = 0$ is unimpaired by the occurrence of recombination, this value of ν_1 implies $j_a/j_c = 6.5$. The left-

⁴ W. Shockley, G. L. Pearson, and J. R. Haynes, *Bell Sys. Tech. Jour.*, this issue.

hand portion of this curve is simply traced from Fig. 5, with a horizontal shift sufficient to give an intercept at $\nu = 0.3$; the right-hand portion was constructed by placing the paper for Fig. 7 over that for Fig. 5, shifting horizontally until the point corresponding to one of the values of s marked on Fig. 5 lay on the axis of ordinates of Fig. 7, marking the position of the point labeled with one plus this value of s , and repeating.

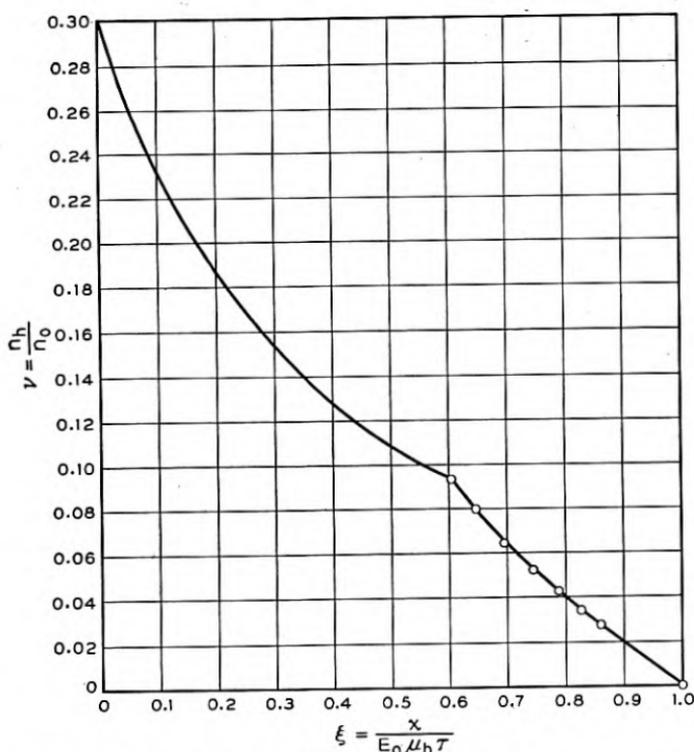


Fig. 7—Variation of hole density n_h with distance x at time $t = \tau$ assuming $n_{h1} = 0.3 n_0$ recombination rate $= n_h n_e / \tau n_0$, and neglecting diffusion.

Figure 8 shows sample plots of ν against s for the same case of ideal volume recombination, with $\nu_1 = 0.3$, for $\xi = 0.5$ and $\xi = 1.0$. Curves for a different ν_1 would start out exactly the same, but rise higher. The rising portion of the curve for $\xi = 0.5$, for example, was constructed from the curve of Fig. 5a by locating various points (ξ, ν) on the latter curve and associating with the ν value of each such point a value of s equal to the difference of the s values marked on the curve of Fig. 5a for the two points abscissae ξ and $(\xi - 0.5)$. As Fig. 5a was prepared entirely by slide rule, the accuracy is not all that can be desired; the individual computed points are shown to give an idea of the magnitude of the computational errors.

For convenience in future calculations the equations will be appended which correspond to (31) to (34) when, instead of n_h , the field E is used as dependent variable in the differential equations. In terms of the dimensionless variable

$$\epsilon = E/E_0 = \frac{1}{1 + \nu(1 + \mu_h/\mu_e)} \quad (36)$$

and the parameter ϵ_0 corresponding to $\nu = \nu_0$, the equations are, for ideal volume recombination (eqs. (31a) etc.),

$$\xi = \left[(1 + \mu_e/\mu_h)\epsilon_0 - \frac{\mu_e^2}{\mu_h} \ln \left(1 + \frac{\mu_h}{\mu_e} \epsilon_0 \right) + \ln (1 - \epsilon_0) \right] \quad (37a)$$

— [same with ϵ instead of ϵ_0]

$$s = \ln \frac{(1 - \epsilon_0)}{\left(1 + \frac{\mu_h}{\mu_e} \epsilon_0 \right)} - \ln \frac{(1 - \epsilon)}{\left(1 + \frac{\mu_h}{\mu_e} \epsilon \right)} \quad (38a)$$

while, for the recombination function leading to eqs. (31b) etc.,

$$\xi = \epsilon_0 - \epsilon + \ln \frac{1 - \epsilon_0}{1 - \epsilon} \quad (37b)$$

$$s = \ln \left[\frac{\frac{1}{\epsilon_0} - 1}{\frac{1}{\epsilon} - 1} \right]. \quad (38b)$$

The electrostatic potential U is

$$U = - \int E dx = - E_0^2 \mu_h \tau \int \epsilon d\xi.$$

In the steady state the relation between ϵ and ξ is given by (37) with ϵ_0 set equal to ϵ_1 which, by (13), is $1 - \frac{(\mu_e + \mu_h)}{\mu_h} \frac{j_0}{(j_a + j_e)}$. For this case

$$\begin{aligned} U &= -E_0^2 \mu_h \tau \left[\epsilon \xi - \int \xi d\epsilon \right] \\ &= -E_0^2 \mu_h \tau \left[\epsilon (\mu_e^2/\mu_h^2 - 1) - \epsilon^2 (1 + \mu_e/\mu_h)/2 \right. \\ &\quad \left. - \ln (1 - \epsilon) - \frac{\mu_e^2}{\mu_h} \ln \left(1 + \frac{\mu_h}{\mu_e} \epsilon \right) \right] + \text{const.} \end{aligned} \quad (39a)$$

for ideal volume recombination; while, for the assumptions leading to eqs. (31b), etc. the relation is

$$U = -E_0^2 \mu_h \tau [\epsilon - \epsilon^2/2 - \ln(1 - \epsilon)] + \text{const.} \quad (39b)$$

Thus, in the steady state, the difference in potential between any two points to the right of $x = 0$ can be obtained by finding the values of ϵ for these two points by (37), and then using these to evaluate the difference in the values of (39) at the two points. To the left of $x = 0$, of course, E is constant and equal to j_a/σ .

5. DIFFUSION EFFECTS

Diffusion will obviously be very important at small values of $\xi = x/E_0 \mu_h \tau$ when n_{h1} is large, because of the tremendous concentration gradients which

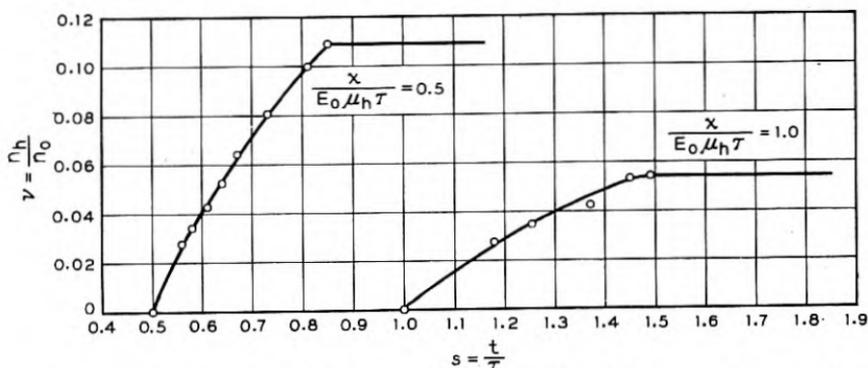


Fig. 8—Transient behavior of n_h with time at position $x/E_0 \mu_h \tau = 0.5$ and 1.0 , assuming $n_{h1} = 0.3 n_0$ and recombination rate $= n_h n_e / \tau n_0$.

Figs. 5 and 6 predict for such cases. Also, of course, diffusion will round off the discontinuities in slope which appear at the front and rear of the transient disturbance as in Fig. 7 and Fig. 8. At other points the importance of diffusion effects can be roughly estimated either by comparing the diffusion current with the drift current or by comparing the divergences of these two contributions to the current, i.e., the last and first terms on the right of (24). Referring to these terms in (24) we have

$$\left[\frac{\text{diffusion current}}{\text{drift current}} \right] = \left(\frac{J}{j} \right)^2 \frac{(1 + 2\nu)}{\nu} \left[\frac{\partial \nu}{\partial \xi} \right] \quad (40)$$

$$\left[\frac{\text{div. diffusion current}}{\text{div. drift current}} \right] = \left(\frac{J^2}{j} \right) \cdot \left[[1 + (1 + \mu_h/\mu_e)\nu][1 + 2\nu] \frac{\partial^2 \nu}{\partial \xi^2} / \frac{\partial \nu}{\partial \xi} + (1 - \mu_h/\mu_e) \frac{\partial \nu}{\partial \xi} \right]. \quad (41)$$

For the steady-state curve approximate values of the expressions (40) and (41) can be computed by evaluating the derivatives of ξ with respect to ν

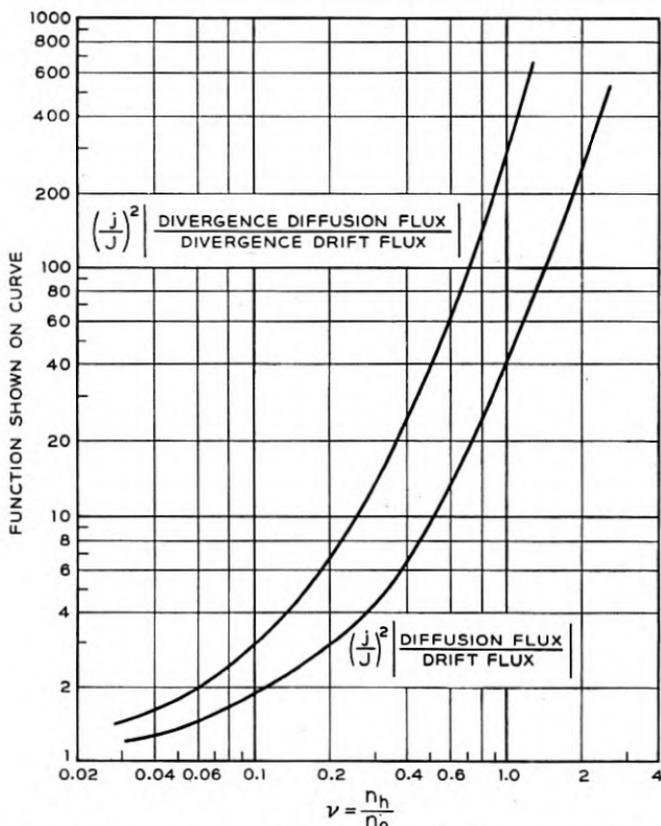


Fig. 9—Asymptotic magnitude of diffusion terms in the steady-state flux of holes, when j/J is large.

from (29) or (31). For the case of ideal volume recombination with $\mu_e/\mu_h = 3/2$ this gives, if the diffusion effects are not too large,

$$\left[\frac{\text{diffusion current}}{\text{drift current}} \right] \approx \left(\frac{J}{j} \right)^2 (1 + \nu)(1 + 2\nu)(1 + \frac{5}{3}\nu)^2 \quad (42)$$

$$\left[\frac{\text{div. diffusion current}}{\text{div. drift current}} \right] \approx \left(\frac{J}{j} \right)^2 (1 + \frac{5}{3}\nu)^2 (1 + \frac{28}{3}\nu + 21\nu^2 + \frac{40}{3}\nu^3). \quad (43)$$

These functions are plotted in Fig. 9. From this figure one can estimate roughly when diffusion will begin to have serious effects other than a slight rounding of the leading and trailing ends of the transient. For example, if

it is desired that the ratio (43) be less than about 0.1 in the steady state for values of ν as high as 0.3, the upper curve of Fig. 9 shows that the current density used must be large enough to make $\left(\frac{j}{j_0}\right)^2 \leq \frac{0.1}{13.6}$, i.e., $j \geq 11.7 J$, where J is given by (25) and (26).

An approximate evaluation of (40) and (41) in the transient region can be performed by graphical or numerical differentiation of a curve such as that of Fig. 7. For example, a rough calculation based on Fig. 7 gives, in the middle of the transient portion ($\xi = 0.75$),

$$\left[\frac{\text{div. diffusion current}}{\text{div. drift current}} \right] \approx 3 \left(\frac{j}{j_0} \right)^2.$$

More important and also more difficult to estimate is the effect of diffusion in rounding off the front and rear edges of the transient. Various ways can be devised to estimate a rough upper limit to the amount of rounding off to be expected. One such is to compute what the diffusive flux just behind the front of the advancing disturbance would be if the distribution of holes were the same as in the absence of diffusion. Under conditions where diffusion is not too serious the time integral of this diffusive flux between any two times can be equated to the increase in rounding of the front, as measured by the area between an ideal curve such as that of Fig. 7 and the actual curve of ν against ξ for the same time s . The integration cannot be extended back to time zero, however, since the integral of the flux diverges logarithmically. The fact that the diffusive flux is actually finite instead of infinite of course arises from the fact that at small times the concentration gradient a short distance behind the front can no longer be approximated by the gradient which would obtain in the absence of diffusion, but instead is very much less. This suggests that an upper limit to the total diffusive flux passing into the region of the front from time 0 to time s can be obtained by taking the flux computed as described above between the times s_0 and s , and adding to it the total number of holes which have left the injection electrode between time 0 and time s_0 . Since this gives an upper limit for any s_0 , one may use the minimum of the resulting sum as s_0 is varied.

The results of some sample calculations of this sort are shown in Fig. 10, which refers to the same time, currents, and recombination function as Fig. 7, viz., $s = t/\tau = 1.0$, $j_0/j_a = 2/13$, ideal volume recombination. The full curve is the transient portion of Fig. 7 replotted on a larger scale. The lower dotted curve is a curve drawn in by hand in such a way as to make the area between it and the full curve equal the upper limit computed in the manner just described, for the case $j = 100J$. The upper curve was drawn

similarly for $j = 31.6J$. Since the true curve of ν against ξ must lie between the dotted curve and the full curve in each case, it can be concluded that for times and current ratios of this order the diffusionless theory of Section 4 gives a useful approximation to the transient when $j \gtrsim 100J$. At the other end, it seems likely that for $j \lesssim 10J$ the theory of Section 4 has no quantitative utility at all in the transient region.

When diffusion effects are sufficiently great, account must also be taken of the fact that the boundary conditions at the injection electrode ($x = 0$)

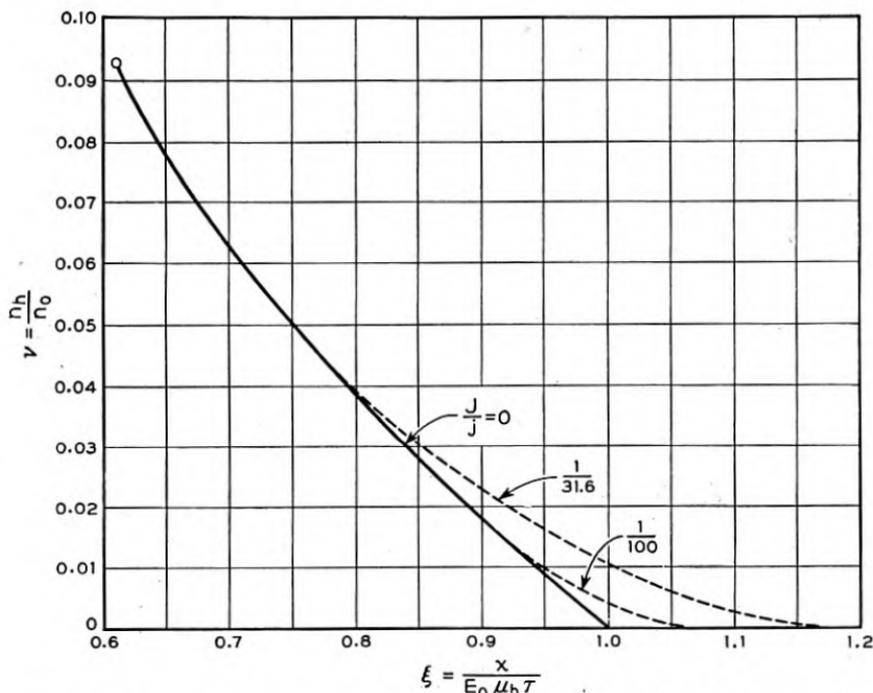


Fig. 10—Approximate magnitudes of the rounding of the front by diffusion for various values of j/J , for the case $t = \tau$, $j_e/j_a = 2/13$, ideal volume recombination. Ordinate is proportional to hole density, abscissa to distance from injection electrode.

take a different form from those in the absence of diffusion. In the absence of diffusion and with the assumption that only holes are injected at $x = 0$, the current just to the right of $x = 0$ must consist of a contribution j_e from holes and a contribution j_a from electrons, while the current just to the left of $x = 0$ is purely electronic and of magnitude j_a . This implies, as we have seen in Section 2, that the hole density be discontinuous, with the value n_{h1} given by (12) just to the right of $x = 0$, and the value zero just to the left. But if diffusion is allowed, the hole density must be continuous. For the

idealized case where holes are injected on the plane $x = 0$ and no electrons are removed there, the equations to be satisfied are

$$D_e \left(\frac{\partial n_e}{\partial x} \right)_+ + n_e \mu_e E_+ = D_e \left(\frac{\partial n_e}{\partial x} \right)_- + n_e \mu_e E_- \quad (44)$$

$$-D_h \left(\frac{\partial n_h}{\partial x} \right)_+ + n_h \mu_h E_+ = \frac{j_e}{e} - D_h \left(\frac{\partial n_h}{\partial x} \right)_- + n_h \mu_h E_- \quad (45)$$

$$D_e \left(\frac{\partial n_e}{\partial x} \right)_- - D_h \left(\frac{\partial n_h}{\partial x} \right)_- + (n_e \mu_e + n_h \mu_h) E_- = j_a/e \quad (46)$$

where subscripts $+$ and $-$ refer to conditions just to the right of $x = 0$ and just to the left, respectively. Using the neutrality condition $n_e = n_0 + n_h$ these are three equations for the five unknowns $\left(\frac{\partial n_h}{\partial x} \right)_\pm$, E_\pm , n_h . To complete the determination of these quantities the differential equation (22) must be solved and the boundary conditions imposed that $n_h \rightarrow 0$ as $x \rightarrow \pm \infty$.

Actually the problem of estimating conditions at $x = 0$ may not be quite as formidable as the preceding paragraph suggests, at least if the diffusion parameter J/j is reasonably small and if j_e/j_a is also not too large. For such cases the "upstream diffusion" of holes into the region of negative x will probably reach a steady state in a very short time. Solutions of the steady state differential equation in this region have been obtained numerically by W. van Roosbroeck (unpublished). Such solutions will give one relation between n_h and $\left(\frac{\partial n_h}{\partial x} \right)_-$; another relation, in the form of a fairly narrow range of limits, is provided by the fact that $\left(\frac{\partial n_e}{\partial x} \right)_+$ will under these conditions be $\ll \left(\frac{\partial n_e}{\partial x} \right)_-$, being in fact probably somewhere between zero and the value for the diffusionless case with the same value of n_{h1} .

Of course, if the mathematical solution for this one-dimensional idealization is to be applied to a case where holes are injected into a filament by a pointed electrode on its boundary, little meaning can be attached to variations in the n_h of the mathematical solution within a range of x values smaller than the diameter of the filament.

6. SUMMARY AND DISCUSSION

There are three principal factors which limit the range of conditions within which the present theory provides a useful approximation to the transient behavior of n_h as a function of t and x . These are diffusion, trapping, and departure from one-dimensional geometry. If the geometry is sufficiently nearly one-dimensional and trapping is negligible, the discussion

of Section 5 shows that the theory of Section 4, with its neglect of diffusion, will give a useful approximation to the truth whenever the field in which the holes migrate is sufficiently strong—e.g., strong enough to make the current density $j \gtrsim 100 J$, where J is given by (25) and (26). The obtaining of “sufficiently strong” fields without excessive heating or other undesirable effects is facilitated by the use of specimens with as long a recombination time τ as possible, and by the use of specimens of low conductivity. However, it is hard to say how low the conductivity can be made without danger that the “no trapping” assumption will break down, since for this assumption to be valid the density of hole traps must be \ll the density of donors.

The numerical predictions of the theory depend upon the way in which the rate of recombination is assumed to depend upon the concentrations of electrons and holes, i.e., upon the form of the function $R(\nu)$ introduced in (17) and (18). The full curves of Figs. 5 and 6 give the steady-state dependence of n_h on x for two simple assumptions regarding $R(\nu)$, the dependence corresponding to any given boundary value n_{h1} at $x = 0$ being simply obtained by a suitable horizontal shift of the curve plotted. When the currents are held constant after their initiation, the auxiliary time scale in these figures can be used to construct the transient disturbance at any time, by the methods described in connection with the examples of Figs. 7 and 8.

These results should hold for a plane-parallel arrangement of electrodes or, to a good approximation, for electrodes placed along the length of a narrow filament, provided the n_h appearing in the equations is interpreted as a cross-sectional average of the hole density and provided the other assumptions given in Section 1 are fulfilled. It is easy to see, however, that practically the same equations apply to cases of cylindrical or spherical geometry, in the approximation where diffusion is neglected. For, in these cases, the original equations (17) and (18) merely have $\frac{\partial}{\partial x}(\dots)$ replaced by $\frac{1}{r} \frac{\partial}{\partial r}(r \dots)$

or $\frac{1}{r^2} \frac{\partial}{\partial r}(r^2 \dots)$; if the diffusion terms are neglected the solution is the same as before with x replaced by $r^2/2$ (cylindrical case) or $r^3/3$ (spherical case) and with j replaced by $I/2\pi d$ (cylindrical case, d = thickness of sample, I = total current) or by $I/4\pi$ (spherical case). However, it may be difficult to realize experimentally conditions approximating cylindrical or spherical geometry which satisfy the requirement that diffusion effects be small.

Another generalization which is easily made is the removal of the assumption that no electrons are withdrawn by the electrode at $x = 0$. As far as conditions to the right of $x = 0$ are concerned (Fig. 1), the only change required in the diffusionless theory is to interpret j_e as the current density leaving the emitter electrode in the form of holes, rather than as the total current from the emitter electrode, and to interpret j_a as the sum of the

current leaving the emitter electrode in the form of electrons and any current to the left of $x = 0$.

It should also be clear that the entire analysis of this paper, though it has for definiteness been formulated for the case where holes are injected into an excess semiconductor, applies just as well to any case where electrons can be injected into a defect semiconductor. For the latter case it is merely necessary to interchange the subscripts e and h in the formulas. Though the types of experiments discussed in this paper have to date only been reported for n -type germanium, the occurrence of similar phenomena in p -type specimens is indicated by the successful use of such specimens in transistors.⁵

An interesting and possibly quite useful phenomenon should occur when, after establishment of a steady state, the current j_e is suddenly turned off. There will result a transient disturbance propagated in the direction of in-

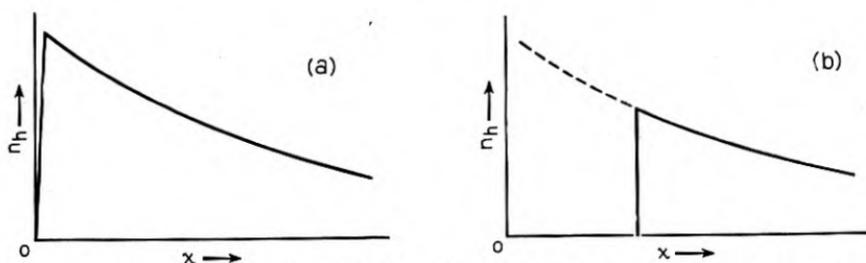


Fig. 11—Schematic variation of hole density n_h with distance x , illustrating formation of a shock wave by quickly decreasing j_e to zero, for the case where $j = j_e + j_a$ is kept constant.

- (a) Immediately after reduction of j_e to zero.
 (b) Later time.

creasing x , which is very much like a shock wave in a gas. This, the most interesting feature of the phenomenon, will occur regardless of whether j_a remains constant when j_e is cut off; however, the simplest example for illustrative purposes is the case where j_a is increased by the amount j_e at the instant when the latter is cut off, so that j remains constant. For this case, illustrated in Fig. 11, the values of n_h ahead of the advancing front will remain the same at each point as in the previous steady state. Just behind the front, n_h must drop abruptly to zero. If j/J is large, where J is given by (25), the drop will be extremely sharp. For the change in the form of the front with time is compounded out of diffusion and propagation with variable velocity along descending curves, as shown schematically in Fig. 4. Since the latter propagation involves a more rapid motion to the right, the smaller n_h , it tends to steepen the front, and this steepening must continue until

⁵ W. G. Pfann and J. H. Scaff, paper presented at the Cambridge Meeting of the American Physical Society, June 16-18, 1949.

the diffusive spreading becomes sufficient to counterbalance it. It is not necessary, for the production of a steep front of this kind, that the decrease of j_e to zero be brought about with corresponding rapidity; even a gradual decrease of j_e will lead to a front which becomes steeper as it advances, and if the decrease of j_e is not too gradual a "shock front" will have developed after a short distance. The order of magnitude of the "shock front thickness" can be estimated by finding the value of the time Δt for which the diffusion distance $\Delta x_D = (2D \Delta t)^{1/2}$ equals the difference Δx_V between the drift distances of the holes at the top and bottom of the front, i.e., $\Delta x_V = [V(0) - V(n_h)]\Delta t$, where V is given by (4) and n_h is the height of the front. For this value of Δt ,

$$\Delta x_D = 2D/[V(0) - V(n_h)] \quad (48)$$

and this is presumably of the order of magnitude of the thickness of the front. If D is interpreted as $D_h = kT\mu_h/e$, which is good enough for the present purpose, this gives

$$\Delta x_D = \frac{2kT}{eE_0} \cdot \frac{1}{\left[1 - \frac{1}{1 + \nu(1 + \mu_h/\mu_e)}\right]^2} \quad (49)$$

Of course, this extremely sharp front can be realized only when the conditions of one-dimensional geometry are accurately fulfilled. When the geometry is made sufficiently ideal, observation of the thickness of the "shock front" can provide a valuable check on the validity of the basic assumptions of the theory such as the neglect of trapping.⁶

The author would like to express his indebtedness to many of his colleagues, and especially to J. Bardeen, J. R. Haynes, and W. van Roosbroeck, for many illuminating discussions of the topics covered in this paper.

⁶ The accompanying paper by W. Shockley, G. L. Pearson and J. R. Haynes describes some observations of this *shock wave* effect, though under conditions where $\nu \ll 1$, so that the thickness of the front as given by (49) is still fairly large.

On the Theory of the A-C. Impedance of a Contact Rectifier

By J. BARDEEN

THE a-c. impedance of the rectifying contact between a metal and a semiconductor is measured by superimposing a small a-c. current on a d-c. bias current. It is generally recognized¹ that an equivalent circuit consists of a parallel resistance and capacitance in series with a resistance as shown in Fig. 1. The parallel components represent the impedance of the barrier layer itself and depend on the d-c. bias current flowing. The series resistance is that of the body of the semiconductor. It has been shown theoretically by Spenke² that under quite general conditions the parallel capacitance and resistance are independent of frequency. Unfortunately Spenke's proof is highly mathematical and is also not readily available. The derivation of the impedance relations which is presented here is in some ways more general and gives more physical insight into the problem.

The method of analysis which is used is similar to that employed by Miss C. C. Dilworth³ for the d-c. case. Except for some obvious differences in sign, the theory is the same for *n*- and *p*-type semiconductors.⁴ We give the theory for the latter because the signs are a little simpler for positively charged holes than for negatively charged conduction electrons. Before the discussion of the theory of the a-c. impedance, a brief outline of Schottky's theory of the barrier layer will be given.

A rough schematic energy level diagram, based on Schottky's theory of the barrier layer at a contact between a metal and a *p*-type semiconductor, is illustrated in Fig. 2. The diagram is plotted upside down from the usual one in order to show the energy of holes increasing upward. The energy of electrons increases downward. In a defect or *p*-type semiconductor, such as Cu₂O, electrons are thermally excited to acceptor levels, charging the acceptors negatively, and leaving missing electrons or holes in the filled band. The holes are mobile and provide the conductivity. Electron states with energies lying above the Fermi level in the diagram, corresponding to lower energies for electrons, have a probability of more than one-half of

¹ For an outline of the theory of contact rectifiers together with references to the earlier literature, see H. C. Torrey and C. A. Whitmer, "Crystal Rectifiers," McGraw-Hill Book Company, Inc., New York, New York (1948).

² Eberhard Spenke, *Wiss. Veroff. Siemen's Konzern*, **20**, 40 (1941).

³ C. C. Dilworth, *Proc. Phys. Soc. London*, **60**, 315 (1947). A similar method was used earlier by H. A. Kramers, *Physica* **1**, 284 (1940), in a discussion of the diffusion of particles over potential barriers.

⁴ We suppose that only one type of carrier takes part in conduction.

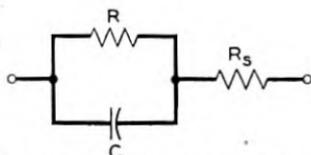


Fig. 1—Equivalent circuit for contact rectifier. The parallel components R and C represent the barrier layer itself and R_s represents the resistance of the body of the semiconductor.

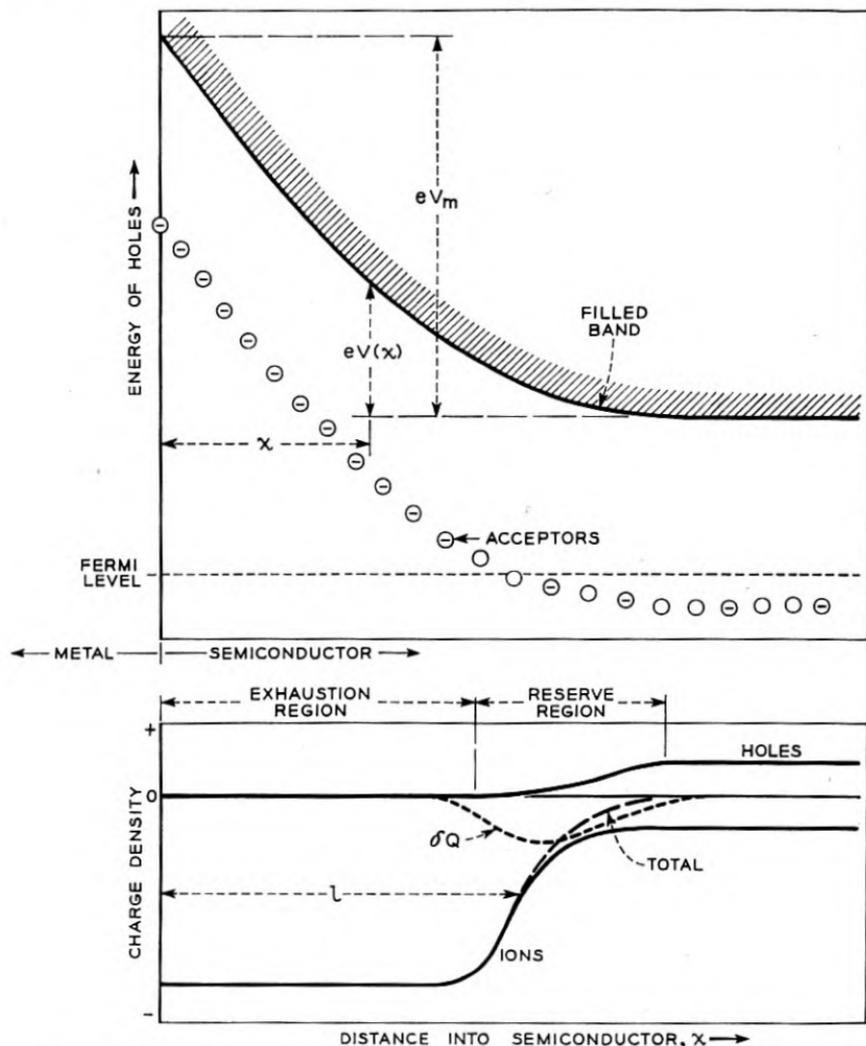


Fig. 2—Schematic energy level diagram of p-type semiconductor in contact with a metal. The diagram is plotted upside down from the usual way in order to show the energy of holes increasing upward. The energy of electrons increases downward. The lower diagram gives the density of charge in the barrier layer. In the body of the semiconductor the space charge of the holes is compensated by the space charge of the negatively charged acceptor ions. Holes are drained out of the barrier layer by the electric field, leaving the negative space charge of the acceptors. The rise in electrostatic potential in the barrier region results from this negative space charge together with the compensating positive charge on the metal. The capacitance of the barrier layer is approximately that of a parallel plate condenser with plate separation ℓ .

being occupied by electrons; those below the Fermi level are most likely unoccupied. Holes are depleted from the barrier layer, leaving the negative space charge of the acceptors. This negative space charge, together with the compensating positive charge on the metal, gives the potential energy barrier which impedes the flow of holes from semiconductor to metal. The thickness of the barrier layer may vary from 10^{-6} to 10^{-4} cm, depending on the materials forming the contact.

In drawing the diagram of Fig. 2 it has been assumed for simplicity that the concentration of acceptors is uniform over the region of interest. In the main body of the semiconductor only a few of the acceptors are charged. Throughout a large part of the barrier layer practically all acceptors are negatively charged and there are very few holes in the filled band. This part of the barrier layer has been called by Schottky the exhaustion region and is in our case a region of uniform space charge, as shown in the lower diagram of Fig. 2. The transition zone in which the concentration of holes is decreasing and the concentration of charged acceptors is increasing is called the reserve region.

In thermal equilibrium, with no applied voltage, the potential drop across the barrier layer, V_m , may be a fraction of a volt. If a voltage is applied in such a direction as to make the semiconductor positive relative to the metal, the effective height of the barrier is reduced and holes flow more easily from the semiconductor to the metal. This is the direction of easy flow. If a voltage is applied in the opposite direction the height of the barrier is increased for holes going from semiconductor to metal and remains unchanged, to a first approximation, for holes going from metal to semiconductor (actually electrons going from the filled band of the semiconductor to the metal). This is the reverse or high resistance direction.

If a voltage is applied in the reverse direction, and equilibrium is established, the thickness of the exhaustion layer increases. The reserve region keeps the same form but moves outward from the metal. A forward voltage decreases the thickness of the space charge layer.

The change in charge density corresponding to a small reverse voltage is shown schematically by the curve marked δQ in the lower diagram of Fig. 2. The maximum of δQ occurs where the total charge density is changing most rapidly with distance. If l is the distance from the metal to this maximum, the effective capacitance C , is approximately that of a parallel plate condenser with plate separation l and with the dielectric constant of the medium equal to that of the semiconductor. The capacitance decreases as l increases with a d-c. bias applied in the reverse direction and the capacitance increases with forward bias. Schottky⁵ has shown that information

⁵ Walter Schottky, *Zeits f. Phys.* **118**, 539 (1942).

about the concentrations of donors and acceptors can be obtained from the variation of capacitance with bias.

In the equivalent circuit of Fig. 1 the capacitance C is in parallel with the differential resistance, R , of the contact, and the parallel components are in series with the resistance R_s of the body of the semiconductor. Spenke showed that R and C are independent of frequency if the frequency is low enough so that the charge density is in equilibrium during the course of a cycle.

If the applied voltage is suddenly changed, it will take time for the charges to adjust to new equilibrium values. The time constant for the readjustment of charge of the carriers (holes in this case) is $\kappa\rho/4\pi$, where ρ is the resistivity (in e.s.u.) of the body of the semiconductor and κ is the dielectric constant, and is $\sim 10^{-10}$ sec. for a resistivity of 100 ohm cm. Even if a larger value of ρ is used, corresponding to a point in the reserve layer, the relaxation time for the carriers is very short.⁶ A much longer time may be required for readjustment of charge on the donor or acceptor ions, giving a variation of R and C at lower frequencies. If the barrier is nonuniform over the contact area, so that much of the current flows through low-resistance patches, the equivalent circuit may consist of a number of circuits like those of Fig. 1 in parallel. In this case, if an attempt is made to represent the contact by a single circuit of this form, it will be found that R and C vary with frequency.

The derivation of the current voltage characteristic for the general case of a time dependent applied voltage follows. The total current per unit area is the sum of contributions from conduction, diffusion, and displacement currents:

$$I(t) = \sigma E - eD(\partial n/\partial x) + (\kappa/4\pi)(\partial E/\partial t), \quad (1)$$

where

- $n(x,t)$ = concentration of holes;
- $\sigma = n(x,t)e\mu$ is the conductivity;
- e = magnitude of electronic charge;
- μ = mobility of holes;
- $D = \mu kT/e$ = diffusion coefficient;
- $V(x,t)$ = electrical potential;
- $E(x,t) = -\partial V(x,t)/\partial x$ = electric field strength.

The coordinate x extends into the semiconductor from the junction. Equation (1) may be written in the form

$$I(t) = ne\mu(-\partial V/\partial x) - \mu kT(\partial n/\partial x) - (\kappa/4\pi)(\partial^2 V/\partial x \partial t) \quad (1')$$

⁶ Another limit is the transit time of carriers through the barrier layer. This time is generally shorter than the relaxation time of the semiconductor.

The potential V is determined from the charge density, q , by Poisson's equation

$$\partial^2 V / \partial x^2 = -4\pi q / \kappa. \quad (2)$$

Since the charge density may be expressed in terms of $n(x,t)$ and the density of fixed charge, these two equations may be used to determine n and V when $I(t)$ is specified. Spenke eliminates the potential V between (1) and (2) and gets a rather complicated equation for n . We prefer to deal with Eq. (1) directly, to treat the potential $V(x,t)$ as a known function, and to solve for the concentration, $n(x,t)$.

The plane $x = 0$ is taken at the interface between metal and semiconductor and the plane $x = x_1$ just beyond the barrier layer in the semiconductor. It is assumed that $V = 0$ at $x = x_1$. Under thermal equilibrium conditions, with no current flowing, the hole concentration in the barrier layer varies as $\exp(-eV/kT)$, taking the values:

$$n = n_0 \text{ at } x = x_1 \quad (3a)$$

$$n = n_m = n_0 \exp(-eV_m/kT) \text{ at } x = 0, \quad (3b)$$

where n_0 is the equilibrium concentration in the body of the semiconductor and V_m is the height of the potential barrier. We suppose that the boundary conditions (3a) and (3b) also hold when a current is flowing and when there is an additional voltage, V_a , across the barrier layer. Our procedure is to solve Eq. (1) for $n(x,t)$, with $V(x,t)$ assumed known, and then to determine $I(t)$ in such a way that the boundary conditions are satisfied. The solution of Eq. (1) which satisfies (3b) is:

$$n(x, t) = n_0 \exp[-e(V - V_a)/kT] - \frac{1}{\mu kT} \int_0^x \left(I + \frac{\kappa}{4\mu} \frac{\partial^2 V'}{\partial x' \partial t} \right) \exp[e(V' - V)/kT] dx' \quad (4)$$

The prime indicates that the variable is x' rather than x . At $x = 0$, V is the sum of V_m and the applied potential, V_a :

$$V = V_a + V_m \text{ at } x = 0 \quad (5)$$

The current $I(t)$ is determined in such a way that (3a) is satisfied. Setting $x = x_1$, using (3a), and solving the resulting equation for $I(t)$, we get:

$$I(t) = \frac{\mu kT [n_0 \exp(eV_a/kT) - n_0 \exp(eV/kT)] - \int_0^{x_1} \frac{\kappa}{4\pi} \frac{\partial^2 V'}{\partial x' \partial t} \exp(eV'/kT) dx'}{\int_0^{x_1} \exp(eV'/kT) dx'} \quad (6)$$

Provided that the barrier height, $V_m + V_a$, is as much as several times kT/e ,⁷ the integrand in both integrals is largest near $x = 0$ and drops rapidly with increase in x . Where the integrand is large we may write to a sufficient approximation:

$$V = V_a + V_m - Fx, \quad (7)$$

where F is the field in the semiconductor at the interface. The approximation (7) may be used if kT/eF is small compared with the thickness of the barrier layer. The value of $\partial^2 V/\partial x \partial t$ is nearly constant over the important part of the integration and may be replaced by its value at $x = 0$ and taken out of the integral. The upper limit x_1 may be replaced by ∞ without appreciable error, so that we get finally:

$$I(t) = I_m(Q)(1 - \exp[-eV_a/kT]) + \partial Q/\partial t, \quad (8)$$

where

$$I_m(Q) = (4\pi e \mu Q n_0/\kappa) \exp[-eV_m/kT] \quad (9)$$

and

$$Q = \kappa F/4\pi \quad (10)$$

is the surface charge density at the metal interface.

The current $I_m(Q)$ has a simple interpretation; it is just the conduction current in the semiconductor at the interface resulting from the field F . In equilibrium, this conduction current is balanced by a diffusion current of equal magnitude and opposite sign. A voltage V_a applied in the reverse direction reduces the diffusion current at the interface as compared with the conduction current by the factor $\exp[-eV_a/kT]$. The current $\partial Q/\partial t$ is the displacement current at the interface.

Actually, the diffusion theory as given above is not complete. The Schottky effect, the lowering of the barrier by the image force, has been neglected. There may be appreciable tunneling through the barrier. There may be a patch field resulting from nonuniformity of the barrier. If the variations in the patch fields are not too large, the modification of current resulting from these factors depends only on the field at the metal and not on the form of the barrier at some distance from the metal. Thus we may expect the form (8) to be generally valid if $I_m(Q)$ is considered to be a general function of Q . Equation (10) is also of the form to be expected from the diode theory.¹ In the latter case, $I_m(Q)$ is the thermionic emission current from metal to semiconductor.

If the current is varying in time it is the instantaneous value of Q at

⁷ The value of kT/e at room temperature is .025 volts.

time t which is to be used in Eq. (10). At high frequencies, the charge at the interface need not be in phase with the applied voltage. If the frequency is low enough so that the charges maintain their equilibrium values during the course of a cycle, Q will be in phase with V and the parallel capacitance for unit area is simply:

$$C = dQ/dV. \quad (12)$$

The barrier layer may be represented by this capacitance in parallel with the d-c. differential resistance, R .

Both R and C may depend on the d-c. bias current flowing. Variations of R and C with frequency at moderate frequencies may result from large scale nonuniformities of the barrier such that the patch fields extend over a large fraction of the thickness of the barrier layer or from charge relaxation times associated with acceptors, donors or trapped carriers. At low frequencies, drift of ions may be involved.

Attempts which have been made to determine the variation of resistivity in the barrier layer from impedance data are invalid. It is not correct to take the impedance of an element of thickness dx to be

$$dx/[\sigma(x) + (j\omega\kappa/4\pi)]$$

and integrate over dx to obtain the impedance of the layer. This procedure omits terms arising from diffusion and changes of concentration in time. It is possible to obtain an integral of Eq. (1') if both sides are divided by $ne\mu$. Integrating over x from $x = 0$ to $x = x_1$, and using the boundary conditions (3a), (3b) and (5), we get

$$V_a = \int_0^{x_1} \frac{I(t) + (\kappa/4\pi)(\partial^2 V/\partial x \partial t)}{ne\mu} dx, \quad (13)$$

which means that the integral of the conduction current over the conductivity gives the applied voltage. This is consistent with the representation of the barrier by a resistance and capacitance in parallel.

ACKNOWLEDGMENTS

The author is indebted to W. Shockley, W. H. Brattain, and P. Debye for stimulating discussions and suggestions.

The Theory of p - n Junctions in Semiconductors and p - n Junction Transistors

By W. SHOCKLEY

In a single crystal of semiconductor the impurity concentration may vary from p -type to n -type producing a mechanically continuous rectifying junction. The theory of potential distribution and rectification for p - n junctions is developed with emphasis on germanium. The currents across the junction are carried by the diffusion of holes in n -type material and electrons in p -type material, resulting in an admittance for a simple case varying as $(1 + i\omega\tau_p)^{1/2}$ where τ_p is the lifetime of a hole in the n -region. Contact potentials across p - n junctions, carrying no current, may develop when hole or electron injection occurs. The principles and theory of a p - n - p transistor are described.

TABLE OF CONTENTS

1. Introduction
2. Potential Distribution and Capacity of Transition Region
 - 2.1 Introduction and Definitions
 - 2.2 Potential Distribution in the Transition Region
 - 2.3 The Transition-Region Capacity
 - 2.4 The Abrupt Transition
3. General Conclusions Concerning the Junction Characteristic
4. Treatment of Particular Models
 - 4.1 Introduction and Assumptions
 - 4.2 Solution for Hole Flow into the n -Region
 - 4.3 D-C. Formulae
 - 4.4 Total Admittance
 - 4.5 Admittance Due to Hole Flow in a Retarding Field
 - 4.6 The Effect of a Region of High Rate of Generation
 - 4.7 Patch Effect in p - n Junctions
 - 4.8 Final Comments
5. Internal Contact Potentials
6. p - n - p Transistors
- Appendix I A Theorem on Junction Resistance
- Appendix II Admittance in a Retarding Field
- Appendix III Admittance for Two Layers
- Appendix IV Time Constant for the Capacity of the Transition Region
- Appendix V The Effect of Surface Recombination
- Appendix VI The Effect of Trapping upon the Diffusion Process
- Appendix VII Solutions of the Space Charge Equation
- Appendix VIII List of Symbols

1. INTRODUCTION

AS IS well known, silicon and germanium may be either n -type or p -type semiconductors, depending on which of the concentrations N_d of donors or N_a of acceptors, is the larger. If, in a single sample, there is a transition from one type to the other, a rectifying photosensitive p - n junction is formed.¹ The theory of such junctions is in contrast to those

¹ For a review of work on silicon and germanium during the war see H. C. Torrey and C. A. Whitmer, *Crystal Rectifiers*, McGraw-Hill Book Company, Inc., New York (1948). P - n junctions were investigated before the war at Bell Telephone Laboratories by R. S. Ohl. Work on p - n junctions in germanium has been published by the group at Purdue

of ordinary rectifying junctions because, on both sides of the junction, both electron flow and hole flow must be considered. In fact, a major portion of the hole current may persist into the n -type region and vice-versa. In later sections we show how this feature has a number of interesting consequences, which we shall describe briefly in this introduction.

A p - n junction may act as an emitter in the transistor sense, since it can inject hole current into n -type material. The a-c. impedance of a p - n junction may exhibit a frequency dependence characterized by this diffusion of holes and of electrons. For high frequencies the admittance varies approximately as $(i\omega)^{1/2}$ and has comparable real and imaginary parts. When a p - n junction makes contact to a piece of n -type material containing a high concentration of injected holes, it acts like a semipermeable membrane and tends to come to a potential which corresponds to the hole concentration.

Although some results can be derived which are valid for all p - n junctions, the diversity of possible situations is so great and the solution of the equations so involved that it is necessary to illustrate them by using a number of special cases as examples. In general we shall consider cases in which the semiconductor may be classified into three parts, as shown in Fig. 1. The meaning of the transition region will become clearer in later sections; in general it extends far enough to either side of the point at which $N_d - N_a = 0$ so that the value of $|N_d - N_a|$ at its boundaries is not much smaller than in the low resistance parts of the specimen. As stated above, appreciable hole currents may flow into the n -region beyond the transition region. For this reason, the rectification process is not restricted to the transition region alone. We shall use the word *junction* to include all the material near the transition region in which significant contributions to the rectification process occur. It has been found that various techniques may be employed to make nonrectifying metallic contacts to the germanium; when this is properly done, the resistance measured between the metal terminals in a suitably proportioned specimen is due almost entirely to the rectifying junction up to current densities of 10^{-1} amp/cm².

directed by K. Lark-Horovitz: S. Benzer, *Phys. Rev.* 72, 1267 (1947); M. Becker and H. Y. Fan, *Phys. Rev.* 75, 1631 (1949); and H. Y. Fan, *Phys. Rev.* 75, 1631 (1949). Similar junctions occur in lead sulfide according to L. Sosnowski, J. Starkiewicz and O. Simpson, *Nature* 159, 818 (1947), L. Sosnowski, *Phys. Rev.* 72, 641 (1947), and L. Sosnowski, B. W. Soole and J. Starkiewicz, *Nature* 160, 471 (1947). The theory described here has been discussed in connection with photoelectric effects in p - n junctions by F. S. Goucher, Meeting of the American Physical Society, Cleveland, March 10-12, 1949 and by W. Shockley, G. L. Pearson and M. Sparks, *Phys. Rev.* 76, 180 (1949). For a general review of conductivity in p - and n -type silicon see G. L. Pearson and J. Bardeen, *Phys. Rev.* 75, 865 (1949), and J. H. Scaff, H. C. Theuerer and E. E. Schumacher, *Jl. of Metals*, 185, 383 (1949) and W. G. Pfann and J. H. Scaff, *Jl. of Metals*, 185, 389 (1949). The latter two papers also discuss photo-voltaic barriers. The most recent and thorough theory for frequency effects in metal semiconductor rectifiers is given elsewhere in this issue (J. Bardeen, *Bell Sys. Tech. Jl.*, July 1949).

Even for distributions of impurities as simple as those shown in part (b) there are two distinctly different types of behavior of the electrostatic potential in the transition region, each of which may be either rectifying or nonrectifying. The requirement that the junction be rectifying can be stated in terms of the current distribution, certain cases of which are shown in (c). The total current, from left to right, is I , the hole and electron currents being

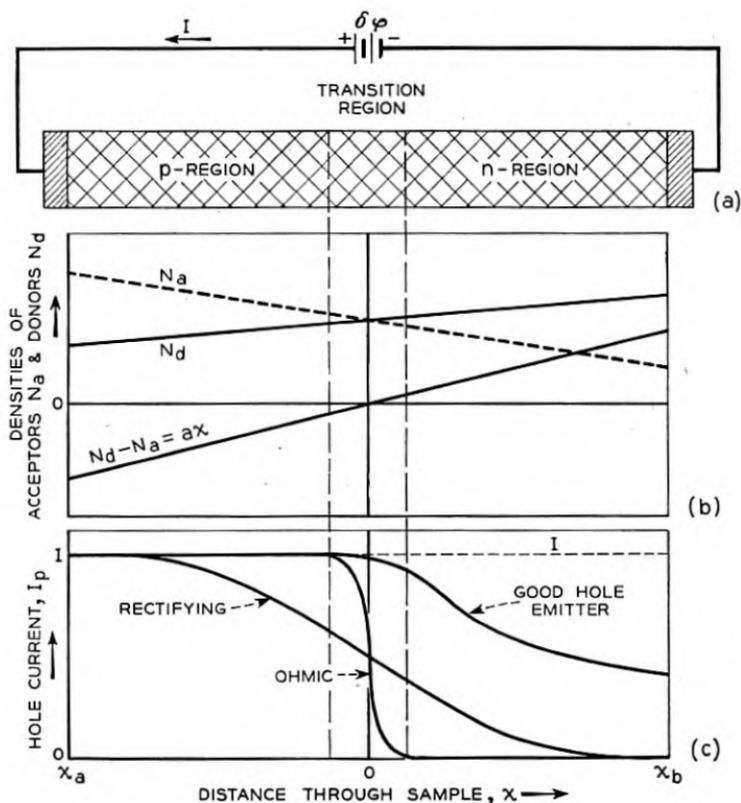


Fig. 1—The *p-n* junction.

(a) Schematic view of specimen, showing non-rectifying end contacts and convention for polarities of current and voltage.

(b) Distribution of donors and acceptors.

(c) Three possible current distributions.

I_p and I_n , with $I = I_p + I_n$. Well away from the junction in the *p*-type material, substantially all of the current is carried by holes and $I_p = I$; similarly, deep in the *n*-type material $I_n = I$ and $I_p = 0$. In general in a nonrectifying junction, the hole current does not penetrate the *n*-type material appreciably whereas in the rectifying junction it does. Under some conditions the major flow across the junction will consist of holes; such

cases are advantageous as emitters in transistor applications using n -type material for the *base*.

Where the hole current flows in relatively low resistance n -type material, it is governed by the diffusion equation and the concentration falls off as $\exp(-x/L_p)$ where L_p is the diffusion length:

$$L = \sqrt{D\tau_p}.$$

Here D is the diffusion constant for holes and τ_p their mean lifetime. The lifetime may be controlled either by surface recombination² or volume recombination. Surface recombination is important if the specimen has a narrow cross-section.

Under a-c. conditions, the diffusion current acquires a reactive component corresponding to a capacity. In addition, a capacitive current is required to produce the changing potential distribution in the transition region itself.

In the following sections we shall consider the behavior of the junction analytically, treating first the potential distribution in the transition region and the charges required change the voltage across it in a pseudo-equilibrium case. We shall then consider d-c. rectification and a-c. admittance.

2. POTENTIAL DISTRIBUTION AND CAPACITY OF TRANSITION REGION

2.1 Introduction and Definitions

We shall suppose in this treatment that all donors and acceptors are ionized (a good approximation for Ge at room temperature) so that we have to deal with four densities as follows:

n = density of electrons in conduction band

p = density of holes in valence-bond band

N_d = density of donors

N_a = density of acceptors

The total charge density is

$$\rho = q(p - n + N_d - N_a), \quad (2.1)$$

where q is the electronic charge. We shall measure electrostatic potential ψ in the crystal, as shown in Fig. 2, from such a point, approximately³ midway in the energy gap, that if the Fermi level φ is equal to ψ , the concentrations of holes and electrons are equal to the concentration $n_i = p_i$ char-

² H. Suhl and W. Shockley *Phys. Rev.* 75 1617 (1949).

³ A difference in effective masses for holes and electrons will cause a shift of ψ from the midpoint between the bands,

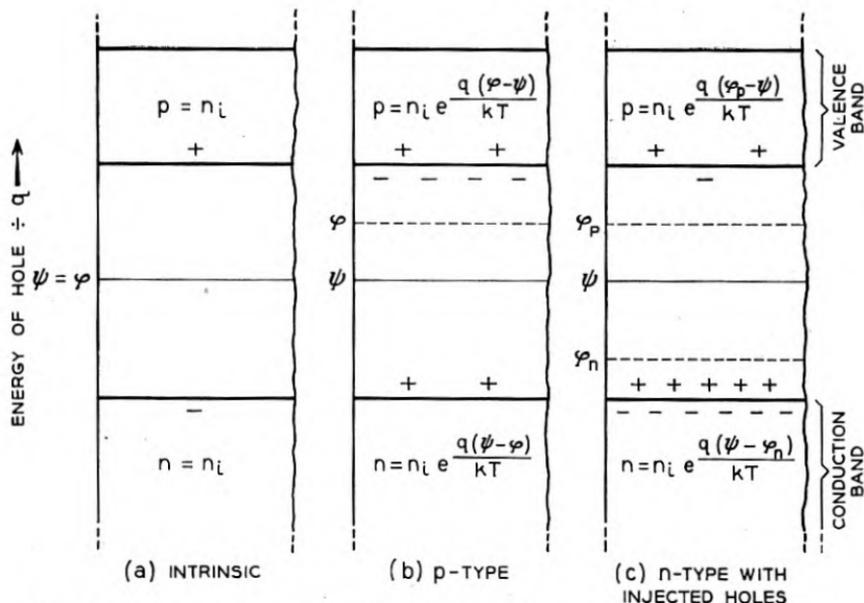


Fig. 2—Electrostatic potential ψ , Fermi level φ and quasi Fermi levels φ_p and φ_n .
 (In order to show electrostatic potential and energies on the same ordinates, the energies of holes, which are minus the energies of electrons, are plotted upwards in the figures in this paper.)

acteristic of a pure sample. For an impurity semi-conductor we shall have, as shown in (b),

$$p = n_i e^{q(\varphi-\psi)/kT} \tag{a}$$

$$n = n_i e^{q(\psi-\varphi)/kT}, \tag{b}$$

where q is the electronic charge. Accordingly,

$$\rho = q\{N_d - N_c + 2n_i \sinh [q(\varphi - \psi)/kT]\}. \tag{2.3}$$

When the hole and electron concentrations do not have their equilibrium values, because of hole or electron injection or production of hole-electron pairs by light, etc., it is advantageous to define two non-equilibrium quasi Fermi levels φ_p and φ_n by the equations

$$p = n_i e^{q(\varphi_p-\psi)/kT} \tag{a}$$

$$n = n_i e^{q(\psi-\varphi_n)/kT} \tag{b}$$

as indicated in Fig. 2 (c). In terms of φ_p and φ_n , the hole and electron currents take the simple forms:

$$I_p = -q[D\nabla p + \mu p \nabla \psi] = -q\mu p \nabla \varphi_p \tag{2.5}$$

$$I_n = bq[D\nabla n - \mu n \nabla \psi] = -qb\mu n \nabla \varphi_n \tag{2.6}$$

where the mobility μ and diffusion constant D for holes are related by Einstein's equation

$$\mu = qD/kT \quad (2.7)$$

and b is the ratio of electron mobility to hole mobility.⁴

Under equilibrium conditions $\varphi_p = \varphi_n = \varphi$ where φ is independent of position. Under those conditions, I_p and I_n are both zero according to equations (2.5) and (2.6). The electrostatic potential ψ , however, will not in general be constant and there will be unbalanced charge densities throughout the semiconductor. We shall consider the nature of the conditions which determine ψ for a general case and will later treat in detail the behavior of ψ for p - n junctions.

For equilibrium conditions, there is no loss in generality in setting φ arbitrarily equal to zero. The charge density expression (2.3) may then be rewritten as

$$\rho = \rho_d - \rho_i \sinh u \quad (2.8)$$

where

$$u \equiv q\psi/kT, \quad \rho_i \equiv 2n_iq, \quad \rho_d \equiv q(N_d - N_a) \quad (2.9)$$

In equation (2.8) ρ_d and u and, consequently, ρ may be functions of position. The potential ψ must satisfy Poisson's equation which leads to the equation

$$\nabla^2 \psi = -4\pi\rho/\kappa \quad (2.10)$$

where κ is the dielectric constant, (2.10) can be rewritten as

$$\nabla^2 u = \frac{4\pi q \rho_i}{kT\kappa} \left(\sinh u - \frac{\rho_d}{\rho_i} \right) \quad (2.11)$$

What this equation requires in physical terms is that the electrostatic potential produces through (2.8) just such a total charge density ρ that this charge density, when used in Poisson's Equation (2.10), in turn produces ψ . It seems intuitively evident that the equation for u will always have a physically meaningful solution; no matter how the charge density ρ_d due to the impurities varies with position, the holes and electrons should be able to distribute themselves so that equilibrium is produced. For a one-dimensional case, it is not difficult to prove that a unique solution exists for $u(x)$ for any $\rho_d(x)$ (Appendix VII).

⁴ We prefer b in comparison to c for this ratio since c for the speed of light also occurs in formulae involving b .

The coefficient in (2.11) has the dimensions of (length)⁻² leading us to define a quantity

$$\begin{aligned} L_D &= \sqrt{\kappa kT/4\pi q\rho_i} = \sqrt{\kappa kT/8\pi q^2 n_i} \\ &= 2.1 \times 10^{-3} \text{ cm for Si with } \kappa = 12.5,^5 n_i = 2 \times 10^{10} \text{ cm}^{-3} \quad (2.12) \\ &= 6.8 \times 10^{-5} \text{ cm for Ge with } \kappa = 19,^6 n_i = 3 \times 10^{13} \text{ cm}^{-3} \end{aligned}$$

where the subscript D for Debye emphasizes the similarity of L_D to the characteristic length in the Debye-Hückel theory of strong electrolytes. The meaning of the Debye length is apparent from the behavior of the solution in a region where ρ_d is constant, and u differs only slightly from the value u_0 which gives neutrality, with $\rho_i \sinh u_0 = \rho_d$. Under these conditions,

$$\frac{d^2 u}{dx^2} = (L_D^{-2} \cosh u_0)(u - u_0) \quad (2.13)$$

so that $u - u_0$ varies as $\exp(\pm x\sqrt{\cosh u_0}/L_D)$. In general, we shall be interested in cases in which the deviation of u from u_0 decays to a small value in one direction. It is evident that the distance required to reduce the deviation to $1/e$ is $L_D/\sqrt{\cosh u_0}$. If only small variations in ρ_d occur within a distance $L_D/\sqrt{\cosh u_0}$, then the semiconductor will be substantially neutral. However, if a large variation of ρ_d occurs in this distance, a region of local space charge will occur. These two cases are illustrated in connection with the potential distribution in a p-n junction.

2.2 Potential Distribution in the Transition Region⁷

We shall discuss the case shown in Fig. 1 for which the charge density due to donors and acceptors is given by

$$N_d - N_a = ax \quad (2.14)$$

This relationship defines a characteristic length L_a given by

$$L_a = n_i/a \quad (2.15)$$

If $L_a \gg L_D$, the condition of electrical neutrality is fulfilled (Appendix VII) and u satisfies the equation

$$\sinh u = \rho_d/\rho_i = ax/2n_i = x/2L_a$$

⁵ J. F. Mullaney, *Phys. Rev.* 66, 326 (1944).

⁶ H. B. Briggs and W. H. Brattain, *Phys. Rev.*, 75, 1705 (1949).

⁷ Potential distributions in rectifying junctions between semiconductors and metals have been discussed by many authors, in particular N. F. Mott, *Proc. Roy. Soc.* 171A, 27 (1939) and W. Schottky *Zeits. f. Physik* 113, 367 (1939) 118, 539 (1942) and elsewhere. A summary in English of Schottky's papers is given by J. Joffe, *Electrical Communications* 22, 217 (1945). All such theories are in principle similar in involving the solution of equations like (2.11). See, for example, H. Y. Fan, *Phys. Rev.* 62, 388 (1942).

On the other hand, if $L_D \gg L_a$, a large change in impurity concentration occurs near $x = 0$ without compensating electron and hole densities occurring. Mathematically, we find that (2.11) can be expressed in the form

$$\frac{d^2 u}{dy^2} = \frac{1}{K^2} (-y + \sinh u) \quad (2.16)$$

and

$$y = x/2L_a, \quad K = L_D/2L_a \quad (2.17)$$

In Appendix VII, it is verified that the appropriate solution for $K \ll 1$ is that giving local neutrality, $u = \sinh^{-1} y$; while for $K \gg 1$, there is space charge as described below.

For $L_D \gg L_a$, or $K \gg 1$, there is a space charge layer in which $N_d - N_a$ is uncompensated. To a first approximation, we can neglect the electron and hole space charge in the layer and obtain, by integrating twice,

$$\psi = -\frac{2\pi q a x^3}{3\kappa} + a_2 x, \quad (2.18)$$

where we have chosen the zero of potential as the value at $x = 0$, a condition required by the symmetry between $+x$ and $-x$ of (2.14). Although the potential rise is steep in the layer, $d\psi/dx$ should be small at the point x_m where the neutral n -type material begins. As an approximation we set $d\psi/dx = 0$ at $x = x_m$:

$$\frac{d\psi}{dx} = -\frac{2\pi q a x_m^2}{\kappa} + a_2 = 0; \quad (2.19)$$

this leads to a value for a_2 which may be inserted in (2.18) to evaluate ψ at x_m :

$$\psi_m = \frac{4\pi q a x_m^3}{3\kappa} = \frac{4\pi q}{3\kappa a^2} (a x_m)^3 = \frac{4\pi q}{3\kappa a^2} n_m^3 \quad (2.20)$$

where $n_m = a x_m$ is the density of electrons required to neutralize $N_d - N_a = a x_m$ at the edge of the space-charge layer. This value of n_m must correspond to that associated with ψ_m by (2.2)

$$n_m = n_i e^{q\psi_m/kT}. \quad (2.21)$$

We thus have two equations relating ψ_m and n_m and the parameter "a." To solve them we plot $\ln \psi_m$ versus $\ln n_m$ as shown in Fig. 3. On this figure the relationship

$$\begin{aligned} \psi_m &= \frac{4\pi q}{3\kappa} \frac{n_m^3}{a^2} \\ &= 3.18 \times 10^{-8} \frac{n_m^3}{a^2} \text{ volts for Ge} \\ &= 4.83 \times 10^{-8} \frac{n_m^3}{a^2} \text{ volts for Si} \end{aligned} \quad (2.22)$$

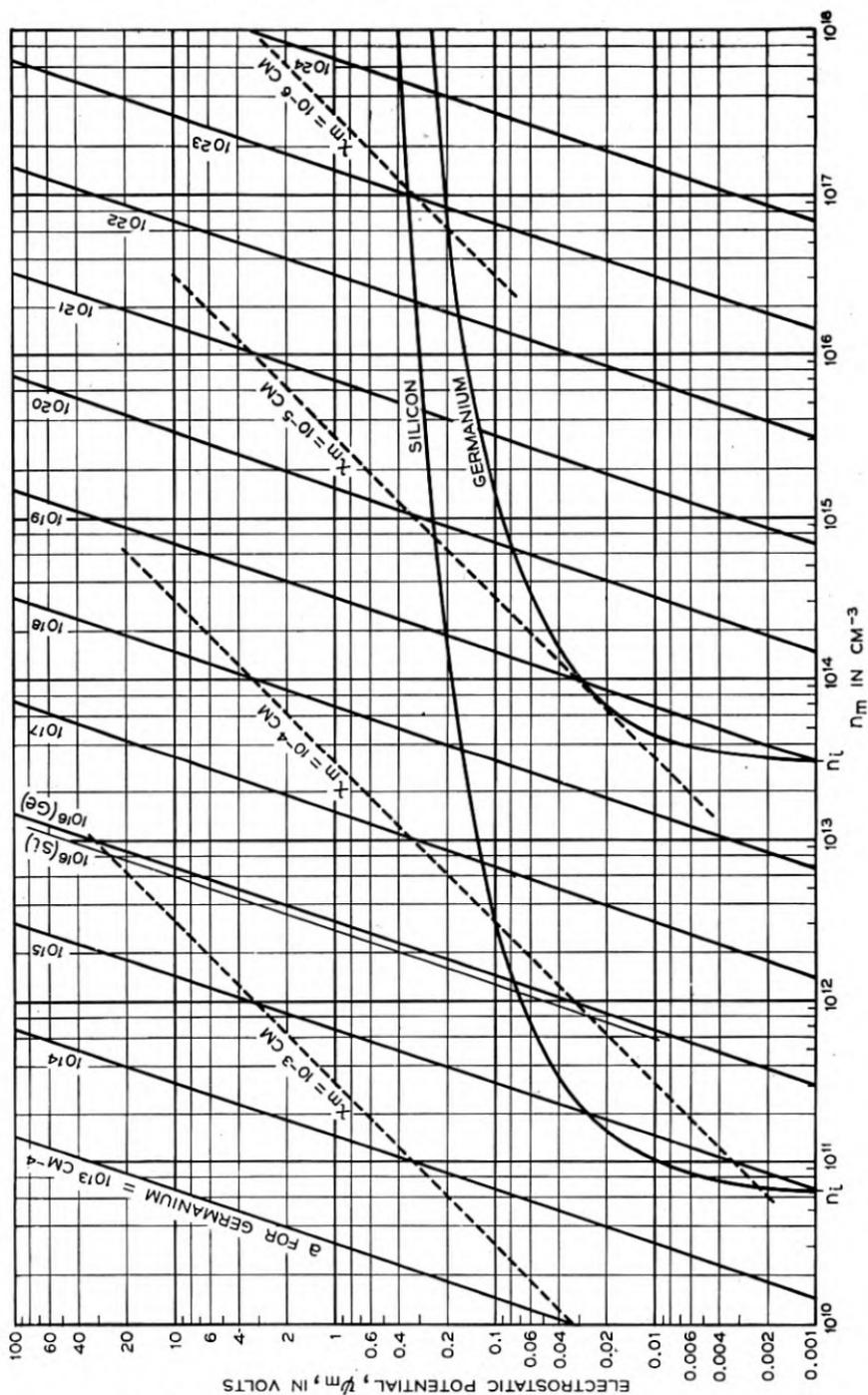


Fig. 3—Solutions for the boundaries of the space-charge region.

becomes a family of straight lines with "a" as a parameter. (Only $a = 10^{16}$ cm^{-4} is shown for Si, all the other lines being for Ge.) The half thickness $x_m (= n_m/a)$ of the space-charge region is also shown. Solutions are obtained when these lines cross the curves $n_m = n_i \exp(q\psi_m/kT)$, which are shown for room temperature. The condition that the intersection lie well to the right on the curve is equivalent to $K \gg 1$. For two Si samples cut from a melt, a was determined from measurements of conductivity⁸ and was about 10^{15} to 10^{16} cm^{-4} . For these, the space charge region has a half-width x_m of more than 10^{-4} cm. For other temperatures, the curves can be appropriately translated.⁹

In Fig. 4(a) we show the limiting potential shapes:

$$ax = 2n_i \sinh \frac{q\psi}{kT} \quad \text{for } K \ll 1 \quad (2.23)$$

$$\psi = (\psi_m/2)(-(x/x_m)^3 + 3(x/x_m)) \quad \text{for } K \gg 1 \quad (2.24)$$

In Fig. 4(b) the charge densities are shown. For the space-charge case, $|N_d - N_a|$ is greater than n or p . For a higher potential rise, i.e. larger ψ_m , the discrepancy would be greater and $N_d - N_a$ would be unneutralized except near x_m .

2.3 The Transition-Region Capacity

When the voltage across the junction is changing, a flow of holes and electrons is required to alter the space charge in the transition region. We shall calculate the charge distribution in the transition region with the aid of a pseudo-equilibrium model in which the following processes are imagined to be prevented: (1) hole and electron recombination, (2) electron flow across the p -region contact at x_a (Fig. 1), (3) hole flow across the n -region boundary at x_b . Under these conditions holes which flow in across x_a must remain in the specimen. If a potential $\delta\varphi$ is applied at the p end, then holes will flow into the specimen until φ_p has increased by $\delta\varphi$ so that the holes inside are in equilibrium with the contact which applies the potential. Since the specimen as a whole remains neutral, an equal electron flow will occur at x_b . When the specimen arrives at its pseudo-equilibrium steady-state, the potential distribution will be modified in the transition region and the number of holes in this region will be different from the number present under conditions of true thermal equilibrium. The added number of holes is proportional to $\delta\varphi$ for small values of $\delta\varphi$ and thus acts like the charge on a condenser. Our problem in this section is to calculate how this charge depends

⁸ Unpublished data of W. H. Brattain and G. L. Pearson.

⁹ The effect of unionized donors and acceptors can also be included by letting n_i include the properly weighted donor states and p_i , the acceptor states.

upon $\delta\phi$ for various types of transition regions and to express the result as a capacity.

The justification for this pseudo-equilibrium treatment is as follows: Under actual a-c. conditions the potential drop in the *p*- and *n*-regions themselves are small because of their high conductivity so that most of the po-

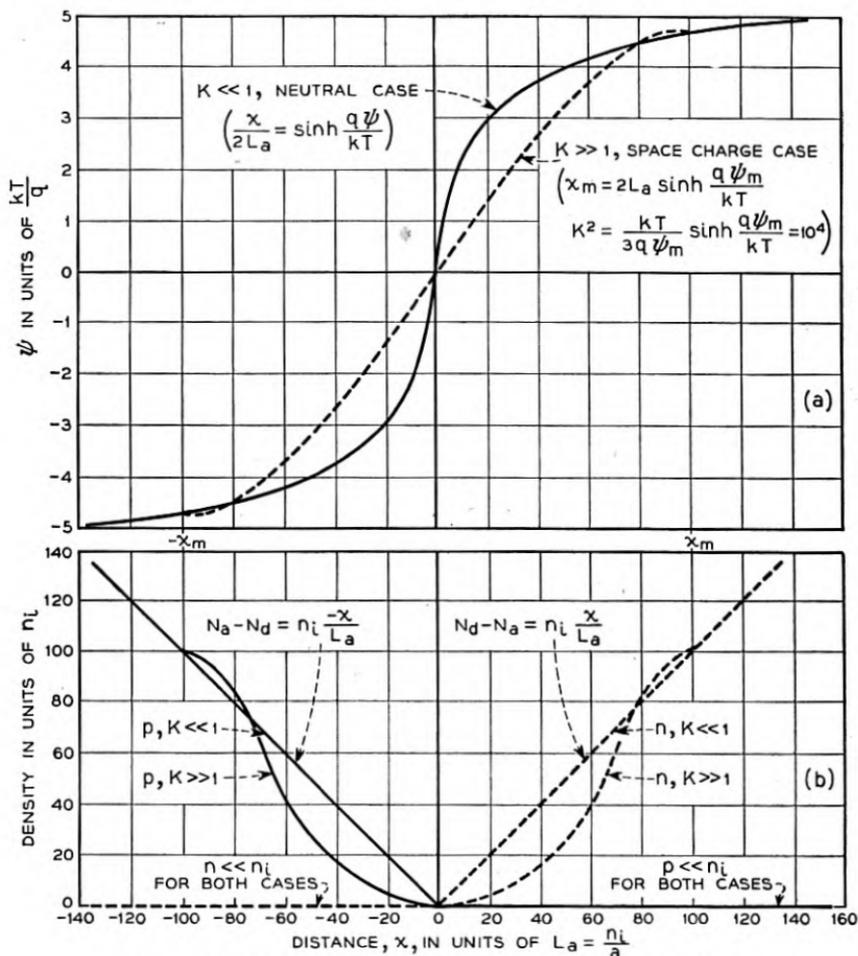


Fig. 4—Electrostatic potential and densities for *p-n* junctions.

tential drop occurs across the transition region. On the *p*-side of the transition region a large supply of holes is available to modify the potential and the fact that a current is flowing across the junction disturbs their concentration negligibly; the electrons on the *n*-side are similarly situated. Hence the distribution of holes and electrons in the transition region will be much the same as for the pseudo-equilibrium case. The question of how the hole

current required to change the potential distribution in the transition region is related to other hole currents is discussed in Section 4.1.

Under our assumptions, after the voltage $\delta\varphi$ is applied, a steady state is reached involving no current hence $\nabla\varphi_p = \nabla\varphi_n = 0$. Consequently, both φ_p and φ_n are constant and

$$\varphi_p - \varphi_n = \delta\varphi \quad (2.25)$$

since the holes are being supplied from a source at a potential $\delta\varphi$ higher than for the electrons.

We shall then have

$$p = n_i e^{q(\varphi_p - \psi)/kT} = n_1 e^{q(\varphi_1 - \psi)/kT} \quad (2.26)$$

$$n = n_i e^{q(\psi - \varphi_n)/kT} = n_1 e^{q(\psi - \varphi_1)/kT} \quad (2.27)$$

where

$$\varphi_1 = (\varphi_p + \varphi_n)/2, \quad \varphi_p = \varphi_1 + \delta\varphi/2, \quad \varphi_n = \varphi_1 - \delta\varphi/2 \quad (2.28)$$

and

$$n_1 = n_i e^{q\delta\varphi/2kT}. \quad (2.29)$$

Thus the effect of applying the potential $\delta\varphi$ in the pseudo-equilibrium case is equivalent to changing n_i to n_1 just as if the energy gap had been reduced by $q\delta\varphi$.

In the p -region, $n \ll p$ and so that $p = -ax$ is a good approximation. Similarly, in the n -region, we set $n = ax$. Hence we have in the p -region

$$\psi = \varphi_1 + (\delta\varphi/2) - (kT/q) \ln(-ax/n_i) \quad (2.30)$$

and in the n -region

$$\psi = \varphi_1 - (\delta\varphi/2) + (kT/q) \ln(ax/n_i). \quad (2.31)$$

Hence the effect of $\delta\varphi$ is to shift ψ in the p -region upwards by $\delta\varphi$ compared to ψ in the n -region. This is an example of the general result that $\psi - \varphi_p$ tends to remain constant at a given point in the p -region no matter what disturbances occur and $\psi - \varphi_n$ tends to remain constant in the n -region.

The Capacity for the Neutral Case $K \ll 1$

For the neutral case, we calculate the total number of holes, P , between x_a and x_b as a function of $\delta\varphi$. The charge of these holes is qP and the effective capacity is $q dP/d\delta\varphi$. As explained above, we are really interested in the change in number of holes in the transition region. However, the value of P is relatively insensitive to the location of the limits x_a and x_b so long as they lie in regions where the conductivity approaches the maximum values in the

p- and *n*-regions. In the following calculations, we shall consider a unit area of the junction so that values of *P* and of capacity are on a unit area bases.

The value of *P* is obtained by integrating *p dx* making use of the neutrality condition to establish the functional relationship between *p* and *x*. The neutrality condition can be written as

$$ax = 2n_1 \sinh \frac{q(\psi - \varphi_1)}{kT} \equiv 2n_1 \sinh u \quad (2.32)$$

where $u \equiv q(\psi - \varphi_1)/kT$ and

$$p = n_1 e^{q(\varphi_1 - \psi)/kT} \equiv n_1 e^{-u} \quad (2.33)$$

$$n = n_1 e^{+u} \quad (2.34)$$

so that the value of *P* can be obtained by changing variables from *x* to *u*:

$$\begin{aligned} P &= \int_{x_a}^{x_b} p \, dx = \int_{u_a}^{u_b} p(2n_1/a) \cosh u \, du \\ &= (n_1^2/a) \int_{u_a}^{u_b} [1 + e^{-2u}] \, du = (n_1^2/a)[u_b - u_a + (e^{-2u_a} - e^{-2u_b})/2] \end{aligned} \quad (2.35)$$

For the cases of practical interest, the value of *p* at *x* = *x_a*, denoted by *p_a*, and the value of *n* at *x* = *x_b*, denoted by *n_b*, will both be large compared to *n₁*. Consequently, we conclude that

$$u_a = -\ln(p_a/n_1) \text{ and } u_b = \ln n_b/n_1$$

are both larger than unity in absolute value but probably less than twenty for a reasonable variation of impurity between *x_a* and *x_b*. (For example for a change in potential of 0.2 volts such as would occur between *p*- and *n*-type germanium, *u_a* and *u_b* would each be about 4 in magnitude.) Hence we obtain for *P*,

$$\begin{aligned} P &= (n_1^2/2a)(2(u_b - u_a) + (p_a/n_1)^2 - (n_1/n_b)^2) \\ &\cong p_a^2/2a + (n_1^2/a)(u_b - u_a) \end{aligned} \quad (2.36)$$

where we have neglected $(n_1/n_b)^2$ which is $\ll 1$ and the negligible compared to $u_b - u_a$. The term $p_a^2/2a$ is simply the integrated acceptor-minus-donor density in the *p*-region, as may be seen as follows:

$$\int_{x_a}^0 (N_a - N_d) \, dx = \int_{x_a}^0 (-ax) \, dx = ax_a^2/2 = p_a^2/2a. \quad (2.37)$$

The second term in (2.36) is essentially the sum of the holes of the right of *x* = 0 plus the electrons to the left of *x* = 0, whose charge is also com-

pensated by holes. The total number of holes can be expressed in terms of $\delta\varphi$ through the dependence of n_1 on $\delta\varphi$. The second term is thus

$$(n_i^2/a)[\ln(n_b/n_1) + \ln(p_a/n_1)] \\ = (n_i^2/a)e^{q\delta\varphi/kT} \cdot [\ln(n_b p_a/n_i^2) - q\delta\varphi/kT] \quad (2.38)$$

Hence for a small change $d\delta\varphi$ in $\delta\varphi$, the change in charge $dQ = q dP$ and the capacity C are given by

$$C = \frac{dQ}{d\delta\varphi} = \frac{q^2}{kT} \frac{n_i^2}{a} [\ln(n_b p_a/n_i^2) - (q\delta\varphi/kT) - 1]. \quad (2.39)$$

This capacity can be reexpressed in terms of the difference in ψ between x_a and x_b : When $\delta\varphi = 0$, corresponding to the thermal equilibrium case, we have

$$p_a n_b = n_i^2 e^{q(\psi_b - \psi_a)/kT} \quad (2.40)$$

Using this together with the definitions of L_D and L_a we obtain

$$C = \frac{\kappa[q(\psi_b - \psi_a - \delta\varphi)/kT - 1] e^{q\delta\varphi/kT}}{4\pi(2L_D^2/L_a)} \quad (2.41)$$

In this expression ψ_a and ψ_b are the potentials when $\delta\varphi = 0$; so that

$$\psi_b - (\psi_a + \delta\varphi)$$

is thus the increase in potential in going from x_a to x_b when $\delta\varphi$ is applied.

For thermal equilibrium, $\delta\varphi = 0$ and, as discussed above, the term in $\psi_b - \psi_a$ will be about 10. Hence, using the definition $K = L_D/2L_a$, we have

$$C \cong \kappa/4\pi(4KL_D/10) \quad (2.42)$$

For $K \ll 1$, the case for which this formula is valid, C will be the capacity of a condenser whose dielectric layer is much less than L_D thick.

Capacity for Space Charge Case, $K \gg 1$

As discussed in connection with (2.30) and (2.31), the applied potential $\delta\varphi$ reduces the increase ($= 2\psi_m$) in ψ between the p -region and the n -region by $\delta\varphi/2$ on each side of $x = 0$. This is accomplished by a narrowing of the space charge layer by δx_m on each side where (according to (2.20))

$$\delta\psi_m = -\delta\varphi/2 = 4\pi q a x_m^2 \delta x_m / \kappa \quad (2.43)$$

The decrease in width δx_m brings with it an increase in number of holes $-ax \delta x_m$ per unit area of the junction on the p -side and an equal number of electrons on the n -side. Thus a charge of holes per unit area of $\delta Q = -qax_m \delta x_m$ must flow in from the left. The capacity per unit area is, therefore,

$$C = \delta Q / \delta\varphi = qax_m \delta x_m / \delta\varphi = \kappa/4\pi 2x_m \quad (2.44)$$

corresponding to a condenser of thickness $2x_m$. It is evident that formula (2.44) will hold for a small change $d\delta\varphi$ superimposed on a large bias $\delta\varphi$ provided that $2x_m$ is the thickness of the space charge region under the conditions when $\delta\varphi$ is applied. If $\psi_{n,0}$ is the value of ψ for $\delta\varphi = 0$, then $\psi_m = \psi_{n,0} - \delta\varphi/2$; and C will vary as

$$C = \kappa[4\pi qa/3\kappa(\psi_{n,0} - \delta\varphi/2)]^{1/3}/8\pi \quad (2.45)$$

so that $1/C^3$ should plot as a straight line versus $\delta\varphi$ with slope

$$- (8\pi/\kappa)^3(3\kappa/8\pi qa) = - \frac{192\pi^2}{\kappa^2 qa}. \quad (2.46)$$

In addition to the holes which flow to account for the change in ψ_m , the concentration of holes in the n -region will be increased by a factor $\exp(q\delta\varphi/kT)$. However, this increase does not lie in the transition region; we shall consider it later, in Section 4, in connection with a-c. admittance.

Comparison of the Two Capacities

It is instructive to compare the two capacities just derived. We suppose that for one value of n_i we have $K \gg 1$ so that the space charge solution is good. For this case we choose $x_a = -x_m$ and $x_b = +x_m$ so as to bound the space charge layer. We then imagine n_i to be increased, either by raising the temperature or by applying a potential difference $\delta\varphi$. The capacity then changes from

$$C_{\text{sp. chg.}} = \kappa/8\pi x_m \text{ to } C_{\text{neut.}} = 5\kappa/8\pi KL_D \quad (2.47)$$

(i.e., from (2.44) to (2.42)) so that the ratio is

$$\frac{C_{\text{neut.}}}{C_{\text{sp. chg.}}} = \frac{5x_m}{KL_D} \quad (2.48)$$

For $K < 1$, this ratio is large, both because of K in the denominator and because $x_m > L_a$ so that $x_m/L_D > L_a/L_D = 1/2 K$.

In Section 4.4 we shall compare these capacities with that due to diffusion of holes and electrons beyond the transition region.

2.4 The Abrupt Transition

For completeness we shall consider the case in which the impurity concentration changes abruptly from p_p to n_n at $x = 0$. For this case the potential in the space-charge layer will be of the parabolic type discussed by Schottky, the potentials varying as

$$\psi = (2\pi/\kappa)q p_p(x - x_p)^2 + \text{constant}, \quad x < 0 \quad (2.49)$$

$$\psi = -(2\pi/\kappa)q n_n(x - x_n)^2 + \text{constant}, \quad x > 0 \quad (2.50)$$

where $x_p < 0$ and $x_n > 0$ are the ends of the space-charge layer in the p - and n -regions. The gradient of potential at $x = 0$ must be equal for the two layers leading to

$$-p_p x_p = n_n x_n \quad (2.51)$$

so that if the total width of the space charge layers is $W = x_n - x_p$, it follows that

$$x_p = -n_n W / (n_n + p_p) \text{ and } x_n = p_p W / (n_n + p_p). \quad (2.52)$$

The potential difference across the layer, which is $\psi_b - \psi_a$ is

$$\psi_b - \psi_a = (2\pi q / \kappa) (p_p x_p^2 + n_n x_n^2) = [2\pi q p_p n_n / \kappa (p_p + n_n)] W^2 \quad (2.53)$$

If $p_p \gg n_n$ this reduces to

$$\psi_b - \psi_a = 2\pi q n_n W^2 / \kappa \quad (2.54)$$

the formula given by Schottky, which should be appreciable in this case, for which all the voltage drop occurs in the n -region.

The capacity for the abrupt transition will be

$$C = \kappa / 4\pi W \quad (2.55)$$

where W is obtained by solving (2.53). For this case $(1/C)^2$ should plot as a straight line versus $\psi_b - \psi_a$:

$$\frac{1}{C^2} = [8\pi(p_p + n_n) / \kappa q p_p n_n] (\psi_b - \psi_a). \quad (2.56)$$

3. GENERAL CONCLUSIONS CONCERNING THE JUNCTION CHARACTERISTIC

In this section we shall consider direct current flow through the junction and shall derive the results quoted in Fig. 1 relating the current distribution to the characteristics of the junction. We shall suppose that holes and electrons are thermally generated in pairs at a rate g and recombine at a rate $rn\dot{p}$ so that the net rate of generation per unit volume is

$$(\text{net rate of generation}) = g - rn\dot{p},$$

which vanishes at equilibrium. Obviously, $g = rn\dot{n}_i^2$. If relatively small concentrations δp and δn of holes and electrons are present in excess of the equilibrium values, the net rate of generation is

$$\delta\dot{p} = \delta\dot{n} = g - r(n + \delta n)(p + \delta p) = -rn\delta p - r\dot{p}\delta n \quad (3.1)$$

This is equivalent to saying that excess holes in an n -type semiconductor,

and excess electrons in a *p*-type semiconductor, respectively, have lifetimes τ_p and τ_n given by

$$\delta \dot{p} = -\delta p / \tau_p = -rn\delta p \text{ or } \tau_p = 1/rn = p/g \tag{3.2}$$

and

$$\delta \dot{n} = -\delta n / \tau_n = -rp\delta n \text{ or } \tau_n = 1/rp = n/g. \tag{3.3}$$

We shall have occasion to use this interpretation later. (We later consider the modifications required when surface recombination occurs, Section 4.2, Appendix V, and the effect of a localized region of high recombination rate, Section 4.6, Appendix III.)

In principle, the steady-state solution can be obtained in terms of the three potentials ψ , φ_p and φ_n . These must satisfy three simultaneous ordinary differential equations, which we shall derive. As discussed in Section 2, we consider all donors and acceptors to be ionized so that Poisson's equation becomes

$$\frac{d^2 \psi}{dx^2} = -\frac{4\pi q}{\kappa} (ax + n_i e^{q(\varphi_p - \psi)/kT} - n_i e^{q(\psi - \varphi_n)/kT}) \tag{3.4}$$

an equation in which the unknowns are the three functions φ_p , φ_n and ψ . The total current density, from left to right, is

$$I = I_p + I_n = -q\mu \left[p \frac{d\varphi_p}{dx} + bn \frac{d\varphi_n}{dx} \right]. \tag{3.5}$$

The elimination of p and n by equation (2.4) results in an equation involving the three unknown functions and I . The divergence of hole current, equal to the net rate of generation of holes, is

$$\begin{aligned} \frac{dI_p}{dx} &= -q\mu p \left[\frac{q}{kT} \left(\frac{d\varphi_p}{dx} \right)^2 - \frac{q}{kT} \frac{d\psi}{dx} \frac{d\varphi_p}{dx} + \frac{d^2 \varphi_p}{dx^2} \right] \\ &= q(g - rn p) = qg(1 - e^{q(\varphi_p - \varphi_n)/kT}), \end{aligned} \tag{3.6}$$

with p in the second term given by (2.4) so that (3.6) is also an equation for the three unknown functions. The equation for dI_n/dx can be derived from the last two and adds nothing new. These three equations can be used to solve for $d^2\psi/dx^2$, $d^2\varphi_p/dx^2$ and $d\varphi_n/dx$ in terms of lower derivatives and I . They thus constitute a set of equations sufficient to solve the problem provided that physically meaningful boundary conditions are imposed. We shall not, however, deal directly with these equations; the main reason for deriving them was to show that the problem in question is, in principle, completely formulated. Instead of attempting to solve the equations, we shall discuss certain general features of the solutions for φ_p and φ_n , using

approximate methods, and in this way bring out the essential features of the theory of rectification.

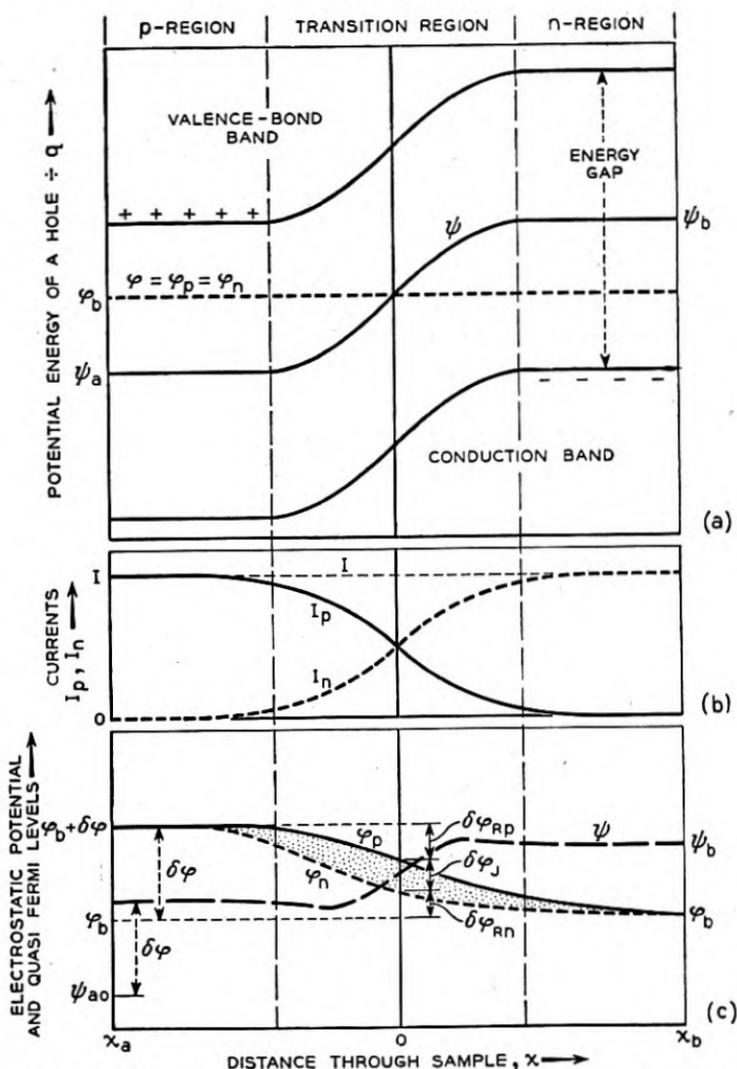


Fig. 5—Potential and current distributions for forward current in p - n junctions.

(a) p - n junction under equilibrium conditions.

(b) Division of current between holes and electrons.

(c) Distribution of potentials for forward current flow showing how the potential $\delta\varphi$ applied at x_a changes φ_p , φ_n and ψ .

In Fig. 5 we represent a general situation which may be used to illustrate the nature of the resistance of the junction. Part (a) corresponds to thermal equilibrium and shows the potential distribution and Fermi level in ac-

cordance with the scheme used in Fig. 2. Part (b) shows the current distribution for a forward current I from left to right and (c) shows the corresponding potential distribution and values of φ_n and φ_p , the total applied potential being $\delta\varphi$. Recombination prevents the hole current from penetrating far into the n -region, the depth of penetration being described by the diffusion length $L_p = \sqrt{D\tau_p} = \sqrt{Dp_n/g}$, where p_n is the hole concentration in the n -region. The electron current similarly is limited by $L_n = \sqrt{bD\tau_n} = \sqrt{bDn_p/g}$. (Diffusion lengths are evaluated for particular models of the junction in Section 4.) Far from the junction, therefore, the hole and electron concentrations have their normal values and consequently $\varphi_p = \varphi_n$ and $\varphi_p - \psi$ has its normal value. This accounts for the equal displacement $\delta\varphi$ for all three curves at $x = x_a$. The curves for φ_p and φ_n have a continuous downward trend which produces the currents

$$I_p = -q\mu p \frac{d\varphi_p}{dx} \quad \text{and} \quad I_n = -qb\mu n \frac{d\varphi_n}{dx}. \quad (3.7)$$

The area between the φ_p and φ_n curves has a special significance: This difference is related to the excess rate of recombination and the integral of this rate over the entire specimen must be sufficient to absorb the hole current $I_p = I$ entering at x_a so that the entire current at x_b is carried by electrons. In terms of $\varphi_p - \varphi_n$ and equation (3.6) we obtain

$$\begin{aligned} I &= I_p(x_a) - I_p(x_b) = \int_{x_a}^{x_b} -dI_p \\ &= gq \int_{x_a}^{x_b} (e^{q(\varphi_p - \varphi_n)/kT} - 1) dx. \end{aligned} \quad (3.8)$$

From (3.8) we conclude that if g is increased indefinitely for a specified current I , then $\varphi_p - \varphi_n$ must approach zero. For this case, in which the rate of recombination and generation is very high, $\varphi_p = \varphi_n$ and

$$I = I_p + I_n = -q\mu(p + bn) d\varphi_p/dx \quad (3.9)$$

and

$$\delta\varphi = -\int_{x_a}^{x_b} d\varphi_p = I \int_{x_a}^{x_b} dx/q\mu(p + bn) \equiv IR_0, \quad (3.10)$$

where R_0 is simply the integral of the local resistivity corresponding to densities p and n . For smaller values of g , I does not divide in the ratio $p:bn$ and $\varphi_p \neq \varphi_n$ and $\delta\varphi > IR_0$.¹⁰

We shall next give an approximate treatment for the case in which $\delta\varphi_J$ (J for junction), the value of $\varphi_p - \varphi_n$ at $x = 0$, is an appreciable fraction of

¹⁰ A general proof that $\delta\varphi > IR_0$ is given in Appendix I.

the total voltage drop. For this purpose we treat $\varphi_p - \varphi_n$ as constant over a range of integration from $x = -L_n$ to $x = +L_p$ obtaining

$$\begin{aligned} I &= gq(L_n + L_p)[e^{(q\delta\varphi_J/kT)} - 1] \\ &= I_s[e^{(q\delta\varphi_J/kT)} - 1] \end{aligned} \quad (3.11)$$

where

$$I_s = gq(L_n + L_p) \quad (3.12)$$

is the current density corresponding to the total rate of generation of hole-electron pairs in a volume $L_n + L_p$ thick. We next consider $\delta\varphi_{Rp} + \delta\varphi_{Rn}$ shown in Fig. 5c, where, as the subscript R implies, these are thought of as resistive terms and are given by the integrals

$$\delta\varphi_{Rp} + \delta\varphi_{Rn} = -\int_{x_a}^0 d\varphi_p - \int_0^{x_b} d\varphi_n = \int_{x_a}^0 I_p dx/q\mu p + \int_0^{x_b} I_n dx/q\mu bn.$$

The denominators are both approximately $q\mu(p + bn)$ which occurs in the integral for R_0 . Furthermore, for most of the first range $I_p = I$ and for most of the second $I_n = I$. Near $x = 0$, I_p or I_n must be at least $I/2$. Hence it is evident that $\delta\varphi_{Rp} + \delta\varphi_{Rn}$ cannot be much less than IR_0 . We shall represent it by IR_1 where $R_0 < 2R_1 < 2R_0$.

In terms of R_1 and I_s , the relationship between current and voltage becomes

$$\delta\varphi = \delta\varphi_{Rp} + \delta\varphi_{Rn} + \delta\varphi_J = R_1 I + \frac{kT}{q} \ln \left(1 + \frac{I}{I_s} \right). \quad (3.13)$$

This corresponds to an ideal rectifier in series with a resistance R_1 . The junction will, therefore, be a good rectifier if the second term represents a much higher resistance.

We shall compare the two resistances for the case corresponding to $K \ll 1$. For this case, we have $p = -ax$ and $n = +ax$ except in the narrow range $|x| < L_a = n_i/a$. The integral R_0 can be approximated by integrating dx/σ for x outside of the range $\pm L_a$ using the approximation $\pm ax$ for p and n and approximating the integral from $-L_a$ to $+L_a$ by $2L_a/\sigma$ (intrinsic). This procedure gives

$$\begin{aligned} R_1 &= \int_{L_a}^{-x_a} dx/q\mu ax + \frac{2L_a}{q\mu n_i(1+b)} + \int_{L_a}^{x_b} dx/q\mu bax \\ &= \frac{L_a}{q\mu n_i} \left(1 + \frac{1}{b} \right) \ln (x_b/L_a) \end{aligned} \quad (3.14)$$

where it is supposed that $-x_a \doteq x_b$ and that $\ln (x_b/L_a)$ is large compared to $2/(b + 2 + 1/b)$. The evaluation of L_p and L_n for use in I_s is more involved

since τ_p and τ_n are both functions of x . We shall obtain an approximate self-consistent diffusion length by assuming that the holes diffuse, on the average, to just such a depth, L_p , that in uniform material of the type found at L_p , their diffusion length would also be L_p . At a depth L_p , the value of n is aL_p so that by (3.2), τ_p is $1/raL_p = n_i^2/gaL_p$. Thus we write

$$L_p^2 = D\tau_p = Dn_i^2/gaL_p. \tag{3.15}$$

We can solve the equation (3.15) for L_p and a similar one for L_n and insert the results in equation (3.13). For small I this gives

$$\begin{aligned} \delta\varphi/I = R_1 + (kT/qI_s) = & \frac{L_a}{q\mu n_i} \left(1 + \frac{1}{b} \right) \ln (x_b/L_a) \\ & + kT/(q^2 g^{2/3} (DL_a n_i)^{1/3} (1 + b^{1/3})). \end{aligned} \tag{3.16}$$

It is seen that for g large, the second term, corresponding to the rectifying resistance, becomes small. For this case, as discussed above, $\varphi_p = \varphi_n$ and the exact integral for R_0 should be used and the junction will give poor rectification.

It is also instructive to consider L_a as a variable. Increasing L_a corresponds to making the transition from p to n more gradual. It is evident that varying L_a changes the two terms of (3.16) in opposite directions so that there will be an intermediate value of L_a for which the resistance of the junction is a minimum. As L_a approaches zero, however, the second term should be modified: If we imagine that in the transition region the concentration ($N_d - N_a$) varies only over a finite range, bounded by fixed values n_n and p_p in the n - and p -regions, then it is clear that the limiting values of L_p and L_n should be given not by (3.15) but by $\sqrt{D\tau_p}$ and $\sqrt{bD\tau_n}$ where τ_p and τ_n are evaluated in the n -region and p -region. This leads to a limiting value for I_s , which is given in equation (4.11) of the following section. In the range for which (3.16) applies, however, the interesting result holds that widening the transition region initially decreases the resistance by furnishing a larger volume in which holes and electrons may combine or be generated.

The condition that $\delta\varphi_j$ dominate the resistance is that the second term of (3.16) be much larger than the first. This leads to the inequality

$$1 \ll \frac{kT}{q^2 g^{2/3} (DL_a n_i)^{1/3}} \cdot \frac{q\mu n_i}{L_a} = (Dn_i/gL_a^2)^{2/3} = (L_{pi}/L_a)^{4/3} \tag{3.17}$$

where we have neglected various factors involving b , which are nearly unity, and $\ln(x_b/L_a)$ (which must be about 4 for Ge since the conductivity at x_b is about $\exp(4)$ times the intrinsic conductivity). The quantity

$$L_{pi} = (Dn_i/g)^{1/2} \tag{3.18}$$

is the diffusion length for holes in the intrinsic region. The inequality states that the diffusion length must be much larger than L_a . This is equivalent to the previous statement that the hole current must penetrate the n -region for the rectifier to have a good characteristic. (If a local region of high recombination is present in the transition region, this result just quoted need not apply. See Section 4.6.)

If the hole current penetrates deeply into the n -region and R_1 is negligible, then we can conclude that the current-voltage characteristic will fit the ideal formula. For these assumptions $\delta\varphi_{np}$ on Fig. 5 will be small and the principal change in φ_p will occur relatively deep in the n -region, at least beyond the transition region. So long as the hole concentration introduced in the n -region is much smaller than n_n , the hole current into the n -region will be a linear function of the value of p at the right edge of the transition region, being zero when p equals p_n , the equilibrium value of p . This leads at once to a hole current proportional to $p - p_n$ and since the shift of φ_p in respect to ψ at the edge of transition region is $\delta\varphi$, $p - p_n$ is equal to $p_n(\exp(a\delta\varphi/kT) - 1)$. (These ideas are discussed in detail in Section 4.) A similar relationship will hold for electrons entering the p -region; hence the total current will vary as $\exp(q\delta\varphi/kT) - 1$. This is a theoretical rectification formula¹¹ giving the maximum rectification for carriers of charge q .

4. TREATMENT OF PARTICULAR MODELS

4.1 Introduction and Assumptions

In this section we shall deal chiefly with good rectifiers so that the IR drop, discussed in connection with R_1 in Section 3, is negligible. We shall deal chiefly with the case for which the transition region is narrow compared to the diffusion length; consequently, there is little change in I_p in traversing the transition region. In Fig. 6(a) we consider a hypothetical junction in which the properties are uniform outside the transition region. The division of the specimen into three parts as shown is seen to be reasonable for germanium: In n -type germanium, the diffusion constant for holes is about $40 \text{ cm}^2/\text{sec}$ and the lifetime is greater than 10^{-6} sec so that the diffusion distance is $L_p = \sqrt{D\tau_p} > 6 \times 10^{-3}$ cm. This is much greater than most transition regions.

The major drop in φ_p must occur to the right of the transition region. This follows from our assumptions: First, we may neglect the IR drop in the p -region; hence φ_p is substantially constant from $x = x_a$ to $x = x_{tp}$. Second, the decrease in φ_p is much less in the transition region than in the n -region; this follows from two considerations: the resistance for hole flow is lower in

¹¹ C. Wagner, *Phys. Zeits.* 32, 641-645 (1931).

the transition region than in the *n*-region; the effective length of flow in the *n*-region, being L_p , is greater than the width of the transition region. Consequently, the variation of φ_p shown in Fig. 6(c) is seen to be reasonable. Similar considerations apply to φ_n . As is shown in Fig. 6(c), the application of $\delta\varphi$ does not alter $\varphi_p - \psi$ in the *p*-region nor $\varphi_n - \psi$ in the *n*-region. The

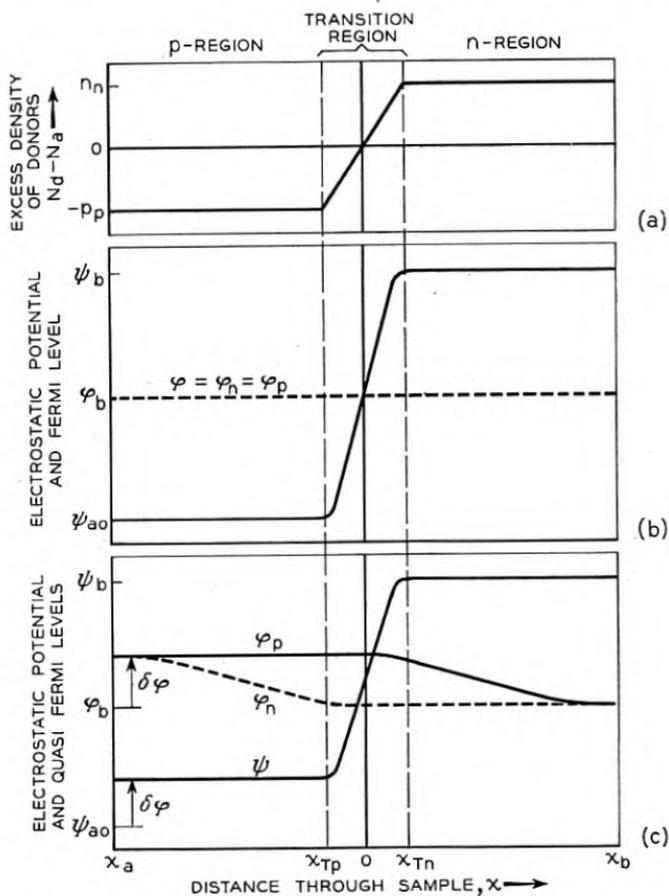


Fig. 6—Simplified model of a *p-n* junction.

- (a) Distribution of donors and acceptors.
- (b) Potentials for thermal equilibrium.
- (c) Effect of $\delta\varphi$ applied potential in forward direction.

reason, as discussed in connection with (2.31), is that in these regions electrical neutrality requires an essentially constant value for the more abundant carrier. Hence the relationships between the φ 's and ψ follow from (2.4).

The nature of the potential distribution in the transition region has no effect in the considerations just discussed. However, as shown in Section 2,

the capacity of the transition region, which we shall denote by C_T in this section, does depend on the nature of the transition region and, consequently, on the value of K .

If the sizes of the p -region and n -region are large compared to the diffusion lengths, we may assume the current at x_a to be substantially I_p only and that at x_b , I_n only. The total current entering at x_a can be accounted for as doing three things: (1) neutralizing the electron current flowing into the p -region across x_{Tp} , (2) contributing to the charge in the transition region (this corresponds to the capacity discussed in Section 2) and (3) contributing a current flow to the right across x_{Tn} .

We have selected the hole current for analysis because the hole has a positive charge and the connection between the algebra and the physical picture is simplified. For the same reason, the text emphasizes forward current, although the equations are equally applicable to reverse currents. Nothing essential is left out by this process; since the sample as a whole remains uncharged, the current I is the same for all values of x and if I_p is known, then $I_n = I - I_p$ is also determined.

4.2 Solution for Hole Flow into the n -region

We shall calculate first the hole current $I_p(x_{Tn})$ flowing across x_{Tn} . It is readily evaluated as follows: The value of $p(x_{Tn})$ is given by

$$\begin{aligned} p(x_{Tn}) &= n_i e^{q(\varphi_b + \delta\varphi - \psi_b)/kT} \\ &= p_n e^{q\delta\varphi/kT} \end{aligned} \quad (4.1)$$

where p_n is the hole concentration in the n -region for thermal equilibrium. If we apply a small a-c. signal superimposed on a d-c. bias so that

$$\delta\varphi = v_0 + v_1 e^{i\omega t} \quad (4.2)$$

where v_1 is an a-c. signal, assumed so small that linear theory may be employed (i.e. $v_1 \ll kT/q$), then

$$p(x_{Tn}) = (p_n e^{qv_0/kT})(1 + (qv_1/kT)e^{i\omega t}).$$

We resolve this density into a d-c. component p_0 and an a-c. component $p_1 e^{i\omega t}$:

$$p(x_{Tn}) = p_n + p_0 + p_1 e^{i\omega t} \quad (4.3)$$

where

$$p_0 = p_n(e^{qv_0/kT} - 1) \quad (4.4)$$

$$p_1 = (qp_n v_1/kT)e^{qv_0/kT}. \quad (4.5)$$

So long as $p(x_{Tn}) \ll n_n$, the normal concentration of electrons in the

n-region, the lifetime τ_p and diffusion constant D for a hole will be substantially unaltered by $\delta\varphi$. Application of the hole-current equation to the hole density $p(x, t)$ gives

$$I_p = -qD \frac{\partial p}{\partial x}. \tag{4.6}$$

Combining this with the recombination equation

$$\frac{\partial p}{\partial t} = \frac{p_n - p}{\tau_p} - \frac{1}{q} \frac{\partial I_p}{\partial x} = \frac{p_n - p}{\tau_p} + D \frac{\partial^2 p}{\partial x^2} \tag{4.7}$$

leads to the solution

$$p = p_n + p_0 e^{(x_{Tn}-x)/\sqrt{D\tau_p}} + p_1 e^{i\omega t + (x_{Tn}-x)(1+i\omega\tau_p)^{1/2}/(D\tau_p)^{1/2}}. \tag{4.8}$$

The quantity $\sqrt{D\tau_p}$ is the diffusion length and is denoted by L_p . (We shall use subscript p for holes in the *n*-region and *n* for electrons in the *p*-region for both L and τ .)

When p is large compared to p_n , but small compared to n_n , the expression for p leads to the following formula for φ_p :

$$\varphi_p = \varphi_n + v_c - (kT/q)(x - x_{Tn})/L_p + v_1 e^{i\omega t - (x-x_{Tn})[(1+i\omega\tau_p)^{1/2}-1]/L_p}. \tag{4.9}$$

This shows that the d-c. part of φ_p varies linearly in the *n*-region, for large forward currents, and decreases by (kT/q) in each diffusion length L_p . The transition from this linear dependence to an exponential decay for φ_p comes when $\varphi_p - \varphi_n = (kT/q)$. This behavior of the d-c. part of φ_p is useful in connection with diagrams of φ_p versus distance. (See Sections 5 and 6.)

The solution just obtained for p gives rise to a current at x_{Tn} of

$$\begin{aligned} I_p(x_{Tn}) &= -qD \frac{\partial p}{\partial x} \\ &= qp_0 D/L_p + qp_1 D e^{i\omega\tau} (1 + i\omega\tau_p)^{1/2}/L_p. \end{aligned} \tag{4.10}$$

The d-c. part is calculated by substituting (4.4) for p_0 :

$$\begin{aligned} I_{p0}(x_{Tn}) &= (qp_n D/L_p)(e^{qv_0/kT} - 1); \\ &\equiv I_{ps}(e^{qv_0/kT} - 1) \end{aligned} \tag{4.11}$$

and the a-c. part is similarly obtained from (4.5) for p_1 :

$$\begin{aligned} I_{p1}(x_{Tn}) &= (qp_n \mu/L_p)[e^{(qv_0/kT)}](1 + i\omega\tau_p)^{1/2} v_1 e^{i\omega t} \\ &\equiv (G_p + iS_p)v_1 e^{i\omega t} \equiv A_p v_1 e^{i\omega t} \end{aligned} \tag{4.12}$$

where A_p is called the admittance (per unit area) for holes diffusing into the *n*-region; its real and imaginary parts are the conductance and suscept-

ance. For $\omega\tau_p$ small, the real term G_p is simply conductance per cm^2 of a layer L_p cm thick with hole conduction corresponding to the density $p_n + f_0$; it is also the differential conductance obtained by differentiating (4.11) in respect to v_0 . For the case of zero bias this establishes the result quoted in Section 1 that the voltage drop is due to hole flow in the n -region where the hole conductivity is low.

In this section we have treated τ_p as arising from body recombination. In a sample whose y and z dimensions are comparable to L_p or L_n , surface recombination may play a dominant role. However, as we show in Appendix V, the theory given here may still apply provided appropriate values for τ_p and τ_n are used.

4.3 D-C. Formulae

The total direct hole current flowing in at x_a is I_{x_0} plus the current required to recombine with electrons in the p -region. This latter current is, of course, equal to the electron current flowing into the p -region. This electron current, denoted by $I_{e,0}$ or $I_{e,0}(x_{Tp})$, is obtained by the same procedure as that leading to (4.11) for $I_{x,0}$ except that bD replaces D and the subscripts of L and τ are now n . Combining the two currents leads to the total direct current:

$$I_0 = I_{p0} + I_{n0} = (qL) \left(\frac{p_n}{L_p} + \frac{bn_p}{L_n} \right) (e^{qv_0/kT} - 1) \quad (4.13)$$

for the direct current per unit area for applied potential v_0 .¹² The algebraic signs are such that $I > 0$ corresponds to current from the p -region to the n -region in the specimen; $v_0 > 0$ corresponds to a plus potential applied to the p -end. The ratio of hole current to electron current across the transition region is

$$\begin{aligned} \frac{I_{p0}}{I_{n0}} &= \frac{p_n}{L_p} \cdot \frac{L_n}{bn_p} = \frac{p_p}{bn_n} \cdot \frac{\sqrt{bD\tau_n}}{\sqrt{D\tau_p}} \\ &= \frac{p_p}{bn_n} \sqrt{\frac{bn_n}{p_p}} = \sqrt{\frac{\sigma_p}{\sigma_n}} \end{aligned} \quad (4.14)$$

where we have used the relationships $n_n p_n = n_p p_p = n_i^2$ from (2.2) and $\tau_p n_n = \tau_n p_p = 1/r$ from (3.2) and (3.3). These results can be summarized by saying that the current flows principally into the material of higher re-

¹² For convenience we repeat the definitions here: $q \equiv$ magnitude of electronic charge; $D \equiv$ diffusion constant for holes; p_n and $n_n \equiv$ thermal equilibrium value of p and n , assumed constant throughout n -region ($x > x_{Tn}$); n_p and $p_p \equiv$ similar values for $x < x_{Tp}$; $L_p \equiv$ diffusion length $\equiv \sqrt{D\tau_p}$ for holes in n -region; $\tau_p \equiv$ lifetime of hole in n -region before recombination; $b =$ electron mobility/hole mobility; L_n and τ_n similar in quantities for electrons in p -region; $\sigma_n = q\mu_n n_n$ and $\sigma_p = q\mu_p p_p$ are the conductivities of the two regions.

sistivity. We can also say that the hole current depends only on the *n*-type material and vice versa. For a *p-n* junction emitter in a transistor with an *n*-type base, it is thus advantageous to use high conductivity *p*-type material so as to suppress an unwanted electron current.

For comparison with experiment, it is advantageous to express the values of p_n and n_p in terms of the conductivities σ_n and σ_p . If the conductivity of the intrinsic material is written as

$$\sigma_i = q\mu n_i(1 + b), \tag{4.15}$$

then, if $p_n \ll n_n$ and $n_p \ll p_p$, we find

$$q\mu p_n = b\sigma_i^2/(1 + b)^2\sigma_n \tag{4.16}$$

$$q\mu n_p = b\sigma_i^2/(1 + b)^2\sigma_p. \tag{4.17}$$

Using these equations, we may rewrite (4.11) and a corresponding equation for electron current into the *p*-region so as to express their dependence on d-c. bias v_0 and the properties of the regions:

$$\begin{aligned} I_{p0}(v_0) &= \frac{b\sigma_i^2}{(1 + b)^2\sigma_n L_p} \cdot \frac{kT}{q} (e^{qv_0/kT} - 1) \\ &\equiv G_{p0} \frac{kT}{q} (e^{qv_0/kT} - 1) \\ &\equiv I_{ps}(e^{qv_0/kT} - 1) \end{aligned} \tag{4.18}$$

$$\begin{aligned} I_{n0}(v_0) &= \frac{b\sigma_i^2}{(1 + b)^2\sigma_p L_n} \cdot \frac{kT}{q} (e^{qv_0/kT} - 1) \\ &\equiv G_{n0} \frac{kT}{q} (e^{qv_0/kT} - 1) \\ &\equiv I_{ns}(e^{qv_0/kT} - 1). \end{aligned} \tag{4.19}$$

The values of G_{p0} and G_{n0} (which are readily seen to be the values of the low-frequency, low-voltage ($v_0 < kT/q$) conductances) and the saturation reverse currents are given by

$$G_{p0} \equiv \frac{b\sigma_i^2}{(1 + b)^2\sigma_n L_p} \equiv \frac{q}{kT} I_{ps} \tag{4.20}$$

$$G_{n0} \equiv \frac{b\sigma_i^2}{(1 + b)^2\sigma_p L_n} \equiv \frac{q}{kT} I_{ns} \tag{4.21}$$

The expression for direct current then becomes

$$\begin{aligned} I_0(v_0) &= [I_{p0} + G_{n0}] \left(\frac{kT}{q} \right) [e^{qv_0/kT} - 1] \\ &= (I_{ps} + I_{ns}) [e^{qv_0/kT} - 1]. \end{aligned} \tag{4.22}$$

4.4 Total Admittance

In order to calculate the alternating current, we must include the capacity of the transition region, discussed in Section 2. Denoting this by C_T , we then find for the total alternating current.

$$I_{ac} = (G_p + iS_p + G_n + iS_n + i\omega C_T) v_1 = A v_1 \quad (4.23)$$

where G_n and S_n are similar to G_p and S_p but apply to electron current into the p -region. The value of the hole and electron admittances can be expressed as

$$A_p = G_p + iS_p = (1 + i\omega\tau_p)^{1/2} G_{p0} e^{qv_0/kT} \quad (4.24)$$

$$A_n = G_n + iS_n = (1 + i\omega\tau_n)^{1/2} G_{n0} e^{qv_0/kT} \quad (4.25)$$

For low frequencies, such that ω is much less than $1/\tau_p$, we can expand $G_p + iS_p$ as follows:

$$G_p + iS_p = G_{p0} e^{qv_0/kT} + i\omega(\tau_p/2)G_{p0} e^{qv_0/kT} \quad (4.26)$$

Hence $(\tau_p/2)G_{p0} e^{qv_0/kT}$ behaves like a capacity.

It is instructive to interpret this capacity for the case of zero bias, $v_0 = 0$, for which we find:

$$C_p = \tau_p G_{p0} / 2 = \tau_p q p_n \mu / 2 L_p = q^2 p_n L_p / 2 k T. \quad (4.27)$$

The last formula, obtained by noting that $\tau_p \mu = q \tau_p D / k T = q L_p^2 / k T$, has a simple interpretation: $q p_n L_p$ is the total charge of holes in a layer L_p thick. For a small change in voltage v , this density should change by a fraction qv/kT so that the change in charge divided by the change in v is $(q/kT)(q p_n L_p)$ which differs from C_p only by a factor of 2, which arises from the nature of the diffusion equation.

This capacity can be compared with $C_{T \text{ neut.}}$, discussed in Section 2, (see equation (2.39) and text for (2.42)) for germanium at room temperature as follows:

$$\frac{C_p}{C_{T \text{ neut.}}} = \frac{q^2 p_n L_p}{2 k T} \cdot \frac{k T a}{10 q^2 n_i^2} = \frac{p_n L_p a}{20 n_i^2}. \quad (4.28)$$

For a structure like Fig. 6(c), the excess of donors over acceptors reaches its maximum value, equal to n_n , at x_{Tn} leading to $n_n = a x_{Tn}$. Consequently $a = n_n / x_{Tn}$. Substituting this value for a in (4.28) and noting that $p_n n_n = n_i^2$ gives

$$\frac{C_p}{C_{T \text{ neut.}}} = \frac{L_p}{20 x_{Tn}} \quad (4.29)$$

As discussed at the beginning of this section, $L_p \doteq 6 \times 10^{-3}$ cm for holes

in germanium. Hence if the transition region is 6×10^{-4} cm thick, the diffusion capacity C_p will dominate the capacitive term in the admittance.

Although A_p simulates a conductance and capacitance in parallel at low frequencies, its high-frequency behavior is quite different. In Fig. 7 the

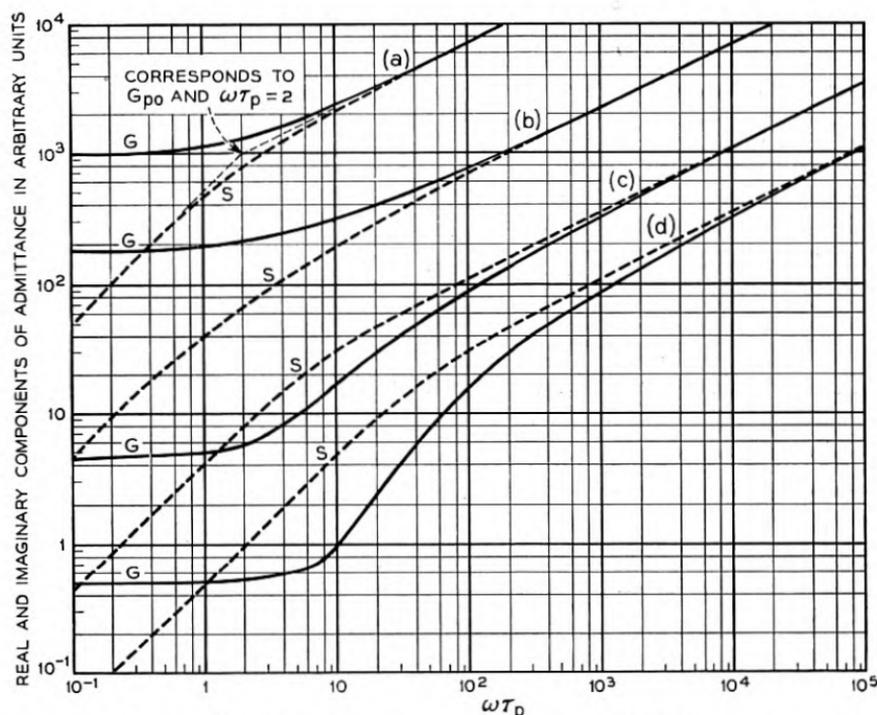


Fig. 7—Real, G , and imaginary, S , components of admittance for hole flow into n -region.

- (a) $10^3 A_p / G_{p0} = 10^3 (1 + i\omega\tau_p)^{1/2}$ corresponding to uniform n -region.
- (b) $10^2 \times$ Formula of Appendix III, corresponding to layer of high recombination rate in front of n -region. This causes G to exceed S at higher frequencies than for (a).
- (c) $10 \times$ Equation (4.33), corresponding to a retarding field in the n -region, with $L_r = L_p / \sqrt{10}$.
- (d) Equation (4.33) with $L_r = L_p / 10$.

behavior of $(1 + i\omega\tau_p)^{1/2} = A_p / G_{p0}$, is shown. For high frequencies G_p and S_p are equal:

$$G_p = S_p = \sqrt{\tau_p/2} G_{p0} \sqrt{\omega} = \frac{b\sigma_i^2 \sqrt{\omega}}{(1+b)^2 \sigma_n \sqrt{2D}} \quad (4.30)$$

Thus for high frequencies the admittance is independent of τ_p and is determined by the diffusion of holes in and out of the n -region. The three straight asymptotes have a common intersection at the point G_{p0} , $\omega\tau = 2$ on Fig. 7, a fact which is useful in estimating the value of τ from such data.

For large ω , S_p varies as $\omega^{1/2}$ as shown in (4.30) whereas $S_T = \omega C_T$. Hence

at very high frequencies C_T will dominate the admittance. At very high frequencies C_T itself will have a frequency dependence; however, for the assumptions on which the treatment of this section is based, the relaxation time for the transition region τ_T is much less than τ_p . This is a consequence of the fact that, although diffusion of holes into the transition region is required for the charging of C_T , the distance is relatively short, being in fact only that fraction of the width $x_{Tn} - x_{Tp}$ of the transition region in which ψ rises by kT/q ; in germanium this will be about one-tenth of $x_{Tn} - x_{Tp}$. Since diffusion times vary as (distance)², the ratio of the times is

$$\frac{\tau_T}{\tau_p} = \frac{(x_{Tn} - x_{Tp})^2}{100L_p^2}. \quad (4.31)$$

Hence if $L_p > x_{Tn} - x_{Tp}$, τ_T will be much less than τ_p .¹³

4.5 Admittance Due to Hole Flow in a Retarding Field

In Appendix II we treat the case in which a potential gradient, due to changing concentration for example, is present in the n - and p -regions. This tends to prevent holes from diffusing deep in the n -region and for this reason the n -region acts partly like a storage tank for holes under a-c. conditions, thus enhancing S_p compared to G_p in A_p . If the electric field is $-d\psi/dx = kT/qL_r$, where L_r is the distance required for an increase of kT/q of potential (i.e. a factor of e increase in n_n), then the value of A_p is

$$A_p = [q\mu p_n/L_p] \frac{(2L_r/L_p)(1 + i\omega\tau_p)}{1 + [1 + (1 + i\omega\tau_p)(2L_r/L_p)^2]^{1/2}} \quad (4.32)$$

For $\omega\tau_p > 1$, this admittance is largely reactive provided $2L_r/L_p$ is sufficiently small.

The dependence of A_p upon ω is shown in Fig. 7 for two values of L_r/L_p . The plot shows the real and imaginary parts of

$$A_p/[2q\mu p_n L_r/L_p^2] = \frac{(1 + i\omega\tau_p)}{1 + [1 + (1 + i\omega\tau_p)(2L_r/L_p)^2]^{1/2}} \quad (4.33)$$

for $L_p/L_r = 10^{1/2}$ and $L_p/L_r = 10$, the two curves being relatively displaced vertically by one decade. The second value implies that the field keeps the holes back so that they penetrate only $\frac{1}{10}$ their possible diffusion length in no field. It is seen that for this case the storage effect is very pronounced and the susceptance S is much larger than G for high frequencies.

The function $(1 + i\omega\tau_p)^{1/2}$, discussed earlier, corresponds to the limiting case of (4.32) for $L_r = \infty$.

¹³ In Appendix IV an analytic treatment of C_T is given.

4.6 The Effect of a Region of High Rate of Generation

There is evidence that imperfections, such as surfaces and cracks, add materially to the rate of generation and recombination of holes and electrons. If there is a localized region of high recombination rate in the transition region, there will be a pronounced modification of the admittance characteristics. In Fig. 8(a) such a layer is represented at $x = 0$. In Fig. 8(b) the customary plot of φ_p and φ_n versus x is shown. If we neglect the effect of the series resistance terms denoted by R_1 in Section 3, the change $\delta\varphi$ will

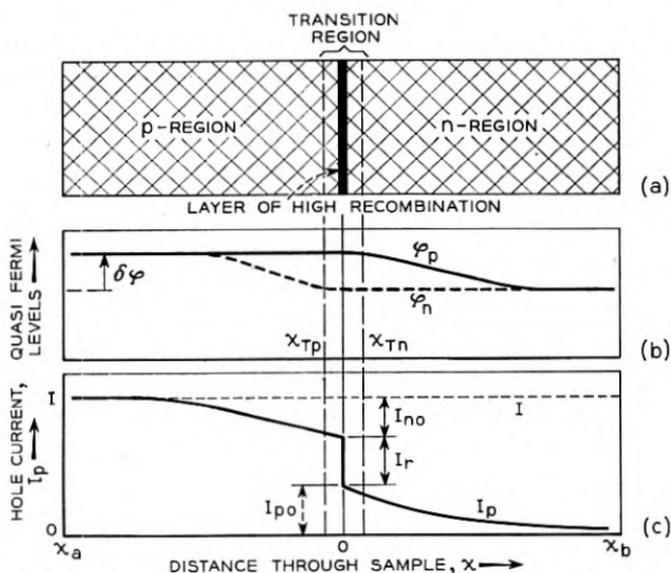


Fig. 8—The effect of a localized layer of high recombination rate on the junction characteristic.

- (a) Location of layer of high recombination rate.
- (b) Quasi Fermi levels.
- (c) Distribution of hole current showing rapid change at layer of high recombination rate.

occur in the *p*-region for φ_n and in the *n*-region for φ_p . The hole current flowing into the *n*-region will thus be the same as before and will be given by equation (4.11) or (4.18) and denoted by $I_{p0}(\delta\varphi)$. Similarly, the electron current will be $I_{n0}(\delta\varphi)$. In the layer we shall suppose that there is a rate of generation of hole electron pairs equal to g_a per unit area of the layer and a rate of recombination proportional to $r_a n p$ per unit area. We suppose, furthermore, that the layer is so thin that n and p are uniform throughout the layer. The net rate of generation is thus

$$g_a - r_a n p = g_a [1 - e^{q(\varphi_p - \varphi_n)/kT}] \tag{4.34}$$

since for equilibrium conditions the rates balance so that $r_a n_i^2 = g_a$. The net hole current recombining in the layer per unit area is thus

$$I_r(\varphi_p - \varphi_n) = qg_a [e^{q(\varphi_p - \varphi_n)/kT} - 1] \quad (4.35)$$

There must, therefore, be a discontinuous decrease of hole current across the layer. The total hole current flowing in at $x = x_a$, which is also the total current I , thus does three things: for $x < x_{Tp}$, it combines with $I_{n0}(\delta\varphi)$; for $x_{Tp} < x < x_{Tn}$, it combines with electrons at rate $I_r(\delta\varphi)$; for $x > x_{Tn}$, it flows into the n -region in amount $I_{p0}(\delta\varphi)$. This leads to

$$I = I_{n0}(\delta\varphi) + I_{p0}(\delta\varphi) + I_r(\delta\varphi). \quad (4.36)$$

In other words the layer of high recombination acts like a rectifier in parallel with $I_{n0}(\delta\varphi) + I_{p0}(\delta\varphi)$. The frequency characteristic of $I_r(\delta\varphi)$, however, will be independent of frequency and will contribute a pure conductance to the admittance of the junction.

If the layer is considered to have finite width, however, it will exhibit frequency effects just as does I_p in the n -region. In Appendix III, we treat a case in which the layer is a part of the n -region itself but has a recombination time different from the main layer. If the time is shorter, a large amount of the hole current may recombine in this layer. For high frequencies, the current may not penetrate the layer, in which case the admittance for hole current is determined by the thin layer rather than by the whole n -type region. A case of this sort is shown in Fig. 7. In this case the thickness of the layer is $\frac{1}{3}$ of its diffusion length and in it the lifetime of a hole τ_l is $\frac{1}{3}$ the value τ_p in the main body of the n -region. The hole current will thus be restricted to this layer when the diffusion distance $\sqrt{D/\omega}$ is less than the layer thickness ($\frac{1}{3}$) $\sqrt{D\tau_l}$; this corresponds to $\omega\tau_l > 9$ or $\omega\tau_p > 81$. The presence of the high rate of combination in the layer is evidenced by the tendency of G to be greater than S at high frequencies. If the layer were infinitely thin, as discussed above, it would simply add a constant conductance to the admittance.

4.7 Patch Effect in p - n Junctions

If there are localized regions of high recombination rate, a "patch effect" may be produced in an n - p junction. As an extreme example, suppose the value of g_a for the layer just considered is allowed to become very large; then the recombination resistance may become small compared to R_1 in Section 3 and the junction will become substantially ohmic. If the region of high rate of recombination is relatively small compared to the area of the rest of the junction, then the behavior of the junction as a whole may be regarded as being that due to two junctions in parallel. Over most of the area,

the currents will flow as if the patch were not present so that one component of the current will be that due to the uniform junction. In addition there will be current due to recombination and generation in the patch. The series resistance to the patch will be relatively high due to the constriction of the current paths. On the other hand, the value of $I_r(\delta\phi)$ associated with the patch may be very high. Hence the current due to the patch will be that of a low impedance ideal rectifier in series with a high resistance; and if the ratio of impedances is high enough, such a series combination amounts essentially to an ohmic leakage path. Thus patches in the *p-n* junction will tend to introduce leakage paths and destroy saturation in the reverse direction.

An extreme example of a region of high rate of recombination would be a particle of metal making a non-rectifying contact to both *p*- and *n*-type germanium. Since holes and electrons are essentially instantly combined in a metal, the boundary condition at the metal surface would be equality of φ_p and φ_n . This would mean that near the metal particle, φ_p and σ_n could not differ by $\delta\varphi$, the condition required, over some parts of the junction at least, in order for ideal rectification to occur.

A common source of imperfection in *p-n* junctions arises from dirt or fragments on the surface which overlap the junction. Even if these do not actually constitute a short circuit across the junction, they may furnish patches of the sort discussed here and modify the junction characteristic.

4.8 Final Comments

Another possible cause for frequency effects may be found in the trapping of holes or electrons.¹⁴ When an added hole concentration is introduced into an *n*-region, a certain fraction of the holes will be captured by acceptors and later re-emitted or else recombined with electrons while trapped. Investigation of this process is given in Appendix VI. One interesting result is that the trapping of holes in a uniform *n*-region cannot produce an effective susceptance (i.e. $i\omega C$) in excess of the conductance, as can a retarding field.

Finally it should be remarked that important and significant variations of the conductivity in the *p*- and *n*-regions may be produced by hole or electron injection. Under these conditions, when the hole concentration approaches n_n , $\psi - \varphi_n$ will vary. Under these conditions R_1 may be appreciably altered. These factors favor the *p-n* junction as a rectifier since they lead to a reduction of series resistance under conditions of forward bias and thus tend to improve the rectification ratio.

¹⁴ Frequency dependent effects in Cu_2O rectifiers have been explained in this way by J. Bardeen and W. H. Brattain, personal communication.

5. INTERNAL CONTACT POTENTIALS

The theory of p - n junctions presented above has interesting consequences when applied to the distribution of potential between two semiconductors

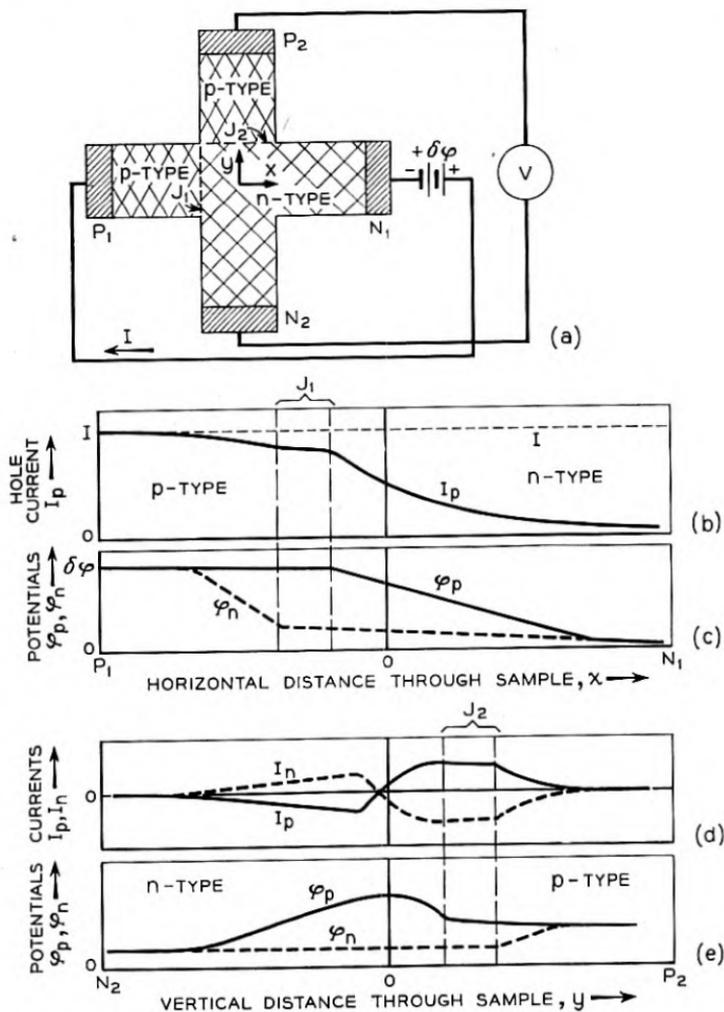


Fig. 9—Internal contact potentials showing how presence of injected holes produces a contact potential across J_2 .

under conditions of hole or electron injection. In Fig. 9 we illustrate an X-shaped structure. A forward current flows across the junction P_1 and out of branch N_1 . If the distance across the intersection is comparable with or small compared to the diffusion length for holes, a potential difference should be measured between P_2 and N_2 . The reason for this is that holes

flow easily into P_2 since the potential distribution there favors their entrance. Since, however, P_2 is open-circuited this hole flow biases J_2 in the forward direction; since J_2 is high resistance, an appreciable bias is developed before the counter current equals the inward hole flow and a steady state is reached. No similar effect occurs in the branch N_2 ; consequently P_2 will be found to be floating (open-circuited) at a more positive potential than N_2 .

Parts (b) to (e) describe this reasoning in more complete terms. We suppose that the p -regions are more highly conducting than the n -regions so that the current across J_1 , shown in (b), is mainly holes. The potentials φ_p and φ_n along the x -axis will be similar to those of Figs. 5 and 6; (c) shows this situation and indicates that the diffusion length for electrons in the p -region is less than for holes in the n -region. Along the y axis φ_p and φ_n vary as shown in (e), the reasoning being as follows: At the origin of coordinates φ_p and φ_n have the same values as for (c). The transverse hole current (d) has a small positive component at $y = 0$ since, as mentioned above, P_2 tends to absorb holes and thus increase diffusion along the plus y -axis. Since the net transverse current is zero, $I_n = -I_p$ in (d). The φ curves of (e) have been drawn to conform to the currents in (d); φ_n is nearly constant in the n -region and φ_p is nearly constant in the p -region. As concluded in connection with Figs. 5 and 6, φ_n and φ_p are also nearly constant across the transition region. These conclusions lead to the shape of φ_n and φ_p for $y > 0$ in (e). For $y < 0$, the reasoning is the same as that used in Sections 3 and 4 and we conclude that φ_n is essentially constant. Hence, a difference in the Fermi levels at P_2 and N_2 will result.

In Fig. 10 we show a structure for which we can make quantitative calculations of the variations of φ_p and φ_n . We assume for this case that the forward current from P_1 to N does not produce an appreciable voltage drop, i.e. change in ψ and φ_n , in region N . This will be a good approximation if the dimensions are suitably proportioned. We shall next solve for the steady-state distribution of p subject to the indicated boundary conditions assuming that p is a function of x only. As we have discussed in Section 4.1, when p is small compared to n in the n -region, we can write

$$p = p_n e^{q(\varphi_p - \varphi_n)/kT} \tag{5.1}$$

In keeping with the treatment in the next section of this structure as a transistor, the terminals are designated emitter, collector and base, the potentials with respect to the base being φ_e and φ_c . The contact to N or the base is such that $\varphi_b = \varphi_n$ in this region. Hence, the boundary conditions at J_1 and J_2 are

$$p_1 = p_n e^{q\varphi_e/kT} \quad x = -w \tag{5.2}$$

$$p_2 = p_n e^{q\varphi_c/kT} \quad x = +w \tag{5.3}$$

The function $p(x)$ which satisfies these boundary conditions and the equation

$$D \frac{d^2 p}{dx^2} - \frac{p - p_n}{\tau_p} = 0 \quad (5.4)$$

is

$$p(x) = p_n + \frac{p_1 + p_2 - 2p_n}{2 \cosh(w/L_p)} \cosh(x/L_p) + \frac{p_2 - p_1}{2 \sinh(w/L_p)} \sinh(x/L_p) \quad (5.5)$$

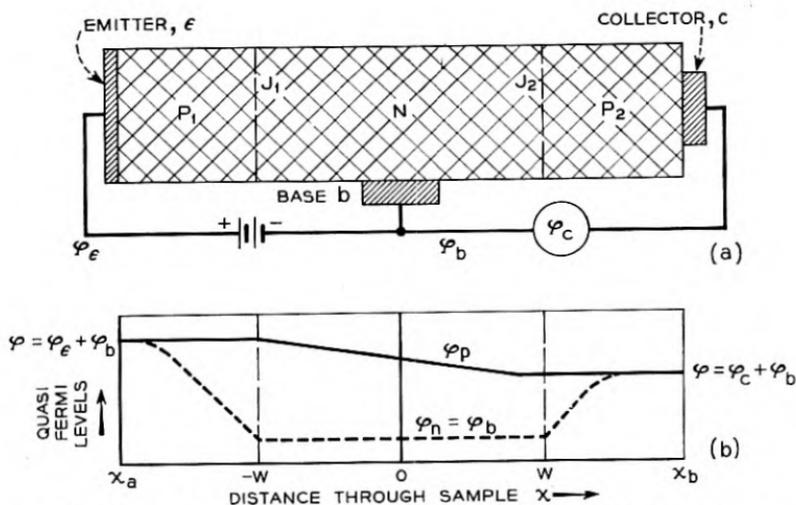


Fig. 10—Model used for calculation of internal contact potential and to illustrate p - n - p transistor.

- (a) Semiconductor with two p - n junctions and ohmic metal contacts.
 (b) Quasi Fermi levels showing internal contact potential between b and c .

which gives rise to a hole current across J_2 into P_2 of amount

$$\begin{aligned} I_p &= -qD \left. \frac{dp}{dx} \right|_{x=w} \\ &= \frac{qD}{2L_p} \left[(p_1 - p_2) \coth \frac{w}{L_p} + (2p_n - p_1 - p_2) \tanh \frac{w}{L_p} \right] \\ &= \frac{qD}{2L_p} \left[(p_1 - p_n) \left(\coth \frac{w}{L_p} - \tanh \frac{w}{L_p} \right) \right. \\ &\quad \left. - (p_2 - p_n) \left(\coth \frac{w}{L_p} + \tanh \frac{w}{L_p} \right) \right] \\ &= + \frac{p_n qD}{L_p} \left[\frac{[e^{q\varphi/kT} - 1]}{\sinh(2w/L_p)} - \frac{(e^{q\varphi_c/kT} - 1)}{\tanh(2w/L_p)} \right] \\ &= \operatorname{csch}(2w/L_p) I_{p0}(\varphi_e) - \coth(2w/L_p) I_{p0}(\varphi_c) \end{aligned} \quad (5.6)$$

where, by $I_{p0}(\varphi)$, we mean the hole current which would flow in the forward direction across either J_1 or J_2 if uninfluenced by the other (i.e. the function of (4.11) or (4.18) and (4.20).) The equation shows that a fraction $\text{csch}(2w/L_p)$ of the current $I_{p0}(\varphi_c)$, which would be injected by φ_c on P_1 in the absence of J_2 , flows into P_2 . The conductance of P_2 across J_2 is increased by the factor $\text{coth}(2w/L_p)$.

The current *into* P_2 carried by electrons will be unaffected by J_1 and can be denoted by $-I_{n0}(\varphi_c)$ the minus sign resulting from the fact that currents *into* P_2 are in the reverse direction. The total current flowing *into* P_2 contains the $-I_{n0}(\varphi_c)$ and $-I_{p0}(\varphi_c)$ terms and must cancel the $+I_{p0}(\varphi_c)$ term for equilibrium. Hence:

$$I_{n0}(\varphi_c) + \text{coth}(2w/L_p) I_{p0}(\varphi_c) = \text{csch}(2w/L_p) I_{p0}(\varphi_c) \quad (5.7)$$

If $p_n \gg n_p$, the I_{n0} term can be neglected compared to $\text{coth}(2w/L_p) I_{p0}$. Hence the value of φ_c must satisfy

$$I_{p0}(\varphi_c) = \text{sech}(2w/L_p) I_{p0}(\varphi_c). \quad (5.8)$$

For $\varphi_c > kT/q$, the exponential approximation may be used for I_{p0} in both terms:

$$\varphi_c = \varphi_c - (kT/q) \ln \cosh(2w/L_p), \quad (5.9)$$

so that, if $(2w/L_p)$ is the order of unity, φ_c should be only about (kT/q) less than φ_c . For $(2w/L_p)$ large, we get

$$\varphi_c = \varphi_c - (kT/q) (2w/L_p) \quad (5.10)$$

corresponding to the linear drop of φ_p , discussed in connection with equation (4.9), across the distance $2w$.

When φ_c is negative, so that we have to deal with reverse current, φ_c will not decrease indefinitely but will reach a minimum value given by

$$[\exp q\varphi_c/kT] - 1 = -\text{sech}(2w/L_p) \quad (5.11)$$

and corresponding to saturation reverse current across J_1 , so that

$$\varphi_c = -(kT/q) \ln [1 + (1/2) \text{csch}^2(w/L_p)]. \quad (5.12)$$

The floating potentials of p -type contacts to n -type material into which holes have been injected (or n -type contacts to p -type material with injected electrons) are reminiscent of probes in gas discharges which tend to become charged negative in respect to the space around them because they catch electrons more easily than positive ions. The situation may also be compared with that producing thermal e.m.f.'s; in fact a "concentration temperature" of the semiconductor with injected holes can be defined by finding the temperature for which $np = n_i^2(T)$. We conclude that, in the

absence of thermal equilibrium, different potentials depending on the nature of the contact are, in general, the rule rather than the exception.

The bias developed on P_2 or c will change its conductance. If we suppose that φ_e and φ_b are held constant, then the current flowing into c is obtained by the same reasoning that led to (5.7) and is

$$I_c(\varphi_c, \varphi_e) = I_{n0}(\varphi_c) + \coth \frac{2w}{L_p} I_{p0}(\varphi_c) - \operatorname{csch} \frac{2w}{L_p} I_{p0}(\varphi_e). \quad (5.13)$$

For an infinitesimal change in φ_c from the value which makes $I_c(\varphi_c, \varphi_e)$ vanish, the admittance to c is readily found from (4.18) and (4.19) to be

$$\begin{aligned} \left(\frac{\partial I_c}{\partial \varphi_c} \right)_{\varphi_e} &= I'_{n0}(\varphi_c) + \coth \frac{2w}{L_p} I'_{p0}(\varphi_c) \\ &= \left[G_{n0} + \coth \frac{2w}{L_p} G_{p0} \right] e^{q\varphi_c/kT} \end{aligned} \quad (5.14)$$

which shows that pronounced variations in admittance should be associated with variations in hole density in N in Fig. 10.¹⁵

6. *p-n-p* TRANSISTORS

The structure shown in Fig. 10 is a transistor with power gain provided the distance w is not too great. As a first approximation, we shall neglect the drop due to currents in the N region. If we use P_2 as the collector and call the collector current, I_c , positive when it flows into P_2 from outside, we shall have from (5.13)

$$I_c = -\operatorname{csch} \frac{2w}{L_p} I_{p0}(\varphi_e) + \coth \frac{2w}{L_p} I_{p0}(\varphi_c) + I_{n0}(\varphi_c). \quad (6.1)$$

The emitter current is similarly

$$I_e = \coth \frac{2w}{L_p} I_{p0}(\varphi_e) - \operatorname{csch} \frac{2w}{L_p} I_{p0}(\varphi_c) + I_{n0}(\varphi_e). \quad (6.2)$$

If $p_n \gg n_p$, then the I_{n0} terms can be neglected. However, the base current will not vanish but will be

$$\begin{aligned} I_b &= -I_e - I_c = \left[\operatorname{csch} \frac{2w}{L_p} - \coth \frac{2w}{L_p} \right] [I_{p0}(\varphi_e) + I_{p0}(\varphi_c)] \\ &= \frac{2 \sinh^2 w/L_p}{\sinh 2w/L_p} [I_{p0}(\varphi_e) + I_{p0}(\varphi_c)]. \end{aligned} \quad (6.3)$$

¹⁵ The variations in admittance discussed in connection with metal point contacts in an accompanying paper in this issue (W. Shockley, G. L. Pearson and J. R. Haynes, *Bell Sys. Tech. J.*, July, 1949), arise from this cause; however, the nature of the contact is not as simple as here.

For w/L_p large, the junctions do not interact and the hyperbolic coefficient becomes unity and $I_b = -[I_{p0}(\varphi_e) + I_{p0}(\varphi_c)]$.

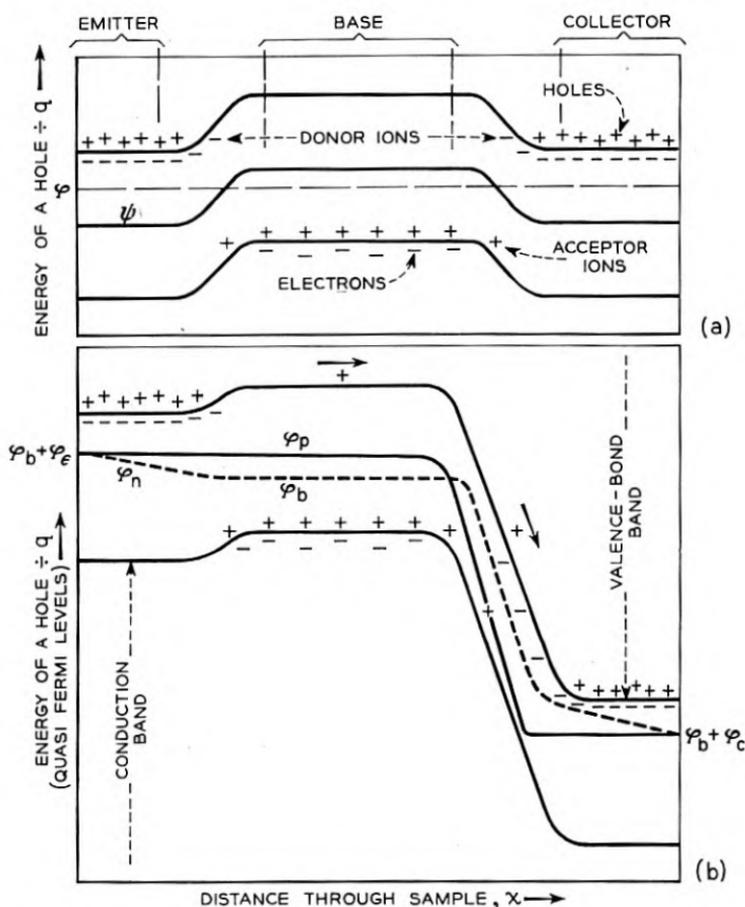


Fig. 11—*p-n-p* transistor.

- (a) Thermal equilibrium.
- (b) Operating condition.

If φ_c is several volts negative, so that $I_{p0}(\varphi_c)$ has its saturation value I_{ps} (see (4.11) and (4.20)), then the ratio $-\delta I_c / \delta I_e \equiv \alpha$ has the value

$$\alpha = -\frac{\delta I_c}{\delta I_e} = \frac{\operatorname{csch} \frac{2w}{L_p}}{\operatorname{coth} \frac{2w}{L_p}} = \operatorname{sech} \frac{2w}{L_p}. \quad (6.4)$$

For $(2w/L_p) = 0.5, 1, 2$ respectively, $\alpha = 0.89, 0.65, 0.27$. Since the output impedance R_{22} will be very high when φ_c is in the reverse direction, and the

input impedance will be low, the power gain formula¹⁶ $\alpha^2 R_{22}/R_{11}$ will yield power gain even when α is less than unity.

In certain ways the structure of Fig. 10 resembles a vacuum tube. In Fig. 11, we show the energy band diagram, with energies of holes plotted upwards so as to be in accord with the convention for voltages. (a) shows the thermal equilibrium distribution and (b) the distribution under operating conditions. It is seen that the potential hill, which holes must climb in reaching the collector, has been reduced by φ_e . The n -region represents in a sense the grid region in a vacuum tube, in which the potential and hence plate current, is varied by the charge on the grid wires. Here the potential in the n -region is varied by the voltage applied between base and emitter. In both cases one current is controlled by another. In the vacuum tube the current which charges the grid wires controls the space current. Because the grid is negative to the cathode, the electrons involved in the space current are kept away from the grid while at the same time the electrons in the grid are kept out of the space by the work function of the grid (provided that the grid does not become overheated.) In Fig. 11, the electrons flowing into the base control the hole current from emitter to collector. In this case the controlled and controlling currents flow in the same space but in different directions because of the opposite signs of their charges.

As this discussion suggests, it may be advantageous to operate the p - n - p transistor like a grounded cathode vacuum tube, with the emitter grounded and the input applied to the base.

The p - n - p transistor has the interesting feature of being calculable to a high degree. One can consider such questions as the relative ratios of width to length of the n -region and the effect of altering impurity contents and scaling the structure to operate in different frequency ranges. However, we shall not pursue these questions of possible applications further here.

ACKNOWLEDGMENT

The writer is indebted to a number of his colleagues for stimulating discussions and encouragement, in particular to H. R. Moore, G. L. Pearson and M. Sparks, whose experimental work, to be described in later publications, suggested development of the theory along the lines presented above. He is also indebted to J. Bardeen, P. Debye, G. Wannier and W. van Roosbroeck for theoretical comments and suggestions and especially to the last and to Mrs. G. V. Smith for valuable assistance in preparing the manuscript.

¹⁶ Physical Principles Involved in Transistor Action, J. Bardeen and W. H. Brattain, *Phys. Rev.* 75, 1208 (1949).

APPENDIX I

A THEOREM ON JUNCTION RESISTANCE

We shall here prove that the junction resistance is never less than the value obtained by integrating the local resistivity $1/q\mu(p + bn)$. This is accomplished by analyzing the following equation which we shall discuss before giving the derivation:

$$I\delta\varphi = \frac{1}{q\mu} \int_{x_a}^{x_b} \left(\frac{I_p^2}{p} + \frac{I_n^2}{bn} \right) dx + qg \int_{x_a}^{x_b} (\varphi_p - \varphi_n) (e^{q(\varphi_p - \varphi_n)/kT} - 1) dx,$$

the meaning of the symbols being that shown in Fig. 5. This expression is valid even if large disturbances in p and n from their equilibrium values occur. The second integral is positive since the integrand is never negative. It may be very large if $\varphi_p - \varphi_n \gg kT/q$ in some regions. If, in the first integral, we consider that I_p and I_n may be varied subject to the restraint $I_p + I_n = I$, we may readily prove that the first integrand takes on a minimum value when

$$I_p = \frac{pI}{p + bn} \quad \text{and} \quad I_n = \frac{bnI}{p + bn}.$$

For this minimum condition, the first integral becomes

$$I^2 \int_{x_a}^{x_b} dx/q\mu(p + bn) = I^2 R_0$$

where R_0 is simply the integrated local resistivity. If I does divide in this way, the second integral is zero, a result which we can see as follows:

$$I_p = -q\mu p \, d\varphi_p/dx$$

$$I_n = -q\mu bn \, d\varphi_n/dx$$

$$\frac{d\varphi_p/dx}{d\varphi_n/dx} = \frac{I_p/p}{I_n/bn}$$

Hence, if the current divides in the ratio of p to bn , then $d\varphi_p = d\varphi_n$ and, since $\varphi_p = \varphi_n$ at x_a , $\varphi_p = \varphi_n$ everywhere and the second integral vanishes.

In general, of course, the conditions governing recombination prevent current division in the ratio $p:bn$ and then $\delta\varphi/I > R_0$.

The equation discussed above is derived as follows: We suppose that

$$\varphi_p(x_a) = \varphi_n(x_a) = \varphi_a$$

$$\varphi_p(x_b) = \varphi_n(x_b) = \varphi_b$$

Then

$$\begin{aligned} I\delta\varphi &= -(I_p\varphi_p + I_n\varphi_n) \Big|_{x_a}^{x_b} \\ &= -\int_{x_a}^{x_b} \frac{d}{dx} (I_p\varphi_p + I_n\varphi_n) dx \\ &= -\int_{x_a}^{x_b} \left(\frac{dI_p}{dx} \varphi_p + \frac{dI_n}{dx} \varphi_n \right) dx - \int_{x_a}^{x_b} \left(I_p \frac{d\varphi_p}{dx} + I_n \frac{d\varphi_n}{dx} \right) dx. \end{aligned}$$

Since

$$\frac{dI_p}{dx} = -\frac{dI_n}{dx} = qg(1 - e^{q(\varphi_p - \varphi_n)/kT})$$

and

$$\frac{d\varphi_p}{dx} = -I_p/q\mu p, \quad \frac{d\varphi_n}{dx} = -I_n/q\mu b n$$

these two integrals are readily transformed into the ones previously discussed.

APPENDIX II

ADMITTANCE IN A RETARDING FIELD

We shall here derive the admittance equation for holes diffusing into a retarding potential $\psi = kTx/qL_r$ in which the potential increases by kT in each distance L_r . The differential equation for the a-c. component of p is

$$i\omega p = -\frac{p}{\tau_p} - \frac{\partial}{\partial x} \left[-D \frac{\partial p}{\partial x} - \mu p \frac{\partial \psi}{\partial x} \right].$$

This equation may be solved by letting $p = p_1 \exp(i\omega t - \gamma x)$ as may be seen by rewriting the equation and substituting this expression for p :

$$\begin{aligned} D \left[\frac{\partial^2 p}{\partial x^2} + \frac{1}{L_r} \frac{\partial p}{\partial x} \right] - \frac{1}{\tau_p} (1 + i\omega\tau_p) p \\ = -\gamma D \left[-\gamma + \frac{1}{L_r} \right] p - \frac{1}{\tau_p} (1 + i\omega\tau_p) p = 0 \end{aligned}$$

leading to

$$\gamma = \frac{1 + [1 + (2L_r/L_p)^2 (1 + i\omega\tau_p)]^{1/2}}{2L_r}.$$

The corresponding current evaluated at $x = 0$ where $p = p_1 \exp(i\omega t) = (p_n q v_1 / kT) \exp(i\omega t)$ is given by

$$\begin{aligned} I &= -q \left[D \frac{\partial p}{\partial x} + \mu p \frac{\partial \psi}{\partial x} \right] \\ &= -qD \left[-\gamma + \frac{1}{L_r} \right] p \\ &= \frac{q(1 + i\omega\tau_p)p}{\gamma\tau_p} \\ &= \frac{p_n q^2 (1 + i\omega t)}{kT\tau_p} \cdot \frac{2L_r}{1 + [1 + (2L_r/L_p)^2(1 + i\omega\tau_p)]^{1/2}} \cdot v_1 e^{i\omega t} \\ &= \frac{q\mu p_n 2L_r}{L_p^2} \cdot \frac{(1 + i\omega\tau_p)}{1 + [1 + (2L_r/L_p)^2(1 + i\omega\tau_p)]^{1/2}} \cdot v_1 e^{i\omega t} \\ &= A_p v_1 e^{i\omega t}. \end{aligned}$$

This is equivalent to (4.32) in Section 4.

APPENDIX III

ADMITTANCE FOR TWO LAYERS

We shall here treat a case in which there is a thin layer on the n -side of the transition region in which recombination occurs much more readily than deeper in the n -layer. The case of an infinitely thin plane, discussed in Section 4, is a limiting case of this model. We shall suppose that the layer extends from $x = -c$ to $x = 0$ while $x > 0$ corresponds to the n -region. We shall suppose that the potential in the layer is uniform with value ψ_1 whereas in the n -region it has value ψ_2 . The lifetimes of holes will be taken τ_1 and τ_2 in the two layers. The solutions for p_1 and p_2 are evidently

$$\begin{aligned} p_1 &= p_{10} + (A e^{-\alpha x} + B e^{+\alpha x}) e^{i\omega t} & x < 0 \\ p_2 &= p_{20} + C e^{-\beta x + i\omega t} & x > 0 \end{aligned}$$

where

$$\begin{aligned} \alpha &= (1 + i\omega\tau_1)^{1/2} / \sqrt{D\tau_1} \equiv (1 + i\omega\tau_1)^{1/2} / L_1 \\ \beta &= (1 + i\omega\tau_2)^{1/2} / \sqrt{D\tau_2} \equiv (1 + i\omega\tau_2)^{1/2} / L_2. \end{aligned}$$

The boundary condition for continuity of φ_p , required to avoid singularity in $\partial\varphi_p/\partial x$, is

$$p_2 e^{q\psi_2/kT} = p_1 e^{q\psi_1/kT}$$

and, for continuity of hole current, is $\partial p_1/\partial x = \partial p_2/\partial x$. Expressing these in terms of A , B , C , α and β for the a-c. components yields:

$$A + B = C e^{q(\psi_1 - \psi_2)/kT} \equiv CF$$

$$\alpha(A - B) = \beta C$$

so that

$$A = (F + \beta/\alpha)C/2.$$

$$B = (F - \beta/\alpha)C/2.$$

Hence the ratio $-\partial p/\partial x/p$ at $x = -c$ is

$$-\frac{\partial \ln p}{\partial x} = \frac{\alpha(A e^{+\alpha c} - B e^{-\alpha c})}{(A e^{+\alpha c} + B e^{-\alpha c})} = \frac{\alpha(F\alpha \sinh \alpha c + \beta \cosh \alpha c)}{F\alpha \cosh \alpha c + \beta \sinh \alpha c}.$$

Since at $x = -c$, the a-c. component of p_1 is $(qv_1/kT)p_{10}e^{i\omega t}$, the admittance is

$$\begin{aligned} A_p &= \frac{-qD\partial p/\partial x}{v_1 e^{i\omega t}} = (q^2 D p_{10}/kT)(-\partial \ln p/\partial x) \\ &= (q\mu p_{10}/L_1)(1 + i\omega\tau_1)^{1/2} \frac{F\alpha \sinh \alpha c + \beta \cosh \alpha c}{F\alpha \cosh \alpha c + \beta \sinh \alpha c}. \end{aligned}$$

For $c \rightarrow 0$, this transforms into

$$(q\mu p_{10}/L_1)(1 + i\omega\tau_1)^{1/2} \beta/F\alpha = (q\mu(p_{10}/F)/L_2)(1 + i\omega\tau_2)^{1/2}$$

which agrees with Section 4, since p_{10}/F then corresponds to p_n .

If c/L_1 and F are not large, an appreciable amount of recombination takes place for $x > 0$ for low frequencies. Dispersive effects will then occur corresponding to τ_2 . The a-c. will not penetrate to $x = 0$, however, if $c(\omega/D)^{1/2} \gg 1$ and the dispersive effects will then be determined by τ_1 .

The frequency-dependent part of the admittance,

$$(1 + i\omega\tau_1) \frac{F\alpha \sinh \alpha c + \beta \cosh \alpha c}{F\alpha \cosh \alpha c + \beta \sinh \alpha c},$$

has been computed and is shown in Fig. 7 for $\tau_p = \tau_2$, $F = 1$, $\tau_1 = \tau_2/9$ and $c/L_1 = \frac{1}{3}$. For these values about half the hole current reaches $x = 0$ for low frequencies. As the time constant for diffusion through the layer is $\tau_p/81$, as discussed in Section 4.6, the layer will act as a largely frequency-independent admittance well above the point for $\omega\tau_p = 1$. This is reflected in the behavior of the curves of Fig. 7 and, for frequencies in the $\sqrt{\omega t}$ range, it is seen that G is larger than S by about 50% of the low-frequency value of G ; this split of $G + iS$ into $(\frac{1}{2})G_0$ plus approximately $(\frac{1}{2})G_0(1 + i\omega\tau_p)^{1/2}$ corresponds to the fact that about half the holes are absorbed in layer 1 for the assumed conditions.

APPENDIX IV

TIME CONSTANT FOR THE CAPACITY OF THE TRANSITION REGION

For this case we shall consider the case of holes in an a-c. field with potential

$$\psi = \frac{kT}{q} \left(\frac{x}{L_r} + \frac{x e^{i\omega t}}{L_1} \right)$$

where the d-c. retarding field is kT/qL_r , and the a-c. field is kT/qL_1 where $1/L_1$ is considered small for the linear theory presented here. The expression for the current of holes is

$$-D \frac{\partial p}{\partial x} - \mu p \frac{\partial \psi}{\partial x} = -D \left[\frac{\partial p}{\partial x} + p \left(\frac{1}{L_r} + \frac{e^{i\omega t}}{L_1} \right) \right]$$

We shall obtain a solution for p by letting

$$p = p_0 e^{-x/L_r} + p_1 [e^{-x/L_r} - e^{-\gamma x}] e^{i\omega t},$$

while neglecting recombination in this region so that p must satisfy the condition $\dot{p} = -\partial$ (hole current)/ ∂x leading to the differential equation

$$D \left[\frac{\partial^2 p}{\partial x^2} + \frac{\partial p}{\partial x} \left(\frac{1}{L_r} + \frac{e^{i\omega t}}{L_1} \right) \right] - \dot{p} = 0$$

There are three separate exponential dependencies of the variables leading to three equations (neglecting terms of order $(1/L_1)^2$)

$$e^{-x/L_r}: \quad D \left[p_0 \frac{1}{L_r^2} - p_0 \frac{1}{L_r^2} \right] = 0$$

$$e^{-x/L_r + i\omega t}: \quad D \left[p_1 \frac{1}{L_r^2} - p_1 \frac{1}{L_r^2} - \frac{1}{L_r L_1} p_0 \right] - i\omega p_1 = 0$$

$$e^{-\gamma x + i\omega t}: \quad D[\gamma^2 - \gamma/L_r] p_1 - i\omega p_1 = 0$$

The first equation is satisfied by the equilibrium distribution and the second by

$$p_1 = -p_0 D/i\omega L_1 L_r$$

and the last by

$$\gamma = \frac{1 + \sqrt{1 + 4i\omega L_r^2/D}}{2L_r}$$

It is evident that dispersive effects set in when

$$\omega = D/4L_r^2$$

This corresponds to the result used in (4.31) in which $(x_{Tn} - x_{Tp})/10$ was used for L_r . For smaller values of ω the current may be calculated and put in simple form by expanding γ up to terms including ω^2 . The resulting expression for the current is

$$I = -i\omega q p_0 L_r (L_r/L_1) e^{i\omega t}$$

This is interpreted as follows: The a-c. voltage across a layer L_r thick is

$$\delta\psi = (kT/q) (L_r/L_1) e^{i\omega t}$$

and, if we consider plus voltage as producing a field from left to right, then the a-c. voltage across L_r is $V = -\delta\psi$. Substituting this for $(L_r/L_1)\exp(i\omega t)$ gives

$$I = i\omega q p_0 L_r (q/kT) V$$

Here $q p_0 L_r$ is the total charge in the layer L_r , (qV/kT) is an average fractional change in this charge for V so that $(q p_0 L_r) (qV/kT) \div V$ is a capacity.

APPENDIX V

THE EFFECT OF SURFACE RECOMBINATION

In this appendix we shall consider the effect of surface recombination upon the characteristics of the p - n junction. As for Section 4 we shall illustrate the theory for the case of holes diffusing into n -type material. For simplicity we shall treat a square cross-section bounded by $y = \pm w$, $z = \pm w$, the current flow being along $+x$.

We shall denote the a-c. component of p as

$$p_1 \equiv p_1(x, y, z, t)$$

At $x = 0$, the edge of the n -region, we shall suppose that φ_p and ψ are independent of y and z so that we shall have

$$p_1(0, y, z, t) = p_{10} e^{i\omega t} = (p_n q v_1/kT) e^{i\omega t}$$

by reasoning similar to that used for equation (4.5). The boundary condition at the surface will be

$$-D \frac{\partial p_1}{\partial y} = s p_1 \quad \text{for } y = +w$$

This states that the recombination per unit area is $s p_1$ and is equal to the diffusion to the surface $-D \partial p_1 / \partial y$. Similar boundary conditions hold for the other surfaces. By standard procedures involving separation of variables we may verify that the solution satisfying the boundary conditions is

$$p_1 = \sum_{i,j=0}^{\infty} a_{ij} e^{-\alpha_i x + i\omega t} \cos \beta_i y \cos \beta_j z$$

where the eigenvalues β_i are determined by the boundary condition

$$\beta_i w \tan \beta_i w = sb/D \equiv \chi.$$

We use $\theta_i = \beta_i w$ for brevity later. Because of the symmetry of the boundary conditions it is not necessary to include sine functions in the sum. The value of α_{ij} is given by

$$\alpha_{ij} = (1 + i\omega\tau_{ij})^{1/2} / (D\tau_{ij})^{1/2}$$

where τ_{ij} is the lifetime of a hole in the eigenfunction $\cos \beta_i y \cos \beta_j z$; i.e. τ_{ij} is the lifetime which makes

$$p = \exp(-t/\tau_{ij}) \cos \beta_i y \cos \beta_j z,$$

a function which satisfies the surface boundary conditions, a solution of the equation

$$\partial p / \partial t = D\nabla^2 p - p/\tau = -D(\beta_i^2 + \beta_j^2)p - p/\tau$$

where to simplify the subsequent expressions we have omitted the subscript p from τ . This equation leads to

$$\frac{1}{\tau_{ij}} = D(\beta_i^2 + \beta_j^2) + \frac{1}{\tau}.$$

The coefficients a_{ij} are readily found since the $\cos \beta_i y$ functions form an orthogonal set (as may be verified by integrating by parts and using the boundary conditions). The values are

$$a_{ij}/p_{10} = 4[\sin \theta_i \sin \theta_j] / \theta_i \theta_j [1 + (1/2\theta_i) \sin 2\theta_i] \cdot [1 + (1/2\theta_j) \sin 2\theta_j]$$

The current corresponding to this solution is

$$I_1 = -qD \iint (\partial p / \partial x) dy dz$$

integrated over the cross section at $x = 0$. This gives

$$I_1 = qDp_{10}e^{i\omega t} \sum \alpha_{ij}(a_{ij}/p_{10})(4w^2/\theta_i\theta_j) \sin \theta_i \sin \theta_j$$

Substituting for a_{ij} and inserting $p_{10} = p_n q v_1 / kT$, we obtain an expression for the admittance $A_p = I_1/V_1 \exp(i\omega t)$:

$$A_p = 4w^2 q\mu p_n \sum_{ij} \alpha_{ij} \frac{4 \sin^2 \theta_i \sin^2 \theta_j}{\theta_i^2 \theta_j^2 \left[1 + \left(\frac{1}{2\theta_i} \right) \sin 2\theta_i \right] \left[1 + \left(\frac{1}{2\theta_j} \right) \sin 2\theta_j \right]}$$

where the sum plays the role formerly taken by $(1 + i\omega\tau)^{1/2} / \sqrt{D\tau}$ in equation (4.12); the factor $4w^2$ is the area of the junction.

We shall analyze the formula for the case in which recombination on the

surface is smaller than diffusion to the surface so that χ is not large. The values of θ_i , over which the sum is to be taken, may be estimated as follows: in each interval of θ_i of the form $n\pi$ to $(n + \frac{1}{2})\pi$, $\theta_i \tan \theta_i$ varies from 0 to ∞ , giving one solution to $\theta_i \tan \theta_i = \chi$. For χ small, the solutions are approximately

$$\begin{aligned}\theta_0 &\doteq \sin \theta_0 \doteq \tan \theta_0 \doteq \sqrt{\chi} \\ \theta_1 &\doteq \pi + \chi/\pi; \quad -\sin \theta_1 \doteq \tan \theta_1 \doteq \chi/\pi \\ &\dots \dots \dots \\ \theta_n &\doteq n\pi + \chi/n\pi; \quad (-1)^n \sin \theta_n \doteq \tan \theta_n \doteq \chi/n\pi\end{aligned}$$

From this we see that the terms in the sum are as follows:

$$\begin{aligned}\alpha_{00} \cdot 4\chi^2/\chi^2 4 &= \alpha_{00} \\ \alpha_{n0} \cdot 2(\chi/n\pi)^2/(n\pi)^2 &= \alpha_{n0} 2\chi^2/n^4 \pi^4 \\ \alpha_{nm} \cdot 4\chi^4/n^4 m^4 \pi^8 &\end{aligned}$$

From this it is evident that unless χ is large, the series converges very rapidly. (This conclusion is not altered when the increase in α_{nm} with $\beta_n \beta_m$ is considered.) Thus the dominant term in the admittance is

$$4w^2 q \mu f_0 (1 + i\omega\tau_{00})^{1/2} / \sqrt{D\tau_{00}}$$

where

$$\begin{aligned}1/\tau_{00} &= 2 \left(\frac{D}{w^2} \right) (\theta_0^2) + 1/\tau \\ &\doteq 2 \left(\frac{D}{w^2} \right) \frac{sw}{D} + 1/\tau \\ &= 2 \left(\frac{s}{w} \right) + 1/\tau\end{aligned}$$

This expression is valid only for sw/D small so that $\theta_0^2 \doteq sw/D$. The term $s/(w/2)$ represents the rate of decay due to holes recombining on the surface, s having the dimensions of velocity. For $\omega \gg 1/\tau_{00}$, the admittance becomes $4w^2 q \mu f_0 (i\omega/D)^{1/2}$, the same value as given in equation (4.12) for large ω and an area $4w^2$.

The conclusion from this appendix is that for χ small, the effect of surface recombination is simply to modify the effective value of τ and otherwise leave the theory of Section 4 unaltered.

For very large values of χ , it is necessary to consider higher terms in the sum and several values of τ will be important. Under these conditions the

approximation is that, at $x = 0$, p_1 is independent of x and y may become a poor one, especially for forward currents, because the transverse currents to the edges will be important. Under these conditions the role of surface recombination will give rise to patch effects of the sort discussed in Section 4.

APPENDIX VI

THE EFFECT OF TRAPPING UPON THE DIFFUSION PROCESS

In this appendix we shall investigate the effect of the trapping of holes upon the impedance. We denote the density of mobile holes in the valence-bond band by p and the density of holes trapped in acceptors by p_a . For thermal equilibrium at room temperature there will be an equilibrium ratio, called α , for p_a/p . For germanium $\alpha \approx 10^{-4}$ and for silicon $\alpha \approx 0.1$ to 0.2 .

We shall consider four processes which occur at rates (per particle per unit time) as follows:

- ν_r direct recombination of a hole with an electron (free or bound to a donor)
- ν_t trapping of a hole by an acceptor
- ν_{ra} recombination of a hole trapped on an acceptor
- ν_e excitation of a trapped hole into the valence-bond band.

Under equilibrium conditions as many holes are being trapped (rate $p\nu_t$) as are being excited ($p_a\nu_e$): hence $\nu_t = \alpha\nu_e$.

We shall study solutions of the customary form for the a-c. components:

$$p_1 = p_{10} e^{i\omega t - \gamma x}$$

$$p_{1a} = p_{1a0} e^{i\omega t - \gamma x}$$

These must satisfy the equations

$$\dot{p}_1 = D\nabla^2 p_1 - (\nu_t + \nu_r)p_1 + \nu_e p_{1a}$$

$$\dot{p}_{1a} = \nu_t p_1 - (\nu_e + \nu_{ra})p_{1a}$$

These lead readily to the equation for γ :

$$D\gamma^2 = i\omega + \nu_r + \nu_t - \nu_e\nu_t/(i\omega + \nu_e + \nu_{ra}) = i\omega$$

$$\cdot \left[1 + \frac{\nu_e\nu_t}{(\nu_e + \nu_{ra})^2 + \omega^2} \right] + \nu_r + \nu_t \left[1 - \frac{\nu_e}{(\nu_e + \nu_{ra}) + \omega^2/(\nu_e + \nu_{ra})} \right]$$

From this equation we can directly reach the important conclusion that the trapping process can never lead to a capacitive term larger than the resistive term. This result is obtained by analyzing the complex phase of γ , the admittance being proportional to γ . In particular, we find that the real term in $D\gamma^2$ is always positive, as may be seen from inspection, so that the complex phase angle of γ is less than 45° .

The form reduces to a simple expression if ν_e and ν_t are very large com-

pared to ν_r , ν_{ra} and ω , a situation which insures local equilibrium between p and p_a . Under these conditions we obtain

$$D\gamma^2 = i\omega[1 - \frac{1}{1+\alpha}] + \nu_r + \alpha\nu_{ra}$$

Dividing by $(1 + \alpha)$ gives

$$[D/(1 + \alpha)]\gamma^2 = [Dp/(p + p_a)]\gamma^2 = i\omega + \frac{p\nu_r + p_a\nu_{ra}}{p + p_a}$$

The interpretation is that the holes diffuse as if their diffusion constant were reduced by the fraction of the time $p/(p + p_a)$ they are free to move and recombine with a properly weighted average of ν_r and ν_{ra} .

APPENDIX VII

SOLUTIONS OF THE SPACE CHARGE EQUATION

We shall first show that the space charge equation (2.11) has a unique solution for the one dimensional case. For simplicity we write (2.11) in the form

$$\frac{d^2u}{dx^2} = \sinh u - f(x) \quad (A7.1)$$

to which it can be readily reduced. We shall deal with the case for which

$$f = f_a \text{ for } x < x_a \quad (A7.2)$$

$$f = f_b \text{ for } x > x_b > x_a \quad (A7.3)$$

so that the interval (x_a, x_b) is bounded by semi-infinite blocks of uniform semiconductor. We shall require that u be finite at $x = \pm \infty$. This boundary condition requires that for large values of $|x|$

$$u = u_a + A_a e^{+\gamma_a x} \quad x \rightarrow -\infty \quad (A7.4)$$

$$u = u_b + A_b e^{-\gamma_b x} \quad x \rightarrow +\infty \quad (A7.5)$$

where

$$\sinh u_a = f_a, \quad \sinh u_b = f_b$$

$$\gamma_a = |(\cosh u_a)^{1/2}|, \quad \gamma_b = |(\cosh u_b)^{1/2}|$$

(If the opposite signs of the γ 's were present, the boundary conditions would not be satisfied.) The exponential solutions are valid for $|u - u_a|$ or $|u - u_b| \ll 1$. For larger values, however, solutions exist which are obtained by integrating (A7.1) to larger or smaller values of x .

For these extended solutions the values of $u(x, A_a)$ and $u'(x, A_a) (= du/dx)$

are monotonically increasing functions of A_a . This may be seen by considering $x = x_a$. For A_a sufficiently small, the value of $u(x_a, A_a)$ and $u'(x_a, A_a)$ are given simply by (A7.4). For larger values of A_a , an exact integral will be required. It is evident, however, that all solutions of the form (A7.4) are related simply by translation for $x < x_a$. Hence increasing A_a is simply equivalent to integrating (A7.1) to larger values of x and it is evident that this increases u and u' monotonically. It may be verified that for a sufficiently large A_a the solution becomes infinite at x_a so that $u(x_a, A_a)$ and $u'(x_a, A_a)$ both vary monotonically and continuously from $-\infty$ to $+\infty$ as A_a varies from negative to positive values. We shall refer to this property of $u(x_a, A_a)$, $u'(x_a, A_a)$ as P_1 .

We next wish to show that $u(x_1, A_a)$, $u'(x_1, A_a)$ has the property P_1 for values of $x_1 > x_a$. To prove this we note that if for any x_1 , $u(x_1, A_a)$ and $u'(x_1, A_a)$ are finite, the solution may be integrated somewhat further to obtain $u(x_2, A_a)$, $u'(x_2, A_a)$ for $x_2 > x_1$. From equation (A7.1) it is evident that an increase in either $u(x_1, a)$ or $u'(x_1, a)$ will result in an increase in d^2u/dx^2 in the interval $x_1 < x < x_2$ so that u and u' at x_2 are monotonically increasing functions of u and u' at x_1 . Hence if u and u' at x_1 have the property P_1 , so do u and u' at x_2 . By extending this argument we conclude that u and u' at any value of x have the property P_1 . (A rigorous proof can easily be completed along these lines provided that $|f(x)|$ is finite.)

Similarly it may be shown, starting from (A7.5), that $u(x, A_b)$ is a monotonically increasing function of A_b and $u'(x, A_b)$ is a monotonically decreasing function of A_b .

In order to have a solution satisfying (A7.4) and (A7.5) we must have, for any selected point x ,

$$u(x, A_a) = u(x, A_b) \quad (\text{A7.6})$$

$$u'(x, A_a) = u'(x, A_b) \quad (\text{A7.7})$$

Now as the equation $u(x, A_a) = u(x, A_b)$ varies from $-\infty$ to $+\infty$, $u'(x, A_a)$ varies from $-\infty$ to $+\infty$ and $u'(x, A_b)$ varies from $+\infty$ to $-\infty$, monotonically and continuously. Hence there is one and only one solution of (A7.1) satisfying (A7.4) and (A7.5).

In order to verify that the solutions discussed in Section 2 are correct for large and for small K , we show schematically in Fig. A1 the solution for a representative K as a dashed line together with the curve $u = u_0(y) = \sinh^{-1} y$. In terms of u_0 , equation (2.16) becomes

$$\frac{d^2u}{dy^2} = \frac{1}{K^2} (\sinh u - \sinh u_0). \quad (\text{A7.8})$$

From the symmetry of the equation, it is evident that u must be an odd function of y and hence that the solution must pass through the origin. The boundary condition in this case will be that $u \rightarrow u_0$ for $y \rightarrow \pm \infty$ so that there will be no space charge far from the junction. We can conveniently use the origin as the point at which the solution from $y = +\infty$ joins that from $y = -\infty$; from symmetry, this requires merely that $u = 0$ when $y = 0$.

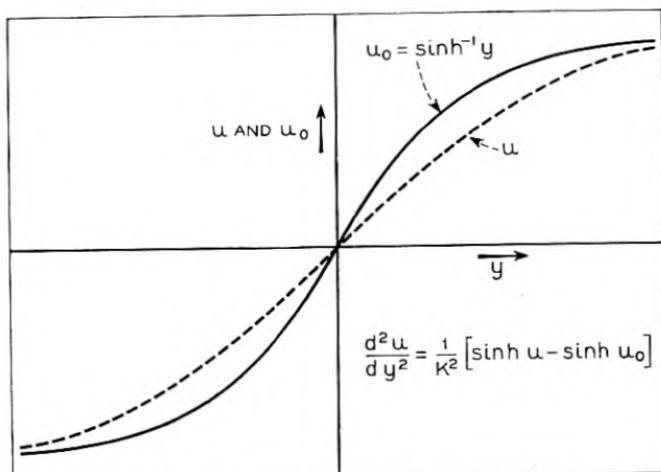


Fig. A1—Behavior of the solution of Equation (2.16) or (A7.8).

For large negative y , $u = \sinh^{-1} y$ and $du/dy = 1/\cosh u_0$ so that du/dy is small. It is at once evident that, for large values of K , u must lie above u_0 so that the integral

$$(1/K^2) \int_{-\infty}^y (\sinh u - \sinh u_0) dy = \frac{du}{dy} \quad (\text{A7.9})$$

will be large enough to make the solution $u(y)$ pass through the origin. If $u - u_0 > 2$ over the region of largest difference, the space charge will be largely uncompensated and the solution will correspond to that used in equation (2.18). On the other hand, as $K \rightarrow 0$, the requirement that $u(y)$ pass through the origin leads to the conclusion that $u - u_0$ must be small for all values of y . The possibility that u oscillates about u_0 need not be considered since it may readily be seen that, if for any negative value of y , say y_1 , both $u(y_1)$ and $u'(y_1)$ are less than $u_0(y_1)$ and $u'_0(y_1)$, then $u(y)$ and $u'(y)$ are progressively less than $u_0(y)$ and $u'_0(y)$ as y increases from y_1 to 0. Hence, if for negative y the u curve goes below the u_0 curve, it cannot pass through the origin.

APPENDIX VIII

LIST OF SYMBOLS

(Numbers in parentheses refer to equations)

$$a = (N_d - N_a)/x \quad (2.14)$$

A = admittance per unit area of junction (4.23)

A_p = component of A due to hole flow into n -region (4.12) (4.24)

A_n = component of A due to electron flow into p -region (4.25)

A_T = component of A due to varying charge distribution in transition region

A also used as a constant coefficient in various appendices

b = ratio of electron mobility to hole mobility

b = symbol for base in Sections 5 and 6

B constant coefficient in various expansions in appendices

c = symbol for collector in Section 6; a length in Appendix III

C = capacity per unit area

C_n, C_p (4.25) (4.27) as for A_n, A_p

C_T (2.42) (2.45) (2.56) as for A_T

D = diffusion constant for holes (bD is the diffusion constant for electrons)

$e = 2.718 \dots$

f see Appendix 7

g = rate of generation of hole-electron pairs per unit volume (3.1)

G = conductance per unit area of junction

G_n, G_p as for A 's

$i = \sqrt{-1}$

I = current density

I_n, I_p = current densities due to electrons and holes (2.5) (2.6) (4.10)

I_{n0}, I_{p0}, I_{p1} (4.11) (4.12) (4.18) (4.19)

I_s, I_{ns}, I_{ps} saturation reverse current densities (4.11) (4.18) (4.21)

I_r see text with (4.35)

J = subscript in Section 3 for junction Fig. 5 equation (3.11)

k = Boltzmann's constant

K = space charge parameter (2.17)

L = length

$L_a = n_i/a$ (2.15)

L_D = Debye length (2.12)

L_n, L_p = diffusion lengths for electron in p -region and holes in n -region (4.8)

L_r = length required for potential increase of kT/q in region of constant field (4.32) Appendices II and IV

L_1 corresponds to a-c. field, Appendix IV

n = density of electrons

- n_n, n_p = equilibrium densities of electrons in n - and p -regions
 p = density of holes
 p_n, p_p = equilibrium densities of holes in n - and p -regions
 p_0 = d-c. component of non-equilibrium hole density (4.3)
 $p_1 \exp(i\omega t)$ = a-c. component of non-equilibrium hole density (4.3)
 P = total number per unit area of holes in specimen (2.35)
 q = electronic charge ($q = |q|$)
 $Q = qP$ = total charge per unit area (2.39)
 r = recombination coefficient for holes and electrons (3.1)
 R = resistance of unit area
 R_0 = resistance of unit area obtained by integrating conductivity (3.10),
 Appendix I
 R_1 = effective series resistance, discussed in connection with (3.13)
 s = rate of recombination per unit area of surface per unit hole density,
 Appendix V
 S = susceptance per unit area (imaginary part of admittance)
 S_p, S_n, S_T as for A 's.
 t = time
 T = temperature in $^{\circ}K$
 T = subscript for transition region
 $u = q\psi/kT$ (2.9), $q(\psi - \varphi_1)/kT$ (2.32), Appendix VII
 v_0 and $v_1 e^{i\omega t}$ = d-c. and a-c. components of voltage applied in forward direc-
 tion (4.2)
 W = width of space charge region in abrupt junction, Section 2.4
 w = half thickness of n -region or transistor base of Sections 5 and 6.
 w = half width of square rod in Appendix V.
 x = coordinate perpendicular to plane of junction
 y, z = transverse coordinates, Appendix V
 y = reduced length (2.17), Appendix VII
 α = current gain factor in transistor (6.4)
 α = parameter in Appendix III and VI
 α_{ij} = parameter in Appendix V
 β_i = parameter in Appendix V
 γ = parameter in Appendices II, IV and VII
 ϵ = symbol for emitter Section 6
 $\theta_i = \beta_i w$ Appendix V
 κ = dielectric constant
 μ = mobility of a hole ($b\mu$ = mobility of electron)
 ν = rates of recombination etc., Appendix VI
 ρ = charge density (2.1)
 σ = conductivity

σ_i = conductivity of intrinsic material (4.15)

σ_n = conductivity of *n*-region $\doteq qb\mu n_n$

σ_p = conductivity of *p*-region $\doteq q\mu p_p$

τ = time

τ_n, τ_p = life times of electrons in *p*-region and holes in *n*-region (3.2) (3.3)
(4.7)

τ_T = relaxation time of transition region, Appendix IV

$\varphi, \varphi_p, \varphi_n$ = Fermi level and quasi Fermi levels (2.2) (2.4)

$\delta\varphi$ = applied voltage across specimen in forward direction, Section 2.3,
(4.2)

χ = sw/D in Appendix V

ψ = electrostatic potential (2.2)

ω = circular frequency of a-c. (4.2)

Band Width and Transmission Performance

By C. B. FELDMAN and W. R. BENNETT

In modern communication theory band width plays an important role as a transmission parameter. The authors discuss the significance of signal band width and frequency occupancy in relation to other transmission factors such as power, noise, interference, and overall performance for certain specific multiplex systems under assumed operating conditions. The intent of the paper is to show how such problems may be attacked rather than to find an unequivocally best system.

The scope of the paper is described by the following table of Headings and Captions.

I. INTRODUCTION

- Fig. 1. Outline of multiplex transmission methods
- 1. Non-simultaneous Load Advantage in FDM
- Table I. Non-Simultaneous Multiplex Load Advantage
- 2. Instantaneous Companding Advantage in Time Division
- Fig. 2. Performance of an experimental instantaneous compander
- 3. Non-simultaneous Load Advantage in Pulse Transmission
- Fig. 3. Quantizing noise in each channel when PCM is applied to an FDM group
- 4. Signal Band Width and Frequency Occupancy
- 5. Regeneration and Re-Shaping
- 6. The Radio Repeater
- Fig. 4. Arrangement of two-way two-frequency repeater of television type showing spacing of bands and antenna discrimination
- Fig. 5. Discrimination of I.F. and R.F. circuits in television type repeater

II. BAND WIDTH CHARACTERISTICS

- Fig. 6. Basic pulse shape and its spectrum
- Fig. 7. Marginal condition in reception of AM pulses and an FM wave in presence of noise.
- Fig. 8. Time allotments in Pulse Position Modulation
- Fig. 9. PPM-AM; fluctuation noise. Relations between band width, power, and signal-to-noise ratio.
- Fig. 10. PPM-AM; CW and similar system interference. Relations between band width and signal-to-interference ratio.
- Fig. 11. PPM-FM; fluctuation noise
- Fig. 12. PPM-FM; CW and similar system interference
- Fig. 13. PAM-FM; fluctuation noise
- Fig. 14. PAM-FM; CW and similar system interference
- Fig. 15. PCM-AM; peak interference
- Fig. 16. PCM-FM; fluctuation noise
- Fig. 17. PCM-FM; CW and similar system interference
- Quantized PPM
- Fig. 18. Comparison of quantized PAM with quantized PPM
- Fig. 19. FDM-FM; fluctuation noise
- Fig. 20. FDM-FM; CW interference

III. BAND WIDTH AND POWER TABLES

- Table II. Optimum Band Widths for Minimum Power for Message Type Circuits
- Table III. Optimum Band Widths for Minimum Power for Program Type Circuits

Table IV. Minimum Band Widths and Corresponding Power Requirements for Message Type Circuits

Table V. Minimum Band Widths and Corresponding Power Requirements for Program Type Circuits

IV. FREQUENCY OCCUPANCY TABLES FOR RADIO RELAY

1. Antenna Characteristics

Fig. 21. Directional selectivity of microwave antenna

Fig. 22. Simplified route patterns for study of selectivity required in congested localities

Table VI. True Frequency Occupancy of Various Message Grade Radio Relay Systems for Congested Routes

Table VII. True Frequency Occupancy of Various Program Type Radio Relay Systems for Congested Routes

2. Conclusions as to Radio

Table VIII. Comparisons of Band Width and Frequency Occupancy for Systems of Equal Ruggedness

V. MORE ABOUT THE NON-SIMULTANEOUS LOAD ADVANTAGE

Fig. 23. Theoretical possibilities of exploiting non-simultaneous load advantage by an elastic PLM-AM system

VI. OVERLOAD DISTORTION AND NOISE THRESHOLD

Fig. 24. Noise threshold and overload ceiling in frequency divided PCM groups

Fig. 25. Overload characteristics of multirepeater systems

VII. PULSES, SPECTRA, AND FILTERS

Fig. 26. Typical pulses and their spectra

1. Pulses for PPM

2. Pulses for PAM

3. Pulses for PCM

4. Optimum Distribution of Selectivity Between Transmitting and Receiving Filters

Fig. 27. Crossfire between frequency divided pulse groups

5. Delay Line Balancing

VIII. TRANSMISSION OVER METALLIC CIRCUITS

Fig. 28. Variation of circuit length with number of repeater sections in an AM system with fixed power capacity and noise figure

Fig. 29. Optimum number of repeater sections and maximum circuit length for metallic AM system with fixed power capacity and noise figure

Fig. 30. Optimum number of repeater sections and maximum circuit length for metallic FM system with limiting only at end of system

Fig. 31. Optimum number of repeater sections and maximum circuit length for metallic PPM-AM system with reshaping at every repeater

Fig. 32. Optimum number of repeater sections and maximum circuit length for metallic FM system with limiting at every repeater

Fig. 33. Relation between circuit length, power, and number of repeaters in radio relay systems

IX. CONCLUSIONS

X. APPENDICES

Appendix I. Noise in PCM Circuits

Fig. 34. Stepping and sampling an audio wave

Fig. 35. Variation of quantizing noise with sampling frequency

Appendix II. Interference Between Two Frequency Modulated Waves

Fig. 36. Geometric solution for resultant phase of two frequency modulated waves

Appendix III. PCM for Band Width Reduction

Appendix IV. Supplementary Details of Derivation of Band-Width Curves
 Appendix V. Sampling a Band of Frequencies Displaced from Zero
 Fig. 37. Minimum sampling frequency for band of width W

LIST OF FREQUENTLY USED SYMBOLS

- B = radio signal band width in megacycles. (Not to be confused with frequency occupancy).
 b = base or radix of PCM system.
 β = peak-to-peak frequency swing of FM systems in megacycles.
 F_b = width of baseband (video band) in megacycles.
 f_r = repetition or sampling frequency in megacycles.
 K = load rating factor (amplitude ratio).
 \log = logarithm to base 10.
 \ln = logarithm to base e .
 N = number of channels in a multiplex system.
 n = number of digits in a PCM system or number of spans in a multirepeater system.
 P = wanted carrier amplitude.
 P_n = mean fluctuation noise power per megacycle.
 Q = interfering carrier amplitude.
 S = span length in miles.
 U = band spacing factor.

I. INTRODUCTION

CARRIER systems for the transmission of many telephone channels on a single metallic circuit have grown to be very important in the telephone network. Since the development of the coaxial cable system in which 480 channels are transmitted in a 2-mc baseband, advances in high frequency techniques, including the war-accelerated microwave art, have inspired efforts to utilize the broad band capabilities of high transmission frequencies. Some of the efforts have related to the wave-guide conductor but mainly they relate to radio relay transmission. As a consequence of these efforts a considerable number of new multiplex methods for use at microwave frequencies have been devised. All of these methods employ bandwidth more liberally than the 4 kc per channel rate associated with single sideband carrier systems, in return for which various transmission advantages are obtained. Theoretically, transmission advantages can be sacrificed to permit bandwidth reduction but the transmission requirements then become very severe. Bandwidth as a transmission parameter has grown to a prominent position in modern communication theory as set forth by Shannon et al.^{1, 2, 3}

The liberal use of bandwidth, employed in an effective way, operates to permit higher noise and distortion within a system and, in the case of radio relay systems, operates to permit higher interfering signals from other radio systems. When all the frequency space necessary to avoid mutual inter-

¹ C. E. Shannon, "A Mathematical Theory of Communication," *Bell Sys. Tech. Jl.*, Vol. 27, pp. 379-423, 623-654, July-Oct. 1948.

² B. M. Oliver, J. R. Pierce and C. E. Shannon, "The Philosophy of PCM," *Proc. I. R. E.*, Vol. 36 (1948), pp. 1324-1331.

³ C. E. Shannon, "Communication in the Presence of Noise," *Proc. I.R.E.*, Vol. 37 (1949), pp. 10-21.

ference between systems in a congested area is taken into account, certain wide-band methods, less vulnerable to interference, may be as or more efficient in the use of frequency space than other narrower band multiplex methods.

The principal purpose of this paper is to examine, for various systems, the relations governing the exchange between frequency space and transmission advantages.

It will be shown that the preferred multiplex method depends in part upon:

1. The grade of facility required; low-grade and high-grade channels lead to different preferences. These preferences also are influenced by the length of circuit.

2. The nature of the transmission obstacle over which advantage is sought. These obstacles may be: (a) intrasystem distortion (phase distortion, overload distortion, etc.) and noise; (b) intersystem interference as between similar radio systems or between different types of radio systems, operating on the same frequency.

Other factors beside the transmission considerations discussed here are likely to be involved in a practical multiplex application; hence the system preferences arrived at in this study may not be the controlling factors in practice.

Before a detailed analysis is undertaken, it may be helpful to examine and comment upon the chart shown in Fig. 1. All of the multiplex methods shown here have been studied sufficiently to permit their approximate evaluation with the aid of some theoretical considerations and subject to certain qualifications as pointed out from time to time. Variations and combinations of these are possible,⁴ some of which will be discussed later.

In addition to the two general classifications of frequency and time division there is a third type based on carrier phase discrimination. A familiar example is the quadrature carrier system,⁵ which is capable of yielding two channels for each double sideband width. In another form⁶ each of N channels is modulated simultaneously on $N/2$ carriers with a different set of carrier phases provided for each channel. Time division multiplex may be regarded as a kind of phase discrimination in which the signal is modulated on harmonic carriers so phased as to balance out except during the channel sampling time intervals. In true phase discrimination,

⁴ A comprehensive listing and discussion of various combinations will be found in a recent paper by V. D. Landon, "Theoretical Analysis of Various Systems of Multiplex Transmission" *R.C.A. Review*, vol. IX, numbers 2 and 3, June-Sept. 1948, pp. 287-351, 438-482.

⁵ H. Nyquist, "Certain Topics in Telegraph Transmission Theory," *A.I.E.E. Trans.*, April, 1928, pp. 617-644.

⁶ W. R. Bennett, "Time Division Multiplex Systems," *Bell Sys. Tech. J.* Vol. 20, pp. 199-221, April, 1941.

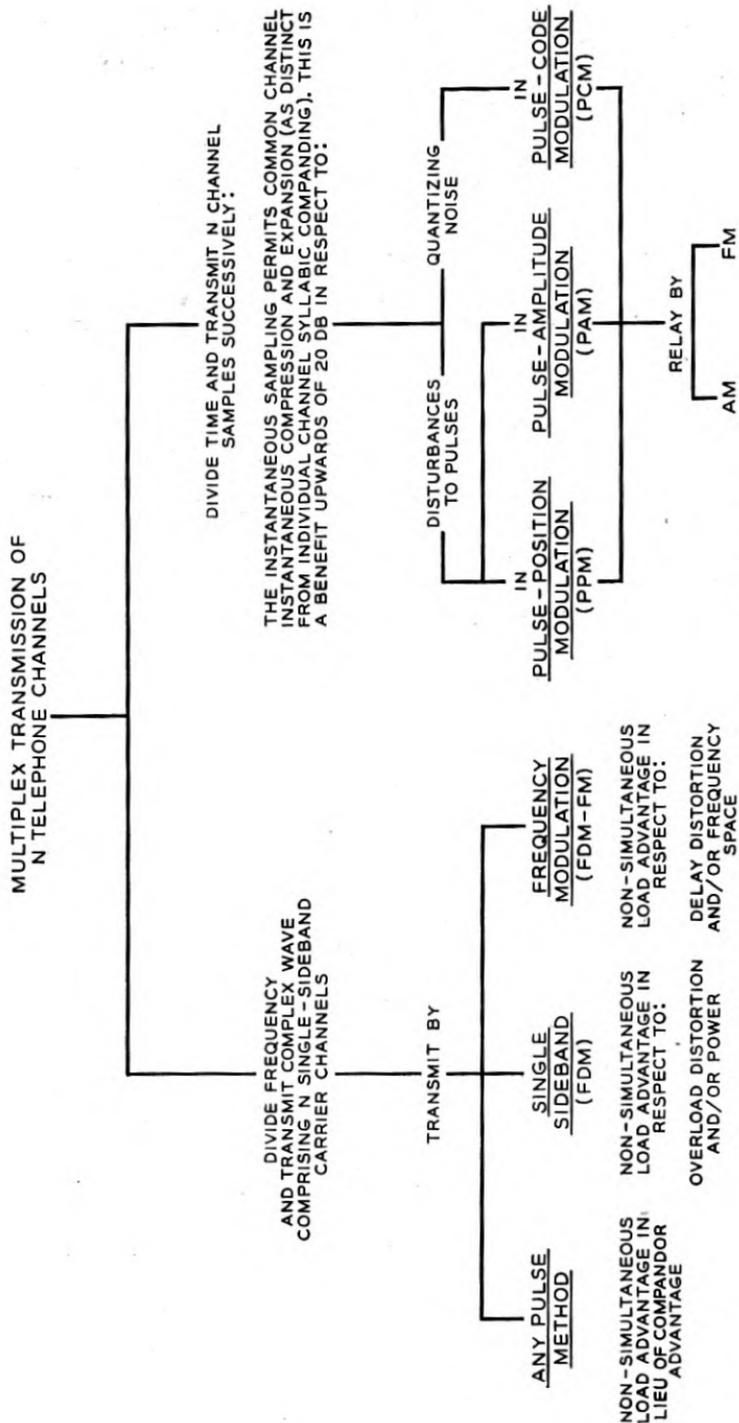


Fig. 1—Outline of multiplex transmission methods.

however, there need be no separation of channels in either time or frequency, and a homodyne detection process is required at the receiver for channel selection. The necessary precision of instrumentation seems in general more difficult to achieve than with either frequency or time division, and only minor prospects appear for exchange of bandwidth for transmission advantage.

Of the systems tabulated, the frequency division method (FDM) with single-sideband suppressed-carrier transmission is the only method in which bandwidth cannot be traded for some transmission advantage.⁷ This system will be used as a standard of comparison. The PAM method, with transmission by AM pulses, can trade upon bandwidth only as a means for reducing interchannel crosstalk. In the other pulse systems, as well as all systems using FM, bandwidth may be expended to gain advantage over noise, intersystem interference, and, generally speaking, intrasystem distortion and noise.

NON-SIMULTANEOUS LOAD ADVANTAGE IN FDM

The *non-simultaneous load advantage* pertaining to frequency division multiplex refers to the fact that the channel sidebands rarely add to an instantaneous value even approaching the value N times the peak value for one channel. This means that the required peak capacity of a relay system transmitting the N channels increases slowly with N . Current toll transmission practice provides for relative power capacity roughly as follows.⁸

TABLE I
NON-SIMULTANEOUS MULTIPLEX LOAD ADVANTAGE

N	Required Relative Power Capacity	Advantage
1	0 db	0
10	+6 db	20 - 6 = 14 db
100	+9 db	40 - 9 = 31 db
500	+13 db	54 - 13 = 41 db
1000	+16 db	60 - 16 = 44 db

To emphasize the strikingly large non-simultaneous load advantage statistically obtainable with conversational speech we may examine Table I and note, for instance, that the capacity of a 1000-channel system is completely

⁷ We have in mind here a system such as Type K or L in which a minimum separation of adjacent channels in frequency is used. It is true that by spreading the channels far apart in frequency, a reduction in cross-modulation falling in individual channels could be obtained, but the resulting amount of improvement is minor compared with that offered by a corresponding band increase in the other systems.

⁸ B. D. Holbrook and J. T. Dixon, "Load Rating Theory for Multichannel Amplifiers," *Bell Sys. Tech. J.*, Vol. 18, pp. 624-644, October, 1939. The values in the table come from curve C, Fig. 7, taking the single channel sine wave power capacity as +9.5 dbm.

used up by peak instantaneous voltage when 994 channels are disconnected and 6 carry full-load tones.

If a group of carrier channels in frequency-division multiplex were translated to microwave frequencies, the overload distortion affecting the transmission would be predominantly of the third-order class. To a first approximation the third order distortion follows a cube law and may be predicted from the single-frequency compression. We assume here that the power capacity of the repeater is the output at which the single frequency compression occurring through the complete system does not exceed 1 db.⁹ This criterion applies roughly to systems of several hundred channels capacity, and to present transmission standards.

INSTANTANEOUS COMPANDING ADVANTAGE IN TIME DIVISION

In time-division systems, as ordinarily understood and known in the current literature, each channel successively is provided with its full-load capacity, and thus a non-simultaneous load advantage does not accrue. However, because of the sampling process, instantaneous compression may be applied at the transmitting terminal before noise and distortion are encountered; when complementary expansion is applied at the receiving terminal the noise is suppressed. The expanded samples derived at the receiving terminal then bear an improved relation to noise, particularly in the case of weak samples. Such an instantaneous companding process applied without sampling to a continuous speech wave requires a greatly increased transmission band between compressor and expander but, in a time division system, no more bandwidth is needed to transmit the speech *samples* after they have been compressed than before. An instantaneous compander currently being used experimentally to handle 12 channels in time division has the noise performance characteristics shown¹⁰ in Fig. 2. It is shown as applied to a telephone system in which the channel noise power (unweighted) would be 45 db down from the power of a sine wave which employs the full load capacity provided for the "loudest talker". Abrupt overloading is assumed to take place when peak amplitudes exceed that of the full-load tone. The location, at -7.5 on the load scale, for the power representing the very loud talker (one in a thousand) conforms approximately to current practice. The speech volumes, referred to the point of zero db transmission level, are shown for the sake of completeness.

⁹ In a multi-repeater system the compression accumulates. This means that each repeater must be restricted to operate approximately $10 \log n$ db below the 1 db compression point of one repeater. (n denotes the number of repeaters.) See Section VI.

¹⁰ Use of the same curve to represent the performance with tone or speech implies an independence of wave form which is not rigorously valid. Calculations based on speech-like signals have indicated that the curve for tone loading is a good approximation when average power is used as the criterion in the manner shown.

The compression and expansion result in a uniform improvement of 26 db for weak signals including the "very weak talker" and a lesser improvement for stronger signals. The noise power in the absence of speech is 71 db

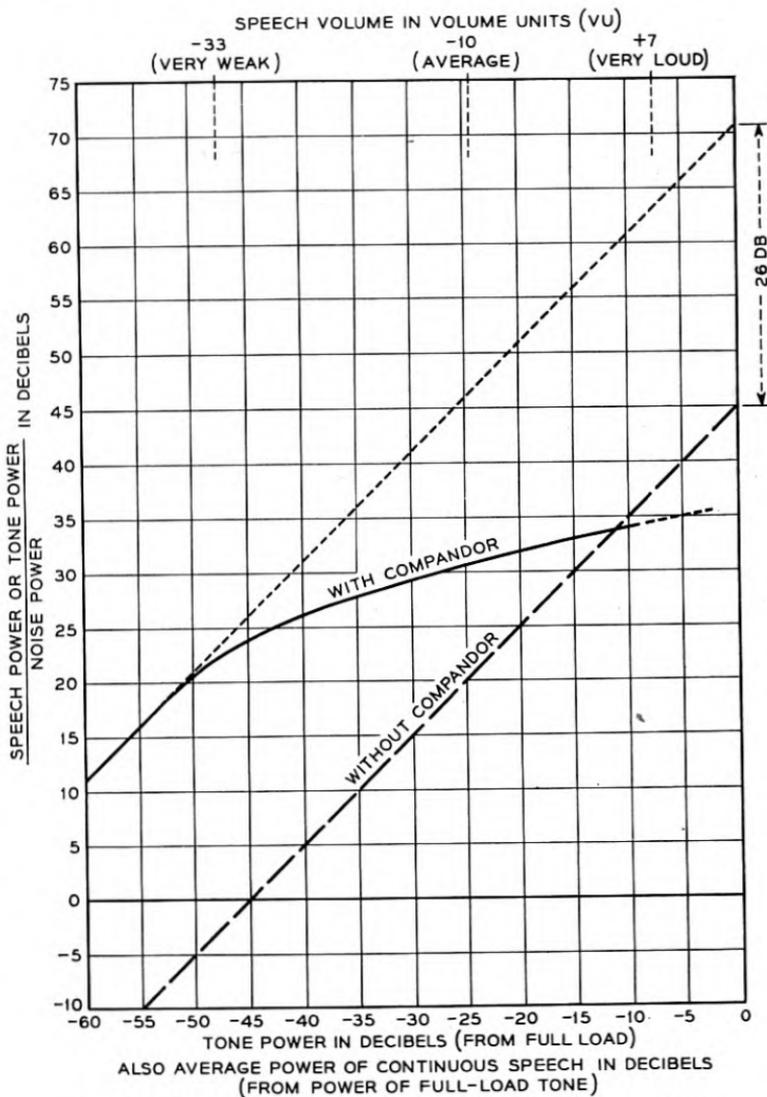


Fig. 2—Performance of an experimental instantaneous compandor.

below the power of the full load sine wave which the channel is designed to handle. The performance is substantially equivalent for telephone purposes to a 71-db circuit without companding, in spite of the fact that for all except the very weak talkers the average noise is greater than with the 71-db circuit.

The noise is increased (over the 71-db value) by the compandor only in the presence of speech and then only in proportion (roughly) to the amplitude, and so becomes masked by the speech. The masking is sufficient to make impairment of medium and loud speech imperceptible provided that the ratio of speech power to noise power is greater than about 22 db. Under these conditions we are justified in defining the equivalent signal-to-noise ratio in terms of the low level noise.

For compandors with a more drastic characteristic, yielding more low-level improvement, the high-level noise increase is enhanced and the limit to this enhancement is controlled by the "uncompanded" signal-to-noise ratio (the ratio without companding.) Thus the amount of low-level improvement that is permissible from the standpoint of high-level performance is determined by the uncompanded signal-to-noise ratio. Experiments have shown that the permissible low-level improvement increases several db for each db increase in uncompanded signal-to-noise ratio. Another way of putting it is that the value of the equivalent signal-to-noise ratio in the speech channel determines the amount of compandor advantage which may be invoked to attain that ratio, and that the permissible compandor contribution increases nearly as fast as the equivalent signal-to-noise ratio. The uncompanded signal-to-noise ratio is thus required to increase only slightly.

For the 45 db uncompanded signal-to-noise ratio of Fig. 2 the compandor could have been designed to yield more than the 26 db low-level improvement shown without impairing high-level performance. In the time-division systems of message grade with which we will deal later, a 22 db compandor advantage is assumed.¹¹

In the quantized systems included in the PCM heading the instantaneous compandor advantage applies to the granularity, or quantizing, noise in the same way as to the common kinds of noise which plague other systems. The compandor of Fig. 2 was actually used in an experimental PCM system.¹² A discussion of quantizing noise appears in Appendix I and a more comprehensive treatment appeared in the Bell System Technical Journal recently.¹³

In transmitting frequency divided groups of channels by pulse methods¹⁴

¹¹ This is the maximum compandor advantage permissible for a circuit equivalent to 57 db signal-to-noise ratio. We will use this figure in connection with power requirements for circuits whose signal-to-noise ratio is intended to be equivalent to 60 db but since we presume that interference or crosstalk may be present in an amount equal to noise and since the compandor acts on interference as on noise we must protect against high level impairment on the basis that the noise is 3 db greater.

¹² L. A. Meacham and E. Peterson, "An Experimental Multichannel Pulse Code Modulation System of Toll Quality", *Bell Sys. Tech. Jl.*, Vol. 27, pp. 1-43, Jan. 1948.

¹³ W. R. Bennett, "Spectra of Quantized Signals," *Bell Sys. Tech. Jl.* Vol. 27, pp. 446-472, July, 1948.

¹⁴ If the group occupies a frequency range extending from zero to F_b , the minimum sampling rate is well known to be $2F_b$. If the group range does not start at zero frequency the minimum sampling rate is not twice the highest frequency of the group but lies between two and four times the width of the band depending on the location of the band. This matter is treated in Appendix V.

the instantaneous compandor advantage is substantially zero because, at full system load, companding actually increases the total noise. In time division the noise is increased at full load by companding but, as discussed earlier, this is permissible because full load occurs only with loud talkers who mask the noise. In a frequency divided group transmitted by pulse methods nearly full load may be produced when a number of loud talkers are momentarily active; the weak talkers then enjoy no improvement due to companding but may, on the contrary, suffer some degradation.

NON-SIMULTANEOUS LOAD ADVANTAGE IN PULSE TRANSMISSION

Transmission of a frequency-divided group by pulse methods does, however, permit the realization of a portion of the non-simultaneous load ad-

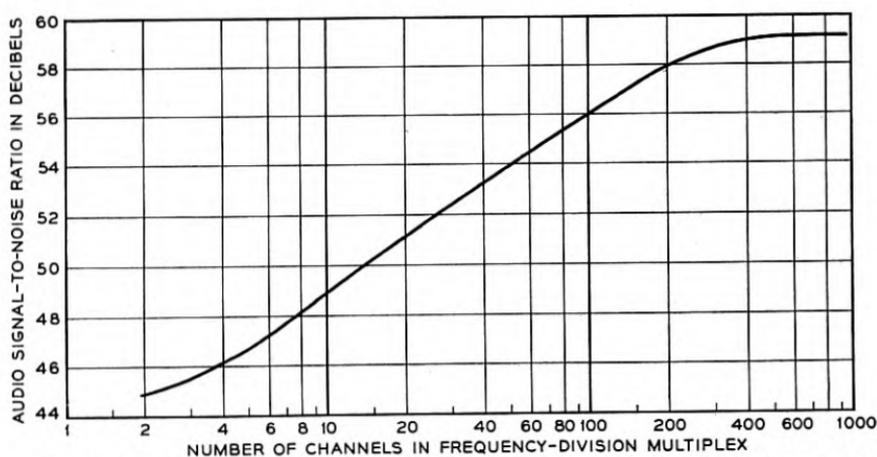


Fig. 3—Quantizing noise in each channel when PCM (128 equal steps) is applied to an FDM group.

vantage in lieu of the instantaneous compandor advantage realizable in time division. A pulse system, designed to carry N channels in time-division with a certain full load signal-to-noise ratio (without companding) may also be used to carry N channels in frequency division. These N channels in frequency division may be treated as a single channel N times wider than the time-divided channels and sampled N times faster. The ratio of the full-load signal in this wide band to the noise in the wide band turns out to be the same as the corresponding ratio in each of the narrow, time-divided channels. This fact makes the transmission of large groups of frequency-divided channels advantageous compared to small groups. For instance, in a 100-channel frequency-divided group, a single channel would have available a total load capacity which is 9 db less than that for the multiplex group. This comes from Table I, which shows that 100 channels require

9 db more range than one channel. This makes the signal-to-noise ratio in a channel 9 db lower than if all of the entire load capacity were devoted to that channel. However, in the 100-channel group a single channel receives only 1% of the noise power in the entire band, so a 20 db improvement accrues on this score. The net improvement is $20 - 9 = 11$ db. Applied to a 128 step PCM system in which the full-load signal-to-noise ratio is 45 db¹⁵ (7 digits binary PCM), the full-load signal-to-noise ratio in one channel of a 100-channel group thus becomes $45 + 11 = 56$ db. With smaller groups than 100 channels the signal-to-noise ratio falls to 45 db while for larger it reaches 59 db, as shown in Fig. 3. Better results than these are obtainable with time division and instantaneous companding, as shown in Fig. 2, but these results may have significance in relation to the transmission of television by a pulse method, such as PCM. If a 128-step system were used, a large frequency-divided group of telephone channels filling the television band could be substituted for television when desired.

A more powerful application of the non-simultaneous load advantage in time division will be discussed later in Section V.

SIGNAL BAND WIDTH AND FREQUENCY OCCUPANCY

We define *signal band width* as the width of the signal spectrum (or more realistically as that portion of the signal spectrum which must be preserved in order to make the signal sufficiently undistorted). *Frequency occupancy* is greater than signal bandwidth in two respects:

First, the frequency range accepted by the receiving filter at the end of each span must be greater than the signal band for reasons of filter imperfection. In all of the pulse or FM systems it would be advantageous from the circuit point of view to make the receiving filter much wider in order thereby to reduce the phase distortion over a small central frequency range occupied by the signal band. The assigned frequency space must include the entire band accepted by the receiving filter. Our comparisons will assume that the filters make use of an appropriate amount of refinement to conserve frequency space.

Second, frequency occupancy must include the multiplication of assignments made necessary to avoid interference between converging or intersecting radio relay routes, between the two directions of a single route, or between a main route and a spur.

Our procedure in evaluating these systems will be to plot for each system certain curves relating power, signal bandwidth and channel signal-to-noise ratio or signal-to-interference ratio for various associated transmission

¹⁵ Appendix I shows that the quantizing noise power at the minimum sampling frequency is the same for wide and narrow signal bands. This illustrates the general principle used here.

conditions. From these curves and other pertinent data we will prepare tables which show the significant frequency occupancy for various radio relay conditions. Such tables will be made for two grades of transmission facilities and for the extremes of signal bandwidth, one corresponding to minimum power and the other to minimum bandwidth. The minimum power condition prevails when the bandwidth has been increased, and the power reduced, to the point where any further increase of bandwidth would require an increase of power to prevent noise from "breaking" either the pulse slicer or the FM limiters.¹⁶ The minimum bandwidth condition occurs when any further band limitation operates to impair the signal too much, assuming that the power is ample to override noise.

REGENERATION AND RE-SHAPING

Two distinct classes of relay operation exist, one applying to the quantized systems (PCM) and the other applying to non-quantized systems. When the transmitted signal is intended to convey a continuous range of values (amplitude, time or frequency) noise and distortion accumulate as the signal progresses from repeater to repeater over a relay route. If, however, a range of values is represented by a discrete (quantized) value, a signal may suffer displacement within the boundaries of that range without altering the information conveyed by the signal. If, therefore, in one span of the relay route the displacement is confined to those boundaries the signal may be regenerated and re-transmitted as good as new. No accumulation of noise and distortion need occur, therefore, as the signal traverses span after span. The most common application of regenerative repeatering is in printing telegraphy where the signal is either a mark or space and, if correctly determined, may be re-transmitted afresh.

In all of the non-quantized systems the repeaters must have low distortion so that a signal may be conveyed through a large number of them (say 133 for a 4000-mile circuit made up of 30-mile spans) without too much mutilation. In spite of good repeater design a signal passing through such a large number of repeaters will accumulate considerable noise, interference from other systems, and distortion characterizing the repeater design limitations. In non-quantized systems there is no escaping accumulation of this sort. In pulse systems, for instance, phase distortion, common in flat band repeaters, may result in tails and the like, while cumulative frequency discrimination (band narrowing), characterizing simple forms of linear phase repeaters, results in cumulative broadening of the pulses. In the former case the tails

¹⁶ In this connection it is of interest to mention that if the objective were a *very low grade* circuit the power required to prevent breaking might be higher than that required by a method having no improvement threshold, and no-power saving could be accomplished by the bandwidth exchange principle. For circuits of telephone grade this situation does not occur.

may eventually grow large enough to break the slicer (if the system employs such a device) while in the latter case the reduced pulse slope and the spreading out of time bounds may also bring about transmission disaster. In both cases these growing distortions successively reduce the margin that it is necessary to provide for noise and interference. To circumvent such effects, the pulses may be reshaped at all or some repeaters. Reshaping consists of measuring the information conveyed by the pulse (in the time or amplitude dimension) and sending out a new pulse of standard shape possessing that measured characteristic in time or amplitude. This process is distinctly different from regeneration as practiced in quantized systems; in general, reshaping can only be counted upon to confine the rate of accumulation of noise, interference and crosstalk to that of power addition from span to span.

In FM systems any distortion which results in amplitude "modulation" of the FM wave may be treated with limiting at each repeater to prevent such amplitude variation from accumulating and breaking the limiter. Like pulse reshaping, this measure does not stop the accumulation of disturbance to the intelligence. Certain kinds of distortion may be combated by double FM.¹⁷

Reshaping (or, in the case of FM, limiting) may be employed to conserve power in the systems having an improvement threshold. Without reshaping, the minimum repeater power is the marginal¹⁸ value for the total noise accumulated from all spans. If reshaping is practiced at each repeater the power need be marginal for the noise from only one span. More bandwidth must be used, then, in exchange for the lower power; and, while this in turn increases the marginal power, the result is a net power saving. Tables II and III of Section III illustrate this point and Section VIII illustrates its application to metallic circuits.

THE RADIO REPEATER

Repeaters for relaying television signals must achieve low distortion and we will take a current design and assume that such a repeater represents a basis for discussing the transmission of multiplex telephony by non-quantizing methods. This repeater employs, in the two-way application, four antennas and two frequencies as shown in Fig. 4. It is proposed to transmit 5-mc video television signals by FM in bands spaced 40 mc. The repeater employs double detection and the band separation is effected mainly by the

¹⁷ Leland E. Thompson, "A Microwave Relay System," *Proc. I.R.E.*, Vol. 34, December, 1946, pp. 936-942.

¹⁸ By marginal power is meant that power which just safely exceeds the improvement threshold power. For a given noise level, minimum power is achieved when the bandwidth improvement factor yields the required signal-to-noise ratio in the channel with the power that is marginal for that bandwidth.

selectivity following conversion to intermediate frequency. Microwave receiving filters afford enough selectivity to divert alternate bands into their correct frequency-converting units without disturbing the other bands; and microwave combining filters serve in the transmitting side of the repeater to

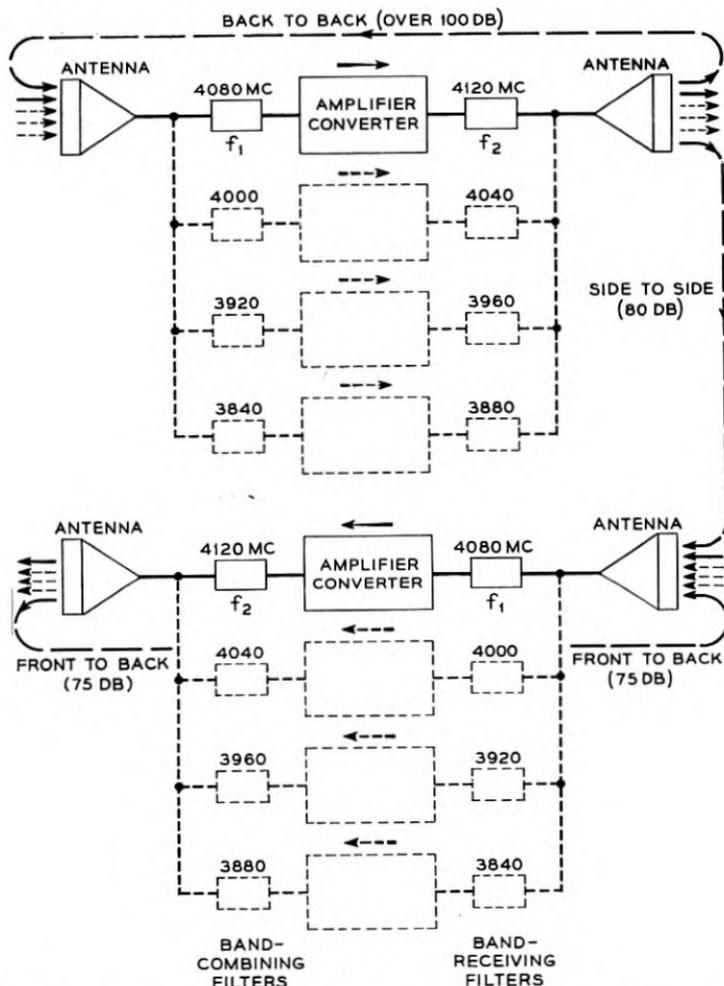


Fig. 4—Arrangement of two-way two-frequency repeater of television type showing spacing of bands and antenna discrimination.

bring the bands into the common antenna with small loss and small mutual disturbance.¹⁹ The combining and separating processes are made easier by the interleaving of transmitting with receiving frequencies. Interleaving

¹⁹ W. D. Lewis and L. C. Tillotson, "A Non-reflecting Branching Filter for Microwaves," *Bell Sys. Tech. J.*, Vol. 27, pp. 83-95, January, 1948.

In all of the systems in which bandwidth may be exchanged for tolerance to interference we restrict, in our comparison tables, the minimum bandwidths to those which provide at least the 44 db tolerance demanded in the two-frequency plan.

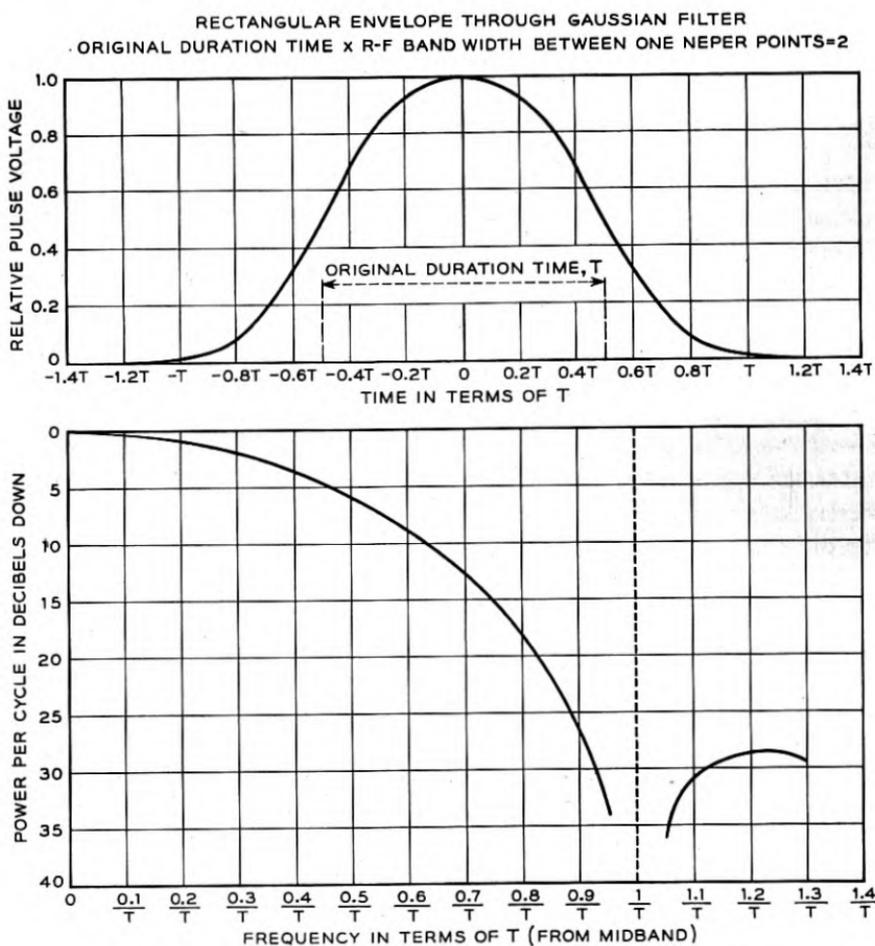


Fig. 6—Basic pulse shape (approximately sinusoidal) and its spectrum.

II. BAND WIDTH CHARACTERISTICS

The type of pulse assumed in the various pulse transmission systems is shown on Fig. 6. As pulse 4 of Fig. 26, it is further discussed in Section VII. The spectral density or distribution of energy vs. frequency associated with such a pulse is also shown. It is evident from this curve that omission of frequencies beyond a baseband width $1/T$ can result in distortion or tails of

only a few per cent of the pulse height. We define signal bandwidth for the pulse systems studied here as $1/T$, or $2/T$ in the r - f medium, i.e., double sideband is assumed in all of the AM pulse cases. In assuming double sideband, we bow to the obvious difficulty of dealing circuit-wise with single sideband and its pulse demodulation problem.

In the FM systems we define the radio signal bandwidth as the peak-to-peak frequency swing, β , plus two times the baseband width, $2F_b$.

We shall consider individually the following types of systems where the meaning of the symbols is explained in Fig. 1.

- | | | | |
|-----------|-----------|-----------|-----------|
| 1. PPM-AM | 3. PAM-AM | 5. PCM-AM | 7. FDM |
| 2. PPM-FM | 4. PAM-FM | 6. PCM-FM | 8. FDM-FM |

The source of disturbance may be either fluctuation noise, a constant-frequency interfering wave (CW), or a similar but independent system operating on the same frequency allocation. CW interference may fall anywhere within the radio signal band. Interference from echoes, which is a special case of similar system interference, is not treated. In certain cases echoes such as might be produced by multiple reflections in waveguide connections to the top of radio towers may be more detrimental than independent system interference of the same amplitude. We assume that such echoes are suppressed sufficiently by good design.

Our first set of curves, Figs. 9-20, exhibits quantitatively the audio signal-to-noise and audio signal-to-interference ratios which can be obtained with increased radio bandwidth in the various systems. Audio signal is taken to be the power of a test tone which fully loads one channel. Audio noise is expressed as the total noise power in the channel. Audio interference is expressed as the power of all of the extraneous frequencies produced in a channel by the assumed interfering signal. The term "radio bandwidth" is intended to mean double-sideband width and does not imply that the transmission is necessarily by radio. Two of the systems, PAM-AM and FDM, are omitted from this study because, as has been pointed out earlier, they do not provide a significant basis of exchange of bandwidth for suppression of noise and interference. The other systems possess this trading property in varying degree as illustrated by the curves. The FDM and PAM-AM systems are entered in Table IV and discussed under Section III. For comparison with the following curves it may be of interest to note here that for 1000 4-kc message channels in FDM, the bandwidth (single sideband) is 4 mc., and the received power for a 60-db audio signal-to-noise ratio must be -77 dbw.²⁰ This is the power in a sine wave which employs

²⁰ Throughout this paper we shall use the abbreviation "dbw" for power expressed in decibels relative to one watt.

the full load capacity in accordance with Table I, at a point where the noise power is -189 dbw per cycle of bandwidth (15 db noise figure, NF).

In most of these curves, plotted for 1000 message channels, the bandwidth scale runs to hundreds of megacycles. We do not mean to imply that the microwave transmission medium can be relied upon to transmit faithfully such wide-band signals or that circuit techniques for producing them are available. As suggested by Fig. 4, the 1000-channel system might be divided into several groups of fewer channels to avoid frequency selective transmission difficulties or circuit limitations. The total frequency occupancy is not altered by such a division, while the required power per group is reduced in proportion to the number of channels.²¹

Curves are shown of audio signal-to-noise ratio as a function of radio signal bandwidth at constant power and at marginal power. Audio signal-to-interference ratios are plotted against radio bandwidth for marginal ratio of radio signal power to interfering signal power. By "marginal power", we mean the radio signal power which just safely exceeds the threshold below which noise or interference causes system failure. In the case of fluctuation noise, any further increment of bandwidth from this point is untenable without an increase in radio signal power. Points on the marginal power curves show as abscissa the bandwidth at which minimum radio power is required to obtain the audio signal-to-noise ratio given by the ordinate. In calculating these curves, we have specified the marginal condition as occurring when the peak disturbance is actually 3 db below the theoretical value which just breaks the system. These relations are shown graphically in Fig. 7. We have in this paper followed the accepted practice of ignoring all fluctuation noise peaks exceeding the rms voltage by more than 12 db. *Radio signal power* is taken as the power averaged over a cycle of the high frequency in the FM wave, or, in the AM pulse case, over a cycle of the high frequency when the pulse is maximum. A curve is included in Fig. 9 showing marginal AM radio pulse power values for various bandwidths of fluctuation noise and a similar curve for FM is shown in Fig. 13. A noise figure of 15 db²² is assumed for the receiver. We have taken the noise bandwidth as equal to the signal bandwidth throughout. This equality cannot be quite attained in actual systems because of the departure of physical filters from ideal characteristics. In practice an allowance for frequency instability would also have to be included.

The relation of the PPM pulse to channel allotment time is shown in Fig. 8. Pulses in channels adjacent in time can just touch when full load signals are impressed on each. The slicer operates at half the pulse height which, for the assumed pulse shape, is also the point of maximum slope. The width of the

²¹ These statements are not exactly true for FDM and FDM-FM, where multiplex load rating is used in the design.

²² This means that the noise power is 189 db below a watt per cycle of bandwidth.

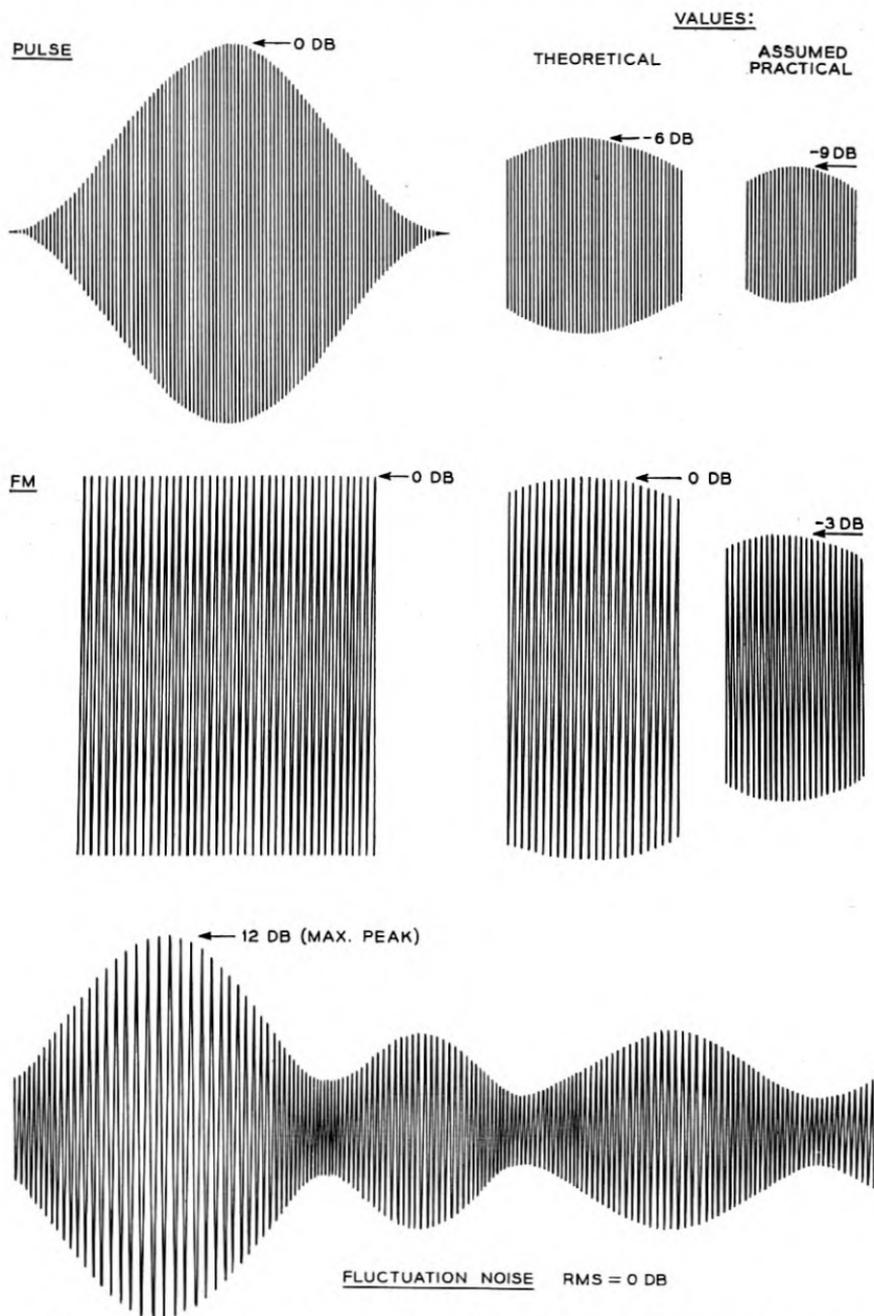


Fig. 7—Marginal condition in reception of AM pulses and an FM wave in presence of noise. Pulse case applies to PCM only if binary.

pulse is inversely proportional to the signal bandwidth. The time available for modulating the pulse position is equal to the channel time minus the pulse duration. The combination of these factors leads to the PPM "slicer"

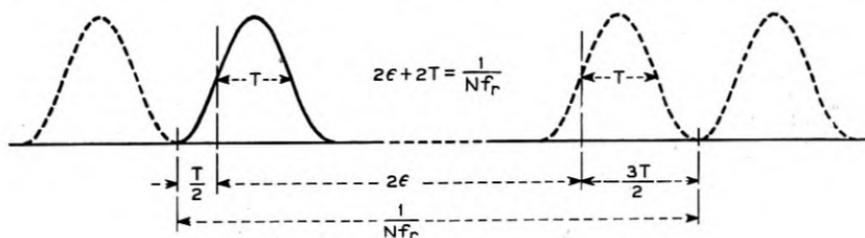


Fig. 8—Time allotments in pulse position modulation.

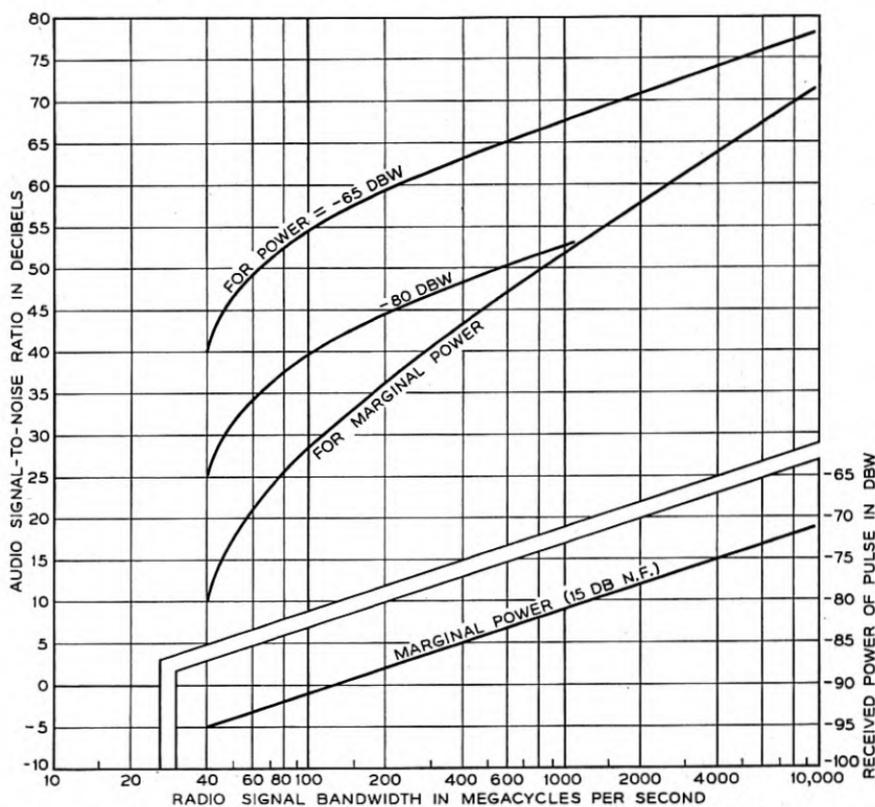


Fig. 9—PPM-AM; performance with respect to fluctuation noise. Relations between bandwidth, power, and audio signal-to-noise ratio for 1000 4-kc channels.

advantage" which, when applied to the r - f pulse-to-noise ratio, gives the full load audio tone-to-noise ratio in each channel. Details of this calculation and others pertaining to various pulse systems are included in Appendix IV.

FIG. 9—PPM-AM, FLUCTUATION NOISE

The curves of Fig. 9 were computed from the slicer advantage derived in Appendix IV. The asymptotic slope of the constant power curves of 3 db per octave of bandwidth reflects the 6 db advantage due to the two-fold greater pulse slope (slicer advantage) diminished by the 3 db increase of noise accepted by the two-fold wider band. In the marginal power case

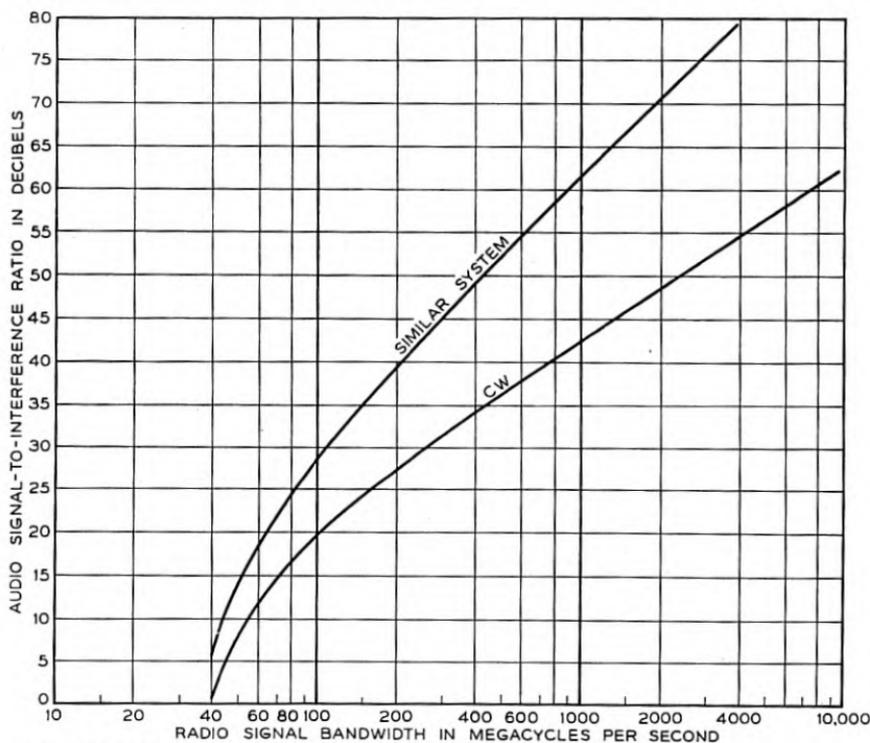


Fig. 10—PPM-AM; performance with respect to CW and similar system interference for 1000 4-kc channels with ratio of pulse to interference marginal. Relations between bandwidth and audio signal-to-interference ratio.

the power is increased with bandwidth so the slicer advantage is preserved and the slope is 6 db per octave.

The sharp reduction of signal-to-noise ratio with reduced bandwidth appearing at the left end of the curves arises from immobilizing the pulse position as the widened pulse uses up more of the total channel time allotment. According to the definition of bandwidth used here and the plan of Fig. 8, no modulation is possible when the bandwidth is $2/T$ and T is half of the channel time. For 1000 channels the channel time is 0.125 microseconds and $T = 0.0625$, which makes the audio signal-to-noise ratio zero at 32 mc.

FIG. 10—PPM-AM, CW AND SIMILAR SYSTEM INTERFERENCE

The curve of Fig. 10 for marginal ratio of pulse power to CW interfering power has the same shape as the corresponding curve of Fig. 9 for marginal power over fluctuation noise. There is a shift of 9 db in ordinates, however, because the peak factor of the CW interference is 9 db less than that of fluctuation noise. The interference from similar systems follows a different law because of the "exposure factor" arising from the finite probability that the interfering pulse does not overlap the wanted pulse. A straightforward probability calculation taking into account the distribution of pulse voltages in an interfering system occupying the same radio frequency band yields the curve shown on the assumption that the repetition frequencies are asynchronous. As the bandwidth is increased the pulses become shorter and their coincidences less frequent, leading asymptotically to a 9 db per octave slope instead of the 6 db per octave of the CW interference.

FIG. 11—PPM-FM, FLUCTUATION NOISE

In the transmission of PPM by FM there are two sources of advantage over noise. One is the ordinary FM advantage and the other is the slicer advantage of PPM acting on the noise remaining in the FM output. There are, likewise, two separate conditions for system failure; one a breaking of the limiter and the other a breaking of the slicer. A certain amount of radio power will result in marginal operation of the limiter for a certain frequency swing. The corresponding deviation ratio is the quotient of the frequency swing and the baseband width; this ratio is maximum when the baseband is least. Except in the region near the minimum PPM band, advantage accrues faster with bandwidth in FM than in PPM. It is apparent, therefore, that most of the radio bandwidth should be devoted to FM advantage. The optimum proportioning occurs when the baseband width has a small value but not so small as to invoke an unsurmountable penalty by not providing for any position modulation. Mathematical analysis given in Appendix IV shows that the optimum baseband for the pulse position modulation varies with radio bandwidth in the manner shown in Fig. 11 by curve 1. Curve 2 shows the audio signal-to-noise ratio vs. radio bandwidth when the baseband width follows curve 1 and the FM limiter is marginal. It is of interest to compare curve 2 with the poorer performance of the dashed curve 3 which is calculated for the case in which both the FM limiter and the PPM slicer are marginal. The baseband width for the double marginal condition follows curve 4. Curves 5 and 6 show audio signal-to-noise ratio vs. bandwidth for constant radio power and optimum baseband. The curve of marginal amount of radio power is not given in Fig. 11, but is the same as the one given later in Fig. 13.

FIG. 12—PPM-FM, CW AND SIMILAR SYSTEM INTERFERENCE

The curve showing interference from a similar system of lower power was based on a calculation of the beat spectrum between two FM waves, both frequency modulated over the same r - f range (β mc) by 8 mc. The phase difference between the 8-mc modulating frequencies was assumed to vary, giving rise to various beat spectra. The power in those beat components accepted by a band zero to F_b was averaged over all 8-mc phase differences

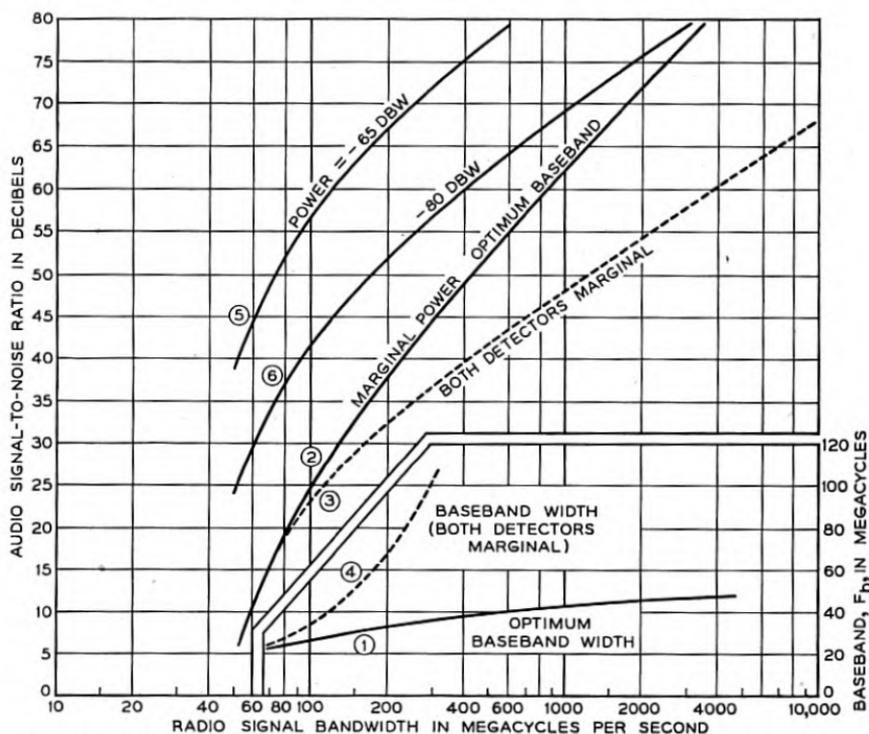


Fig. 11—PPM-FM; performance with respect to fluctuation noise. Relations between bandwidth, power, and audio signal-to-noise ratio for 1000 4-kc channels.

and this average power was taken as a measure of the interference to which the baseband signal is subjected. As outlined here, this procedure is valid for interference between idle PAM-FM systems in which the FM waves are frequency modulated as assumed above. For the PPM-FM case in which we are here interested, we take the position suggested by the transient viewpoint that the effect of interference from spaced pulses will not be much different because of their spacing and so we apply the slicer advantage possessed by the wanted system to the total interference calculated above and obtain the curve shown. At the left hand end where the pulse spacing is only slight

and the sinusoidal frequency modulation is nearly the correct representation, the above procedure is not subject to much suspicion. The validity of the right-hand portion of the curve is upheld by the fact that it is about 12 db lower than the marginal fluctuation noise curve of Fig. 11. If the wide swing FM interfering wave had a spectrum much like fluctuation noise of the same power as the FM wave the difference would be 9 db.

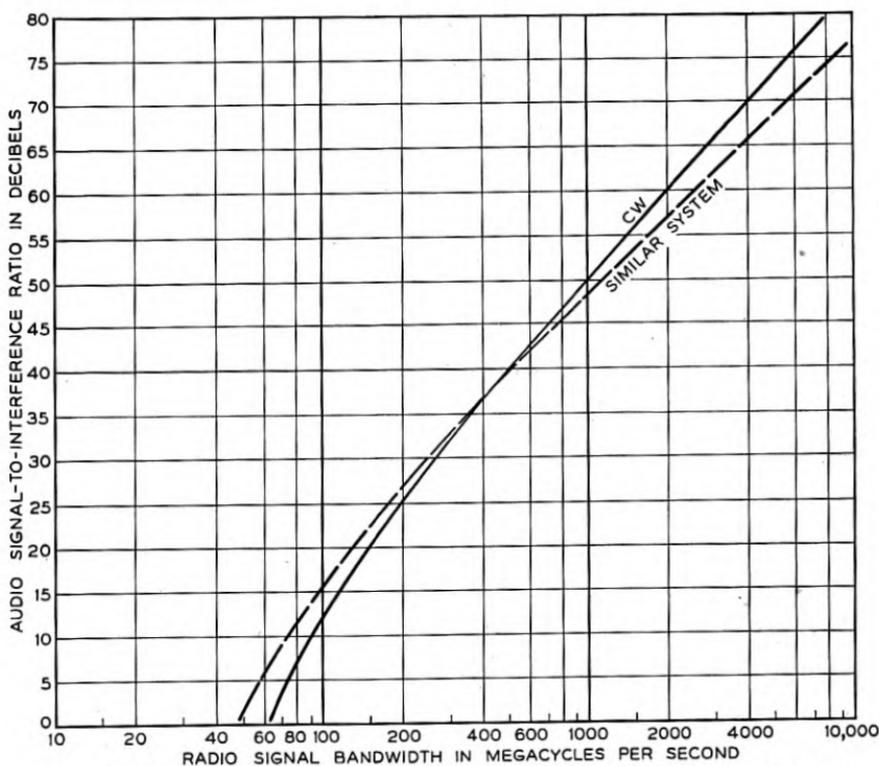


Fig. 12—PPM-FM; performance with respect to CW and similar system interference for 1000 4-kc channels with ratio of FM wave to interference marginal. Relations between bandwidth and audio signal-to-interference ratio when baseband is optimum for suppression of fluctuation noise.

It has been explicitly assumed that both systems are idle, but we see no reason to believe that if either or both were normally active the interference would be significantly different for our purposes.

Audible interference from a CW wave is caused by a disturbance to the frequency of the FM wave. Let us first assume that the CW frequency lies near the middle of the frequency swing range. No disturbance to the FM wave occurs as its frequency passes through coincidence with that of the CW but, as the frequencies diverge, the magnitude of the disturbance as well as

the frequency of the disturbance increases linearly. The baseband filter is excited only during the time the difference is less than F_b . Thus, the disturbance results from a series of perturbations to the otherwise smooth frequency variation of the FM wave. The time during which these perturbations can affect the baseband filter is short compared with the shortest pulse the baseband filter can pass, except when the baseband width is greater than half of the swing. This occurs at the extreme left-hand end of the curve. We have not attempted to calculate the response to these transients except to note that the response is a pulse which extends roughly $2T$ from its point of origin, peaking somewhere near the center of this interval. If we assume that the PPM pulses are closely spaced ($\epsilon = 0$) so that they result in a wave frequency modulated by 8 mc, there are two such evenly spaced disturbance pulses per cycle of modulation (two per $2T$ interval) and therefore there is an almost continuous disturbance wave in the base band filter output whose amplitude does not greatly exceed its RMS value. We have accordingly calculated the power sum of all the extraneous frequencies passed by the baseband filter, assuming the FM wave to be sinusoidally modulated. The location of the CW frequency giving greatest interference power was used in these calculations except in the wide band cases where the worst frequency appeared to be near the edge of the band. Here the transient viewpoint indicated that the resulting interference in the baseband would be greater if the CW frequency were nearer the center.

If the trailing edge is used to measure the time of the pulse, the principal disturbance of this time arises from the perturbation produced at the leading edge of the same pulse, and so the calculation for close-spaced pulses is not greatly in error when applied to wider-spaced pulses. If the leading edge were used the worst CW frequency for widely spaced pulses would be one differing from the rest frequency by F_b and the interference would be worse, we think, than that arising from the frequency worst for trailing edge operation.

It has been explicitly assumed that the system is idle, but we see no reason to believe that the interference would be significantly different with normal activity.

FIG. 13—PAM-FM, FLUCTUATION NOISE

Fluctuation noise in a PAM-FM system produces the sloped noise spectrum characteristic of FM in the output of the frequency detector. The noise power per cycle is zero at zero frequency and increases with the square of the frequency. The baseband filter accepts only the portion of the spectrum between zero and F_b . If instantaneous sampling of the signal values is used, all noise frequencies in this range are equally effective as causes of errors. Use of a channel gate of maximum permissible duration

consistent with a satisfactory margin over crosstalk from adjacent channels furnishes a practical method of discriminating against the influence of noise components near the top of the baseband where the noise spectrum is strongest. The exact shape of the gate is not very critical. The curves have been calculated for a rectangular gate coincident with the channel allotment time, which is just possible without crosstalk in the case of non-

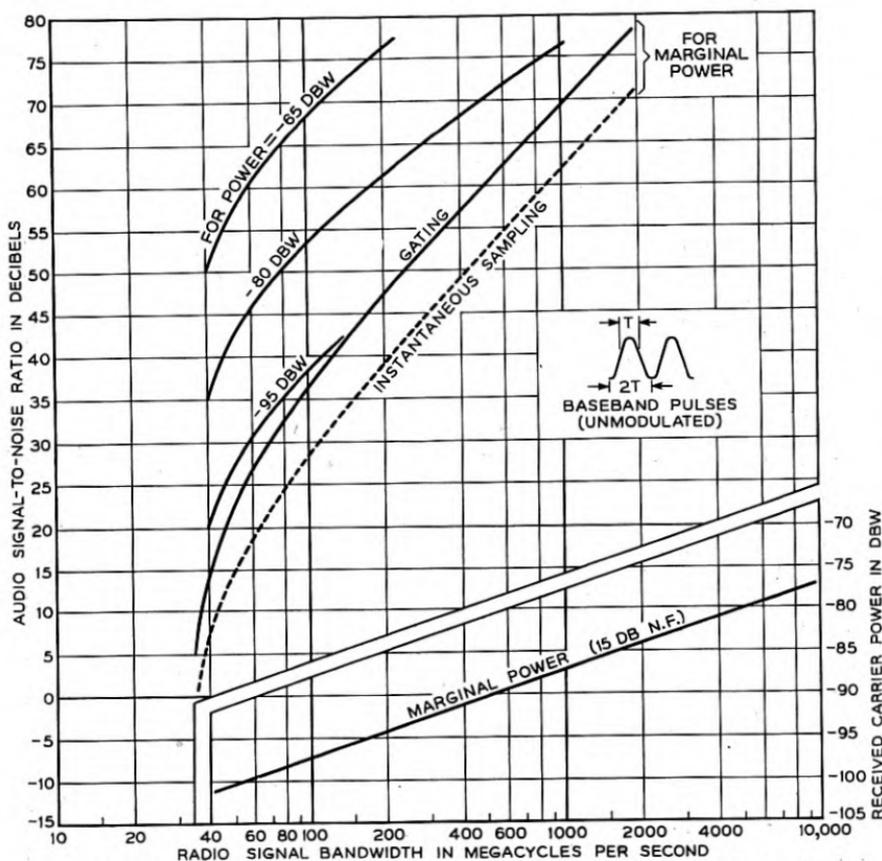


Fig. 13—PAM-FM; performance with respect to fluctuation noise. Relations between bandwidth, power, and audio signal-to-noise ratio for 1000 4-kc channels.

overlapping sinusoidal pulses. A somewhat shorter rectangular gate or a gate of sinusoidal shape leads to very nearly the same results. The advantage of gating as compared to instantaneous sampling is approximately 8 db. Calculation of the gated noise is a straightforward process if based on the concept of the FM noise spectrum acting as signal on a product demodulator in which the carrier consists of the harmonics of the gating function. Each harmonic demodulates the spectrum centered about the harmonic

frequency and contributes audio power proportional to the product of harmonic power and spectral density. The channel filter accepts only the demodulated noise falling in the audio signal range.

The marginal power curve has been drawn for a 3 db ratio of peak carrier to peak interference or 12 db ratio of mean carrier power to mean fluctuation noise power. Curves for specific amounts of received power are included as

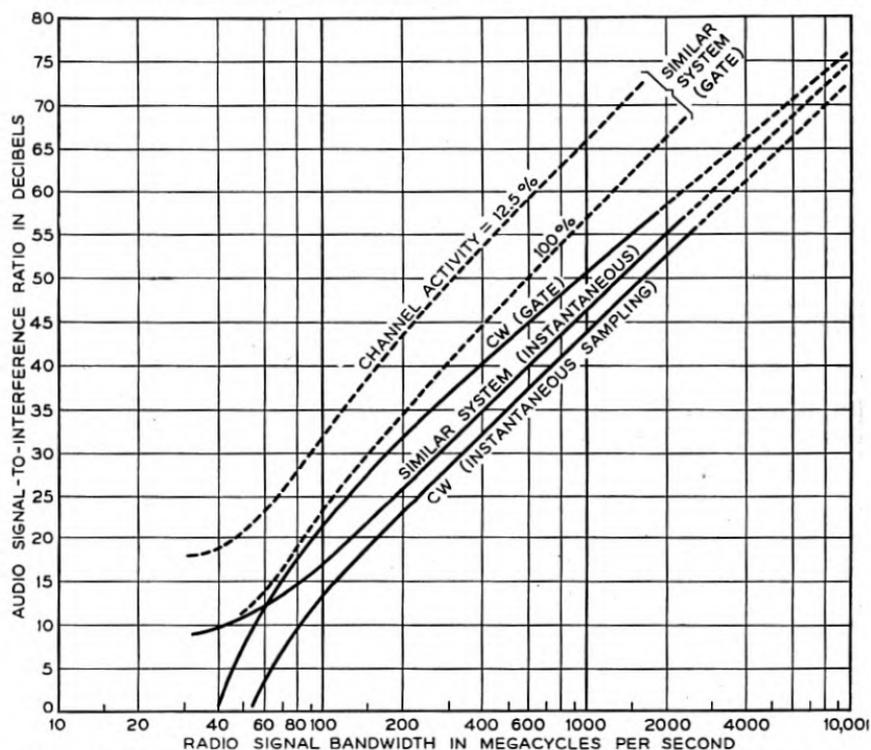


Fig. 14—PAM-FM; performance with respect to CW and similar system interference for 1000 4 kc channels with ratio of FM wave to interference marginal. Relations between bandwidth and audio signal-to-interference ratio.

well as the curve of marginal received power vs. radio signal bandwidth for a receiver with 15 db noise figure.

FIG. 14—PAM-FM, CW AND SIMILAR SYSTEM INTERFERENCE

CW interference can be calculated conveniently by assuming all channels idle and determining at what frequency within the radio signal band a CW component of fixed power produces maximum disturbance of an audio channel. This worst possible amount of disturbance is then assumed not to be much affected by the various channel loading conditions existing during normal operation of the system. When all channels are idle, the transmitted

carrier of a PAM-FM system using sinusoidal pulses assumes the particularly simple form of an FM wave modulated by a sinusoidal signal having frequency Nf_r (8 mc for 1000 channels) and total frequency swing $\beta/2$. Hence the rigorous steady state solution for interference between CW and sinusoidally modulated FM was calculated and the interfering components falling in the baseband range selected. The gating function was then applied to these components in the same way as described above for fluctuation noise, and the resulting products falling in the audio channel range evaluated. The signal-to-interference ratio was expressed as the ratio of rms signal power received from a full-load channel test tone to the rms value of the audio interference. A range of frequency locations for the CW interference was investigated for each radio signal bandwidth and the one giving maximum audio interference used for the point on the curve. The worst position of the CW was usually found to be near the extremities of the idle channel frequency swing. Curves are shown for a rectangular gate of maximum duration and for instantaneous sampling.

When the source of interference is a similar system, we assume that the midband frequencies differ only slightly. With both systems idle, we have two sinusoidally modulated FM waves which are identical except for (1) a small variable difference between mean carrier frequencies and (2) a variable phase shift between the two modulating frequencies. The interference falling in the baseband consists of steady state components which are approximately harmonics of the channel slot frequency Nf_r . As is characteristic of FM interference, the amplitude at the m th harmonic contains a factor proportional to m ; and the component near zero frequency, the approximate zeroth harmonic, is very small. If we gate this interference with a rectangular gate of duration $1/Nf_r$, we find that the gated output vanishes for input components at $Nf_r, 2Nf_r, \dots$, because these frequencies are located at the infinite loss points of the aperture admittance. The gate would transmit the zeroth harmonic, but this component tends toward zero amplitude. Our conclusion is that two idle PAM-FM systems accurately lined up to occupy the same frequency range are balanced against interference from each other when a rectangular channel gate of full channel allotment time is used. The balance tends to disappear as the channels are loaded because the interference then spreads throughout the base band instead of being concentrated at the blind spots of the aperture. Thus, for the first time in our consideration of pulsed systems, we are obliged to take account of channel loading conditions.²³

²³ A wave could be frequency modulated about a central frequency by PAM pulses of plus and minus sign and an idle system would thus consist of a wave of constant frequency. The weaker of two such idle systems aligned in frequency would produce no (or very little) interference in the other, using either channel gating or instantaneous sampling. The susceptibility to CW interference would be greater than in the biased modulation assumed above, however.

We note that the balance disappears if instantaneous sampling is used instead of gating because there is no longer any aperture discrimination. The curve for instantaneous sampling is plotted on Fig. 14. Calculation of this curve brings out the fact that the amplitude of the interfering components also depend on the framing phase difference between the two systems. When the framing frequencies are in phase, the two waves have a constant frequency difference, and the interference vanishes. We assume the phase difference as equally likely to fall anywhere within a complete cycle and average the received interference power over all phases. The curve is found to approach an asymptotic ordinate of 9 db at minimum band width as the frequency swings on the two systems approach zero together. The 9 db limit is compounded of 3 db from the marginal ratio between the two carriers, 3 db from averaging over the carrier phase difference, and 3 db from averaging over the framing phase difference. When wide bands and large swings are used, the curve approaches parallelism with the dashed one of Fig. 13 for fluctuation noise but about 15 db lower. Of this difference 9 db is accounted for by the higher marginal mean power level. The remainder is assignable to differences in spectral distribution; in particular the *r-f* spectrum of idle similar system interference is concentrated in half of the band instead of being uniformly spread as in fluctuation noise.

The curve for gated similar system interference has been estimated by assuming that, with all channels loaded and a wide frequency swing, the performance is like that with fluctuation noise except for a 3 db correction allowed for the more concentrated spectrum. This gives an asymptote on the right parallel to the solid marginal curve of Fig. 13 and 12 db lower. At the left the curve must approach the same asymptote as the one for instantaneous sampling. It then seems reasonable to assume that the interference power is directly proportional to the number of active channels and the curve for an average of one eighth of the channels loaded is obtained by raising the full load curve 9 db.

FIG. 15—PCM-AM

The curves on the left show how the audio signal-to-noise ratio varies with bandwidth, the audio noise being quantizing, or granularity, noise as discussed in Appendix I. The number of digits per code symbol is n and the number of digit values (including zero), i.e., the base, is b . Bandwidth is $2/T$ where T is the time per pulse and is therefore 16 mc per digit. The curves are, of course, a set of discrete points rather than continuous as shown. The steep rise is to be contrasted with the 3 to 9 db per octave slopes of the curves previously presented.

The curves on the right plot the maximum values (with a 3 db allowance included) of peak noise or interference, referred to the highest pulse value,

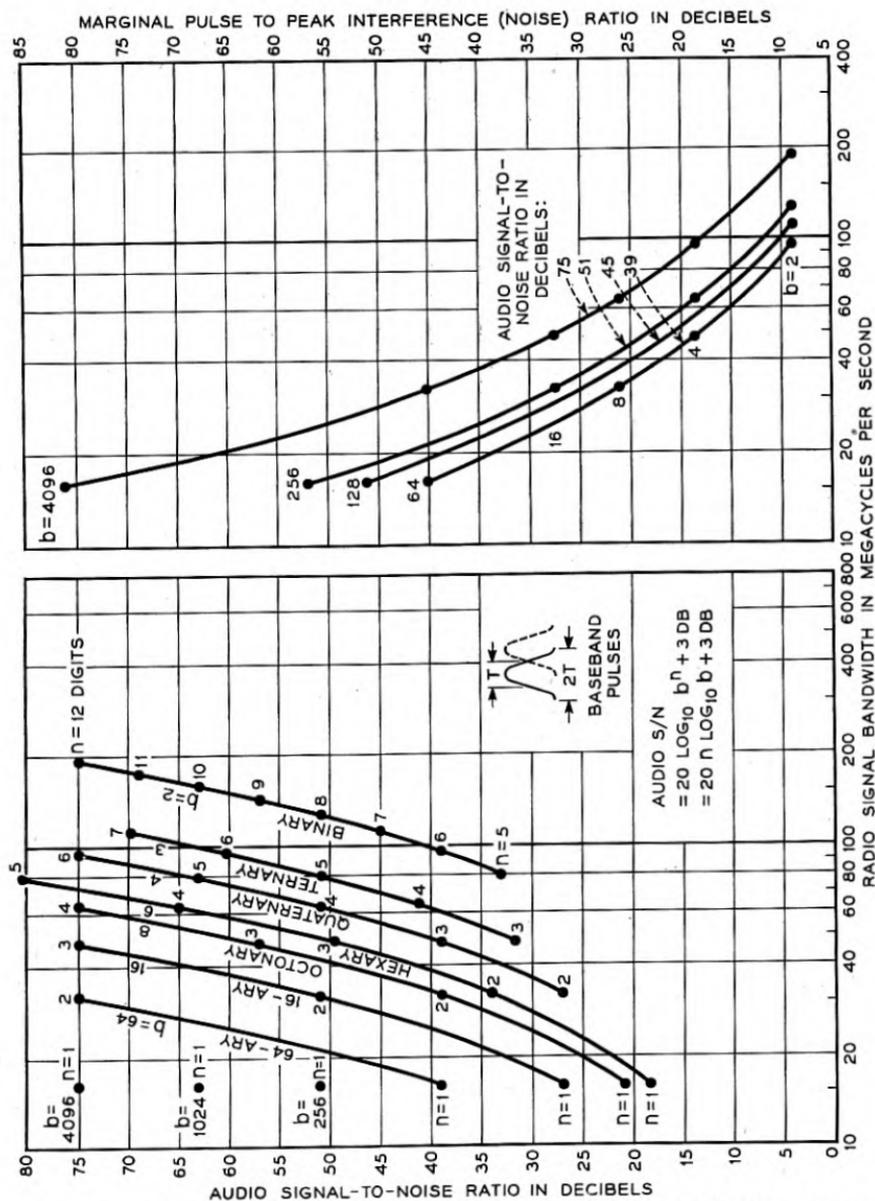


Fig. 15—PCM-AM; performance with respect to number of steps. Relations between bandwidth and audio signal-to-noise ratio and between bandwidth and required pulse-to-interference ratio for 1000 4-k channels.

permitted by the system, versus bandwidth for three different audio signal-to-noise ratios. The lower boundary of the curves corresponds to binary PCM while the left-hand boundary corresponds to the other extreme, namely

quantized PAM having the number of steps necessary to yield the specified signal-to-noise ratio. The quantized PAM bandwidth of 16 mc assumes the use of overlapping sinusoidal pulses as in binary PCM. Actually, such an overlap would be hazardous in the higher base systems; and quantized PAM, like unquantized PAM, should perhaps be assigned more time per pulse but not as much as $2T$ because regeneration could be employed to prevent accumulation of interchannel crosstalk. The tables presented later do not include the bandwidth increase that would follow such an increase in time per channel.

The curves at the right in Fig. 15 are terminated at 16 mc corresponding to one pulse per channel. In accordance with the principles of Appendix III more than one channel per pulse can be transmitted, theoretically. To include such a hypothetical case of less than one digit per channel, the curves could have been extended upward to the left. The 39 db signal-to-noise ratio curve would have reached an ordinate of 81 db at 8 mc on the bandwidth axis.

It is of interest to compare the audio signal-to-noise ratio of unquantized PAM with that of quantized PAM for the interference ratios demanded by quantized PAM. In the case of marginal CW interference the audio noise (evaluating the audio disturbance as noise of equivalent power) turns out to be the same as the quantizing noise and so, in a circuit of more than one span, quantized PAM is advantageous from a transmission point of view. With fluctuation noise the unquantized PAM audio noise would be 9 db lower than the quantizing noise and so, in a circuit of more than 9 spans of equal loss, the quantized PAM would be preferred.

FIG. 16—PCM-FM, FLUCTUATION NOISE

Here FM advantage is employed to permit operation in the presence of more noise than is possible with AM. It seems more illuminating to explain these curves by checking their correctness rather than by deriving them.

In all cases, a baseband signal-to-noise ratio giving the same margin over noise peaks as for AM (Fig. 15) is obtained by FM advantage. For the solid curves the FM limiter is assumed to be marginal (12 db radio signal-to-noise ratio), and for the dashed curves the radio signal-to-noise ratio is assumed to be the same as the marginal requirement for binary PCM-AM (18 db). The FM advantage with respect to an FM wave of the same power as in the peak AM pulse is, in db

$$20 \log \frac{\beta}{F_b} + 4.8 = 20 \log \left(\frac{B}{F_b} - 2 \right) + 4.8$$

However, the FM power is greater than the peak AM pulse power by 10

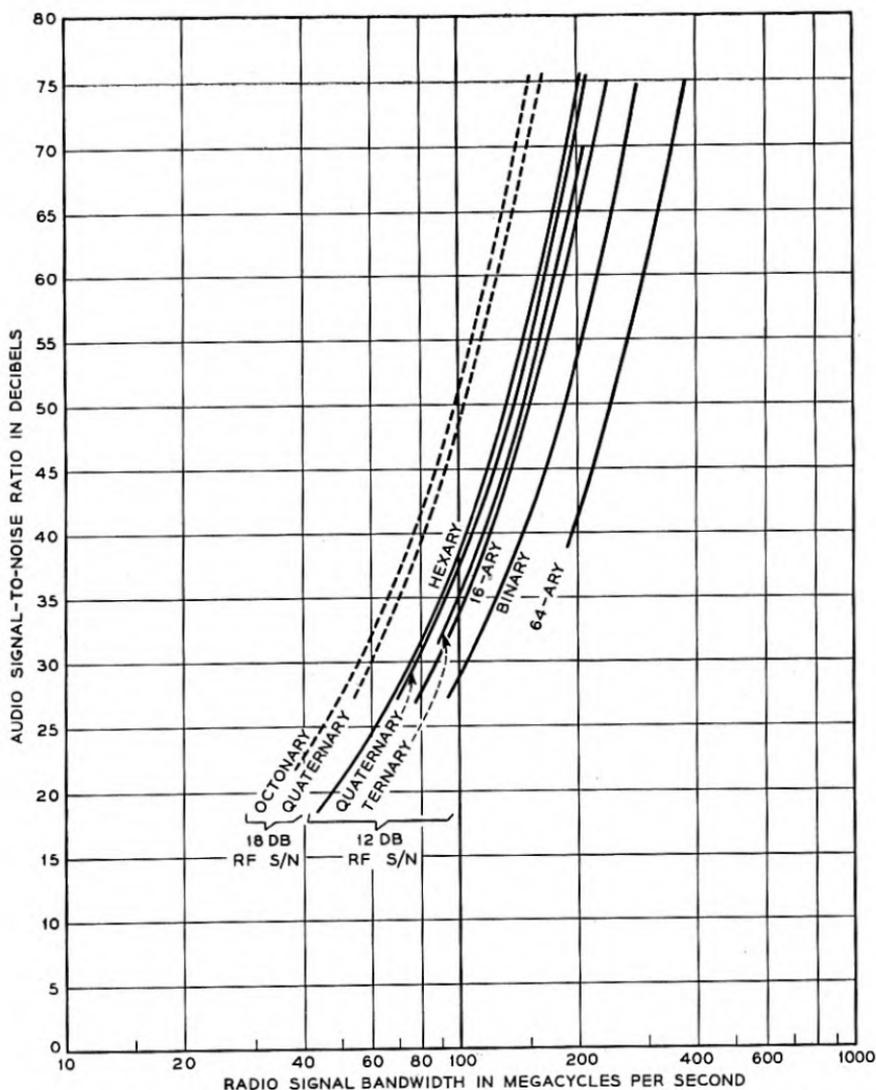


Fig. 16—PCM-FM; performance with respect to fluctuation noise. Relations between bandwidth, power, and audio signal-to-noise for 1000 4-kc channels, showing optimum bases for different ratios of FM wave to noise.

$\log \frac{B}{2F_b}$ because the FM power-to-noise ratio is maintained regardless of bandwidth. The total advantage of the FM case is therefore

$$20 \log \left(\frac{B}{F_b} - 2 \right) + 4.8 + 10 \log \frac{B}{2F_b}$$

and this must just make up for the difference between the 12 db (or 18 db) FM wave-to-noise ratio and the pulse-to-noise ratios of 18 db + 20 log ($b - 1$) required in the AM case. Substituting values from the curves will show that this is so.

These curves show a minimum bandwidth for an optimum PCM base. This is to be expected since two different rates of exchange between bandwidth and advantage are involved. One is the advantage growing out of PCM of reduced base while the other is the conventional FM advantage. An analogous situation was found in PPM-FM.

It is of interest to examine the PCM-FM situation when the FM circuit is as tolerant of noise as the most tolerant AM case, namely when the r - f signal-to-noise ratio is 18 db. The optimum PCM base is octonary and the corresponding minimum bandwidth (as we define it) is actually 20% less than for binary AM. This apparent advantage of PCM-FM is not obtained when tolerance to CW and similar systems is considered. Figure 17, which follows, shows that when allowance is made for a 9 db r - f signal-to-interference ratio (as in binary PCM-AM), the minimum FM bandwidth is greater by about 30% than for binary AM and the optimum base is ternary or quaternary. If the 3 db interference tolerance possible in FM is required, it is obtained, as shown in Fig. 17, with ternary PCM-FM, at a cost of approximately twice the bandwidth required in binary PCM-AM, which has a tolerance of 9 db. We should point out here that binary PCM transmitted by single sideband and detected by a local carrier has a tolerance of 3 db and requires half the bandwidth shown in Fig. 16. PCM-FM requires a bandwidth 3.8 times that of single sideband binary PCM for the same 3 db tolerance.

FIG. 17—PCM-FM, CW AND SIMILAR SYSTEM INTERFERENCE

In PCM, sequences of several pulses of the same amplitude may occur. The FM signal then consists of a steady frequency. A steady beat frequency persisting for several pulse periods will be produced by CW interference.²⁴ If this beat frequency is F_b the maximum interfering amplitude will be produced. The amplitude is $(Q/P) F_b$ while the step interval is $\beta/(b - 1)$. To confine the interference to a half step (with 3 db margin) requires that

$$\beta/(b - 1) \geq 2(Q/P) \sqrt{2} F_b$$

For $Q/P = 0.707$,

$$\beta \geq 2(b - 1) F_b$$

²⁴ The general solution of the problem of frequency error produced by superimposing a sine wave on an unmodulated carrier is given in Appendix II.

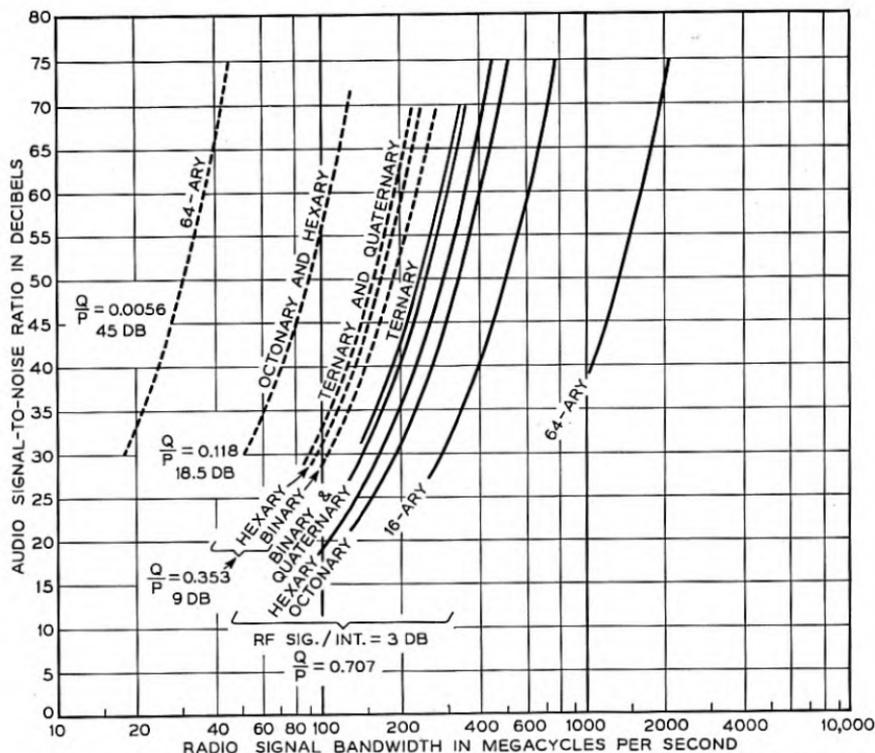


Fig. 17—PCM FM; performance with respect to CW and similar system interference for 1000 4-kc channels. Relations between bandwidth and audio signal-to-noise ratio showing optimum bases for different ratios of FM wave to interference.

and the minimum band width becomes

$$B = \beta + 2 F_b = 2bF_b$$

For $Q/P = 0.3535$ (9 db) the minimum required value of β is halved so that

$$B = (b + 1) F_b$$

For $Q/P = 0.118$ (18.5 db) the minimum required value of β is further reduced threefold so that

$$B = \frac{b + 5}{3} F_b$$

The curves of Fig. 17 are calculated from these relations.

With interference from a similar system a number of necessary conditions must be met. If the systems are similar in PCM base and in radio frequency, the sustained beat frequencies that may occur are $\beta/(b - 1)$ and

multiples thereof. The amplitude of the lowest of these frequencies is $(Q/P)\beta/(b-1)$. For $Q/P = 0.707$, this frequency must be suppressed by the baseband filter since otherwise the threshold would be exceeded. Thus,

$$\begin{aligned}\beta &\geq (b-1)F_b \\ B &= \beta + 2F_b = (b+1)F_b\end{aligned}$$

For $Q/P = 0.353$, the lowest beat frequency need not be suppressed but the $2\beta/(b-1)$ frequency must be suppressed; thus $2\beta/(b-1) \geq F_b$.

$$\begin{aligned}\beta &\geq \frac{b-1}{2}F_b \\ B &= \beta + 2F_b = \frac{b+3}{2}F_b\end{aligned}$$

Comparing these bandwidth values with those required for CW shows that the above requirements are more lenient than for the corresponding CW cases, particularly for the higher values of b where the above bandwidth values approach one-half of those obtained for CW.

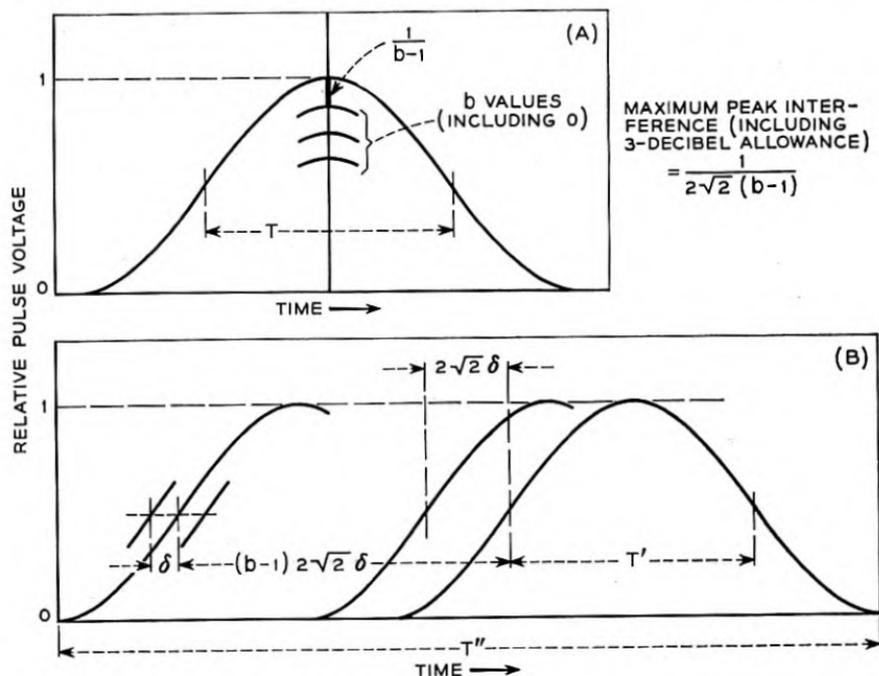
However, the above requirements are not quite sufficient. Transitions between adjacent frequency values, occurring in one system, will produce varying beat frequencies which pass through all values. This case differs from the CW case in that the beat frequency is not sustained and that the baseband filter output will not be as high as in the CW case. Calculations show that the bandwidth requirements are intermediate between those for the CW case and the similar system case considered previously.

When we remember that for low base systems (binary) the requirements for similar system and CW interference are nearly alike, while for high base systems a small frequency difference in frequency alignment can produce similar system interference completely equivalent to CW interference, we may regard Fig. 17 as applying to both, practically. Such a conclusion also makes the curves apply to interference between systems of different base.

QUANTIZED PPM

PCM pulses, including the limiting case of quantized PAM pulses, may be transmitted by time modulation instead of frequency modulation, i.e., by "quantized PPM." In this case, as in PCM-FM, bandwidth may be used to increase the tolerance to noise and interference. Figure 18 illustrates this case. At (A) is shown a PCM pulse having b values including zero, the highest amplitude being unity. The maximum tolerable peak interference is $1/2\sqrt{2}(b-1)$ and the time per pulse is taken as T . (For high base systems the time per pulse should be greater, perhaps $2T$ as pointed out in the discussion of PCM-AM). At (B) is shown the quantized PPM

pulse of the same amplitude. The time per pulse is taken to be T'' and the duration of the pulse at half height is T' . If T'' is put equal to T the ratio of the PPM bandwidth to the PCM bandwidth is T/T' . The time shift



$$\delta = \frac{\text{INTERFERENCE}}{\text{SLOPE OF PULSE}}; \quad \text{SLOPE} = \frac{\pi}{2T'}$$

CASE 1

$$\text{INTERFERENCE} = \frac{1}{2\sqrt{2}(b-1)}$$

$$\delta = \frac{2T'}{2\pi\sqrt{2}(b-1)}$$

$$T'' = T' \left(2 + \frac{2}{\pi} \right)$$

CASE 2

$$\text{INTERFERENCE} = 0.353 \text{ (9 DB)}$$

$$\delta = \frac{2T' \times 0.353}{\pi} = \frac{T'}{\sqrt{2}\pi}$$

$$T'' = T' \left[2 + \frac{2(b-1)}{\pi} \right]$$

IF WE TAKE THE TIME PER PULSE TO BE T AND T'' AND MAKE THEM EQUAL, THE BANDWIDTH RATIOS BECOME

$$2 + \frac{2}{\pi} = 2.64$$

$$2 + \frac{2(b-1)}{\pi}$$

Fig. 18—Comparison of quantized PAM with quantized PPM.

produced by peak interference is represented by δ . Our system is marginal when the smallest quantized step produces a shift of $2\sqrt{2}\delta$. The peak time shift produced by a signal is then $(b-1)2\sqrt{2}\delta$. Two cases are considered. In one the PPM system operates in the presence of the same peak interference as in the PCM case. The bandwidth ratio is 2.64. The other case

assumes that the peak interference is marginal for all bandwidths (9 db down). The bandwidth ratio is then $2\left(1 + \frac{b-1}{\pi}\right)$. It was previously found that in PCM-FM, with peak interference 9 db down, the radio bandwidth must be $(b+1)F_b$. Since the radio bandwidth of PCM-AM is $2F_b$, the bandwidth ratio is $(b+1)/2$. Comparing these bandwidth ratios we see that the PPM bandwidth required to operate in the presence of marginal interference is nearly two times that required in PCM-FM. Furthermore, this PPM bandwidth ratio applies to marginal fluctuation noise whereas in PCM-FM a more favorable result was obtained.

FIG. 19—FDM-FM, FLUCTUATION NOISE

When a group of channels in frequency division is transmitted by frequency modulation, the addition of channel voltages is translated to an addition of instantaneous frequency shift. The non-simultaneous load advantage applicable to a multichannel amplifier for frequency divided channels thus becomes an advantage in reduction of total frequency swing as compared to the sum of the individual peak frequency swings of the channels. The numerical db increments versus number of channels listed in Table I should, however, be modified for the following reason: The fluctuation noise spectrum in the output of an FM detector is not uniform with frequency, and hence the noise is unequally distributed among the channels. In order to obtain the same noise in all channels it is necessary to taper the signal levels in such a way that the full load frequency swing produced by one channel is proportional to the frequency of the channel. The frequency swing corresponding to full load in the top channel is therefore a larger part of the maximum instantaneous swing required for the group than the swings corresponding to lower channels. The result is, in effect, phase modulation. The multiplex addition factors for tapered level channels have not been determined experimentally. We have assumed here a 3 db reduction in the power capacity values listed in Table I. These reduced values then give the incremental capacity referred to *full load on the top channel*. Curves are shown for 100, 500 and 1000 channels. On account of the multiplex addition factor, it is not possible to obtain results for other numbers of channels from one curve by simply changing the frequency scale.

The derivation of these curves is straightforward but leads to an expression for the required bandwidth as a root of a cubic equation. As in the case of Fig. 16 we shall discuss the FDM-FM curves by checking them numerically. We have assumed that the channels are tapered in level and that we have, in fact, phase modulation with its consequent flat base-band noise distribution. To check the 60-db point on the 1000-channel

curve, we calculate that the noise in the entire baseband must be 43 db below the power in a sine wave which employs the full system load capacity. This figure comes from reducing the 60 db full load *channel* ratio by 30 db because of the 1000-fold greater baseband width and increasing it by the amount, $16 - 3 = 13$ db, by which the full system load must exceed full load in the top channel. Thus $60 - 30 + 13 = 43$ db. An FM advantage of 31 db must be obtained to permit the marginal *r-f* signal-

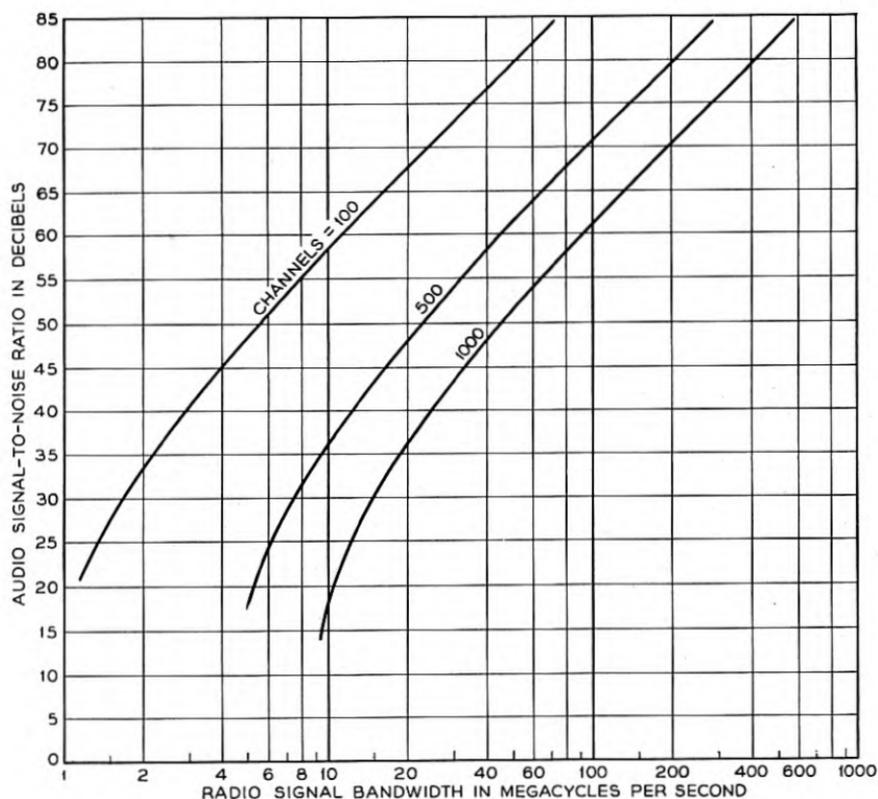


Fig. 19—FDM-FM; performance with respect to fluctuation noise. Relations between bandwidth and audio signal-to-noise ratio for marginal power; 4-kc channels.

to-noise ratio of 12 db to satisfy the above requirement, ($43 - 12 = 31$). We get this advantage in part by phase modulation gain given by $20 \log \frac{\beta}{F_b} - 6$ db. This gain is referred to 100% modulation AM whose unmodulated carrier power is the same as the FM wave power. This means that the reference is a system in which the FDM baseband appears as upper and lower sidebands which, when demodulated, yield a baseband

signal-to-noise ratio equal to the ratio of unmodulated carrier power to the noise power in the double width radio band. Since we keep the FM power marginal for all bandwidths an additional bandwidth improvement of $10 \log \frac{B}{2F_b}$ accrues. Substituting $B = 92$ mc, $F_b = 4$ mc, and $\beta = 92 - 8 = 84$ mc, will show that the above gains total 31 db.

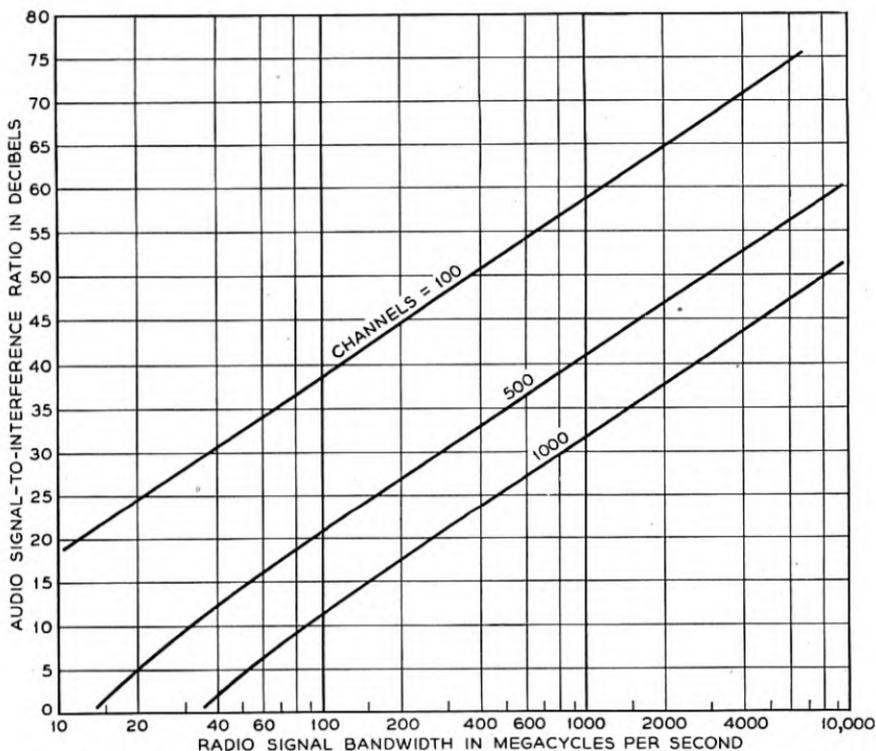


Fig. 20—FDM-FM; performance with respect to CW interference. Relations between bandwidth and audio signal-to-interference ratio for marginal ratio of FM wave to interference; 4-kc channels.

FIG. 20—FDM-FM, CW INTERFERENCE

The disturbance produced by CW is most readily evaluated when all channels are idle for then we have only the frequency error produced by a sine wave of relatively small amplitude superimposed on the steady sinusoidal carrier wave. To a first approximation (see Appendix II) the error has a frequency equal to the difference between the carrier and CW frequencies and an amplitude equal to this frequency difference multiplied by the ratio of the CW to the carrier amplitude. The error thus increases linearly with the frequency of the channel in which it falls but, since the channel levels

are also tapered in the same way, the signal-to-interference ratio is independent of the frequency of the disturbing CW. Varying the CW frequency only changes the number of the channel into which the interference falls. Loading the channels distributes the interference over several channels instead of concentrating it in one, but we have plotted in Fig. 20 the more severe case in which all channels are idle.

We have not undertaken to compute curves for similar system interference in the case of FDM-FM, but estimates for two extreme conditions can be made. In the case of low index FM systems the carrier frequency component of the spectrum is not affected by the modulating signal and the FM wave is, in fact, like an AM wave with the carrier displaced 90 degrees in phase. A similar interfering FM wave combines with the wanted FM to produce frequency or amplitude variations and does this cyclically as the r - f phase between the systems varies. When the phases are appropriate for the production of frequency variation, crosstalk appears in the wanted reception at a level lower by the ratio of FM wave amplitude. Averaging over all r - f phases should reduce the crosstalk by 3 db. The actual amount of interference received in a channel is less than would be predicted from replacement of the interfering FM wave by fluctuation noise of the same mean power spread over the r - f band, because the bulk of the interfering power is contained in the carrier component located at a frequency which does no harm. Increase of the frequency swing in both systems produces significant reduction in crosstalk when the carrier amplitude diminishes appreciably and important higher order sidebands appear, i.e. when the interfering system has its spectrum spread out more or less uniformly, like noise. Systems designed for wide swings under full load may, however, operate with only a few channels active; in such cases the low index situation may exist and the received interference will be down approximately by the ratio of the FM waves, without the benefit of FM advantage. While in this situation the bulk of the interfering power is again contained in the harmless carrier, the received interference is concentrated in a few channels and is greater than if the interfering wave power were spread, like noise, thinly over the r - f band, which in this case is many times wider than the band occupied by the low index signal. For such adverse loading conditions, the curve for similar system interference, while starting at the left above the corresponding noise curve of Fig. 19, may actually cross over and finally approach it from the lower side.

In the case of systems of very wide swing such as are involved in Table II we regard the interfering system as equivalent, under all common load conditions, to noise spread uniformly over the bandwidth and having the same power as the interfering wave. The entry in Table II is obtained by reading

the curve of Fig. 19 at 69 db audio signal-to-noise ratio which would be appropriate to yield 60 db when the "noise" is marginal at 3 db below the FM wave power instead of 12 db. A different procedure is required for the narrow band entries of Table IV. Here the emphasis on conservation of bandwidth leads to a two-frequency repeater plan with tolerance of similar system interference 44 db down. A 60 db audio signal-to-interference ratio can be met under these conditions with moderate swings for which the equivalent noise representation of the interference is not valid. The result is considerably influenced by the channel loading and we have no impeccable method of calculating the necessary bandwidth. We estimate that a bandwidth of 22.5 mc., with $\beta = 14.5$ mc., will satisfy the requirements for all except unusually adverse loading conditions.

III. BAND WIDTH AND POWER TABLES

The information contained in the curves of Fig. 9-20 has been used in preparing Tables II and III, which show what can be done with the various systems when bandwidth is used freely. The prime objective studied here is the conservation of peak transmitted power. In Table II the audio channel must meet message circuit requirements²⁵ while, in Table III, a much better grade of performance—more than sufficient for transmission of high fidelity musical programs—is stipulated. We have prepared Table III (as well as Table V) on the basis of replacing the 1000 4-kc. message channels of Table II with 250 16-kc. channels. Since we have available established load rating theory only for message circuits, we have omitted FDM and FDM-FM from Table III (and Table V). The values of Table II are based on a nominal 60 db ratio of signal-to-noise, but it is assumed permissible to meet this in the pulse systems by using 22 db of instantaneous companding so that only 38 db signal-to-noise ratio is actually required within the compandor. The PCM systems provide for a 39 db circuit within the compandor, corresponding to 6 binary digits, 3 quaternary digits, 2 octonary digits or one 64-ary digit. No allowance is made for the accumulation of quantizing noise arising when several PCM links are connected in tandem at voice frequency. In practice, 7 binary digits might be used. This would provide for several links and would permit slightly more companding. Table III assumes a 75 db signal-to-noise ratio and no companding and is referred to here as a "program" circuit. We use such a high-grade circuit to illustrate more emphatically how the system preferences depend

²⁵ We do not pretend to deal fully with the involved matter of system requirements distinguishing between kinds of noise and interference or crosstalk that appear in message channels. We merely assume that the power of the separate types of disturbances considered must be individually 60 db below that of a full load test tone under the worst specified transmission condition.

TABLE II
OPTIMUM BAND WIDTHS FOR MINIMUM POWER FOR MESSAGE TYPE CIRCUITS
133 30-mi spans, 1000 4-kc channels, 15 db NF, 75 db span loss.

SYSTEM	S/N ⁽¹⁾ IN DB	NUMBER OF RESHAPINGS: (4)						ONE EXPOSURE TO MARGINAL INTERFERENCE			MARGINAL PEAK INTERFERENCE IN DB		
		133		5(2)		1(3)		BAND WIDTH IN MC		POWER IN WATTS		CW	SIMILAR SYSTEM
		BAND WIDTH IN MC	POWER IN WATTS	BAND WIDTH IN MC	POWER IN WATTS	BAND WIDTH IN MC	POWER IN WATTS						
PPM-AM	38		10		10		10						-9
			5		5		5						
PPM-FM	38		1.0		1.0		1.0						-3
			0.5		0.5		0.5						
PAM-FM	38		1.0		1.0		1.0						-3
			0.5		0.5		0.5						
FDM-FM	60		1.0		1.0		1.0						-3
			0.5		0.5		0.5						

(1) AUDIO SIGNAL TO NOISE (OR INTERFERENCE) RATIO
(2) RESHAPING EVERY 27 SPANS

(4) REGENERATION IN THE CASE OF PCM, LIMITING IN THE CASE OF FM

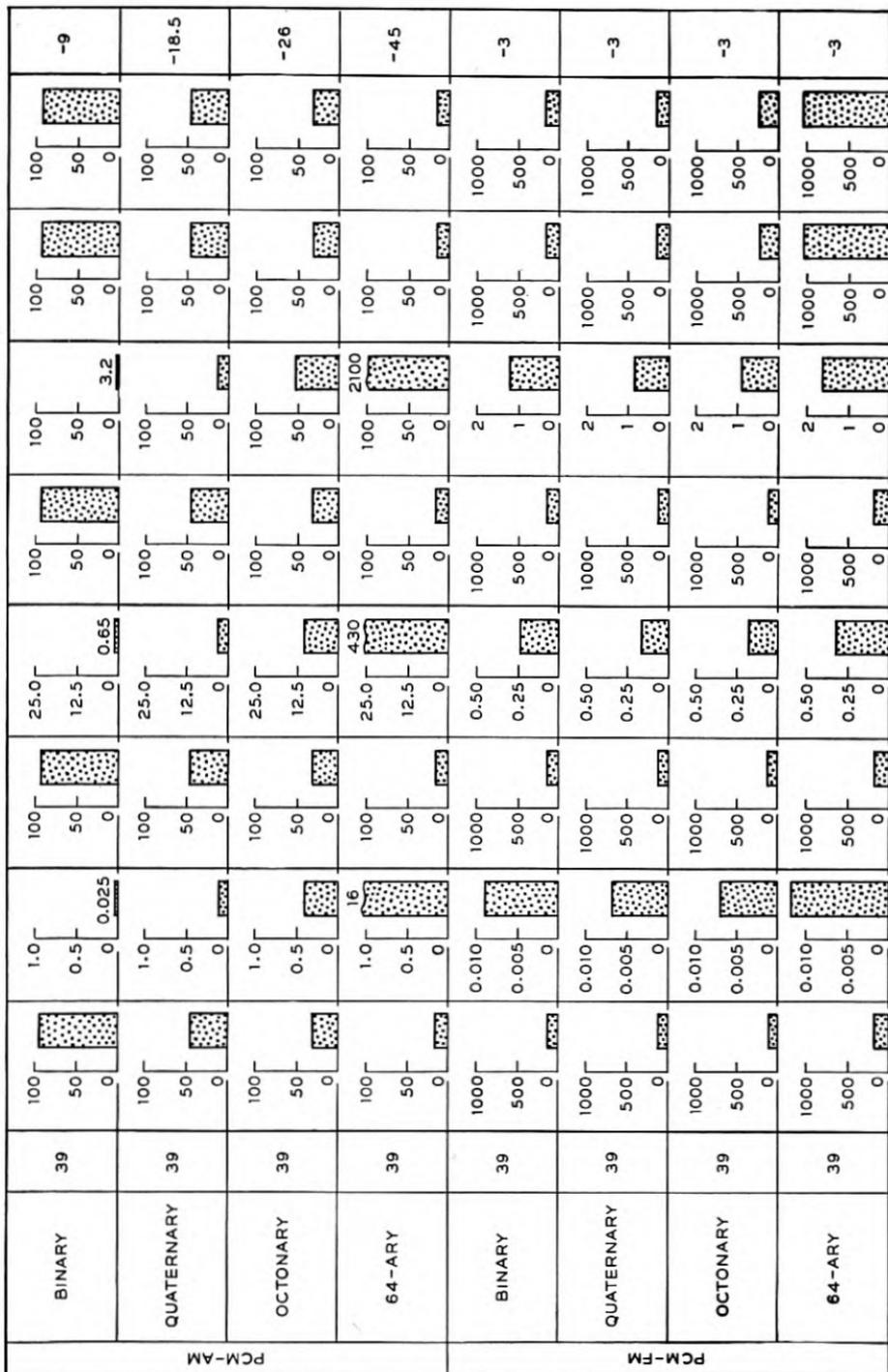


TABLE III
OPTIMUM BAND WIDTHS FOR MINIMUM POWER FOR PROGRAM TYPE CIRCUITS
133 30-mi spans, 250 16-kc channels, 15 db NF, 75 db span loss.

SYSTEM	S/N(1) IN DB	NUMBER OF RESHAPINGS: (4)						MARGINAL INTERFERENCE		MARGINAL PEAK INTERFERENCE IN DB
		133		5(2)		1(3)		CW	SIMILAR SYSTEM	
		BAND WIDTH IN MC	POWER IN WATTS	BAND WIDTH IN MC	POWER IN WATTS	BAND WIDTH IN MC	POWER IN WATTS			
PPM-AM	75						-9			
PPM-FM	75						-3			
PAM-FM	75						-3			

(1) AUDIO SIGNAL TO NOISE (OR INTERFERENCE) RATIO

(2) RESHAPING EVERY 27 SPANS

(3) RESHAPING EVERY 27 SPANS

(4) REGENERATION IN THE CASE OF PCM LIMITING IN
THE CASE OF FM

(5)

upon signal-to-noise ratio. Both tables apply to long systems comprising 133 repeaters spaced 30 miles apart. The span loss is 75 db. Three reshaping or regeneration plans are shown: reshaping at every repeater, reshaping five times in the complete system, and no reshaping within the system.

In the case of PCM, reshaping becomes a true regeneration which completely removes noise accumulated in the transmission link; while, in the non-quantized pulse systems, reshaping restores the original pulse shape but retains timing or amplitude errors. In the case of FM systems, reshaping is accomplished by the amplitude limiter which removes envelope variations arising from noise, but does not suppress the accompanying frequency shifts. Reshaping, in contrast to regeneration, is only a partial prevention of cumulative effects but, as shown by the tables, it has definite value in enabling the use of wider transmission bands with corresponding smaller amounts of power than permissible without reshaping. Regeneration completely removes errors in both the amplitude and in the time. The maximum bandwidth which can be used is independent of the number of regenerations, but the signal power must be increased in proportion to the number of spans covered before regeneration.

In the case of non-regenerative radio transmission systems, we may regard the 75 db span loss as 60 db free space loss plus a fading allowance of 15 db. This allowance is intended to cover the increment of noise caused by fading in the entire system. Available data on distribution of fading are too meager to permit generalization, but indicate that on some routes, at least, the total degradation suffered would rarely be worse than that produced by 15 db simultaneous fades on all spans; and hence a design based on 75 db span loss should satisfy the noise requirements except for an extremely small fraction of the time. In other words a few spans of the non-regenerative system may fade deeply but, in regard to total accumulated noise, the system is credited with the higher signal-to-noise ratio occurring on spans which are not simultaneously in fading minima. In regenerative PCM systems no credit accrues from a higher signal-to-noise ratio occurring between points of regeneration, and protection must be provided against the worst condition that is likely to occur in any section included between regeneration points.

However, the values of minimum power obtained by lavish use of bandwidth exhibited in Tables II and III may be particularly significant for wave guide transmission systems; and hence the assumption that all spans have the same loss is appropriate here.

The outstanding features of Tables II and III are the extremely small amounts of power needed in the PCM systems with only moderate expenditures of bandwidth as compared with the non-quantized system. These

results illustrate the properties of PCM as a means of converting bandwidth into transmission advantage.

The PCM-FM entries are taken from the curves of Fig. 16 plotted for noise 12 db down. Curves are also given in Fig. 16 for noise 18 db down. It will be noted that the bandwidths indicated become smaller when the noise is required to be farther down, but that the power requirements become greater because the bandwidth reduction factor is less than the factor multiplying the r - f signal-to-noise ratio.

The columns at the right in Tables II and III show the bandwidths which must be employed in order to attain a 60 db signal-to-interference ratio in the presence of one source of interference whose amplitude is just marginal for the type of system concerned.

Tables IV and V are prepared from another point of view—that of conserving bandwidth²⁶ instead of power. The systems particularly suited for narrow bands such as FDM and PAM-AM have been added to the list. The actual minimum bandwidths are, in many cases, determined by engineering judgment; smaller values than those tabulated may be possible at the expense of greatly increased power and precision requirements. Thus in the case of PPM-AM we have arbitrarily chosen 40 mc as necessary for 1000 4-kc channels. According to our initial postulate the audio signal-to-noise ratio vanishes at 32 mc, and indefinitely great signal power would be required as we approach this limit. In PAM-AM we have assumed that pulses in adjacent channels just touch, thereby setting the bandwidth at 32 mc. Smaller bandwidths could be used if the pulses were allowed to overlap. This would reduce the allowable duration of the channel gate and deprive the system of some of its tolerance to similar system interference as well as noise. The maximum pulse power required for 100% modulation is tabulated. If instantaneous sampling were used this would be 6 db above the unmodulated pulse power which is, in turn, 38 db above the mean total fluctuation power accumulated in a 32-mc band from 133 spans. We have reduced the value of power thus computed by 1.7 db to allow for a calculated improvement in signal-to-noise ratio obtainable by gating at the channel input with a time function of the same shape as the signal pulse.

The FM systems listed are of two kinds. The first is a relatively narrow-band type in which advantages such as relative immunity to gain fluctuation and amplitude non-linearity are sought with small increment in bandwidth over AM. Since these objectives are not sufficient in themselves to fix the actual bandwidth needed, an arbitrary additional requirement has

²⁶ We do not here entertain the idea of using certain exchange methods to permit use of less band width than the conventional minimum of 4 kc per channel, but rather to use modest amounts of additional band width. Appendix III discusses briefly a band reduction principle.

TABLE IV
 MINIMUM BAND WIDTHS AND CORRESPONDING POWER REQUIREMENTS FOR MESSAGE TYPE
 CIRCUITS

133 30-mi spans, 1000 4-kc channels, 15 db NF.

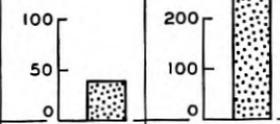
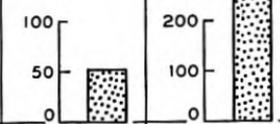
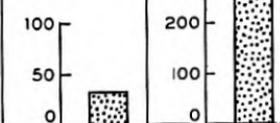
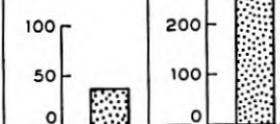
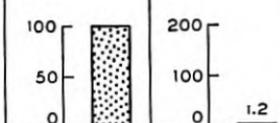
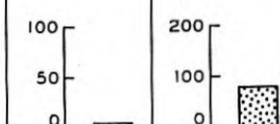
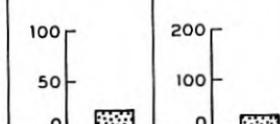
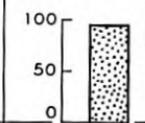
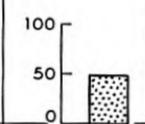
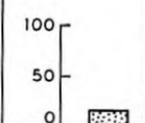
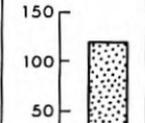
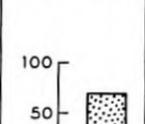
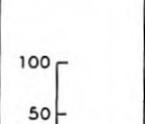
SYSTEM	S/N IN DB	BAND WIDTH IN MC	POWER IN WATTS	TOLERABLE INTERFERENCE RATIO IN DB		SPAN LOSS IN DB	
				CW	SIMILAR SYSTEM		
PPM-AM	38		850	46	41 ⁽¹⁾	75	
PPM-FM	38		850	49	39 ⁽¹⁾	75	
PAM-AM	38		850	44	40 ⁽¹⁾	75	
PAM-FM	NARROW BAND	38		850	46	23 ⁽²⁾	75
	WIDE BAND	38		1.2	20	9 ⁽²⁾	75
FDM	60		850	76	60	75	
FDM-FM	60		850	68	44	75	

TABLE IV—Concluded

SYSTEM	S/N IN DB	BAND WIDTH IN MC	POWER IN WATTS	TOLERABLE INTERFERENCE RATIO IN DB		SPAN LOSS IN DB
				CW	SIMILAR SYSTEM	
PCM-AM	BINARY	39		9	9	85
	QUATERNARY	39		18.5	18.5	85
	64-ARY	39		45	45	85
PCM-FM	QUATERNARY	39		9	9	85
	OCTONARY	39		18.5	18.5	85
	64-ARY	39		45	45	85

REGENERATE AT EVERY REPEATER IN PCM

- (1) INTERFERING SYSTEM IDLE; ACTIVITY HAS SMALL EFFECT.
 (2) ASSUMING 12.5% CHANNEL ACTIVITY.

TABLE V
MINIMUM BAND WIDTHS AND CORRESPONDING POWER REQUIREMENTS FOR PROGRAM TYPE
CIRCUITS

133 30-mi spans, 250 16-kc channels, 15 db NF.

SYSTEM	S/N IN DB	BAND WIDTH IN MC	POWER IN WATTS	TOLERABLE INTERFERENCE RATIO IN DB		SPAN LOSS IN DB	
				CW	SIMILAR SYSTEM		
PPM-AM	75			83	78 ⁽¹⁾	75	
PPM-FM	75			86	76 ⁽¹⁾	75	
PAM-AM	75			81	77 ⁽¹⁾	75	
PAM-FM	NARROW BAND	75			83	60 ⁽²⁾	75
	WIDE BAND	75			25	9 ⁽²⁾	75
PCM-AM	BINARY	75			9	9	85
	QUATERNARY	75			18.5	18.5	85
PCM-FM	QUATERNARY	75			9	9	85
	OCTONARY	75			18.5	18.5	85

REGENERATE AT EVERY REPEATER IN PCM

(1) INTERFERING SYSTEM IDLE; ACTIVITY HAS SMALL EFFECT.

(2) ASSUMING 12.5% CHANNEL ACTIVITY.

been imposed that the transmitted power in the FM system should be equal to that required in the corresponding AM system. The resulting values of bandwidth are found to be reasonable ones for low index FM. An exception is made in the case of FDM-FM systems of message type where it is found that a net economy of frequency occupancy can be obtained by increasing the frequency swing sufficiently to tolerate similar system interference 44 db down. This enables the two-frequency repeater plan of Fig. 4, as discussed in Section I, to be used and substantially reduces the frequency occupancy over that of a lower index FM system more vulnerable to antenna crosstalk and therefore requiring more frequency assignments. We have estimated that a radio signal bandwidth of 22.5 mc achieves the required 44 db tolerance. The second type of FM is a wide-band system designed for specified tolerances of interference from similar systems. Data on this second type will be used later in our study of inter-route interference, where it will appear that ruggedness is a more important criterion of frequency occupancy than the minimum bandwidth needed for transmission.

No curves have been furnished to determine the FDM entries, since there is no variation with radio bandwidth to consider. The band required is merely the number of channels multiplied by the width of a channel. The power required for message channels is determined by calculating the amount of power in one channel to give a 60 db margin over mean fluctuation noise power in a 4-kc band and applying the multiplex addition factor of Table I. Similar system interference is simply linear crosstalk and must be, we say, 60 db down. CW interference referred to maximum system power must be down an additional amount equal to the multiplex addition factor of Table I in order to meet 60 db suppression in the disturbed channel. Since the two-frequency plan of Fig. 4 does not suppress interference between the two directions of a single route by 60 db, we must use twice as many frequency assignments as there shown. This duplication will appear in Table VI. FDM is the only system of Table IV for which such duplication is necessary, since the others do not require more than 44 db suppression. In the program type systems of Table V, however, the first four listed would need duplicated frequency assignments.

The PCM-AM systems of course do not use any smaller bandwidths than those given in Tables II and III and would, therefore, be expected to show disadvantageously in a bandwidth comparison with the other systems. On the other hand they make, relatively, a good showing in power requirements and in tolerance to CW and similar system interference.

In the next section we shall show that economy of bandwidth may, in fact, be illusory because of the greater susceptibility to intra- and inter-system interference associated with narrow band methods. It is not the bandwidth actually needed for transmission that is important, but the

tightness with which the bands may be packed into the frequency range without mutual interference. Because of the ruggedness of low-base PCM, neighboring frequency bands can actually be allowed to overlap. Introduction of the proper spacing factors for satisfactory separation in frequency between adjacent bands causes the PCM system to overtake the other methods in effective utilization of frequency space, especially when intersecting routes are involved.

The PCM-FM systems listed in Tables IV and V are of the second class listed above, in which equivalent ruggedness rather than equivalent power as compared to AM is the criterion. Thus, binary PCM-AM is compared with a PCM-FM system having the same 9-db tolerance of interference. The curves of Fig. 17 show that, with such a tolerance, the minimum PCM-FM bandwidth is secured when the base is either three or four. We choose the quaternary case here because the signal-to-noise ratios obtainable coincide with those of the binary. Likewise, either octonary or hexary PCM-FM furnishes the optimum base for the 18.5 db tolerance possessed by quaternary PCM-AM and, of the two, octonary is more suitable for our tabulation. In determining the power required to override fluctuation noise in a PCM-FM system designed for a specified tolerance of similar system interference, we must make sure that both the limiter and slicer are protected against breaking.

The values of repeater power capacity shown in Tables IV and V will satisfy the noise requirements on a 133-span non-regenerated circuit with 75 db loss on all spans. For spans of 60 db free space loss the tabulated power thus provides for 15 db fades simultaneously on all spans, or for 13 db simultaneous fades of 25 db, or for a single fade of 36 db. PCM systems employing regeneration on every span must be powered for the deepest fade that is likely to be encountered. We have arbitrarily taken this to be 25 db making the span loss 85 db. This is probably not a sufficient allowance for some situations but will serve for illustrative purposes. If regeneration were not practiced the power would be $25 - 15 = 10$ db lower from the fading allowance standpoint but would have to be increased $10 \log 133 = 21$ db for noise accumulation, so regeneration results in a power saving of 11 db. If, with regeneration, we were to protect each span against the deepest single fade (36 db) permitted by the power provided for non-regenerative operation, the power advantage of regeneration would disappear ($36 - 15 - 21 = 0$ db).

In general, when the power without regeneration protects against simultaneous fades upwards of just a few db there is little or no power advantage in regeneration if we then protect each span against the deepest single fade permitted when regeneration is not practiced. This is true even for large numbers of spans. There remain, however, important advantages for

regeneration in preventing accumulation of disturbances that are not much affected by the distribution of fading.

IV. FREQUENCY OCCUPANCY TABLES FOR RADIO RELAY

The frequency space occupancy for a single two-way route is, according to principles laid down in the introduction, a frequency block $2U$ times the signal bandwidth.²⁷ Our problem, as stated in the introduction, is to examine the situations arising when a number of 1000-channel routes converge toward a terminal city, assuming all of the routes to be of the same kind. We will determine the number of times the above frequency blocks must be repeated in the spectrum in order to keep interference within tolerable bounds. The sum of these blocks then really defines the frequency occupancy and determines the space which must be allocated or, conversely, determines the number of routes a given allocation will accommodate. We will use the tolerable ratios of similar system interference taken from Tables IV and V, together with appropriate antenna directivity, to determine the number of these blocks.

ANTENNA CHARACTERISTICS

The directional discrimination afforded by the antennas is obviously an important factor in frequency economy. For our present study, we employ an antenna having a directional pattern slightly superior to that of the 4000-mc shielded lens antenna in use on the New York-Boston radio relay circuit. Figure 21 shows the assumed directional characteristic omitting "nulls" between the minor lobes. Of importance also are the nearby discrimination characteristics of the antennas as given in Fig. 4.

The situations arising at a point where a number of routes converge (or cross) or where a route is equipped with a spur connection are variations of that occurring at a single repeater point. In fact, the situation in which two routes converge from approximately opposite directions occurs at every repeater point in a straight route, while a repeater point at which the route bends sharply is like a terminal point at which two routes converge at a small angle.

The crosstalk in our assumed two-frequency long distance repeater system has been estimated in Section I (under "The Radio Repeater") and was found equivalent to a single source of similar system interference 44 db down. A system which possesses just enough tolerance to withstand the accumulated crosstalk on a long straight repeater system is not capable of meeting another such system at an angle unless additional frequencies are

²⁷ In the case of very tender systems, such as FDM, the factor $2U$ is replaced by $4U$ because a four-frequency plan is needed for a two-way repeater. U is the band spacing factor discussed under "The Radio Repeater."

invoked. The FDM-FM system of "minimum band width" listed in Table IV is intended to possess this 44 db tolerance. To illustrate the requirement of new frequencies let us take the case of several FDM-FM routes

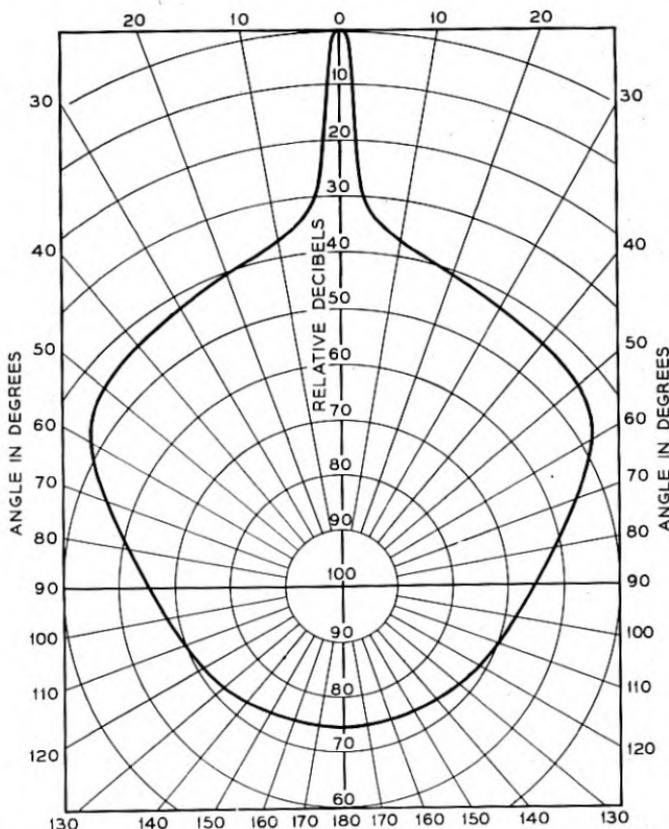


Fig. 21—Assumed directional selectivity of 10 ft. x 10 ft. antenna at 4000 mc.

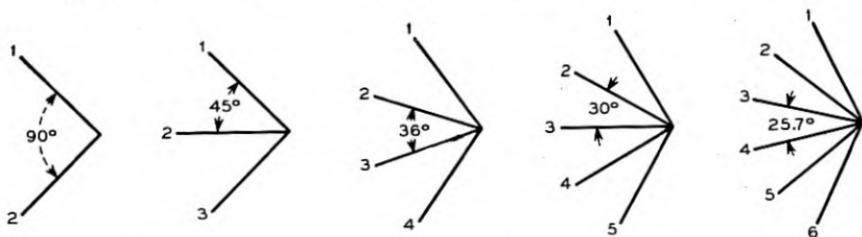


Fig. 22—Simplified route pattern for study of selectivity required in congested localities.

converging in the manner suggested by any of the diagrams of Fig. 22. A critical situation occurs in regard to the frequencies used for receiving at the point toward which the systems converge. If the same receiving

frequency were to be used on two or more routes, receiving directional discrimination amounting to 75 db would have to be secured:

1. Required interference ratio	44 db from Table IV FDM-FM
2. Allowance for repeater crosstalk	1 (51 db down)
3. Differential fading allowance ²⁸	30
	—
	75 db

The repeater crosstalk is here taken to be equivalent to one source 51 db down which is the value corresponding to no differential fading on adjacent spans as calculated in Section I. It will be remembered that allowance was made for a single differential fade of 30 db occurring somewhere along the route. Here we assume that this differential fade may occur between two of the converging paths and we demand that the receiving directional discrimination shall protect the system against such an occurrence. In this case the required directional discrimination turns out to be equal to the 75 db front-back ratio from which the 44 db figure was obtained. This is manifestly impossible with the assumed directivity characteristic²⁹ and the angles involved. Therefore, different receiving frequencies are required on each route. These same frequencies may be used for transmitting at the junction, provided the disposition of terminals is such as to provide enough directional discrimination and physical separation to permit operation at the low received level in the face of the high transmitted level on the same frequency. The interference path loss plus antenna discrimination must be, for the case involving the longest span:

1. Required interference ratio	44 db (FDM-FM, Table IV)
2. Allowance for repeater crosstalk	1
3. Free space span loss	60
4. Fading allowance	25
	—
	130 db

We continue our discussion of the converging routes of Fig. 22 by assuming that:

1. Conditions encountered elsewhere on the routes do not restrict the freedom to switch the frequencies among the routes.
2. The disposition of terminals at the junction is such that inter-terminal interference is not a controlling factor.

Under the above assumptions the directional discrimination of the terminal

²⁸ This differential fading allowance corresponds to a fade of 25 db below free space on one route and a 5 db increase over free space on the other.

²⁹ The use of perpendicular polarizations cannot, we assume, be counted on to give further discrimination when the directional discrimination is already 40 or more decibels.

antennas alone determines the number of different frequencies. The same receiving frequency may be used at the junction on two routes separated by an angle sufficient to yield the required antenna discrimination. All intervening routes must employ different frequencies. The frequencies so determined may be used for transmitting from the junction if they are staggered with respect to the receiving frequencies. Take, for instance, the five-route plan shown in Fig. 22. Suppose the directional discrimination needs to be 60 db for a particular system. The directional pattern shows that this requirement is met at 85 degrees. Thus, routes 1 and 4 may use frequency A, say. Routes 2 and 3 then must use different frequencies, B and C. Thus, we have:

Route	1	2	3	4	5
Trans. Freq.	A	B	C	A	B
Rec. Freq.	B	C	A	B	C
or	(C	A	B	C	A)

While the treatment of the route congestion problem outlined above is oversimplified it enables us to make a broad survey having some significance.

Table VI for 1000 4-kc message channels and Table VII for 250 16-kc "program" channels were derived on the above basis. The decibel figures at the head of each column are the allowable interference ratios from Tables IV and V increased by 30 db for differential fading.³⁰ A single source of interference of the values given in the table is supposed to degrade the circuit to the minimum requirements for a long circuit. In regenerative PCM there is no accumulation of degradation due to interference occurring on various spans. In non-quantized systems such degradations are cumulative. However, when protection to the above values is provided, with no allowance of 30 db for differential fading of the desired and interfering signals, the occurrence of *simultaneous* additional degradations is extremely unlikely. Protection against this severe fading at one point alternately protects against the simultaneous occurrence of several less severe fades.

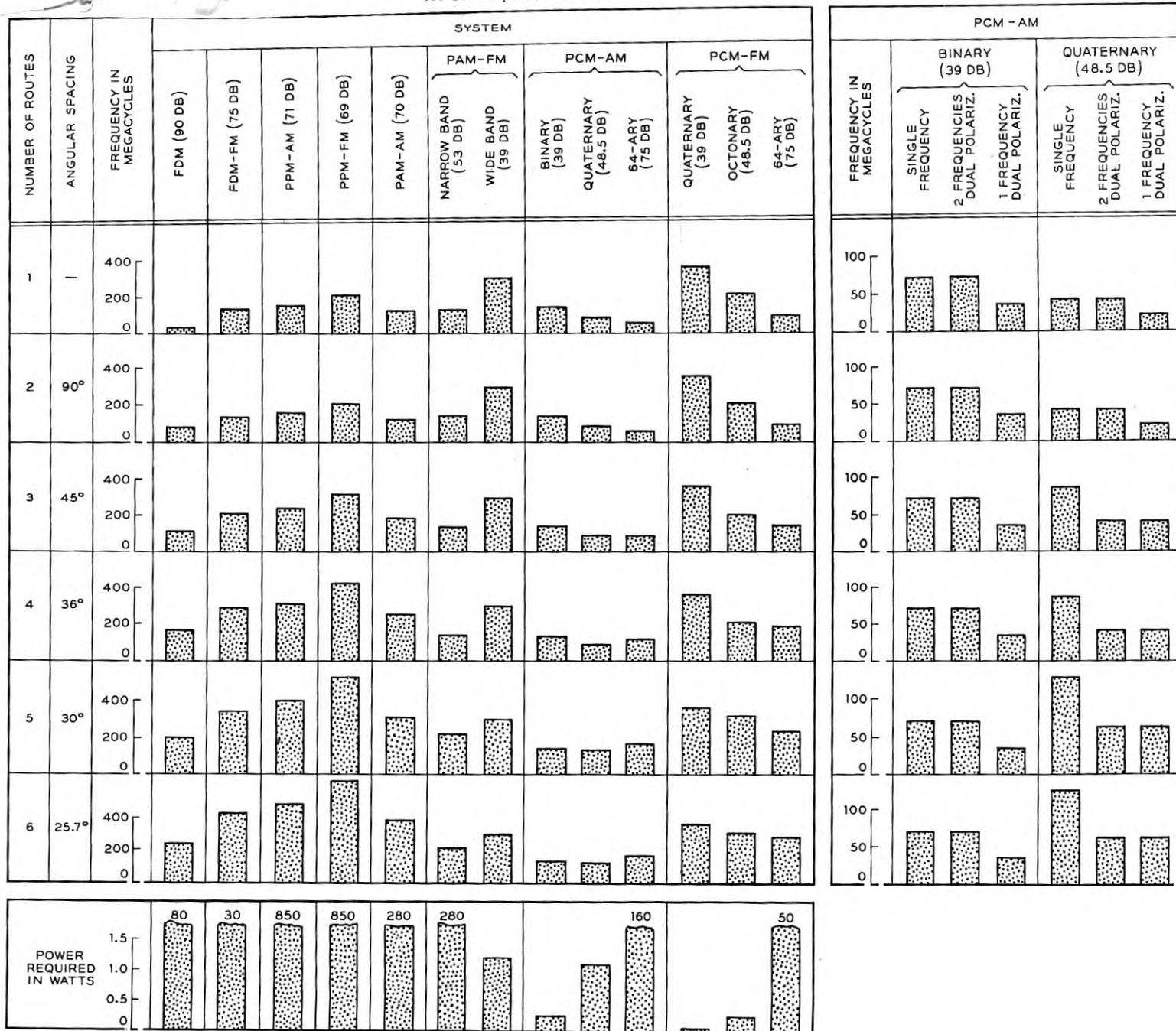
The values of repeater power capacity shown in the table will satisfy the noise requirements on a 133-span transcontinental nonregenerated circuit with 15 db fades simultaneously on all spans. This is equivalent to providing for 13 simultaneous fades of 25 db or, statistically, for the fading that is not likely to be exceeded except during a small fraction of the time. The PCM systems employing regeneration are powered for 25 db fades on any or all spans.³¹ The free space loss is 60 db.

In computing the frequency occupancy we take cognizance of the fact

³⁰ The FDM-FM entry provides 1 db additional allowance for repeater crosstalk as mentioned before. Repeater crosstalk is negligible in the other systems.

³¹ No such distinction was made in Tables II and III. There, provision was made for 75 db span loss in all cases.

TABLE VI
TRUE FREQUENCY OCCUPANCY OF VARIOUS MESSAGE GRADE RADIO RELAY SYSTEMS FOR CONGESTED ROUTES
 133 30-mi spans, 1000 4-kc channels, 15 db NF.



that systems of different vulnerability require different guard space to protect against adjacent-band crosstalk. In Section VII it is concluded that with binary PCM-AM the band spacing could perhaps be as small as $1.5/T_0$ with realizable filters and gates. T_0 is the time allotted to one pulse. Underlying the meaning of B as used heretofore in connection with PCM-AM, is the relation $T_0 = 2/B$. If the band spacing is to be taken as $1.5/T_0$, the band spacing may be expressed in terms of B as $0.75 B$. In other words, the band spacing factor U may be reduced to 0.75.

In the case of ideal *FM* systems the receiving frequency discrimination need not suppress adjacent radio signal bands to anywhere near the degree required of co-channel interference, provided the near edges of the adjacent signal bands differ by more than the width of the baseband filter. We do not, of course, assume ideal apparatus and have been rather liberal in guard space allowance.

The following band spacing factors, to be used with the bandwidths of Tables IV and V, are considered realizable and consistent with the shape of the spectrum to be transmitted. A reduction to practice would likely lead to somewhat different factors but these will suffice for our illustrative Tables VI and VII. In the non-regenerative systems, the spacing factor required to protect against interference from new frequencies required at a junction is, perhaps, less than is required to protect against interference at every repeater on a long route. A small economy in occupancy could properly be invoked on this account in some cases but, in the interest of simplicity, this has been neglected, and the same factor U will be associated with every frequency required.

SYSTEM OF TABLE VI	FACTOR U
FDM	2.5
FDM-FM	3
PPM-AM	2
PPM-FM	2
PAM-AM	2
PAM-FM (narrow band)	2
PAM-FM (wide band)	1.5
PCM-AM (64-ary)	2
PCM-AM (quaternary)	0.9
PCM-AM (binary) ¹²	0.75
PCM-FM (64-ary)	2
PCM-FM (octonary)	1.5
PCM-FM (quaternary)	1.5

¹² The experimental system described by Meacham and Peterson (loc. cit.) employs a spacing factor of 1.12.

TABLE VII
TRUE FREQUENCY OCCUPANCY OF VARIOUS PROGRAM TYPE RADIO RELAY SYSTEMS FOR
CONGESTED ROUTES

133 30-mi spans, 250 16-kc channels, 15 db NF.

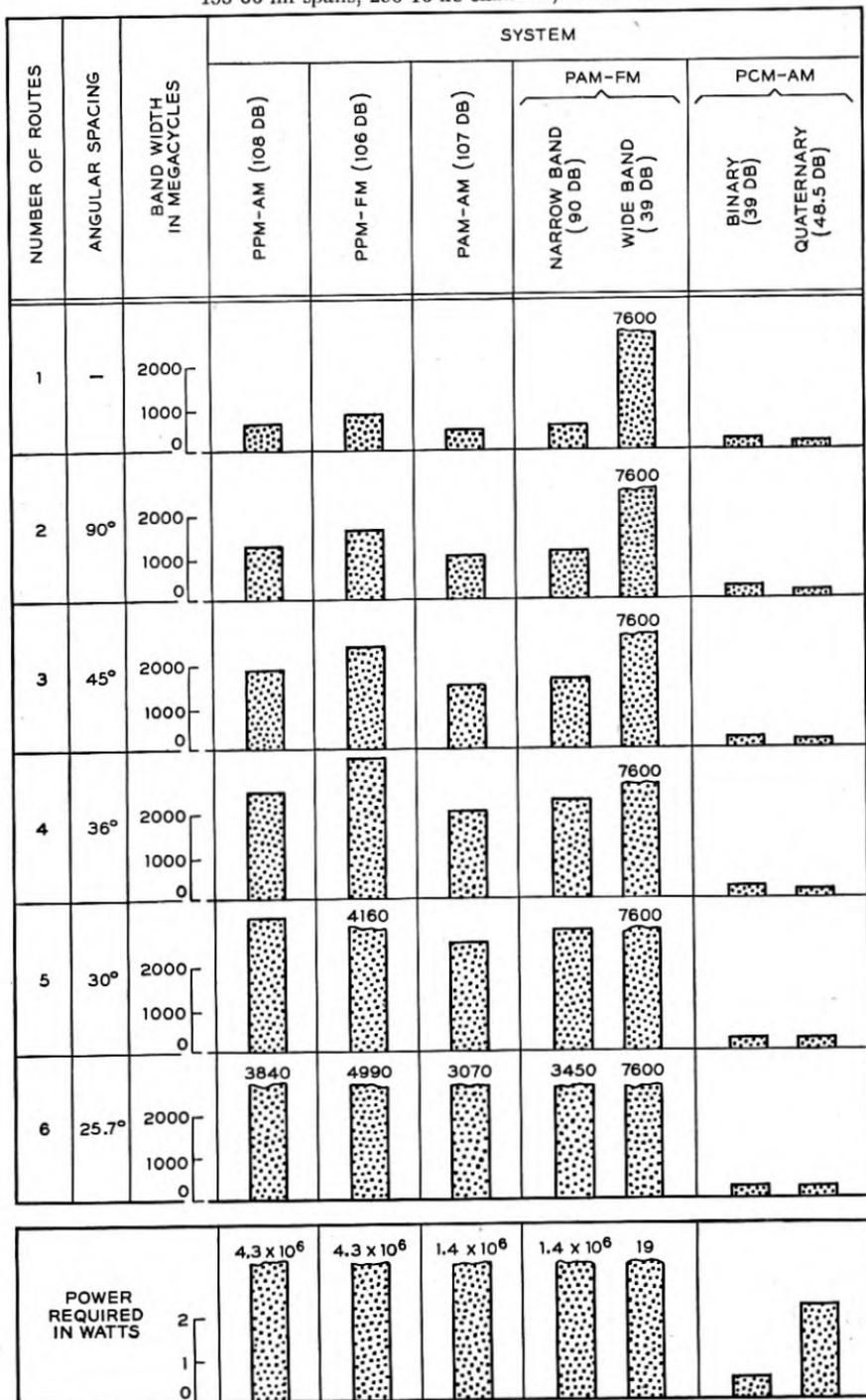


TABLE VII—Concluded

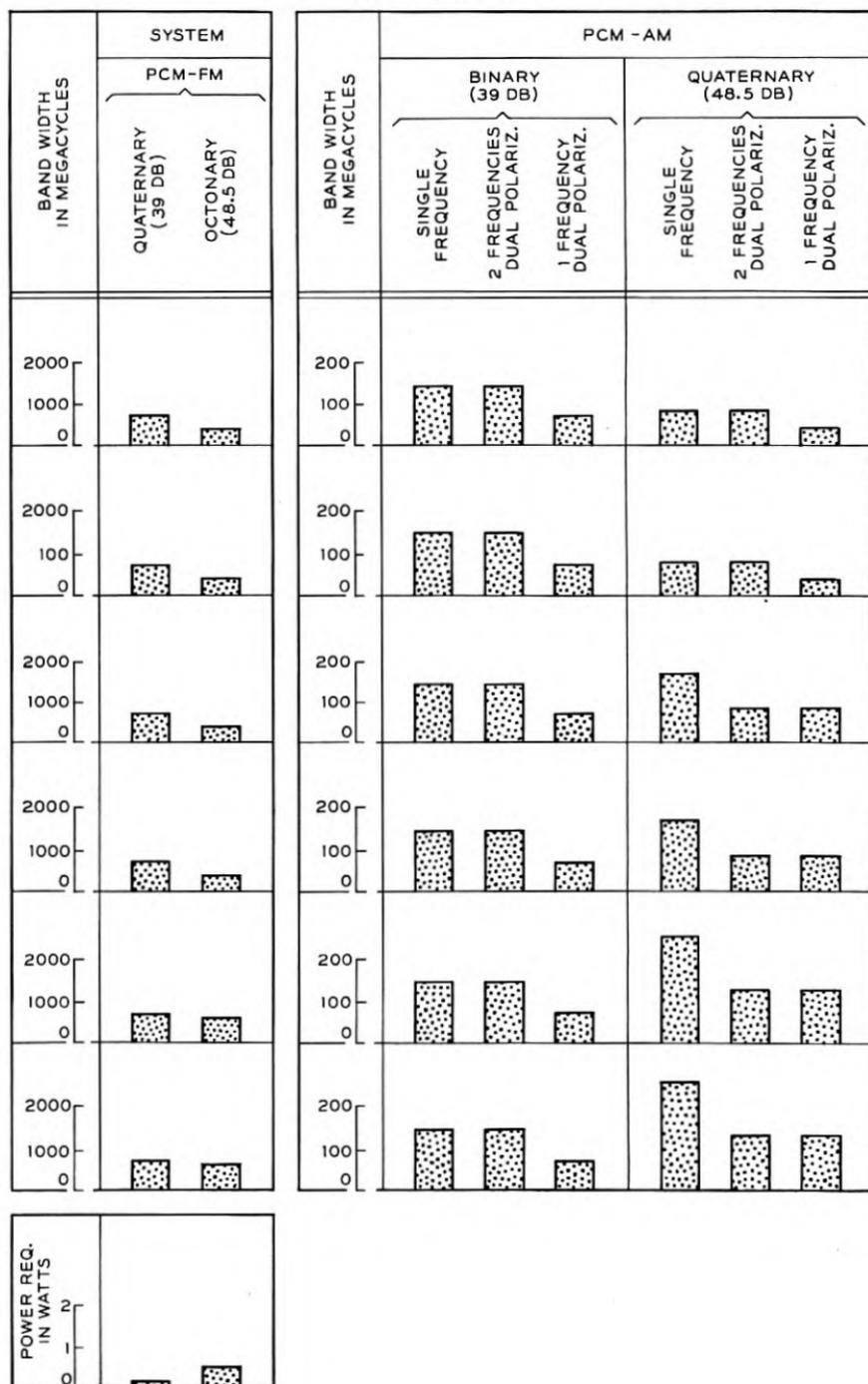


TABLE VIII
 COMPARISONS OF BAND WIDTH AND FREQUENCY OCCUPANCY FOR SYSTEMS OF EQUAL ROUGHNESS
 Dotted bars show bandwidth; crosshatched bars show relative frequency occupancy.

SIMILAR SYSTEM INTERFERENCE RATIO IN DB	BAND WIDTH IN MC	1000 MESSAGE-TYPE CHANNELS						250 PROGRAM-TYPE CHANNELS			
		PCM-AM	PCM-FM	PPM-AM	PPM-FM	PAM-FM (12.5% ACTIVITY)	FDM-FM**	PCM-AM	PCM-FM		
18.5		0.9	1.5	1.3	1.5	1.5	1.5	0.9	1.5		
		OCCUPANCY RATIO, U								1.5	1.5
		0.75	1.5	1.1	1.5	1.5	1.5	0.75	1.5		
9		0.75	1.5	1.1	1.5	1.5	1.5	0.75	1.5		
		OCCUPANCY RATIO, U								1.5	1.5
		0.75	1.5	1.1	1.5	1.5	1.5	0.75	1.5		
3		0.75	1.5	*	1.5	1.5	1.5	*	1.5		
		OCCUPANCY RATIO, U								1.5	1.5
		0.75	1.5	1.1	1.5	1.5	1.5	0.75	1.5		

* SUPPRESSED CARRIER, HOMODYNE DETECTION

** INTERFERENCE CALCULATED AS NOISE OF SAME POWER

These factors were multiplied by the product of bandwidth and number of frequencies to obtain the dotted bars in Table VI.

In regard to the "program grade" of circuit we must be more liberal in our allowance for guard space. Our estimates for the band spacing factor are:

SYSTEM OF TABLE VII	FACTOR U
PPM-AM	4
PPM-FM	4
PAM-AM	4
PAM-FM (narrow band)	4
PAM-FM (wide band)	3
PCM-AM (quaternary)	0.9
PCM-AM (binary)	0.75
PCM-FM (octonary)	1.5
PCM-FM (quaternary)	1.5

These factors were used to compute the dotted bars in Table VII.

If transmission on two polarizations can be accomplished with mutual cross-fire suppressed to a sufficient degree, half of the channels could be transmitted by each polarization, on the same frequency, thus halving the frequency occupancy. A probably unattainable cross-fire ratio seems necessary to meet the requirements in the non-regenerative systems, if we remember that the interference produced by cross-fire accumulates from span to span; but a suppression likely to be attainable, of the order of 15-20 db, makes this frequency saving feasible in the rugged systems such as binary PCM-AM or PCM-FM. The tables show entries for binary and quaternary PCM-AM, assuming dual polarization transmission.

If antennas could be improved to insure nearby discrimination ratios adequate to allow use of the same frequency in and out and west and east, the single-route occupancy would be halved again; with such a one-frequency repeater plan the occupancy in a congested area is not, however, always halved. Whenever the frequency requirements, as determined by the terminal antenna directivity, result in two or more frequencies, A, B, C . . . etc., there is no saving accruing from a one-frequency repeater plan, because two-frequency routes can be accommodated with no additional frequencies by suitably switching frequencies. It is only in the case of a system so rugged that the terminal antenna directivity permits a single frequency, A, to be used that the occupancy is reduced and it is then halved. With PCM of low base this is a possibility and the tables include entries for this case.

As to achieving antenna characteristics suitable for one-frequency operation, it may be noted that reflection from a heavy rainfall in front of the

antennas limits the attainable side-to-side ratio.³² Reflection from aircraft may also impose a practical limitation. Spacing the antennas laterally (on two towers) would achieve freedom from these limitations. Another way of coping with the antenna discrimination obstacle is to use short spans in congested areas. This reduces the discrimination requirements particularly because fading is reduced by shortening the spans.

CONCLUSIONS AS TO RADIO

Of the systems included in Table VI we find that, for six routes, binary PCM-AM, even without the potential frequency economy of dual polarization and/or single-frequency repeaters, has come close to being the most economical of frequency space; quaternary PCM-AM shows a slight advantage (which would be lost if the route spacing were less than fifteen degrees). Even without dual polarization or single-frequency repeaters, the binary PCM-AM occupancy is less, for more than 3 routes, than the *occupancy* required by FDM whose *band width* is 4 kc per channel. There is here an excellent illustration of the possibility of a net saving in frequency space through the use of tough wide-band systems.

The power requirements also favor the low-base PCM systems. It should be noted, in particular, that the linearity requirements with FDM demand that the tabulated power of 80 watts be a very light load on the repeaters.

Inspection of Table VII brings out the effectiveness of the coding principle if very high-grade channels are required. Only with PCM (of low base, as shown) are the occupancy and power requirements both within the practical realm. The non-PCM methods that achieve small occupancy, comparable with that of low-base PCM, all require colossal amounts of power. When the power requirement is reduced and the ruggedness increased by use of band width, the occupancy becomes, in turn, colossal. This is illustrated by the two entries for PAM-FM.

As route congestion increases without limit, any type of system that permits exchange between bandwidth and ruggedness will always achieve the minimum occupancy when bandwidth has been used to secure the degree of ruggedness that avoids multiplying the frequency assignments. Our studies have shown that, with the assumptions made, this result is valid for channels of message grade when the congestion has reached a degree that is by no means fantastic. We have accordingly prepared Table VIII in which the dotted bars show the bandwidths (taken from Fig. 9-19) of the various systems when their interference tolerances are alike and have values of 18.5, 9, and 3 db.³³ While these systems, having the same tolerance, all

³² Measurements made at the BTL radio laboratory at Holmdel, N. J. indicate that this limit to side-to-side ratio is of the order of 85 to 90 db.

³³ The AM pulse systems are here assumed to achieve the 6 db increase in tolerance by suppressing the carrier.

fare alike in respect to frequency requirements imposed by antenna directivity, the bandwidth figures do not adequately reflect the merits of the systems. This is because the band-spacing factors are different and, in addition, only the regenerative systems can be expected to achieve the halving of occupancy accruing from dual polarization and from one-frequency routes. The crosshatched bars of Table VIII include the effect of multiplying by the estimated band-spacing factors shown beneath the bars. These band spacing factors are in some cases smaller than those previously tabulated for the less rugged systems of Tables IV and V. Only the PCM methods are shown for the case of very high-grade channels, since the non-PCM methods are so strikingly less effective here.

These conclusions depend for validity on the assumptions made and particularly on those concerning antennas, route disposition and fading, and apply when the converging systems *are of the same kind*. In a real situation, departures from the assumed conditions could markedly affect the conclusions. For instance, the meritorious showing of PCM in respect to efficient utilization of frequency space in the face of route congestion depends heavily on the assumption that all routes in the occupied space employ PCM. Any routes employing a modulation method that is highly vulnerable to interference like some of the narrower bandwidth methods would have to employ higher power to operate in the face of interference from the PCM routes. This higher power, concentrated in a narrower band, could destroy the PCM routes. In some cases it would obviously be impossible to assign values of power which permit the two kinds of routes to share the same frequency band.

Our calculations should be taken to illustrate the factors involved and the philosophy by which such problems may be approached rather than to find an unequivocally best system.

V. MORE ABOUT THE NON-SIMULTANEOUS LOAD ADVANTAGE

The transmission advantage enjoyed by multiplexing many single sideband telephone channels in frequency division, discussed in the introduction, stems from several factors:

1. During the busiest period, only a small percentage (of the order of 12 to 15%) of the channels are actually transmitting speech ("talk spurts") at one time, on the average.
2. There are only a few *loud* talkers; the remaining ones range downward to a volume 35 to 40 db lower.
3. In the addition of the sideband voltages representing the talkers actually producing talk spurts, only a fraction of the grand maximum occurs often enough to be significant.

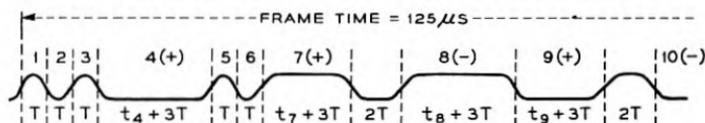
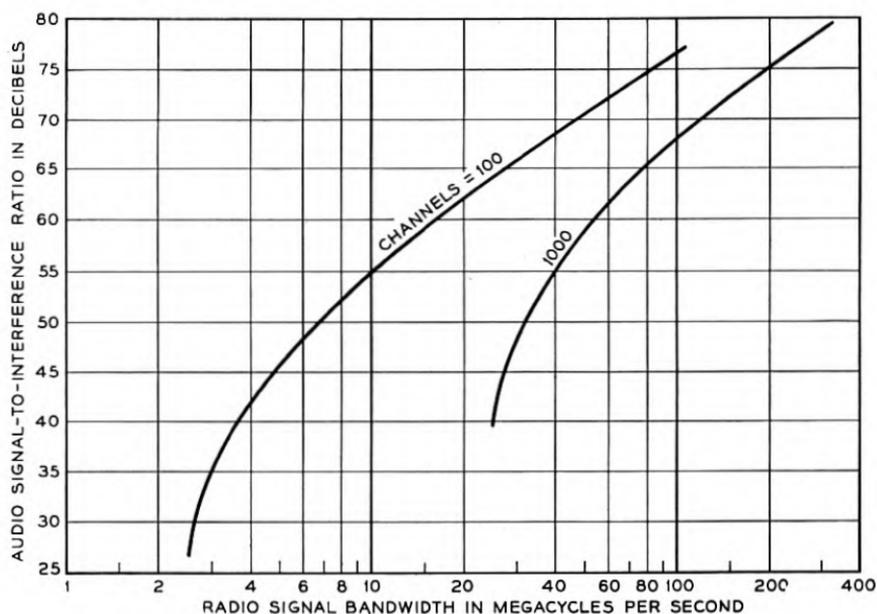
With frequency division all of these factors jointly contribute in a natural

and automatic manner to the low peak load ratings given in Table I. In time division, complicated instrumentation is needed to obtain such a low load rating (in time, now, not power capacity) and the saving is in bandwidth (time). Savings accruing from item (1) above are theoretically obtainable in all time-division systems (and, in fact, in nonmultiplexed multipair cable transmission systems) by having automatic devices which skip the channels that are momentarily inactive and which advise the receiver of the skipping. It is possible also to benefit from items (2) and (3) above in systems which transmit a time interval to represent an amplitude. The amplitudes may be sent as absolute magnitudes together with a polarity indication. If this is done the channel time allotments actually required in a given multiplex frame appear piled up end to end, and many more channels can be handled than if provision were made for full amplitude on all. PPM is one such system, and PCM is another if the code symbols containing fewest digits are used to represent the smallest absolute magnitudes.

The use of instantaneous companding, which tends to make all talkers contribute equally to the system load, reduces the advantage represented by (2) above, but does not basically affect (1) which represents a substantial part of the total multiplex advantage.

It is illuminating to compute the performance of a pulse length modulation system (PLM) employing the elastic time allotment and assuming that the load ratings of Table I apply. We imagine a system working on the principles illustrated in Fig. 23. There we assign a time $T (= 2/B)$ to each inactive channel. Active channels whose absolute amplitudes are described by t , are assigned $t + 3T$ and those that are negative are preceded by a $2T$ pulse to designate that they are negative. If the interference is no greater than marginal (9 db down) the receiver can distinguish between (a), the T intervals which count off the channels that are skipped and (b), the $2T$ polarity indications and (c), the $3T$ minimum signal intervals. The frame time of 125 microseconds must include the sum of these intervals plus Kt_0 where t_0 is the time shift for a full-load tone in a single channel which gives the required signal-to-noise ratio for the bandwidth $B (= 2/T)$. The load rating factor is K , expressed as an amplitude ratio. The relations used to plot the two curves of Fig. 23 are shown in the insert. Little or no instantaneous companding could be used to advantage so that a signal-to-interference ratio of 50 to 60 db would be required and for 1000 channels the bandwidth would be between 30 and 50 mc, which is some two or three-fold less than in binary PCM-AM, both systems being equally tolerant to a single source of CW interference. The elastic principle could presumably be applied to PCM also to achieve a several-fold bandwidth reduction, but no experience has been obtained with any of these elastic systems. While this paper has avoided for the most part questions of instru-

mentation, it should be pointed out that the elastic schemes tend to become complex apparatus-wise. If one chooses to discount this on the grounds



N = NUMBER OF CHANNELS

a = NUMBER OF ACTIVE CHANNELS

$$125 = N \left(1 - \frac{a}{N} \right) T + N \frac{a}{N} 3T + \frac{1}{2} N \frac{a}{N} 2T + t_0 K = \frac{2N}{B} \left(1 + 3 \frac{a}{N} \right) + t_0 K$$

$$t_0 = \frac{125 - \frac{N}{B} \left(2 + \frac{6a}{N} \right)}{K}$$

$$\text{SLICER ADVANTAGE} = 20 \log_{10} \frac{\pi}{2} \frac{t_0}{T} - 3 \text{ DB}$$

$$\frac{S}{I} = \text{RF PULSE TO INTERFERENCE RATIO} + 20 \log_{10} \frac{\pi}{2} \frac{t_0}{T} - 3$$

CALCULATED FOR: RF RATIO = 9 DB (MARGINAL)

$N = 1000$ $K = 6.3$ (16 DB)

$N = 100$ $K = 2.83$ (9 DB)

$$\frac{a}{N} = \frac{1}{8}$$

Fig. 23—Theoretical possibilities of exploiting non-simultaneous load advantage by an elastic PLM-AM system.

that future developments may resolve the complexity, there remains the objection that any system designed to take advantage of the multiplex load rating counts heavily on being used almost exclusively for conversational

speech under present operating procedures. The extensive use of telephone channels for nontelephone purposes is thus curtailed.

VI. OVERLOAD DISTORTION AND NOISE THRESHOLD

In designing a microwave system for a large number of channels the power required to override noise may exceed the power capacity of available amplifiers. Also, the bandwidth may exceed the limit imposed by microwave transmission phenomena or circuit techniques. In either case, the remedy is to divide the channels into several groups of fewer channels and transmit the groups in adjacent narrower bands spaced by the proper factor U , and separable with filters for individual amplification, reshaping or regeneration. The power requirement falls off linearly with bandwidth. The filter problem for AM pulse transmission is considered in Section VII. The total frequency occupancy is no greater for this division since the same percentage "guard band" is involved if, in both cases, the neighboring, foreign signals are of the same kind as the wanted signals. In case the neighboring signals are of a different kind, the multiple band arrangement is in fact likely to represent a smaller occupancy because the occupancy is in general more sharply defined when made up of several narrower bands.

When considering a multiple group arrangement, it may be economical to provide for a substantial amount of common amplification prior to separation into the several bands which receive individual treatment. The non-linearity of the common amplifier then sets a limit to the common amplification. Experiments bearing on this overload limit were made with the PCM equipment described by Meacham and Peterson.¹² Two- and eight-frequency groups were employed and the amplifier load was increased until the effects of distortion began to appear. The distortion was measured in terms of the maximum amount of CW interference which, when added to the amplifier output, resulted in no audible effect in the PCM channels. The right-hand part of Fig. 24 plots the results. For eight bands (six of which were not pulsed but were left on as unmodulated carriers) it is seen that the margin provided against CW interference begins to shrink rapidly when the single group load is 20 db below the output at which 1 db compression occurs. The margin is completely used up (the channels begin to show noise) when the load is 13 db higher. The left-hand part of Fig. 24 plots the manner in which the low level limitation (noise) was found to appear. Margin against CW interference shows a reduction for a pulse-to-noise ratio of 28 db and is completely used up at a ratio of 18 db.

The overload occurred in the 4000-mc power amplifier associated with the repeater, and the noise originated in the receiver.

In non-reshaping amplitude-modulated systems, the effect of compression

¹² Loc. cit.

occurring in the repeaters is cumulative. In microwave repeaters second-order distortion products fall outside the band and third-order distortion is likely to be predominant. We assume in what follows that the distortion arises solely from a cubic term. When the low-level gains of the repeaters are maintained equal to the preceding span losses, it can be shown that the single-frequency compression characteristic at the end of n spans is approxi-

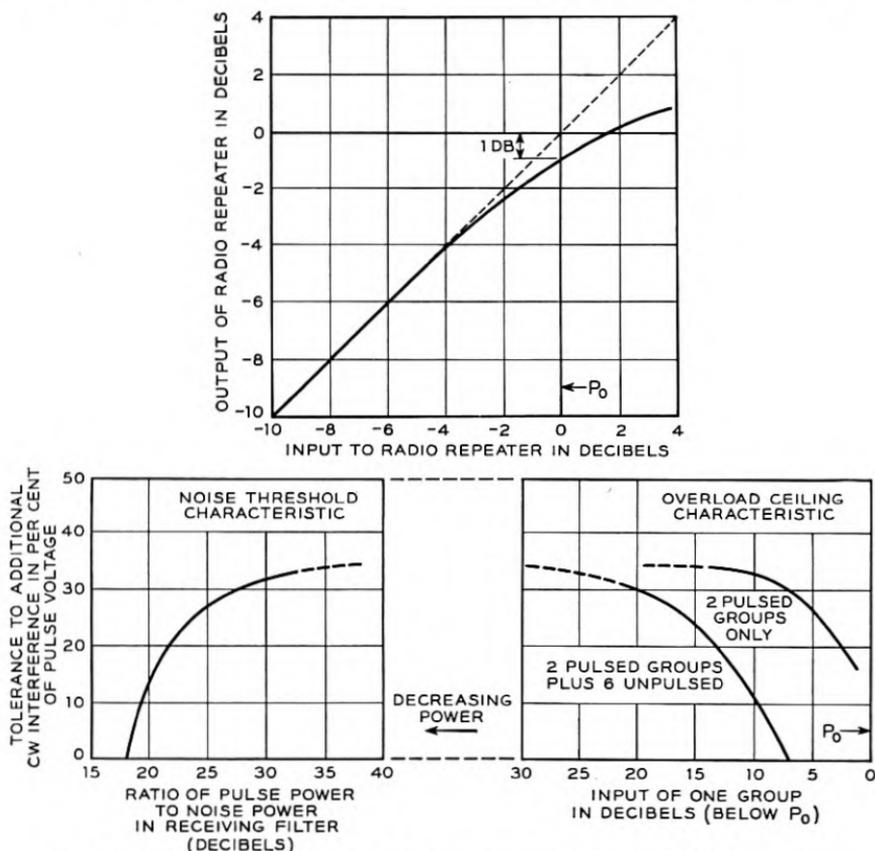


Fig. 24—Noise threshold and overload ceiling in frequency divided PCM groups.

mately the same as for one span but occurs at a power level $10 \log n$ db lower. This approximation becomes more exact as the over-all distortion involved becomes less (as by lower input power). Fig. 25 shows a third order compression curve for one span and the resulting curves, obtained graphically, for 2, 4 and 10 spans. Examination of these curves shows that the curves are substantially the same as for the single span but displaced, 3, 6 and 10 db respectively. This is illustrated by line A, which intersects all of the curves at the same compression value (1.7 db). The points

marking the intersections are seen to be displaced from the intersection with the curve for one span by approximately 3, 6 and 10 db. If the phase of the repeaters is as linear as it must be in pulse systems, this single frequency characteristic can be applied for the entire signal band as if it resulted from a single source of third-order distortion. The effect of this distortion is

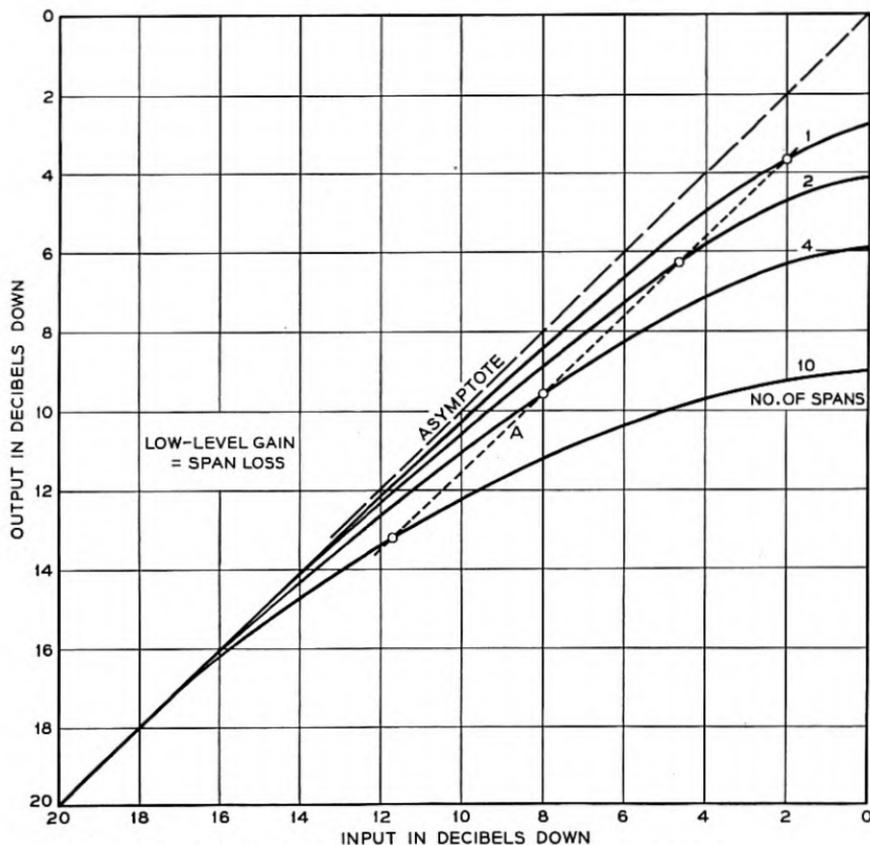


Fig. 25—Overload characteristics of multi-repeater systems.

serious in multiband PCM repeaters, as illustrated by the measurements on Fig. 24, but is, generally speaking, less important in single band pulse repeaters. For instance³⁴ PPM-AM pulses and binary PCM-AM pulses might operate on the flat part of curve 10 (Fig. 25). With PCM of higher

³⁴ On the grounds that pulse slicers themselves include the compression function to a high degree, one might not see the harm of compression in repeaters. If all of the compression occurred *after* the noise had been acquired there would be no fundamental compression penalty in slicing pulse systems. The penalty comes about because as the pulses progress from span to span they shrink and become more vulnerable to noise.

base as well as with PAM, the repeater loading would have to be sharply reduced, however.

More power on all spans could be obtained by making the repeater gain greater than the span loss. This very quickly defeats its purpose, however, because the excess low-level gain raises low-level noise between pulses to a high level status as it progresses from span to span.

Reshaping of AM pulses (and of course regeneration in PCM-AM) at all repeaters avoids the cumulative effect of compression by permitting the repeater gain to be greater than the span loss by the amount of compression on one span.

When a signal is transmitted by FM, the phase curve of the transmission circuit plays a role somewhat analogous to the amplitude characteristic of an AM system. The correspondence is not complete, however, for we find that modulation products arising from even-order phase distortion as well as from odd fall in the signal band even though the FM band is located in a frequency range very high compared with the baseband width. For amplitude modulated signals in the baseband, we can replace the FM phase distortion effects by an equivalent non-linear baseband amplifier characteristic which has the same shape with respect to zero voltage input as the phase characteristic has with respect to the midband frequency of the FM range. If the distortion is small, the square and cube law approximations obtained by expanding the phase-shift function about the mid frequency may be applied as in conventional multichannel cross-modulation theory.³⁵ We shall not here attempt to discuss the accumulation of phase distortion in a multi-repeater FM system.

VII. PULSES, SPECTRA, AND FILTERS

In this section, we will consider: (1) pulse shapes in relation to the particular pulse modulation method employing them, (2) the shaping filters by which they may be obtained and (3) the transmitting and receiving filters employed in systems comprising a multiplicity of adjacent frequency bands each carrying pulse signals.

Column A of Fig. 26 shows various pulse shapes which can be approximated (with the exception of shapes 8 and 9) by fairly simple circuits, both in the baseband and radio spectrum. Pulse 1 is an "unshaped" rectangular pulse. A good approximation to it can be obtained in wide-band circuits accommodating the extensive spectrum it possesses, i.e., in circuits having rise and decay times short compared with the duration T_0 . Such a pulse when transmitted through Gaussian filters of the various widths shown in

³⁵ W. R. Bennett, "Cross-Modulation in Multichannel Amplifiers" *Bell System Technical Journal*, Vol. 19, pp. 587-610, October, 1940.

column C emerges with smooth transitions as shown in 2, 3 and 4. These pulses rise and fall in a nearly sinusoidal manner. The width between half-amplitude points is T_0 . Shortening the rectangular pulse ("curbing") and narrowing the shaping filter can be made to result in pulses 5 and 6 which have the same width between, say, 3% points (at t_1) as pulse 4.

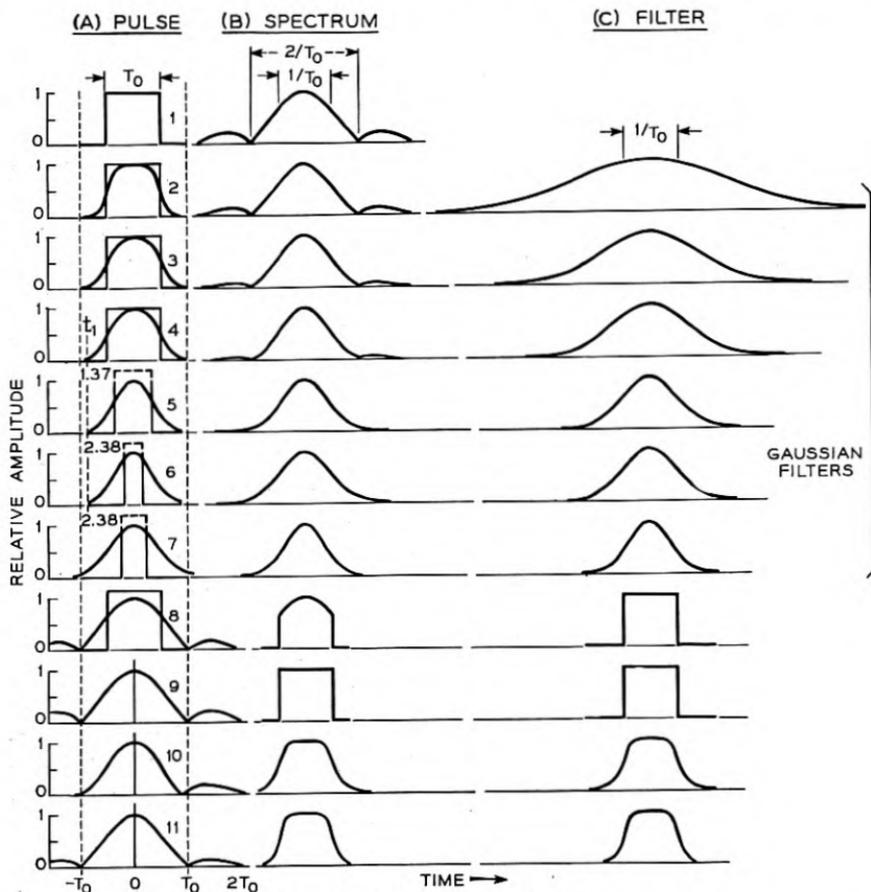


Fig. 26—Typical pulses and their spectra.

Pulses 5 and 6 are then shorter than 4 between half-amplitude points. If the half-amplitude width is made the same as in pulse 4 the width between lower amplitude points is greater than in pulse 4. This is illustrated in pulse 7.

Gaussian filters³⁶ as defined here are naturally linear phase networks

³⁶ Gaussian filters are networks whose transfer admittance follows the error law as a function of frequency. A decibel plot of a bandpass Gaussian filter is accordingly a parabola in shape.

and we have assumed linear phase in computing pulses 2 to 7. A good approximation to the Gaussian filter can be obtained both as to phase and amplitude with a number of tuned circuits in tandem, coupled through buffers. A fair approximation can also be obtained by combining a 3- or 4-section maximally flat filter³⁷ with a tuned circuit through a buffer.

Rectangular or near-rectangular shaping filters produce pulses with overshoot as shown by pulses 8 to 11. The filter corresponding to pulses 8 and 9 is assumed to have rectangular shape and linear phase. Filters of this sort have no simple approximation in practice and are included for comparison with filters 10 and 11 which are made up of simple maximally flat networks. In pulses 9, 10 and 11 the "unshaped pulse" is assumed to be very narrow and of amplitude sufficient to yield pulses of the heights shown.

Let us now regard these pulses as received pulses and compare them in respect to shape for use in various kinds of pulse systems.

PPM. In PPM the pulses may occupy any time position in the assigned interval and so the tails of pulses 8 to 11 may "crosstalk" into time assigned to an adjacent channel. To allow guard time for the train of tails or to design for satisfactory operation in the presence of the tails is uneconomical of frequency space. It follows that pulses which are more definitely bounded in time such as those obtained with Gaussian filters are more desirable and likely to be more economical of frequency space in general despite their wider spectrum.

In PPM where the trailing (or leading) edge of a pulse is used to convey the information a flat top pulse such as pulse 2 is no better than one in which the flat portion is absent and the two transitions brought together.³⁸ The latter pulse would, in fact, be superior since more time would then be available for additional channels or for greater swing.

We are thus led to conclude that one of the pulses in the 4 to 6 group is the preferred shape for PPM. We chose pulse 4 in our illustrative calculations and defined bandwidth as $2/T_0$, but pulses 5 or 6 would have given substantially the same results.

PAM. In PAM the pulses occur at standardized, regular times so that if pulse 9 were used the accompanying tails, which disappear completely at instants T_0 , $2T_0$, etc., from the pulse peak, need not theoretically produce crosstalk between channels if the channels are spaced T_0 and the pulse amplitudes are measured *instantaneously* at the time the nulls occur. As a practical matter both the precise pulse shape and the instantaneous measure-

³⁷ W. W. Mumford, "Maximally-Flat Filters in Wave Guide," *Bell Sys. Tech. Jl.*, Vol. 27, October, 1948, pp. 684-713.

³⁸ Such a pulse would look like pulse 4 if the latter were shrunk to occupy 0.6 of the time shown in the plot. The spectrum would accordingly be that of pulse 4 expanded by the factor 1.7 but would not include more significant band width than is necessary to form pulse 2 as shown. This deduction follows from the fact that the rise time of pulse 2 is the same with or without the flat top.

ments at precise instants are probably not realizable to a degree which would keep the crosstalk within tolerable limits, so that one of the smooth pulse shapes is preferred. Pulse 4 with a spacing of T_0 is feasible from the sampling precision point of view but a spacing of $2T_0$ provides margin against crosstalk arising from small imperfections in any realizable approximation to the theoretical pulse.

It is to be noted that, if an instantaneous sample is taken of a PAM pulse, the measured magnitude is affected directly by the instantaneous value of noise present in the entire band occupied by the pulse. No frequency selectivity can be applied afterward to remove the influence of any part of the noise band because the error, even though caused by wide-band components, is exactly the same as could have been produced by a uniquely determined wave wholly confined to the signal band itself. The best signal-to-noise ratio obtainable with instantaneous sampling is that associated with minimum bandwidth for the pulse (i.e., pulse 9) and the corresponding maximum stringency of synchronization requirements on the sampling and pulse distortion. The same signal-to-noise ratio can, however, be approached with a wider band provided that we allow a finite segment of the received pulse to enter the channel filter. An averaging out of higher-frequency disturbances produced by wide-band noise is thus attained.

PCM. In PCM a short sample taken near the center of a pulse serves to determine correctly the presence or absence of a pulse even in the presence of interference at or near the breaking point of the slicer. Thus, pulse 4 may be used with a spacing of T_0 , and if a gate pulse 25% of T_0 is used, it need not be aligned with an inordinate precision to obtain good operation.³⁹ Greater tolerance in the matter of sampling would be obtained with pulse 2 but the frequency extravagance could scarcely be countenanced. As stated we assume pulse 4 in our PCM bandwidth curves but employ pulse 11 in Tables VI-VIII. Use of pulse 11 is a frequency conservation measure that seems feasible only with PCM and is attractive only with binary PCM.

OPTIMUM DISTRIBUTION OF SELECTIVITY BETWEEN TRANSMITTING AND RECEIVING FILTERS

In a regenerative repeater system both the receiving and transmitting filters may be Gaussian without suffering cumulative narrowing of the system bandwidth since each span commences with a freshly shaped pulse. In this case, the transmitting filter of one repeater and the receiving filter of the succeeding repeater combine, as Gaussian filters do, to make another Gaussian filter. The resulting pulse may be one of the series 2 to 6 of Fig. 26. On the assumption that one of these shapes is desired and that the trans-

³⁹ This is the pulse shape approximated in the experimental system described by Meacham and Peterson (loc. cit.).

mitting and receiving filters are to be Gaussian, a problem arises as to how to divide the total selectivity (in column C) between them. If most of the pulse shaping is done at the transmitter, the Gaussian receiving filter must be extremely broad, with the result that discrimination between pulses in adjacent bands is poor and the bands must be spaced widely in order to keep cross-fire down. If, on the other hand, all of the shaping is done at the receiver the wide spectrum of the unshaped transmitted pulse spills over into neighboring bands unless the bands are widely spaced. Clearly, an optimum proportioning of selectivity exists and it is interesting and enlightening to analyze this problem. Such an analysis was made for pulses 4, 5 and 6. This analysis pertains only to crossfire and not to signal-to-noise ratio as influenced by curbing (shortening of the rectangular pulses) and by the division of selectivity between transmitting and receiving filters. Wide receiving filters accept more noise and narrow ones may prevent the transmitted pulse from attaining full height in the receiving filter output if curbing is used. If the curbing is pronounced, as in pulses 6 to 11, amplification may have to follow the transmitting filter to establish the desired transmitted power level. For divisions of selectivity close to the optimum, the receiving filter selectivity appreciably reduces the transmitted pulse height in the case of pulse 5 and seriously reduces it in the case of pulse 6.

Crossfire from a pulse in an unwanted band appears as a transient in the wanted band. In some circumstances, this transient has peaks which occur while the crossfiring pulse is rising and falling and has a minimum between which sometimes dips below the level fixed with the steady-state discrimination to the crossfiring carrier. If the pulses in the crossfiring band are synchronized with those in the wanted band as they might be in PAM and PCM only the minimum, central, crossfire might be significant. If, as in PPM, the pulses cannot be synchronized, the peak crossfire is significant. Curves for two values of band separation are shown in Fig. 27, one appropriate to yield minimum crossfire in the 25 to 35 db range and the other to yield peak crossfire in that range. This is the range that is sufficient for binary PCM. The steady state discrimination is also shown. We conclude from this study that pulses 4 or 5 are about equally good in respect to minimum central crossfire and that pulse 4 is slightly preferable in that the trough and the crest are more symmetrical. For PCM in which the pulse spacing is made equal to T_0 this symmetry means that there is the same margin for misalignment of the gating pulse, as regards correctly interpreting a space or a mark. Pulses 5 and 6 appear to be about equally good in respect to peak crossfire but both (and particularly pulse 6) incur a signal-to-noise penalty because the receiving filter does not permit the transmitted pulse to attain full height.

In practice, the approximations to Gaussian filters have shown worse

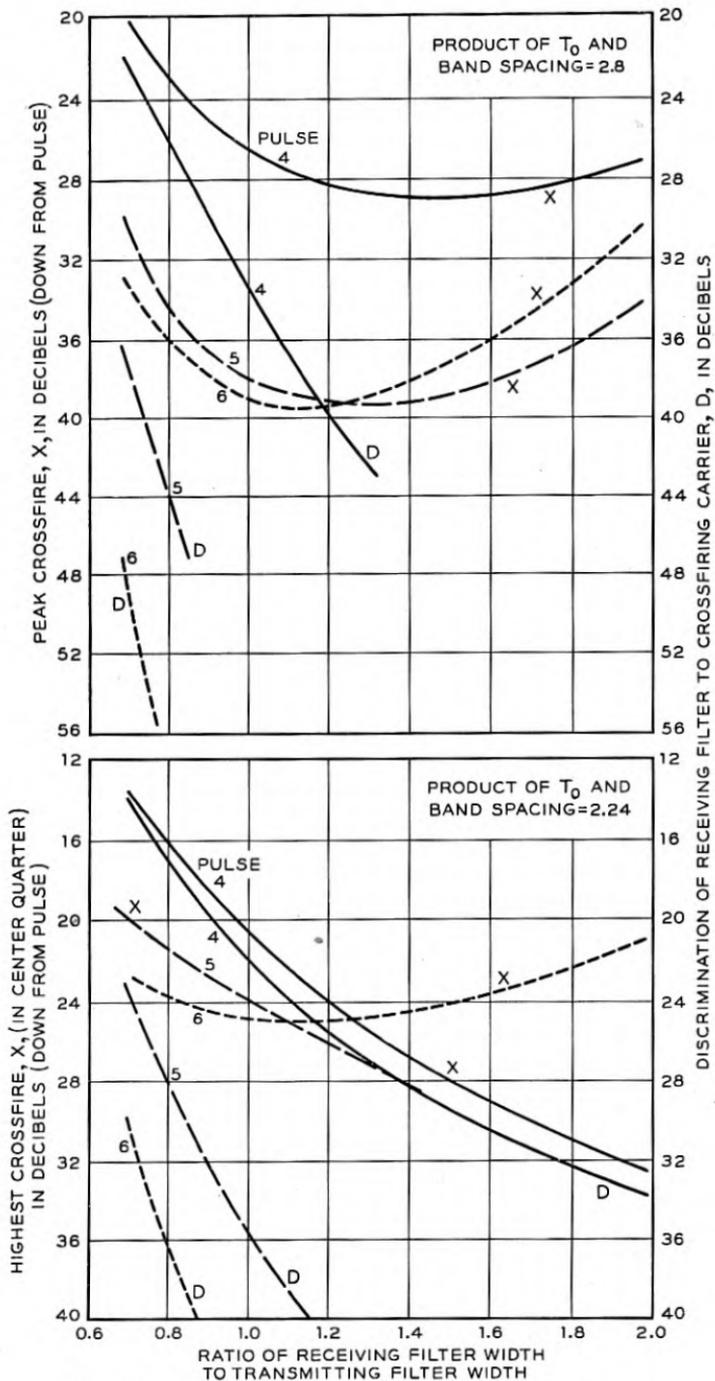


Fig. 27—Crossfire between frequency divided pulse groups.

pulse crossfire than the curves predict. This is particularly so for cases in which the curves show crossfire 30 or more decibels down. Approximations usually possess less rapidly falling attenuation skirts and possess phase distortion, both of which prevent realization of the calculated crossfire values.

Because regeneration (or reshaping) *permits* the use of Gaussian receiving filters, it does not follow that *flat-topped* filters are inferior as receiving filters. Calculations were made for maximally flat receiving filters of about the same overall complexity as was involved in the Gaussian approximations. They showed that when the transmitted pulse has the shape 4 and the flat filter is scaled to transmit such a pulse without much distortion, values of peak crossfire of the order of 30 db can be obtained when the product of band spacing and T_0 is 2.8. It was also found that the crossfire in that case consists of a single peak (not unlike the main pulse) nearly coincident in time with the crossfiring pulse. Our Gaussian approximations gave peak crossfire of this same order, for band spacing times $T_0 = 2.8$. The maximally flat receiving filter accepts roughly twice the noise power accepted by the optimum Gaussian filter, so the favor remains with the Gaussian filter and pulses 4 or 5.

The main conclusion from all of this is that, if smooth pulses, like numbers 4 or 5, are employed, band spacings of the order of $2.8/T_0$ (perhaps $2.5/T_0$) can be used with crossfire entirely suitable for binary PCM, as well as for PPM systems with sufficient swing ratio. Larger spacings would be required for PCM using multi-valued digits, and for PAM.

As mentioned earlier, the use of pulses 10 or 11 spaced by T_0 is possible in binary PCM, with small penalty, if very short accurately aligned gate pulses are used. The spectrum of these pulses is more sharply defined and includes a band only slightly wider than $1/T_0$. Rectangular receiving filters of that width could be used side-by-side so that the band spacing would be only slightly greater than $1/T_0$. This is the "theoretical minimum" and in telegraph parlance would be specified as a band spacing of twice the dot frequency.

Pulse 10 results from transmitting a very short pulse through a 4-section maximally flat filter whose response is shown in Column C. The phase distortion characteristics of such a filter produces asymmetry in the pulse. Pulse 11 is produced by the filter shown, assuming that the distortion is corrected. Most of the pulse shaping is assumed to reside in the transmitting filter. The assumed receiving filter is a 4-section maximally flat filter, and therefore has the shape of the filter shown for pulse 10, but is about 30% wider than shown. When two such bands are spaced $1.5/T_0$ the maximum crossfire is about 26 db down.

With shaping and receiving filters of reasonable complexity a band spacing

of $1.5/T_0$ to $1.7/T_0$ can be expected to have satisfactorily small crossfire for binary PCM. Pulse 11 and a spacing of $1.5/T_0$ were assumed for binary PCM-AM in Tables VI to VIII.

Figure 26 shows the envelopes of r.f. pulses produced by passing flat-topped r.f. pulses or r.f. *spikes* through r.f. filters. These envelopes are the baseband shapes produced by wide-band envelope detectors. If baseband pulses are shaped by baseband filters the resulting pulses are the same as shown for pulses 1 to 7, but for pulses 8 to 11 the tails turn out to be overshoots passing smoothly through zero instead of reaching zero cusp-wise. If these pulses are used to modulate the amplitude of a carrier in a product modulator, the cusps in the envelope are produced as shown, but if they are used to modulate the frequency of a carrier the baseband pulses produced by frequency detection retain their smooth transition through zero. In PAM-FM relatively wide gate pulses could be centered at time zero, T_0 , $2T_0$, etc., and the inter-pulse crosstalk would be partially balanced out by partial cancellation of positive and negative contributions. By the use of biases in the AM case a similar result could be obtained. Our tables, assuming pulse 4 spaced $2T_0$, do not reflect this possibility of operation.

DELAY LINE BALANCING

Techniques have been developed^{4, 40} in which the received pulse train is split into two or more branches, after detection to the baseband, and recombined with suitable delays, attenuations and polarity reversals. Such a procedure is effective in reducing the pulse tails or hangover and its use has been especially valuable in experimental PAM and PAM-FM systems. While this device may be regarded as a kind of phase and amplitude equalizer (comprising as it does only linear, passive elements) the result may be a pulse shape slightly more desirable than those obtained from simple but "ideal" networks, shown in Fig. 26. Our judgment that pulses of shape 4, spaced $2T_0$, should be used in PAM-FM may be slightly pessimistic if this kind of balancing is used.

More significant reductions of inter-pulse interference may be sought by the method suggested by MacColl⁴¹ (which is more than "equalizing") but this method, like the PCM method of Appendix III soon makes preposterous demands on the transmission medium and upon the transmitted power.

VIII. TRANSMISSION OVER METALLIC CIRCUITS

In radio relay transmission we have assumed a span length of 30 miles and have assumed span losses in keeping with the microwave antenna art

⁴ V. D. Landon, loc. cit.

⁴⁰ W. D. Boothroyd and E. M. Creamer, Jr., "A Time Division Multiplexing System," Paper presented at winter general meeting, *A.I.E.E.*, New York, Jan. 31, 1949.

⁴¹ *U. S. Patent No. 2,056, 284* Oct. 6, 1936 issued to L. A. MacColl.

and the relatively meager propagation experience now available. The high cost of the towers, power facilities and access roads involved in repeaters as we know them points to the desirability of few repeaters and long spans. Topography usually permits spans of a few tens of miles without requiring towers of excessive height. Very much longer spans can rarely be had without tremendous towers and are questionable because of the increase of fading depth with distance.

In wave guide (or other metallic conductor) transmission entirely different considerations apply and we will discuss some electrical relationships which seem significant in this case.

Let us consider a microwave repeater having a noise figure NF and a power capacity PC. The overload characteristic, together with the amount of nonlinear distortion that the signal can stand, determines the maximum output power. This maximum power is the power capacity. These two characteristics, PC and NF, thus determine the amount of attenuation that may be introduced between the transmitting half of a repeater regarded now as a transmitting terminal and the receiving half regarded as a receiving terminal. This amount of attenuation expressed in decibels, which we will designate as M , is available to be used up by the loss of one span plus accumulation of noise from n repeaters and may be regarded as a figure of merit of the repeater.

Five different relationships apply as follows:

$$\text{AM Systems: } M = \text{span loss}_{db} + 20 \log n \quad (1)$$

$$\text{FM Systems: } M = \text{span loss}_{db} + 10 \log n \quad (2)$$

PPM-AM Systems with

$$\begin{aligned} \text{reshaping repeaters: } M &= \text{span loss}_{db} + 5 \log n & (3) \\ \text{Band increased } &^2\sqrt{n} \text{ referred to (1)} \end{aligned}$$

FM Systems with

$$\begin{aligned} \text{limiting repeaters: } M &= \text{span loss}_{db} + 3.33 \log n & (4) \\ \text{Band increased } &^3\sqrt{n} \text{ referred to (2)} \end{aligned}$$

PCM Transmission with

$$\text{regenerative repeaters: } M = \text{span loss}_{db} + \text{zero} \quad (5)$$

In (1) the $20 \log n$ term includes $10 \log n$ for noise accumulation plus $10 \log n$ for cumulative compression. In microwave amplifiers only odd-order terms contribute to the distortion and the third order term predominates for moderate degrees of overload. This results in the well-known compression characteristic such as appears in Figs. 24 and 25 previously discussed. A significant approximation for the over-all compression when

n such amplifiers are connected in tandem is that the compression characteristic is the same as with one amplifier but occurs at outputs $10 \log n$ db lower. Thus, the power level must be reduced $10 \log n$ db and this penalty accrues over and above the noise accumulation penalty.

In (2) only noise accumulation occurs.

In (3) and (4) it is assumed that minimum power conditions are attained and the operation has reached the straight part of the minimum (marginal) power curves. Without reshaping, the system must be powered so that, at the final repeater, the accumulation of noise does not exceed the marginal value. With reshaping at each of the n repeaters each span may be marginal. Making each span marginal with the same bandwidth would be accomplished with $10 \log n$ db less power and would make the signal-to-noise ratio $10 \log n$ db lower. This can be made up by using more bandwidth. In marginal PPM-AM the signal-to-noise ratio improvement occurs at the rate of 20 db per decade of bandwidth and thus the bandwidth must be increased by $^2\sqrt{n}$. This requires, to keep the operation marginal, an increase in power of $10 \log n^{1/2} = 5 \log n$ db. In the case of marginal FM, signal-to-noise ratio is improved at the rate of 30 db per decade, and the bandwidth must accordingly be increased by $^3\sqrt{n}$. To keep the operation marginal, the power must be increased $10 \log n^{1/3} = 3.33 \log n$ db. The entries in Tables II and III invoke these relationships. There, n may be thought of as having values 1, 5 or 133.

Equation (5) reflects the fact that where PCM regenerative repeaters are employed no accumulation of noise occurs with number of spans.

With metallic conductors, the span loss in decibels is proportional to the length of span. If A denotes the span loss in decibels per mile and S denotes span length in miles, the circuit length $L = nS$ and

$$M = \frac{AL}{n} + x \log n \quad (6)$$

or,

$$L = \frac{n}{A} M - \frac{nx}{A} \log n \quad (7)$$

where x is the appropriate coefficient, 20, 10, 5 or 3.33. In this expression there is an optimum value of n corresponding to a maximum value of circuit length L . Figure 28 is a plot of circuit length for $x = 20$ (Eq. 1), showing the maxima. Figures 29 to 32 show the optimum values of n and the resulting maximum circuit lengths for each of the relations expressed in equations (1), (2), (3), (4).

Considerations affecting transmission over metallic circuits are different from those affecting radio relay in at least the following four ways:

1. Interference from other routes substantially vanishes with coaxial and

wave-guide conductors. This diminishes the premium on ruggedness provided sufficient power is available so that ruggedness with respect to noise is not critical.

2. Since there is no fading and all spans can be of approximately equal length, all spans will possess the same loss, approximately. This situation

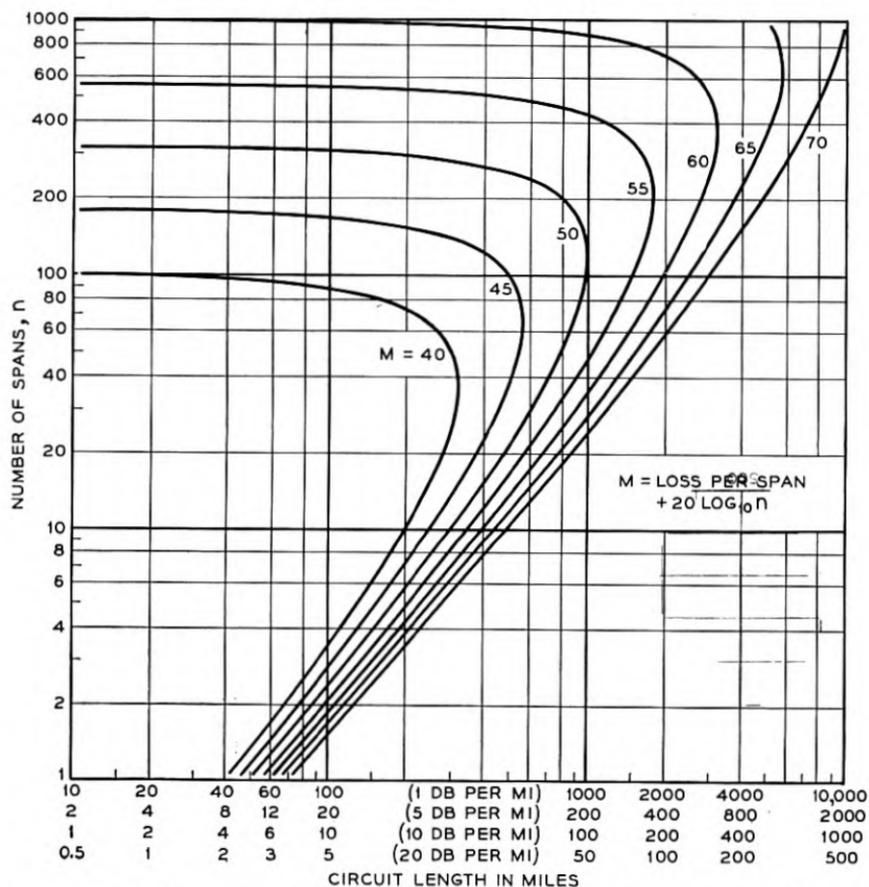


Fig. 28—Variation of circuit length with number of repeater sections in an AM system with fixed power capacity and noise figure.

is favorable to all systems but is most favorable to PCM, which gets no credit for low loss spans.

3. In the case of wave guide, frequency space may be much less precious than in coaxial or radio relay transmission.

4. There is a possibility that many small repeaters should replace the few higher powered repeaters used in radio relay.

These different considerations may lead to a different evaluation of the

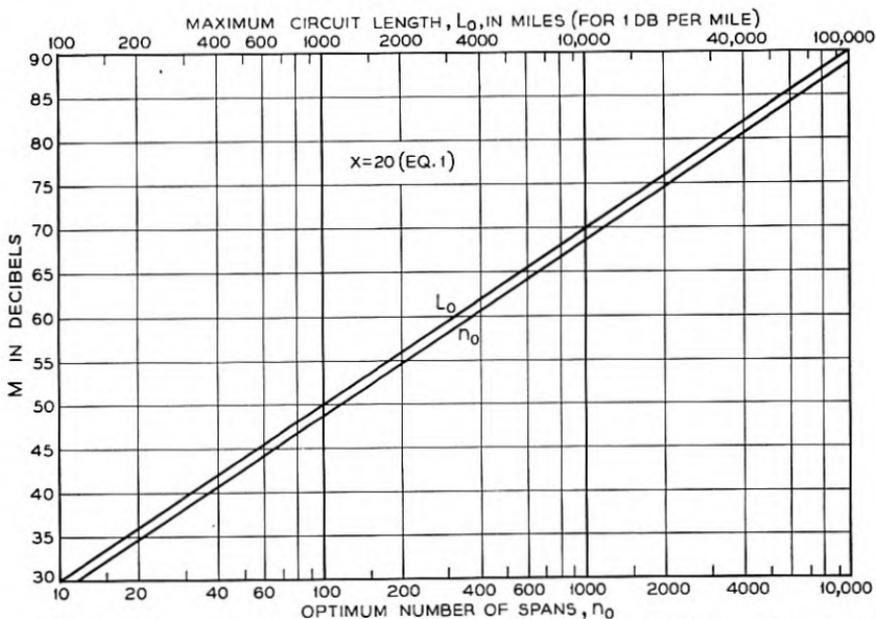


Fig. 29—Optimum number of repeater sections and maximum circuit length for metallic AM system with fixed power capacity and noise figure.

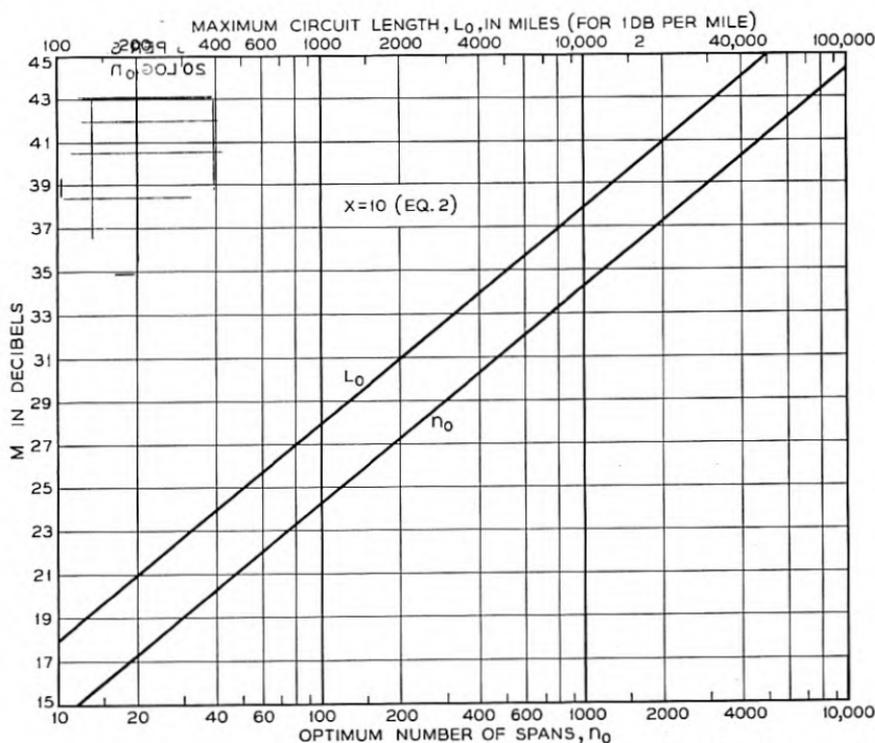


Fig. 30—Optimum number of repeater sections and maximum circuit length for metallic FM system with limiting only at end of system.

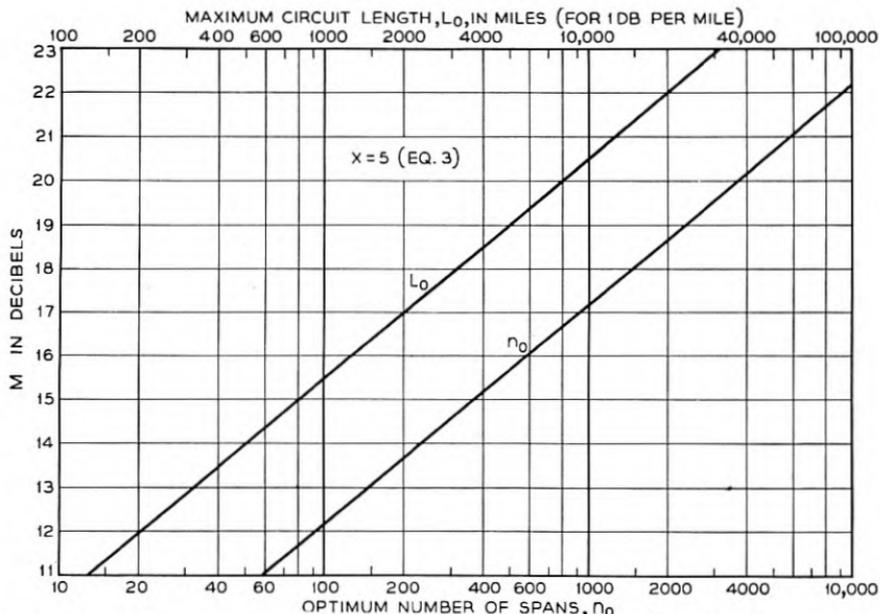


Fig. 31—Optimum number of repeater sections and maximum circuit length for metallic PPM-AM system with reshaping at every repeater.

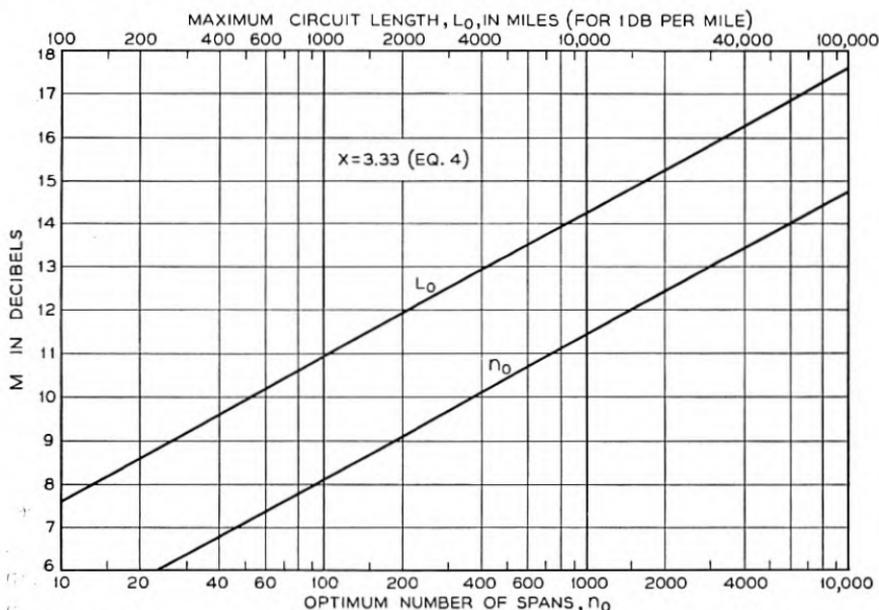


Fig. 32—Optimum number of repeater sections and maximum circuit length for metallic FM system with limiting at every repeater.

modulation methods discussed in this paper. We will not attempt to make such a re-evaluation here.

It is of interest now to return to radio relay transmission and examine the relations derived for metallic conductors, but now assuming that the span attenuation is that associated with an inverse k -power of distance law ($k = 2$ for free space attenuation). If we use the symbol E to denote the excess power capacity (in decibels) of the repeater over that required for a unit span of, say, one mile, we get the relation

$$10k \log L = E + (10k - x) \log n \quad (8)$$

where $x = 20, 10, 5, 3.33, 0$ for the cases described by equations (1), (2), (3), (4), (5) respectively. The equation shows no optimum number of spans

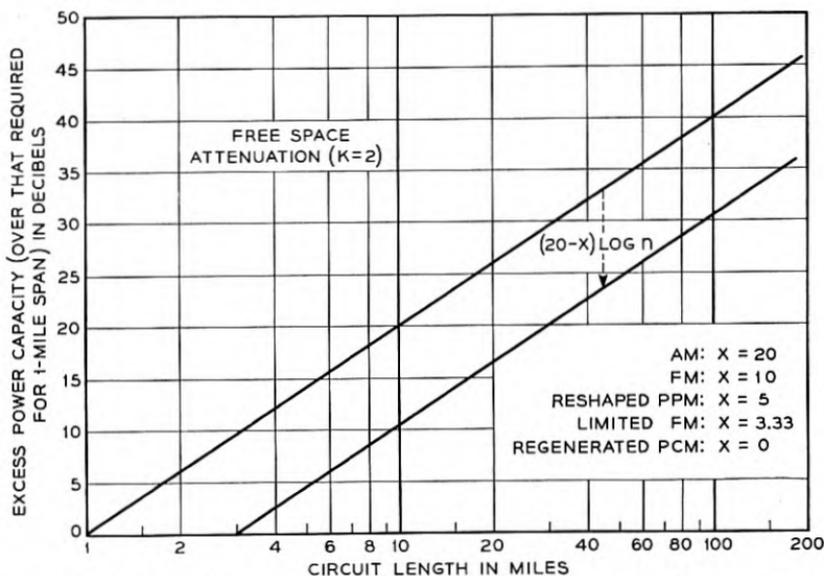


Fig. 33—Relation between circuit length, power, and number of repeaters in radio relay systems.

corresponding to a maximum circuit length. It also shows that when x is less than $10k$ the circuit length can be increased indefinitely by adding spans although the spans become shorter with increased circuit length. When $x = 10k$ the circuit length can not be increased beyond the maximum single span, i.e., it depends solely upon E and is not affected by the number of spans. If x is greater than $10k$ the circuit length again cannot be increased beyond the maximum single span and is reduced by employing more than one span. This last case does not occur for free space attenuation. In Fig. 33 is plotted the relationship between L , E and n for the free space attenuation law ($k = 2$). The curve passing through zero decibels excess power capacity at one mile circuit length applies to one span for any value

of x or to any number of spans when $x = 20$. In other words the maximum circuit length for $x = 20$ is the length of span corresponding to the excess power capacity as noted above for $x = 10k$. For all smaller values of x any circuit length can be achieved with any value of excess power capacity if a sufficient number of spans is employed. The number of spans required for a given circuit length is obtained by moving the curve downward until it intersects the desired length at the appropriate excess power ordinate, and equating $(20 - x) \log n$ to the downward shift in decibels.

Notwithstanding the present radio outlook in which large towers and antennas seem indicated, it is of interest to imagine small repeaters powered for a one-mile span, say. Using FM with limiting at every repeater, a 100-mile circuit could be obtained with 250 repeaters spaced 0.4 miles. This result comes from Fig. 33 with excess power = zero db and $x = 3.33$. A difficulty with such a case might be multiple paths produced by one repeater output overreaching into other spans.

The inverse k power attenuation does not accurately describe propagation over long spans; fading then occurs and is greater for long spans than short spans. This introduces a term in the span loss similar to that of the metallic conductor case in which the span loss is proportional to span length.

IX. CONCLUSIONS

We have, in this paper, examined some of the relations governing the exchange of bandwidth for advantages in transmission that grow out of the liberal use of bandwidth. While we have not dealt specifically with the instrumentation involved in the application of the various exchange methods, we have taken cognizance of certain basic obstacles in circuit design such as overload distortion, phase distortion and discrimination characteristics of selective networks and the limitations of microwave antennas. Not having, in most cases, a wealth of experience bearing on the manner in which these obstacles affect the transmission problem, we have been obliged to estimate their effect in many cases. Considerable unreliability in these estimates would not, however, much affect the broad purpose of the paper. The economic factor that is involved in achieving reliable operation of apparatus has been largely ignored, although methods that seem to lead to fantastic instrumentation have not been given much attention.

Ruggedness of the transmitted signal, which is obtained at the cost of increased *bandwidth* can, properly handled, be made to conserve *frequency occupancy* in two ways: (a) ruggedness reduces the required "guard space" between one band and neighboring bands carrying other signals; (b) ruggedness reduces the multiplication of frequency assignments necessary in congested radio route situations.

For wave guide systems, the inter-route interference problem arising from

route congestion disappears but ruggedness is still a valuable feature. As to PCM, regeneration is an outstanding asset applicable also in wave guide transmission. In the case of very high-grade channels the unique advantage of PCM that stems from the coding principle is presumably valuable in any transmission medium. We have shown that, theoretically, PCM methods can achieve lower power requirements than any of the other methods considered and can do so with considerably less frequency space.

While this paper is primarily concerned with the transmission of multiplex telephony, it seems appropriate to dwell briefly on the transmission of television signals by radio relay. The repeater plan of Fig. 4 is capable of handling long distance transmission of a 5-mc (video) television signal (by FM). The frequency occupancy of a single two-way route is 80 mc. The occupancy for 1000 4-kc telephone channels is 72 mc from Table VI for binary PCM-AM with dual polarization. At this rate a 5-mc video television band would require 90 mc assuming that the 39 db signal-to-quantizing noise ratio is satisfactory for television.⁴² Remembering that route congestion can lead to a greater occupancy than 80 mc in the FM case and perhaps to no increase over 90 mc in the PCM case, we conclude that on these assumptions PCM might be a desirable method for long television relay routes. In the event that a better signal-to-noise ratio is found necessary, binary PCM provides 6 db improvement for each additional digit.

These conclusions relate to the transmission problem under the assumed conditions, and do not reflect the impact of many factors that may grow out of an application to a real situation. As has been said before, this paper should be taken to illustrate the way in which the transmission factors are interrelated, and the philosophy by which the problem is approached, rather than to find an unequivocally best system.

In preparing this paper the authors have, of course, drawn on the general transmission background of the Bell Telephone Laboratories. Nourishment has come particularly from W. M. Goodall, A. L. Durkee, H. S. Black, D. H. Ring, J. C. Schelleng and F. B. Llewellyn in addition to those mentioned specifically in the paper.

We wish specifically to thank Mr. R. K. Potter, whose broad transmission concepts were responsible for initiating the work.

APPENDIX I

NOISE IN PCM CIRCUITS

In the transmission of speech by PCM the kinds of noise and distortion which are acquired by other systems in transmission are completely missing.

⁴² W. M. Goodall, "Television by Pulse-Code Modulation." Paper presented at 1949 IRE National Convention, New York, March 9, 1949.

Instead, a special kind of impairment is incurred at the terminals, because of the fact that the speech wave is transmitted by quantized amplitude samples of the wave. Transmitting samples of a wave results in a received wave having no impairment, provided the samples are not subjected to time or amplitude distortion. In PCM, since the samples are telegraphed, their reception is inaccurate by the quantization imposed by the code. These errors in the samples constitute the sole inherent impairment in transmission.

Strictly speaking, the transmission impairment in PCM is manifested only when a signal is being transmitted. An imaginary telephone circuit with the transmitting side completely devoid of any kind of signal, except that from the talker, could be transmitted by PCM from coast to coast and would sound completely silent if the talker were silent. In any real situation, however, some background noise (room noise, breath noise or line noise) is always present in the subscriber's circuit. This background noise is usually comparable to or greater than the weak parts of weak speech. In order to transmit the speech of weak talkers the size of the discrete amplitude steps must be small with the result that at least a few steps are always brought into play by background noise.

Being thus enabled to rule out the case of no *signal* we are able to ascribe a basic signal impairment to a PCM system. This impairment is, strictly speaking, a result of non-linear distortion inherent in the quantizing, but because of its very complex nature it behaves, and sounds, much like thermal noise and we have accordingly called it quantizing *noise*. A PCM circuit can be regarded as a source of noise whose rms value is simply related to the size of the quantizing step and the sampling frequency, as follows:

In a low-pass band extending to approximately 40% of the sampling frequency the basic noise power is related to the power of a sine wave signal by

$$\frac{\text{Signal power}}{\text{Noise power}} = 20 \log_{10} \frac{\text{peak-to-peak signal voltage}}{\text{step voltage}} + 3 \text{ db}$$

This band of noise has an amplitude distribution somewhat different from thermal noise, and a spectral distribution which depends somewhat upon the spectral distribution of the signal and upon its amplitude and disposition with respect to the step boundaries.

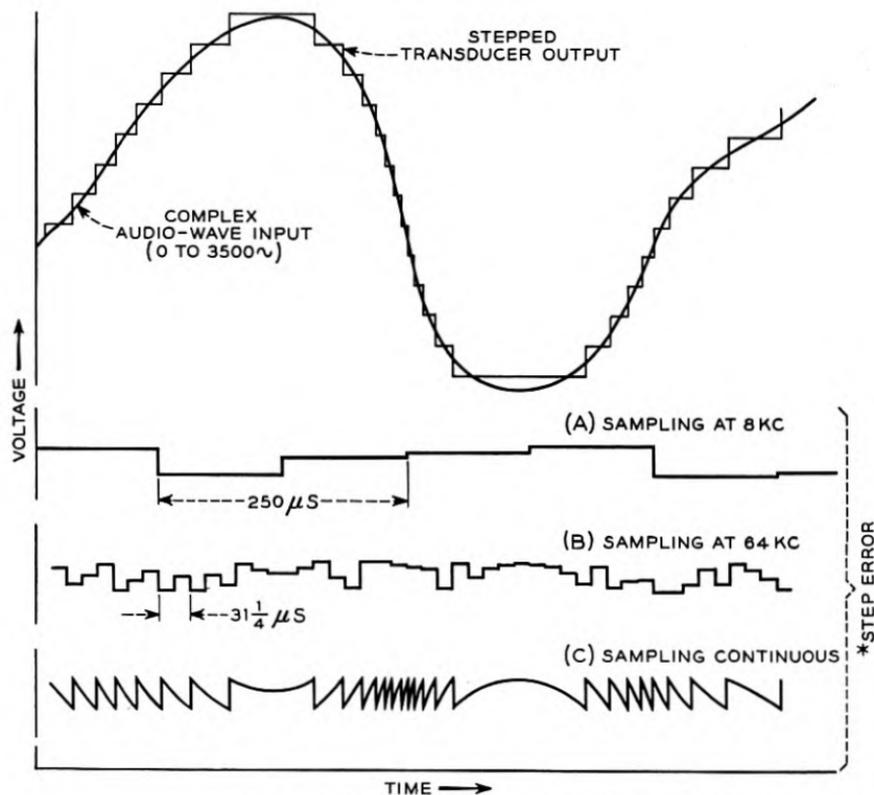
For a sine wave signal the noise spectrum is characterized by a number of prominent components rising above a diffuse background of numerous smaller components. The outstanding components may be either harmonics of the signal frequency or differences between harmonics of the sampling frequency and harmonics of the signal frequency. The background thus consists of an array of various orders of cross-products between the signal and the sampling rate. When the amplitude of the signal is comparable to one step in the quantizing process, a few components may contain a substan-

tial portion of the total power in the distortion spectrum. If the signal is not only weak but has its frequency near the low edge of the band, the distortion spectrum has a decided downward slope on the frequency scale with a major part of the distortion power concentrated in the lower harmonics of the signal frequency. Similarly, a weak signal at the upper edge of the band may cause a few scattered difference products to be outstanding. Stronger signals with more centrally located frequencies give practically a uniform distribution of distortion power throughout the signal band. For all except the extreme cases of low amplitudes and frequencies near the edges of the band, the weighting network used to evaluate the telephonic interfering effect of noise gives a reading equal to that obtained with a flat band of thermal noise of the same mean power. The exceptional cases show a spread in the readings which are sensitive to amplitude, frequency and disposition with respect to step boundaries. The spread is reduced when complex signal waves are applied. An operationally significant case is that in which the noise is produced by residual power hum in the equipment. In such a case, weighted noise readings range from approximately the value obtained for flat noise of the same mean power, to several db lower. Connecting even a short subscriber's loop to the input usually adds enough miscellaneous noise, if the steps are as small as they need to be, to remove the variability and to yield a reading within one db of the equivalent flat noise case.

Thus, a PCM system, like any other transmission system, possesses a noise source and experiments show that this noise combines by power addition with that from another system connected to the input or output of the PCM system. In tandem connections of PCM systems in which successive quantizations may occur, the quantizing noise also adds like power, from system to system, and soon becomes almost indistinguishable from thermal noise.

The quantizing noise consists of distortion products which may be classified as two kinds. One class includes those products which would be produced by transmitting the wave through a transducer whose input-output characteristic is stepped like a staircase. If such a transducer were actually used the PCM process would be equivalent to sampling its output at a regular rate and transmitting the step designations by code. This sampling process, applied to the stepped transducer output, produces the other class of distortion (or noise) and is illustrated in Fig. 34. Let us consider the sampled value as the sum of the true value plus the step error, and focus attention on the step error which is responsible for the distortion. At minimum permissible sampling frequency (twice the highest signal frequency), the step errors in consecutive samples are practically unrelated to each other. The low-pass output filter passes most of the power in this sequence of random errors

when they occur at a frequency only twice the filter cutoff frequency. See A in Fig. 34. If the sampling frequency were increased from the minimum permissible value, the consecutive step errors would still be unrelated to each other, and more and more of the step error spectrum (noise) would fall above the low-pass filter. This is shown in B.



* THE QUANTIZING NOISE CONSISTS OF THE RESPONSE OF A 3500-CYCLE LOW-PASS FILTER TO THE STEP ERROR

Fig. 34—Stepping and sampling an audio wave.

Reduction of noise would occur in this way until the sampling frequency became so high that a considerable number of samples are taken while the wave crosses a step interval. Correlation between successive step errors then begins to be apparent. When the interval between samples becomes vanishingly small, the process is equivalent to transmitting the stepped transducer output directly. This case appears in C.

In an alternate line of thinking, one may regard the stepped transducer

output as the signal wave plus a wide spectrum of distortion frequencies representing the effect of the steps. From this point of view, it is clear that only a high sampling frequency prevents lower sidebands associated with the sampling frequency and its harmonics from overlapping the signal band.

Quantizing noise decreases with increase of sampling frequency at an initial rate of approximately 3 db per octave and continues until correlation of successive errors becomes appreciable. This occurs at a sampling frequency which is dependent upon the spectral distribution of the signal, being lower for signals having a predominately low-frequency spectral density. An increase of step size also reduces the lowest sampling frequency at which effects of correlation are observed. Figure 35 shows curves calculated for an

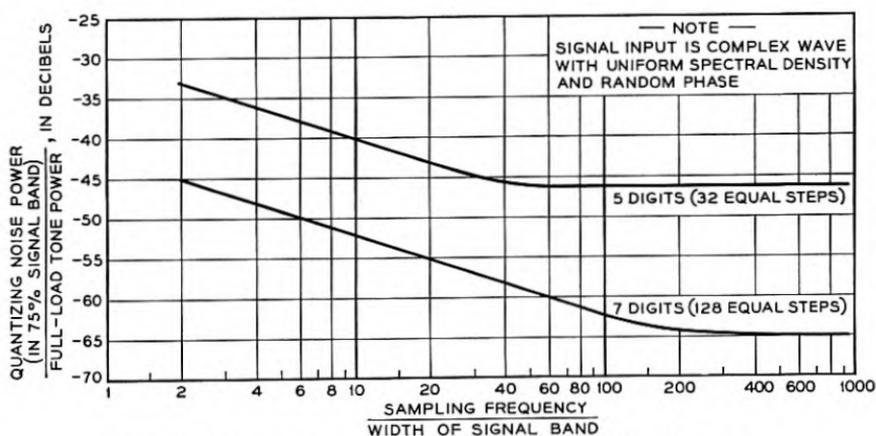


Fig. 35—Variation of quantizing noise with sampling frequency.

input consisting, in fact, of thermal noise. Such an input is a rough approximation to a speech wave.

The asymptotic values shown for five and seven digits represent the quantizing noise corresponding to transmission of the thermal noise signal through stepped transducers having 32 and 128 steps, respectively. The curves suggest that sampling is a penalty such that 32-step granularity without sampling is about equivalent⁴³ to 128-step granularity with sampling at the minimum rate. However, sending information which designates the irregular instants of time at which the signal enters and leaves each step interval is far less efficient than designating the steps at the regular instants of the minimum sampling rate.

⁴³ The equivalence would be in terms of total noise power; the properties of the asymptotic noise are different than were described earlier in this appendix, for sampling at the minimum rate.

APPENDIX II

INTERFERENCE BETWEEN TWO FREQUENCY MODULATED WAVES

This problem occurs so frequently in the present paper that its solution is appended here for reference. Figure 36 shows a geometric figure from which the phase of a two-component wave can be calculated. We write

$$P \cos \theta + Q \cos \varphi = R \cos \psi$$

where

$$R^2 = P^2 + Q^2 + 2PQ \cos (\theta - \varphi)$$

$$\tan \psi = \frac{P \sin \theta + Q \sin \varphi}{P \cos \theta + Q \cos \varphi}$$

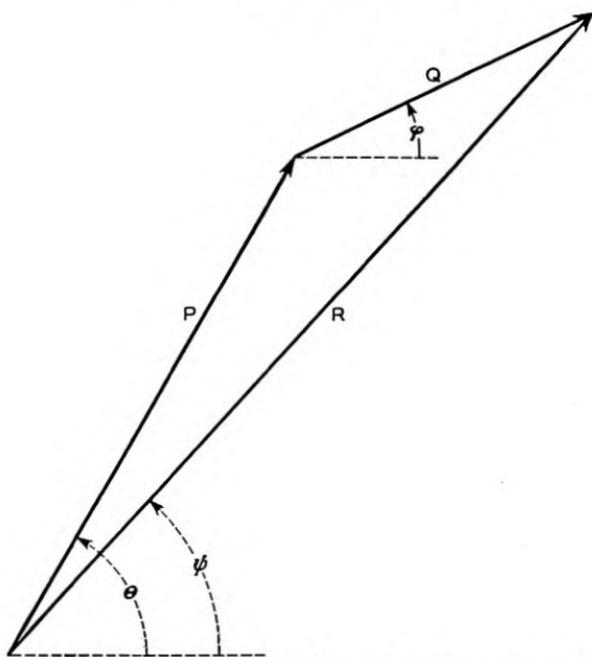


Fig36. —Geometric solution for resultant phase of two frequency modulated waves.

The response of a perfect frequency detector in radians/sec. is given by

$$\begin{aligned} \Omega &= d\psi/dt = \frac{d}{dt} \left(\text{arc tan } \frac{P \sin \theta + Q \sin \varphi}{P \cos \theta + Q \cos \varphi} \right) \\ &= \frac{1}{2}(\theta' + \varphi') + \frac{\theta' - \varphi'}{2} \frac{P^2 - Q^2}{P^2 + Q^2 + 2PQ \cos (\theta - \varphi)} \end{aligned}$$

In the above expression, the primes represent derivatives with respect to time. If $Q/P < 1$, we expand in Fourier series, obtaining

$$\Omega = \theta' + (\theta' - \varphi') \sum_{m=1}^{\infty} \left(-\frac{Q}{P}\right)^m \cos m(\theta - \varphi).$$

When Q/P is small, we retain only the term proportional to Q/P as the error, which may be written in the compact form:

$$\Omega - \theta' \doteq \frac{Q}{P} \frac{d}{dt} \sin(\varphi - \theta)$$

If the waves are unmodulated, $\theta = pt$ and $\varphi = qt$, giving

$$\Omega - \theta' \doteq \frac{Q}{P} (q - p) \cos(q - p)t$$

APPENDIX III

PCM FOR BANDWIDTH REDUCTION

We have treated PCM as a means of increasing bandwidth beyond the value corresponding to one pulse per sample per channel (quantized PAM) and have studied the transmission advantages that accrue therefrom. The PCM method can, in principle, serve to reduce bandwidth. An example of bandwidth reduction,^{44, 45} suggested to the writers by C. E. Shannon, is as follows:

Any number, say N , of 4 kc telephone channels can be transmitted in the form of one quantized pulse per 125 microseconds, by sampling all channels in the usual way, encoding each sample into a code symbol having, say, 64 possible values, assembling all code pulses into one new group and decoding this group at the transmitter. If only one channel were to be transmitted the decoded signal would have 64 possible amplitudes; for two channels it would have 64^2 possible amplitudes, and for N channels, 64^N . Now, if a single quantized pulse conveying these amplitudes could be transmitted without an error as large as one step, the receiver could encode the quantized pulse, disassemble the resulting code pulses into groups according to channels and decode the groups to obtain the N channel samples. The requirements on transmission circuits capable of the precision required to transmit even two channels in place of one are very severe, however.

In the event the signals to be transmitted were not speech signals but a very elemental kind of signal such as a black and white pattern requiring for

⁴⁴ A paper "Reducing Transmission Bandwidth" by Bailey and Singleton *Electronics*, Aug. 1948 gives a somewhat different example of reduction.

⁴⁵ An early disclosure of a system theoretically capable of any desired amount of bandwidth reduction is contained in U. S. Patent No. 2,056,284, Oct. 6, 1936, issued to L. A. MacColl. As in the current proposals, the decreased band is obtained at the expense of a vastly greater signal-to-noise ratio requirement and the necessity for precise synchronism between transmitter and receiver.

its description not 64 values but only 2, the number of possible amplitudes would be 2^N . With some of the better transmission circuits in existence, as many as 10 such channels could be multiplexed in the same bandwidth required by one channel, by this adaptation of the PCM method.

The above considerations show that PCM offers the means of matching the transmission signal to the capabilities of the transmission circuit in order to transmit the maximum amount of information. As has been shown, with microwave telephone systems the economical balance seems to come well over on the wide band side, permitting operation with low transmitted power through relatively strong interference.

APPENDIX IV

SUPPLEMENTARY DETAILS OF DERIVATION OF BANDWIDTH-CURVES

PPM-AM

The diagram of Fig. 8 shows that the maximum time deviation is assumed to be

$$\epsilon = \frac{1}{2Nf_r} - T = \frac{1}{2Nf_r} - \frac{1}{F_b} \quad (1)$$

The shift in time produced by an interfering voltage of magnitude E_n at the slicing instant is

$$\Delta t = E_n/s \quad (2)$$

where s is the slope of the signal pulse at the slicing instant. For small noise the slicing instant occurs near half the peak, E , of the pulse and the slope of the assumed sinusoidal pulse (Fig. 6) is:

$$s = \frac{\pi}{2T} E = \frac{\pi F_b E}{2} \quad (3)$$

Hence

$$\Delta t = \frac{2E_n}{\pi F_b E} \quad (4)$$

The signal-to-noise power ratio in the channel is the ratio of mean square values of signal deviation σ and Δt . For thermal noise we assume that the root mean square value is one fourth the peak and place the peak at $1/\sqrt{2}$ times $E/2$ for marginal operation. We write, therefore,

$$E_n = E/2\sqrt{2}, \quad \overline{\Delta t^2} = (4E_n^2/\pi^2 F_b^2 E^2)/16 \quad (5)$$

$$\overline{\sigma^2} = \frac{\epsilon^2}{2} = \frac{1}{2} \left(\frac{1}{2Nf_r} - \frac{1}{F_b} \right)^2 = \frac{1}{2F_b^2} \left(\frac{F_b}{2Nf_r} - 1 \right)^2 \quad (6)$$

$$S/N = 4\pi \left(\frac{F_b}{2Nf_r} - 1 \right) = 4\pi \left(\frac{B}{4Nf_r} - 1 \right) \quad (7)$$

Here S/N is the ratio of root mean square audio signal and noise voltages. The formula illustrates a general principle common to all the pulse systems in that the marginal signal-to-noise ratio is a function of B/Nf_r . The axes of the curves have been labeled for $N = 1000$ and $f_r = 8$ kc, but can be read for any other values of N and f_r , by changing B accordingly. Equation (7) was used to plot the marginal power curve of Fig. 9. We note that the ratio of rms pulse voltage at the top of the pulse to the rms noise voltage is $E/\sqrt{2}$ divided by $E_n/4$ which leads to a value of eight when (5) is substituted. The "slicer advantage" is thus the right-hand member of (7) divided by eight.

For CW interference in a PPM-AM system the procedure used above applies except that the root mean square interference is now $1/\sqrt{2}$ times the peak instead of one fourth. The marginal ratio of rms audio signal to rms audio interference ratio is therefore poorer by a factor of $2\sqrt{2}$, or

$$S/I = \pi \sqrt{2} \left(\frac{B}{4Nf_r} - 1 \right). \quad (8)$$

When the interference is from a similar system, we calculate the distribution of the disturbance as follows. The probability that there is an interfering pulse present during slicing is the ratio of the pulse duration to the channel allotment, or

$$p_0 = 2TNf_r = \frac{2Nf_r}{F_b}. \quad (9)$$

The interfering carrier will not, in general, be exactly synchronous with the wanted carrier, and hence the actual interference is a beat frequency with envelope having a voltage distribution calculable from the pulse shape. For a sinusoidal pulse of height A , the probability $p(y) dy$ that the instantaneous magnitude of the interfering envelope is in the interval dy at y is

$$p(y) dy = \frac{p_0 dy}{\pi \sqrt{y(A-y)}}, \quad 0 < y < A. \quad (10)$$

Since the relative phase of the two carriers drifts with time, the mean square interfering voltage is half the mean square interfering envelope, or

$$E_n^2 = \frac{1}{2} \int_0^A y^2 p(y) dy = \frac{3Nf_r A^2}{9F_b} = \frac{3Nf_r A^2}{4B}. \quad (11)$$

Hence

$$\overline{\Delta t^2} = \frac{12Nf_r A^2}{\pi^2 B^3 E^2} \quad (12)$$

and

$$\frac{S}{I} = \frac{\pi E}{A} \left(\frac{B}{4Nf_r} - 1 \right) \sqrt{\frac{B}{6Nf_r}}. \quad (13)$$

For marginal interference $E = 2\sqrt{2}A$ and

$$\frac{S}{I} = 2\pi \left(\frac{B}{4Nf_r} - 1 \right) \sqrt{\frac{B}{3Nf_r}}. \quad (14)$$

This equation shows that S/I varies as $(B/Nf_r)^{3/2}$ for large bandwidths giving 9 db improvement per octave of bandwidth. The curves of Fig. 10 were plotted from equations (8) and (14).

PPM-FM

The pulse here is transmitted by a change in frequency from f_1 to $f_1 + \beta$ and back again. The total frequency swing β corresponds to the pulse height E in the AM case. The frequency detector delivers a pulse of height β to the baseband filter. Associated with the pulse is the error caused by noise or interference in the rf-band. In the case of fluctuation noise having mean power P_n per cycle in the rf-medium, a baseband filter of width F_b accepts the familiar triangular voltage distribution of noise with frequency resulting⁴⁶ in a mean square integrated magnitude expressed on a frequency scale as:

$$E_n^2 = P_n F_b^3 / 3W_c \quad (15)$$

where W_c is the mean carrier power. Then, on substituting β for E , and the above expression for $\overline{E_n^2}$ in the equation for Δt :

$$\overline{\Delta t^2} = \frac{4P_n F_b}{3\pi^2 \beta^2 W_c} \quad (16)$$

Taking the ratio of $\overline{\epsilon^2}$ to $\overline{\Delta t^2}$,

$$(S/N)^2 = \frac{3\pi^2 \beta^2 W_c}{8P_n F_b^3} \left(\frac{F_b}{2Nf_r} - 1 \right)^2. \quad (17)$$

The radio signal bandwidth B is approximately equal to the frequency swing plus a sideband at each end or

$$B = \beta + 2F_b \quad (18)$$

Using this relation to eliminate β , we have

$$(S/N)^2 = \frac{3\pi^2 (B - 2F_b)^2 W_c}{8P_n F_b^3} \left(\frac{F_b}{2Nf_r} - 1 \right)^2. \quad (19)$$

For marginal operation of the FM limiter:

$$W_c = kP_n B \quad (20)$$

where we shall assume $k = 16$ in numerical calculations.

⁴⁶ An elementary component of interference $Q \cos qt$ produces a frequency error $(Q/P)f \cos 2\pi ft$ where f is the difference between the interfering and carrier frequencies. The corresponding mean square frequency error is $f^2 Q^2 / 2P^2$. But $Q^2/2 = P_n df$ and there are equal contributions from upper and lower sidebands centered around the carrier. Also replacing $P^2/2$ by W_c , we get a mean square frequency error in band df at f equal to $P_n f^2 df / W_c$. Integrating over frequencies from 0 to F_b gives the above result.

Then

$$(S/N)^2 = \frac{3k\pi^2}{8} \left(\frac{B}{F_b}\right)^3 \left(1 - 2\frac{F_b}{B}\right)^2 \left(\frac{F_b}{2Nf_r} - 1\right)^2. \quad (21)$$

The signal-to-noise ratio is found to be maximum when

$$\frac{F_b}{Nf_r} = \sqrt{\left(\frac{B}{4Nf_r} + 1\right)^2 + \frac{3B}{Nf_r}} - \frac{B}{4Nf_r} - 1. \quad (22)$$

To calculate the CW interference with an idle channel we assume that the carrier wave of the system is represented by

$$V_1(t) = P \cos [2\pi f_1 t + \phi(t)] \quad (23)$$

where

$$\phi(t) = \left\{ \begin{array}{ll} \pi\beta \left(t + \frac{1}{\pi F_b} \sin \pi F_b t \right), & 0 < t < \frac{1}{F_b} \\ \pi\beta/F_b, & \frac{1}{F_b} < t < \frac{1}{2Nf_r} \end{array} \right\} \quad (24)$$

$$\phi(-t) = -\phi(t), \quad \phi\left(t \pm \frac{2m}{F_b}\right) = \phi(t), \quad m = 1, 2, \dots \quad (25)$$

We have chosen $\phi(t)$ so that the phase is a continuous function of time with a derivative representing the correct frequency. This gives a sinusoidal change in the instantaneous frequency $\phi'(t)/2\pi$ starting with the value f_1 at $t = -\frac{1}{F_b}$, reaching the peak $f_1 + \beta$ at $t = 0$, and subsiding to f_1 at $t = \frac{1}{F_b}$. By making the wave repeat with frequency Nf_r , we assume all channels of the system are idle since all pulses are at their central positions. It seems reasonable to neglect the effect of variations in adjacent channel loading on CW interference in one channel. The interfering CW wave is represented by

$$V_2(t) = Q \cos (2\pi f_2 t + \theta) \quad (26)$$

To a first approximation the resulting error in frequency at the output of the frequency detector is:

$$\delta(t) = \frac{Q}{2\pi P} \frac{d}{dt} \sin [2\pi(f_1 - f_2)t + \phi(t) - \theta] \quad (27)$$

By straightforward Fourier series expansion and differentiation:

$$\delta(t) = \frac{Q}{P} F_b \sum_{n=-\infty}^{\infty} (c + n\lambda) A_n \cos [2\pi F(c + n\lambda)t - \theta] \quad (28)$$

where:

$$c = (f_1 - f_2)/F_b, \quad \lambda = Nf_r/F_b, \quad y = \beta/F_b, \quad (29)$$

$$A_n = 2\lambda \mathcal{F}_{2n\lambda - y}(y) - \frac{1}{n\pi} [(-)^n + \sin(2n\lambda - y)\pi], \quad (30)$$

$$A_0 = 2\lambda \mathcal{F}_{-y}(y) + (1 - 2\lambda) \cos \pi y. \quad (31)$$

The function $\mathcal{F}_\nu(y)$ is Anger's function:⁴⁷

$$\mathcal{F}_\nu(y) = \frac{1}{\pi} \int_0^\pi \cos(\nu\theta - y \sin \theta) d\theta \quad (32)$$

It is equal to $J_\nu(y)$, the more familiar Bessel function of the first kind, only when ν is an integer. The values of $\nu = 2n\lambda - y$ appearing in this solution are in general not integers and hence the ordinary tables of Bessel functions are inapplicable.

The baseband filter accepts the components of the error which have frequencies in the range:

$$-F_b \leq F_b(c + n\lambda) \leq F_b \quad (33)$$

or

$$-\frac{1+c}{\lambda} \leq n \leq \frac{1-c}{\lambda}. \quad (34)$$

The interfering wave in the baseband filter output is then

$$\delta_0(t) = \frac{Q}{P} F_b \sum_{n=n_1}^{n_2} (c + n\lambda) A_n \cos[2\pi F_b(c + n\lambda)t - \theta] \quad (35)$$

where n_1 is the smallest integer not less than $-(1+c)/\lambda$ and n_2 is the largest integer not greater than $(1-c)/\lambda$. It would be convenient at this point to assume that Δl^2 is expressible directly in terms of $\bar{\delta}_0^2(t)$. However, there is reason to believe that such an assumption is pessimistic especially at the higher bandwidths where the disturbance $\delta_0(t)$ may never reach its maximum values in the neighborhood of the actual slicing instant. A complete investigation requires a study of the instantaneous wave form of $\delta_0(t)$ in the neighborhood of the slicing instant. We note that if the slicer operates at the trailing edge, the unperturbed slicing instant is $t = \frac{1}{2F_b} + m/f_r$, and the value of the disturbance at that instant is:

$$\delta_0\left(\frac{1}{2F_b} + m/f_r\right) = \frac{Q}{P} F_b \sum_{n=n_1}^{n_2} (c + n\lambda) A_n$$

⁴⁷ Watson, Theory of Bessel Functions, Chapter X.

$$\cos \left[2\pi(c + n\lambda) \left(\frac{1}{2} + m \frac{F_b}{f_r} \right) - \theta \right], \quad m = 0, \pm 1, \pm 2, \dots \quad (36)$$

Averaging the square over all values of $\left[2m\pi \frac{F_b}{f_r} - \theta + c\pi \right]$, which may be treated as a randomly distributed angle, we find an expression for $\overline{\delta_0^2}$ averaged over the values at the slicing instants and not over all time, viz.:

$$\overline{\delta_0^2} = \frac{Q^2}{2P^2} F_b^2 (R^2 + X^2) \quad (37)$$

$$\text{where} \quad R = \sum_{n=n_1}^{\infty} (c + n\lambda) A_n \cos \left[n\pi\lambda \left(1 + \frac{2mF_b}{f_r} \right) \right] \quad (38)$$

$$X = \sum_{n=n_1}^{\infty} (c + n\lambda) A_n \sin \left[n\pi\lambda \left(1 + \frac{2mF_b}{f_r} \right) \right] \quad (39)$$

Then

$$\overline{\Delta t^2} = \frac{2Q^2(R^2 + X^2)}{\pi^2 P^2 \beta^2} \quad (40)$$

and

$$\frac{S}{I} = \frac{\pi P}{Q} \left(\frac{B}{2F_b} - 1 \right) \left(\frac{F_r}{2Nf_r} - 1 \right) (R^2 + X^2)^{-1/2} \quad (41)$$

For each value of B , the value of S/I should be computed over a range of values of c ; and the lowest value of S/I , corresponding to the most unfavorable allocation of the CW frequency, plotted as a point on the curve. The curve of Fig. 12 was not calculated in this way but estimated from the simpler solution existing when the pulses are contiguous.

When the interference is from a similar system, we substitute for the interfering wave:

$$V_2(t) = Q \cos [2\pi f_1 t + \phi(t - \tau) + \theta] \quad (42)$$

This gives

$$\delta(t) = \frac{Q}{2\pi P} \frac{d}{dt} \sin [\phi(t) - \phi(t - \tau) - \theta + 2\pi(f_1 - f_2)t] \quad (43)$$

We distinguish between the cases of overlapping and non-overlapping pulses.

If the pulses do not overlap, we take the origin of time midway between pulses and write

$$\frac{\phi(t) - \phi(t - \tau)}{\pi\beta} = \left\{ \begin{array}{l} -\frac{1}{F_b}, -\frac{1}{2Nf_r} < t < -\frac{\tau}{2} - \frac{1}{F_b} \\ t + \frac{\tau}{2} + \frac{1}{\pi F_b} \sin \pi F_b \left(t + \frac{\tau}{2} \right), -\frac{1}{F_b} < t + \frac{\tau}{2} < \frac{1}{F_b} \\ \frac{1}{F_b}, \frac{1}{F_b} - \frac{\tau}{2} < t < \frac{\tau}{2} - \frac{1}{F_b} \\ -t + \frac{\tau}{2} - \frac{1}{\pi F_b} \sin \pi F_b \left(t - \frac{\tau}{2} \right), -\frac{1}{F_b} < t - \frac{\tau}{2} < \frac{1}{F_b} \\ -\frac{1}{F_b}, \frac{1}{F_b} + \frac{\tau}{2} < t < \frac{1}{2Nf_r} \end{array} \right\} \quad (44)$$

If the pulses overlap, we also take the origin midway between pulses, but we then obtain

$$\frac{\phi(t) - \phi(t - \tau)}{\pi\beta} = \left\{ \begin{array}{l} -\frac{1}{F_b}, -\frac{1}{2Nf_r} < t < -\frac{\tau}{2} - \frac{1}{F_b} \\ t + \frac{\tau}{2} + \frac{1}{\pi F_b} \sin \pi F_b \left(t + \frac{\tau}{2} \right), -\frac{1}{F_b} - \frac{\tau}{2} < t < \frac{\tau}{2} - \frac{1}{F_b} \\ t - \frac{1}{F_b} + \frac{2}{\pi F_b} \sin \pi \frac{F_b \tau}{2} \cos \pi F_b t, \frac{\tau}{2} - \frac{1}{F_b} < t < \frac{1}{F_b} - \frac{\tau}{2} \\ -t + \frac{\tau}{2} - \frac{1}{\pi F_b} \sin \pi F_b \left(t - \frac{\tau}{2} \right), \frac{1}{F_b} - \frac{\tau}{2} < t < \frac{1}{F_b} + \frac{\tau}{2} \\ -\frac{1}{F_b}, \frac{1}{F_b} + \frac{\tau}{2} < t < \frac{1}{2Nf_r} \end{array} \right\} \quad (45)$$

In both cases the right-hand members are even functions of t . Hence

$$\sin [\phi(t) - \phi(t - \tau) - \theta] = \sum_{m=1}^{\infty} B_m \cos 2m\pi Nf_r t \quad (46)$$

$$\cos [\phi(t) - \phi(t - \tau) - \theta] = \frac{A_0}{2} + \sum_{m=1}^{\infty} A_m \cos 2m\pi N f_r t \quad (47)$$

The coefficients A_m and B_m are considerably more complicated than in the corresponding CW case.

PAM-FM

An idle N -channel system sends sinusoidal pulses of duration $2/F_b = 1/Nf_r$ which merge into a continuous sinusoidal variation of frequency expressible by

$$f = f_1 + \frac{\beta}{4} \cos \pi F_b t \quad (48)$$

where β is the peak-to-peak frequency swing and f_1 is the mid-frequency. A full load audio tone frequency $\frac{q}{2\pi}$ impressed on one channel produces a series of sinusoidal pulses of varying heights expressed with sufficient accuracy for a large number of channels by

$$F = f_1 + \frac{\beta}{2} g_0(t) \cos qt \quad (49)$$

where $g_0(t)$ represents a pulse of unit height and duration $\frac{2}{F_b}$ repeated periodically at the frame frequency f_r . Aperture effect is neglected in this approximation. We may expand $g_0(t)$ in a Fourier series:

$$g_0(t) = \frac{G_0}{2} + \sum_{m=1}^{\infty} G_m \cos 2m\pi f_r t \quad (50)$$

where

$$G_m = 2f_r \int_{-1/2Nf_r}^{1/2Nf_r} g_0(t) \cos 2m\pi f_r t dt \quad (51)$$

When a rectangular gate of full-channel duration $1/Nf_r$ is used between the baseband filter and the audio output filter for the channel, F represents the input to the channel filter. The only term passed by the latter is

$$F_q = \frac{\beta}{4} G_0 \cos qt \quad (52)$$

$$G_0 = 2f_r \int_{-1/2Nf_r}^{1/2Nf_r} \frac{1}{2}(1 + \cos 2\pi N f_r t) dt = \frac{1}{N} \quad (53)$$

Therefore the peak sine wave channel output is $\beta/4N$, and the mean square value is $\beta^2/32N^2$. If a gating function $g_1(t)$ is used instead of a rectangular gate, we replace $g_0(t)$ by $g_0(t)g_1(t)$ in the calculation of G_0 .

When the interference is fluctuation noise of mean power $P_n df$ in bandwidth df , the mean square frequency error in the output of the frequency detector is $w_n(f)df$ at frequency f , where

$$w_n(f) = P_n f^2 / W_c \quad (54)$$

The baseband filter accepts the portion of this spectrum between $f = 0$ and $f = F_b$.

The action of the rectangular gate on this spectrum may be calculated by multiplication of the typical spectral component by the gating function:

$$G(t) = \frac{a_0}{2} + \sum_{m=1}^{\infty} a_m \cos 2\pi m f_r t \quad (55)$$

where for $G(t) = 1$ throughout the channel allotment time,

$$a_m = 2f_r \int_{-1/2Nf_r}^{1/2Nf_r} \cos 2\pi m f_r t \, dt = \frac{2 \sin \frac{m\pi}{N}}{m\pi} \quad (56)$$

Each harmonic of $G(t)$ beats with the noise spectrum on either side to produce audio components which sum up to total mean square audio noise:

$$W_n = \frac{1}{4} \int_0^{f_r/2} \left[a_0^2 w_n(f) + \sum_{m=1}^{2N} a_m^2 w_n(mf_r + f) + \sum_{m=1}^{2N} a_m^2 w_n(mf_r - f) \right] df \quad (57)$$

The summations stop at $2N$ because the baseband filter cuts off at $f = F_b = 2Nf_r$. The contribution of the a_0 term is negligible. Then

$$W_n = \frac{P_n}{\pi^2 W_c} \sum_{m=1}^{2N} \frac{\sin^2 \frac{m\pi}{N}}{m^2} \int_0^{f_r/2} [(mf_r + f)^2 + (mf_r - f)^2] df \quad (58)$$

$$= \frac{P_n f_r^3}{\pi^2 W_c} \sum_{m=1}^{2N} \left(1 + \frac{1}{12m^2} \right) \sin^2 \frac{m\pi}{N}.$$

When N is large, the sum approaches

$$W_n \cong \frac{NP_n f_r^3}{\pi^2 W_c} \quad (59)$$

and

$$(S/N)^2 = \frac{\pi^2 W_c}{32N^3 P_n f_r} \left(\frac{\beta}{f_r} \right)^2 \quad (60)$$

$$S/N = \frac{\pi}{4} \left(\frac{B}{Nf_r} - 4 \right) \sqrt{\frac{kB}{2Nf_r}} \quad (61)$$

on substituting $W_c = kP_n B$, $\beta = B - 2F_b$, and $F_b = 2Nf_r$. Equation (61) with $k = 16$ was used to obtain the marginal power curve of Fig. 13. The result may be compared with that given by Rauch⁴⁸ (See also Landon⁴) for a rectangular pulse and rectangular gate, which requires a higher value for F_b . The two systems give the same signal-to-noise ratio when the rectangular pulse and gate of Rauch's system endure for one half of the total channel allotment time. The value of F_b necessary for such a case was estimated by Rauch to be $3.5Nf_r$. A calculation made as above, except that the gate was assumed sinusoidal instead of rectangular, showed very nearly the same signal-to-noise ratio.

The case of CW interference with all channels idle is represented by the r.f. wave:

$$V(t) = P \cos \left(2\pi f_1 t + \frac{\beta}{2F_b} \sin \pi F_b t \right) + Q \cos 2\pi f_2 t. \quad (62)$$

When Q/P is small, the detected frequency is

$$\begin{aligned} f &= f_1 + \frac{\beta}{4} \cos \pi F_b t - \frac{Q}{2\pi P} \frac{d}{dt} \sin \left[2\pi (f_1 - f_2) t + \frac{\beta}{2F_b} \sin \pi F_b t \right] \\ &= f_1 + \frac{\beta}{4} \cos \pi F_b t - \delta(t). \end{aligned} \quad (63)$$

By Fourier series expansion followed by differentiation, the error $\delta(t)$ may be written as:

$$\delta(t) = \frac{Q}{P} \sum_{m=-\infty}^{\infty} \left(f_1 - f_2 + \frac{mF_b}{2} \right) J_m(x) \cos 2\pi \left(f_1 - f_2 + \frac{mF_b}{2} \right) t, \quad (64)$$

where $x = \beta/2F_b$. The baseband filter passes only those components of $\delta(t)$ which have frequencies in the range $-F_b$ to F_b . Writing $c = (f_1 - f_2)F_b$, we find:

$$\delta_0(t) = \frac{QF_b}{P} \sum_{m=m_1}^{m_2} \left(c + \frac{m}{2} \right) J_m(x) \cos 2\pi F_b \left(c + \frac{m}{2} \right) t \quad (65)$$

where m_2 is the largest integer which does not exceed $2(1 - c)$ and m_1 is the smallest integer which is not exceeded by $-2(1 + c)$. The term integer is here understood to include zero and both positive and negative integers. The wave $\delta_0(t)$ is next multiplied by the gating function $G(t)$ and the components falling in the audio band $-f_r/2$ to $f_r/2$ selected. We find:

$$G(t)\delta_0(t) = \frac{F_b Q}{2P} \sum_{m=m_1}^{m_2} \sum_{n=-\infty}^{\infty} a_n \left(c + \frac{m}{2} \right) J_m(x) \cos 2\pi [(2c + m)N + n] f_r t \quad (66)$$

with $a_n = a_{-n}$.

⁴ Loc. cit.

⁴⁸ L. L. Rauch, Fluctuation Noise in Pulse-Height Multiplex Radio Links, Proc. I.R.E., Vol. 35, Nov. 1947, pp. 1192-1197.

For each value of m , there is only one value of n satisfying the audio filter inequality, which may be written:

$$-\frac{1}{2} - (2c + m)N < n < \frac{1}{2} - (2c + m)N \quad (67)$$

Hence interference accepted by the channel filter is:

$$I(t) = \frac{F_b Q}{2P} \sum_{m=m_1}^{m_2} a_n \left(c + \frac{m}{2} \right) J_m(x) \cos 2\pi[(2c + m)N + n]f_r t \quad (68)$$

The mean square value of interference is

$$\overline{I^2(t)} = \frac{F_b^2 Q^2}{8P^2} \sum_{m=m_1}^{m_2} a_n^2 \left(c + \frac{m}{2} \right)^2 J_m^2(x). \quad (69)$$

The signal-to-interference ratio is

$$(S/I)^2 = \frac{G_0^2 \beta}{32} [\overline{I^2(t)}]^{-1} \quad (70)$$

or

$$S/I = \frac{G_0 \beta P}{2F_b Q} \left[\sum_{m=m_1}^{m_2} a_n^2 \left(c + \frac{m}{2} \right)^2 J_m^2(x) \right]^{-1/2}. \quad (71)$$

When a rectangular gate of duration equal to the full channel allotment is used, we substitute $a_n = 2(\sin n\pi/N)/n\pi$. We then find that the largest values of mean square interference occur when c is an odd multiple of one fourth. If we set

$$c = -(2r + 1)/4, r = 0, \pm 1, \pm 2, \dots \quad (72)$$

it follows that if N is an even integer,

$$n = (r + \frac{1}{2} - m)N, \quad (73)$$

$$\sin \frac{n\pi}{N} = \sin (r + \frac{1}{2} - m)\pi, \quad (74)$$

$$\left| \sin \frac{n\pi}{N} \right| = 1. \quad (75)$$

Substituting these values in the expression for S/I , we find

$$S/I = \frac{\pi \beta P}{2F_b Q} \left[\sum_{m=r-1}^{r+2} J_m^2(x) \right]^{-1/2}. \quad (76)$$

The value of r is to be chosen as the integer which makes S/I a minimum; i.e., we place the CW frequency at that part of the band where it does the most damage. The curve marked CW(Gate) of Fig. 14 was obtained in this way.

When instantaneous sampling is used instead of a gate, the value of a_n becomes a constant for all values of n of interest and is equal to a_0 . We then find:

$$S/I = \frac{\beta P}{2F_b Q} \left[\sum_{m \geq -2(1+c)}^{\leq 2(1-c)} \left(c + \frac{m}{2} \right)^2 J_m^2(x) \right]^{-1/2}. \quad (77)$$

Here c is to be selected to make S/I minimum for each value of x . The poorest values of S/I occur when Bessel functions of comparable order and argument appear in the summation, which means that c is in the neighborhood of $-x/2$. The corresponding difference between the CW and mid-carrier frequencies is one-fourth the peak-to-peak swing.

To calculate the interference between two similar idle systems, we set

$$V(t) = P \cos \left(2\pi f_1 t + \frac{\beta}{2F_b} \sin \pi F_b t \right) + Q \cos \left[2\pi f_2 t + \frac{\beta}{2F_b} \sin (\pi F_b t - \theta) \right]. \quad (78)$$

The frequency error registered in the first system is then

$$\begin{aligned} \delta(t) &= \frac{Q}{2\pi P} \frac{d}{dt} \sin [2\pi(f_1 - f_2)t + x \sin \pi F_b t - x \sin (\pi F_b t - \theta)] \\ &= \frac{Q}{2\pi P} \frac{d}{dt} \sin \left[2\pi(f_1 - f_2)t + 2x \sin \frac{\theta}{2} \cos \left(\pi F_b t - \frac{\theta}{2} \right) \right] \\ &= \frac{Q}{P} \sum_{m=-\infty}^{\infty} \left(f_1 - f_2 + \frac{mF_b}{2} \right) J_m \left(2x \sin \frac{\theta}{2} \right) \\ &\quad \cdot \cos \left[2\pi \left(f_1 - f_2 + \frac{mF_b}{2} \right) t - \frac{m\theta}{2} \right]. \end{aligned} \quad (79)$$

It follows that the response of the baseband filter is

$$\delta_0(t) = \frac{QF_b}{P} \sum_{m=m_1}^{m_2} \left(c + \frac{m}{2} \right) J_m \left(2x \sin \frac{\theta}{2} \right) \cos \left[2\pi F_b \left(c + \frac{m}{2} \right) t - m \frac{\theta}{2} \right]. \quad (80)$$

The effects of the channel gate and filter are computed in the same way as for CW, giving the audio interference:

$$I(t) = \frac{F_b Q}{2P} \sum_{m=m_1}^{m_2} a_n \left(c + \frac{m}{2} \right) J_m \left(2x \sin \frac{\theta}{2} \right) \cdot \cos \left(2\pi[(2c + m)N + n]f_r t - m \frac{\theta}{2} \right) \quad (81)$$

Since the two systems occupy the same frequency assignment, we assume that c is not greatly different from zero. Then for fixed θ :

$$\overline{I^2(t)} = \frac{F_b^2 Q^2}{32P^2} \sum_{m=m_1}^{m_2} m^2 a_n^2 J_m^2 \left(2x \sin \frac{\theta}{2} \right) \quad (82)$$

Since θ , the frame phase difference, is a random angle we average over its possible values by setting:

$$\begin{aligned} A_m(x) &= \frac{1}{2\pi} \int_0^{2\pi} J_m^2 \left(2x \sin \frac{\theta}{2} \right) d\theta = \frac{1}{\pi} \int_0^\pi J_m^2(2x \sin \phi) d\phi \\ &= \frac{\Gamma^2(m + \frac{1}{2})(2x)^{2m}}{\pi(2m)!(m!)^2} \\ &\quad \cdot {}_2F_3\left(m + \frac{1}{2}, m + \frac{1}{2}; 2m + 1, m + 1, m + 1; -4x^2\right), \\ &\quad m \geq 0 \end{aligned} \quad (83)$$

$$A_{-m}(x) = A_m(x)$$

Noting further that for $c = 0$, $m_1 = -2$, $m_2 = 2$, and $n = -mN$, we have then:

$$I^2(t) = \frac{F_b^2 Q^2}{16P^2} [a_N^2 A_1(x) + 4a_{2N}^2 A_2(x)] \quad (84)$$

and

$$S/I = \frac{G_0 \beta P}{F_b Q \sqrt{2}} [a_N^2 A_1(x) + 4a_{2N}^2 A_2(x)]^{-1/2}. \quad (85)$$

For a sinusoidal pulse and rectangular gate of full channel allotment time in duration, $a_N = a_{2N} = 0$, $I^2(t) = 0$, and S/I is infinite. If instantaneous sampling is used,

$$G_0 = a_N = a_{2N} \quad (86)$$

and

$$S/I = \frac{\beta P}{\sqrt{2} F_b Q} [A_1(x) + 4A_2(x)]^{-1/2} \quad (87)$$

The curve for similar system interference with instantaneous sampling, Fig. 14, was calculated from Eq. (87).

For small values of x , we may use the ascending power series:

$$A_1(x) = \frac{x^2}{2} \left[1 - \frac{3^2}{3 \cdot 2^2 \cdot 1!} x^2 + \frac{3^2 \cdot 5^2}{3 \cdot 4 \cdot 2^2 \cdot 3^2 \cdot 2!} x^4 - \dots \right] \quad (88)$$

$$A_2(x) = \frac{3x^4}{32} \left[1 - \frac{5^2}{5 \cdot 3^2 \cdot 1!} x^2 + \frac{5^2 \cdot 7^2}{5 \cdot 6 \cdot 3^2 \cdot 4^2 \cdot 2!} x^4 - \dots \right] \quad (89)$$

For large values of x , the following asymptotic formula has been derived by Mr. S. O. Rice by use of the Mellin-Barnes contour integral representation and the method of steepest descents:

$$\pi^2 x A_m(x) \sim \ln x + \ln 32 + \gamma - 2 \left(1 + \frac{1}{3} + \frac{1}{5} + \cdots + \frac{1}{2m-1} \right) - (-)^m \frac{\pi^2}{4} \sqrt{\frac{\pi}{2x}} \cos \left(4x + \frac{\pi}{4} \right) + \cdots \quad (90)$$

$\gamma = \text{Euler's constant} = 0.5772 \dots$

As x approaches zero, S/I approaches $2P/Q$, which is to be expected because the frequency deviation of the unwanted carrier is represented by a pair of first order sidebands P/Q times as great as those on the wanted carrier. Averaging over the random carrier phase difference brings in a factor $\sqrt{2}$, and averaging over all frame phases accounts for another.

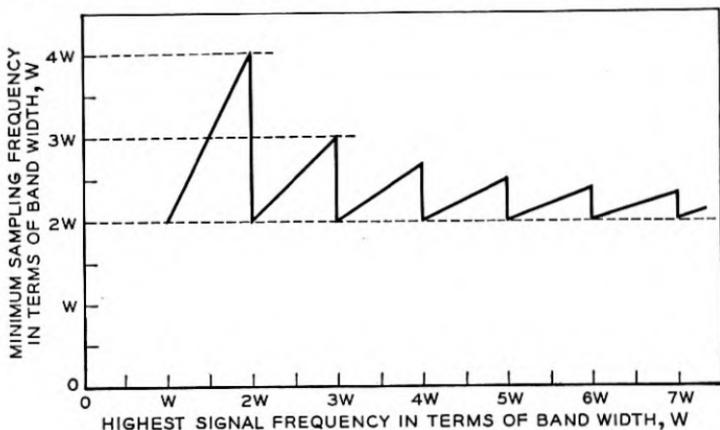


Fig. 37—Minimum sampling frequency for band of width W .

APPENDIX V

SAMPLING A BAND OF FREQUENCIES DISPLACED FROM ZERO

It is often necessary to transmit a signal band which does not extend all the way down to zero frequency. For example, a group of channels in FDM may be based on a set of carrier frequencies remote from zero. When we consider the application of pulse methods to transmit such a signal, the question of what sampling rate is needed immediately arises. A band extending from f_1 to $f_1 + W$ could of course be translated to the range 0 to W by standard modulation techniques, sampled at a rate $2W$, and restored to the original range by an inverse translation at the receiver. The frequency shifting apparatus required includes modulators, carrier generators, band separating filters, and possibly amplifiers to make up the inevitable losses.

A direct sampling process which avoids shifting the band may therefore be preferred. A useful theorem for uniformly spaced samples is that the minimum sampling frequency is not in general twice the highest frequency in the band but is given by the formula:

$$f_r = 2W \left(1 + \frac{k}{m} \right), \quad (1)$$

where:

f_r = minimum sampling frequency

W = width of band

f_2 = highest frequency in band

m = largest integer not exceeding f_2/W

$$k = \frac{f_2}{W} - m$$

The value of k in (1) varies between zero and unity. When the band is located between adjacent multiples of W , we have $k = 0$ and it follows that $f_r = 2W$ no matter how high the frequency range of the signal may be. As k increases from zero to unity the sampling rate increases from $2W$ to $2W \left(1 + \frac{1}{m} \right)$. The curve of minimum sampling rate versus the highest frequency in a band of constant width thus becomes a series of sawteeth of successively decreasing height as shown in Fig. 37. The highest sampling rate is required when $m = 1$ and k approaches unity. This is the case of a signal band lying between $W - \Delta f$ and $2W - \Delta f$ with Δf small. The sampling rate needed is $2(2W - \Delta f)$ which approaches the value $4W$ as Δf approaches zero. Actually when $\Delta f = 0$, we change to the case of $m = 2$, $k = 0$, and $f_r = 2W$. The next maximum on the curve is $3W$, which is approached when f_2 nears $3W$. The successive maxima decrease toward the limit $2W$ as f_2 increases. The sampling theorem contained in Eq. (1) may be verified from steady state modulation theory by noting that the first order sidebands on harmonics of $2W$ do not overlap the signal when the equation is satisfied.

Abstracts of Technical Articles by Bell System Authors

*The Transistor—A New Semiconductor Amplifier.*¹ J. A. BECKER and J. N. SHRIVE. This article describes the construction, characteristics, and behavior of the newly discovered device, the transistor. Used as a semiconductor amplifier, it works on an entirely different principle and is capable of performing the same tasks now done by the vacuum tube triode.

*A Review of Magnetic Materials.*² R. A. CHEGWIDDEN. Significant advances have been made within recent years in the development of new and better magnetic materials, and in the theories of magnetism. High permeability materials that may be classed as non-conductors, materials with greatly improved initial permeabilities, and permanent magnet alloys capable of storing four or five times as much energy as those obtainable ten years ago are now available. Descriptions of some of these developments are given. The paper gives compilations of data and curve sheets showing some of the typical characteristics of many of these materials.

*Ratio of Frequency Swing to Phase Swing in Phase- and Frequency-Modulation Systems Transmitting Speech.*³ D. K. GANNETT and W. R. YOUNG. Computed and measured data are presented bearing on the relation between the phase and the frequency swing in phase- and frequency-modulation systems when transmitting speech. The results were found to vary with different voices, with the microphone and circuit characteristics, and with the kind of volume regulation used. With a particular carbon microphone, it was found that a phase deviation of 10 radians corresponds to a frequency deviation of between 11 and 15 kc in a phase-modulation system, and between 6 and 12 kc in a frequency-modulation system, depending on conditions.

*Design and Performance of Ethylene Diamine Tartrate Crystal Units.*⁴ J. P. GRIFFIN and E. S. PENNELL. A research program on synthetic crystals has resulted in the development and adoption of EDT for carrier telephone filters. Some unusual physical properties of the crystalline material give rise to novelty in the processing methods and mechanical design of the units. These properties include anisotropic expansion coefficients, fragility, natural cleavages and water solubility. The electrical properties of EDT result in filters with wider pass bands and lower impedance levels than commercially obtainable with quartz.

¹ *Electrical Engineering*, v. 68, pp. 215-221, March 1949.

² *Metal Progress*, November 1948.

³ *Proc. I. R. E.*, v. 37, pp. 258-263, March 1949.

⁴ *A. I. E. E. Transactions*, v. 67, pt. 1, pp. 557-561, 1948.

*Recent Improvements in Loading Apparatus for Telephone Cables.*⁵ S. G. HALE, A. L. QUINLAN, and J. E. RANGES. Through the use of improved materials, manufacturing techniques, and designs, a series of exchange-area loading coils has been provided which is equivalent electrically to the superseded types but requires one-third less copper and has considerably smaller overall dimensions. Similarly, 3-coil toll cable loading units have been provided with a saving of one-half in both copper and core material, with a small sacrifice in electrical behavior as compared with superseded types. The reduced size of the new coils and units, together with improved assembly arrangements, made possible a 65 per cent saving in the volume and weight of the cases housing them.

*The Coaxial Transistor.*⁶ WINSTON E. KOCK and R. L. WALLACE, JR. The success of the earlier types of transistors led to the exploration of other forms of similar amplifiers, one of which is the coaxial transistor. A description of its construction, characteristics, and many advantages is contained in this article.

*Paralleled-Resonator Filters.*⁷ J. R. PIERCE. This paper describes a class of microwave filters in which input and output waveguides are connected by a number of resonators, each coupled directly to both guides. Signal components of different frequencies can pass from the input to the output largely through different resonators. This type of filter is a realization of a lattice network. An experimental filter is described.

*A Broad-Band Microwave Relay System between New York and Boston.*⁸ G. N. THAYER, A. A. ROETKEN, R. W. FRIIS, and A. L. DURKEE. This paper describes the principal features of a broad-band microwave relay system which has recently been installed between New York and Boston. The system operates at frequencies around 4,000 Mc and provides two two-way channels, each accommodating a signal-frequency band extending from 30 cps to 4.5 Mc. Noise and distortion characteristics are satisfactory for the transmission of several hundred simultaneous telephone conversations or a standard black-and-white television program.

*Growing Crystals of Ethylene Diamine Tartrate.*⁹ A. C. WALKER and G. T. KOHMAN. The need for a synthetic piezoelectric crystal to relieve the critical quartz supply situation has resulted in the development by the Bell Telephone Laboratories of a new organic salt crystal, ethylene diamine tartrate, which is being used in place of quartz in telephone circuits.

This crystal is grown from a supersaturated aqueous solution of its salt by an entirely new method known as the constant temperature process. It

⁵ *A. I. E. E. Transactions*, v. 67, pt. 1, pp. 385-392, 1948.

⁶ *Electrical Engineering*, v. 68, pp. 222-223, March 1949.

⁷ *Proc. I. R. E.*, v. 37, pp. 152-155, February 1949.

⁸ *Proc. I. R. E.—Waves and Electrons Section*, v. 37, pp. 183-188, February 1949.

⁹ *A. I. E. E. Transactions*, v. 67, pt. 1, pp. 565-570, 1948.

differs from previous methods used for growing large single crystals from solution, in that the solution saturated at one temperature is continuously fed into a crystallizer tank maintained at a slightly lower temperature, thus providing the supersaturation condition necessary for crystal growth. Further, the solution is circulated in such a manner that the partially impoverished mother liquor overflows from the growing tank back into the saturator where it is refortified and filtered. It is then heated and returned to the growing tank in such a way as to avoid the formation of undesirable crystal nuclei.

The paper contains a description of the new method which is now in commercial operation, together with a general discussion of some of the important principles involved in the successful growth of large single crystals of water soluble salts.

*Crystal Filters Using Ethylene Diamine Tartrate in Place of Quartz.*¹⁰ E. S. WILLIS. Ethylene diamine tartrate (EDT) crystal filters were developed to replace the earlier quartz type channel filters in the broad-band carrier telephone systems, because of the threatened scarcity of quartz. These new filters give performance comparable to that of the earlier design. The growth of the EDT crystals from seeds and their fabrication into crystal units for use in filters are covered in companion papers on "Design and Performance of Ethylene Diamine Tartrate Crystal Units" and "Growing Crystals of Ethylene Diamine Tartrate" in this same volume of the *Transactions*

¹⁰ A. I. E. E. *Transactions*, v. 67, pt. 1, pp. 552-556, 1948

Contributors to This Issue

JOHN BARDEEN, University of Wisconsin, B.S. in E.E., 1928; M.S., 1930. Gulf Research and Development Corporation, 1930-33; Princeton University, 1933-35, Ph.D. in Math. Phys., 1936; Junior Fellow, Society of Fellows, Harvard University, 1935-38; Assistant Professor of Physics, University of Minnesota, 1938-41; Prin. Phys., Naval Ordnance Laboratory, 1941-45. Bell Telephone Laboratories, 1945-. Dr. Bardeen is engaged in theoretical problems related to semiconductors.

W. R. BENNETT, B.S., Oregon State College, 1925; A.M., Columbia University, 1928. Bell Telephone Laboratories, 1925-. Mr. Bennett has been active in the design and testing of multichannel communication systems, particularly with regard to modulation processes and the effects of nonlinear distortion. He is now engaged in research on various transmission problems.

C. B. FELDMAN, University of Minnesota, B.S. in Electrical Engineering, 1926; M.S., 1928. Bell Telephone Laboratories, 1928-. As Transmission Research Engineer, Mr. Feldman has charge of a group studying new transmission methods. He is a Fellow of the Institute of Radio Engineers.

J. R. HAYNES, B.S. in Physics, University of Kentucky, 1930. Bell Telephone Laboratories, 1930-. Mr. Haynes is in the Physical Research Department, engaged in solid state studies.

CONYERS HERRING, A.B., University of Kansas, 1933; Ph.D., Princeton University, 1937; National Research Fellow, Massachusetts Institute of Technology, 1937-39; Instructor in Mathematics and Research Associate in Mathematical Physics, Princeton University, 1939-40; Instructor in Physics, University of Missouri, 1940-41; Columbia University Division of War Research, 1941-45; Professor of Applied Mathematics, University of Texas, 1946. Bell Telephone Laboratories, 1945-. Dr. Herring has been engaged in theoretical problems in the fields of solid state physics and electron emission.

R. J. KIRCHER, B.S. in E.E., California Institute of Technology, 1929; M.S., Stevens Institute of Technology, 1941. Bell Telephone Laboratories,

1929-. Mr. Kircher was engaged in radar and counter measures projects during the war. Electronic Apparatus Development Department since 1944. Transistor Development Group since 1948.

G. L. PEARSON, A.B., Willamette University, 1926; M.A. in Physics, Stanford University, 1929. Bell Telephone Laboratories, 1929-. Mr. Pearson is in the Physical Research Department where he has been engaged in the study of noise in electric circuits and the properties of electronic semi-conductors.

ROBERT M. RYDER, Yale University, B.S. in Physics, 1937; Ph.D., 1940. Bell Telephone Laboratories, 1940-. Dr. Ryder joined the Laboratories to work on microwave amplifier circuits, and during most of the war was a member of a group engaged in studying the signal-to-noise performance of radars. In 1945 he transferred to the Electronic Development Department to work on microwave oscillator and amplifier tubes for radar and radio relay applications. He is now in a group engaged in the development of transistors.

W. SHOCKLEY, B.Sc., California Institute of Technology, 1932; Ph.D., Massachusetts Institute of Technology, 1936. Bell Telephone Laboratories, 1936-. Dr. Shockley's work in the Laboratories has been concerned with problems in solid state physics.