THE TRANSISTOR

SELECTED REFERENCE MATERIAL

ON

CHARACTERISTICS & APPLICATIONS

PREPARED BY BELL TELEPHONE LABORATORIES, INC.

FOR WESTERN ELECTRIC CO., INC.

NEW YORK, N. Y.

J. S. Breuzler

THE TRANSISTOR

The transistor, the spidery object at the right, has been called the first serious rival of the vacuum tube. With amplifications of 40-50 db the transistor occupies 1/400 as much space as a vacuum tube performing the same job and requires about a millionth of the power supply.

ORRADRE LIBRARY
SANTA CLARA UNIVERSITY
SANTA CLARA, CALIFORNIA

THE TRANSISTOR

SELECTED REFERENCE MATERIAL

ON

CHARACTERISTICS & APPLICATIONS

CONTRACT DA 36-039 SC-5589 (TASK 3)

PREPARED BY BELL TELEPHONE LABORATORIES, INC.

FOR WESTERN ELECTRIC CO., INC.

NEW YORK, N. Y.

PRINTED IN THE UNITED STATES OF AMERICA.

Copyright 1951 By
Bell Telephone Laboratories, Inc.

Articles on pages
7, 17, 25, 81, 83, 97, 103, 111, 127, 165 and 171
are reproduced by permission of the copyright owners.

TK
7872
.T3
B4
1951

FOREWORD

The material included herein covers the same ground and supplements the talks given at a symposium on transistors, which was held at the Bell Telephone Laboratories during the week beginning September 17, 1951.

In the three odd years since the transistor was announced, work has been going on to study and perfect this device in order to make it a useful and reliable tool for general application. With ultimate success along these lines now seemingly assured, it appeared desirable to spread the available information widely in order to aid the military effort.

In this volume, as in the symposium, the treatment follows the general sequence of: theory, transistor properties, circuit design principles, application and, finally, characteristics of different transistor types now under development. In the short time since the invention of the transistor, an enormous amount of information on physics, device properties and circuit applications has evolved. It is the purpose of this volume to provide enough representative material to give a picture of the current status. Included are published technical articles, reports of work carried on under contracts with the military services, and supplementary memoranda. In the symposium, some of the theory presented was drawn from Dr. William Shockley's book "Electrons and Holes in Semiconductors". The pertinent chapters have not been included as they were considered to be too voluminous.

It will be recognized that, in a broad sense, this constitutes but a progress report. The art is moving rapidly. There are still many problems to be solved and some of the material given may be soon out of date. It is too early for a full appraisal of the extent to which the transistor may find its place in the field of electronics. Just now it is a promising infant trying to shed its swaddling clothes. It is hoped that this reference material will assist in this effort.

Bell Telephone Laboratories, Inc.
Murray Hill, N. J.
November 15, 1951

THE TRANSISTOR

SELECTED REFERENCE MATERIAL ON CHARACTERISTICS & APPLICATIONS

CONTENTS

		Page
FOREWORD		
INTRODUCTION.	J. A. Morton	1
HOLES AND ELECTRONS	William Shockley	7
THE PHYSICS OF ELECTRONIC SEMI-CONDUCTORS.	G. L. Pearson	17
THE THEORY OF p-n JUNCTIONS IN SEMI-CONDUCTORS AND p-n JUNCTION TRANSISTORS	W. Shockley	25
THEORY AND EXPERIMENT FOR A GERMANIUM p-n JUNCTION	F. S. Goucher G. L. Pearson, M. Sparks, G. K. Teal, W. Shockley	81
THE p-n JUNCTION TRANSISTORS. . . .	W. Shockley M. Sparks, G. K. Teal	83
THE p GERMANIUM TRANSISTOR. . . .	W. G. Pfann J. H. Scaff	97
THE TRANSISTOR - A NEW SEMI-CONDUCTOR AMPLIFIER	J. A. Becker J. N. Shive	103
THE COAXIAL TRANSISTOR.	W. E. Kock R. L. Wallace, Jr.	111
THE PHOTOTRANSISTOR	J. N. Shive	115
AGING AND TEMPERATURE RESPONSE OF POINT-CONTACT TRANSISTORS	R. M. Ryder	123
DUALITY AS A GUIDE IN TRANSISTOR CIRCUIT DESIGN.	R. L. Wallace, Jr. G. Raisbeck	127

CONTENTS CONT'D

		Page
SOME CIRCUIT ASPECTS OF THE TRANSISTOR	R. M. Ryder R. J. Kircher	165
SOME CIRCUIT PROPERTIES AND APPLICATIONS OF n-p-n TRANSISTORS.	R. J. Wallace, Jr. W. J. Pietenpol	201
GENERAL CONSIDERATIONS CONCERNING NON-LINEAR CIRCUITS AND NEGATIVE RESISTANCE.	B. G. Farley	237
IDEALIZED NEGATIVE RESISTANCE CHARACTERISTICS OF THE TRANSISTORS	R. L. Trent	249
SOME SWITCHING ASPECTS OF TRANSISTORS.	A. E. Anderson	283
SOME SYSTEM APPLICATIONS OF TRANSISTOR AMPLIFIERS.	R. S. Caruthers	335
A NEW HIGH-EFFICIENCY LINEAR AMPLIFIER USING A TRANSISTOR AND A VACUUM TUBE.	G. Raisbeck	377
A HIGH-EFFICIENCY UNTUNED AMPLIFIER.	G. Raisbeck	389
SOME EXPERIMENTAL AND PRACTICAL APPLICATIONS OF TRANSISTOR OSCILLATORS.	R. S. Caruthers	397
MODULATORS IN CARRIER TELEPHONE SYSTEMS.	R. S. Caruthers	415
A STABILIZED TRANSISTOR DELAY AND SWITCHING CIRCUIT.	A. E. Anderson	429
A TRANSISTOR PACKAGED CIRCUIT FOR MULTIFUNCTIONAL SWITCHING APPLICATIONS	R. L. Trent	437

CONTENTS CONT'D

		Page
STABILIZED SINGLE TRANSISTOR BINARY COUNTER	R. L. Trent	463
A STABILIZED GENERAL PURPOSE TWO-TRANSISTOR BINARY COUNTER	R. L. Trent	469
A TRANSISTOR REVERSIBLE BINARY COUNTER	R. L. Trent	483
FOUR STAGE BINARY COUNTER	R. L. Trent	501
DIODE GATES	L. W. Hussey	507
ANTICOINCIDENCE OR "NOT AND" GATE	L. W. Hussey	517
STORAGE AND ADDITION OF BINARY NUMBERS	J. R. Harris	523
A PACKAGED ANGULAR POSITION ENCODER	H. G. Follingstad	583
A GRAY TO BINARY TRANSLATOR AND SHIFT REGISTER.	R. E. Yaeger	611
NOTES ON THE DESIGN OF HIGH SPEED DIGITAL COMPUTERS USING TRANSISTORS	J. H. Felker	627
DATA ON EXPERIMENTAL TRANSISTOR TYPES	J. J. Kleimack R. M. Ryder, R. J. Kircher, John N. Shive	753

I N T R O D U C T I O N

J. A. MORTON

INTRODUCTION

When the transistor was announced about three years ago, it was felt that a new departure in communication techniques had come into view. Here was a mechanically simple device which could perform many of the amplification functions over which the electron tube had long held a near monopoly. The device was small, required no heater power, and was potentially very rugged; moreover, it consisted of materials which might be expected to last indefinitely long, and it did not appear to be too complicated to make.

However, as might be expected for a newly invented electron device, the practical realization of these promises still required the overcoming of considerable obstacles. While the operation of the first devices was well understood in a general way, several items were limiting and puzzling, for example:

a) Units intended to be alike varied considerably from each other - the reproducibility was bad.

b) In an uncomfortably large fraction of the exploratory devices, the properties changed suddenly and inexplicably with time and temperature whereas others exhibited extremely stable characteristics - the reliability was poor.

c) It was difficult to use the theory and the then-existing undeveloped technology to design and develop devices to a varied range of electrical characteristics needed for different circuit functions - performance characteristics were limited with respect to gain, noise figure, frequency range and power, - the designability was poor.

Before the transistor could be regarded as a commercially practical circuit element, it was necessary to find out the causes of these limitations, to understand the theory and develop the technology further in order to produce and control more desirable characteristics.

Over the past year and one-half, measurable progress has been made in reducing, but not eliminating, the three mentioned limitations.

These advances have been obtained through an improved understanding, improved processes and improved germanium materials. As a result:

a) the beginnings of method have evolved in the use of the theory to explain and predict the network characteristics of transistors in terms of physical structure and material properties.

b) It is now possible to evaluate some of the effects and physical meaning of empirically derived processes and to thereby devise better methods subject to control.

c) As a result, on an exploratory development level, it is now possible to make transistors in the laboratory to several sets of prescribed characteristics with usable tolerances and satisfactory yields.

d) Such transistors are greatly improved over the old ones in so far as life and ruggedness and some reduction in temperature dependence has been achieved. However, it is not to be inferred that all reliability problems are solved.

e) It has become possible in the laboratory to explore some of the consequences of the theory experimentally such that point contact devices with new ranges of performance are indicated. Even more importantly, new p-n junction devices have been built in the laboratory and these junction devices have indicated an extension in performance characteristics in some respects.

f) By having interchangeable and reliable devices with a wider range of characteristics, it has become possible to carry on exploratory circuit and system applications on a more realistic basis.

This course represents the belief that enough advance has been made in reducing the original limitations to warrant further serious exploratory circuit applications. It does not imply that these limitations have been completely reduced nor that the full range of possible performance characteristics has been achieved. But it will become apparent from the other reference material that a considerable range of circuit applications is now possible. It is believed that more rapid development will be possible under the stimulus of a critical evaluation arising from a still wider range of circuit and system work.

Systems Aspects of Transistors

In the last analysis the only reasons for considering transistors over vacuum tubes for any electronic applications have to be a combination of superior performance, miniaturization, reliability or economy. To a great extent it is much too early to get a clean cut answer to such a comparison for we are comparing a 40-year-old giant with a 3-year-old child. On the other hand, unless an attempt is made to answer such questions now on the basis of extrapolating present reality to future potentialities, the child may never grow or develop to these possible future potentialities. What we are really asking is: Can we now build, or can we expect after suitable development to be able to build, electronic communication and control systems with transistors which are superior to tube systems from the point of view of a) actual performance, b) miniaturization in space and power, c) reliability, d) economy.

In CW transmission systems the following functions and figures of merit are of prime importance:

Amplification

Figure of Merit	Tube	Transistor
Gain/Stage	20-40 db.	20-40 db.
Noise Figure	0 to 30 db.	10 to 50 db at 1 kc
Frequency Limit	0 to 60,000 mc	0 to 30 mc
Power Output	0 to kilowatts	0 to 2 watts
Class A Efficiency	35%	35-49%
Class B Efficiency	79%	> 80%

Generation

Figure of Merit	Tube	Transistor
f_o max.	0 to 60,000 mc	0 to 300 mc
Efficiency	60-70%	>70%
P_o	0 to 10^5 watts	0 to 10 watts

On a straight performance basis then, CW transmission functions are clearly performed better by tubes for frequencies greater than 30 mc. Below some 10-20 mc, however, it seems that transistors are the equals of tubes in most categories and in some places superior except for extremely high powers. In pulse transmission and control (including switching and computing) quantitative figures of merit are not available for either tubes or transistors simply because they have not yet been defined for this field of systems. However, in so far as comparisons can be made for such functions as pulse generation, pulse counting, pulse storage, gating, pulse amplification, and pulse delay it seems that transistors can perform equally well to tubes for pulse lengths >0.5 μ sec. and repetition rates up to 10^6/sec. In systems utilizing semiconductor diodes for logical operations transistors have some system advantages over tubes because they are basically of the same physical nature as semiconductor diodes and coupling and switching problems are somewhat simplified.

In so far as the actual volume occupied by the electron devices is concerned we note the following comparisons:

Tubes	Transistors
1 to 1/8 cu. in.	1/50 to 1/2000 cu. in.

This, of course, is not the whole story and only further study of tube and transistor systems can answer this, but indications are that at least an overall saving of from 5-10 to one can be currently realized.

The relative power supplies for small signal applications are indicated below:

Tubes			Point Contact Transistors	Junction Transistors
Heater	Plate	Total	Total	Total
25 to 1000 mw.	25 to 1000 mw.	50 to 2000 mw.	4 mw to 50 mw	1 μw to 100 μw

With regard to reliability, extrapolation of aging curves indicate a life of the order of 75,000 - 100,000 hours, well above most vacuum tube types. Transistors will stand shocks in excess of 30,000 g and there is no indication of modulation in vibration tests up to 1000 g. Temperature effects are, however, greater in the case of transistors thereby limiting their use to temperatures less than 80°C.

Tentative Conclusions

In pulse systems (< 1 mc repetition rate) transistors should be considered seriously because of their marked superiority in miniature space and power, bearing in mind that in some reliability figures they are superior whereas in the matter of temperature dependence they are inferior (though certainly as good as single contact diodes and moreover they generate less heat).

In cw applications at frequencies less than 10 megacycles, essentially the same results are now possible primarily because of junction transistors. However, in some cases point contact transistors are adequate. In the range from 10-100 mc tubes are currently superior in every performance figure except perhaps noise and bandwidth so that for transistors to be considered seriously for such systems much greater premium must be placed on miniaturization and reliability than for the other aforementioned systems applications.

Thus, it is believed that even though there are many outstanding problems of a circuit and device nature yet to be solved, it is time for the circuit and systems engineers to explore seriously the application possibilities of transistors not only in the hope of building better systems but also to influence transistor development towards those most important systems needs for which their intrinsic potentialities best fit them.

It will develop that transistors, because of their essentially different physical structure, turn out to be unique in their circuit properties. It will become apparent that it is not often possible to simply substitute for a vacuum tube a suitable transistor and have even a working circuit let alone an equivalent system. Accordingly, it becomes necessary to re-examine requirements in terms of our communication techniques by going back to the starting point and asking anew:

1) What are the essential system functions to be performed and what requirements do they place upon the circuit blocks which produce them.

2) Being fully aware of the circuit characteristics of transistors, what kind of network arrays are required and what specifications do these circuits demand from transistors.

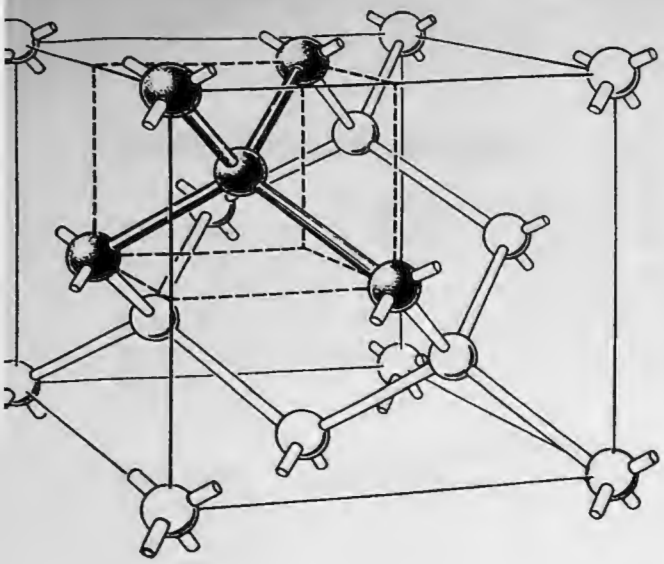
3) What circuit characteristics of transistors are currently available and how do they further satisfy the reliability requirements of modern communications systems.

H O L E S A N D E L E C T R O N S

WILLIAM SHOCKLEY

PHYSICS TODAY VOL.3 pp.16-24 OCTOBER 1950

Produced outside the work performed under this contract



Germanium, the crystal of the transistor, has the crystal structure of diamond.
(Illustration courtesy Bell System Technical Journal, July, 1949, p. 338.)

Holes and Electrons

by ***William Shockley***

Some new experiments in transistor electronics are described here in which concepts suggested by theory have been verified directly by experiment.

Recently a far more intimate and detailed view of the basic processes by which electrons carry current has been made possible by new experiments based upon the invention of the transistor.

Electronic conduction, in which only the electrons move while the atoms or ions remain in fixed positions, is preferred in most electrical applications since it does not cause the gross chemical or physical changes produced by ionic conduction. Aside from their intrinsic usefulness, the new experiments provide another example of the intimate connection between the basic and the applied in science, for here fundamental scientific knowledge, which has stimulated technological developments, has in turn been helped directly by the developments it stimulated.

The transistor was invented by J. Bardeen and W. H. Brattain of the Bell Telephone Laboratories. They showed that when two rectifying contacts, in close proximity to one another, are made to a germanium crystal, the electronic structure in the germanium may be changed by passing current forward through one of the points, and that this change may be detected by drawing reverse current through the other point. The effects are so pronounced that, as is well known now, power gain may result and the device may be used as an amplifier or for many of the other purposes for which a vacuum tube is usually employed. The description of the transistor as a useful circuit element has been given so many times that I shall not repeat it here (see, for example, *Physics Today* for August, 1948, page 22), but work on this device has made possible descriptive experiments on various phases of electronic conduction. These experiments have experimentally supported parts of the physical picture of electronic conduction which were based previously on theoretical considerations alone.

This paper was the subject of the 19th Joseph Henry Lecture delivered before the Philosophical Society of Washington on May 13, 1950. It presents in a simplified form a portion of the author's book "Electrons and Holes in Semiconductors, with Applications to Transistor Electronics" which is planned for publication this fall as one of the Bell Telephone Laboratories Series. *William Shockley* acquired his interest in solid state physics at MIT while carrying out his doctorate research under J. C. Slater. He has continued in this field at Bell Telephone Laboratories except for interruptions for war work. At the end of World War II, the Solid State Group at Bell Telephone Laboratories was organized with Dr. Shockley as a cosupervisor. Special emphasis was placed on semiconductor research because of the possibility of making amplifiers, an aim which was achieved with the transistor and which led to the experiments in transistor electronics described in this article.

The physical picture of electronic conduction resulting from investigations of silicon and germanium in connection with the development of crystal rectifiers during the war, and still earlier theories, describes the conduction process in considerable detail. Prior to the new experiments (and I shall consider only germanium for which the story is most complete) only certain over-all conclusions from the theory could be verified directly. In P. W. Bridgman's sense of the word "operational" there were many conceptual aspects of the picture which had no operational verification. The new experiments give a real operational meaning to a large number of the concepts which formerly entered only as intermediate steps in the calculations before theory was compared with experiment. In addition to giving this deeper and more intimate insight into the basic mechanisms the new results furnish information of direct applicability in problems of understanding and designing transistor devices.

The Physical Picture

Like carbon, germanium is tetravalent and crystallizes in the diamond structure, in which each atom is surrounded by four neighbors with which it forms electron-pair bonds, all making equal angles with each other. The valence-bond system may be regarded as a rigid structure which holds the atoms in place and provides a role for each of the four valence electrons of the germanium atoms. Since the electrons are occupying all of the possible bond wave functions, there is no possibility of producing current by displacing them while keeping the valence-bond structure intact.

If a germanium crystal or a diamond crystal is illuminated by light with sufficiently energetic photons, then the valence-bond structure may be temporarily disturbed. A high energy photon will eject an electron from one of the bonds and this electron will then be free to move about the crystal. The hole left behind can also move since electrons in adjoining bonds may move into it, thus producing a net displacement of the hole.

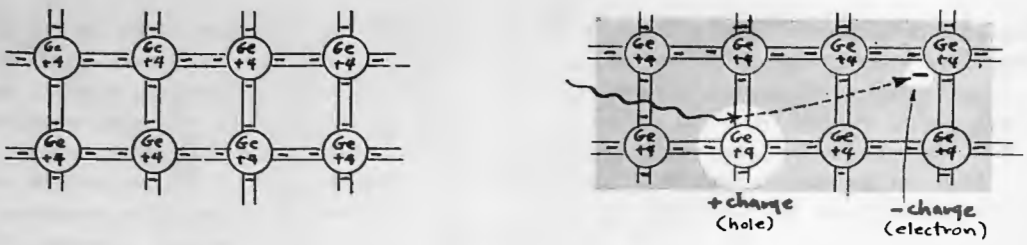
According to the accepted theoretical pictures the excess electron in a semiconductor, or simply the electron, as it is usually called, can be thought of as behaving much like an electron in free space. It can move through the crystal lattice for great distances without being appreciably deflected by the fields due to the atoms of the lattice. The reason that it is not deflected each time it travels through a unit cell is that it moves as a wave. If the crystal were perfectly periodic, this wave could travel indefinitely without being scattered; but after traveling a few thousand lattice constants the electron is almost certain to be scattered in germanium at room temperature because thermal vibrations disturb the perfectly periodic arrangement of the atoms.

Precisely similar theoretical conclusions are reached about the behavior of a hole. Thus the physical picture of an electron is that of a more or less localized wave packet having minus one electronic unit of charge which moves in a random fashion under the influence of thermal agitation and changes its direction of motion, on the average, each time it has traveled about one thousand lattice constants. The picture of the behavior of a hole is precisely the same except that the change is plus one electronic charge. (This is not an obvious result and can be justified only by lengthy theoretical arguments.)

Expected Effects

It should be pointed out that other types of behavior are perfectly conceivable. For example, it might be possible that the effect of light falling upon the crystal would be to destroy entirely one electron-pair bond. Under these conditions the hole would have a charge of + 2 electronic units and would otherwise behave the same as the hole, as we have already described. Thus, one objective of experiment is to verify that the charges which move in the crystal are − 1 and + 1 electronic units of charge respectively. Another objective of experiment is to show that the random diffusive motion actually takes place.

Other effects must be consistent with the physical picture. Under the influence of an electric field a steady drift is superimposed on the random motions and as a result, in addition to diffusing away from any given point, an electron tends to flow through the crystal in a direction opposite to the electric field—opposite since its charge is negative. Similarly a hole tends to flow in the direction of the applied field. The localized disturbance of the valence electrons, which is what the hole really is, moves just as would a particle and drifts in the direction of an applied field. This displacement of the disturbance in the direction of the applied field



Conduction in the perfectly bonded structure may be produced by photon absorption which produces hole-electron pairs.

has been directly observed as a result of experiments on transistor action. If both electric and magnetic fields are simultaneously present then the drift motion produced by the electric field results in a sidewise thrust in the presence of the magnetic field.

It is interesting and important to note that the same electric and magnetic fields will exert a side thrust on holes and electrons in the same direction rather than in opposite directions. This can be understood by realizing that the force due to motion in a magnetic field is dependent both upon the direction of motion and upon the charge. For the case of an electron the motion is opposite to that of a hole and the sign of the charge is also opposite. As a result the sidethrust of the magnetic field is in the same direction for both carriers. There are, of course, deflections due to the random motion as well as the steady drifting motion of the particles, but these cancel out on the average so that for purposes of calculating the deflection it suffices to consider the average motion. One of the new experiments shows that these sidewise deflections actually occur, and under favorable conditions result in concentrating both holes and electrons on one side of a conductor.

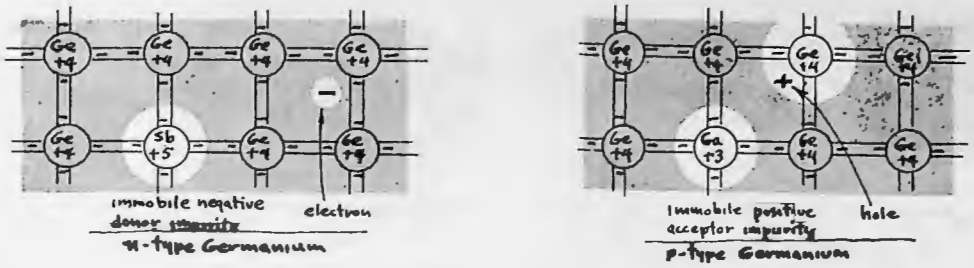
A Fortunate Property

In germanium at room temperature holes and electrons are created by other processes than light absorption. Thermal agitation of the lattice causes the spontaneous breaking of bonds and the generation of holes and electrons. The reciprocal process goes on at a compensating rate under conditions of thermal equilibrium so that hole electron pairs are being continually created and recombined. In germanium at room temperature these processes would reach a balance when the concentration of holes and electrons is sufficient to give the sample a resistivity of approximately 60 ohm cm.

If it were necessary to deal exclusively with semiconductors having equal numbers of holes and electrons, the problem of unraveling the behavior of the individual types of current carriers would be much more difficult. Fortunately, samples having conductivity by electrons alone or by holes alone can be prepared in controlled ways.

A specimen which conducts by electrons alone is called n-type since the current carriers are negative. Its conductivity arises from the presence of chemical impurities called donors which give electrons to the conduction band. A typical donor atom for germanium is antimony which has a valence of 5. This atom forms four electron-pair bonds with its neighbors using for this purpose four of its five valence electrons. The fifth valence electron cannot fit into the bond structure and becomes free to wander through the crystal. The addition of antimony does not produce a net negative charge in the crystal, however, because the antimony atom itself represents a fixed immobile positive charge, its share of the four valence bonds around it being insufficient to neutralize the charge of + 5 units on its core. At low temperatures the attraction between this plus charge and the electron is sufficient to bind the electron to the antimony ion in a wave function similar to that of an electron in a hydrogen atom. The binding is much weaker than that of hydrogen, however, because the charge of the antimony ion is imbedded in a dielectric consisting of the germanium atoms, and as a result the electron is so weakly bound that thermal agitation readily frees it. At room temperature, in a sample of n-type germanium, a negligible number of electrons are bound to the donor ions.

An entirely similar situation occurs for a sample containing impurity of valence 3, of which gallium is an example. In such a crystal the electron-pair bonds around the gallium atom are completed by the stealing of an electron from an electron-pair bond somewhere else. This results in the presence



Adding donors with valence 5 produces an n-type semiconductor with conduction by electrons alone. Acceptors with valence 3 similarly produce p-type with conduction by holes.

of a mobile hole, and the gallium atom with its surrounding share of the valence bonds becomes a negative ion. Because of its thieving nature an atom of valence 3 is politely called an acceptor.

If both acceptors and donors are incorporated in germanium in equal numbers, the resulting conductivity is not the sum of the conductivities of the two types but is substantially the same as it would be if neither were present. The reason is that the electrons produced by the donors combine with the holes produced by the acceptors until finally an equilibrium density of holes and electrons is produced just equal to that which would occur as a result of spontaneous thermal generation. Consequently, the conductivity of a sample having donors and acceptors is determined by the more abundant type of impurity and is the same as it would be if simply the excess density of that type alone were present.

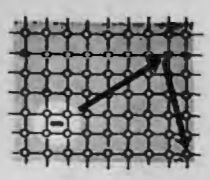
The basis of the development of transistor electronics at Bell Telephone Laboratories both from the applied and theoretical sides has been the preparation by J. H. Scaff and his colleagues of germanium of high purity with controlled additions of donors and acceptors. Another vital aid to research has been the growth by G. K. Teal of relatively large and perfect single crystals. And still another essential ingredient has been the development by W. L. Bond and G. L. Pearson of microtechniques for adjusting point contacts and cutting small specimens or filaments.

We may summarize the physical picture of the principal attributes of holes and electrons as follows: they behave like particles having charges equal respectively to minus the charge of the electron and plus the charge of the electron. They undergo random thermal motions so that they tend to diffuse away from any point where they are initially located. Under the influence of an electric field an electron drifts opposite to the field and a hole with the field. These drift velocities are, according to theory, proportional to the electric field, the proportionality constant being called the mobility. According to convention, mobility is always given as a positive number so that it gives merely the ratio of speed to electric field and is plus for electrons as well as for holes. Under the influence of electric and magnetic fields both electrons and holes tend to be thrust towards the same side of the specimen. The physical picture also includes the concepts of electron and hole densities usually expressed in number per cubic centimeter and written as n and p.

Conductivity and Hall Effect

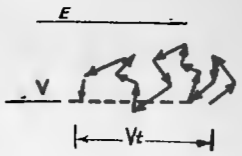
The pretransistor measurements of the behavior of holes and electrons consisted chiefly of measurements of conductivity and Hall effect. Conductivity was measured simply by preparing a specimen of suitable shape, preferably with a rectangular or circular cross section and much longer than it was wide. Current was then passed through this specimen and the conductivity measured using Ohm's law. For an n-type sample of germanium this conductivity gave information about the product only of three of the quantities of the physical picture: the size of the charge, the density of electrons, and their mobility. By itself it verified nothing about sign of the charge of the carrier since the theory requires that the charge and the direction of motion are coupled so that ordinary positive resistances occur.

The other measurement which gave information on the nature of the conduction process was based on the Hall effect. When electrons flow down a filament in the presence of electric and magnetic fields, they tend to be deflected to one side. As a result of this deflection there is a small accumulation of electrons on one side of the specimen and a corresponding deficit on the other side. These charges produce an additional electric field which



The excess electron wave pocket moves many lattice constants between deflections. The resultant motion is diffusive.

An electric field superimposes a drifting motion with average velocity V on top of the random motion. The ratio V/E is the mobility μ.



tends to neutralize the sidewise thrust of the magnetic field. According to one somewhat oversimplified way of considering the Hall effect, the steady state is reached when the transverse electric field due to the accumulation of charge sets up a sidewise thrust which exactly balances the sidewise thrust due to the magnetic field.

It is evident that these sidewise electrical thrusts, which exert forces in the same direction for holes as for electrons, must correspond to fields of opposite signs when the carriers are considered separately. The same result can be seen by considering the transient effect which occurs when the magnetic field is being applied. As the magnetic field builds up, electrons in an n-type sample would be deflected towards the top; and in a p-type sample holes would similarly be deflected towards the top so that the transverse electric fields, which can be measured in Hall effect experiments, would have opposite signs for n-type and p-type. This difference in sign allowed the conductivity type to be determined by the Hall effect.

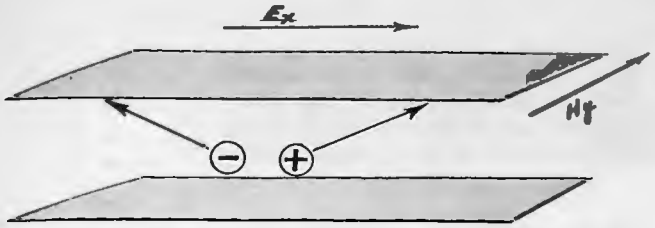
Furthermore, by equating the two thrusts it was found that a measurement of the ratio of transverse and longitudinal and electric fields at a known magnetic field is equivalent to a measurement of the mobility. This ratio of transverse to longitudinal electric fields is the angle in radians between the direction in which the current flows and the direction of the resultant or total electric field. In terms of a more detailed picture of the statistics of the process involved this angle has an interesting interpretation: if the electron were moving in a magnetic field alone, its path would be an arc of a circle. The Hall angle is the average angle of the arc between collisions.

So the Hall effect may be considered as a means of measuring mobility. Actually it is a very indirect means since it does not involve measuring times and distances in order to determine the drift velocities associated with the mobility. Time enters the mobility deduced from the Hall effect through the calibration of the instruments which measure the magnetic field. This indirectness has a theoretical counterpart as well, arising from the neglect of momentum transfer during collisions with the atoms; thus theory indicates that only under very special conditions can the mobility deduced from the Hall effect be considered as equivalent to the true or drift mobility. For germanium it appears at present that these simplifying assumptions are almost but not exactly correct. In principle, however, it is possible to have electrons and holes behave in such anomalous ways that the Hall effect gives the wrong sign.

Thus, from the operational viewpoint, the Hall effect cannot be regarded as giving a demonstration that particles drift with the predicted velocities nor does it show that they are deflected by the sidewise thrust as predicted by the physical picture. Practically, however, there is evidence that the Hall effect is generally reliable and most of the quantitative aspects of the theory are based on its use. Even accepting the mobility as given by the Hall effect, however, data from it and from the conductivity measurement give only two relationships among three quantities (the charge of the carrier, the density of the carriers, and the mobility).

Radioactive Antimony

The concentrations of donors and acceptors which produce the conductivities usually observed in germanium samples are extremely small, of the order of one atom in one hundred million, so small that conventional chemical methods of analysis are quite inadequate. If the impurity contains a radioactive isotope, however, even these small concentrations may be measured. Samples of germanium containing radioactive antimony have been studied by Pearson, Struthers, and Theuerer and the effective concentration of added donor atoms has been directly determined. On the basis of the theoretical picture, the concentration of donors should be equal to the concentration of excess electrons so that in these

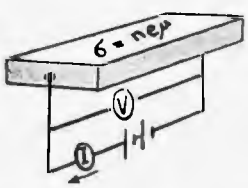


The drift velocity of electrons and holes is deflected upwards (+Z) by electric field E_x and magnetic field H_y.

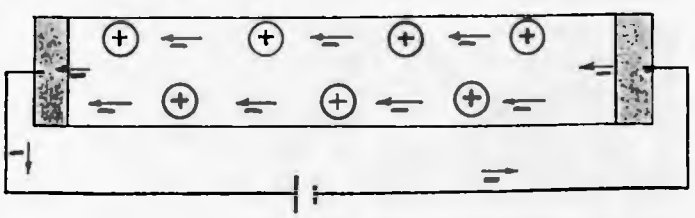
samples we may consider that the concentration is known. Even if no antimony is deliberately added, some impurity is always present and this must be estimated by control experiments. Combining this information with the knowledge of the mobility obtained from the Hall effect one is able to deduce an effective charge for the electron. This charge appears to be somewhat greater than the charge of the free electron but the accuracy of the data does not, at present, warrant a positive conclusion on this point. From the theoretical point of view as discussed above, the mobility of the particles and a correction factor may have to be introduced. Fortunately, the charge of the carriers can now be determined by a more direct means.

The Sign of the Charge—Carrier Injection

One of the basic new phenomena resulting from the transistor studies is that of hole injection. The phenomenon occurs when a point contact is made to a sample of n-type germanium. (A corresponding process of electron injection occurs when p-type germanium is used and the two types are grouped together under the phrase "carrier injection".) This process is so different from that occurring in ordinary conduction process that it is helpful to give first a description of the ordinary conduction process in more detail than usual in order to contrast the two.



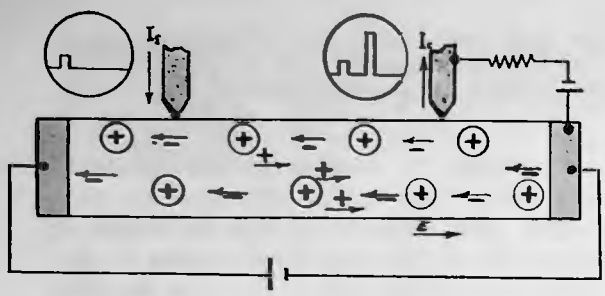
In an n-type semiconductor normal conduction corresponds to an incompressible flow of electrons since any accumulation would produce enormous space charges.



If electrodes are placed on the two ends of an n-type sample, and suitable precautions are taken to make them nonrectifying, then the passage of current through the sample involves a flow of electrons in at one of the terminals and out at the other. The total number of electrons in the sample stays constant during the process, however, because if any appreciable deviation in the total number of electrons occurred, the specimen would become electrically charged, and a very small fractional deviation would set up enormous electric fields. Thus the electron flow is like that of an incompressible fluid.

Another important feature of the ordinary conduction process is that the velocity of an electrical signal down a long wire of germanium is much greater than the velocity with which the electrons drift down the wire. If a coaxial line were made with a central conductor of germanium, for example, and a negative voltage were suddenly applied to one end, electrons would flow into that end almost instantaneously. This flow would produce an electric field which would be transmitted down the coaxial conductor with approximately the speed of light. This electric field would set in motion electrons at successively more distant points along the filament so that very quickly all of the electrons in the filament would be drifting in the same direction. It would not be necessary to wait until electrons injected at one end drifted to another point in the filament to detect the signal produced by applying the voltage.

The phenomenon of hole injection stands in great contrast to that of ordinary conduction. It is necessary to wait until the holes drift to another point to detect the signal. In the case of hole injection, a rectifying contact (of transistor emitter type) is made to the n-type germanium. When this rectifying contact carries a positive current into the germanium, the electronic equilibrium is upset. The reason is that the electrons drawn out of the germanium, when the positive current flows through the contact, are taken from the valence bonds rather than from the excess electrons of the conduction band. The theory for this behavior can be understood in terms of transistor action; for our purposes we may consider it simply as an established fact.



A pulse of holes injected by an emitter point is neutralized by extra electrons flowing in at the end terminals. The hole pulse drifts in the direction of the applied field.

Assuming that this process occurs, we see that when the point is positive, it generates holes within the germanium which can then diffuse away from the contact. If an electric field is present these holes will be carried away in the direction of the electric field.

Another point contact is made to the same sample and is biased in the negative direction, which is the reverse or high resistance direction for n-type germanium. If holes injected by the first point arrive at the second point, the high resistance of second point will be greatly reduced—this being one of the important transistor principles. Thus the second or collector point is a detector for holes and can be used to time the arrival of holes after they have been injected by the emitter point.

J. R. Haynes has used these effects to make precision measurements of the drift mobilities of holes and electrons: a thin rod of n-type germanium is caused to carry a steady current from end to end by large inert electrodes which do not inject holes; two point contacts are made to the rod and a pulse of holes is injected into the rod by one of the points; the other point is connected to an oscilloscope and the arrival of holes added is detected by a change in its reverse current. As soon as the emitter current flows a signal is detected at the collector, this being the ordinary Ohm law signal, discussed above, which is transmitted with the speed of light. It serves the useful purpose of giving a reference time to tell when the holes were injected. At an appreciable time later there is another signal on the collector point, which shows the arrival of the holes.

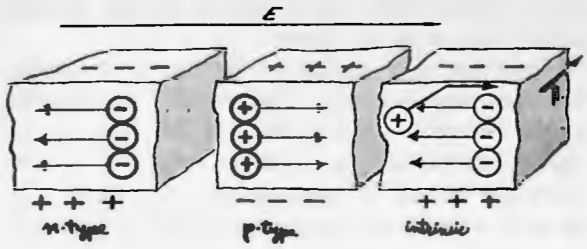
Haynes has found that the disturbance set up at the emitter point moves in the direction expected for a positive charge and that its drift velocity is proportional to the electric field. This experiment thus observes directly the drift velocity, which was a concept in the original theory not supported directly by experiment, and it also verifies the fact that the disturbance introduced has a positive sign. He has carried out similar experiments for p-type germanium into which electrons are injected.

A hole injected in these experiments will combine with an electron unless it is drawn out of the end of the filament first. Experiments using filaments with different cross sections show that this recombination takes place largely on the surface of the filament and is describable in terms of an average lifetime for a hole. Values ranging from less than one to over one hundred microseconds have been measured for both holes and electrons.

Unlike the case of ordinary conduction, electrostatic neutrality does not limit the number of holes which can be injected into a sample of n-type germanium. The reason is that as the holes are injected, they set up an electric field which attracts electrons. There is a practically unlimited supply of these electrons available to flow in through wires connecting the end terminals of the germanium sample to the battery, which is used for injecting the holes. Thus, very large concentrations of holes and electrons may be produced. The presence of the added holes and electrons under these conditions increases the conductivity of the sample and this increase can be used to measure the added concentration. The resulting conductivity modulation can be used so as to obtain power gain. Transistors using this principle, called filamentary transitors, have been shown by Pearson and Shockley to operate in accordance with the theory based on the physical picture and have been used to study some of the quantities entering the theory.

An Observation of Diffusion

An experiment which measures not the drift velocity in an applied electric field but measures instead simply the diffusion of added holes has been carried out by Goucher. No sweeping field is applied to the specimen and the emitter point is replaced by a small spot of light. (Goucher has shown by an independent experiment that every photon absorbed in the germanium over a wavelength range of 1.0 to 1.8 microns generates one hole-electron pair.) Since no net current is produced when hole-electron pairs are generated by the light, the electric fields



In the Hall effect, charges build up until the transverse field causes the current flow to run parallel to the specimen. In the Suhl effect, which requires strong electric and magnetic fields, concentration of carriers occurs since space charge neutralization of opposite types takes place: the case shown corresponds to injecting a few holes in an n-type specimen.

are so small that they contribute negligibly to the motion of the holes; and the holes thus move substantially under the influence of diffusion alone.

For such a case the manner in which the density of holes decreases at distances from the light source can be predicted directly from theory. The lifetime of a hole, which enters the calculation, has been determined by experiments of the sort carried out by Haynes. The presence of the holes can be detected by the collector point on the filament without drawing any current because when the holes arrive at this collector point they cause an internal contact potential difference to develop. By these means Goucher has detected the presence of holes which have diffused more than one millimeter away from the light source and has found that the hole density depends on distance from the light source in precisely the way one would expect, based on the picture of diffusion discussed above.

An Observation of Magnetic Deflection—The Suhl Effect

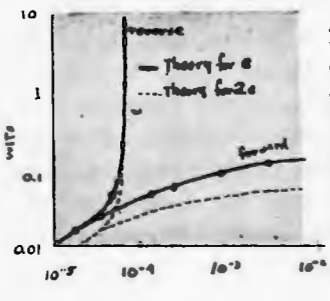
Direct evidence that holes and electrons are deflected by combined electric and magnetic fields in the predicted directions has been obtained by H. Suhl. In his experiment both holes and electrons are simultaneously present. This condition is very different from that occurring when holes and electrons are present individually. In the normal Hall effect, for one type of carrier alone, the carriers flow down the filament producing a current which is uniform across the cross section and the carriers have a uniform density. The uniform density is required by the fact that deviations from it produce a net space charge.

The Suhl effect is probably most easily described in terms of an extreme case corresponding to an intrinsic semiconductor (one in which holes and electrons are present in equal numbers). Here the combined action of electric and magnetic fields will tend to push the holes and the electrons both towards the same side. If the electric and magnetic fields are both sufficiently large, a situation made possible by germanium filament techniques, the holes and electrons may be concentrated on one side of the filament. Since the carriers are equally abundant, the combined densities of holes and electrons will produce no net space charge.

In Suhl's experiment only small concentrations of holes were injected into an n-type filament. These small hole currents were then found to be deflected by the magnetic field so that the holes were found all on one side of the filament. The means of detection was again a collector point operated at a low voltage so that the current which it drew produced only a small disturbance in the hole flow that was being measured. Suhl found that the concentration of holes as measured by the collector point obeyed precisely the theoretical laws expected if holes were actually subjected to the side thrust predicted by the theory. Furthermore, he found that the lifetime of the holes in the filament was greatly reduced by this process and this reduction could be explained by supposing that the holes recombined with electrons chiefly on the surface of the filament, a conclusion in agreement with Haynes' findings.

The Magnitude of the Charge—Rectification

The magnitudes of the charges of the holes and electrons can both be shown to be equal to the



The rectification curve for a p-n junction fits well if the carriers are assumed to have the charge of one electron and badly if another charge, such as 2e, is assumed.

electronic charge by studying the rectification curves of p-n junctions. A p-n junction occurs when the donor and acceptor densities vary in a single piece of germanium so that one part where the donors predominate is n-type while another part is p-type. A rectifier may be made by making low resistance contacts to the p- and n-parts, making such contacts being somewhat of an art.

Current flow across the junction between the p- and n-type parts has a very nonlinear current voltage relationship. The easy direction of current flow is that in which the p-side is biased positive so that holes tend to flow from it into the n-side while electrons flow in the opposite direction and on both sides of the junction carrier injection occurs. The behavior of holes and electrons in such junctions has been analysed and it is found that the theoretical relationship between current and voltage is the same as the theoretical formula for other types of rectifiers.

However, the greater perfection of the junctions made in germanium leads to a much closer approach of the experimental curve to the theoretical curve than has been observed before. The formula for rectification involves directly the effective charge of the carriers and it is found that the formula is best fitted for a p-n junction of high perfection by choosing the charge on the mobile carriers as that of the electron.

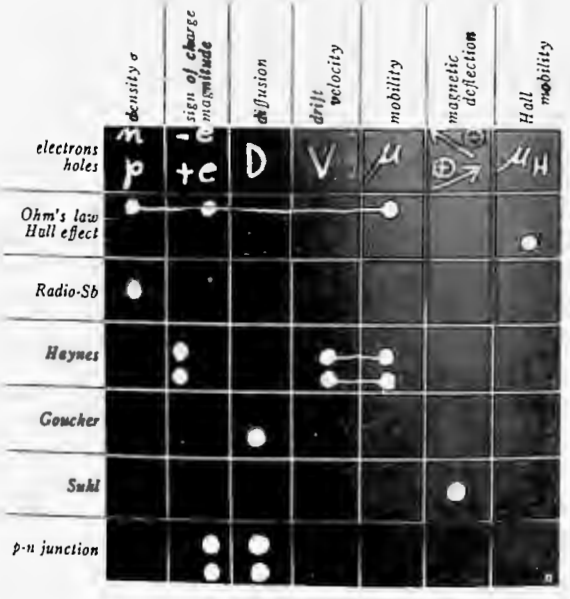
The quantitative fit of Suhl effect data also indicates that the charge is one electronic unit, but the available data are not so accurate as that for p-n junctions and the theory somewhat more involved.

Conclusions

Although the experiments just described have verified many features of the physical picture of the behavior of holes and electrons, the picture itself is not complete. The verified features, however, are those which are of most importance in practical applications. Applications in transistor electronics require knowledge of mobilities, drift velocities, diffusion, the magnitude of the charge, and the deflections of magnetic fields. These quantities have now been given an operational meaning (in Bridgman's sense) in situations closely resembling those in which they will be used for design purposes. For this reason these features of the physical picture have a reality which can hardly be affected by any future changes in the theory.

There are other aspects of the theory that have had some degree of test. The velocity distribution of the electrons is predictable and has been investigated theoretically through its influence on the scattering process. A more serious problem has to do with whether the electron waves move through the crystal like pressure waves in a gas, which can be only longitudinal, or like mechanical waves in a solid which have three sound velocities, this latter case corresponding to a degenerate state at the edge of the band. The increase in resistance in magnetic fields seems to require the more complicated case but the details have not yet been worked out. Even with these lacks, however, germanium has become one of the best understood of all electronic conductors and may well become the best understood as a result of studies centered around its application in transistor electronics.

In the pretransistor experiments only conductivity and Hall effect were measured giving two relations among three quantities (or among four quantities if μ and μ_H are distinguished). The new experiments give direct operational significance to all the remaining quantities of the table.



THE PHYSICS OF ELECTRONIC SEMICONDUCTORS

G. L. PEARSON

TRANS. OF AIEE VOL.66 pp.209-214 1947

Produced outside the work performed under this contract.

The Physics of Electronic Semiconductors

G. L. PEARSON
NONMEMBER AIEE

SEMICONDUCTORS have a number of unique properties which make them useful in the design of electric circuit elements. Their electrical resistivities near room temperature fall in the range 10^{-2} to 10^{-6} ohm-centimeters. In comparison, the resistivities of typical metals in this temperature range are around 10^{-5} ohm-centimeters and those of insulators between 10^{6} and 10^{15} ohm-centimeters or higher. The electrical resistivities of semiconductors and insulators decrease rapidly with rise in temperature, while those of metals increase relatively slowly. The resistivities of dry rectifiers, which consist of a metal in contact with a semiconductor, may change by a factor of 1,000 as the sign of the applied potential is reversed. The purpose of this paper is to outline the present theories relating to semiconductors and to correlate these with quantitative experimental data obtained on some typical semiconducting samples.

Semiconductors may be classified on the basis of their current carriers into electronic, ionic, and mixed conductors. In ionic and mixed conductors, ions are transported through the solid, thus changing the composition of the material. As this leads to unstable characteristics, only those substances in which the current is carried solely by electrons will be discussed here.

As early as 1900 Drude[1] suggested that the high electrical conductivity of metals (low resistivity) was due to the large number of so-called free electrons. These electrons are free in the sense that under the action of an electric field they move through the crystal lattice, thus carrying a current. Their number is of the order of one per atom or approximately 10^{22} per cubic centimeter in most metals, and this density does not change appreciably with change in temperature. In the modern theory of semiconductors and insulators as proposed by Wilson[2-4] around 1930, it is shown that most of the electrons in these substances are bound to their respective atoms, thus leaving only a small fraction of the electrons (1 per 10^{3} to 10^{9} atoms) free to conduct current. Furthermore, the number of such electrons increases rapidly with increase in temperature, thus accounting for the high negative temperature coefficient of resistance in semiconductors and insulators.

The physical picture of a metal consists of an electron gas formed by the free electrons in which the positive ions float. As the ions repel each other, they tend to arrange themselves so as to use their space to the best advantage, which requires that they take up some close-packed crystal arrangement such as body-centered cubic, face-centered cubic, or close-packed hexagonal. By means of X-ray analysis Grimm[5] has determined, as shown in Figure 1*B*, the electron charge density in the metal magnesium for the base plane of the close-packed hexagonal lattice. This figure is drawn to scale, the unit being angstroms (1 angstrom = 10^{-8} centimeter). It is seen that the valence electrons give a uniform negative charge in which the positive magnesium ions are imbedded. The heavy density near the lattice points is from the electrons bound to the ions. Figure 1*A* gives the electron density in the insulator diamond as determined by Grimm for the 110 plane of the face-centered cubic lattice. Although the positive carbon ions are imbedded in the negative charge of the valence electrons, it is seen that spherical symmetry is lacking and the electrons are concentrated along the shortest paths between atoms.

Band Theory of Solids

If all of the charge in the electron gas were free to move through the lattice under the influence of an electric field, Figures 1*A* and 1*B* indicate that diamond and magnesium should be equally good conductors. Measurements show, however, that at room temperature magnesium is more conducting by a factor of 10^{19} and therefore we must add something to our picture to explain this enormous difference. This was one of the major problems of the physics of solids which had to wait until the advent of quantum mechanics around 1930 for a solution. Although the problem is now completely solved, the solution is so wrapped up in mathematical language that it can not be presented here. It can be said that solids, like free atoms, are pictured as having a large number of discrete energy states with one and only one electron allowed in each individual state. These energy states arrange themselves into bands, the structure of which determines the properties of the solid. Just as automobiles in a parking lot are unable to move when they are parked bumper to bumper, so the electrons in a solid are immobile when every energy state in the band is filled with an electron. When there are unoccupied states in the band, the electrons move under the action of an electric field and the material is a conductor. A semiconductor has only a few electrons in a conduction band (a band with energy states only partially filled) at room temperature, but this number increases with increase in temperature. These ideas are represented graphically in Figure 2 where the black areas represent

Paper 47-34, recommended by the AIEE committee on basic sciences for presentation at the AIEE winter meeting, New York, N. Y., January 27-31, 1947. Manuscript submitted October 31, 1946; made available for printing December 5, 1946.

G. L. PEARSON is with Bell Telephone Laboratories, Inc., Murray Hill, N. J.

Acknowledgment is made to J. H. Scaff and H. C. Theuerer for furnishing the silicon samples described in this paper and to E. F. Kingsbury for contributing Figure 5*A*. A great deal of helpful criticism was received from J. Bardeen during the preparation of this paper.

allowed energy levels filled with electrons, the white areas represent forbidden regions where no energy states exist, and the shaded areas represent conduction bands. Energy is plotted vertically on an arbitrary scale. The width of the forbidden region may vary from zero to ten electron volts (one electron volt is the energy required to move one electron through a potential difference of one volt). Figure 2*A* indicates the conditions existing in an insulator where all the electrons are in filled bands and the forbidden region is so wide, several electron volts, that thermal agitation is incapable of exciting electrons from the filled band into the conduction band. In a semiconductor (Figure 2*B*) the energy gap between the filled and conduction bands is small, one electron volt or less, so that electrons from the completely filled band may be thermally excited to the conduction band, thus producing a conductivity which increases with temperature. In some metals such as magnesium, the energy gap between the filled and unfilled bands is zero as shown in Figure 2*C*, and in others such as the alkali metals the lower band is only half filled with electrons (Figure 2*D*). In either case conduction is good at any temperature.

Semiconductor Classifications

More detailed representation of the electron arrangements in the energy bands of chemically pure semiconductors and of those containing small amounts of different types of impurities are shown in Figure 3. Figure 3*A* is for pure semiconductors which have been termed intrinsic. At any temperature T an average number n of electrons are excited thermally from the lower filled band to the upper band. Conduction then can take place both by means of the electrons in the upper band and holes in the lower band, which is now not quite full. In the latter mechanism the sign of the carrier charge appears to be positive.

In most semiconductors the conductivity is highly dependent on the impurities present in the sample. By impurities we mean added foreign atoms, physical defects in the lattice, or differences from stoichiometric composition. It is assumed that the impurities introduce extra energy states in the normally forbidden region between the bands as shown in Figures 3*B* and 3*C*. Electrons in these impurity states are localized in the neighborhood of the impurity and cannot take part in conduction. Two types of impurity are possible. In the donator or *N*-type shown in Figure 3*B* the impurities carry extra electrons with them and usually have energies near the conduction band. Since the energy gap E_1 is much less than E, electrons are thermally excited at lower temperatures than in an intrinsic semiconductor. At very low temperatures the extra electrons are all bound to the impurity atoms; at higher temperatures these electrons are thermally excited to the upper band where they may conduct. At still higher temperatures the donator electrons may all become excited so that the number of conduction electrons remains constant with further increase in temperature until the thermal energy becomes great enough to excite intrinsic electrons from the lower band. An impurity semiconductor turns into an intrinsic semiconductor at high temperatures. At low temperatures the number of excited impurity states and, therefore, the conductivity at a given temperature increase with the amount of impurity present.

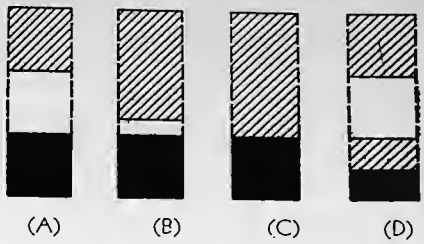


Figure 2. Energy level diagrams for solids

The clear regions are forbidden, the dark areas are allowed and full, the shaded areas area allowed but empty

A. An insulator
B. A semiconductor
C. A metal with overlapping filled and conduction band
D. A metal with half filled band

In the acceptor or *P*-type impurity semiconductor shown in Figure 3*C*, the extra states lie near the filled band. At very low temperatures these states are all empty, but on heating, electrons from the lower band are excited into them. Although the excited electrons are bound, hole conduction is possible in the lower band.

In an intrinsic semiconductor the conduction is by an equal number of electrons and holes, in an *N*-type conduction is solely by electrons at low temperatures and by electrons and holes at high temperature, and in a *P*-type solely by holes at low temperatures and by electrons and holes at high temperatures. In addition to explaining the magnitude of the conductivity and temperature coefficient, we will see that this picture clarifies certain anomalies in the Hall and thermoelectric effects in semiconductors.

Electrical Resistivity

Wilson's theory shows that for intrinsic semi-conductors the specific resistance

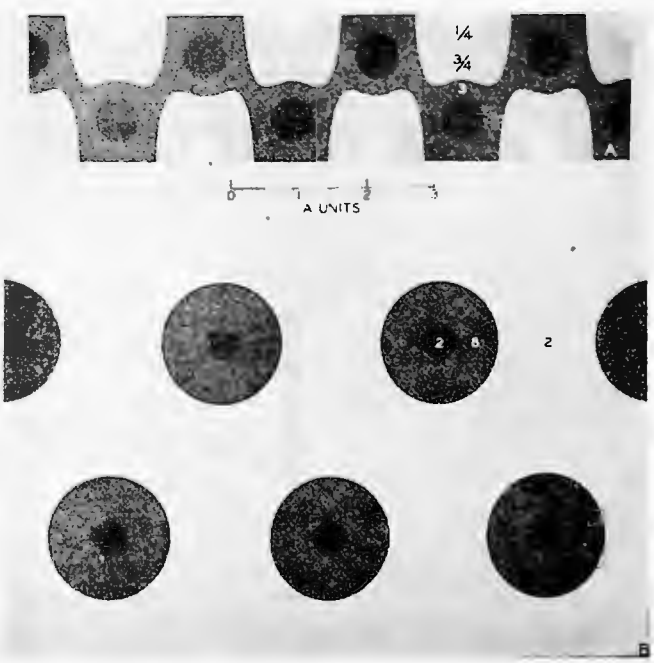


Figure 1. Electron charge densities in an insulator and a metal

The numerical values give the number of electrons per atom in the space corresponding to each intensity of shading

A. Diamond
B. Magnesium

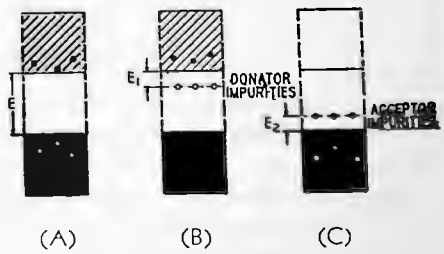


Figure 3. Energy level diagrams for semiconductors

A. Intrinsic semiconductor
B. *N*-type impurity semiconductor
C. *P*-type impurity semiconductor

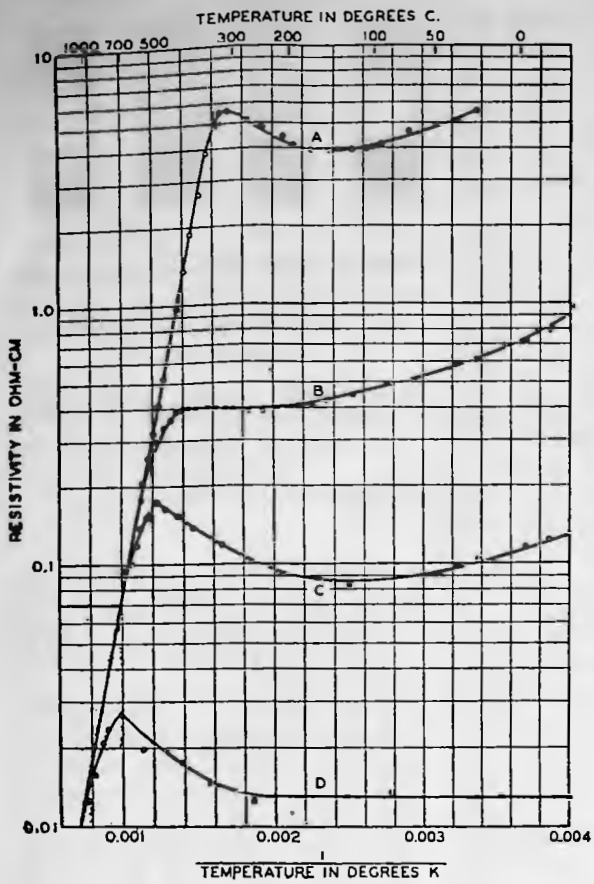


Figure 4. Logarithm of the resistivity of various specimens of silicon as a function of inverse absolute temperature

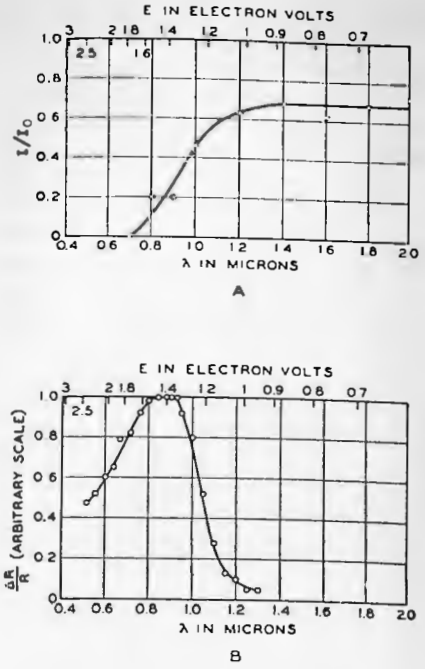


Figure 5

A. Photo transmission in a thin silicon film as a function of wave length

B. Photo conductivity in a thin silicon film as a function of wave length

ρ in ohm-centimeters is

$$\rho = A e^{E/2kT} \qquad (1)$$

where A is a factor which varies relatively slowly with temperature, T is the temperature in degrees Kelvin, k is Boltzman's constant (8.69×10^{-5} electron volts per degree), and E is the energy gap between the two bands in electron volts. From equation 1 it can be seen that a plot of log ρ versus $1/T$ for an intrinsic semiconductor should give a straight line the slope of which is equal to $E/2k$ or 5,750 times E in electron volts. Figure 4 is such a plot for silicon containing different amounts of impurity. Curve A was obtained from measurements made on a block of silicon prepared from a melt of high purity material obtained from du Pont. Curve D is from a block of the same material to which 0.03 atomic per cent of boron had been added. Curves B and C were obtained from silicon samples containing intermediate amounts of unknown impurities. At the higher temperatures the resistivities of all the samples fall on the same straight line in accordance with our picture of semiconductors in the intrinsic range. The slope of this line indicates that the energy gap between the filled and unfilled bands in silicon is 1.2 electron volts. At lower temperatures the resistivities do not continue to follow the intrinsic line, but branch off at increasingly higher temperatures with increase in impurity. In the branch curves, the number of intrinsic electrons and holes becomes small compared with the number excited from impurity states and the resistivity is determined by the amount of impurity present. The decrease in resistivity just beyond the branch point is not associated with a change in the number of excited impurity states, but rather arises from a decrease in mobility resulting from the heat motion of the ions in the lattice.

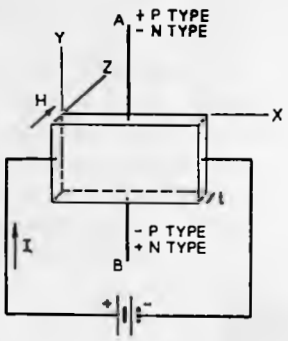


Figure 6. Schematic circuit arrangement for the measurement of Hall coefficient in semiconductors

Sign conventions are as shown

This is a complicated story and will not be discussed here. On going to still lower temperatures the impurity resistivity increases and an estimate of the energy gaps E_1 or E_2 can be made from the slopes.

Although sample A is the purest silicon we know how to prepare, even this material must contain some impurity, as defined above. At room temperature the addition of three boron atoms for each 10,000 atoms of silicon in sample D reduced the resistivity by a factor of 500 as compared with sample A. This indicates that A is pure to about one atom in five million so far as conductivity is concerned. If absolutely pure silicon, which followed the intrinsic line to room temperature, could be obtained its resistance would be

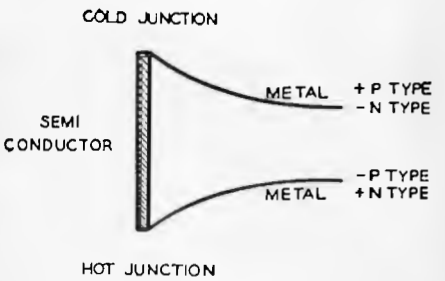


Figure 7. Schematic circuit arrangement for the measurement of thermoelectric power in semiconductors

Sign conventions are as shown

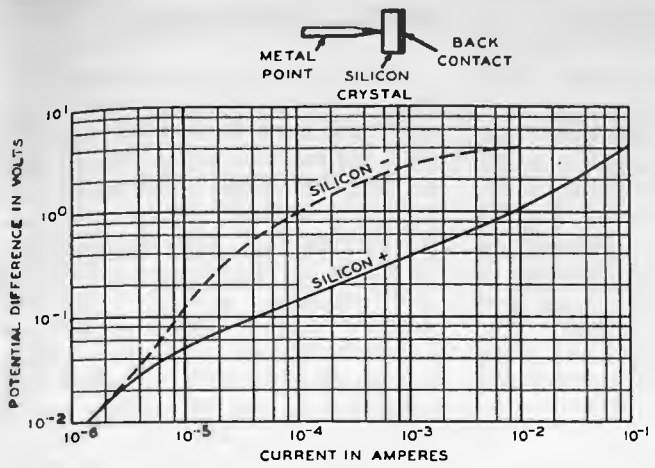


Figure 8. Logarithmic plot of the current versus voltage characteristic for a *P*-type silicon point contact rectifier

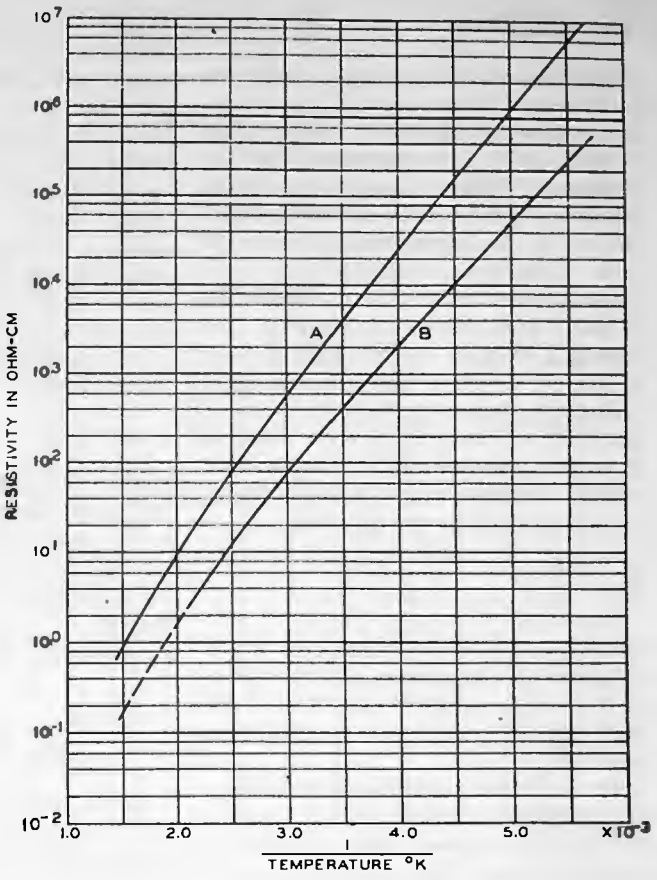


Figure 9 (right) Logarithm of the resistivity of two thermistor materials as a function of inverse absolute temperature

greater than that of sample *D* by a factor of 10^7 at room temperature. This would require impurity of less than one atom for each 10^{11} silicon atoms.

Resistivity measurements such as those shown in Figure 4 have been made on other materials which have intrinsic characteristics. The values of the energy gap *E* in electron volts are shown in Table I. Diamond, of course, is an insulator at ordinary temperatures but it, like most other insulators, has a high negative temperature coefficient of resistance and shows intrinsic characteristics at very high temperatures.

The values of the energy gaps E_1 and E_2 due to impurities can be deduced in like manner from measurements made at low temperatures. As these vary both with the variety and amount of impurity, no quantitative values will be given, although in general they lie between 0.001 and 0.3 electron volt.

Photo Effect

In addition to an increase in conductivity with rise in temperature, many semiconductors show increased conductivity on exposure to light of the proper wave length. This arises from the fact that just as electrons can acquire enough energy thermally to excite them from the filled band to the conduction band, so they may acquire sufficient energy optically to produce the same result. The fundamental equation giving the energy of a quantum of light is

$$E = \frac{hc}{\lambda} \qquad (2)$$

where *E* is the energy in electron volts, *h* is Planck's constant (4.1×10^{-15} electron volt per second), *c* is the velocity of light (3×10^{10} centimeters per second), and λ is the wave length of the optical radiation in centimeters. The energy of a light quantum in electron volts according to equation 2 is 1.236 divided by the wave length of the radiation expressed in microns (1 micron $= 10^{-4}$ centimeter). The wave length of the sodium *D* line in the yellow part of the visible is 0.589 micron and the quantum energy is 2.1 electron volts, which is sufficient to excite electrons from the filled to the conduction band in germanium, silicon, cuprous oxide, and boron, but is insufficient to excite electrons in diamond. Radiation of 1.5 microns, situated in the infrared, has a quantum energy of 0.82 volt and cannot excite any of the semiconductors shown in Table I excepting germanium. By equation 2 only wave lengths shorter than 1.05 microns have energy equal to or greater than the 1.2 electron volt energy gap in intrinsic silicon. Light longer than this wave length should pass through silicon without absorption but shorter wave lengths should be absorbed and cause photoconductivity. Figure 5*A* is a plot of the ratio of the light transmitted to the light incident on a thin film of silicon as a function of the wave length λ expressed in microns. At wave lengths longer than 1.2 microns the transmission is about 67 per cent, the remainder being reflected. At some wave length between 1.0 and 1.2 microns, absorption sets in and the transmission falls off with decrease in wave length. Figure 5*B*, which was obtained by Teal[9] and his collaborators, gives the ratio of the change in resistance to the dark resistance when light of constant energy per square centimeter but varying wave length falls on a film of high purity silicon. It is seen that the resistivity is little affected by long wave-length radiation, but photoconductivity sets in around 1.2 microns and increases sharply below 1.0 microns. The decrease in photoconduction below 0.8 microns is not completely understood, but it is thought the large absorption in this wave-length region keeps the excited electrons near the surface where they recombine with the ions before producing photoconduction.

Table I

Material	E in Electron Volts
Diamond[4]	7
Boron[8]	2
Cuprous oxide[7] (Cu_2O)	1.4
Silicon	1.2
Germanium[8]	0.76

The Hall Effect

Fundamental information about the mechanism of conduction in semiconductors can be obtained from studies of the Hall effect. In 1879 Hall[9] discovered that a potential difference may be produced across a strip of metal which is carrying a current by placing the strip in a magnetic field. Consider a rectangular plate in the plane Z = zero, with edges parallel to the X and Y axes as shown in Figure 6. If a current I from an external battery flows in the X direction and the sample is placed in a magnetic field H in the Z direction, a transverse electric potential V is produced between the terminals A and B. The magnitude of this potential in volts is given by the equation

$$V=\frac{10^{-8}RIH}{t} \qquad (3)$$

where I is the current in amperes, H the magnetic field in gausses, t the thickness of the sample in centimeters, and R the Hall coefficient inc entimeters cubed per coulomb.

The Hall effect occurs because the magnetic field causes the electrons in the sample to travel in curved paths, thus charging up the sides of the conductor until a transverse electric field exists of just the right magnitude to cancel the effect of the magnetic field and make the electrons travel through the solid undeviated. Under the conditions shown in the figure, the terminal A should be negative with respect to terminal B. In many metals and semiconductors this is the case, but in others the sign is reversed. The so-called anomolous Hall effect (reversal of sign in some materials) was not understood until the advent of Wilson's band theory of solids and semiconductors. In the light of this theory, N-type samples give the correct sign since conduction is by free electrons in the upper band but P-type samples give the reverse sign as conduction is by holes (equivalent to an electron with a positive charge) in the lower filled band.

The number n of free electrons or holes per cubic centimeter in semiconductors can be deduced from Hall measurements because the theory shows that

$$n=\pm 3\pi/8eR \qquad (4)$$

where R is the Hall constant in centimeters cubed per coulomb, e is the electronic charge (1.59×10^{-19} coulomb), the minus sign is for electrons, and the plus sign for holes. Putting in the values of the constants we find that n is equal to 7.4×10^{-15} divided by R. Hall measurements on the silicon samples C and D of Figure 4 at room temperature gave R values of $+6.5$ and $+0.5$ centimeter cubed per coulomb respectively, which indicates 1.1×10^{15} and 1.5×10^{19} holes per cubic centimeter. Although the impurity of sample C is unknown, that in sample D is 3 boron atoms for each 10,000 silicon atoms. As there are 5.2×10^{22} silicon atoms per cubic centimeter, this means that 1.7×10^{19} atoms of boron are present in this sample and that each boron produces approximately one hole.

The conductivity σ of an electrical conductor is given by

$$\sigma=nev \qquad (5)$$

where n is the number of current carriers per cubic centimeter, e is their charge, and v their mobility in centimeters per second per volt per centimeter. Combining equations 4 and 5 we see that the mobility of the holes or electrons in a semiconductor is

$$v=\frac{8R\sigma}{3\pi}=0.85R/\rho \qquad (6)$$

The mobilities of the holes in silicon samples C and D at room temperature are, respectively, 55 and 33 centimeters per second per volt per centimeter. In intrinsic semiconductors where electrons and holes are both present at the same time, the Hall coefficient is[2]

$$R=-\frac{3\pi(n_1C^2-n_2)}{8e(n_1C+n_2)^2} \qquad (7)$$

where n_1 is the electron density, n_2 the hole density, and C the ratio of the electron mobility to the hole mobility. As C is generally greater than one, intrinsic semiconductors show the negative sign.

The value of the mean free path of the current carriers is shown from the theory to be

$$l=\frac{2\times10^{-7}R(2\pi mkT)^{1/2}}{\pi e\rho} \qquad (8)$$

where l is the mean free path in centimeters, M is the mass of the carrier ($m=9\times10^{-28}$ grams for electrons and is assumed to be the same for holes), and the remainder of the constants are as previously defined. Substituting in these values we see that

$$l=3.5\times10^{-10}T^{1/2}R/\rho \qquad (9)$$

At room temperature the mean free paths of silicon samples C and D are, respectively, 4×10^{-7} and 2.3×10^{-7} centimeter. These are approximately 10 times greater than the lattice constant of silicon which is 5.4×10^{-8} centimeter.

The Thermoelectric Effect

When a semiconductor is connected between two metal wires as shown in Figure 7, a thermoelectric voltage is produced when the junctions are at different temperatures. The sign of the potential is dependent on whether the semiconductor is N- or P-type, the polarity being as shown in the figure. Intrinsic semiconductors give the same sign as N-type. The thermoelectric power for some semiconductors may be as high as one millivolt per degree centigrade, which is greater than that obtained from most metal thermo-junctions by a factor of 200. For pure hole or pure electron conduction the magnitude of the thermoelectric power Q in volts per degree centigrade difference in temperature between the junctions is[10]

$$Q=\frac{k}{e}\left[\log_e n/T^{3/2}-\frac{\log_e 2(2\pi mk)^{3/2}}{h^3}\right] \qquad (10)$$

Putting in the appropriate values for the constants we see that

$$Q=6.3\times10^{-4}-8.6\times10^{-5}\log_e RT^{3/2} \qquad (11)$$

Reference to the original paper should be made for a treatment of simultaneous electron and hole conduction.[10] Experimental determinations of the thermoelectric powers in silicon samples C and D at room temperature gave 0.92 and 0.44 millivolt per degree centigrade, respectively.

Rectification

When a metal point is placed against an impurity semiconductor as shown in Figure 8, the combination has a nonlinear current versus voltage characteristic and acts as a rectifier. For a P-type sample the current is large when the semiconductor is made positive and the metal point is negative, but is small when these polarities are reversed. For an N-type sample the sign of the effect is reversed so that the current is large when the semiconductor is negative. The plot in Figure 8 was obtained with a tungsten point pressed against silicon sample D which, according to the Hall measurements described, is P-type.

It is impractical within the scope of this paper to give a thorough exposition of rectifying action; however, the following simplified description will convey some of the general features of the theory.[11] Electrons which are thermally excited from the filled band to the acceptor impurity levels produce positive holes in the filled band and charge the acceptor centers negatively. The positive holes in the contact area drift away, thus leaving next

to the metal a negatively charged barrier layer from 10^{-6} to 10^{-4} centimeter in thickness. This barrier layer is a region of high resistivity so that when an external voltage is applied between the metal point and the back electrode, almost the entire drop is across the layer. If the external voltage is poled to make the semiconductor negative, the barrier layer increases in thickness and the resistance goes up. If the semiconductor is made positive the barrier layer becomes thinner and disappears entirely for a sufficiently high applied voltage. This is the direction of low resistance and high current. The case of an *N*-type semiconductor in contact with a metal is analogous. When electrons are thermally excited from donator levels to the conduction band, the donator levels are left positively charged. Electrons are depleted from the boundary layer, leaving an excess of positively charged donators. A positive external voltage applied to the semiconductor increases the thickness of the barrier and a negative voltage decreases the thickness, thereby permitting a large flow of current.

Thermistors

Thermistors, or *therm*ally sensitive re*sistors*, are circuit elements made of semiconducting materials whose electrical resistances vary rapidly with temperature.[12] The semiconducting materials used in the construction of thermistors include sintered mixtures of nickel oxide (NiO), manganese oxide (Mn_2O_3), and cobalt oxide (Co_2O_3). Figure 9 is a plot of log ρ versus $1/T$ for two compositions which have been extensively used to date. Curve *A* is for a binary mixture of manganese and nickel oxides and curve *B* is for a ternary mixture of manganese, nickel and cobalt oxides. These semiconducting materials are unique in that small changes in their composition and impurity content have relatively little effect on their electrical characteristics. This is a particularly useful property as it permits the large scale manufacture of a reproducible product.

The specific resistances at room temperature of the materials *A* and *B* are 250 and 2,000 ohm-centimeters respectively, and the corresponding values of *E* as calculated by equation 1 are 0.68 and 0.60 electron volts. It must be pointed out, however, that the energy band picture of solids is not applicable to these thermistor materials as each of the ingredient oxides have partially filled conduction bands like the metals of Figure 2*D* and therefore, according to the band theory, should be metallic conductors. This fundamental discrepancy was first pointed out by De-Boer and Verwey.[13] Although these authors have derived a theory different from that of Wilson to explain the electrical properties of this class of materials, it will not be discussed here.

Conclusion

The band theory of solids is capable of explaining such fundamental properties of electronic semiconductors as

(1). The dependency of specific resistance on impurity content.

(2). The negative temperature coefficient of resistance.

(3). The sign of the Hall and thermoelectric effects.

(4). The direction of rectification.

Measurements of the specific resistance and the Hall constant enable one to calculate the density, mobility, and mean free path of the electric carriers as a function of temperature and impurity.

References

1. Zur Elektronentheorie der Metalle, P. Drude. *Annalen der Physik* (Leipzig, Germany), volume 1, 1900, pages 566–613.

2. Semiconductors and Metals (book), A. H. Wilson. The University Press, Cambridge, England, 1939.

3. The Modern Theory of Solids (book), Frederick Seitz. McGraw-Hill Book Company, New York, N. Y., 1940.

4. Electronic Processes in Ionic Crystals (book), N. F. Mott, R. W. Gurney. The Clarendon Press, Oxford, England, 1940.

5. Anorganische Riesenmoleküle, H. G. Grimm. *Die Naturwissenschaften* (Berlin, Germany) volume 27, January 6, 1939, pages 1–11.

6. Modern Theory of Solids, Frederick Seitz, R. P. Johnson. *Journal of Applied Physics* (New York, N. Y.), volume 8, March 1937, pages 186–99.

7. Zur Elektrischen Leitfahigkeit von Kupfercxydul, W. P. Juse, R. W. Kurtschatow. *Physikalische Zeitschrift Der Sowjetunion* (Moscow, U. S. S. R.), volume 2, 1932, pages 453–67.

8. Electrical Properties of Germanium Alloys—I. Electrical Conductivity and Hall Effect, K. Lark-Horovitz, A. E. Middleton, E. P. Miller, I. Walerstein. *Physical Review* (New York, N. Y.), volume 69, March 1946, page 258.

9. A New Bridge Photo Cell Employing A Photo Conductive Effect in Silicon. Some Properties of High Purity Silicon, G. K. Teal, J. R. Fisher, A. W. Treptow. *Physical Review*, (New York, N. Y.), volume 69, June 1946, page 686.

10. Theory of Thermoelectric Power in Germanium, V. A. Johnson, K. Lark-Horovitz. *Physical Review* (New York, N. Y.), volume 69, March 1946, page 259.

11. Schottky's Theories of Dry Solid Rectifiers, J. Joffe. *Electrical Communication* (New York, N. Y.), volume 22, 1945, pages 217–25.

12. Properties and Uses of Thermistors—Thermally Sensitive Resistors, J. A. Becker, C. B. Green, G. L. Pearson. AIEE Transactions, volume 65, 1946, November section, pages 711–25.

13. Semiconductors With Partially and Completely Filled 3d-Bands, J. H. DeBoer, E. J. W. Verwey. *The Proceedings of the Physical Society* (London, England), volume 49, August 31, 1937, pages 59–71.

RESISTIVITY IN OHM-CM

p

p=

w
ti
tl
is
el
e
b
a
s
I
I
i
c
c
I
s
c
a
c
l
l

THE THEORY OF p – n JUNCTIONS IN SEMI-CONDUCTORS AND p – n JUNCTION TRANSISTORS

W. SHOCKLEY

BSTJ VOL.XXVIII pp.435-489 JULY 1949

Produced outside the work performed under this contract.

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/cm^2.

---

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

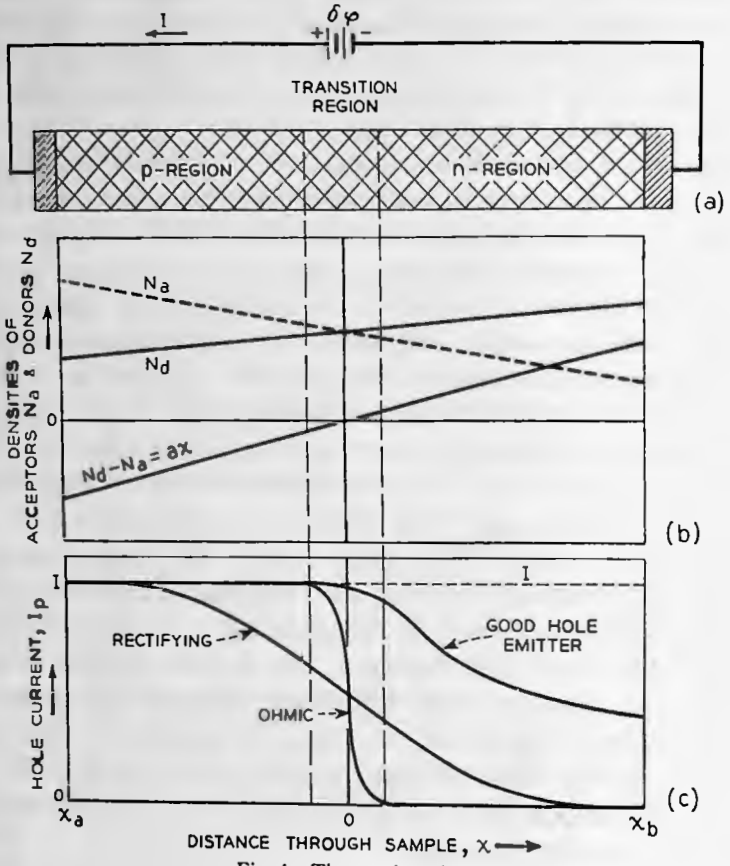


Fig. 1—The p-n junction.

(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 recombination[2] 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 capacitative 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

$$\rho = q\,(p - n + N_d - N_a), \tag{2.1}$$

where q is the electronic charge. We shall measure electrostatic potential ψ in the crystal, as shown in Fig. 2, from such a point, approximately[3] 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.

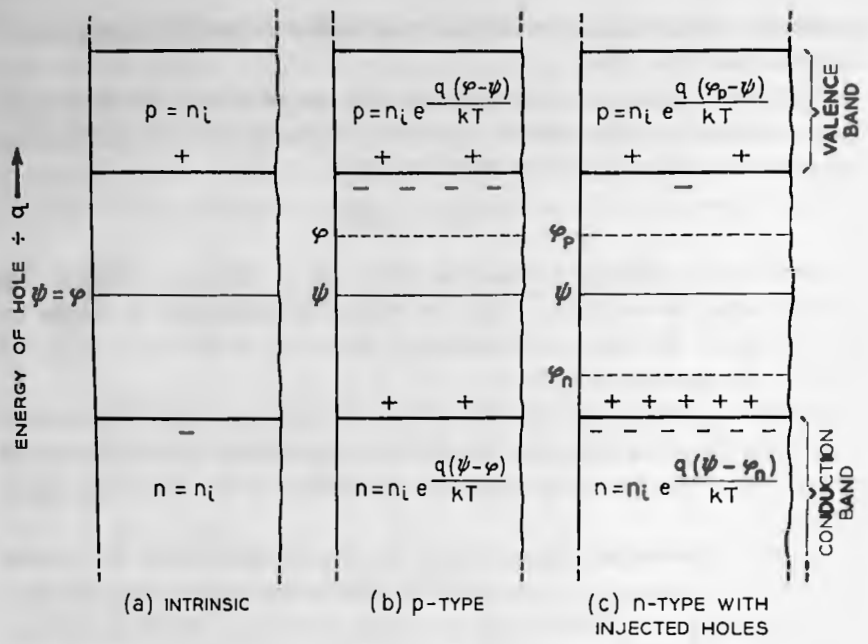


Fig. 2—Electrostatic potential ψ, Fermi level φ and quasi Fermi levels φ_p and φ_n.
(In order to show electrostatic potential and energies on the same ordinates, the energies of holes, which are minus the energies of electrons, are plotted upwards in the figures in this paper.)

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

$$\begin{aligned} p &= n_i e^{q(\varphi-\psi)/kT} && \text{(a)} \\ n &= n_i e^{q(\psi-\varphi)/kT}, && \text{(b)} \end{aligned} \tag{2.2}$$

where q is the electronic charge. Accordingly,

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

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

$$\begin{aligned} p &= n_i e^{q(\varphi_p-\psi)/kT} && \text{(a)} \\ n &= n_i e^{q(\psi-\varphi_n)/kT} && \text{(b)} \end{aligned} \tag{2.4}$$

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

$$\mu = qD/kT \tag{2.7}$$

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

$$\rho = \rho_d - \rho_i \sinh u \tag{2.8}$$

where

$$u \equiv q\psi/kT, \qquad \rho_i \equiv 2n_i q, \qquad \rho_d \equiv q(N_d - N_a) \tag{2.9}$$

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

$$\nabla^2\psi = -4\pi\rho/\kappa \tag{2.10}$$

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

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

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

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

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,

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

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 Region[7]

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

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

This relationship defines a characteristic length L_a given by

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

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

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

and

$$y = x/2L_a, \qquad K = L_D/2L_a \tag{2.17}$$

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,

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

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 $x = x_m$:

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

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

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

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)

$$n_m = n_i e^{q\psi_m/kT}. \tag{2.21}$$

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

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

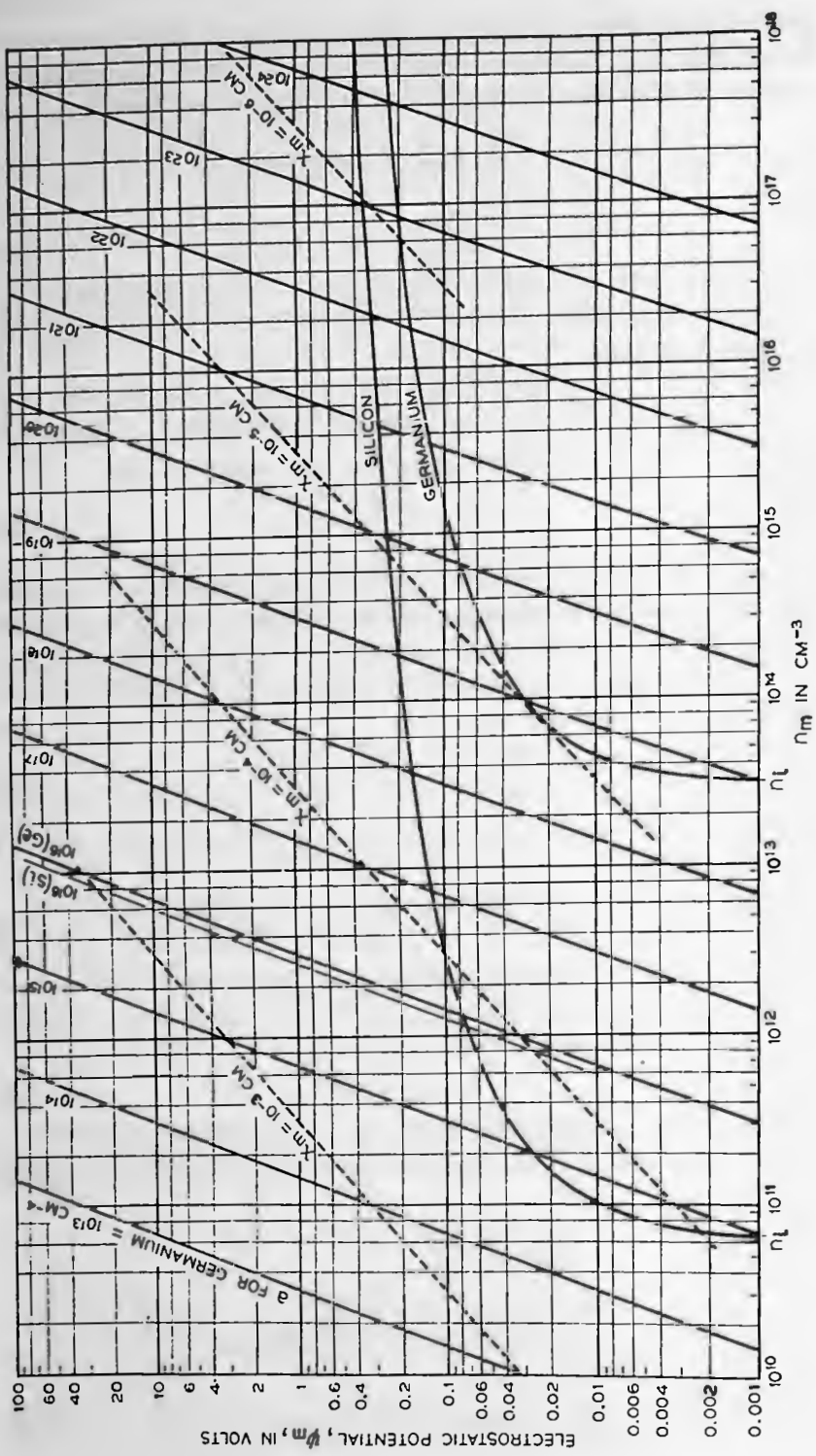


Fig. 3—Solutions for the boundaries of the space-charge region.

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 conductivity[8] and was about 10^{15} to 10^{16} cm^{-4}. For these, the space change 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:

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

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

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 $\delta\varphi$ 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-

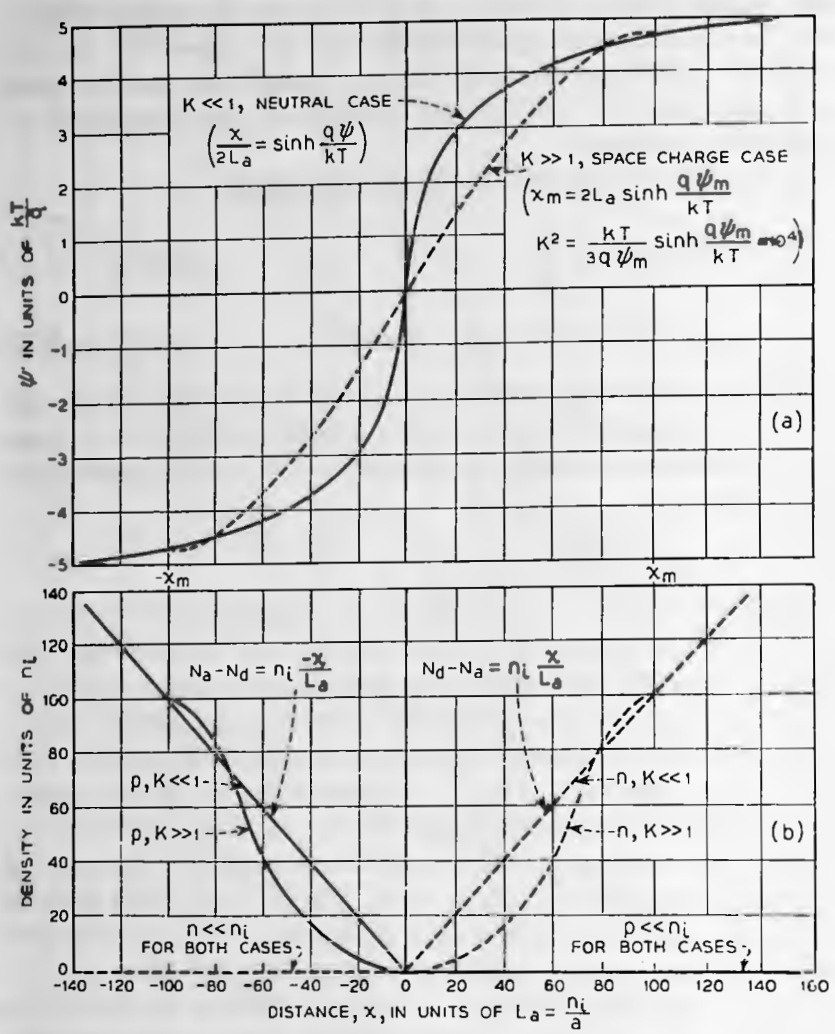


Fig. 4—Electrostatic potential and densities for p-n junctions.

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

$$\varphi_p - \varphi_n = \delta\varphi \tag{2.25}$$

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

We shall then have

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

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

where

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

and

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

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

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

and in the n-region

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

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

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

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

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

$$n = n_1 e^{+u} \tag{2.34}$$

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

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

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 n_1. 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,

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

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:

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

The second term in (2.36) is essentially the sum of the holes to 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

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

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

$$C = \frac{dQ}{d\delta\varphi} = \frac{q^2}{kT}\frac{n_1^2}{a}[\ln(n_b p_a/n_i^2) - (q\delta\varphi/kT) - 1]. \qquad (2.39)$$

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

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

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

$$C = \frac{\kappa[q(\psi_b - \psi_a - \delta\varphi)/kT - 1]\, e^{q\delta\varphi/kT}}{4\pi(2L_D^2/L_a)} \qquad (2.41)$$

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

$$C \cong \kappa/4\pi(4KL_D/10) \qquad (2.42)$$

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))

$$\delta\psi_m = -\delta\varphi/2 = 4\pi q a x_m^2 \delta x_m/\kappa \qquad (2.43)$$

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,

$$C = \delta Q/\delta\varphi = qax_m\delta x_m/\delta\varphi = \kappa/4\pi 2x_m \qquad (2.44)$$

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 ψ_{m0} is the value of ψ for $\delta\varphi = 0$, then $\psi_m = \psi_{m0} - \delta\varphi/2$; and C will vary as

$$C = \kappa[4\pi qa/3\kappa(\psi_{m0} - \delta\varphi/2)]^{1/3}/8\pi \tag{2.45}$$

so that $1/C^3$ should plot as a straight line versus $\delta\varphi$ with slope

$$-(8\pi/\kappa)^3(3\kappa/8\pi qa) = -\frac{192\pi^2}{\kappa^2 qa}. \tag{2.46}$$

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

$$C_{\text{sp. chg.}} = \kappa/8\pi x_m \text{ to } C_{\text{neut.}} = 5\kappa/8\pi KL_D \tag{2.47}$$

(i.e., from (2.44) to (2.42)) so that the ratio is

$$\frac{C_{\text{neut.}}}{C_{\text{sp. chg.}}} = \frac{5x_m}{KL_D} \tag{2.48}$$

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

$$\psi = (2\pi/\kappa)q\,p_p(x - x_p)^2 + \text{constant}, \qquad x < 0 \tag{2.49}$$

$$\psi = -(2\pi/\kappa)q\,n_n(x - x_n)^2 + \text{constant}, \qquad x > 0 \tag{2.50}$$

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

$$-p_p x_p = n_n x_n \tag{2.51}$$

so that if the total width of the space charge layers is $W = x_n - x_p$, it follows that

$$x_p = -n_n W/(n_n + p_p) \text{ and } x_n = p_p W/(n_n + p_p). \tag{2.52}$$

The potential difference across the layer, which is $\psi_b - \psi_a$ is

$$\psi_b - \psi_a = (2\pi q/\kappa)(p_p x_p^2 + n_n x_n^2) = [2\pi q\ p_p n_n/\kappa(p_p + n_n)]W^2 \tag{2.53}$$

If $p_p \gg n_n$ this reduces to

$$\psi_b - \psi_a = 2\pi q\ n_n W^2/\kappa \tag{2.54}$$

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

$$C = \kappa/4\pi W \tag{2.55}$$

where W is obtained by solving (2.53). For this case $(1/C)^2$ should plot as a straight line versus $\psi_b - \psi_a$:

$$\frac{1}{C^2} = [8\pi(p_p + n_n)/\kappa q\ p_p\ n_n](\psi_b - \psi_a). \tag{2.56}$$

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 rnp so that the net rate of generation per unit volume is

$$\text{(net rate of generation)} = g - rnp,$$

which vanishes at equilibrum. Obviously, $g = rn_i^2$. 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

$$\delta\dot{p} = \delta\dot{n} = g - r(n + \delta n)(p + \delta p) = -rn\delta p - rp\delta n \tag{3.1}$$

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

$$\frac{d^2\psi}{dx^2} = -\frac{4\pi q}{\kappa}\left(ax + n_i e^{q(\varphi_p-\psi)/kT} - n_i e^{q(\psi-\varphi_n)/kT}\right) \tag{3.4}$$

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

$$\begin{aligned}\frac{dI_p}{dx} &= -q\mu p\left[\frac{q}{kT}\left(\frac{d\varphi_p}{dx}\right)^2 - \frac{q}{kT}\frac{d\psi}{dx}\frac{d\varphi_p}{dx} + \frac{d^2\varphi_p}{dx^2}\right] \\ &= q(g - rnp) = qg(1 - e^{q(\varphi_p-\varphi_n)/kT}),\end{aligned} \tag{3.6}$$

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.

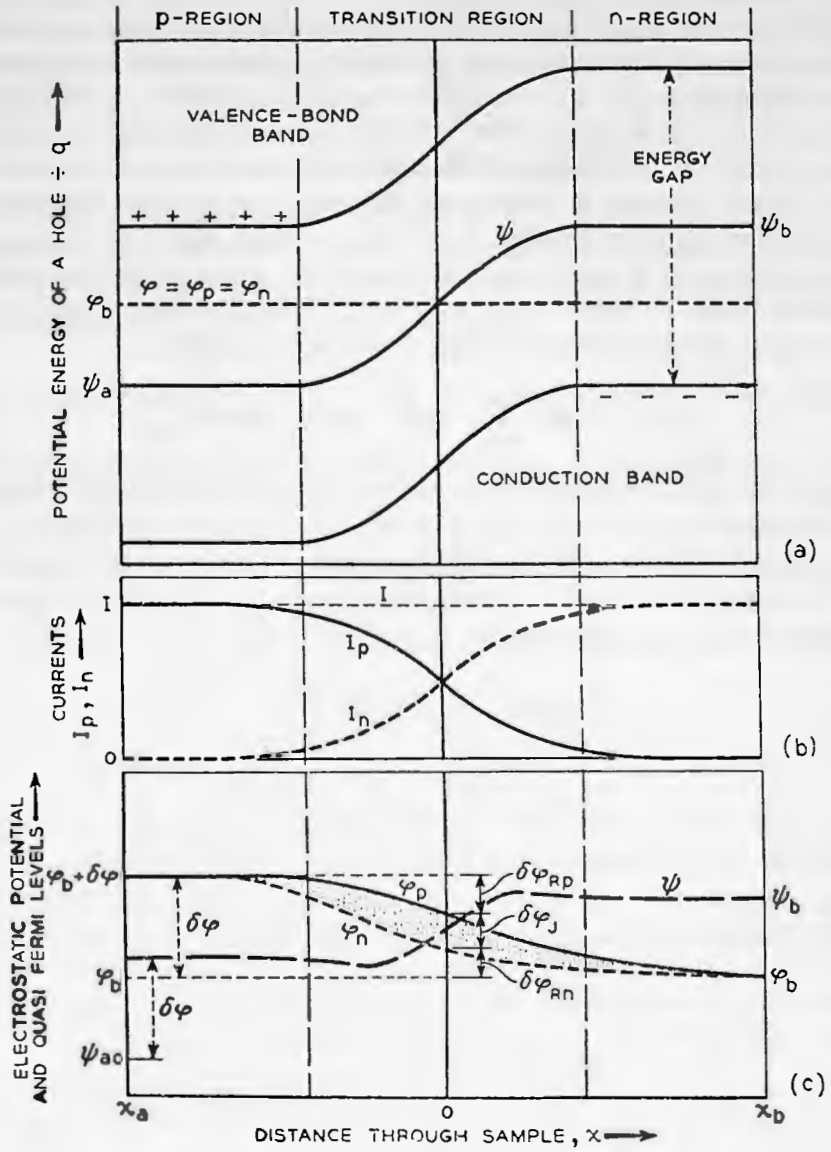


Fig. 5—Potential and current distributions for forward current in *p-n* junctions.
(a) *p-n* junction under equilibrium conditions.
(b) Division of current between holes and electrons.
(c) Distribution of potentials for forward current flow showing how the potential $\delta\varphi$ applied at x_a changes 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

$$I_p = -q\mu p \frac{d\varphi_p}{dx} \quad \text{and} \quad I_n = -qb\mu n \frac{d\varphi_n}{dx}. \tag{3.7}$$

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

$$\begin{aligned} I = I_p(x_a) - I_p(x_b) &= \int_{x_a}^{x_b} - dI_p \\ &= gq \int_{x_a}^{x_b} (e^{q(\varphi_p - \varphi_n)/kT} - 1)\, dx. \end{aligned} \tag{3.8}$$

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

$$I = I_p + I_n = -q\mu(p + bn)\, d\varphi_p/dx \tag{3.9}$$

and

$$\delta\varphi = -\int_{x_a}^{x_b} d\varphi_p = I \int_{x_a}^{x_b} dx/q\mu(p + bn) = IR_0, \tag{3.10}$$

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

$$\begin{aligned} I &= gq(L_n + L_p)[e^{(q\delta\varphi_J/kT)} - 1] \\ &= I_s[e^{(q\delta\varphi_J/kT)} - 1] \end{aligned} \tag{3.11}$$

where

$$I_s = gq(L_n + L_p) \tag{3.12}$$

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

$$\delta\varphi = \delta\varphi_{Rp} + \delta\varphi_{Rn} + \delta\varphi_J = R_1 I + \frac{kT}{q} \ln\left(1 + \frac{I}{I_s}\right). \tag{3.13}$$

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

$$\begin{aligned} R_1 &= \int_{L_a}^{-x_a} dx/q\mu ax + \frac{2L_a}{q\mu n_i(1 + b)} + \int_{L_a}^{x_b} dx/q\mu bax \\ &= \frac{L_a}{q\mu n_i}\left(1 + \frac{1}{b}\right) \ln\,(x_b/L_a) \end{aligned} \tag{3.14}$$

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

$$\delta\varphi/I = R_1 + (kT/qI_s) = \frac{L_a}{q\mu n_i}\left(1 + \frac{1}{b}\right)\ln(x_b/L_a) + kT/(q^2 g^{2/3}(DL_a n_i)^{1/3}(1 + b^{1/3})). \tag{3.16}$$

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 $\delta\varphi_{Rp}$ 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 formula[11] 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 cm^2/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

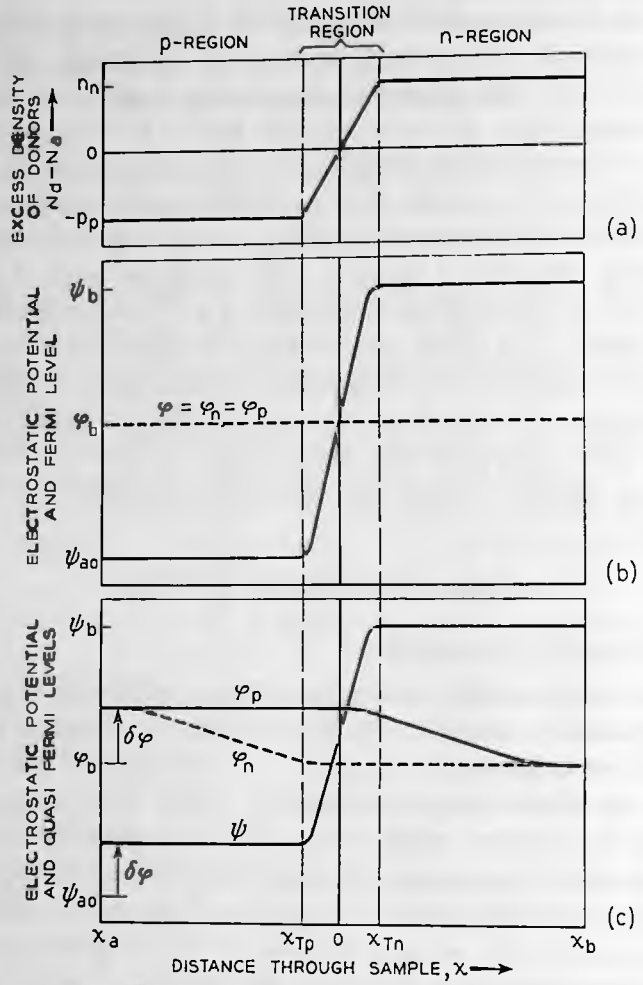


Fig. 6—Simplified model of a p-n junction.
(a) Distribution of donors and acceptors.
(b) Potentials for thermal equilibrium.
(c) Effect of $\delta\varphi$ applied potential in forward direction.

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

$$\begin{aligned} p(x_{Tn}) &= n_i e^{q(\varphi_b+\delta\varphi-\psi_b)/kT} \\ &= p_n e^{q\delta\varphi/kT} \end{aligned} \tag{4.1}$$

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

$$\delta\varphi = v_0 + v_1 e^{i\omega t} \tag{4.2}$$

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 $p_1 e^{i\omega t}$:

$$p(x_{Tn}) = p_n + p_0 + p_1 e^{i\omega t} \tag{4.3}$$

where

$$p_0 = p_n(e^{qv_0/kT} - 1) \tag{4.4}$$

$$p_1 = (qp_n v_1/kT)e^{qv_0/kT}. \tag{4.5}$$

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

$$p = p_n + p_0 e^{(x_{Tn}-x)/\sqrt{D\tau_p}} + p_1 e^{t\omega t+(x_{Tn}-x)(1+i\omega\tau_p)^{1/2}/(D\tau_p)^{1/2}}. \tag{4.8}$$

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 :

$$\varphi_p = \varphi_n + v_c - (kT/q)(x - x_{Tn})/L_p + \tau_1 e^{i\omega t-(x-x_{Tn})[(1+i\omega\tau_p)^{1/2}-1]/L_p}. \tag{4.9}$$

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

$$\begin{aligned} I_p(x_{Tn}) &= -qD\frac{\partial p}{\partial x} \\ &= qp_0 D/L_p + qp_1 De^{i\omega\tau}(1 + i\omega\tau_p)^{1/2}/L_p. \end{aligned} \tag{4.10}$$

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 :

$$\begin{aligned} I_{p1}(x_{Tn}) &= (qp_n\mu/L_p)[e^{(qv_0/kT)}](1 + i\omega\tau_p)^{1/2} v_1 e^{i\omega t} \\ &\equiv (G_p + iS_p)v_1 e^{i\omega t} \equiv A_p v_1 e^{t\omega t} \end{aligned} \tag{4.12}$$

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 cm² of a layer L_p cm thick with hole conduction corresponding to the density $p_n + p_0$; 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_{p0} 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_{n0} or $I_{n0}(x_{\tau p})$, is obtained by the same procedure as that leading to (4.11) for I_{p0} 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:

$$I_0 = I_{p0} + I_{n0} = (qD)\left(\frac{p_n}{L_p} + \frac{bn_p}{L_n}\right)(e^{qv_0/kT} - 1) \qquad (4.13)$$

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

$$\begin{aligned} \frac{I_{p0}}{I_{n0}} &= \frac{p_n}{L_p} \cdot \frac{L_n}{bn_p} = \frac{p_p}{bn_n} \cdot \frac{\sqrt{bD\tau_n}}{\sqrt{D\tau_p}} \\ &= \frac{p_p}{bn_n}\sqrt{\frac{bn_n}{p_p}} = \sqrt{\frac{\sigma_p}{\sigma_n}} \end{aligned} \qquad (4.14)$$

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; $\sigma_n = q\mu bn_n$ and $\sigma_p = q\mu p_p$ 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

$$q\mu p_n = b\sigma_i^2/(1 + b)^2\sigma_n \tag{4.16}$$

$$q\mu b n_p = b\sigma_i^2/(1 + b)^2\sigma_p . \tag{4.17}$$

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

$$\begin{aligned} I_0(v_0) &= [G_{p0} + G_{n0}]\left(\frac{kT}{q}\right)[e^{qv_0/kT} - 1] \\ &= (I_{ps} + I_{ns})[e^{qv_0/kT} - 1]. \end{aligned} \tag{4.22}$$

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.

$$I_{ac} = (G_p + iS_p + G_n + iS_n + i\omega C_T)\, v_1 = A v_1 \tag{4.23}$$

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

$$A_p = G_p + iS_p = (1 + i\omega\tau_p)^{1/2} G_{p0}\, e^{qv_0/kT} \tag{4.24}$$

$$A_n = G_n + iS_n = (1 + i\omega\tau_n)^{1/2} G_{n0}\, e^{qv_0/kT} \tag{4.25}$$

For low frequencies, such that ω is much less then $1/\tau_p$, we can expand $G_p + iS_p$ as follows:

$$G_p + iS_p = G_{p0}\, e^{qv_0/kT} + i\omega(\tau_p/2) G_{p0}\, e^{qv_0/kT} \tag{4.26}$$

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:

$$C_p = \tau_p G_{p0}/2 = \tau_p q p_n \mu/2L_p = q^2 p_n L_p/2kT. \tag{4.27}$$

The last formula, obtained by noting that $\tau_p\mu = q\tau_p D/kT = qL_p^2/kT$, has a simple interpretation: qp_nL_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 $(q/kT)(qp_nL_p)$ 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:

$$\frac{C_p}{C_{T\,\text{neut.}}} = \frac{q^2 p_n L_p}{2kT} \cdot \frac{kTa}{10q^2 n_i^2} = \frac{p_n L_p a}{20 n_i^2}. \tag{4.28}$$

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

$$\frac{C_p}{C_{T\,\text{neut.}}} = \frac{L_p}{20 x_{Tn}} \tag{4.29}$$

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 capacitative 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

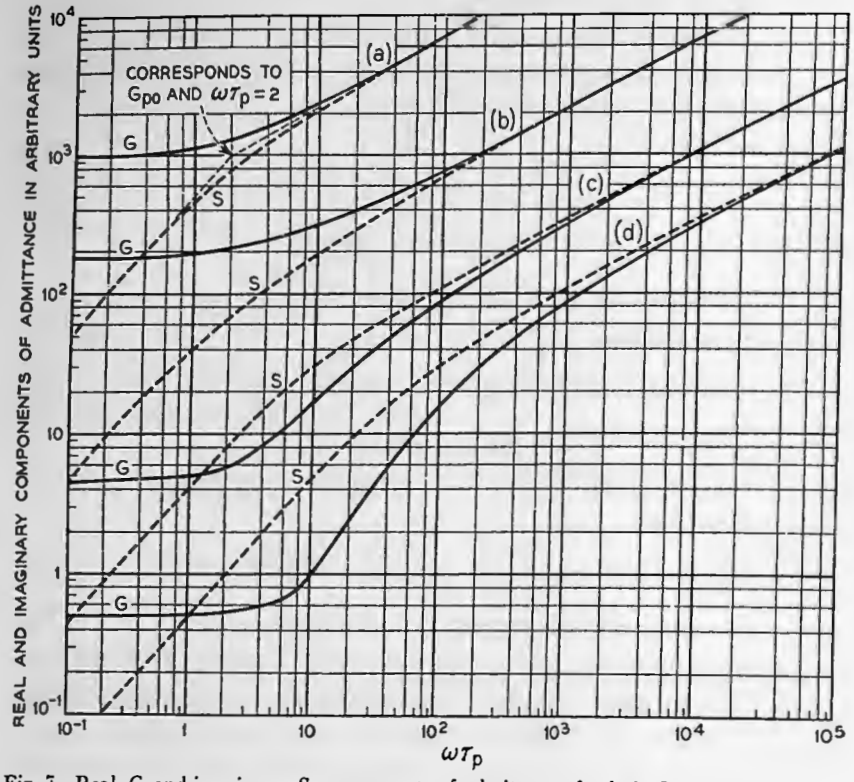


Fig. 7—Real, G, and imaginary, S, components of admittance for hole flow into n-region.
(a) $10^3 A_p/G_{p0} = 10^3(1 + i\omega\tau_p)^{1/2}$ corresponding to uniform n-region.
(b) $10^2 \times$ Formula of Appendix III, corresponding to layer of high recombination rate in front of n-region. This causes 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:

$$G_p = S_p = \sqrt{\tau_p/2}\, G_{p0} \sqrt{\omega} = \frac{b\sigma_i^2 \sqrt{\omega}}{(1+b)^2 \sigma_n \sqrt{2D}} \qquad (4.30)$$

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 is ωC_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

$$\frac{\tau_T}{\tau_p} = \frac{(x_{Tn} - x_{pn})^2}{100 L_p^2}. \tag{4.31}$$

Hence if $L_p > x_{Tn} - x_{pn}$, 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

$$A_p = [q\mu p_n/L_p] \frac{(2L_r/L_p)(1 + i\omega\tau_p)}{1 + [1 + (1 + i\omega\tau_p)(2L_r/L_p)^2]^{1/2}} \tag{4.32}$$

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

$$A_p/[2q\mu p_n L_r/L_p^2] = \frac{(1 + i\omega\tau_p)}{1 + [1 + (1 + i\omega\tau_p)(2L_r/L_p)^2]^{1/2}} \tag{4.33}$$

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

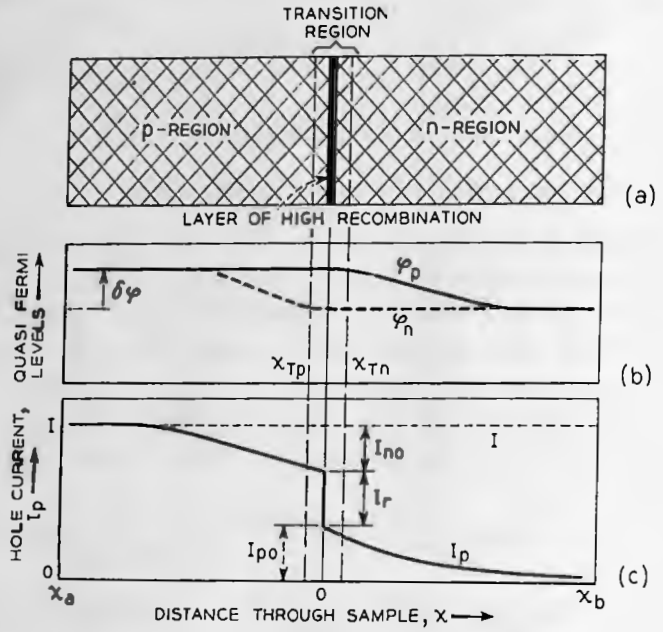


Fig. 8—The effect of a localized layer of high recombination rate on the junction characteristic.
(a) Location of layer of high recombination rate.
(b) Quasi Fermi levels.
(c) Distribution of hole current showing rapid change at layer of high recombination rate.

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 $r_a np$ 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

$$g_a - r_a np = g_a [1 - e^{q(\varphi_p - \varphi_n)/kT}] \tag{4.34}$$

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

$$I_r(\varphi_p - \varphi_n) = q g_a [e^{q(\varphi_p - \varphi_n)/kT} - 1] \tag{4.35}$$

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 $x_{Tp} < x < xT_n$, 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

$$I = I_{n0}(\delta\varphi) + I_{p0}(\delta\varphi) + I_r(\delta\varphi). \tag{4.36}$$

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 $\frac{1}{9}$ 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 $(\frac{1}{3})\sqrt{D\tau_\ell}$; this corresponds to $\omega\tau_\ell > 9$ 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

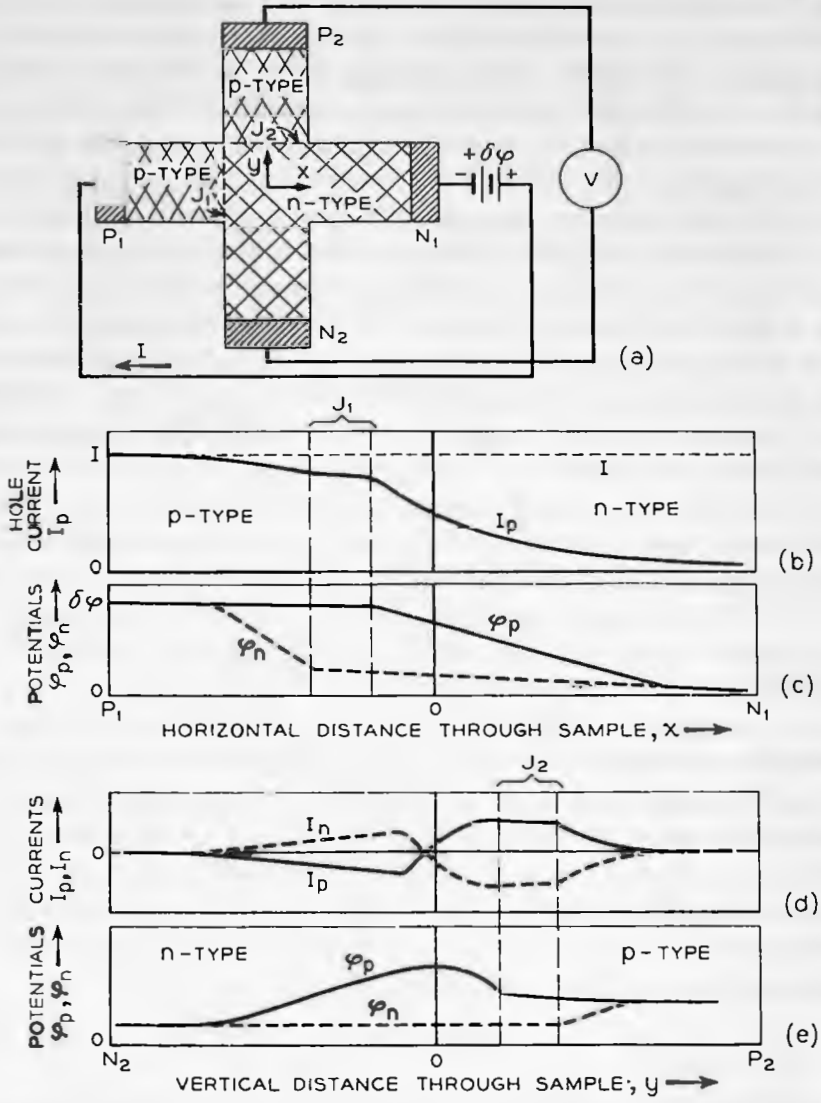


Fig. 9—Internal contact potentials showing how presence of injected holes produces a contact potential across J_2.

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

$$p_1 = p_n e^{q\varphi_e/kT} \qquad x = -w \tag{5.2}$$

$$p_2 = p_n e^{q\varphi_c/kT} \qquad x = +w \tag{5.3}$$

The function $p(x)$ which satisfies these boundary conditions and the equation

$$D\frac{d^2p}{dx^2} - \frac{p - p_n}{\tau_p} = 0 \tag{5.4}$$

is

$$p(x) = p_n + \frac{p_1 + p_2 - 2p_n}{2\cosh(w/L_p)}\cosh(x/L_p) + \frac{p_2 - p_1}{2\sinh(w/L_p)}\sinh(x/L_p) \tag{5.5}$$

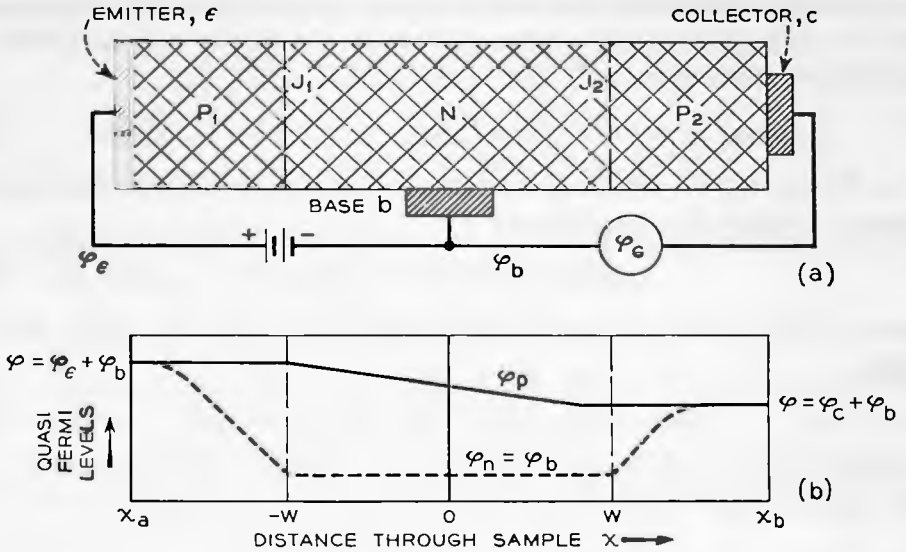


Fig. 10—Model used for calculation of internal contact potential and to illustrate *p-n-p* transistor.
(a) Semiconductor with two *p-n* junctions and ohmic metal contacts.
(b) Quasi Fermi levels showing internal contact potential between *b* and *c*.

which gives rise to a hole current across J_2 into P_2 of amount

$$\begin{aligned} I_p &= -qD\frac{dp}{dx}\Big|_{x=w} \\ &= \frac{qD}{2L_p}\left[(p_1 - p_2)\coth\frac{w}{L_p} + (2p_n - p_1 - p_2)\tanh\frac{w}{L_p}\right] \\ &= \frac{qD}{2L_p}\left[(p_1 - p_n)\left(\coth\frac{w}{L_p} - \tanh\frac{w}{L_p}\right)\right. \\ &\qquad \left. - (p_2 - p_n)\left(\coth\frac{w}{L_p} + \tanh\frac{w}{L_p}\right)\right] \\ &= +\frac{p_n qD}{L_p}\left[\frac{[e^{q\varphi/kT} - 1]}{\sinh(2w/L_p)} - \frac{(e^{q\varphi_c/kT} - 1)}{\tanh(2w/L_p)}\right] \\ &= \operatorname{csch}(2w/L_p)I_{p0}(\varphi_e) - \coth(2wL_p)I_{p0}(\varphi_c) \end{aligned} \tag{5.6}$$

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 csch $(2w/L_p)$ of the current $I_{p0}(\varphi_e)$, which would be injected by φ_e 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 $\coth(2w/L_p)$.

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 $+I_{p0}(\varphi_e)$ term for equilibrium. Hence:

$$I_{n0}(\varphi_c) + \coth\,(2w/L_p)\, I_{p0}(\varphi_c) = \operatorname{csch}\,(2w/L_p)\, I_{p0}(\varphi_e) \tag{5.7}$$

If $p_n \gg n_p$, the I_{n0} term can be neglected compared to $\coth\,(2w/L_p)\, I_{p0}$. Hence the value of φ_c must satisfy

$$I_{p0}(\varphi_c) = \operatorname{sech}\,(2w/L_p)\, I_{p0}(\varphi_e). \tag{5.8}$$

For $\varphi_e > kT/q$, the exponential approximation may be used for I_{p0} in both terms:

$$\varphi_c = \varphi_e - (kT/q) \ln \cosh\,(2w/L_p), \tag{5.9}$$

so that, if $(2w/L_p)$ is the order of unity, φ_c should be only about (kT/q) less than φ_e. For $(2w/L_p)$ large, we get

$$\varphi_c = \varphi_e - (kT/q)\,(2w/L_p) \tag{5.10}$$

corresponding to the linear drop of φ_p, discussed in connection with equation (4.9), across the distance $2w$.

When φ_e 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

$$[\exp q\varphi_c/kT] - 1 = -\operatorname{sech}\,(2w/L_p) \tag{5.11}$$

and corresponding to saturation reverse current across J_1, so that

$$\varphi_c = -(kT/q) \ln [1 + (1/2) \operatorname{csch}^2(w/L_p)]. \tag{5.12}$$

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

$$I_c(\varphi_c, \varphi_e) = I_{nc}(\varphi_c) + \coth \frac{2w}{L_p} I_{pc}(\varphi_c) - \operatorname{csch} \frac{2w}{L_p} I_{p0}(\varphi_e). \quad (5.13)$$

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

$$\begin{aligned} I_b = -I_e - I_c &= \left[\operatorname{csch} \frac{2w}{L_p} - \coth \frac{2w}{L_p}\right] [I_{p0}(\varphi_e) + I_{pc}(\varphi_c)] \\ &= \frac{2 \sinh^2 w/L_p}{\sinh 2w/L_p} [I_{p0}(\varphi_e) + I_{p0}(\varphi_c)]. \end{aligned} \quad (6.3)$$

[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. Jl.*, 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 $I_b = -[I_{p0}(\varphi_e) + I_{p0}(\varphi_c)]$.

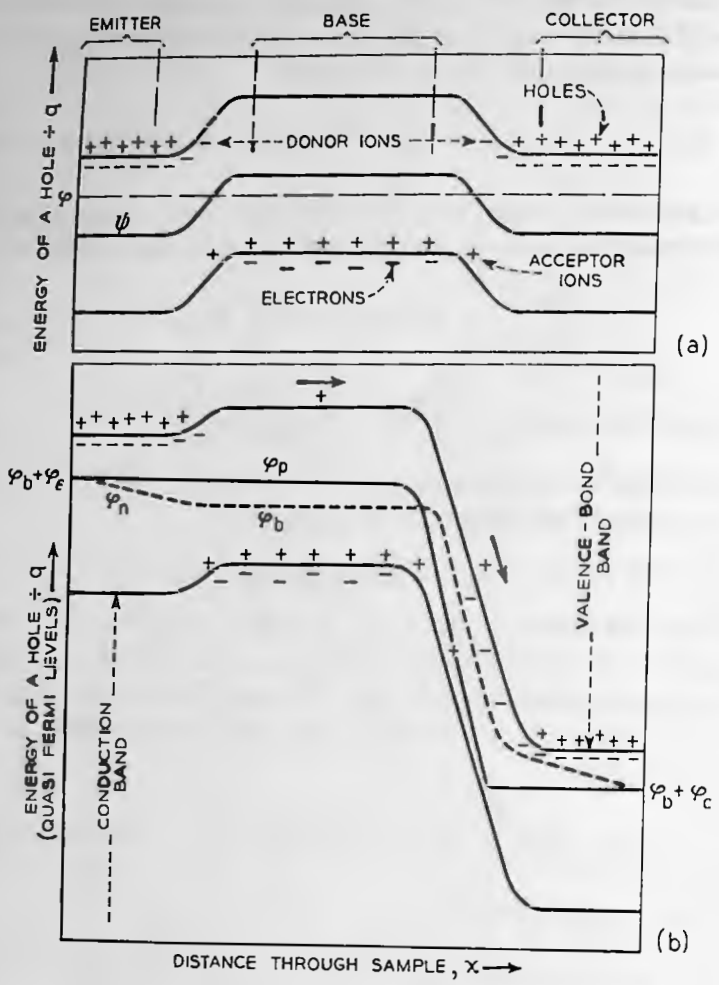


Fig. 11—p-n p transistor.
(a) Thermal equilibrium
(b) Operating condition.

If φ_c is several volts negative, so that $I_{p0}(\varphi_c)$ has its saturation va ue I_{p_s} (see (4.11) and (4.20)), then the ratio $-\delta I_c/\delta I_e = \alpha$ has the value

$$\alpha = -\frac{\delta I_c}{\delta I_e} = \frac{\operatorname{csch} \dfrac{2w}{L_p}}{\coth \dfrac{2w}{L_p}} = \operatorname{sech} \frac{2w}{L_p}. \tag{6.4}$$

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 formula[16] $\alpha^2 R_{22}/R_{11}$ 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_ϵ. 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

$$I\delta\varphi = -(I_p\varphi_p + I_n\varphi_n)\Big|_{x_a}^{x_b}$$

$$= -\int_{x_a}^{x_b} \frac{d}{dx}(I_p\varphi_p + I_n\varphi_n)\,dx$$

$$= -\int_{x_a}^{x_b}\left(\frac{dI_p}{dx}\varphi_p + \frac{dI_n}{dx}\varphi_n\right)dx - \int_{x_a}^{x_b}\left(I_p\frac{d\varphi_p}{dx} + I_n\frac{d\varphi_n}{dx}\right)dx.$$

Since

$$\frac{dI_p}{dx} = -\frac{dI_n}{dx} = qg(1 - e^{q(\varphi_p-\varphi_n)/kT})$$

and

$$\frac{d\varphi_p}{dx} = -I_p/q\mu p, \qquad \frac{d\varphi_n}{dx} = -I_n/q\mu n$$

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:

$$D\left[\frac{\partial^2 p}{\partial x^2} + \frac{1}{L_r}\frac{\partial p}{\partial x}\right] - \frac{1}{\tau_p}(1 + i\omega\tau_p)p$$

$$= -\gamma D\left[-\gamma + \frac{1}{L_r}\right]p - \frac{1}{\tau_p}(1 + i\omega\tau_p)p = 0$$

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 $p = p_1 \exp(i\omega) = (p_n q v_1/kT)\exp(i\omega t)$ 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

$$
\begin{aligned}
p_1 &= p_{10} + (A\, e^{-\alpha x} + B\, e^{+\alpha x})\, e^{i\omega t} \qquad x < 0 \\
p_2 &= p_{20} + C\, e^{-\beta x + i\omega t} \qquad x > 0
\end{aligned}
$$

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:

$$A + B = Ce^{q(\psi_1-\psi_2)/kT} \equiv CF$$

$$\alpha(A - B) = \beta C$$

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

$$-\frac{\partial \ln p}{\partial x} = \frac{\alpha(Ae^{+\alpha c} - Be^{-\alpha c})}{(Ae^{+\alpha c} + Be^{-\alpha c})} = \frac{\alpha(F\alpha \sinh \alpha c + \beta \cosh \alpha c)}{F\alpha \cosh \alpha c + \beta \sinh \alpha c}.$$

Since at $x = -c$, the a-c. component of p_1 is $(qv_1/kT)p_{10}e^{i\omega t}$, the admittance is

$$A_p = \frac{-qD\partial p/\partial x}{v_1 e^{i\omega t}} = (q^2 Dp_{10}/kT)(-\partial \ln p/\partial x)$$

$$= (q\mu p_{10}/L_1)(1 + i\omega\tau_1)^{1/2} \frac{F\alpha \sinh \alpha c + \beta \cosh \alpha c}{F\alpha \cosh \alpha c + \beta \sinh \alpha c}.$$

For $c \to 0$, 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

$$\psi = \frac{kT}{q}\left(\frac{x}{L_r} + \frac{xe^{i\omega t}}{L_1}\right)$$

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)$/\partial x$ 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}: $$D\left[p_0\frac{1}{L_r^2} - p_0\frac{1}{L_r^2}\right] = 0$$

$e^{-x/L_r + i\omega t}$: $$D\left[p_1\frac{1}{L_r^2} - p_1\frac{1}{L_r^2} - \frac{1}{L_r L_1}p_0\right] - i\omega p_1 = 0$$

$e^{-\gamma x + i\omega t}$: $$D[\gamma^2 - \gamma/L_r]p_1 - i\omega p_1 = 0$$

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 qp_0L_r is the total charge in the layer L_r, (qV/kT) is an average fractional change in this charge for V so that $(qp_0L_r)\ (qV/kT) \div V$ 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

$$p_1(0, y, z, t) = p_{10}\, e^{i\omega t} = (p_n\, qv_1/kT)e^{i\omega t}$$

by reasoning similar to that used for equation (4.5). The boundary condition at the surface will be

$$-D\,\frac{\partial p_1}{\partial y} = sp_1 \qquad \text{for } y = +w$$

This states that the recombination per unit area is sp_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

$$p_1 = \sum_{i,j=0}^{\infty} a_{ij}\, e^{-\alpha_{ij}x + i\omega t} \cos\beta_i y \cos\beta_j z$$

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

$$\alpha_{ij} = (1 + i\omega\tau_{ij})^{\frac{1}{2}}/(D\tau_{ij})^{\frac{1}{2}}$$

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

$$I_1 = qDp_{10}e^{i\omega t} \Sigma \alpha_{ij}(a_{ij}/p_{10})(4w^2/\theta_i\theta_j) \sin \theta_i \sin \theta_j$$

Substituting for a_{ij} and inserting $p_{10} = p_n q v_1/kT$, we obtain an expression for the admittance $A_p = I_1/V_1 \exp(i\omega t)$:

$$A_p = 4w^2 q\mu p_n \Sigma_{ij} \alpha_{ij} \frac{4 \sin^2 \theta_i \sin^2 \theta_j}{\theta_i^2 \theta_j^2 \left[1 + \left(\frac{1}{2\theta_i}\right) \sin 2\theta_i\right]\left[1 + \left(\frac{1}{2\theta_j}\right) \sin 2\theta_j\right]}$$

where the sum plays the role formerly taken by $(1 + i\omega t)^{1/2}/\sqrt{D\tau}$ 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

$$\theta_0 \doteq \sin \theta_0 \doteq \tan \theta_0 \doteq \sqrt{\chi}$$

$$\theta_1 \doteq \pi + \chi/\pi; \; -\sin \theta_1 \doteq \tan \theta_1 \doteq \chi/\pi$$

. . . . . .

$$\theta_n \doteq n\pi + \chi/n\pi; \; (-1)^n \sin \theta_n \doteq \tan \theta_n \doteq \chi/n\pi$$

From this we see that the terms in the sum are as follows:

$$\alpha_{00} \cdot 4\chi^2/\chi^2 4 = \alpha_{00}$$

$$\alpha_{n0} \cdot 2(\chi/n\pi)^2/(n\pi)^2 = \alpha_{n0} 2\chi^2/n^4\pi^4$$

$$\alpha_{nm} \cdot 4\chi^4/n^4 m^4 \pi^8$$

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

$$4w^2 q\mu p_0 \, (1 + i\omega\tau_{00})^{1/2}/\sqrt{D\tau_{00}}$$

where

$$1/\tau_{00} = 2\left(\frac{D}{w^2}\right)(\theta_0^2) + 1/\tau$$

$$\doteq 2\left(\frac{D}{w^2}\right)\frac{sw}{D} + 1/\tau$$

$$= 2\left(\frac{s}{w}\right) + 1/\tau$$

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 $4w^2 q\mu p_0 (i\omega/D)^{1/2}$, 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 $\alpha \doteq 10^{-4}$ and for silicon $\alpha \doteq 0.1$ 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 capacitative 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

$$D\gamma^2 = i\omega[1 + \alpha] + \nu_r + \alpha\nu_{ra}$$

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

$$\frac{d^2u}{dx^2} = \sinh u - f(x) \tag{A7.1}$$

to which it can be readily reduced. We shall deal with the case for which

$$f = f_a \text{ for } x < x_a \tag{A7.2}$$

$$f = f_b \text{ for } x > x_b > x_a \tag{A7.3}$$

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 $|x|$

$$u = u_a + A_a e^{+\gamma_a x} \qquad x \longrightarrow -\infty \tag{A7.4}$$

$$u = u_b + A_b e^{-\gamma_b x} \qquad x \longrightarrow +\infty \tag{A7.5}$$

where

$$\sinh u_a = f_a, \qquad \sinh u_b = f_b$$

$$\gamma_a = |(\cosh u_a)^{1/2}|, \qquad \gamma_b = |(\cosh u_b)^{1/2}|$$

(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 $u'(x, A_\alpha)$ $(= du/dx)$

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 $u(x_a, A_a)$ $u'(x_a, A_a)$ 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,

$$u(x, A_a) = u(x, A_b) \tag{A7.6}$$

$$u'(x, A_a) = u'(x, A_b) \tag{A7.7}$$

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

$$\frac{d^2u}{dy^2} = \frac{1}{K^2}(\sinh u - \sinh u_0). \tag{A7.8}$$

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 $u \to u_0$ for $y \to \pm\infty$ 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 $y = 0$.

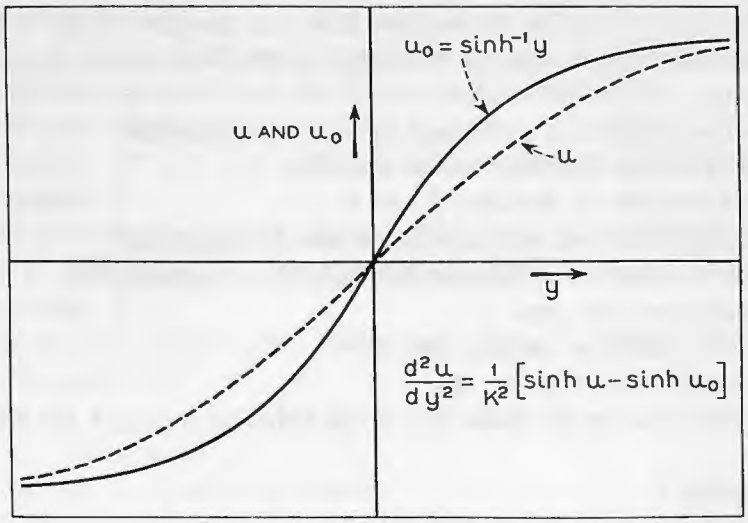


Fig. A1—Behavior of the solution of Equation (2.16) or (A7.8).

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

$$(1/K^2) \int_{-\infty}^{y} (\sinh u - \sinh u_0)\, dy = \frac{du}{dy} \qquad \text{(A7.9)}$$

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 $K \to 0$, 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 $u'(y_1)$, 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)

$a = (N_d - N_a)/x$ (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)

$e = 2.718\ldots$

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 °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 direction (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)

$\chi = sw/D$ in Appendix V

ψ = electrostatic potential (2.2)

ω = circular frequency of a-c. (4.2)

THEORY AND EXPERIMENT FOR A GERMANIUM p - n JUNCTION

F. S. GOUCHER, G. L. PEARSON, M. SPARKS
G. K. TEAL, W. SHOCKLEY

PHYSICAL REVIEW VOL.81 NO.4 pp.637-638 1951

Produced outside the work performed under this contract.

Theory and Experiment for a Germanium $p-n$ Junction

F. S. Goucher, G. L. Pearson, M. Sparks, G. K. Teal, and W. Shockley
Bell Telephone Laboratories, Murray Hill, New Jersey
December 21, 1950

RECTIFYING $p-n$ junctions in germanium have been produced in which the approach to the idealized conditions is so close that most of the expected theoretical features can be

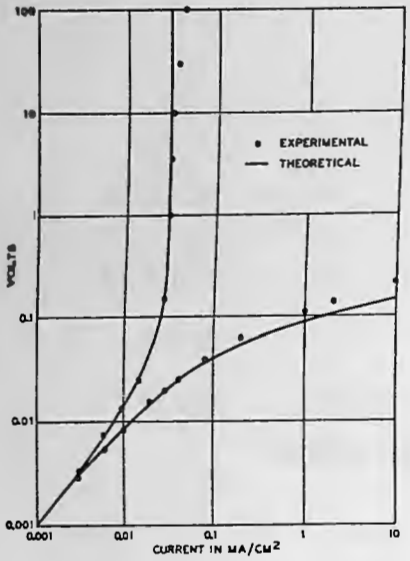


FIG. 1. Rectification characteristics of a $p-n$ junction.

quantitatively verified experimentally, thus putting these junctions in a unique position in the field of solid rectifiers. The junctions discussed in this letter were produced in high back voltage n-type germanium by the addition of gallium so that one portion of the single crystal[1] was p-type. The unit reported here was 0.65×0.6 cm in cross section.

According to theory[2] the rectification curve for such a junction should be of Wagner[3] form $I=I_s[\exp(eV/kT)-1]$. Figure 1 shows a comparison between theory and experiment, the theoretical value 39 volt^{-1} being used for e/kT. The voltage values were measured between zero current probes near to the junction on each side and employed to eliminate ohmic series resistance.

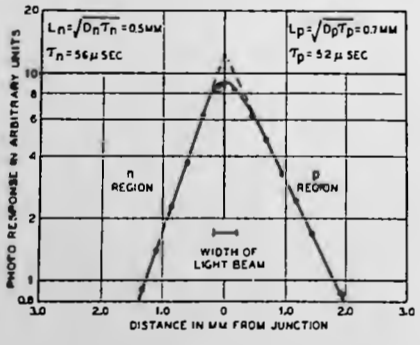


FIG. 2. Photo-response *vs* distance from junction.

The conductance of the junction at low voltages, eI_s/kT, is due to hole flow in the n-region in parallel with electron flow in the p-region. It can be calculated in terms of the intrinsic conductivity σ_i taken as 0.0165 ohm^{-1} cm^{-1}, the conductivities of the two regions σ_p and σ_n, the lifetimes of injected carriers τ_p for holes and τ_n for electrons, and the diffusion constants D_p and D_n, and their ratio $b=D_p/D_n$. The lifetimes were measured by scanning with a slit of light of wavelength 1.85 microns, which penetrated deeply into the specimen. The abnormal carriers so produced should diffuse[4] a distance x to the junction with decay factors of $\exp(-x/L)$, where $L=(D\tau)^{\frac{1}{2}}$ is the diffusion length, a prediction which is in agreement with Fig. 2. Except for the decay due to finite τ, it was determined that each photon absorbed resulted in a transfer of one electron across the junction, in agreement with the previous finding of unit quantum efficiency.[5] This treatment of light as a voltage-independent current generator is in agreement with the findings of Pietenpol[6] and differs in emphasizing diffusion and photo-current rather than photo-voltage which is usually discussed.[7]

From the lifetimes of Fig. 2 and values of $\sigma_p=2.5$ and $\sigma_n=0.188$ determined by probe measurements, a value of 0.007 for eI_s/kT was found, compared with a dc low voltage value of 0.011 ohm^{-1} cm^{-1}.

A further check of the lifetime results was obtained from the impedance characteristic at zero bias, Fig. 3, which is in good agreement with the theoretical form based on diffusion theory.[2]

For reverse biases, the diffusion term is suppressed and a capacity corresponding to a concentration gradient in the transition region was found. From Eq. (2.46) of reference 2 a value of 10^{15} cm^{-4} was obtained.

The dependence of admittance on bias for intermediate biases has also been found to be in agreement with theory as has the effect of temperature. The dependence of the photo-response on

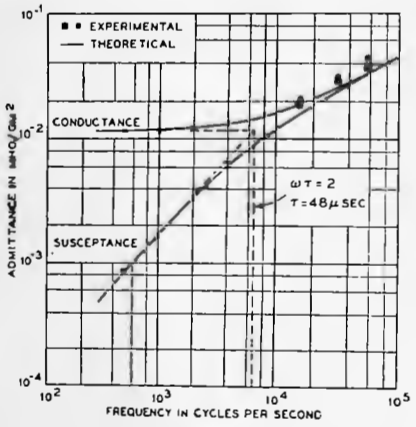


FIG. 3. Admittance of $p-n$ junction *vs* frequency.

wavelength has been used for the study of surface recombination, which is more important at the short wavelengths which do not penetrate deeply.

We are indebted to W. L. Bond and H. R. Moore for aid with the experimental techniques, and to J. A. Morton for encouragement in connection with the germanium single-crystal program.

[1] G. K. Teal and J. B. Little, Phys. Rev. 78, 647 (1950).
[2] W. Shockley, Bell System Tech. J. 28, 435–489 (1949). W. Shockley, *Electrons and Holes in Semiconductors* (D. Van Nostrand Company, New York, 1950).
[3] C. Wagner, Physik. Z. 32, 641 (1931).
[4] F. S. Goucher, Phys. Rev. 81, 475 (1951).
[5] F. S. Goucher, Phys. Rev. 78, 816 (1950).
[6] Personal communication.
[7] H. Y. Fan, Phys. Rev. 75, 1631 (1949) and reference 3 for additional publications.

THE p - n JUNCTION TRANSISTORS

W. SHOCKLEY, M. SPARKS, G. K. TEAL

PHYSICAL REVIEW VOL.83 NO.1 pp.151-162 JULY, 1951

Produced outside the work performed under this contract

p-n Junction Transistors

W. Shockley, M. Sparks, and G. K. Teal
Bell Telephone Laboratories, Inc., Murray Hill, New Jersey
(Received May 24, 1951)

The effects of diffusion of electrons through a thin *p*-type layer of germanium have been studied in specimens consisting of two *n*-type regions with the *p*-type region interposed. It is found that potentials applied to one *n*-type region are transmitted by diffusing electrons through the *p*-type layer although the latter is grounded through an ohmic contact. When one of the *p-n* junctions is biased to saturation, power gain can be obtained through the device. Used as "*n-p-n* transistors" these units will operate on currents as low as 10 microamperes and voltages as low as 0.1 volt, have power gains of 50 db, and noise figures of about 10 db at 1000 cps. Their current-voltage characteristics are in good agreement with the diffusion theory.

I. INTRODUCTION

IN this article we shall consider the phenomena which occur when voltages are applied to a semiconductor consisting of several regions of different conductivity types. Structures of this sort, consisting in particular of two regions of one conductivity type separated by another region of the opposite type, are of great practical interest in transistor electronics. Such structures can also be used to exhibit the behavior of hole and electron diffusion in rather impressive ways. In particular, the phenomenon of "internal contact potentials" can be strikingly demonstrated with such structures.

Transistors in which the nonlinear effects originate within the germanium as a result of the relationships of *p*-type and *n*-type regions are called "*p-n* junction transistors" to distinguish them from point-contact types, in which the metal semiconductor contact often plays an essential role. There are a number of possible *p-n* junction transistor structures: The *p-n-p* transistor has been discussed previously from a theoretical viewpoint.[1,2] In this article we shall consider chiefly the *n-p-n* transistor,[3] the *n*-type phototransistor with a *p-n* hook collector and a *p-n-p-n* transistor with *p*-type emitter and *p-n* hook collector.

In the following sections we shall describe first in simple terms the basic phenomena and effects with which we are concerned. We shall next describe the actual physical structure of several of the *n-p-n* transistors and their electrical characteristics. The theoretical principles will then be put in quantitative form and the current-voltage relationships derived for certain particular models. Finally a direct comparison between theory and experiment will be presented.

II. THE *n-p-n* STRUCTURE AS A TRANSISTOR AND AS A "HOOK MULTIPLIER"

In Fig. 1 we represent an *n-p-n* structure and indicate how it may be used as a transistor. Like the type-A transistor, the current paths between emitter terminal and base and between collector terminal and base have rectifying junctions. Unlike the type-A transistor, however, the rectification arises in the interior of the germanium and not at the contacts between metal leads and the germanium; which are substantially ohmic. There are other important differences between the *n-p-n* and the type-A: In the *n-p-n* the flow of injected carriers takes place chiefly by diffusion rather than by drift in an electric field; the current multiplication at the collector, which makes possible the positive feedback instability of the type-A transistor, is lacking in the *n-p-n* transistor.

In this section we shall give a brief resumé of the theory of the operation of the transistor. Additional details will be found in the cited references. In Secs. V, *et seq.*, some aspects of the theory will be treated analytically. The analysis is simplified by the use of the following assumptions:[4]

(1) The donors and acceptors are fully ionized (this is a good assumption for germanium at room temperature).

(2) The density of minority carriers is much smaller than the density of majority carriers in each region.

(3) The net rate of recombination in any region is linear in the deviation of the minority carrier density from its thermal equilibrium value. (Assumptions (2) and (3) permit us to use linear equations in dealing with the currents arising from carrier injection.)

(4) Space charge is negligible except at the space-charge regions in the *p-n* junctions themselves.

In Fig. 1 we show the energy band diagram for the structure under consideration for zero bias and for biases applied in such a way that the unit becomes an amplifying transistor. Under the latter condition the junction J_c on the right of the figure is biased in the reverse direction. This direction is such that electrons in the *n*-type collector region have low potential energy and cannot climb the potential energy hill to the base region; similarly holes are held in the base region. Electrons in the emitter region, however, may climb the small potential hill into the base region and once in this region may diffuse so that some of them will arrive at the right-hand junction. The flow over the hill depends on the height of the hill and this height may

[1] W. Shockley, Bell System Tech. J. 28, 435–489 (1949).
[2] W. Shockley, *Electrons and Holes in Semiconductors* (D. van Nostrand Company, Inc., New York, 1950).
[3] R. L. Wallace and W. J. Pietenpol, Bell System Tech. J. July, 1951. (This article, to which we shall refer a number of times, deals with a number of practical features which we do not consider here.) It is also scheduled for the Proc. Inst. Radio Engrs. July, 1951.

[4] These assumptions are discussed further in references 1 and 2.

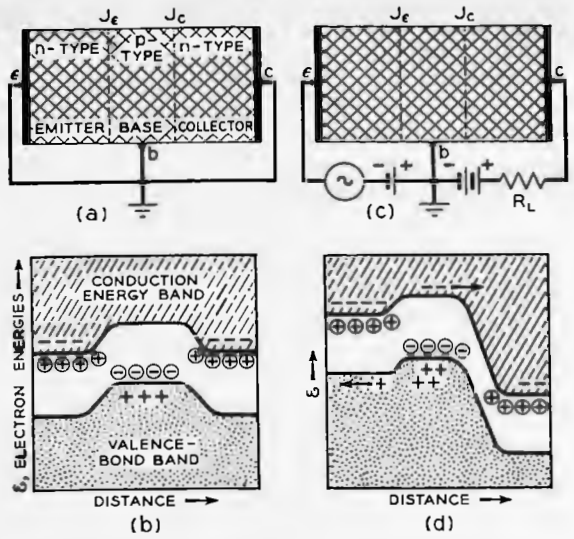


FIG. 1. The *n-p-n* structure and the energy level scheme. (a) and (b) Thermal equilibrium. (c) and (d) Biased as an amplifier.

be varied by applying a variable potential to the emitter while maintaining the base at constant potential. If the base region is very thin, few of the electrons will recombine with holes in it and, as a result, efficient transmission of electron current through the layer will occur. Furthermore, if the emitter region is more highly conducting than the base region, there will be many more electrons available to climb the hill than there are holes to climb the corresponding hill in the opposite direction. As a result, most of the current across the left junction will consist of electrons. Under these conditions the behavior of this device is closely analogous to that of a vacuum tube: the emitter region corresponds to the cathode, the base to the region around the grid wires, and the collector region corresponds to the plate. In favorably designed units the controlled electron current flowing through the base region may be very much larger than the current furnished the base region for control purposes so that the transistor has a current amplification factor that is very high. It may, therefore, be operated like a grounded cathode triode with the emitter region grounded and the signal applied to the base. In Sec. IV some current-voltage characteristics for transistors are shown and for them it may be seen that the current transmission is nearly perfect.

It is interesting to note that both for the vacuum tube and for the transistor, control is accomplished by the interaction of two forms of electron flow. In the vacuum tube, metallic flow in the grid wire controls the flow of thermionic electrons in the space between grid wires. In the transistor, hole flow in the base changes the base-emitter voltage and controls electron flow through the base layer.

The structure shown in Fig. 1 with the same operating biases may be used as a collector with high multiplication in a transistor.[5] It may also be used as a phototransistor.[6] We shall consider the application as a collector in a transistor later and shall describe here how multiplication of photocurrents can occur. For this purpose there need be no electrode connected to the base layer. If light shines on the germanium near the junctions, then the hole electron pairs generated will be separated by the field of the junctions and consequently a current of holes will flow into the base layer. These holes will accumulate in the layer and will charge it positively and thus reduce its potential energy for electrons. As a result more electrons will be able to climb the hill and flow to the collector. The effect of the added holes will die away after the light is removed due to the diffusion of holes into the left region where they combine with electrons and also due to recombination with electrons which diffuse into the base layer. If the layer is very thin, however, and the density of electrons in the left region is very high, then a very large number of electrons will be able to climb over the hill for each hole that is able to enter the emitter region and recombine. In Sec. VII we shall show that the current amplification obtained in this way is proportional to the ratio of the conductivities of the two layers and is inversely proportional to the thickness of the base layer.

An interesting consequence of the diffusion of electrons through the base layer is the occurrence of "internal contact potentials."[7] In order to illustrate these we shall suppose that the base layer is grounded and that a potential is applied to the emitter. If an additional ohmic contact is also made to the base region, it will of course show ground potential. If the contact is rectifying, however, and in particular if it carries most of its current in the form of electrons, its potential will be determined by the electron density in the base layer rather than by the potential established by the grounding contact. The *n*-region on the right represents such a contact. It is found both theoretically and experimentally that if the base layer is grounded, then potentials applied to the region on the left are transmitted through the base layer, although its electrostatic potential is practically unaltered, and exhibit themselves in the region to the right, which tends to "float" (when no current is drawn from it) at a potential approximately equal to that of the left region—at least over a certain range of voltages. Theory and experiment related to this phenomenon are given in Secs. VII and VIII.

III. DESCRIPTION OF EXPERIMENTAL UNITS

The experimental units were made of a piece of single crystal germanium in which a thin *p*-type layer is

[5] W. Shockley, Phys. Rev. **78**, 294 (1950).

[6] J. N. Shive who has developed the phototransistor [Phys. Rev. **76**, 575 (1949)] has proposed this use of the *p-n* hook.

[7] These were first discussed in reference 1. Internal contact potentials for point contacts have been measured by G. L. Pearson and analyzed by J. Bardeen [Bell System Tech. J. **29**, 469–495 (1950)].

interposed between two n-type sections. Units of a variety of conductivity values have been prepared. Typical values for an n-p-n structure as shown in Fig. 1 are: emitter, 100 (ohm-cm)$^{-1}$; base, 1 (ohm-cm)$^{-1}$; and collector, 0.1 (ohm-cm)$^{-1}$. Typical values of lifetime of the minority carrier in the collector section are 300–400 microseconds. The experimental evidence is that lifetimes in the other sections, in which direct measurement of lifetimes is more difficult, do not differ greatly from this. The leads to the three sections are mechanically strong and ohmic in character. The structure thus operates through conditions arising within the interior of the single crystal and not because of phenomena arising at the contacts between the leads and the germanium.

IV. PERFORMANCE CHARACTERISTICS

The performance characteristics of n-p-n transistors will be considered briefly.[8] Since these devices operate uniformly over the surfaces of p-n junctions they may be greatly altered as to the size of the active area, in contrast to point-contact devices. Thus it is possible to increase power output without corresponding increases in current density. One of the larger n-p-n transistors[9] studied as an amplifier had a junction area of 0.3 sq cm, a base layer thickness of about 0.07 cm, and delivered 2.0 watts of undistorted output in class A operation. Its frequency cutoff was about 10,000 cps, this limit being in general agreement with the effect of diffusion through the p-layer as discussed in Sec. IX.

The low power potentialities of these structures have been more thoroughly investigated. A junction area of about 0.01 sq cm and a base layer thickness of about 1.5×10^{-3} cm are typical dimensions for these units. They have operated with gains of 50 db and noise figures of about 10 db to 15 db at 1000 cycles per second. Each of these quantities is an improvement of several orders of magnitude over point-contact transistors. For low signal levels they give essentially full gain with collector voltages higher than 0.1 volt and are, therefore, exceptionally good very low power amplifiers. At some sacrifice in gain they may be operated at efficiencies of 48 to 49 percent out of a theoretical maximum of 50 percent for class A. An oscillator has been constructed by R. L. Wallace, Jr., and D. E. Thomas of this laboratory which operates on 0.6-microwatt input. The same small transistors will also operate as amplifiers with maximum output powers of several hundred milliwatts. A set of operating curves is given in Fig. 2. The frequency cutoff of these units in high gain circuits is determined chiefly by the capacitance of the collector junction and is much lower than the limit set by diffusion through the base layer. A discussion of the capacitative effect will be found in the article by Wallace and Pietenpol.

[8] An extensive presentation of circuit properties is given in reference 3.

[9] The performance of this transistor was discussed at the June, 1950, Inst. Radio Engrs. Conference on Electron Devices at the University of Michigan and also at the July, 1950, Conference on Semiconductors at Reading.

V. THEORETICAL PRINCIPLES AND BOUNDARY CONDITIONS

In this and following sections we shall give analytic form to the ideas discussed in Sec. II. The principal symbols used in dealing with the theory are shown in the accompanying table of notation. In this section and the next we shall deal with the n-p-n structure in general terms and for this reason shall use subscripts "l" and "r" standing for "left" and "right" for the two n-type regions. This permits us to consider impartially cases in which either region may be biased as a collector or as an emitter.

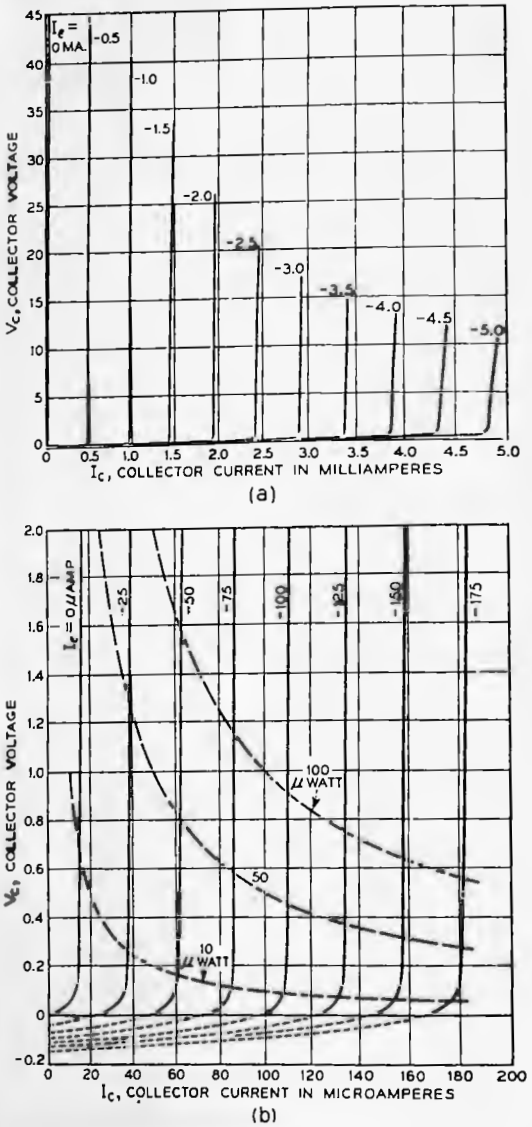


FIG. 2. Current-voltage relationships for an n-p-n transistor.

We shall first discuss the boundary conditions at the junctions when potentials are applied across them.[10] The electrostatic potential ψ in the interior of the semiconductor may have its zero chosen arbitrarily. For our purpose the zero is so chosen that $-q\psi$ is approximately the energy of an electron at a level of energy midway in the energy gap. The exact value of ψ is such that

$$n = n_i \exp q(\psi - \varphi)/kT \tag{5.1}$$

$$p = n_i \exp q(\varphi - \psi)/kT \tag{5.2}$$

under equilibrium conditions. Under nonequilibrium conditions, similar equations determine the "imrefs"[11] φ_n and φ_p in terms of ψ, n, and p as follows:

$$n = n_i \exp q(\psi - \varphi_n)/kT \tag{5.3}$$

$$p = n_i \exp q(\varphi_p - \psi)/kT. \tag{5.4}$$

NOTATION

A = cross-sectional area of unit;
$b = \mu_n/\mu_p = 2.1$;
$B_l = (kT/q)[1 - \exp(-q\varphi_l/kT)]$;
$B_r = (kT/q)[1 - \exp(-q\varphi_r/kT)]$;
e = base of naperian logarithms;
D_n, D_p = diffusion constants for electrons and holes;
G_{ll}, G_{lr}, etc. = conductances at zero voltage;
n = density of electrons;
$n_b = n$ in the base layer;
$n_i = n$ in an intrinsic specimen;
n_1 = deviation of n from its thermal equilibrium value;
N_d, N_a = density of donors, acceptors;
p = density of holes;
q = charge of a hole = $-$charge of an electron;
V, v = dc and ac components of voltage;
W = thickness of base layer;
μ_n, μ_p = mobilities of electrons, holes = 3600, 1700 cm^2/volt sec;
σ_i = intrinsic conductivity;
τ_{pl}, τ_{pr}, τ_{nb} = lifetimes of minority carriers;
φ = (Fermi level)/$(-q)$;
φ_n = imref for electrons;
φ_p = imref for holes;
φ_l, φ_b, φ_r = voltages of the three regions; and
ψ = electrostatic potential.

The imrefs or "quasi-Fermi levels" are introduced for convenience in discussing boundary conditions at the junctions and the meaning of applied voltages. In terms of the imrefs the current densities assume a particularly simple form:

$$I_n = qD_n \nabla n + q\mu_n nE = -q\mu_n n \nabla \varphi_n \tag{5.5}$$

$$I_p = -qD_p \nabla p + q\mu_p pE = -q\mu_p p \nabla \varphi_p. \tag{5.6}$$

These equations show that the current densities are those corresponding to materials with the conductivity appropriate for the electron and hole densities, respectively, and to electric fields derived from the imrefs as potentials. From these relationships it is evident that a given electron current will produce much bigger changes of the imrefs when it flows in a p-type region, where the electron density is small, than it will in an n-type region. In fact the ratio of conductivity by electrons is so great between the two regions that the imref for electrons can be regarded as substantially constant in the n-type region. In accordance with the assumption that the minority carrier density is small compared to the majority density and the assumption that the space charge is negligible, it follows that the potential ψ is substantially uniform in the interior of each region also. If the contacts on regions l and r are so far from the junctions that no injected carriers reach them and are substantially ohmic, then it also follows that the imrefs for electrons in these regions are simply the voltage applied to the two contacts. In accordance with the assumption that current to the base contact is carried by holes, it also follows that the imref for holes in this region is equal to φ_b.

We shall now apply the aforementioned conclusions to the boundary condition at J_l. For simplicity we shall assume that the base is grounded so that we shall consider in general cases in which

$$\varphi_l \neq 0, \quad \varphi_b = 0, \quad \varphi_r \neq 0. \tag{5.7}$$

By the reasoning of the preceding paragraph, the imref for electrons is continuous across the junction J_l and in fact has its largest gradient only after the interior of the base region is reached. Consequently we may take φ_n in the base region near J_l as substantially equal to φ_l. The electron density in the base region near J_l is thus given by

$$n(\text{in } b \text{ near } J_l) = n_b \exp(-q\varphi_l/kT), \tag{5.8}$$

where n_b is the thermal equilibrium concentration of electrons at the corresponding point. (This equation follows directly from (5.3) together with the conclusion previously reached that ψ and φ_p in the base layer are unaffected by the applied potentials of (5.7).)

We shall be chiefly concerned with deviations of the densities from their equilibrium values and shall use the subscript 1 to indicate such densities. The deviation corresponding to (5.8) is

$$n_1 = n_b[\exp(-q\varphi_l/kT) - 1] = -n_b q B_l/kT. \tag{5.9}$$

In this expression we have introduced the quantity B_l defined as

$$B_l \equiv (kT/q)[1 - \exp(-q\varphi_l/kT)]. \tag{5.10}$$

This symbol is introduced since all the currents with which we shall be concerned depend functionally upon the voltages in the form (5.10). The coefficient kT/q is introduced so that B_l has the dimensions of a voltage and for small values of φ_l, B_l is in fact approximately equal to φ_l.

[10] The notation used here is similar to that of references 1 and 2 and the analysis is substantially an abreviation of that of reference 1.

[11] We are indebted to the most appropriate authority for suggesting this modified name for the quasi-Fermi levels.

Entirely similar reasoning leads to corresponding relationships for the hole density in the l region:

$$p(\text{in } l \text{ near } J_l) = p_l \exp(-q\varphi_l/kT) \tag{5.11}$$

$$p_l = p_l[\exp(-q\varphi_l/kT) - 1] \cong -p_l q B_l/kT. \tag{5.12}$$

VI. THE CURRENT-VOLTAGE RELATIONSHIPS

The analysis of the last section indicates that near J_l the deviations of both hole and electron densities are proportional to B_l. These deviations lead to diffusion currents, which would vanish for the case of thermal equilibrium with $B_l = 0$. As a result of the linear approximations discussed in Sec. II, these currents will be proportional to B_l. In Fig. 3 the conventions selected for signs of current are shown. We shall accordingly denote the current into region b due to hole flow across J_l as follows:

$$I_{lp} = -G_{llp} B_l. \tag{6.1}$$

For the case of a uniform cross section of area A and material of uniform conductivity and uniform lifetime τ_{pl}, the value of the coefficient may be easily derived:[12]

$$G_{llp} = q\mu_p p_{nl} A/L_{pl} = \sigma_i^2 bA/(1+b)^2 \sigma_l L_{pl} \tag{6.2}$$

in which the diffusion length is given by the equation,

$$L_{pl} = (D_p \tau_{pl})^{\frac{1}{2}}. \tag{6.3}$$

The last form of (6.2) expresses the conductance G_{llp} in terms of the conductivity σ_i of an intrinsic sample and the actual conductivity of the n-type region. The quantity b occurring in the equation is the ratio of mobilities:

$$b = \mu_n/\mu_p. \tag{6.4}$$

The electron current flowing across J_l can also be directly evaluated for the case of $B_r = 0$. Even if no potential is applied across J_r, some of the electrons injected across J_l will arrive in the r-region. This is a consequence of the fact that the deviation n_l is required to be zero at J_r when φ_r is zero. The electron currents across the two junctions are found to be[13]

$$I_{ln} = [(q\mu_n n_b A/L_n) \coth(W/L_n)] B_l \equiv G_{lln} B_l \tag{6.5}$$

$$I_{rn} = -[(q\mu_n n_b A/L_n) \operatorname{csch}(W/L_n)] B_l \equiv G_{rln} B_l. \tag{6.6}$$

The conductance G_{lln} may be expressed in terms of the properties of the base layer as follows:

$$G_{lln} = [\sigma_i^2 bA/(1+b)^2 \sigma_b L_n] \coth(W/L_n). \tag{6.7}$$

A similar treatment for J_r leads to a corresponding set of equations. In terms of the G's and the B's the current-voltage relationship may be written as follows:

$$I_l = G_{ll} B_l + G_{lr} B_r \tag{6.8}$$

$$I_r = G_{rl} B_l + G_{rr} B_r \tag{6.9}$$

$$G_{ll} = G_{lln} + G_{llp}, \quad G_{lr} = G_{lrn} \tag{6.10}$$

$$G_{rl} = G_{rln}, \quad G_{rr} = G_{rrn} + G_{rrp}. \tag{6.11}$$

For low voltages we have the approximate relationship,

$$B_l \doteq \varphi_l, \quad B_r \doteq \varphi_r \quad \text{for} \quad |\varphi_{l,r}| \ll kT/q, \tag{6.12}$$

so that the coefficients in Eqs. (6.8) and (6.9) are simply the low voltage conductance components. If the potentials consist of a bias plus a small ac component so that we may write

$$\varphi_l = V_l + v_l, \quad \varphi_r = V_r + v_r, \tag{6.13}$$

then the small signal equations become

$$i_l = g_{ll} v_l + g_{lr} v_r \tag{6.14}$$

$$i_r = g_{rl} v_l + g_{rr} v_r, \tag{6.15}$$

where the relationship between the small g's and the large G's is

$$g_{ll}/G_{ll} = g_{rl}/G_{rl} = \exp(-qV_l/kT) \tag{6.16}$$

while a similar equation applies for the other two coefficients.

From these relationships one can also derive the result that each g is proportional to the deviation of its corresponding GB term from its saturation value corresponding to $B = (kT/q)$. This may be expressed symbolically as follows:

$$g = (\text{deviation of } GB \text{ from saturation}) \cdot (q/kT). \tag{6.17}$$

For the model discussed in connection with Eqs. (6.5) and (6.6), it is evident that symmetry leads to the equation,

$$G_{lr} = G_{rl}. \tag{6.18}$$

If the conductivity varies in an unsymmetrical way, however, in the middle layer or if the lifetime is greater at one side of the layer than the other, then we cannot reach the conclusion that the two G's are equal from symmetry arguments. It can be shown, however, that it is a consequence of the linear assumptions described in Sec. II that the symmetry relationship holds no

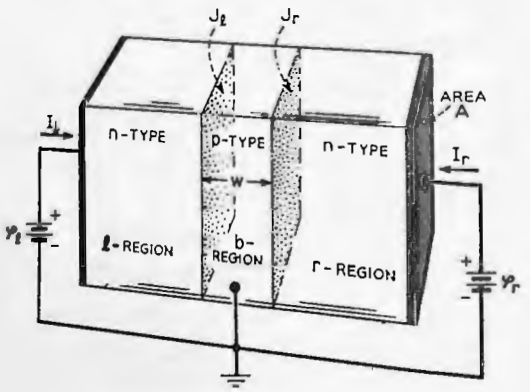


FIG. 3. Dimensions and conventions for voltage and current for an n-p-n structure.

[12] Reference 1, Eqs. (4.20) and (4.21) or reference 2, p. 316. In a specimen of finite cross section, the recombination must be described in terms of a set of normal modes. For the cross sections of the small transistors, the lowest mode dominates and its lifetime may be used in the formulas derived for the one-dimensional case. See reference 1, Appendix V.

[13] Reference 1, Eq. (5.6), modified for electron diffusion.

matter what the geometry of the middle layer. The proof of this result is given in an appendix.[14] This leads us to the conclusion that for the idealized sort of model in which all of the currents are linear in the injected carrier density, the behavior of the transistor is described by three parameters, at least so far as low frequencies are concerned. These parameters are the four coefficients in Eqs. (6.8) and (6.9) with the relationship (6.18) between two of them.

In terms of these coefficients we may define the "alpha" of the unit using either junction as an emitter. The two α's need not be equal and the α for the left junction is given by the relationship,

$$\alpha_l \equiv -(\partial I_r/\partial I_l) \quad [\text{for} \quad \varphi_r \text{ const}] = -G_{rl}/G_{ll}$$
$$= -G_{rln}/(G_{lln}+G_{llp})$$
$$= (-G_{rln}/G_{lln})[G_{lln}/(G_{lln}+G_{llp})] = \gamma_l\beta_l. \quad (6.19)$$

The definition of γ in transistor terminology is the fraction of the current at the emitter junction produced by emitter voltage that is carried by minority carriers in the base; evidently

$$\gamma_l = G_{lln}/(G_{lln}+G_{llp}). \quad (6.20)$$

The fraction of these injected carriers that reaches the collector is defined as β:

$$\beta_l = -G_{rln}/G_{lln} = \operatorname{sech}(W/L_n). \quad (6.21)$$

The "intrinsic α" or α^* of the collector junction is defined as the ratio of change in total current per unit minority carrier current arriving at it. For a simple p-n junction collector, α^* is unity. For a "hook collector," which we treat in the next section, the arrival of injected current provokes a flow of carriers and α^* may be 100 or more.

If the middle layer is thin so that the hyberbolic functions of (6.5) and (6.6) may be approximated by their first terms, we may write

$$\beta_l \doteq 1-\tfrac{1}{2}(W/L_n)^2 \doteq 1 \quad (6.22)$$

$$\gamma_l \doteq 1/[1+(\sigma_b W/\sigma_l L_{pl})] \quad (6.23)$$

$$\alpha_l \doteq 1/[1+(\sigma_b W/\sigma_l L_{pl})] \quad (6.24)$$

$$G_{lln} \doteq \sigma_1{}^2 bA/(1+b)^2\sigma_b W. \quad (6.25)$$

Entirely similar expressions may be written for J_r simply by interchanging r and l.

[14] This symmetry result may be derived in a more general way from the reciprocity principle of electrical conduction provided no magnetic field is present. The necessary theorem is proved in section 5 of H. B. G. Casimir [Revs. Modern Phys. 17, 343 1945)]. The proof there shows that in the linear range of conductivity, G_{lr} must equal G_{rl}. Since we have shown that the currents are linear functions of the B's by an independent argument, it follows that we may take $G_{lr}=G_{rl}$ in general. The method used by Casimir is based on Onsager's principle of microscopic reversibility and has an unnecessarily abstract flavor so far as the needs of this article are concerned. The desired theorem can be proved by straightforward analytical methods as is shown in the Appendix.

VII. SPECIAL OPERATING CONDITIONS

We shall next consider the consequences of the equations derived in the previous section for several limiting cases of voltages and currents.

A. Operating as an Amplifier

In order to get into the range of linear behavior it is necessary to apply a sufficiently large reverse bias to the collector so that its current saturates. This condition corresponds to the straight parts of the characteristics shown in Fig. 2. The voltage required to get into this range must exceed about $4kT/q$. After this point is reached the emitter current and collector current may be approximated by

$$I_l = G_{ll}B_l - \alpha_l G_{ll} kT/q \quad (7.1a)$$

$$I_r = -\alpha_l G_{ll} B_l + G_{rr} kT/q$$
$$= -\alpha_l I_l + (G_{rr}+\alpha_l{}^2 G_{ll})(kT/q). \quad (7.1b)$$

Equation (7.1b) accounts for the parallel lines with currents increasing linearly in I_l shown in Fig. 2 for high collector voltages.

For applications in circuit theory, it is important to know the emitter admittance g_{ll}. To the approximation emphasized in this paper in which series ohmic resistances are neglected, this admittance is $1/r_e$ of the equivalent circuit.[3] The range of interest is usually such that $-\varphi_l > 4kT/q$ so that the application of the reasoning of (6.17) leads to

$$1/r_e = g_{ll} = \partial I_l/\partial\varphi_l = G_{ll}\exp(-q\varphi_l/kT)$$
$$\doteq I_l q/kT \doteq 40 I_l \text{ mho}. \quad (7.2)$$

If the unit is operated with grounded emitter and with the collector current saturated, then the input admittance is

$$dI_b/d\varphi_b = d(-I_l-I_r)/d(-\varphi_b)$$
$$= (1-\alpha_l)G_{ll}\exp(-q\varphi_l/kT) \quad (7.3)$$

while the transconductance is $1/(1-\alpha_l)$ larger. Thus if $\alpha_l = 0.99$, there will be a current gain of 100-fold, if the collector current is saturated. The theoretical power gain would be infinite if the collector impedance were infinite corresponding to the ideal saturation of (7.1b). Actually collector resistances of 10^7 to 10^8 ohms are observed. For currents large compared to saturation currents, (7.3) leads to an admittance of $(1-\alpha_l)I_r q/kT = 10^{-2}\times 10^{-4}\times 40 = 4\times 10^{-5}$ mhos for $I_r = 10^{-4}$ amp, a typical value for high gain performance. The power gain will be roughly the ratio of output to input impedances times the square of the current gain and will thus be about $10^7\times 4\times 10^{-5}\times 10^4 \doteq 66$ db for a matched load. As discussed in Sec. IV, gains as high as 50 db have been obtained in practical circuits.

B. Hook Collector in p-n-p-n Transistor

Point contact transistors are frequently observed to have current multiplication in the sense that at a

fixed collector bias, changes in emitter current produce changes in collector current several fold larger. One explanation is that a "p-n hook" is formed at the collector junction so that the "intrinsic α" or α^* of the collector is high. We shall illustrate this theory for the case of an n-type transistor in which holes are injected by a p-n junction emitter and the collector consists of n and p layers.

This structure is represented in Fig. 4 and is shown for bias conditions similar to those discussed for Fig. 1 except for the added emitter region and the absence of a contact to the b layer. Regarded as a transistor, the emitter, base, and collector are e', b', and c'. The operating biases put reverse voltage across J_r and forward across J_l. The base layer b will float at a potential such that the net current to it is zero:

$$I_r+I_l=G_{ll}(1-\alpha_l)B_l+G_{rr}(1-\alpha_r)B_r=0. \quad (7.4)$$

For large reverse biases B_r will be positive and B_l negative and

$$\varphi_l-\varphi_r=V_{c'}, \quad (7.5)$$

where $V_{c'}$ is the voltage on the collector. For large $V_{c'}$, B_r will saturate and φ_l will be determined by

$$B_l=-G_{rr}(1-\alpha_r)(kT/q)/G_{ll}(1-\alpha_l) \quad (7.6)$$

so that the saturation current is

$$I_r(\text{sat.})=(1-\alpha_r\alpha_l)G_{rr}kT/q(1-\alpha_l). \quad (7.7)$$

If holes are injected by e' so that a hole current I_p^* arrives at J_r, then the condition,

$$I_r+I_l+I_p^*=0 \quad (7.8)$$

leads to a change

$$\Delta B_l=-I_p^*/G_{ll}(1-\alpha_l) \quad (7.9)$$

and this produces an increased electron current across J_r which leads to

$$\Delta I_{nr}=G_{rl}\Delta B_l=\alpha_l I_p^*/(1-\alpha_l). \quad (7.10)$$

The "intrinsic α" or α^* of the composite structure is

$$\alpha^*=(\Delta I_{nr}+I_p^*)/I_p^*=1/(1-\alpha_l). \quad (7.11)$$

For the thin layer approximation we see that

$$\alpha^*\doteq 1+(\sigma_l L_{pl}/\sigma_b W). \quad (7.12)$$

Thus for thin layers and for highly conducting l regions, α^* may be made very large.

The reason for calling the structure a "p-n hook" is illustrated in Fig. 4b. The high energy potential for electrons in layer b is low potential for holes. Holes injected by e' become caught in this hook and bias J_l forward so as to provoke the enhanced electron flow.

C. Phototransistor

The structure in Fig. 4 can act as a phototransistor if the hole injection by the emitter junction is simply replaced by hole electron pair generation by light. For this application, the structure has only two terminals, b' and c' of Fig. 4, the region e' being absent. If the hole electron pairs are generated in the neighborhood of J_r, then they will be separated by the field in the junction with a current equivalent to the passage of one hole across J_r for each pair so separated. The hole current across the junction will under these conditions be multiplied just as for the case of the p-n-p-n transistor so that the apparent quantum efficiency for hole electron pairs generated at J_r is

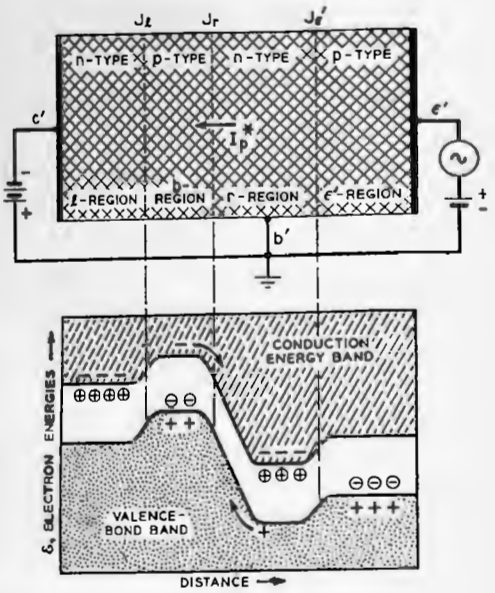


FIG. 4. A p-n-p-n transistor with a p-n hook collector.

apparent quantum efficiency

$$=\alpha^*\doteq 1+(\sigma_l L_{pl}/\sigma_b W) \quad (7.13)$$

by the reasoning that leads to Eq. (7.12). In the following section we shall discuss some values of α^* determined by measurements of photocurrents.

D. Internal Contact Potentials

In order to discuss internal contact potentials we return to a consideration of the three region device of Fig. 3. If b is grounded and region r is allowed to seek its own potential, then potentials applied to region l will produce potentials on region r although the potential measured with an ohmic contact to b would everywhere be zero. The floating potential of region r is determined by setting the total current equal to zero so that we have

$$0=I_r=G_{rl}B_l+G_{rr}B_r \quad (7.14)$$

$$B_r=(-G_{rl}/G_{rr})B_l=(-G_{lr}/G_{rr})B_l=\alpha_r B_l \quad (7.15)$$

$$\exp(-q\varphi_r/kT)=1-\alpha_r+\alpha_r\exp(-q\varphi_l/kT). \quad (7.16)$$

Equation (7.16) expresses φ_r as a function of φ_l and α_r. It takes simple limiting form for extreme bias conditions:

Forward bias:

$$-q\varphi_l/kT \gg 1$$

$$\varphi_r = \varphi_l - (kT/q)\ln\alpha_r \tag{7.17a}$$

$$\alpha_r = \exp q(\varphi_l - \varphi_r)/kT. \tag{7.17b}$$

Zero bias:

$$|q\varphi_l/kT| \ll 1$$

$$\varphi_r = \alpha_r\varphi_l. \tag{7.18}$$

Reverse bias:

$$q\varphi_l/kT \gg 1$$

$$\varphi_r = -(kT/q)\ln(1-\alpha_r) \tag{7.19a}$$

$$\alpha_r = 1-\exp(-q\varphi_r/kT). \tag{7.19b}$$

In the following section, we shall show that these expressions are approximately satisfied.

VIII. COMPARISON WITH EXPERIMENT

In this section we shall be chiefly concerned with an analysis of data on an *n-p-n* transistor and with showing that it may be interpreted on the basis of the theory discussed above. The data were taken under two sets of conditions: In the first the voltages were small compared to (kT/q) and from these the G's were determined using the approximation (6.12). In the second set, a wide range of voltages were used; for all conditions, however, one of the two B's was taken as independent and the other B was either constant or else proportional to the independent B. Consequently, each current is of the form

$$I = c + mB = I_S + m[B - (kT/q)]$$
$$= I_S - I_V \exp(-qV/kT), \tag{8.1}$$

where V is the voltage upon which B depends and

$$I_S = c + (mkT/q) \tag{8.2}$$

is the "saturation" value of the current for large positive values of V and

$$I_V = mkT/q. \tag{8.3}$$

Both I_S and I_V are readily calculated in terms of the G's. In the analysis of the data the measured I_S and I_V values are compared with values computed from the G's, and the dependence of I upon V is investigated.

It should be pointed out that the values of the G's are strongly dependent on temperature. Consider, for example,

$$G_{rl} = b\sigma_i^2 A/(1+b)^2\sigma_p W. \tag{8.4}$$

TABLE 8.1. Zero bias conductances for an *n-p-n* transistor (conductances in micromhos at $T = 22°C$).

	Measured	Calculated
G_{11}	8.8 ±0.5	8.8
G_{22}	33.3 ±0.5	33.3
$G_{11}(1-\alpha_1\alpha_2)$	6.9 ±0.5	6.9
$G_{22}(1-\alpha_1\alpha_2)$	26.5 ±0.5	26.4
$\alpha_2 = (V_c/V_e)$ for $I_c = 0$	0.86±0.02	0.89
$\alpha_1 = (V_e/V_c)$ for $I_e = 0$	0.22±0.02	0.23
$G_{11}+G_{12}-2G_{21}$	26.2 ±0.5	26.5

In this expression[15]

$$\sigma_i^2 \propto \exp(-\epsilon_G/kT) \tag{8.5}$$

where ϵ_G is the energy gap and

$$\sigma_p \propto T^{-1} \tag{8.6}$$

Consequently the value of G_{rl} at $T_0 + \Delta T = 300°K + \Delta T$ is approximately

$$G_{rl}(T_0+\Delta T) \doteq G_{rl}(T_0)\exp[(\epsilon_G/kT_0^2) + (3/2T_0)]\Delta T$$
$$= G_{rl}(T_0)\exp(0.095+0.005)\Delta T \tag{8.7}$$

so that G_{rl} increases approximately 10 percent per degree C. This increase arises chiefly from σ_i^2 and will be approximately the same for all the G's.

The fact that the G's have large temperature coefficients implies that at fixed voltages the currents will be very sensitive to temperature. This does not mean, however, that in a properly designed circuit, the behavior of the unit will be highly sensitive to temperature. The value of α_1, for example, as shown in (6.24) involves only $\sigma_b W/\sigma_e L_{pe}$ and this has only a small temperature coefficient. At a fixed emitter current, the emitter resistance is proportional to T in °K. Thus the most important quantities from a circuit point of view have small temperature coefficients.

Since in this section we are dealing with a transistor designed to have one terminal as emitter and one as collector, we shall abandon the "left" and "right" terminology and use subscript "1" for the emitter and "2" for the collector as is customary for transistors. The large signal equations are then

$$I_e = G_{11}B_1 + G_{12}B_2 \tag{8.8}$$

$$I_c = G_{21}B_1 + G_{22}B_2 \tag{8.9}$$

and the small signal equations are

$$i_e = g_{11}v_e + g_{12}v_c \tag{8.10}$$

$$i_c = g_{21}v_e + g_{22}v_c \tag{8.11}$$

where each g gepends on its corresponding voltage in the form,

$$g_{ij} = G_{ij}\exp(-qV_j/kT). \tag{8.12}$$

The low voltage conductances were measured at 2 millivolts. For this small voltage

$$B(+2\text{ mv}) = \quad 1.95\text{ mv} \tag{8.13a}$$

$$B(-2\text{ mv}) = -2.05\text{ mv} \tag{8.13b}$$

so that the currents are nearly linear in this range, and the nonlinearity is practically eliminated by averaging the two polarities. With the collector grounded, I_e was measured and G_{11} computed from I_e/V_e. With the collector open circuited, I_e and V_c were measured; for this case $I_e/V_e = G_{11}(1-\alpha_1\alpha_2)$ and $V_c/V_e = -G_{21}/G_{22} = \alpha_2$. Similar data were taken with voltage applied to the

[15] n_i^2 varies as $T^3\exp(-\epsilon_G/kT)$, see reference 2, page 475, and $\mu_n\mu_p$ as T^{-3}, see page 287.

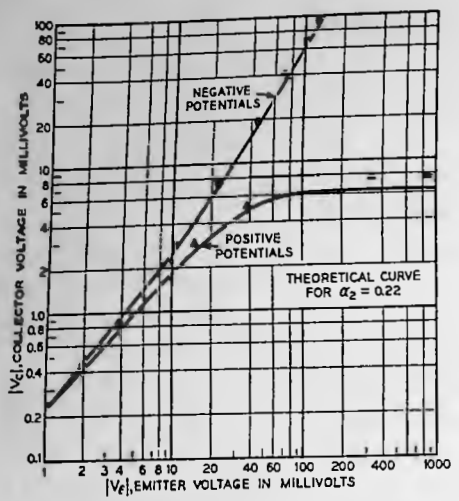


FIG. 5. Internal contact potential developed on the open circuited collector.

collector. Finally collector and emitter were connected together and the combined currents measured; this gives

$$(I_e+I_c)/V_e=G_{11}+G_{22}-G_{12}-G_{21}. \quad (8.14)$$

The values selected for the G's were: Conductances in micromhos at 22°C

$$G_{11}=8.8,\ G_{22}=33.3,\ G_{12}=G_{21}=-7.8. \quad (8.15)$$

These three values fit the seven measurements within the limits of experimental accuracy as shown in Table 8.1. The fact that the fit can be achieved with three constants is not a real test of the theory of Sec. VI, however, since any passive three terminal device in the absence of magnetic fields should satisfy the reciprocity condition and be described by three constants. What the table shows essentially is the consistency of the measurements.

Accurate values for the constants for the base layer were not available for the unit studied. However, the orders of magnitude of the G's are in reasonable agreement with values expected for a structure with constants lying in the ranges expected from the method of fabrication. We shall not attempt to obtain a perfect fit[16] but will choose as an example a structure with $A=0.003$ cm², $W=2.5\times10^{-3}$ cm, $\sigma_e=20$, $\sigma_b=10$, $\sigma_c=0.1$, $\sigma_i=(1/60)$ ohm^{-1} cm^{-1}, and lifetimes $\tau_{pc}=40$ and $\tau_{pe}=10$ microseconds. The resulting values for the G's based on Eqs. (6.2), (6.22), and (6.25) are

$$G_{11n}=7.3 \text{ micromhos} \quad (8.16a)$$

$$G_{11p}=0.5 \text{ micromhos} \quad (8.16b)$$

$$G_{22p}=45 \text{ micromhos}. \quad (8.16c)$$

[16] For the simpler case of a p-n junction, the electrical properties have been predicted with an accuracy of about 20 percent from the independently measured constants describing the junction.

The value of β is greater than 0.99 for $\tau_n=40$ microseconds and may be taken as unity so far as the G's are concerned. These values lead to

$$G_{11}=G_{11n}+G_{11p}=7.8 \quad (8.17a)$$

$$-G_{12}=-G_{12n}=\beta_1 G_{11n}=\beta_2 G_{22n}=7.3 \quad (8.17b)$$

$$G_{22}=G_{22n}+G_{22p}=7.3+45=52. \quad (8.17c)$$

According to this interpretation the failure of α_1 to be unity is due chiefly to hole flow across J_e and a similar condition is true of α_2. The base current arises almost entirely from these hole flows with recombination in the base being nearly negligible.

Figures 5 and 6 show the internal contact potential effect. In Fig. 5 the potential V_e is applied and V_c measured while zero current flows to the collector terminal. The data are seen to be in general agreement with the theory and with the value of α_2, corresponding to the collector junction, obtained at low voltage for Table 8.1. Figure 6 shows similar data with the voltage applied to the collector. The data are seen to differ slightly from the theoretical curves for reverse biases. The values of α obtained by applying (7.19b) to these data are

$$\alpha_2=0.226 \quad \text{and} \quad \alpha_2=0.894, \quad (8.18)$$

which shows that the fit is very sensitive to small variations in value of α. The test of Eq. (7.17b), which applies to forward bias, cannot be carried out as satisfactorily because under these conditions the currents are relatively large and the voltage drops across the series resistances of the specimen are not negligible compared to the effects studied. The values obtained are, however, approximately the same as those of Table 8.1.

In Fig. 7, the collector is biased to saturation and I_c and I_e are plotted as functions of V_e. For this case the

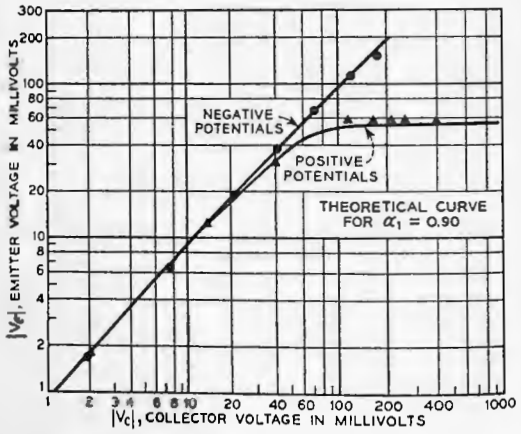


FIG. 6. Internal contact potential developed on the open circuited emitter.

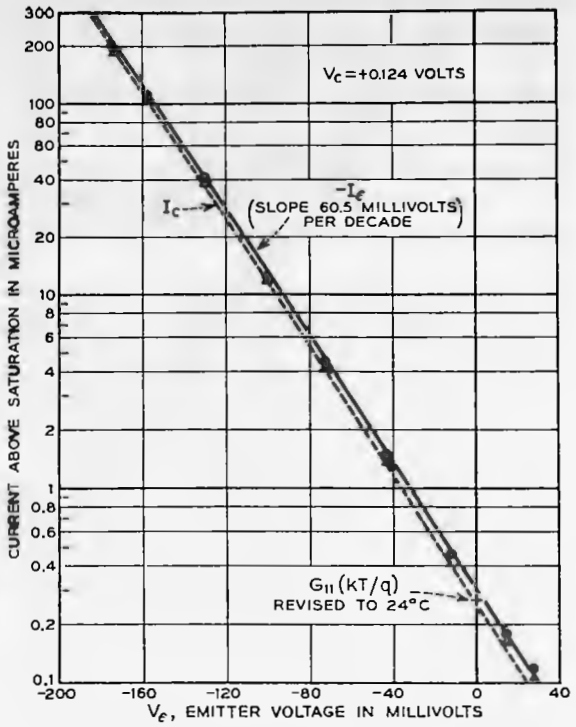


FIG. 7. Emitter and collector currents for the collector biased to saturation.

formulas should be

$$I_e = G_{11}B_1 + G_{12}(kT/q)$$
$$= G_{11}(1-\alpha_1)kT/q - (G_{11}kT/q)\exp(-qV_e/kT) \quad (8.19a)$$
$$I_c = G_{12}B_1 + G_{22}(kT/q)$$
$$= G_{22}(1-\alpha_2)(kT/q)$$
$$+ (\alpha_1 G_{11}kT/q)\exp(-qV_e/kT). \quad (8.19b)$$

It is seen that the lines agree well with the exponential forms and, furthermore, that the slope is in good agreement with theory which requires that for one decade of change in the current the voltage change should be

$$\Delta V = 2.30 \times kT/q = 2.30/39.4 = 59.0 \text{ mv} \quad (8.20)$$

for $T = 297°K$, the temperature at which the data were taken. The values of I_S and I_V deduced from Table 8.1 (corrected for a ΔT of 2°C) and from the data on which Fig. 7 was based are:

	Fig. 7	Table 8.1
$G_{11}(1-\alpha_1)kT/q$	0.021 μa	0.025 μa
$G_{11}(kT/q)$	0.30 μa	0.27 μa
$G_{22}(1-\alpha_2)kT/q$	0.88 μa	0.78 μa
$\alpha_1 G_{11}kT/q$	0.27 μa	0.24 μa.

The slope terms are simply the values for $V_e = 0$ in Fig. 7. The saturation values were deduced directly from the data. In the voltage range used, the saturation for the collector was not perfect and the collector saturation values were corrected for a "leakage" term of about 1 megohm. (The origin of this leakage effect is not clear and it tends to saturate at higher reverse biases.)

In Fig. 8 the dependence of emitter current upon emitter voltage is again shown. The unit was at somewhat higher temperature and was measured at a higher collector voltage in one case and with the collector floating in the other. The ratio of the two terms should be

$$G_{11}/G_{11}(1-\alpha_1\alpha_2) = 1.25. \quad (8.21)$$

The observed ratio is 1.30 at $V_e = 0$ which is satisfactory agreement.

In Fig. 9 the dependence of collector current upon collector voltage is shown for two cases similar to those of Fig. 8. The ratio of the two values is again 1.30 in good agreement with the prediction. It should be noted, however, that the slope requires 74 millivolts per decade, a value appropriate to an unreasonably high temperature of 103°C. This slope is established for such low currents that it seems difficult to explain it by spurious effects of series resistances.

There is an important difference in the nature of the currents of Fig. 7, which fit the theoretical slope, and those of Fig. 9, which do not. The currents of Fig. 7 consists chiefly of electrons which diffuse through the base layer and arrive at the collector; the evidence for

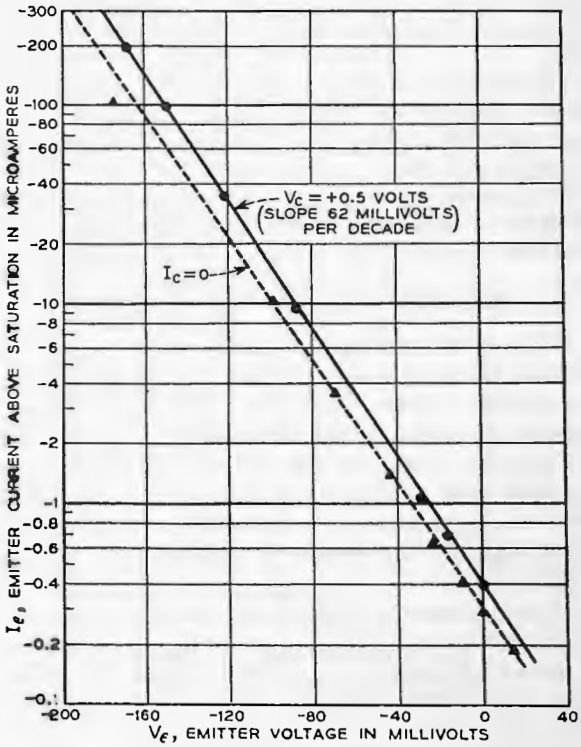


FIG. 8. Emitter current for two conditions of collectors.

this conclusion is the value 0.89 for α_1, which implies that only 11 percent of the current is carried by electrons recombining in the base layer or holes diffusing into and recombining in the emitter body (or the reverse processes, depending on the polarity). On the other hand, most of the collector current is probably carried by holes which recombine in the collector body. (This reasoning is in agreement with the attempt to interpret the G's in terms of the structure discussed in connection with (8.17a, b, c).) It is to be expected theoretically that if the recombination process involves trapping on recombination centers,[17] then the rate of recombination will increase less rapidly than linearly with injected carrier density because at high densities the traps tend to saturate. This view has received some experimental support from the work of F. S. Goucher and J. R. Haynes who find an apparent increase in lifetime with increasing carrier density. It may be that this mechanism accounts for failure of the currents to increase as rapidly as they should with increasing voltage in Fig. 9.

Further evidence for nonlinearity in the recombination of holes in the emitter is furnished by the dependence of α_1 upon emitter current. This can be seen in Fig. 7. As the emitter current increases, the ratio of collector to emitter current (above saturation) increases from 0.9 to about 0.95. This is interpreted as being due to the failure of hole current to increase as rapidly with voltage as does emitter current.

The tendency of α to increase with emitter current appears to be a general feature of n-p-n transistors. Wallace and Pietenpol[3] report values as high as 0.995 for α.

Phototransistors made of n-p-n structures are extremely responsive to light and exhibit apparent quantum efficiencies of at least several hundreds. In accordance with the interpretation of Sec. VII C, these efficiencies would lead to α_1 values deduced from $1-(1/\alpha^*)$ comparable to or larger than the highest observed in n-p-n transistors.

IX. SOME DESIGN CONSIDERATIONS

It has been the principal purpose of the preceding sections to examine the consequences of the diffusion theory and compare them with experiment. For this purpose the experimental conditions considered were the simplest: small currents and zero frequency. For practical applications high frequency and larger currents are also of interest. In this section we shall discuss briefly some of the factors of importance in design considerations.

Of great interest is the frequency cutoff. This may be determined by the external circuit or by one or another of several internal features of the transistor. The most fundamental of these latter is that set by the diffusion

[17] See reference 2, page 342.

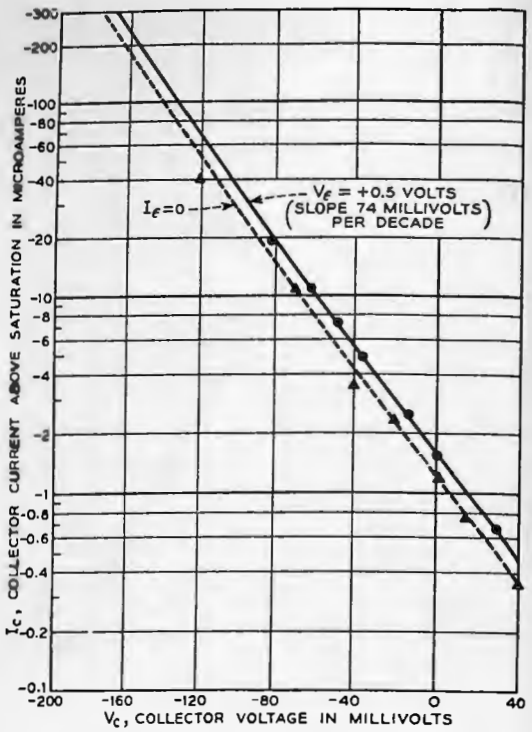


FIG. 9. Collector current for two conditions of emitter.

time through the base layer. This time is

$$\tau_D = W^2/D_n. \tag{9.1}$$

At a circular frequency ω, solutions for n_1 in the base layer are of the form,

$$\exp(i\omega t \pm x(1+i\omega\tau_n)^{\frac{1}{2}}/L_n) \tag{9.2}$$

and this leads to a β value of

$$\beta = \operatorname{sech}(1+i\omega\tau_n)^{\frac{1}{2}} W/L_n. \tag{9.3}$$

For frequencies such that $\omega\tau_n \gg 1$, this reduces to

$$B = 2/[\exp(1+i)(\omega\tau_D/2)^{\frac{1}{2}} + \exp(-1-i)(\omega\tau_D/2)^{\frac{1}{2}}]. \tag{9.4}$$

From this it is evident that for $\omega\tau_D \gg 2$ there is a phase lag of $(\omega\tau_D/2)^{\frac{1}{2}}$ radians and an equal attenuation in nepers. The power gain, which is proportional to β^2 in many cases, drops about 3 db when $\omega\tau_D = 2$ or $f = D_n/\pi W^2 \doteq 30/W^2$. For $W = 10^{-3}$ inch or 2.5×10^{-3} cm this is about 5×10^6 cps.

In addition to this fundamental limitation, there may be limitations due to capacitance and ohmic resistances. We shall illustrate this by considering the grounded emitter form of circuit, which is analogous to a grounded cathode vacuum tube circuit. For this case ac signals are applied to the base and ac voltages are developed on the collector. These voltages charge the capacitance of the base collector junction and this charging current must be furnished by hole flow in the base layer. For a layer

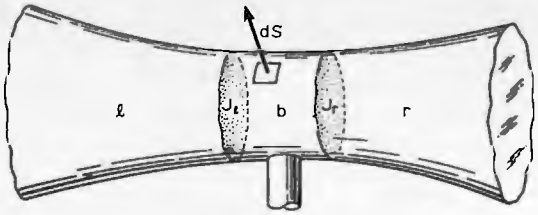


FIG. 10. The base region over whose surface the integration is made.

with $\sigma_b = 10$ ohm^{-1} cm^{-1} and $W = 2.5\times10^{-3}$ cm, the resistance from edge to opposite edge of a unit square is 40 ohms. If the capacitance is 2000 $\mu\mu$f/cm^2 and the voltage gain is 10, then the input signal will dominate the voltage on the base layer only if the width of the unit is less than

$$(10\times\omega\times2\times10^{-9}\times40)^{-\frac{1}{2}} = (\omega8\times10^{-7})^{-\frac{1}{2}}.$$

For a frequency of 5×10^6 cps, this leads to a width of 2 mm. The current through the base layer must also charge the emitter capacitance which may be much larger than the collector capacitance due to the effects of diffusion.

If the unit is too wide, the signal applied at the electrode on the base layer will be attenuated so that only a portion of the base layer will be operative in the desired way. The remainder of the base layer will be dominated by the capacitative voltages and these will induce currents which will lead to large "active" capacitances appearing between emitter and collector.

It is evident that the effects involved will lead to rather detailed calculations for any particular case but that the physical principles required to design n-p-n transistors are simply extensions of those well established for p-n junctions and n-p-n transistors for the low bias, low frequency conditions.

The low noise figures of these transistors cannot be said to be explained in the absence of an established theory of noise generation. They are, however, in rough agreement with a theory based on noise modulation of the recombination mechanism.[17] This theory predicts that each element of volume is a source of (noise current)^2 proportional to the square of the deviation of minority carrier density from its normal value. Applying this criterion to the n-p-n structure and comparing it to the type A indicates that the observed difference of 40 db or more between noise figures can be accounted for in terms of the change in current densities and geometries.

X. ACKNOWLEDGMENTS

We are indebted to a number of our colleagues for encouragement and assistance. In particular, we recognize the contribution of J. A. Morton of the transistor development group, whose encouragement resulted in the techniques that made good p-n transistors possible. We are grateful to E. Buehler and R. M. Mikulyak who processed the germanium, to W. J. Pietenpol who prepared the particular unit studied, to R. L. Wallace for help with the measurements and the manuscript, and to W. van Roosbroeck for a critical reading of the manuscript.

APPENDIX

Proof of the Equality $G_{lr} = G_{rl}$

We shall prove this equality subject to the assumption that the electron density in the base layer is small compared to the thermal equilibrium hole density. Under these conditions, the effect of injected electrons on the potential distribution may be neglected so that the variation of the electron density from its equilibrium value may be treated by linear equations. In Fig. 10 we represent the situation considered. We shall denote by $I_{nl}(B_l, B_r)$ the current across J_l into the base carried by electrons. The symmetry relation $G_{lr} = G_{rl}$ is then established by proving that

$$I_{nl}(0, B) = I_{nr}(B, 0). \tag{A1}$$

In the base layer we shall suppose that the electrostatic potential ψ and the lifetime τ are arbitrary functions of position. The boundary condition on the external surfaces and at the metal contact will be taken as

$$I_n \cdot dS = -qn_1 s|dS|, \tag{A2}$$

where dS is the outward normal, s the surface recombination velocity and

$$n_1 = n - n_p \tag{A3}$$

is the deviation of n from the thermal equilibrium value.

We shall denote the solutions corresponding to potentials applied to the two junctions as follows:

$$B_l = B, \quad B_r = 0 \qquad n_1, \quad I_n \tag{A4}$$

$$B_l = 0, \quad B_r = B \qquad n_1', \quad I_n'. \tag{A5}$$

The currents in question are then

$$I_{nl}(0, B) = -\int_{J_l} I_n' \cdot dS \tag{A6}$$

$$I_{nr}(B, 0) = -\int_{J_r} I_n \cdot dS. \tag{A7}$$

The desired theorem is proved by considering the vector A:

$$A = (n_1/n_b)I_n' - (n_1'/n_b)I_n. \tag{A8}$$

Since $I_n \cdot dS$ is proportional to n_1 on the external surfaces, $A \cdot ds = 0$ on these surfaces. Hence, the integral of $A \cdot dS$ over the surface of the base region is

$$\int A \cdot dS = (qB/kT)[-I_{nl}(0, B) + I_{nr}(B, 0)] \tag{A9}$$

since on J_l and J_r we have

$$\begin{aligned} n_1/n_b &= (qB/kT) \text{ on } J_l \\ &= 0 \text{ on } J_r \end{aligned} \tag{A10}$$

$$\begin{aligned} n_1'/n_b &= 0 \text{ on } J_l \\ &= (qB/kT) \text{ on } J_r. \end{aligned} \tag{A11}$$

Furthermore, it can be shown that $\nabla \cdot A = 0$ in the base region and hence that (A9) is zero by Gauss's theorem so that (A1) is proved. The proof that $\nabla \cdot A = 0$ is accomplished by showing that $\nabla \cdot (n_1/n_b)I_n'$ is symmetrical in n_1 and n_1' so that the two terms in (A8) have cancelling divergences:

$$\begin{aligned} &\nabla \cdot (n_1/n_b)I_n' \\ &\quad = (1/n_b)(\nabla n_1 - n_1 \nabla \ln n_b) \cdot I_n' + (n_1/n_b)\nabla \cdot I_n' \\ &\quad = (1/n_b)[(\nabla n_1 + n_1 qE/kT) \cdot (q\mu_n n_1' E + qD_n \nabla n_1') \\ &\qquad\qquad - q(n_1 n_1'/\tau)], \end{aligned} \tag{A12}$$

which is seen to be symmetrical.

THE p-GERMANIUM TRANSISTOR

W. G. PFANN J. H. SCAFF

PROC. IRE VOL.38 pp.1151-1154 OCTOBER 1950

Produced outside the work performed under this contract.

The *p*-Germanium Transistor*

W. G. PFANN† AND J. H. SCAFF†

Summary—The transistor effect in *p*-type germanium is discussed and some properties are given for *p*-germanium transistors made in the laboratory. These exhibit higher cutoff frequency and somewhat lower current multiplication than their *n*-germanium counterparts. Under certain conditions a negative resistance "snap" effect is observed which is apparently peculiar to *p*-type germanium. Both types of transistor are governed by the same physical principles but they differ in the signs of the emitted carriers and of the bias voltages.

IN THE TRANSISTOR described by Bardeen and Brattain[1,2] two adjacent metallic points make rectifying contact to one surface of a small block of germanium. A third large-area contact comprises the base electrode. The point electrodes are designated emitter and collector and form the input and output terminals, respectively, of the transistor when connected as an amplifier. The basic material of the transistor described by Bardeen and Brattain is the electronic semiconductor, *n*-germanium.

The transistor effect has also been produced in the semiconductor *p*-germanium and a number of experimental *p*-germanium transistors have been made in the laboratory.[3] The preparation and properties of such *p*-germanium transistors are described herein, following a brief description of some characteristics of the semiconductors, *n*- and *p*-germanium.[4]

Germanium is an electronic semiconductor. Its electrical conductivity lies between those of metals and insulators. The conductivity can be increased by the addition of certain elements to the germanium. Such addition elements produce carriers of electrical charge which contribute directly to the conductivity. Addition elements fall into one of two classes, depending on the type of conductivity which they produce. Elements in Group V of the Periodic System donate conduction electrons to the germanium, thereby producing *n*-type (*n* for negative) conductivity. Elements of Group III accept an electron from the germanium lattice, leaving therein a net positive charge, or "hole," which can move freely through the lattice, producing *p*-type (*p* for positive) conductivity.[5] The existence of negative or positive carriers in semiconductors such as germanium can be demonstrated experimentally by measurement of the Hall voltage, which is a transverse electromotive force produced by a magnetic field in a current-carrying conductor, the sign of the Hall voltage depending upon the sign of the moving carriers.

In actual practice, relatively pure *n*- or *p*-germanium of resistivity on the order of 10 ohm-centimeters, has been used. Residual impurities may produce sufficient *n*-type conductivity for the requirements of transistors and high back-voltage rectifiers. Furthermore it has been shown by Scaff and Theuerer that such *n*-germanium can be thermally converted to *p*-germanium by heating at about 600°C or higher followed by quenching.[6,7] Such *p*-germanium, as well as *p*-germanium produced by additions of aluminum, have been used for the *p*-type transistors discussed herein.

While a good rectifying junction can be made between a metallic point and either *n*- or *p*-germanium, it should be noted that the polarity of the rectifying junction depends on the conductivity type of the semiconductor. Table I shows the polarity of the point with respect to the semiconductor in the two cases.

TABLE I

Conductivity Type of the Semiconductor	Polarity of Point	
	Forward Direction	Reverse Direction
n type	+	−
p type	−	+

In either type of transistor the collector is biased in the reverse direction and the emitter is usually biased in the forward direction. Operation of the *n*-germanium transistor may be explained on the basis that the current which passes through the emitter consists largely of holes, i.e., of carriers of opposite sign to those normally in excess in the body of the germanium. The holes are attracted to the negatively biased collector, so that a large part of the emitter current, introduced at low impedance, flows into the collector circuit and through a high impedance load, as may be seen in Fig. 1.

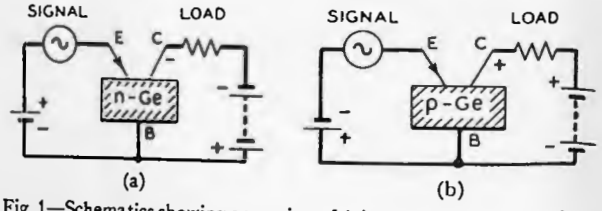


Fig. 1—Schematics showing operation of (a) *n*-germanium transistor, and (b) *p*-germanium transistor.

* Decimal classification: R282.12. Original manuscript received by the Institute, May 25, 1950.

† Bell Telephone Laboratories, Inc., Murray Hill, N. J.

[1] J. Bardeen and W. H. Brattain, "The transistor, a semi-conductor triode," *Phys. Rev.*, vol. 74, p. 230; July 15, 1948.

[2] J. Bardeen and W. H. Brattain, "Physical principles involved in transitor action," *Phys. Rev.*, vol. 75, pp. 1208–1225; April 15, 1949. And *Bell Sys. Tech. Jour.*, vol. 28, pp. 239–277; April, 1949.

[3] W. G. Pfann and J. H. Scaff, "The *p*-germanium transistor," *Phys. Rev.*, vol. 76, p. 459; August, 1949.

[4] For further information on electronic conduction in germanium, see footnote reference 2.

[5] J. H. Scaff, H. C. Theuerer, and E. E. Schumacher, "*P*-type and *N*-type silicon and the formation of the photovoltaic barrier in silicon ingots," *Jour. Met.*, vol. 185, pp. 383–388; June, 1949.

[6] J. H. Scaff and H. C. Theuerer, "Preparation of High Back Voltage Germanium Rectifiers," National Defense Research Council, NDRC 14–555, October 24, 1945.

[7] H. C. Torrey and C. A. Whitmer, "Crystal Rectifiers," McGraw-Hill Book Co., New York, N. Y., chap. 12; 1948.

The result is a voltage gain and a power gain of an input signal. There may be an amplification of current as well.

In the p-germanium transistor a large part of the emitter current consists of electrons, which enter the germanium and are drawn to the positively biased collector. Voltage gain and current gain occur, just as in the n-type transistor.

The characteristics of the two types of transistor are such that in both, the output voltage is in phase with the input voltage for the grounded-base connection shown in Fig. 1. The input and output currents, however, are 180° out of phase in both types.

A transistor construction which has been found useful for laboratory studies is shown in Fig. 2. The germanium wafer assembly and the contact spring assembly are force-fitted into a metallic sleeve which serves also as the base electrode. The germanium surface is etched as has been described[6,7] for high back-voltage rectifiers, the etchant not being particularly critical. The contact springs are mounted on supporting pins which have previously been molded in a bakelite compound. The contact springs are phosphor bronze wires, 0.002 inch in diameter. Their tips are usually several mils apart; the wider spacing shown in Fig. 2 was used for clarity of illustration.

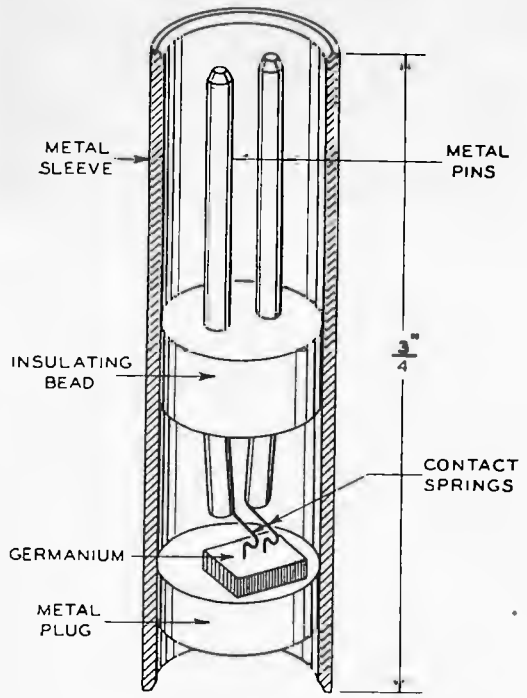


Fig. 2—Transistor construction used for laboratory studies.

The ability of the emitter to inject electrons into p-germanium is greatly increased by an electrical forming treatment.[8] If desired, both point electrodes may be pulsed, thereby enabling either point to be used as an emitter.[9]

The ac small signal performance of the p-germanium transistor at relatively low frequencies may be described with the help of the equivalent circuit of Fig. 3. The impedance associated with the emitter contact is r_e, and because the emitter junction is usually biased in the forward direction, r_e is small, of the order of hundreds of ohms. The impedance associated with the collector junction r_c is relatively high because the collector is biased in the reverse direction. The impedance in the semiconductor common to emitter and collector currents, r_b, represents a positive feedback. The transfer impedance r_m represents the active properties of the network. This network corresponds to the grounded-base connection.

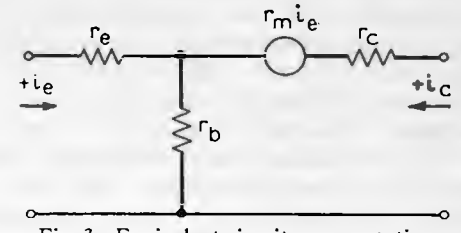


Fig. 3—Equivalent circuit representation of the transistor.

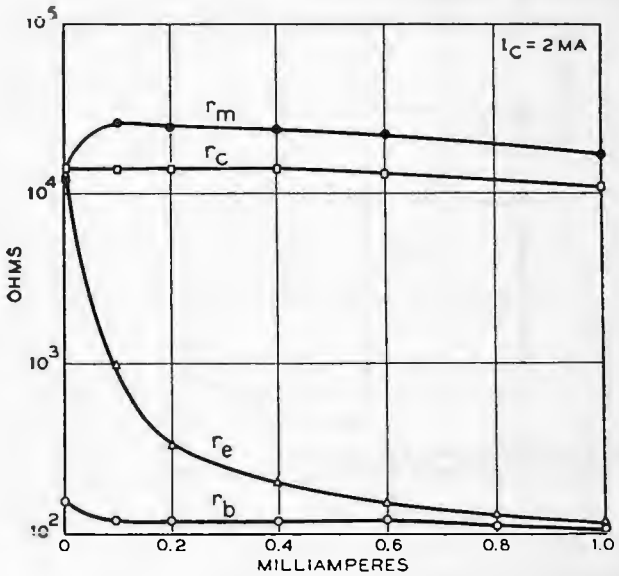


Fig. 4—Small-signal self- and transfer-impedances of a representative p-germanium transistor.

Some properties of a representative p-germanium transistor are shown in Fig. 4 in terms of these impedances. The r's are plotted as functions of emitter current for a constant collector current of +2.0 milliamperes. At values of emitter current above a few tenths of a milliampere the impedances have the following approximate values:

$r_e \sim 200$ ohms $r_c \sim 12{,}500$ ohms
$r_b \sim 100$ ohms $r_m \sim 22{,}000$ ohms

[8] J. Bardeen and W. G. Pfann, "Effects of electrical forming on the rectifying barriers of n- and p-germanium transistors," *Phys. Rev.*, vol. 77, pp. 401–402; December, 1950.

[9] W. G. Pfann, correspondence, "The transistor as a reversible amplifier," PROC. I.R.E., this issue, p. 1222.

The ratio r_m/r_e is approximately equal to α, the current multiplication factor and as may be seen, α is greater than 1, usually being from about 1.5 to 2. The significance of $\alpha > 1$ in terms of the physical operation of the transistor is that a given change in emitter current causes a larger change in the current through the collector.

As in the case of *n*-germanium, the gain of the *p*-germanium transistor falls off at frequencies in the megacycle range. Although we do not have an entirely complete comparison of the *n* and *p* devices, in which all quantities except frequency response are held constant, it does appear that the *p*-germanium transistor will operate at higher frequencies. Cutoff frequencies above 15 megacycles are common. (The cutoff frequency is that at which α is down 6 db.) Measurements of cutoff frequency have been made by Rack and are plotted against distance between emitter and collector for both types of transistor in Fig. 5.

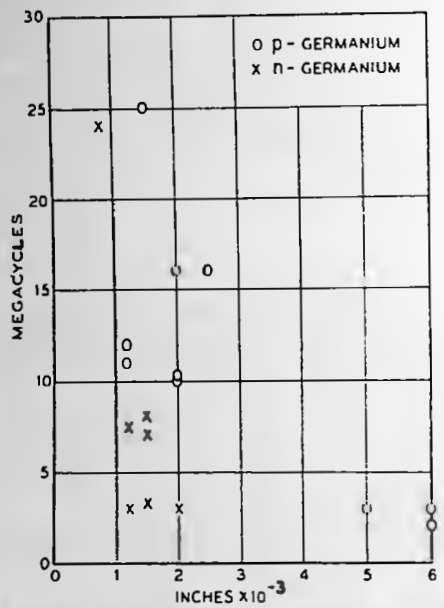


Fig. 5—Cutoff frequency of *n*-germanium and *p*-germanium transistors as a function of distance between emitter and collector.

An inherent feature of the *p*-germanium transistor may be at least partly responsible for its superiority in frequency response. It has been shown by recent measurements[10] that the mobility of electrons in germanium exceeds that of holes by a factor of about 1.5. Since the emitted carriers in the *p*-germanium transistor are electrons, and since it is dispersion in transit times of emitted carriers which limits the operating frequency, the observed superiority of the *p*-germanium transistor is at least partly to be expected.

The *p*-germanium transistor has a peculiarity which is found when the germanium is of particularly high resistivity. In such instances the forward current-voltage characteristic of the emitter has a negative resistance region of the voltage-maximum type, the peak voltage being on the order of several tenths of a volt. When the emitter bias exceeds this peak voltage, and if the series resistance in the emitter circuit is low, a sudden increase in I_e occurs which causes a corresponding change in the collector circuit. This action has been called the snap effect and may be seen in Fig. 6, which is a photograph of an oscilloscope trace in which the current-voltage characteristic of the collector junction is traced out at a frequency of 60 cycles per second, the voltage axis being horizontal, current axis vertical. The nearly horizontal portion of the trace represents the

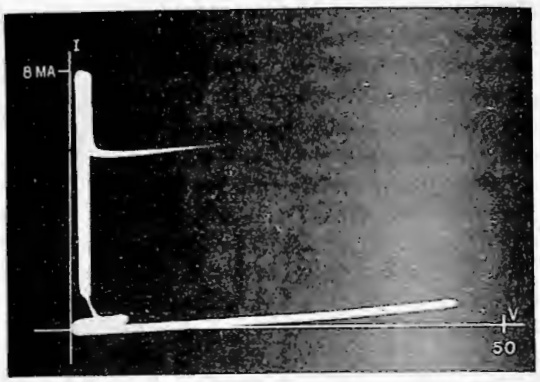


Fig. 6—Oscillograph pattern showing "snap" effect in *p*-germanium transistor.

reverse current through the collector for an emitter bias which is barely below the voltage maximum. At the extreme tip of the characteristic a small amount of positive feedback due to increase in collector current has caused the emitter bias to just exceed the voltage maximum. As a result the collector trace snaps suddenly to the high-current position as indicated by the vertical trace. The return trace remains in the low resistance position until the current falls to a small value, whereupon the characteristic again "snaps" back to the high-resistance position.

It is emphasized that the negative resistance in the forward current-voltage characteristic of the emitter is a diode effect, which can occur in a single point contact to *p*-germanium. It should not be confused with a similar effect which can be produced in an ordinary transistor by placing positive feedback resistance in the common base lead. However, the finite value of r_b does enter into the explanation of Fig. 6.

On the basis of a limited number of measurements there does not appear to be a significant difference in noise performance of *n*- and *p*-germanium transistors. Noise figures for *n*-germanium transistors are given in the literature.[11]

[10] W. Shockley, G. L. Pearson, and J. R. Haynes, "Hole injection in germanium," *Bell Sys. Tech. Jour.*, vol. 28, pp. 344–366; July, 1949.

[11] R. M. Ryder and R. J. Kircher, "Some circuit aspects of the transistor," *Bell Sys. Tech. Jour.*, vol. 28, pp. 367–400; July, 1949.

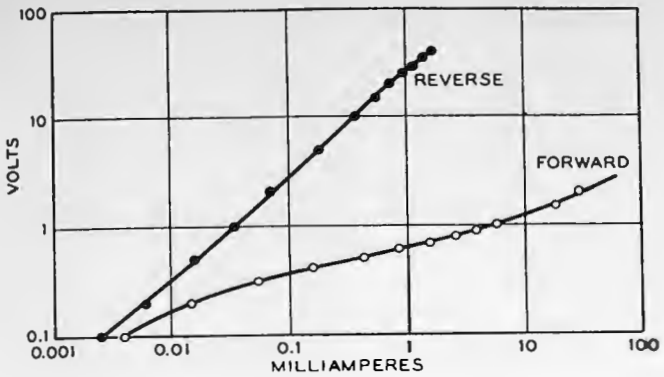


Fig. 7—Current-voltage characteristic of collector junction in a representative *p*-germanium transistor.

While it does not appear to have been recognized generally,[12,13] useful point contact rectifiers can be made with *p*-germanium. We have found that rectifier properties approaching those of high back-voltage *n*-germanium can be obtained by pulsing high-resistivity, thermally-converted *p*-germanium. The forward resistance before pulsing is high, on the order of thousands of ohms at one volt, but is greatly decreased by pulsing, to values on the order of 100 ohms or less. The reverse resistance is essentially unaffected by pulsing, and peak reverse voltages above 100 volts are common. A theoretical interpretation of electrical forming effects in *n*- and *p*-germanium has been made by Bardeen and Pfann.[8]

In Fig. 7 are shown the current-voltage characteristics of the collector junction in a representative *p*-germanium *transistor* at zero emitter current. Here, the reverse impedance is lower than the highest which can be obtained for it has been found, in common with the *n*-germanium transistor, that too high a reverse impedance is incompatible with best collector performance.

The attainment of transistor action in both *n*- and *p*-germanium provides further illustration of the dual nature of conduction processes in semiconductors such as germanium. In the *n*-type transistor, positive holes are injected into a semiconductor normally having an excess of electrons; in the *p*-type transistor, electrons are injected and the semiconductor normally has an excess of holes. In general, the properties of the *n*- and *p*-germanium transistors are similar, although in devices made up to the present time the *n*-germanium device appears to be superior in α and inferior in frequency response to the *p*-germanium transistor. The present differences may be as much a result of the early status of transistor technology as of fundamental differences between the semiconductors themselves.

[12] See p. 63 of footnote reference 7.
[13] W. C. Dunlap, Jr., and E. F. Hennelly, "Non-rectifying," *Phys. Rev.*, vol. 74, p. 976; October, 1948.

THE TRANSISTOR
A NEW SEMICONDUCTOR AMPLIFIER

J. A. BECKER J. N. SHIVE

ELECTRICAL ENGINEERING VOL.68 pp.215-221 MARCH 1949

Produced outside the work performed under this contract

The Transistor—A New Semiconductor Amplifier

J. A. BECKER
FELLOW AIEE

J. N. SHIVE

"TRANSISTOR" is the name which has been given to the semiconductor amplifier triode recently developed at the Bell Telephone Laboratories. The prefix "trans" designates the translational property of the device, while the root "istor" classifies it as a solid circuit element in the same general family with resistor, varistor, and thermistor.

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 transistor differs advantageously from a vacuum tube in several important respects. It has no vacuum. It has no filament; consequently, it consumes no filament power and requires no warm-up time. It is both smaller and lighter than any commercially available vacuum tube. Within the present limitations of their power handling capacity, noise, and frequency response, transistors can perform many of the tasks now performed by vacuum tube triodes. They have been successfully demonstrated in radio-frequency, intermediate-frequency, and audio-frequency amplifiers, oscillators, mixers, and pulse generators.

The transistor principle was discovered by J. Bardeen, W. H. Brattain, and W. Shockley[1,2,3] in the course of theoretical and experimental studies on the possibility of controlling the resistances of thin semiconductor layers by the application of electric fields strong enough to penetrate their surfaces. In the course of these investigations Bardeen and Brattain found that when two suitably arranged contacts are made to the surface of a semiconductor, the current-voltage relationship of one of them can be altered by the passage of current through the other. The extent of this interaction is such that a small signal introduced into a circuit containing the control contact is reproduced in amplified form in the circuit containing the main contact.

This article describes the characteristics and behavior of two kinds of transistor which have received considerable study at the Bell Telephone Laboratories and which differ from each other in appearance, construction, and electrode geometry. The first of these, the type-*A* transistor, has been made in considerable numbers, and is available for experimental use. The second kind, the double-surface transistor, is still in early-stage development.

Acknowledgments. We wish to thank our many colleagues at the Bell Telephone Laboratories who have made contributions of material and advice in the course of our preparation of this article. In particular, we are pleased to acknowledge our indebtedness to J. A. Lehans for his measurements of the static characteristics of Figs. 5 to 8; to H. C. Montgomery for the data on transistor noise, including the information presented in Figs. 10 and 11; to C. O. Mallinckrodt for the results of his studies on the frequency dependance of α appearing in Fig. 12; to A. J. Rack for his findings on the frequency dependence of r_e, r_c, r_b and r_m, and to J. A. Morton and R. M. Ryder for critical discussions on network analysis. We wish to point out that the design and processing of the type A transistor follow closely those of an earlier model which was developed by W. G. Pfann. This model was subsequently modified by R. J. Kircher to adapt it to the present socketing scheme.

CONSTRUCTION OF TRANSISTORS

The type-*A* transistor consists of a small wafer of the semiconductor germanium onto one surface of which two point contacts are made, side by side and close together. These contacts, together with their immediate electric connections, are called the "emitter" and the "collector" respectively, for reasons which will appear later. A large-area contact to the opposite surface of the wafer is called the "base electrode." These elements are assembled into a cartridge which is pictured in full and cutaway photographs in Figure 1. To assist in the understanding of the relationship of the various parts, the longitudinal section drawings of Figure 2 are presented.

In the fabrication of a type-*A* unit a slice of germanium cut from a high-back-voltage ingot is ground flat on both sides, copper-plated and tinned on one side, and diced into small squares with a diamond wheel. One of these squares then is sweated onto the brass base plug which, after the administration of the proper surface treatment[4,5] to the germanium wafer, is force-fitted into the base end of the cartridge.

The contact assembly is made by molding the two support pins into a cylindrical plug of insulating material. The 0.005-inch phosphor bronze wires, which previously have been pointed at one end by bevel-grinding and polishing, are spot-welded to the ends of these pins.

After welding, the wires are bent[4,5] in a jig into the

Full text of conference paper, "The Transistor, a New Solid State Amplifier," recommended by the AIEE technical program committee for presentation at the AIEE winter general meeting, New York, N. Y., January 31–February 4, 1949.

J. A. Becker and J. N. Shive are with the Bell Telephone Laboratories, Inc., Murray Hill, N. J.

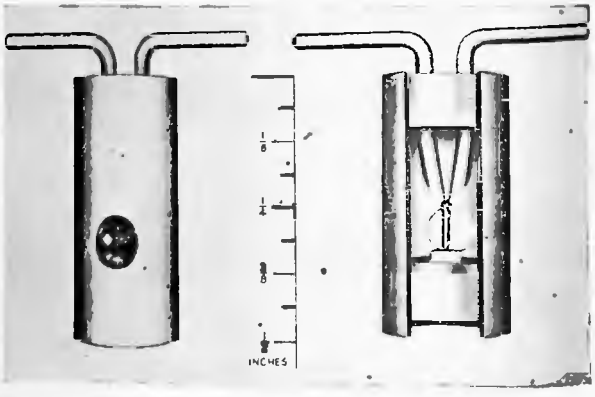


Figure 1. Full-view and cutaway photographs of the type-A transistor

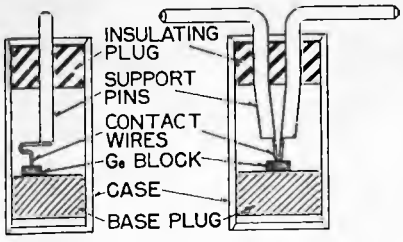


Figure 2. Longitudinal section drawings of the type-A transistor

cantilever form shown in Figures 1 and 2. After a final manual adjustment of the points to equal vertical clearance and specified lateral separation, the assembly is forced into the other end of the cartridge until contact is made between the points and the germanium surface. An additional 0.002 inch push compresses the springs enough to insure that the contact will be permanent under conditions of ordinary use. The cartridge then is wax-filled[4,5] to improve its mechanical stability. After a final electrical forming treatment[4,5] of one or both contacts, the emitter-collector polarity is selected, and the electrode pins are appropriately bent according to the scheme of Figure 3. The metal case is the base connection of the unit. The size of the type-*A* transistor can be seen by comparison with the scale in Figure 1. Its weight is 1.3 grams.

The double-surface transistor[6] differs from the type-*A* device in that the emitter and collector points bear on the opposite faces of a thin piece of semiconductor, to which a large-area base contact also is provided. This form of transistor first was produced with germanium in the shape of an acutely tapered wedge, with the two contacts bearing oppositely near the thin edge. A mounted form of the double-surface transistor is described in a companion article by W. E. Kock and R. L. Wallace, Jr., appearing elsewhere in this issue.[7] In this model the germanium is cut into a pill-shaped cylindrical wafer with a dimple ground into the center of one or both sides, so that the thickness at the center is only a few thousandths of an inch. Here the emitter and collector bear oppositely upon the semiconductor in a coaxially arranged cartridge mounting.

CONVENTION REGARDING SIGN OF CURRENTS AND VOLTAGES

Since transistors undoubtedly will receive considerable attention in the engineering literature, it is desirable that all engineers use the same convention regarding direction of voltages and currents. This matter has been investigated at the Bell Telephone Laboratories by a group of engineers who have made the tentative proposals that: all voltages are to be referred to a common reference, which in this article will be taken as zero for the potential of the base; all currents are positive if they flow in the direction that positive voltages would produce. This convention is illustrated in Figure 4, which shows a schematic diagram of a transistor circuit in which the arrows for currents and voltages indicate positive directions. In most transistors made at the present time E_1, V_1, and I_1 are positive, while E_2, V_2, and I_2 are negative when the transistor is properly biased for class-*A* amplification. Cases have arisen, however, in which this is not true and more such cases unquestionably will arise in the future. This convention, which is consistent with the new proposal for electron tubes being made by the Institute of Radio Engineers, may afford a simple and unambiguous method of describing these cases.

STATIC CHARACTERISTICS OF TRANSISTORS

In a transistor both the emitter and the collector contacts are rectifiers. The emitter contact is operated in the forward or low resistance direction while the collector is operated in the reverse or high resistance direction. When the points are far apart the emitter currents and voltages do not affect the collector currents and voltages. However, when the contacts are near each other, the collector voltage versus current relation is greatly affected by the values of the emitter current or voltage. This interaction is shown for a representative type-*A* unit in Figure 5. This is a plot of V_2 versus I_2 for a series of constant values of I_1 (solid lines) and for a series of constant values of V_1 (dashed lines).

Current is chosen as the independent variable because such a choice simplifies the physical interpretation of the actions in the transistor, the voltages are single valued functions of the currents, and in certain regions of their characteristics, transistors sometimes become unstable when the external terminals are short-circuited to alternating current.

In Figure 5 the solid line for which $I_1=0$ is merely the voltage versus current curve for the collector as a diode when the emitter circuit is open. If a constant positive current is made to flow in the emitter circuit, the resistance in the collector circuit decreases and the curves shift toward larger negative currents. Over a considerable region of the plot, the increase in the collector current for a constant collector voltage is greater than the change in the emitter current producing it. This means that the transistor is a current amplifier with a current amplification factor $\alpha \equiv \frac{\Delta I_2}{\Delta I_1}$ greater than unity. In most units α is approximately 2.0. Occasionally values higher than ten are found. Even if α is only 1.0 or less, power gains still are obtained since the current is transferred from a low resistance circuit to a high resistance circuit.

If the emitter contact is directly connected to the base the V_2 versus I_2 relationship is shown by the dashed curve, marked $V_1=0$. Similarly, if the emitter is connected to the base through a battery having zero internal resistance and a small positive or negative electromotive force, the

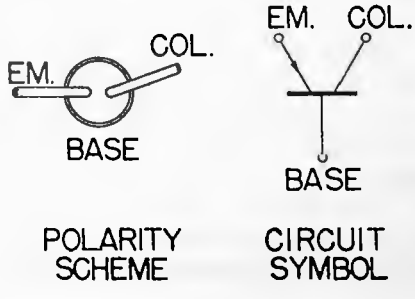


Figure 3. Polarity convention and circuit symbol for type-A transistor

series of dashed curves is obtained. Note that these dashed curves may have negative slopes, indicating negative variational resistance when V_1 and I_2 have relatively large values. This means that for low input and output resistances the transistor may be unstable. Stability can be secured by inserting a sufficiently high resistance in the output circuit. The same result can be accomplished by including a high resistance in the input circuit.

Note that any two of the four quantities I_1, V_1, I_2, V_2 determine the other two. The interrelationship between these quantities can be presented profitably in three other plots describing the same unit. The data for these three plots all are contained in Figure 5. Figure 6 shows curves of V_1 versus I_1 for a series of constant values of I_2. The curve for $I_2=0$ is the forward voltage-current diode characteristic of the emitter contact. This figure shows the influence of various collector currents on the forward characteristics of the emitter. In this case, for a constant V_1, the increase in I_1 is always less than the corresponding increase in I_2; that is, $\frac{\Delta I_1}{\Delta I_2}<1$.

Figure 7 shows how the collector voltage varies with the emitter current for constant collector currents. From these curves it is seen that the collector voltage for a constant collector current decreases rapidly for rather small increases in emitter current. This must mean that the resistance near the collector decreases rapidly for an increase in emitter current. As will appear more clearly later, the slopes of these curves measure the forward transfer resistance for a-c currents in the emitter circuit.

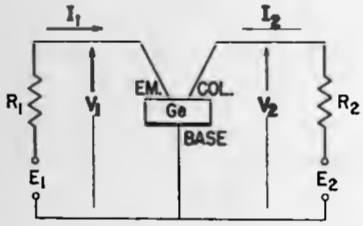


Figure 4. Convention for directions of voltages and currents in transistor circuits

Figure 8 shows how the emitter voltage varies with the collector current for a series of constant emitter currents. For $I_1=0$, the emitter is "floating" and this curve measures the floating potential of the emitter for various values of the collector current. For positive emitter currents, the decrease of V_1 with I_2 is less rapid. The slopes of these curves measure the feedback resistance for small a-c currents impressed in the collector circuit.

LARGE SIGNAL PERFORMANCE

From the data in Figures 5 and 6 it is possible to predict the input current and output current for any electromotive force and external resistance in the input circuit coupled with any electromotive force and load resistance in the output circuit. To do this one proceeds as follows. Draw a load line across the curves of Figure 5 with voltage intercept equal to the electromotive force in the collector circuit and negative slope equal to the load resistance. Draw another load line across the curves of Figure 6 appropriate to the values of external resistance and electromotive force in the emitter circuit. Replot this emitter load line on the collector characteristics of Figure 5. The point of its intersection with the collector load line will be an operating point of the transistor. Changing the signal voltage in the emitter circuit will cause a change in the position of the emitter load line.

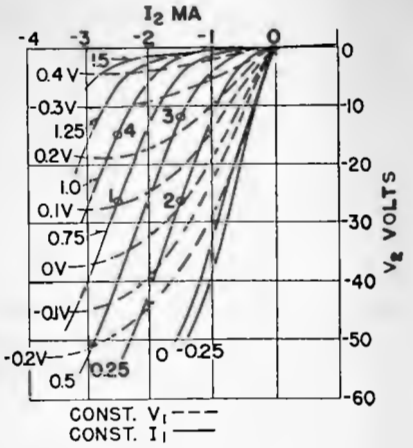


Figure 5. Family of collector characteristics for type-A transistor

By following through with this procedure the corresponding successive instantaneous operating points of the transistor can be obtained. The large signal performance thus can be deduced in a way analogous to that used for vacuum tube performance.

SMALL SIGNAL PERFORMANCE

The small signal performance about any particular d-c operating point can be deduced easily by analysis with reference to an equivalent circuit. From this equivalent circuit the a-c behavior of the transistor can be calculated for any input and output conditions. In particular, equations can be derived for input impedance, output impedance, and power gain.

Since we wish to consider currents as independent variables, we can write

$$V_1=f_1(I_1, I_2) \quad (1)$$

$$V_2=f_2(I_1, I_2) \quad (2)$$

where f_1 and f_2 are functions determined by the experiments from which Figures 6 and 5 are plotted.

By expanding these functions in a Taylor series about

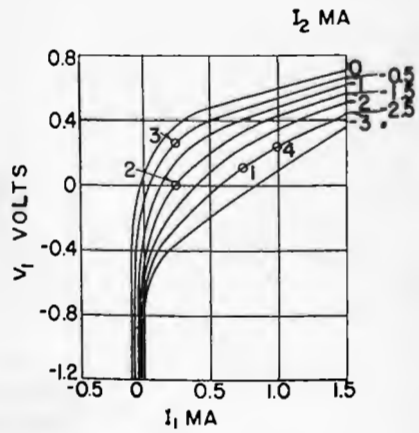


Figure 6. Family of emitter characteristics for type-A transistor

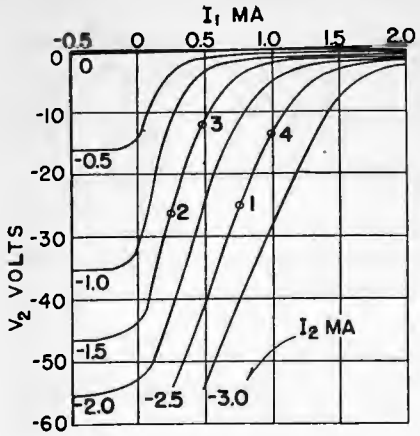


Figure 7. Family of forward transfer characteristics for type-A transistor

any d-c operating point and considering only first order terms, we obtain the following equations:

$$\Delta V_1 = \frac{\partial f_1}{\partial I_1}\,\Delta I_1 + \frac{\partial f_1}{\partial I_2}\Delta I_2$$

$$\Delta V_2 = \frac{\partial f_2}{\partial I_1}\,\Delta I_1 + \frac{\partial f_2}{\partial I_2}\,\Delta I_2$$

In these equations the partial derivatives with respect to one current imply that the other current is kept constant. The partial derivatives have the dimensions of impedances. Therefore, we may write

$\frac{\partial f_1}{\partial I_1} = \frac{\partial V_1}{\partial I_1} = R'_{11}$, the input impedance for a-c open-circuited output

$\frac{\partial f_1}{\partial I_2} = \frac{\partial V_1}{\partial I_2} = R'_{12}$, the backward transfer impedance for a-c open-circuited input

$\frac{\partial f_2}{\partial I_1} = \frac{\partial V_2}{\partial I_1} = R'_{21}$, the forward transfer impedance for a-c open-circuited output

$\frac{\partial f_2}{\partial I_2} = \frac{\partial V_2}{\partial I_2} = R'_{22}$, the output impedance for a-c open-circuited input

In terms of the static characteristics already presented

R'_{11} = slopes of curves of Figure 6
R'_{12} = slopes of curves of Figure 8
R'_{21} = slopes of curves of Figure 7
R'_{22} = slopes of curves of Figure 5

Well-known linear network theory[8] leads to a number of possible equivalent circuits containing resistances and internal generators. A particularly useful form of equivalent circuit is shown inside the dashed box in Figure 9. In this circuit there are three resistances: r_e associated with the emitter contact, r_c with the collector contact, and r_b with the base contact. The active property of the network is described by the inclusion of an internal generator $r_m\Delta I_1$ whose electromotive force is the product of the mutual transfer resistance r_m and the variation of the input current ΔI_1. The convention for positive currents and voltages is as shown by the arrows in this figure. The quantities ΔI_1 and ΔI_2 can be replaced by alternating currents i_1 and i_2 whose value is small compared with the bias currents I_1 and I_2. The figure does not show the external batteries and bias currents. v_g is an alternating voltage signal generator in the input circuit.

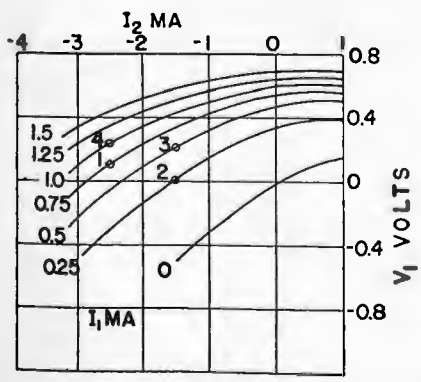


Figure 8. Family of backward transfer characteristics for type-A transistor

It is easy to deduce then, that the four resistances in this equivalent circuit are related to the open-circuit input, the output, and the transfer resistances by the following equations:

$R'_{11} = r_e + r_b$
$R'_{12} = r_b$
$R'_{21} = r_m + r_b$
$R'_{22} = r_c + r_b$

From Figures 5 to 8 it follows that the open-circuit resistances and hence the parameters in the equivalent circuit will depend on the values of the bias currents I_1 and I_2. They also will vary from one unit to the next. For a representative type-*A* transistor in a good operating region the equivalent circuit resistances might be: $r_e = 250$ ohms; $r_b = 250$ ohms; $r_c = 20{,}000$ ohms; $r_m = 40{,}000$ ohms. Since transistors are in an early stage of development, appreciable variations from unit to unit are to be expected. As the transistor art develops these variations undoubtedly will be reduced.

The following expressions can be derived from circuit analysis of the equivalent circuit. Numerical examples are given for $R_1 = 500$ ohms, $R_2 = 10{,}000$ ohms, and the typical equivalent circuit values have been given in the foregoing paragraph:

$$\text{Input impedance } R_{11} = R'_{11} - \frac{R'_{12}R'_{21}}{R'_{22} + R_2}$$

$$= (r_e + r_b) - \frac{r_b(r_m + r_b)}{(r_c + r_b) + R_2} \simeq 500 - 330 = 170 \text{ ohms}$$

$$\text{Output impedance } R_{22} = R'_{22} - \frac{R'_{12}R'_{21}}{R'_{11} + R_1}$$

$$= (r_c + r_b) - \frac{r_b(r_m + r_b)}{r_e + r_b + R_1} \simeq 20{,}000 - 10{,}000 = 10{,}000 \text{ ohms}$$

$$\text{Available gain} = \frac{\text{Power to matched load}}{\text{Maximum power available from generator}}$$

$$= \frac{R_1}{R_{22}}\left(\frac{R'_{21}}{R'_{11} + R_1}\right)^2 = \frac{R_1}{R_{22}}\left(\frac{r_m + r_b}{r_e + r_b + R_1}\right)^2$$

$\simeq 80$ (or 19 decibels).

Maximum available gain = Available gain when $R_1 = R_{11}$

$$= \frac{R_{11}}{R_{22}}\left(\frac{R'_{21}}{R'_{11} + R_{11}}\right) \simeq 130 \text{ (or 21 decibels)}$$

If r_b is reduced to zero these expressions for impedance and gain become much simpler. Also, the gains become

smaller because of the elimination of the positive feedback supplied by finite r_b. Thus:

$R_{11} = R'_{11} = r_e = 250$ ohms
$R_{22} = R'_{22} = r_c = 20{,}000$ ohms

$$\text{Available gain} = \frac{R_1}{R'_{22}}\left(\frac{R'_{21}}{R'_{11}+R_1}\right)^2 \simeq \frac{R_1}{r_c}\left(\frac{r_m}{r_e+R_1}\right)^2 \simeq 71$$

$$\text{Maximum available gain} = \frac{R'_{11}}{R'_{22}}\left(\frac{R'_{21}}{2R'_{11}}\right)^2 = \frac{(R'_{21})^2}{4R'_{22}R'_{11}} = \frac{1}{4}\frac{r_m r_m}{r_c r_e} \simeq 80$$

In transistor production it is considered desirable to keep r_b as small as possible, despite the reduction in available gain. This provides a greater margin against instability. Feedback always can be supplied externally when desired.

A family of collector characteristics for a double-surface unit is presented in Figure 4 of the afore-mentioned Kock-Wallace article. The chief point of difference between these curves and those for the type-*A* unit of Figure 5 is in the relative positions of the dashed lines for constant V_1. This difference can be accounted for by the fact that the feedback resistance r_b is about 250 ohms for type-*A* units, while for double-surface units it is around 500 ohms.

Table I. Comparison of D-C and A-C Measurements of Transistor Circuit Constants

Point	I_1 (Milliamperes)	I_2 (Milliamperes)	Circuit Constant	A-C Measurement (Ohms)	Slope of Static Characteristic (Ohms)
1	0.75	2.5	R'_{11}	575	525
			R'_{12}	272	288
			R'_{21}	50,000	53,500
			R'_{22}	28,000	25,600
2	0.25	1.5	R'_{11}	925	1,040
			R'_{12}	278	270
			R'_{21}	67,000	73,000
			R'_{22}	31,000	28,000
3	0.5	1.5	R'_{11}	625	580
			R'_{12}	252	240
			R'_{21}	43,400	45,300
			R'_{22}	22,900	20,800
4	1.0	2.5	R'_{11}	520	480
			R'_{12}	266	240
			R'_{21}	46,500	39,500
			R'_{22}	26,400	23,000

COMPARISON BETWEEN PARAMETERS DETERMINED FROM STATIC CHARACTERISTICS AND FROM A-C METHOD

Table I compares the values of R'_{11}, R'_{12}, R'_{21}, and R'_{22} determined from the slopes of the d-c static characteristics with the values determined by an a-c test equipment designed by J. H. Felker for the direct measurement of these quantities. In the latter case the a-c test current was 10^{-4} ampere at a frequency of 5 kc per second. The comparison is given for the four bias conditions indicated by circles in Figures 5 to 8. The agreement is as good as the uncertainty in the determination of the slopes of the curves. This fact justifies the assumptions underlying the network analysis and shows that the equivalent circuit constants are the same at 5 kc per second as at zero cycles per second.

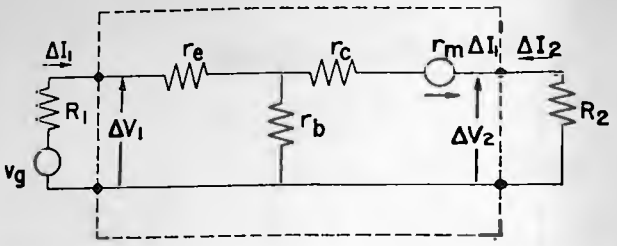


Figure 9. Equivalent T network for describing transistor small signal performance

POWER-HANDLING CAPACITY

A complete analysis of the power delivering ability of a transistor can be made from data of the sort presented in Figure 5. As with vacuum tubes, the maximum useful power obtainable from a transistor is a compromise with allowable distortion. The transistor of Figure 5, operated on a 25,000-ohm class-*A* load line about the biasing point represented by circle number 1, would deliver six to eight milliwatts of useful power with a distortion tolerable in the final stage of an ordinary audio amplifier. In class-*B* operation, typical units of both types deliver 25 to 30 milliwatts per push-pull pair. Occasionally units are found from which 50 or 60 milliwatts of reasonably undistorted power can be obtained per pair.

FORMING TREATMENT

When the two point contacts are pressed first against the germanium they are likely to have similar characteristics, and either point may be used as the collector. In some cases, the emitter current in the forward direction may have only a small influence on the collector characteristics. Transistor action usually can be improved by a forming treatment.[1]

One such treatment comprises passing a relatively large current through the collector point in the reverse direction. The effect of this process is to reduce the reverse resistance of the collector and to increase the influence of the emitter current on the collector characteristics.

NOISE IN TRANSISTORS

In discussing transistor noise, use will be made of a theorem[9] which states that the noise in a 4-terminal network can be considered as originating in two noise generators

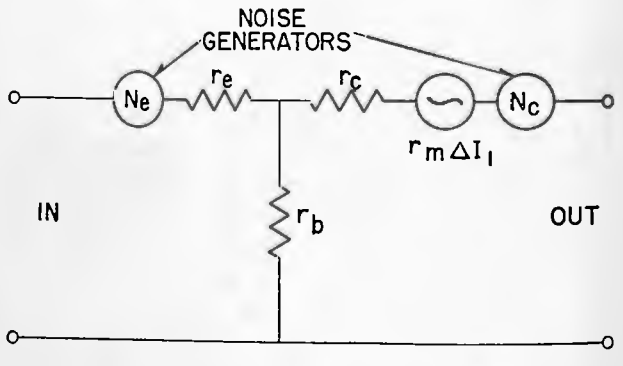


Figure 10. Transistor noise equivalent circuit

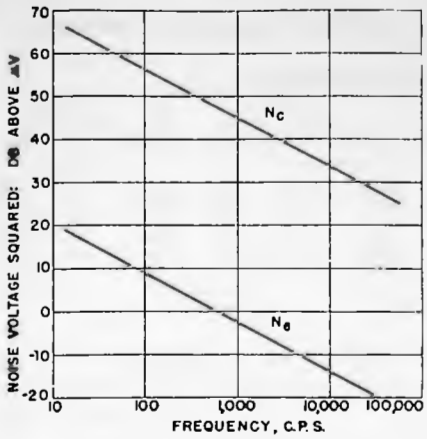


Figure 11. Noise versus frequency plot for a typical transistor

included in two of the arms of the network, whose other components are regarded then as ideally noiseless. In the transistor equivalent network already developed, the two generators will be considered as voltage generators, placed in the emitter and collector arms of the circuit, as shown in Figure 10. A generally satisfactory specification of transistor noise now can be made by giving the noise voltage N_e and N_c of these two generators per unit band width, the dependence of these voltages on frequency, and the magnitude and phase of any correlation existing between the voltages.

In making measurements of these quantities, the transistor under test is biased to a d-c operating point previously selected for good amplifier gain performance. The biases are applied through chokes having practically infinite impedances at all frequencies of testing interest. The open-circuit noise voltages then are measured directly by placing a narrow-band voltmeter, such as a General Radio wave analyzer, across the input and output terminals in turn. The frequency dependence of the noise voltages is obtained by sweeping the frequency spectrum with the pass band of the voltmeter. Measurements on the correlation between N_e and N_c have been based so far on the challengeable assumption that, at least at low noise frequencies (~1 kc per second), any correlation between the two would be either wholly in phase or wholly in opposition. The voltages N_c and N_e are amplified until their rms amplitudes

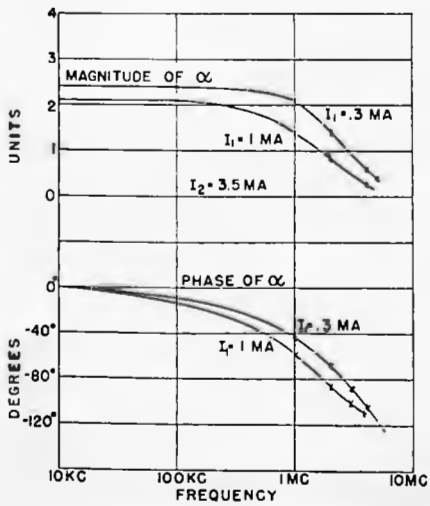


Figure 12. Current gain versus frequency plot for a type-A transistor

are equal; the two instantaneous voltages then are added, under conditions such that the impedance to each is the same as when each was measured separately, and the resulting rms voltage is observed. If N_e and N_c are completely random and no correlation exists, the voltage sum would be $\sqrt{2}$ times either one alone. For 100 per cent correlation with zero phase difference the voltage sum would be twice either voltage alone, and for 100 per cent correlation with 180 degree phase difference the sum would vanish.

Measurements of some of the foregoing noise quantities have been made for about three dozen transistors, including individuals of both types. Typical values for the noise voltages are $N_e \sim 1$ microvolt and $N_c \sim 100$ microvolts in a band of unit width at 1,000 cycles per second Most of the units measured have noise voltages falling between one-third and three times these values. A generally accepted measure of noise performance in an amplifier is its noise figure, defined as the ratio of the noise power actually available from the output terminals to the noise power which would be available if N_e and N_c were both zero and the only contributing noise source were thermal noise in the input signal generator. For transistors with typical impedances and gain properties operated in typical circuits, the noise figures for a narrow band centered at 1,000 cycles per second run from 55 to 70 decibels. Various ways of reducing the collector noise of transistors have been suggested and tried, some with success. It is expected that, by resorting to procedures by which noise has been successfully brought down in the crystal diode art, less noisy transistors will be produced.

In the frequency variation of both N_e and N_c, it has been found approximately that the noise voltage squared per unit band width is inversely proportional to frequency. Figure 11 shows a typical noise versus frequency plot in which this relationship is demonstrated for N_e and N_c. The frequency dependence has been examined in detail from 20 cycles per second to 16 kc per second for some of the transistors investigated and determined by spot checks at 450 kc per second and one megacycle per second for occasional transistors.

In most transistors, correlation between N_e and N_c is found to exist, although of unpredictable magnitude. Within the assumed restrictions on phase already stated, negative correlations up to 80 per cent have been found.

FREQUENCY LIMITATIONS

The highest frequency at which a transistor can be used in a given application is a matter of interest to the engineer. The information at present available on this subject indicates that useful transistor performance is limited to frequencies below about ten megacycles.

Because of the comparative ease of its measurement as a ratio of two alternating currents, the current amplification factor α has been studied in some detail as a function of frequency. Figure 12 shows plots of α versus frequency for a type-*A* transistor at constant collector bias current and two different values of emitter bias current. The phase angle of α is plotted in the lower part of the figure for the same biasing conditions.

Approximate measurements, on a limited number of transistors, of the individual elements of the equivalent circuit show that r_b, r_c and r_e are reasonably constant up to at least ten megacycles per second, that r_m alone changes with frequency, and that the changes in r_m are sufficient to account for the variations in α already presented. These findings are in accord with the existing knowledge that the forward and reverse resistances of germanium diodes are practically constant in the frequency range depicted. The variation of r_m with frequency can be attributed to a transit-time dispersion associated with the interaction between the emitter and collector currents.

Some tentative verification of this transit-time supposition has been obtained in a set of α versus frequency measurements on a series of double-surface units having different germanium thicknesses at the contact site. For units with thicker germanium the α curves fall off at lower frequencies, while the correspnding phase lags are greater at any given frequency in the dispersion region.

Double-surface transistors do not differ greatly from type-*A* units in the frequency dependence of α.

Table II. Equivalent Circuit Parameters and Their Temperature Coefficients

I_1 and V_2 Are Kept Fixed for Each Unit

Transistor	r_b (Degrees Centigrade)	$\beta \times 10^4$ 0–25	$\beta \times 10^4$ 25–50	r_e (Degrees Centigrade)	$\beta \times 10^4$ 0–25	$\beta \times 10^4$ 25–50	r_c (Degrees Centigrade)	$\beta \times 10^4$ 0–25	$\beta \times 10^4$ 25–50	r_m (Degrees Centigrade)	$\beta \times 10^4$ 0–25	$\beta \times 10^4$ 25–50
A1790	140	56	28	280	28	28	16,000	−240	−48	41,000	+60	−28
A1791	108	44	−28	220	28	−20	36,000	+100	−160	40,000	−112	−120
A2200	107	32	32	180	20	20	9,300	+12	−48	12,000	+64	+16
B89	210	16	−28	147	18	21	30,000	−40	−110	52,000	−40	−40
B90	500	0	0	315	−7	84	14,500	−160	+12	29,000	−80	−12

Table III. Insertion Gain in Decibels Versus Ambient Temperature

Transistor	I_1 and V_2 Fixed for Each Unit 0	25	50 (Degrees Centigrade)	E_1, R_1, E_2, R_2 Fixed for Each Unit 0	25	50 (Degrees Centigrade)
A1790	26.2	26.2	26.2	26.6	27.3	26.7
A1791	23.3	24.3	25.0	22.4	23.5	22.9
A2200	18.9	19.5	19.9	20.1	20.6	20.8
B89	26.2	26.5		26.0	26.8	27.2
B90	22.1	24.0	23 8	21.6	23.3	23.8

AMBIENT TEMPERATURE EFFECTS

In planning for the practical application of the transistor it is necessary to know the effect of the changes in ambient temperature which the device may experience in use. Experiments were performed to obtain such information.

The characteristics of three type-*A* and two double-surface transistors were measured successively at 25, 50, 25, 0, and 25 degrees centigrade. The performance characteristics chosen were the equivalent circuit parameters and the insertion gain. Since the d-c bias conditions change with ambient temperature it is necessary to decide which bias conditions are to be kept constant. For one set of tests, I_1 and V_2 were chosen for each unit at 25 degrees centigrade so as to maximize the gain and minimize the distortion. These bias conditions then were kept constant for succeeding ambient temperatures. The results are given in Table II and the left part of Table III. In Table II, the quantity β is the temperature coefficient for the corresponding resistance. Values of β are given for 0 to 25 degrees centigrade and for 25 to 50 degrees centigrade.

It is to be noted that the temperature coefficients are moderate and that they may be either positive or negative. The coefficients for r_b and r_e are smaller than those for r_c and r_m. The largest values occur in r_c and may amount to about one per cent per centigrade degree.

In another test the external electromotive forces and resistances were kept constant for each unit. R_1 was 500 ohms for all units. R_2 was 20,000 ohms for type-*A* and 10,000 ohms for double-surface units. For each unit, E_1 and E_2 were chosen to maximize the gain and minimize the distortion at 25 degrees centigrade and then were kept constant at succeeding temperatures. In this test only the gains were measured. The results are given in the right side of Table III. The gain changes are similar to those in the left side of Table III.

In most cases the gain increases with temperature. The increase may amount to one decibel in 25 degrees centigrade. Between 25 and 50 degrees centigrade the gain may decrease by a fraction of a decibel. These changes in gain are less than might be expected from the changes in equivalent circuit constants, probably due to the fact that changes in circuit constants tend to cancel each other.

In the course of these tests it was found that the values of the circuit constants determined at 25 degrees centigrade after a treatment at either 50 or 0 degrees centigrade differed from the original values by as much as a factor of two either up or down. Similarly the gain might permanently increase or decrease by about two decibels. Such changes are due to slight shifts in either the collector or emitter points resulting from stresses caused by the different expansion coefficients of the materials used.

REFERENCES

1. The Transistor, A Semiconductor Triode, J. Bardeen, W. H. Brattain. *Physical Review* (New York, N. Y.), volume 74, 1948, page 230.

2. Nature of the Forward Current in Germanium Point Contacts, W. H. Brattain, J. Bardeen. *Physical Review* (New York, N. Y.), volume 74, 1948, page 231.

3. Modulation of Conductance of Thin Films of Semi-Conductors by Surface Charges, W. Shockley, G. L. Pearson. *Physical Review* (New York, N. Y.), volume 74, 1948, page 232.

4. Crystal Rectifiers (book), H. C. Torrey, C. A. Whitmer. McGraw-Hill Book Company, New York, N. Y., 1948.

5. Development of Silicon Crystal Rectifiers, J. H. Scaff, R. S. Ohl. *Bell System Technical Journal* (New York, N. Y.), volume 26, 1947, pages 1–30.

6. The Double—Surface Transistor, J. N. Shive. *Physical Review* (New York, N.Y.), volume 75, 1949, page 69.

7. The Coaxial Transistor, Winston E. Koch, R. L. Wallace, Jr. *ELECTRICAL ENGINEERING*, March 1949, pages 222–3.

8. Equivalent Circuits of Linear Active 4-Terminal Networks, Lisa C. Peterson. *Proceedings*, Institute of Radio Engineers (New York, N. Y.), volume 27, 1948, page 593.

9. Space Charge, Signal, and Noise in Microwave Tetrodes, Lisa C. Peterson. *Proceedings*, Institute of Radio Engineers (New York, N. Y.), volume 35, 1947, page 1264.

THE COAXIAL TRANSISTOR

W. E. KOCK R. L. WALLACE, JR.

ELECTRICAL ENGINEERING VOL.68 pp.222-223 MARCH 1949

Produced outside the work performed under this contract

The Coaxial Transistor

WINSTON E. KOCK 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 are contained in this article.

THE TRANSISTOR is a semiconductor amplifier using two point contacts pressing against a small block of germanium. It has been described by J. Bardeen and W. H. Brattain.[1] When the two point contacts are placed close together on the surface, and proper bias potentials are applied, there is a mutual influence between one contact, called the "emitter," and the other, called the "collector," which makes it possible to use the device to amplify electric signals.

Shortly after the discovery of the transistor, J. N. Shive of the Bell Telephone Laboratories observed that amplification also could be obtained with a germanium wedge when the emitter and collector points were placed on opposite sides of the wedge.[2] In this construction the germanium wedge is narrowed down to a sharp edge and the point contacts are placed on opposite sides at a point where the wedge is only a few thousandths of an inch thick.

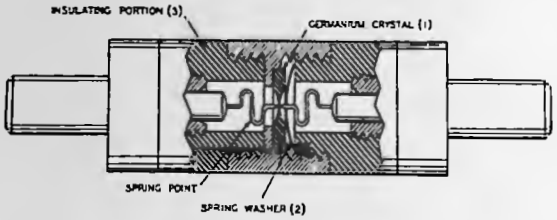


Figure 1. Cutaway view of first experimental coaxial transistor

The success of this unit led to the exploration of other forms of such amplifiers, one of which will be described here.

Investigation of the wedge device seemed to indicate that the current passing between emitter and collector actually was passing through the semiconductor and not around the wedge surface. Thus the effect was apparently not a surface phenomenon but rather a current amplification process occurring within the semiconductor itself. If this were true, transistor action also should be possible by reproducing the wedge geometry in circularly symmetrical form, thereby providing complete shielding between emitter and collector points. These considerations, when taken into account led to the development of the coaxial transistor.

CONSTRUCTION

A disk of germanium, one-eighth inch in diameter and 20 mils thick, was cut from a thin slab of germanium by means of a hole saw. Two dish-shaped depressions were ground and lapped into the faces of the disk by means of a spherical tool. The germanium wafer was placed in a mount shown in Figure 1 and held in place by spring pressure. In the cutaway view shown in Figure 1, (*1*) indicates the germanium disk, and (*2*) the spring washer, and the point contacts are seen to press against opposite sides of the disk. Most of the parts in this assembly are standard piece-parts used in the manufacture of single point rectifiers. For the experimental models, the insulating portions, (*3*) in Figure 1, were made of lucite for ease of fabrication. A photograph of this construction is shown in Figure 2 with the individual components spread out for observation.

The germanium disk, normally grounded electrically is seen to provide an electrostatic shield between the emitter and collector points, and all three parts—emitter point, collector point, and germanium disk—are seen to be coaxial.

PROCESSING PROCEDURE

It was found experimentally that high polish of the active surfaces of the germanium allowed the passage of higher collector currents before burnout occurred. Therefore, the spherical surfaces were polished with a very fine grade of diamond lapping compound and followed with an electropolish operation. Where maximum current capacity is not required, the usual transistor treatment of etching and later electrically forming the collector point by passing

Full text of conference paper, "The Coaxial Transistor," recommended by the AIEE technical program committee for presentation at the AIEE winter general meeting, New York, N. Y., January 31-February 4, 1949.

Winston E. Kock and R. L. Wallace, Jr. are with the Bell Telephone Laboratories, Inc., Murray Hill, N. J.

The authors wish to acknowledge the aid of their associates. They particularly are indebted to J. N. Shive and R. A. Ehrhardt.

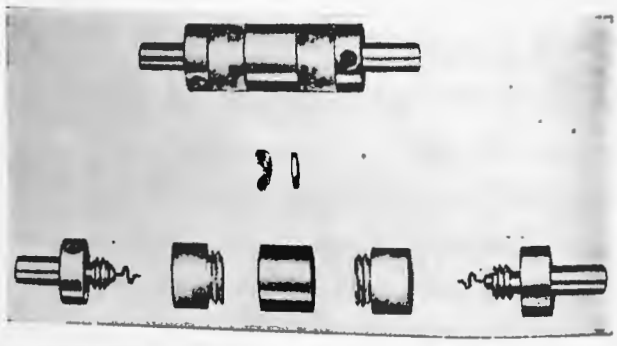


Figure 2. Exploded view of model in Figure 1

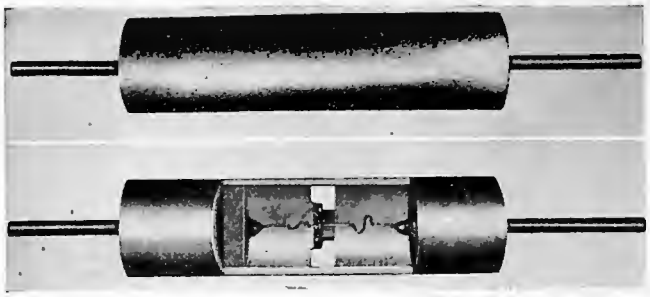


Figure 3. Cutaway view of proposed production model

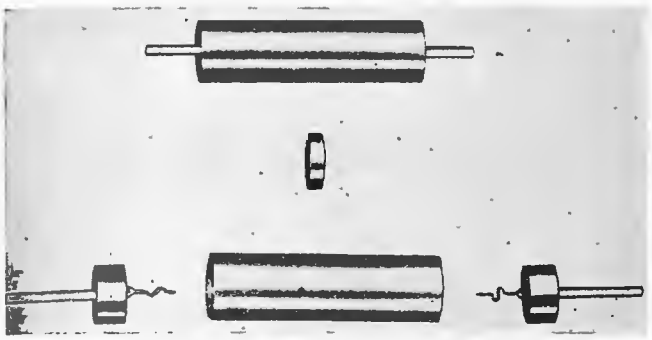


Figure 4. Exploded view of proposed production model

large currents in the reverse direction[1] also yields satisfactory results.

With the highly polished surfaces desirable for the higher current operation, it was found that points which were not perpendicular to the surface often had a sufficient tangential component of force to cause them to slip as they were pressed against the surface. In the coaxial construction described here, the two contact points are on opposite sides and thus can be exactly perpendicular to their contacting surfaces.

SMALL SCALE PRODUCTION MODEL

A construction which permitted moderate quantity production of these units is shown in Figures 3 and 4. Here a metal case was used to support the germanium wafer, and dielectric plugs carrying the emitter and collector points were forced into opposite ends of the cylindrical metal case until contact was made.

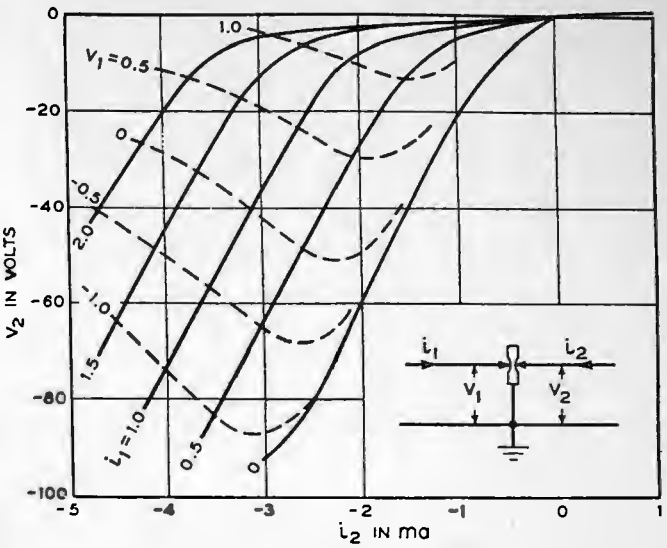


Figure 5. Static characteristics

ELECTRICAL CHARACTERISTICS

The d-c characteristics of a typical sample of the coaxial transistor shown in Figures 3 and 4 are given in Figure 5. The characteristics compare moderately well with those of the semiconductor triode having both points on the same side of the germanium.

ADVANTAGES

The advantages which might be attributed to the coaxial construction are improved stability of the points, especially on highly polished surfaces; electrostatic shielding between input and output circuits; and the avoidance of construction problems involving the placing of two spring contacts within a few thousandths of an inch of one another. Constructional tolerances are not completely avoided by this design, however, since the two points should be accurately in line on opposite sides of the germanium for most satisfactory operation.

REFERENCES

1. The Transistor, A Semiconductor Triode, J. Bardeen, W. H. Brattain. *Physical Review* (New York, N. Y.), volume 74, 1948, page 230.

2. The Double-Surface Transistor, J. N. Shive, *Physical Review* (New York, N. Y.), volume 75, 1949, page 69.

THE PHOTOTRANSISTOR

J. N. SHIVE

BELL LAB. RECORD VOL.XXVIII NO.8 pp.337-342 AUGUST 1950

Produced under contract W 36-039 sc-44497

The Phototransistor

Developments that result in the production of new devices often suggest studies not anticipated at the start of the original development. Such is the case with the Transistor.* Shortly after it was found that amplification could be obtained with the emitter and collector wires touching the same surface of the germanium block, comparable characteristics were obtained by placing these electrodes on opposite sides of a thin wedge of germanium only a few thousandths of an inch thick. The Coaxial Transistor† represents one embodiment utilizing this principle.

It was found that the exciting signal can be charge-produced in the semi-conductor by the absorption of light. Experiments have resulted in the production of a new photoconductivity cell, called the "Phototransistor." A photograph of several of these is shown in Figure 1.

This photocell differs from most such devices in its appearance, in its electrode arrangement, and in some features of its behavior. A longitudinal section drawing is shown in Figure 2. The heart of the device is a pill-shaped wafer of germanium having a spherical "dimple" ground in one side so that the thickness of the wafer at the center is about 0.003 in. The wafer is finally etched in the same way as in the manufacture of germanium diodes, to obtain clean surfaces. The wafer is secured at its periphery in a retaining ring which is force-fitted into one

*RECORD, *March*, 1949, *page* 89.
†RECORD, *April*, 1949, *page* 129.

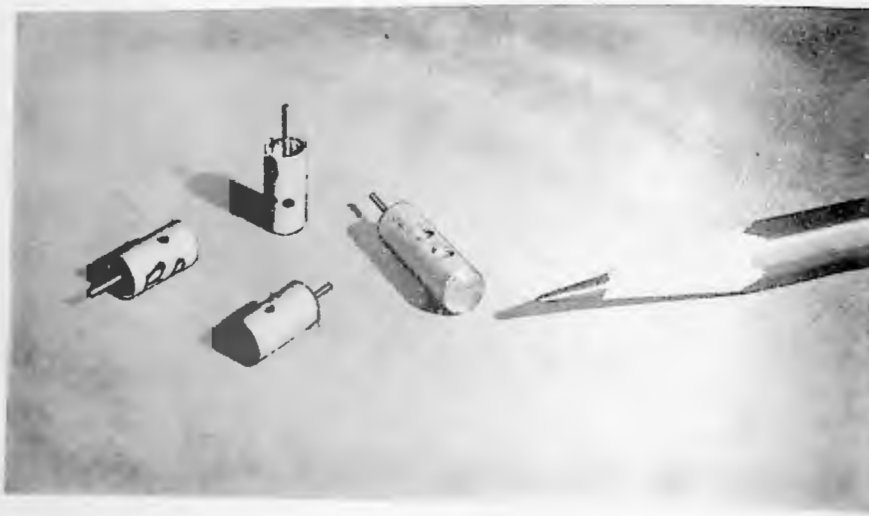


Fig. 1 – The Phototransistor has a high power output for a photoelectric device and gives good response to a rapidly fluctuating light source.

end of a metal cartridge. A pointed wire electrode of 0.005 in. phosphor bronze bears upon the germanium at the center of the dimple; this electrode is called the collector. The wire is fastened to a metal pin embedded in an insulating plug which is force fitted into the other end of the cartridge.

Operation of the cell depends upon the decrease in resistance of the germanium element between the peripheral contact and the collector when light is absorbed by the germanium. In a simple d-c circuit in which the Phototransistor is connected in series with a battery and load resistance, a small current will flow even when the photocell is in the dark. When light falls on the photocell, the current in the circuit increases in proportion to the amount of light falling on the cell. The current increment produced by the light is called the photocurrent.

One of the distinguishing features of the Phototransistor is that a substantial photocurrent response is obtained only when light

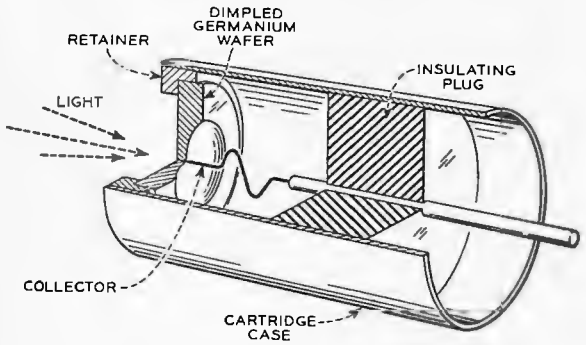


Fig. 2—Longitudinal section of the Phototransistor.

falls in the immediate neighborhood of the collector on the illuminated surface of the germanium. Figure 3 shows the photocurrent increment as a function of distance when a tiny spot of light is moved across the center of the responsive area of a typical Phototransistor. When this response curve is corrected for the diameter of the spot of light, the breadth of the sensitive area at half maximum is about 0.008 in.

It is apparent that allowing light to fall uniformly all over the germanium surface not only wastes a large fraction of the light, but avoids the realization of the spatial resolving power of which this device is inherently capable. Consequently, it is desirable to provide the cells with lenses to focus the available light into the responsive areas. One way of doing this is to fit a small lens into one end of the cartridge mounting —as shown in the cell adjacent to the pencil point in Figure 1. The cartridge in this instance is made longer than the others so as to accommodate the focal distance of the lens. The small size of the responsive area of the Phototransistor endows this device with a high spatial resolving power that can be obtained in other photoelectric cells only by a special masking arrangement or by some other optical means.

Figure 4 presents a family of collector current versus collector voltage curves for a number of different values of steady light flux. The experiment from which these curves were obtained was performed in such a way that the light fluxes indicated were focused substantially within the responsive area of the cell under test. The photocurrent is best observed and most profitably used when the dark current on which it is superposed is small rather than large. The dark current is made small by biasing the cell as a crystal diode in the reverse direction of its rectification characteristic. This corresponds, for n-type° germanium, to collector voltages negative with respect to the cell cartridge. The currents and voltages are therefore presented, in accordance with convention, in the third quadrant of rectangular coordinates. The particular curve designated "dark" will be recognized simply as the reverse characteristic of a germanium diode, while the other curves show the modification of this characteristic produced by light of different fluxes.

In a typical operating condition for use with time-varying light flux, the cell is connected in series with a load, which for good matching may be from 10,000 to 30,000 ohms. Performance of such a circuit can be deduced by analysis familiar in electronics. Referring to Figure 4, the dashed line shown as 20,000 ohms is called the "load line" for that value of series resistance. The inter-

°An n-type semi-conductor in contact with a metal electrode passes rectified current in the direction of easy flow when the semi-conductor is negative with respect to the metal electrode.

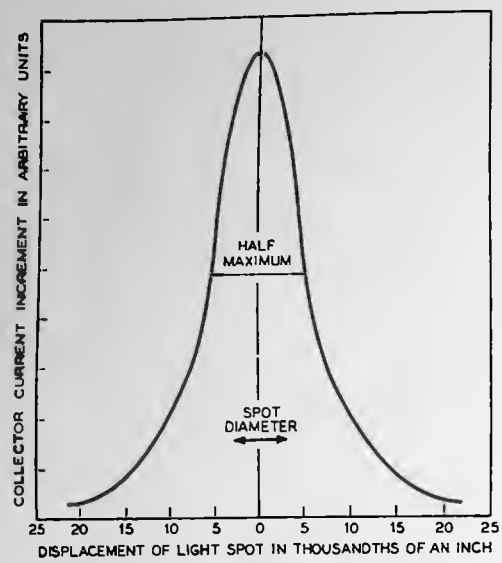


Fig. 3—Profile of the responsive area.

sections of the load line with the curves for the several values of light falling on the cell indicate the collector voltage and collector current for each light value. The collector voltage is the difference between the applied voltage and the voltage drop across the 20,000-ohm load. Calculations for this example yield a current output response of about 0.07 milliampere per millilumen. For a light fluctuation from dark to 20 millilumens, it can be shown that this is equivalent to an alternating current output of about 5 milliwatts.

In operation with light fluctuating at various different frequencies, the output of the cell is substantially flat up to at least 200 kilocycles per second, the highest frequency so far studied. This frequency range is amply wide for applications in the communication field.

Theory of the Photoconductivity Effect

The mechanism of photoconductivity, on which the operation of this cell depends, can be explained by the application of the concepts of modern physics. An atom of any substance consists of a positively charged nucleus and a number of electrons which revolve about the nucleus in orbits of various sizes and shapes. The electrons are prevented from flying away from the nucleus by the electrostatic attraction between the electrons and the nucleus. This attraction leads to a funnel-shaped potential profile, which is centered at the nucleus. This potential is sketched in Figure 5(a). Electronic energy is plotted vertically, while distance, right and left from the nucleus, is plotted horizontally. Another way of describing the situation of the electrons is to say that they are trapped down in the potential funnel, from which they cannot escape unless some external agency endows them with enough additional energy to climb up the sides of the funnel.

Electrons belonging to the various orbits of the atom reside on a number of discrete energy levels in the interior of the funnel, each energy level corresponding to a particular electron orbit. The energy levels are represented by the horizontal lines in Figure 5(a). Positions and spacings of the energy

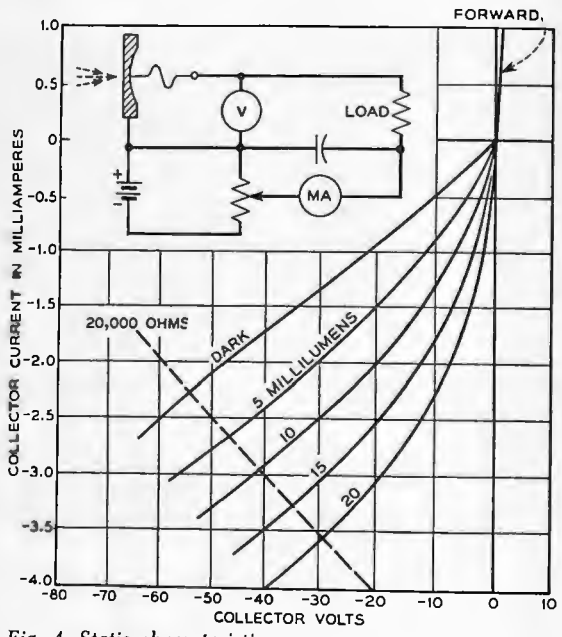


Fig. 4—Static characteristics.

levels are defined by physical laws, which provide also that each level may accommodate a specified small number of electrons, and no more. The orbital electrons belong-

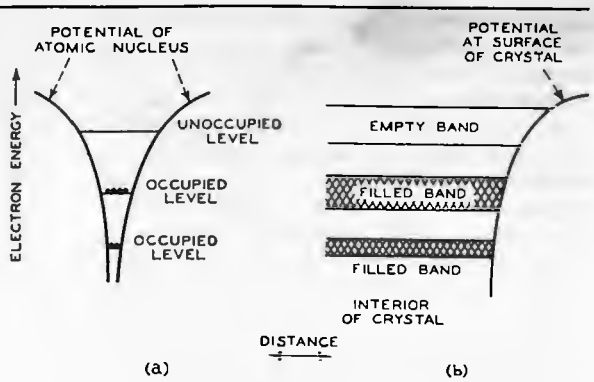


Fig 5—Energy levels for electrons—(a) in an isolated atom and (b) in a crystalline solid.

ing to the atom fill up these levels, starting with the lowest, which corresponds to the innermost orbit. When all the electrons have been apportioned among the lower-lying energy levels, there will be a number of higher-lying levels, of which only one is shown in the figure, which are normally unoccupied. For some elements, moreover, the uppermost occupied level may have fewer electrons than would be required to fill it completely. Although the dispositions and populations of the atomic energy levels differ from element to element, the same general scheme described above is common to all elements.

When the atoms of a particular substance coalesce to form a crystalline solid, the electronic energy levels, which in isolated atoms are discrete and very narrow, broaden into bands, sometimes of several electron volts' height from top to bottom. Figure 5(b) shows how the energy bands of a solid develop from the energy levels of isolated atoms. The lowest filled band in the solid corresponds to the lowest occupied level of the isolated atom, and so on. If there are N atoms in our specimen, each band can accommodate N times the number of electrons which the corresponding energy level can accommodate in the isolated atom.

In some materials the uppermost occupied band is completely filled with electrons; in other materials it is only partly filled. The motions of the electrons in an energy band are governed by the laws of quantum physics. When a band is completely filled, the electrons in that band cannot participate in the conduction of an electric current in response to an applied electric field. It is only when a band is partly filled or partly empty that its electron population may carry a net current. Evidently, then, insulators are substances whose occupied energy bands are completely filled with electrons, while good conductors are substances for which one or more of the bands in the energy level scheme are only partly filled. Semi-conductors are substances in which one or more of the bands are almost, but not quite completely filled, or else almost, but not quite completely empty.

If germanium could be obtained absolutely pure, its energy band scheme would be similar to that sketched in Figure 5(b). All of its occupied bands would be completely filled, all of its non-occupied bands would be completely empty, and the substance would be an insulator according to our definition. However, at ordinary temperatures the specimen contains enough thermal energy to transfer a few electrons from the uppermost filled band to the empty band. This thermal transfer produces a small partial filling of one band and a small partial emptying of the other, thus endowing the specimen with a small conductivity. In addition, all germanium specimens contain small amounts of various impurities which either contribute a few more electrons to the normally empty

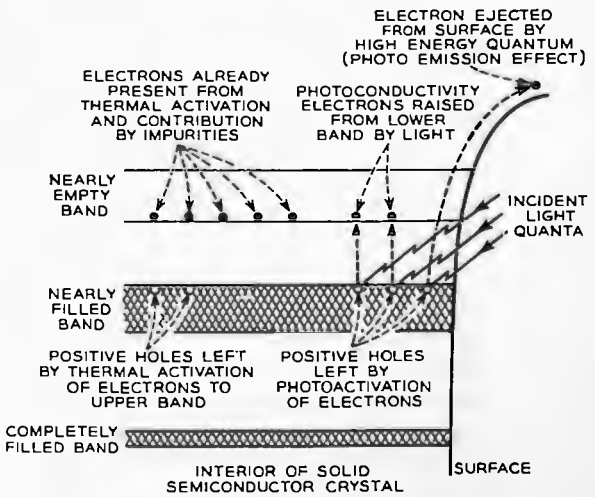


Fig. 6—Mechanism of the photoconductivity effect.

band or rob a few electrons from the normally filled band, thus further increasing the conductivity of the specimen. These concepts explain a fact well known in semiconductor technology, that the conductivity of a semiconductor depends markedly both on the temperature and on the impurity content of the material.

A semiconductor which owes its electrical properties predominantly to free electrons in an otherwise empty band is called an n-type semiconductor, while one which owes its electrical properties predominantly to vacant places in an otherwise filled band is called a p-type semiconductor. These vacant places are called "positive holes." They behave to all intents and purposes like particles having charge, mass, and mobility approximately equal to those of an electron, but with charge of positive sign. They move from atom to atom in the direction of an applied electric field, and by their motion they produce a current, just as do the free electrons by their movement in the nominally empty band.

The energy level picture for an n-type semiconductor is shown in Figure 6. The nearly filled band contains a few positive holes left by the thermal activation of electrons to the upper, nearly empty band, while the latter contains additional electrons contributed by the impurities which make the specimen n-type.

A light beam is a stream of radiant energy proceeding through space by the mechanism of wave motion. The energy of the beam resides in discrete energy packets, called quanta, which ride along with the waves. The energy of a single quantum is extremely small. In the visible light from an ordinary 100-watt lamp, and at a distance of three feet from it, there are about 20,000 quanta per cubic centimeter.

When light of the proper quantum energy is absorbed by a semiconductor, the quanta expend their energy in raising electrons from the filled band to the empty band, thus producing additional free electrons and positive holes. This mechanism is illustrated in Figure 6. The photocurrent increment in a semiconductor photoconductivity cell is due to the liberation of these extra charges and their subsequent movement in the applied electric field. When the light is turned off, recombination of the electrons with the holes quickly restores the conductivity of the sample to its original dark value.

With the recombination process continually taking place, even while the light is on, it is evident why the germanium Phototransistor should have a responsive area limited to the immediate neighborhood of the collector contact in the center of the germanium wafer. The distribution of bias current flow lines in the semiconductor is such as to concentrate the electric field near the collector. When electrons and positive holes are liberated here, the charges are separated and collected before recombination can occur. If, however, such pairs are produced near the periphery of the wafer where the field is weak, the charges effectively disappear by recombination before they can be separated and collected.

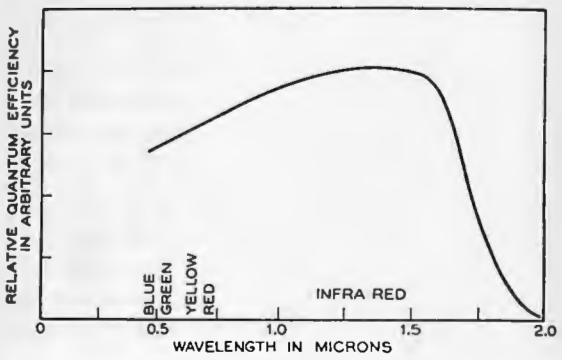


Fig. 7—Relative quantum efficiency versus spectral wave length for the Phototransistor.

In order to induce photoconductivity, the absorbed light quanta must deliver at least enough activation energy to raise the electrons across the energy gap between the top of the nominally filled band and the bottom of the nominally empty band. Since the energy of a light quantum varies inversely with the wave length of the light, it follows that the photoconductivity response depends upon the wavelength of the incident light. Figure 7 shows this dependence for the germanium Phototransistor.* The response in this case is presented in terms of

*These spectral response determinations were made in collaboration with Howard B. Briggs.

relative quantum efficiency, which is the photocurrent increment in the external circuit when a given number of quanta per second are incident upon the responsive area of the cell.

The response curve of Figure 7 has a long wave length limit in the vicinity of 2.0 microns. This wave length corresponds to a quantum energy of 0.61 electron volt, which represents the least energy that can create a free electron-hole pair in germanium. It may therefore be considered as the activation energy for the photoprocess, being equal to the energy gap between the top of the nearly filled band and the bottom of the nearly empty band. This value is in good agreement with the figure of 0.75 electron volt† obtained independently from measurements of the temperature variation of the specific resistance of germanium at high temperatures.

Referring again to Figure 6 one may deduce that if the quantum energy of the incident light is sufficiently great, electrons may be given enough energy to enable them to escape completely from the sample against the work function field at its surface. The photoemission effect comes about in just this way. For most substances, however, the quantum energies required correspond to wavelengths in the blue, violet, and ultra violet regions of the spectrum, and emission type photocells made from even the best of these materials are not suited to take the most effective advantage of light from incandescent filament sources, for example, from which the emission at these wavelengths is comparatively meager. Since the induction of photoconductivity requires less quantum energy than the induction of photoemission, it follows that, in general, photoconductivity cells are responsive to light of longer wavelengths than are photoemission devices. For the Phototransistor cell in particular, as can be seen from Figure 7, the sensitivity is greatest in the spectral region from 1.0 to 1.5 microns. This is the same region in which the quantum emission from ordinary incandescent light sources is greatest. The germanium cell is therefore a particularly suitable device for use with such sources of light.

†W. H. Brattain and J. Bardeen: *Physical Review*, Vol. 74, 1949, page 231.

AGING AND TEMPERATURE RESPONSE OF POINT-CONTACT TRANSISTORS

R. M. RYDER

Produced outside the work performed under this contract

AGING AND TEMPERATURE RESPONSE OF POINT-CONTACT TRANSISTORS

The reliability of transistors depends on their constancy irrespective of external conditions; in particular, they should not change too much with time, temperature, shock, or vibration. Information on such behavior follows.

When point-contact transistors similar to the M1729 are operated for some time under conditions representative of amplifier use, the chief variation is a slow decrease in the collector resistance r_c and the current amplification factor alpha. The accompanying figure 1 gives quantitative data which are the average of a number of units life tested up to about 7500 hours. There is an initial change of the order of 5% in the first few score hours, followed by a slow drift which persists at least a few thousand hours.

The effect of such a change on an amplifier using the units would be a gradual reduction in gain, the exact value of which would depend on the circuit parameters used. For order of magnitude, one may note that in a matched grounded-base circuit with negligible feedback, the power gain is $\alpha^2 r_c / 4 r_e$, so that to lose a db of gain from a circuit which was originally matched would require a reduction of either 11% in alpha or 20% in r_c. The changes in α and r_c are in the same direction (additive).

The temperature response is given in the accompanying Figure 2. For these transistors, above about 65°C the collector resistance r_c begins to drop rapidly, and the current amplification factor alpha begins to rise rapidly. For other transistors such as the 1729, the critical temperature where rapid variation sets in amy be as high as 80°C. It should be noted that these temperature effects are reversible, so that on returning to room temperature the original properties are restored. Also, the effects on amplifier gain are opposite in direction so that they tend to cancel; in other words, rise in temperature of the unit at first has rather small effect on the gain, tending principally to increase the collector current drawn at constant voltage and constant emitter current.

Since many practical operating environments are somewhat warm, this temperature sensitivity of transistors usually requires some attention from the circuit designer. Two features mitigate the temperature effects somewhat. Since the a-c gain is at first little affected, an a-c coupled amplifier is often superior in this respect to a d-c amplifier, especially if the power supply circuitry is properly arranged to hold a suitable operating point. Second, since transistors can operate with very low power drains one or more orders of magnitude below tubes, their self-heating effects may be much less than usual tube practice, even with greater miniaturization. Here again, however, for best results the system designer needs to pay attention to obtaining free flow of air around the transistor circuits.

To shock and vibration, transistors are very resistant. Shocks up to the order of 10,000 g do not damage the unit, while vibration up to at least 100 g produces no observable effect on the electrical characteristics. One would expect good behavior in these mechanical respects, because of the very low inertia of the transistor parts.

Similar data on junction units are not yet available. However, life data on junction diodes do not show the slow aging effects noted above, from which one might hope that triode life might be even longer than for point contacts.

In short, transistors are very reliable in the sense that their life is long and they can stand considerable amounts of shock and vibration. Their operating temperature should be held below 60° to 80° C, the precise value depending upon the particular type of unit.

EFFECT OF AGING

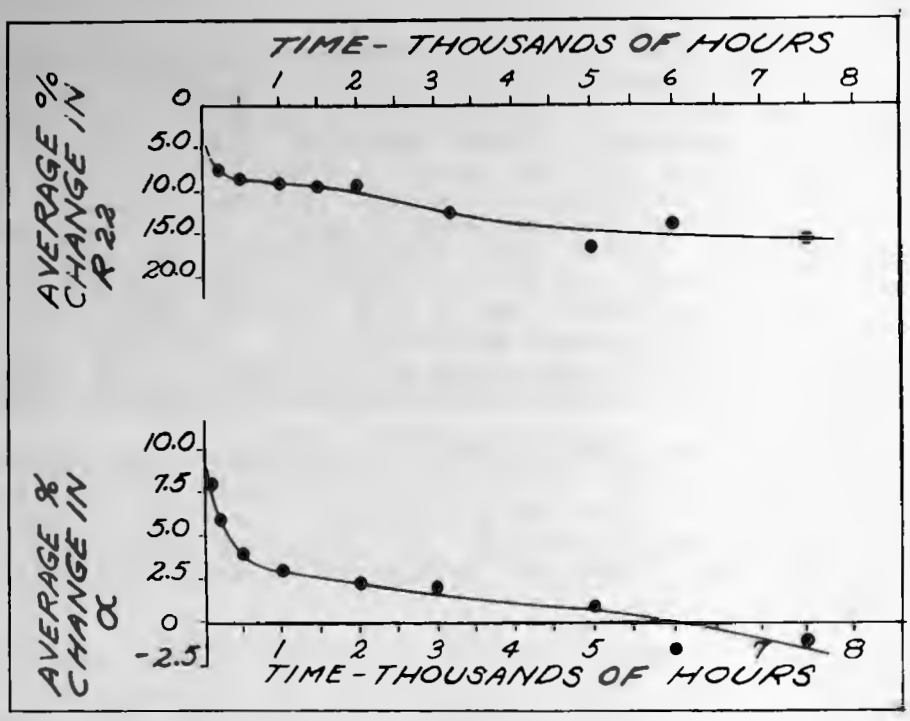


FIGURE 1

EFFECT OF TEMPERATURE

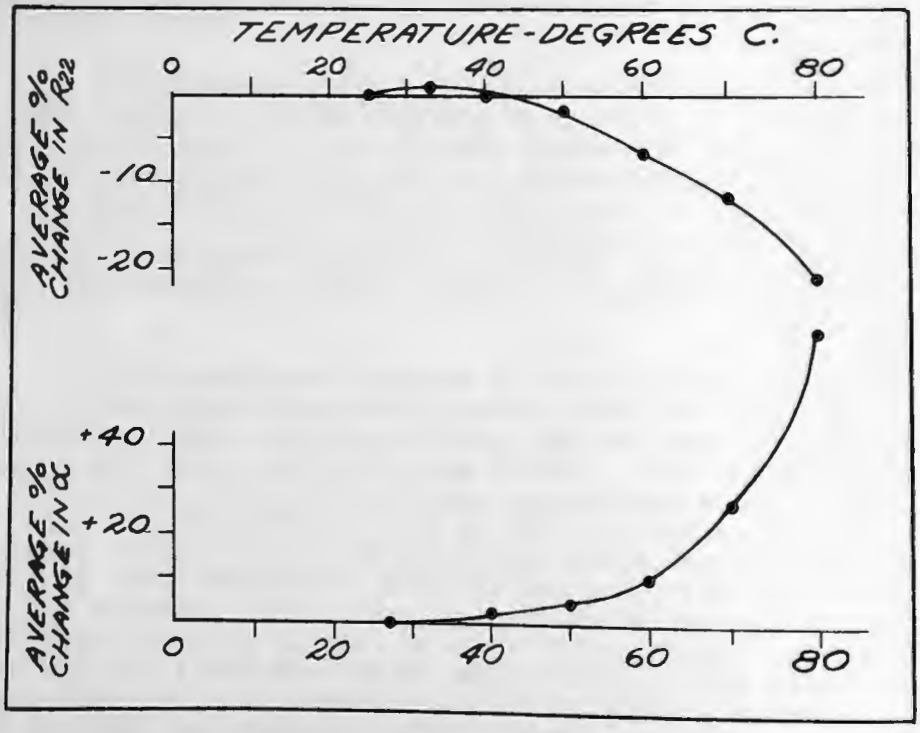


FIGURE 2

DUALITY AS A GUIDE IN TRANSISTOR CIRCUIT DESIGN

R. L. WALLACE, JR. G. RAISBECK

B.S.T.J. VOL.XXX pp.381-418 APRIL 1951

Produced outside the work performed under this contract.

Duality as a Guide in Transistor Circuit Design

By R. L. WALLACE, JR. and G. RAISBECK

(*Manuscript Received Sept. 26, 1950*)

Because of a relationship which exists between the properties of a vacuum tube triode and those of a transistor, it is possible to start with a known vacuum tube circuit and to transform it into a completely different circuit suitable for use with transistors. The nature of this transformation is discussed and a number of examples are given.

Introduction

SINCE the invention of the transistor there has been a natural tendency to compare its properties with those of a vacuum tube triode. This comparison indicates that the two devices are different in many important respects. For example, the grounded cathode vacuum tube is essentially a voltage amplifying device with a high input impedance and a relatively low output impedance, while the grounded base transistor is essentially a current amplifying device with a low input impedance and a relatively high output impedance. Furthermore, high gain vacuum tubes tend to be unstable with open circuit terminations, while high gain transistors tend, on the other hand, to be unstable with short circuit terminations.

The properties of the two devices are, in fact, so radically different that the development of the transistor has posed an entirely new set of circuit design problems. If the vacuum tubes in a known circuit are simply replaced by transistors (and appropriate changes are made in the supply voltages), it is usually found that the transistor is not used to best advantage and the circuit performance is not satisfactory. For this reason, circuit designers heretofore have exercised considerable ingenuity in devising new circuits which take into account the peculiarities of the transistor and use them to best advantage. It turns out that some of these circuits bear little resemblance to vacuum tube circuits designed to perform the same function.

Although there is a great difference between the electrical properties of transistors and vacuum tubes, there is a very simple approximate relationship between them. It is the purpose of this paper to show how it is possible, taking this relationship into account, to start with a known vacuum tube circuit and transform it into a completely different circuit suitable for use with transistors. Circuits derived in this way tend to take advantage of the peculiarities of the transistor, and in a number of cases have shown exceptionally good performance.

The Relation between Vacuum Tube and Transistor Properties

It is the purpose of this section to show that the properties of a transistor are related to those of a vacuum tube triode through an interchange of current and voltage, and that transistor currents behave like vacuum tube voltages and vice versa. The discussion is aimed particularly at the large-signal properties of the two devices and is restricted to the frequency range in which static characteristics are sufficient to determine circuit performance.

Consider first the grid-cathode input terminals of a triode as compared to the emitter-base input terminals of a transistor. With respect to these terminals each device behaves as a diode rectifier the properties of which are relatively unaffected by biases applied to the third electrode (plate or collector). The grid conducts when biased in the forward direction and fails to conduct when biased in the reverse direction. A similar statement can be made about the emitter. Furthermore, either device behaves as a low impedance when biased in the forward direction and as a relatively high impedance when biased in the reverse direction.

The difference between the emitter circuit and the grid circuit comes about in the following way: The vacuum tube is most effective as an amplifier when the grid is biased in the *reverse* direction, while the transistor is most effective when the emitter is biased in the *forward* direction. With respect to these input terminals, then, the essential difference between the two devices amounts to the difference between "forward" and "reverse". But this, in turn, amounts to an interchange of current and voltage.

Whatever qualitative statements can be made about emitter current and voltage can also be made about grid voltage and current, respectively. For example, the grid is normally given a moderate voltage bias at which the grid current is essentially zero, while the emitter is normally given a moderate current bias at which the emitter voltage is essentially zero. Furthermore, the principal non-linearity in the grid circuit occurs when the grid voltage is allowed to swing through zero with the result that grid current begins to rise, while the principal non-linearity in the emitter circuit occurs when the emitter current is allowed to swing through zero with the result that emitter voltage begins to increase.

The comparison between the plate-cathode output circuit of the triode and the collector-base output circuit of the transistor is somewhat complicated by the effects of grid and emitter biases. Consider first the situation in which zero bias is applied to the input circuits ($v_g = 0$ and $i_e = 0$). In this case, both the plate and the collector behave like diode rectifiers, conducting readily when biased in the forward direction and conducting relatively poorly when biased in the reverse direction. When input biases

are applied, however, the principal difference between the two devices becomes apparent and turns out again to be associated with the difference between forward and reverse. This is because biases applied to the grid affect only the forward part of the plate circuit characteristic while biases applied to the emitter affect only the reverse part of the collector circuit characteristic.

Thus the grid and plate are normally biased in the reverse and forward directions, respectively, with the result that the vacuum tube input impedance is high and the output impedance is relatively low. The emitter and collector, on the other hand, are normally biased in the forward and reverse directions, respectively, with the result that the transistor input impedance is low and the output impedance is relatively high.

The comparison of vacuum tube and transistor properties can be carried further with the help of Fig. 1(a) which shows the plate circuit characteristics of a particular vacuum tube triode and Fig. 1(b) which shows the collector circuit characteristics of a particular transistor. The axes in these two figures have been chosen to facilitate comparison of transistor currents with vacuum tube voltages and vice versa. The result is that the two families look quite similar. It is seen that the quantities to be compared are

r_p with $-i_c$,
i_p with $-v_c$,
$-v_g$ with i_e, and, though not shown,
$-i_g$ with v_e.

The consistent difference in sign between vacuum tube and transistor quantities holds only when the transistor is made from an N-type semiconductor. If the transistor is made of P-type material[1] there is no difference in sign between corresponding transistor and vacuum tube quantities.

By referring to Fig. 1(a) it can be seen that to a first approximation the effect of applying a negative voltage bias to the grid is simply to shift the plate circuit characteristic to the right along the v_p axis. The number of volts shift caused by a change of one volt on the grid is called the voltage amplification factor, μ, of the triode. Similarly, it can be seen from Fig. 1(b) that the principal effect of applying a positive current bias to the emitter is simply to shift the collector circuit characteristic to the right along the $-i_c$ acis. The number of milliamperes shift caused by a change in emitter current of one milliampere is called the current amplification factor, α, of the transistor. Thus, α of the transistor corresponds to μ of the vacuum tube.

[1] The p-Germanium Transistor. W. G. Pfann and J. H. Scaff, *Proc. I.R.E.*, 38, 1151.

It is interesting to note that the gross non-linearities in the vacuum tube plate circuit have their counterparts in the transistor collector circuit. For example, the counterpart of plate *current* cutoff is collector *voltage* cutoff.

The relationship between vacuum tubes and transistors is not only qualitative, but can be made quantitative as well provided a suitable vacuum tube is chosen for comparison with the transistor. The requirements are that the vacuum tube and transistor have similar dissipation ratings and that μ be equal to α. These conditions are roughly satisfied by the vacuum tube and transistor of Fig. 1(a) and Fig. 1(b). By comparing the axes in these two figures it may be seen that one milliampere in the transistor corresponds

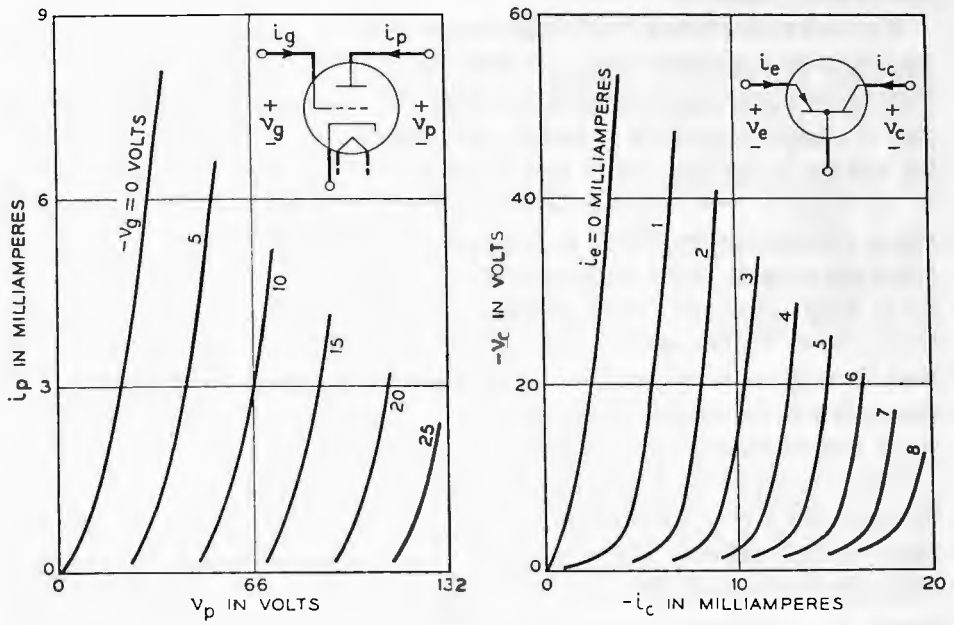


Fig. 1—The static characteristics of a transistor look like those of a vacuum tube triode provided transistor currents are compared with vacuum tube voltages and vice versa.

to 6.6 volts in the vacuum tube and vice versa. It follows that, in this case, transistor currents are related to vacuum tube voltages through a "transformation resistance," r, given by

$$r = \frac{6.6 \text{ volts}}{(10)^{-3} \text{ amps}} = 6{,}600 \text{ ohms.} \tag{1}$$

Circuit Considerations

The internal structure of a vacuum tube imposes a particular set of relationships between the vacuum tube currents and potentials. At low frequencies these relationships are given by the static characteristics which

show that over a fairly wide range of values the tube currents are roughly linear functions of the voltages. When the tube is connected into an external circuit, the circuit imposes a second set of algebraic relationships between vacuum tube currents and potentials and the performance of the circuit as a whole represents a simultaneous solution of these two sets of relationships. Now if the vacuum tube is replaced by a transistor and the external circuit is left unchanged, then the relationships internally imposed are markedly changed while the relationships imposed by the external circuit are left unaltered. Ordinarily this will lead to a completely different simultaneous solution for the two sets of conditions and hence to completely different circuit performance.

If circuit performance (with respect to the terminals of the tube or transistor) is to be maintained after substituting a transistor for a vacuum tube then the external circuit must be modified. One might suppose, for example, that it should be possible to find a new external circuit such that the collector voltage in the new circuit would behave exactly as did the plate voltage in the original circuit. To a certain extent this is possible, but this procedure meets with a serious difficulty. Although transistor voltages are fairly well behaved, roughly linear single-valued functions of transistor currents over fairly wide ranges of values, transistor currents are relatively more non-linear, often double valued, functions of the voltages. This means at once that if circuit performance is to be maintained for large signals, non-linear elements will be needed in the external circuit. This approach seems very much less promising than another to which we now come.

The new approach is to seek a transistor circuit in which every current behaves like a corresponding voltage in a known vacuum tube circuit and every voltage behaves like a corresponding current. This approach is relatively simple because, as has already been shown, half the problem is solved simply by exchanging transistor for vacuum tube. The remaining part of the problem is to find an external circuit which will impose the same relation between transistor potentials and currents as the original circuit imposed between vacuum tube currents and potentials. This amounts to saying that if the vacuum tube is to be replaced by a device in which the roles of currents and voltages are just interchanged then the external network should also be replaced by a new network which accomplishes this same interchange.

Networks in which this sort of interchange is accomplished are known as duals,[2] one of the other. It has been shown in the literature that it is possible to find and to realize physically the duals of most practical circuits. The total number of circuit elements in a network is ordinarily preserved when the

[2] Communication Networks, E. A. Guillemin, Vol. 2, pp. 246-254, John Wiley & Sons (1935); Network Analysis and Feedback Amplifier Design, H. W. Bode, p. 196, Van Nostrand, (1945). What Bode calls inverse we have called dual.

dual transformation is performed, each element being transformed into a new element which is its dual. The transformed elements are not, however, connected together in the same way as were the original ones. Elements in parallel are transformed into elements in series and vice versa. Nodes transform into loops and loops into nodes.

There are cases when finding the dual of a network is not as straightforward as the reader might infer from the above. Complications may arise when the network contains mutual inductance or non-linear elements, or if the network cannot be drawn on a flat surface without crossovers. Some of these questions are discussed by Bode.[2]

Duality

Table 1 shows side by side a number of network elements and the duals of these elements related through the transformation resistance r. The table also shows the duals of a few simple networks. It is the purpose of this section to show by means of examples how these dual relationships are established.

One network element is the dual of another provided the role of current in one is played by voltage in the other, and vice versa. Consider what this means in the case of a capacitance in which current and voltage are related by the equation

$$e = \frac{1}{jC\omega} i. \tag{2}$$

Interchanging the roles of current and voltage means replacing e in this equation by $i'r$ and replacing i by e'/r. The value of r determines how many volts across the condenser are to correspond to an ampere through its dual. Making the indicated substitutions gives

$$i' = \frac{1}{jr^2C\omega} e'. \tag{3}$$

This, however, is the kind of equation which relates the current through an inductance to the voltage across it. It is seen, therefore, that the dual of a capacitance C is an inductance of value given by

$$L' = r^2C. \tag{4}$$

In the dual transformation of a network every capacitance in the original network will be transformed in this way into an inductance in the dual network. Also, if e_c and i_c represent the voltage across a capacitance and

[2] Bode, loc. cit.

the current through it in the Kirchoff equations of the original network, these quantities will be replaced by $i_{L'}$, and $e_{L'}$, respectively, in the Kirchoff equations of the dual network. The quantities $i_{L'}$, and $e_{L'}$, represent the current through an inductance of value L' given by (4) and the voltage across it.

The argument just given can equally well be interpreted to mean that the dual of an inductance L is a capacitance C', the value of which is given by

$$C' = L/r^2, \tag{5}$$

so that every e_L and i_L in the Kirchoff equations of the original network are replaced by $i_{C'}$, and $e_{C'}$, respectively, in the Kirchoff equations of the dual network.

The dual of a resistance R is found in the same way. The equation applicable to a resistance is

$$e = Ri, \tag{6}$$

which, with the substitution of ri' for e and e'/r for i, becomes

$$i' = e'/(r^2/R). \tag{7}$$

Thus it is seen that a resistance R transforms into a resistance R' where

$$R' = r^2/R. \tag{8}$$

Also, e_R and i_R in the Kirchoff equations of the original network are replaced by $i_{R'}$, and $e_{R'}$, respectively, in the Kirchoff equations of the dual network.

The dual of a temperature sensitive resistance which changes value with changes in average signal level can be found by exchanging the labels on the axes of an *E-I* plot of its characteristic. This shows that the dual of a resistance which has a positive temperature coefficient, and hence increases in resistance with increase in signal level, is a resistance with a negative temperature coefficient of resistance which decreases in resistance with increase in signal level. Similarly, the dual of a short-circuit-stable negative resistance is an open-circuit-stable negative resistance.

The equations applicable to an ideal transformer of impedance ratio $1:\alpha^2$ are

$$\begin{aligned} e_2 &= \alpha e_1, \text{ and} \\ i_2 &= i_1/\alpha. \end{aligned} \tag{9}$$

Making the substitutions previously indicated leads to

$$\begin{aligned} i_2 &= \alpha i_1, \text{ and} \\ e_2 &= e_1/\alpha. \end{aligned} \tag{10}$$

TABLE I

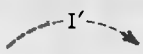
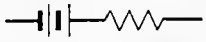
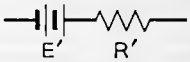
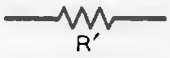
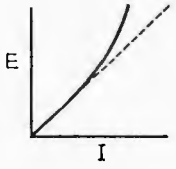
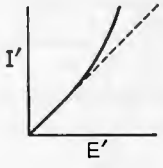
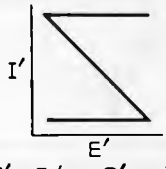
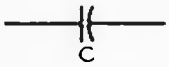
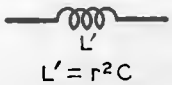
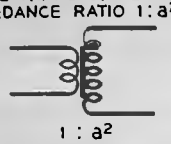
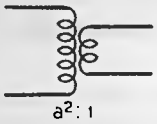
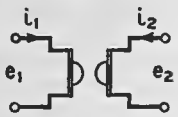
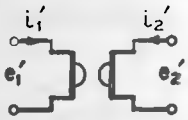
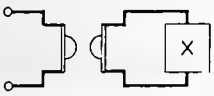
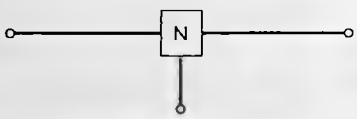
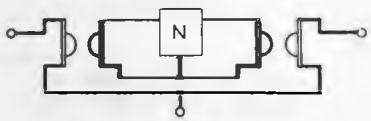
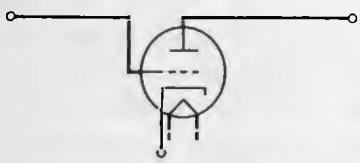
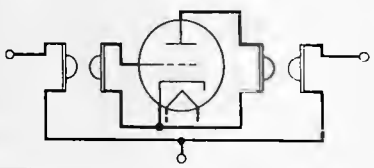
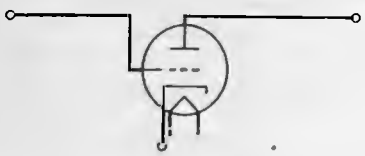
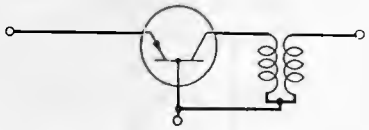
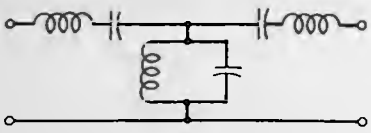
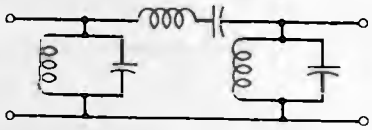
(1a) CONSTANT VOLTAGE SUPPLY 	(1b) CONSTANT CURRENT SUPPLY  $I' = E/r$
(2a) SERIES BATTERY AND RESISTANCE  	(2b) CONSTANT CURRENT SUPPLY AND RESISTANCE IN PARALLEL $I' = E/r, \quad R' = r^2/R$
(3a) SERIES BATTERY AND RESISTANCE 	(3b) SERIES BATTERY AND RESISTANCE  $E' = (r/R)E, \quad R' = r^2/R$ (EQUIVALENT TO (2b), BY THEVENIN'S THEOREM)
(4a) RESISTANCE 	(4b) RESISTANCE  $R' = r^2/R$
(5a) POWER-SENSITIVE RESISTANCE WITH POSITIVE TEMPERATURE COEFFICIENT  	(5b) POWER-SENSITIVE RESISTANCE WITH NEGATIVE TEMPERATURE COEFFICIENT  $I' = E/r, \quad E' = rI$
(6a) SHORT-CIRCUIT-STABLE NEGATIVE RESISTANCE 	(6b) OPEN-CIRCUIT-STABLE NEGATIVE RESISTANCE  $I' = E/r, \quad E' = rI$
(7a) CAPACITANCE  	(7b) INDUCTANCE  $L' = r^2C$

TABLE I—Continued

(8a) IDEAL TRANSFORMER OF IMPEDANCE RATIO $1:a^2$  $1 : a^2$	(8b) IDEAL TRANSFORMER OF IMPEDANCE RATIO $a^2:1$  $a^2:1$
(9a) IDEAL GYRATOR OF FORWARD TRANSFER IMPEDANCE R  $e_1 = -Ri_2$ $e_2 = Ri_1$	(9b) IDEAL GYRATOR OF FORWARD TRANSFER IMPEDANCE $-r^2/R$  $e_1' = (r^2/R)\,i_2'$ $e_2' = -(r^2/R)\,i_1'$
(10a) ANY TWO-TERMINAL NETWORK X 	(10b) THE SAME TWO-TERMINAL NETWORK, X, PLUS AN IDEAL GYRATOR OF TRANSFER IMPEDANCE r 
(11a) ANY THREE-TERMINAL NETWORK N 	(11b) THE SAME THREE-TERMINAL NETWORK, N, PLUS TWO IDEAL GYRATORS 
(12a) VACUUM TUBE TRIODE 	(12b) VACUUM TUBE TRIODE PLUS TWO GYRATORS 
(13a) SUITABLE VACUUM TUBE TRIODE 	(13b) TRANSISTOR PLUS IDEAL PHASE REVERSING TRANSFORMER  NOTE THAT (13b) AND (12b) ARE EQUIVALENT
(14a) ANY MID-SERIES TERMINATED CONSTANT-K FILTER SECTION OF DESIGN RESISTANCE R 	(14b) THE SAME CONSTANT-K FILTER SECTION MID-SHUNT TERMINATED BUT WITH DESIGN RESISTANCE CHANGED TO r^2/R 

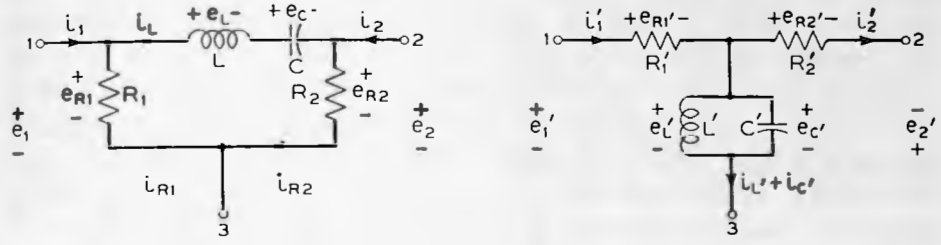
This indicates that the dual of an ideal transformer of impedance ratio $1:\alpha^2$ is another ideal transformer of impedance ratio $\alpha^2:1$. It follows that the dual of a 1:1 ideal transformer is a 1:1 ideal transformer.

The dual of a constant voltage supply E is, of course, a current supply which maintains a constant current equal to

$$I' = E/r, \tag{11}$$

and the dual of a constant current supply I is a supply of constant voltage equal to

$$E' = Ir. \tag{12}$$



$i_1 - i_L - i_{R1} = 0$
$i_2 + i_L - i_{R2} = 0$
$e_{R1} - e_C - e_L - e_{R2} = 0$
$e_1 = e_{R1},\ e_2 = e_{R2}$
$i_L = i_C$ (a)

$e_1' - e_{C'} - e_{R1'} = 0$
$e_2' + e_{C'} - e_{R2'} = 0$
$i_{R1'} - i_{L'} - i_{C'} - i_{R2'} = 0$
$i_1' = i_{R1'},\ i_2' = i_{R2'}$
$e_{C'} = e_{L'}$ (b)

Fig. 2—The dual of a network is found by transforming the Kirchoff equations.

The procedure of substitution used in all the examples above can be used in a straightforward way to find the dual of a more complicated network, but, in view of what has already been said some labor can be saved by writing the Kirchoff equations in the abbreviated notation used in Fig. 2. The equations on the left corresponding to the original network are then transformed into the equations of the dual network by making the following substitutions:

$$\begin{aligned} e_1 &= i_{1'} \\ i_1 &= e_{1'} \\ e_{R_1} &= i_{R_1'} \\ i_{R_1} &= e_{R_1'} \\ e_L &= i_{C'} \\ i_L &= e_{C'}, \text{ etc.} \end{aligned}$$

From these transformed equations, shown on the right hand side of Fig. 2, the dual network shown above them can be drawn by inspection.

From the example given in this figure, it is seen that a ladder network is transformed into another ladder network with each series branch of the original network being transformed into a shunt branch in the dual network and vice versa. Note also that a series combination of L and C is transformed into a shunt combination of C' and L'. The effect of a short circuit between terminals 1 and 2 in the original network (which makes $e_1 = e_2$) is an open circuit at terminal 3 in the dual network (which makes $i_1 = i_2$).

The Dual of an Ideal Vacuum Tube Triode

In a previous section it was shown that transistor currents behave approximately like vacuum tube voltages and vice versa. In view of what has been said about duality it might be assumed that, as three-terminal networks, the transistor and the vacuum tube triode are approximate duals. It is the purpose of this section to examine the relationships between the two devices in detail and to show that they fail to be duals one of the other principally because of a sign. What it amounts to is that signals transmitted through the dual of a vacuum tube suffer a phase reversal while, on the other hand, signals are transmitted through a transistor without change of phase.

A convenient way of proceeding is to start with the 4-pole equations of an ideal vacuum tube triode and transform them, by the methods already presented, into the equations of the dual. These transformed equations will then be compared with the 4-pole equations of a transistor.

The small signal behavior of a vacuum tube triode is represented by the equations,

$$\begin{aligned} i_g &= (0)v_g + (0)v_p\,, \\ i_p &= g_m v_g + k_p v_p \\ &= k_p(v_p + \mu v_g), \end{aligned} \tag{13}$$

where $\mu = g_m/k_p\,,$

and $k_p = 1/r_p\,.$

These equations apply when the positive directions of current and voltage are as indicated in Fig. 3. The equations corresponding to the dual of the ideal vacuum tube triode are found by substituting in equations (13),

$$\begin{aligned} i_g &= v_1/r, \\ i_p &= v_2/r, \\ v_g &= r i_1\,, \text{ and} \\ v_p &= r i_2\,. \end{aligned}$$

The quantities i_1 and v_1 will then represent the current and voltage at the input terminals of the dual device and i_2 and v_2 will represent the cur-

rent and voltage at the output terminals. Making these substitutions leads to

$$
\begin{aligned}
v_1 &= (0)i_1 + (0)i_2\,, \\
v_2 &= r^2 g_m i_1 + r^2 k_p i_2 \\
&= r^2 k_p (i_2 + \mu i_1).
\end{aligned} \tag{14}
$$

It remains to be seen how the directions of current and voltage must be assigned in Fig. 3. If the directions of v_1 and i_1 are arbitrarily assigned as indicated in the figure, then the directions of v_2 and i_2 can be found by an argument like that used in connection with the passive three-terminal network just discussed. The dual of making $v_p = v_g$ by placing a short circuit between plate and grid is making $i_1 = i_2$ by opening terminal 3 of the dual. This says that the positive direction of i_2 is as shown in Fig. 3. Similarly, the dual of making $i_g = -i_p$ by opening the cathode connection to the vacuum tube is making $v_1 = -v_2$ by placing a short circuit between terminals 1 and 2 of the dual. This requires that positive values of v_1 and v_2 have opposite signs when measured with respect to terminal 3 and so fixes the positive direction of v_2 as shown in Fig. 3.

The 4-pole equations for a transistor[3] are

$$
\begin{aligned}
v_e &= r_{11} i_e + r_{12} i_c\,, \\
v_c &= r_{21} i_e + r_{22} i_c \\
&= r_{22}(i_c + \alpha i_e),
\end{aligned} \tag{15}
$$

where

$$\alpha = r_{21}/r_{22}\,.$$

These equations are similar in form to equations (14) which correspond to the dual of an ideal vacuum tube triode. Comparing the two sets of equations shows that the following transistor and vacuum tube quantities correspond to each other:

$r^2 g_m$ and r_{21},
$r^2 k_p$ and r_{22}, and
μ and α.

Comparing the first of equations (14) with the first of equations (15) shows that the transistor quantities r_{11} and r_{12} should be zero if the transistor is to be an accurate dual of the vacuum tube triode. These quantities are small in present day transistors and there is hope that they may be made still smaller in the future. In the transistor of Fig. 1(b), for example, r_{11} is approximately 200 ohms. This corresponds to a grid-to-cathode leakage resistance in the triode which can be computed from

$$r_g = r^2/r_{11}\,.$$

[3] Some Circuit Aspects of the Transistor, R. M. Ryder and R. J. Kircher, *Bell Sys. Tech. Jl.*, *28*, 367 (July 1949).

Since $r = 6600$ ohms for the transistor and vacuum tube of Fig. 1, r_g amounts to 218,000 ohms. This is large compared to r_p and would not seriously impair the operation of the tube for many purposes.

What has been said indicates that transistor currents and voltages are fairly accurate duals of vacuum tube voltages and currents. As a three-terminal network, however, the transistor fails to be the dual of a vacuum tube because the values of i_c and v_c which behave as duals of v_p and i_p are measured with a convention of signs which is not consistent with Fig. 3. This can be seen by comparing the directions of i_2 and v_2 in the dual of a vacuum tube (Fig. 3) with the convention of signs for the transistor indicated in Fig. 1(b). A transistor like present day ones in all respects except for a reversal in sign of i_c and v_c would be a fairly good dual for a vacuum tube triode. This discrepancy in sign means, of course, that the grounded base transistor fails to give the phase reversal which would be given by the dual of a vacuum tube. This does not mean that the duals of vacuum tube circuits cannot be

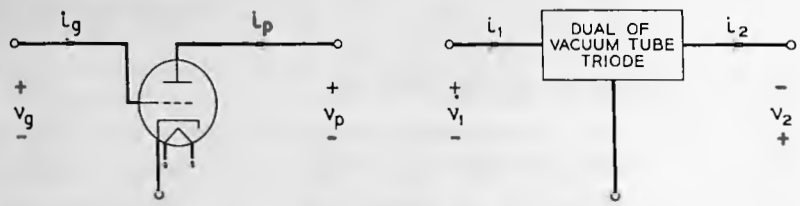


Fig. 3—The right-hand figure shows the convention of signs which must be used with the transformed equations (14).

found and used to advantage with transistors. It simply means that if the circuits are to be strictly dual an ideal transformer or some other means must be used to supply the phase reversal.

In finding the dual of a vacuum tube circuit there are several equally satisfactory ways of proceeding. Perhaps the simplest is to begin by treating the transistor as though it were a perfect dual of a vacuum tube triode. In this case, the transistor is substituted for the vacuum tube—emitter for grid, base for cathode, and collector for plate—and then the remaining part of the vacuum tube circuit is replaced by its dual. The resulting circuit fails to be a dual of the original only with respect to a phase reversal which can be corrected by inserting a phase reversing ideal transformer at the most convenient appropriate place.

Another procedure, which is perhaps more straightforward but which may also require more work, takes care of the phase reversal automatically. The first step in this case is to write down the Kirchoff equations for the vacuum tube circuit and then transform them into a new set of equations

in the manner illustrated in a previous section (see Fig. 2). In doing this, replace

$$v_p \text{ by } -i_c\,,$$
$$i_p \text{ by } -v_c\,,$$
$$v_g \text{ by } -i_e\,, \text{ and}$$
$$i_g \text{ by } -v_e\,.$$

This gives a set of Kirchoff equations which apply to the dual circuit and it remains only to find, by inspection, a circuit which satisfies them. It will often be found that, in order to satisfy these equations, it is necessary to introduce a phase reversing transformer. Several examples of this method of procedure will be given in sections to follow. In the appendix a great many more examples of vacuum tube circuits and their transistor duals are shown.

Gyrators and Duality

Tellegen[4] has shown that in principle it is possible to make a new kind of passive 4-pole circuit element to which he has given the name "ideal gyrator". This device is characterized by the 4-pole equations

$$e_1 = Ri_2 \text{ and}$$
$$e_2 = -Ri_1\,.$$

Though such a device is not known to have been realized in a practical physical form as yet, its properties are so closely related to duality as to be worth mentioning here.

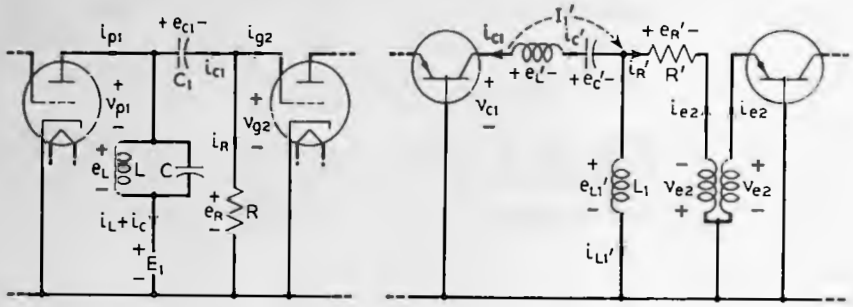
The following interesting properties can readily be deduced from the equations above. First, signals are transmitted through the device in one direction without phase reversal, while signals transmitted in the other direction are reversed in phase. Second, if an impedance Z is connected across the output terminals of an ideal gyrator, the impedance seen at the input terminals is R^2/Z. This means that the ideal gyrator has the property of transforming any two-terminal network into its dual. Also a three-terminal network can be converted into its dual by connecting one gyrator to the input terminals of the network and another to the output terminals. These gyrators must be so poled that no phase reversal is produced in either direction by the action of the two together.

This means that the dual of a vacuum tube triode can be obtained by using a triode plus two gyrators as shown in Table I and, of course, the dual of a transistor can be obtained by using a transistor plus two gyrators. Also,

[4] The Gyrator, A New Electric Network Element, B.D.H. Tellegen, *Phillips Research Reports*, 1948, pp. 81–101.

since the gyrator gives a phase reversal or not, depending on the direction of transmission, a transistor plus two gyrators can be made the *equivalent* of a vacuum tube triode by poling the gyrators in such a way as to take care of the phase reversal.

Ideal gyrators are not yet available but passive circuit elements having very similar properties over a narrow frequency range are available and are used in certain vacuum tube circuits. A quarter-wave line or its lumped-constant equivalent (which amounts to a full section of low- or high-pass filter) has the property of impedance inversion at a single frequency. Instead of giving zero or 180° phase change as does the ideal gyrator discussed above, this single frequency gyrator can be designed to give either $+90°$ or $-90°$ phase change. In either case, the phase shift is independent of the



$v_{p1} - e_L - E_1 = 0$

$E_1 + e_L - e_{c1} - e_R = 0$

$v_{g2} - e_R = 0$

$e_L = e_c$

$i_{p1} + i_{c1} + i_L + i_c = 0$

$i_{g2} + i_R - i_{c1} = 0$ (a)

$-i_{c1} - i_{c'} - I_1 = 0$

$I_1 + i_{c'} - i_{L1'} - i_{R'} = 0$

$-i_{e2} - i_R = 0$

$i_{c'} = i_{L'}$

$-v_{c1} + e_{L1'} + e_{c'} + e_{L'} = 0$

$-v_{e2} + e_{R'} - e_{L1'} = 0$ (b)

Fig. 4—A tuned amplifier stage and the transistor dual.

direction of transmission. This leads to the possibility of exchanging a transistor for a vacuum tube plus two quarter wave lines. An application of this sort will be discussed in a later section.

The Dual of an Amplifier with Shunt Tuned Interstage

Figure 4(a) shows a vacuum tube amplifier with the Kirchoff equations which apply to it. Figure 4(b) shows the transformed equations and a transistor circuit which satisfies them.

The ideal transformer in this circuit has no effect on any aspect of the circuit performance except the phase of the output and, if this is not important, the transformer may be left out. The curved arrows represent constant current supplies. All the element values in the transistor circuit may be

computed from the values in the vacuum tube circuit by means of the relations of Table 1.

Figure 5, (a) and (b), shows a different arrangement of power supplies in the vacuum tube circuit and the resulting more convenient arrangement of power supplies in the transistor circuit. In this case, the ideal transformer has been omitted but in all other respects the circuits are duals.

The application of duality in this case has led to the use of a series tuned circuit in series with the load instead of the shunt tuned circuit in shunt with the load, which the vacuum tube circuit might otherwise have suggested. The series tuned circuit has the advantage when used with short-circuit unstable transistors of insuring stability outside the pass band.

The Dual of a Push-pull Class B Amplifier

Figure 6(a) shows the circuit of a push-pull amplifier and the Kirchoff equations which apply to it. Figure 6(b) shows the transformed equations

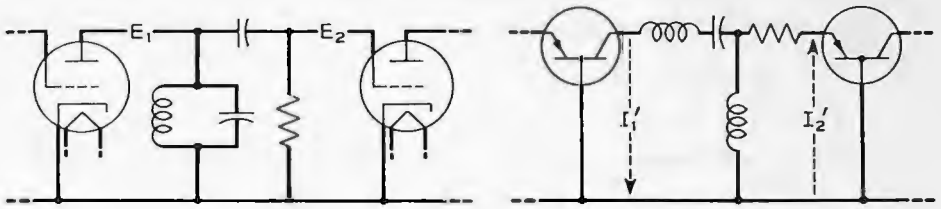
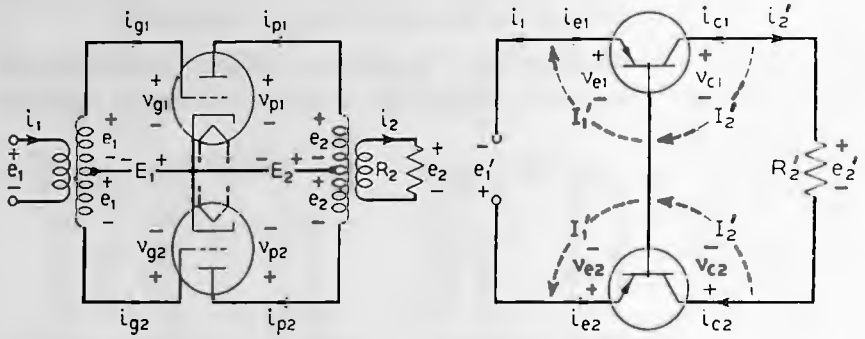


Fig. 5—An unconventional arrangement of power supplies in the vacuum tube circuit leads to a convenient arrangement for transistors. The phase reversing transformer which would make the transistor circuit strictly dual has been omitted.

and the dual transistor circuit. In this case, not only the circuit configuration but also the choice of operating point is important. For class B operation the two tubes are given a high negative grid bias, so that in the absence of an input signal the two plate currents are nearly zero while the plate voltages are quite high. In the transistor case, the dual situation is that the emitters are given a high positive emitter current bias so that in the absence of an input signal the two collector *voltages* are nearly zero while the collector *currents* are quite high. During a positive half cycle of input voltage the upper vacuum tube plate circuit begins to conduct and the plate current of the upper tube goes through a positive half cycle while the plate current in the lower tube remains essentially at zero. During this half cycle the plate current of the upper tube is coupled through the output transformer to the load while the lower tube contributes nothing, behaving simply as an open circuit element in shunt with the load and with the upper tube. The corresponding set of events in the transistor amplifier is that, in response to

a positive half cycle of input current, the collector voltage of the upper transistor goes through a negative half cycle while the collector voltage of the lower transistor remains essentially zero. All the collector voltage swing of the upper transistor is impressed directly on the load because, during this half cycle, the lower transistor serves as a short circuit element in series with the load and with the upper transistor. The next half cycle is, of course, like the first except that the lower tube and the lower transistor assume the active roles.

It was this circuit which first showed the great advantage which can sometimes be achieved through the use of duality. Using two type A transistors in the circuit of Fig. 6(b), it has been possible to obtain 400 milliwatts of



$$v_{g1} - e_1 + E_1 = 0$$
$$v_{g2} + e_1 + E_1 = 0$$
$$v_{p1} - e_2 - E_2 = 0$$
$$v_{p2} + e_2 - E_2 = 0$$
$$i_{g1} - i_{g2} - i_1 = 0$$
$$i_{p1} - i_{p2} + i_2 = 0 \qquad \text{(a)}$$

$$-i_{e1} - i_1' + I_1' = 0$$
$$-i_{e2} + i_1' + I_1' = 0$$
$$-i_{c1} - i_2' - I_2' = 0$$
$$-i_{c2} + i_2' - I_2' = 0$$
$$-v_{e1} + v_{e2} - e_1' = 0$$
$$-v_{c1} + v_{c2} + e_2' = 0 \qquad \text{(b)}$$

Fig. 6—A class B amplifier and the transistor dual.

audio output with a collector circuit efficiency of 60%. The same two transistors which gave this result could not be made to deliver more than 25 milliwatts output when used as grounded base amplifiers in a conventional circuit like that of Fig. 6(a).

The Dual of a Bridge Stabilized Oscillator

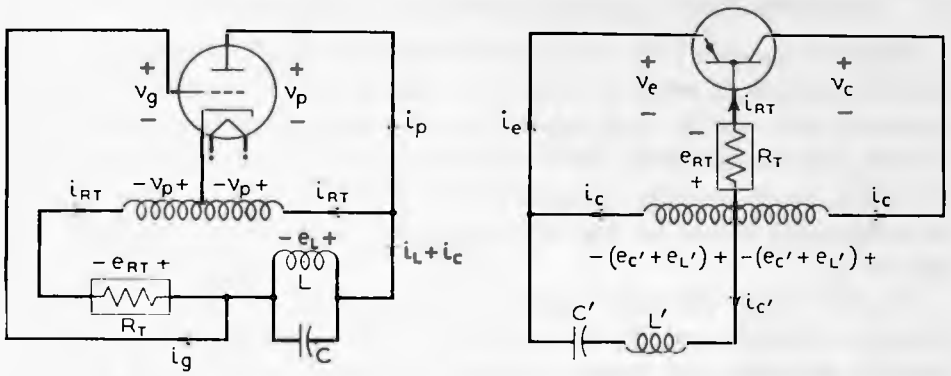
Figure 7(a) shows the circuit of a bridge stabilized oscillator due to Meacham,[5] in which amplitude stabilization can be achieved through the action of a temperature sensitive resistance R_T which has a positive tem-

[5] The Bridge-Stabilized Oscillator, L.A. Meacham, *Proc. I.R.E.*26, 1938, p. 1278; *Bell Sys. Tech. Jl.* 17, 1938, p. 574; *U. S. Pat.* 2,303,485.

perature coefficient of resistance. At the resonant frequency of the tuned circuit

$$\frac{v_0}{v_p} = \frac{\frac{R_T}{R_x} - 1}{\frac{R_T}{R_x} + 1}$$

where R_x is the equivalent resistance of the tuned circuit at resonance. The circuit values are chosen so that R_T is smaller than R_x and therefore the feedback is positive. As the amplitude of oscillation builds up, the increasing signal level across R_T increases its temperature thereby increasing its



$$v_g + v_p - e_{RT} = 0$$
$$2v_p - e_L - e_{RT} = 0$$
$$e_L = e_c$$
$$i_p + i_{RT} + i_L + i_c = 0$$
$$i_g + i_{RT} - i_L - i_c = 0 \qquad \text{(a)}$$

$$-i_e - i_c - i_{RT'} = 0$$
$$-2i_c - i_{c'} - i_{RT'} = 0$$
$$i_{c'} = i_{L'}$$
$$-v_c + e_{RT'} + e_{c'} + e_{L'} = 0$$
$$-v_e + e_{RT'} - e_{c'} - e_{L'} = 0 \qquad \text{(b)}$$

Fig. 7—A bridge-stabilized oscillator and the transistor dual.

resistance and bringing the bridge nearer to balance, so that the amount of positive feedback is reduced. This results in a stable amplitude of oscillation sufficient to make R_T slightly smaller than R_x.

Figure 7(b) shows the transformed equations and the dual transistor oscillator. In this case, $R_{T'}$ is a thermistor with a negative temperature coefficient of resistance. At the resonant frequency of the series tuned circuit

$$\frac{i_e}{i_c} = \frac{1 - \frac{R_{T'}}{R_{x'}}}{1 + \frac{R_{T'}}{R_{x'}}},$$

where $R_{x'}$ is the effective series resistance of the tuned circuit at resonance. The circuit values are so chosen that $R_{T'}$ is greater than $R_{x'}$ and therefore

the feedback is positive. As the amplitude of oscillation increases, $R_{T'}$ is heated so that its resistance decreases and brings the bridge more nearly into balance. This reduces the amount of positive feedback until a stable amplitude of oscillation is reached with $R_{T'}$ only a little greater than $R_{x'}$.

Meacham has shown that the stability of such an oscillator increases as the gain of the amplifier is increased. Since the vacuum tube amplifier of Fig. 7(a) can be made to give more gain than can be obtained from a single transistor, the transistor oscillator of Fig. 7(b) is not as stable as its vacuum tube dual. If increased stability is desired, it can be obtained by using a two-stage transistor amplifier instead of the single transistor shown.

Circuits Using Vacuum Tubes and Transistors Together

Since the vacuum tube and the transistor are basically different kinds of circuit elements, it seems reasonable to suppose that there may be circuits in which both can be used together to advantage. Two examples of such circuits will be discussed. The first has to do with a very ingenious high efficiency linear amplifier designed by Mr. W. H. Doherty.[6] This amplifier is particularly suited for use with amplitude modulated radio frequency inputs.

Figure 8 shows the basic features of one form of the Doherty amplifier. The networks N_1 and N_2 are impedance inverting networks of the type already discussed and amount to ideal gyrators for frequencies near the carrier frequency. Tube T_1 is biased nearly to cutoff and works, for small rf inputs, as a linear class B amplifier; while T_2 is biased well below cutoff and is inactive except when the rf input is higher in level than the unmodulated carrier. Downward swings of modulation are amplified by T_1 alone, which sees an effective load impedance just twice the value into which it could deliver maximum power. Under these conditions the peak voltage swing of T_1 begins to approach the supply voltage just as the rf input reaches a value corresponding to the unmodulated carrier. For greater input signals T_1, if acting alone, would begin to distort. But as the input signal is increased above the value corresponding to the unmodulated carrier, T_2 comes into action and contributes in two different ways to increasing the output signal linearly. First, T_2 acts as a class C amplifier and delivers power to the load and second, through the action of the impedance inverting network N_2, T_2 acts in such a way as to lower the effective load impedance seen by T_1. This makes it possible for T_1 to deliver more power to the load without an increase in plate voltage swing. The result of all this, which is discussed in much greater detail in Doherty's papers, is a linear amplifier of unusually high efficiency.

[6] A New High-Efficiency Power Amplifier for Modulated Waves, W. H. Doherty, *Proc. I.R.E.*, *24*, 1163 (September, 1936).

The part of the circuit of Fig. 8 shown inside the dotted box amounts to a vacuum tube plus two gyrators, which is just the dual of a vacuum tube. Apart from a phase shift of 180° this is *equivalent* to a transistor. This part of the circuit can therefore be replaced by a transistor plus a phase reversing transformer to obtain the basic transistor-vacuum-tube circuit of Fig. 9. This results in a considerable simplification because the impedance inverting networks are no longer needed.

The operation of the circuit of Fig. 9 is exactly similar to that of Fig. 8 except that the transistor operates as the dual of T_1. This means that the transistor is given a large forward emitter bias so that collector voltage is

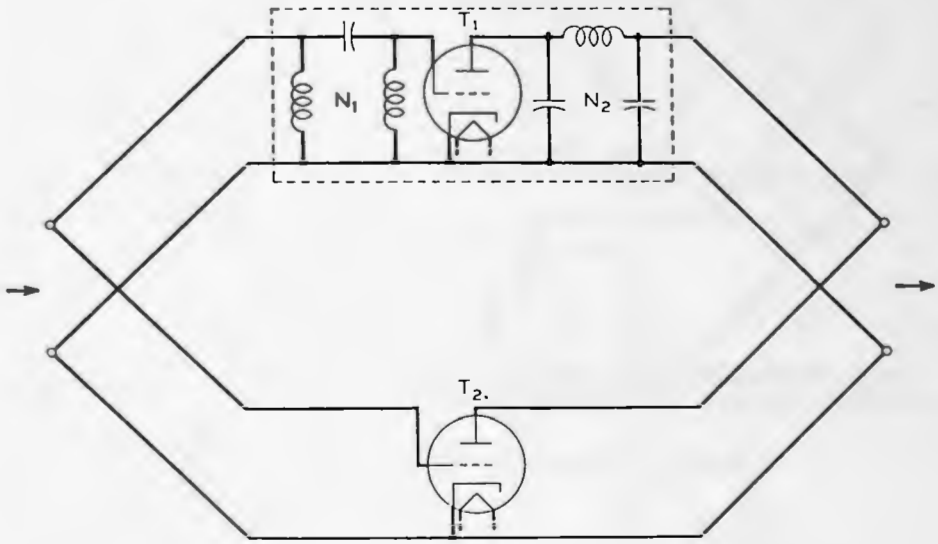


Fig. 8—The basic arrangement of a Doherty amplifier.

almost cut off. Under these circumstances, it is capable of operation as a linear amplifier. The load resistance is just half that into which the transistor could deliver maximum power. The transistor acts alone to amplify downward swings of modulation (T_2 being biased well below cutoff as before) but as the input signal exceeds that of the unmodulated carrier the collector current swing begins to approach the maximum value permitted by the (current) supply and T_2 begins to contribute to the output in just the ways it did in the circuit of Fig. 8. First, it acts as a class C amplifier delivering power directly to the load and second, it behaves as a negative resistance bridged across the load and thereby increases the impedance into which the transistor works. This permits the transistor to deliver more power without increasing the collector current swing.

Just as the basic Doherty circuit of Fig. 8 needs tank circuits to suppress

carrier harmonics, so also does the circuit of Fig. 9. When these are added the circuit of Fig. 10 is obtained.

Doherty shows that there are two basic circuit arrangements for obtaining the high efficiency linear amplifier action which he describes. One of them has been discussed above. By starting with Doherty's other arrangement,

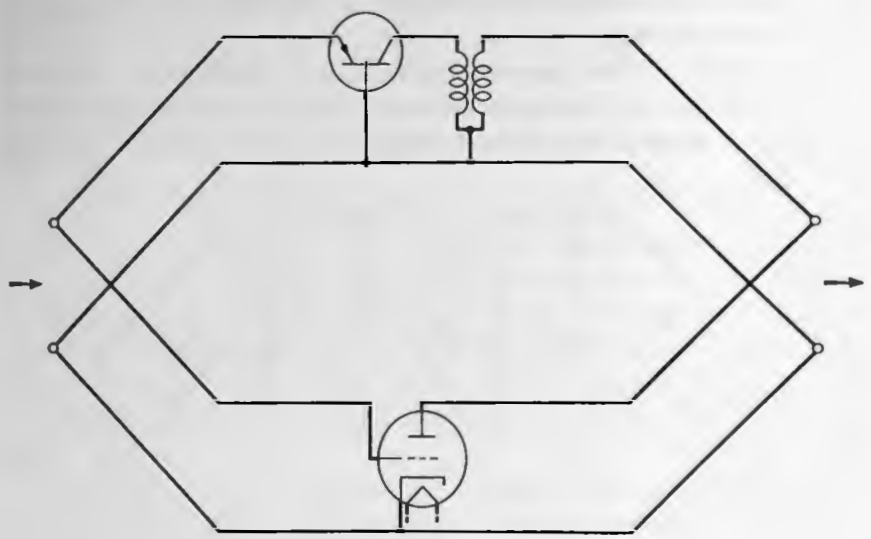


Fig. 9—The basic circuit of a Doherty-type amplifier using a transistor to replace a vacuum tube and two impedance inverting networks.

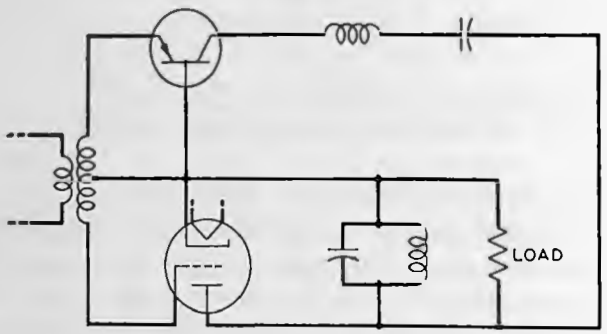


Fig. 10—A Doherty-type amplifier in which low level signals are amplified by the transistor alone.

one arrives at the circuit of Fig. 11. In this case, it is the vacuum tube which operates class *B* and the transistor which helps to supply the peaks by class *C* operation. At low input levels the transistor behaves as a short circuit and the vacuum tube works into an impedance just twice the value into which it can deliver maximum power. As the input signal increases above the carrier level the transistor begins to operate, contributing in two ways

to increasing the power output. First, it delivers power directly to the load and, secondly, it behaves as a negative resistance in series with the load, thereby decreasing the impedance into which the vacuum tube works and permitting it to deliver more power without increasing its plate voltage swing.

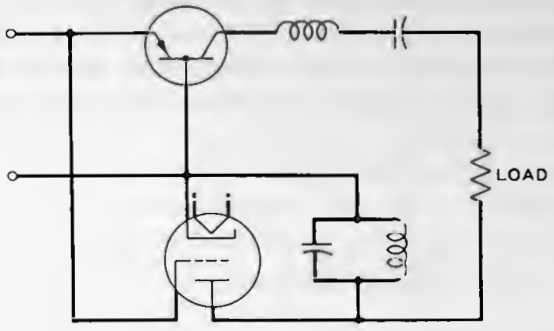


Fig. 11—A Doherty-type amplifier in which low level signals are amplified by the vacuum tube alone.

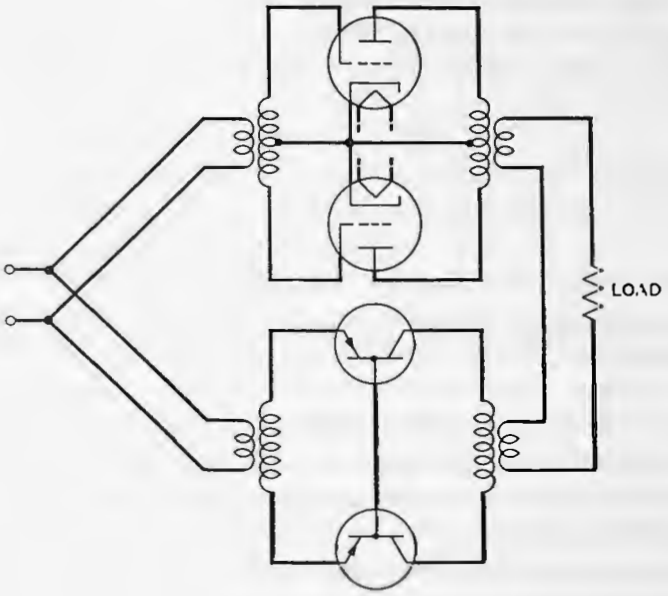


Fig. 12—A high efficiency untuned amplifier in which small signals are amplified by the vacuum tubes alone.

The Doherty amplifier is limited to narrow band operation only because the networks N_1 and N_2 will behave as gyrators only over a narrow range of frequencies. Apart from a phase change of 180°, however, the transistor behaves as a vacuum tube plus two *ideal* gyrators and is therefore capable of acting as the dual of a vacuum tube over a wide band of frequencies. This

leads to the possibility of an entirely new, wide band, high efficiency amplifier which operates on the same principles as the Doherty circuit.

Both the transistor part and the vacuum tube part of the amplifier must be made push-pull in order that both halves of the input wave be amplified equally. In the circuit of Fig. 12 the vacuum tubes are biased for class *B* operation, while the transistors are given a large forward emitter current bias so that they are operated well below collector voltage cutoff. For small input signals the transistors are inactive, serving simply as short circuit elements in series with the load. As the input signal reaches half the peak

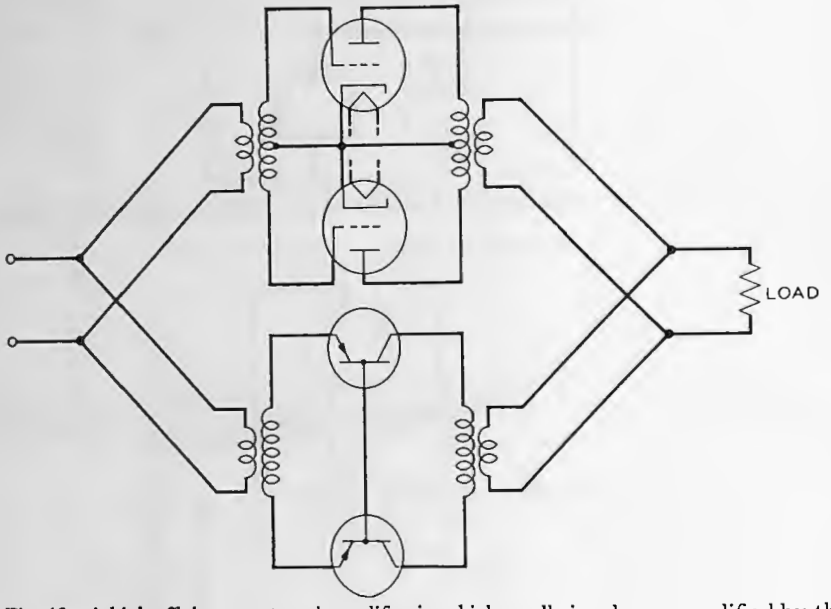


Fig. 13—A high efficiency untuned amplifier in which small signals are amplified by the transistors alone.

permissible value the vacuum tubes begin to distort because their voltage swings approach the supply voltage. At this point the transistors begin to operate in two separate ways, just as in the Doherty amplifier. First, they work as class *C* amplifiers delivering power directly to the load and second they behave as a negative resistance in series with the load thereby serving to reduce the impedance into which the vacuum tubes work. This permits the vacuum tubes to deliver more power without increasing their plate voltage swing.

Just as there are two forms of the Doherty amplifier, there are also two forms of this wide band arrangement. In the second form shown in Fig. 13 the transistors are biased for class *B* operation (near collector voltage cutoff) while the vacuum tubes are biased well below cutoff. For small signals the transistors act alone as class *B* amplifiers and the vacuum tubes act simply

as open circuit elements in shunt with the load. As the instantaneous input signal reaches half the permissible peak value, the transistors begin to distort because the collector current swing begins to approach the value of the (current) supply. At this point the vacuum tubes begin to operate in two separate ways to increase the power output. First they act as class *C* amplifiers delivering power directly to the load and second, they behave as negative resistance elements in shunt with the load and thereby increase the impedance into which the transistors work. This permits the transistors to deliver more power without increased collector current swing.

The circuits of Figs. 12 and 13 can both be adjusted to give reasonably linear performance. Perhaps the most interesting aspect of these circuits is that the theoretical maximum efficiency (for sinusoidal signals) is 93%. This should be a matter of importance in applications where the greatest possible power output is desired from transistors and tubes of limited dissipation rating.

It has been pointed out by Ryder and Kircher[3] that a transistor with α just equal to unity behaves like a vacuum tube triode when operated with the emitter grounded. If transistors can be made to operate satisfactorily in this way with large signal swings then all the vacuum tubes in the circuits discussed in this section can be replaced by grounded-emitter transistors.

General Comments

It is obvious that not all useful transistor circuits can be found in the manner presented in this paper and, furthermore, not all of the circuits found through the application of duality are useful.

One limitation of the method is imposed by the fact that present day transistors correspond to rather low μ vacuum tubes. On this account, vacuum tube circuits which require high μ tubes for satisfactory performance will lead to inferior transistor circuits. If further development of the transistor produces higher values of α, this limitation will be reduced.

Another limitation of the method comes from the failure of the transistor to produce a phase reversal. Although this is not important in many cases, and in other cases in which it is important a transformer provides a satisfactory solution, still the fact remains that transformers do not respond at d.c. and because of this fact some transistor dual circuits are useless.

In spite of these limitations, the methods presented in this paper have led to a number of useful transistor circuits and may be expected to yield still more in the future.

Acknowledgment

The authors wish to express their gratitude for the keen interest in this work shown by Mr. W. E. Kock and Mr. R. K. Potter under whose direction

[3] Ryder and Kircher, loc. cit.

it has been carried out. We are indebted to Mr. B. McMillan and also to Messrs. Harold Barney, R. J. Kircher, L. A. Meacham, J. A. Morton, L. C. Peterson and R. M. Ryder for encouragement and helpful comments.

APPENDIX

The appendix is a brief account of some of the circuits which have been investigated with the aid of the methods described in the main text. The circuits shown do not exhaust all possibilities, and the specific configurations shown are not to be regarded as optimal choices.

Figure 14 shows a one-stage resistance-capacitance coupled amplifier and its dual. In the input circuit of 14(a), C is a series element which passes alternating currents and blocks direct currents. The dual element, L, in the input circuit of 14(b) is a shunt element which is a short circuit to direct voltages, but not to alternating voltages. The resistance R_1 is a shunt element which provides a path through the battery without creating a short circuit to ground for the alternating signal. Correspondingly, the resistance R_1' provides a path around the current supply, which otherwise would be an open circuit for the alternating signal.

The passive elements in the input circuit are also capable of acting as a source of self-bias. Suppose, for example, that R_1 be connected directly to ground with no battery interposed and that the emitter current supply be removed. The resulting vacuum tube circuit is familiar. The usual explanation of its behavior is that when the grid is driven positive and draws grid current, the condenser C becomes charged, and that subsequently the condenser discharges through the resistance R_1, supposed large enough to assure a long discharge time constant. In this way the condenser is kept charged so that grid current flows only a small portion of the time.

The behavior of the dual circuit is exactly analogous, but is much harder to explain simply because words and expressions dual to those used above do not exist or are not in current use. For example, we speak of a condenser as "charged" when there is a potential between the terminals. There is no corresponding term for an inductor with a current passing through it. The explanation, nevertheless, might be as follows: The emitter normally presents a low impedance to positive input currents, and a high impedance to negative input currents. When the input current is negative, the high impedance of the emitter blocks the current and a current is therefore drawn through the inductor L, in an upward direction as the figure is drawn. Subsequently, when the input current becomes positive, the emitter presents a low impedance, and the current in the inductor is free to pass through R_1' and the emitter. It is supposed that the inductance is large enough so that the decay time constant of the inductor through R_1' and the emitter is large. Then the current through the inductor will be approximately constant over a short period and will be a bias current. This current will regulate itself

so that the resultant emitter current is positive most of the time, becoming negative only long enough to keep the inductor "charged."

The output circuit is easier to explain. The resistance R_2 provides a path through the battery which is not a short circuit. The resistance R_2' provides a path around the collector current supply which does not have infinite impedance to the signal. The loads Z_L and Z_L' are the ultimate receivers of the amplified signal. The duality of the loads may be emphasized by pointing out that the condition corresponding to $Z_L = \infty$ is $Z_L' = 0$.

If the circuits are analyzed with the aid of the equivalent circuits discussed in the text, the voltage amplification of the vacuum tube circuit will be found to be

$$\frac{g_m}{\frac{1}{r_p} + \frac{1}{R_2} + \frac{1}{Z_L}},$$

while the current amplification of the transistor circuit is found to be

$$\frac{r_m}{r_c + R_2' + Z_L'}.$$

These expressions are obviously duals

Transistor amplifiers like the one shown in Fig. 14 can be connected in cascade. Three examples are shown in Fig. 15 (b) and (c) and (d). Figure 15(b) is the most obvious connection, and 15(c) and 15(d) have provisions for correcting the relative phase inversion that occurs in the transistor circuit. If the circuit equations for the three examples are written out, it will be discovered that only 15(c) and 15(d) are duals (in the sense defined in the text) of the vacuum tube circuit 15(a). The remaining example, 15(b), is the dual of a peculiar looking circuit with one vacuum tube inverted.

In the range where operation is nearly linear, the three cascaded amplifiers behave much alike; and 15(c) and 15(d) can be regarded as pedantic attempts to make the signs come out "right." As soon as non-linear operation is encountered, however, the differences between the circuits become pronounced. This will be clearer when multivibrators are considered.

Figure 16 shows a variation of one of the circuits of Fig. 15 designed to operate on a single power supply. A circuit like this with four cascaded stages has been built and tested, and was found to work satisfactorily with selected matched transistors. The two extra resistors in each stage, R_1 and R_2, are voltage dropping resistors, chosen to balance the voltage drops in the emitter and collector circuits respectively.

Figure 17 shows a multivibrator, conveniently illustrated as a two-stage RC coupled amplifier with its output connected to its input. Below are shown three circuits, of which the first is almost a dual and the other two are duals

of the vacuum tube circuit. The first transistor circuit, 17(b), fails to be a dual in that besides having positive feedback around two stages, it has positive feedback in each stage separately. This is avoided in 17(c) by an isolating

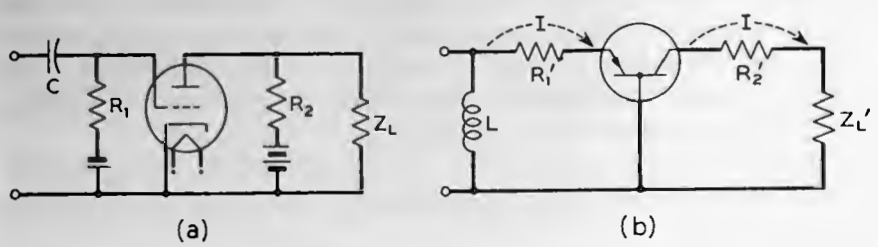


Fig. 14—Resistance—capacitance coupled amplifier and dual.

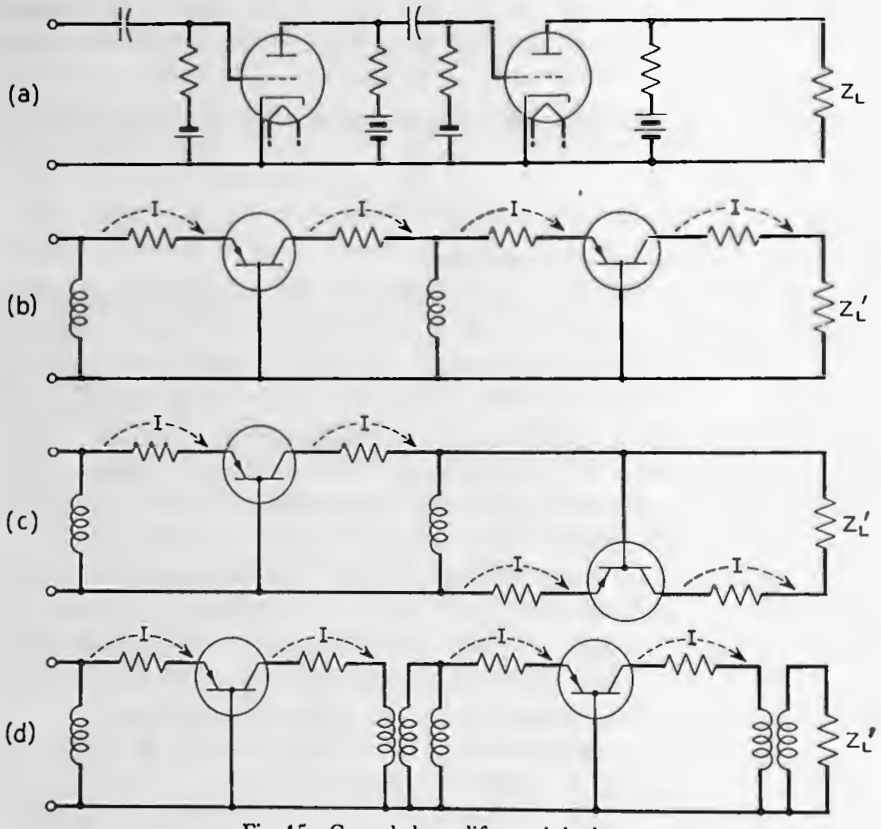


Fig. 15—Cascaded amplifier and duals.

transformer. It has been shown nevertheless by practical tests that 17(b) acts as a multivibrator, and is perhaps even a better circuit than the others. It has the interesting characteristic that the two inductors are in parallel, and hence may be replaced by a single inductor.

One explanation of the operation of a vacuum tube multivibrator is this.

Suppose one grid is at a large negative potential, cutting off that tube, and the other is at a positive potential or at zero potential. The potential of the

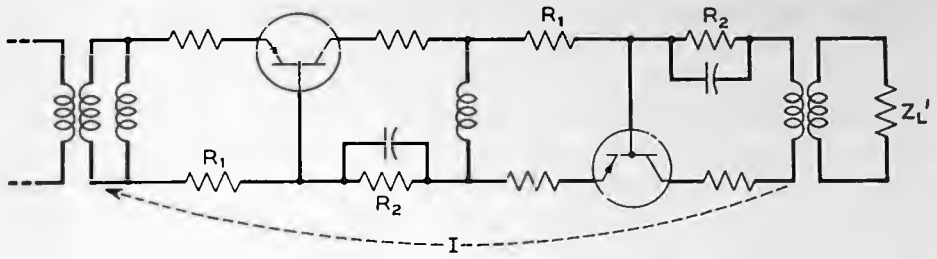


Fig. 16—Two-stage transistor amplifier using a single power supply.

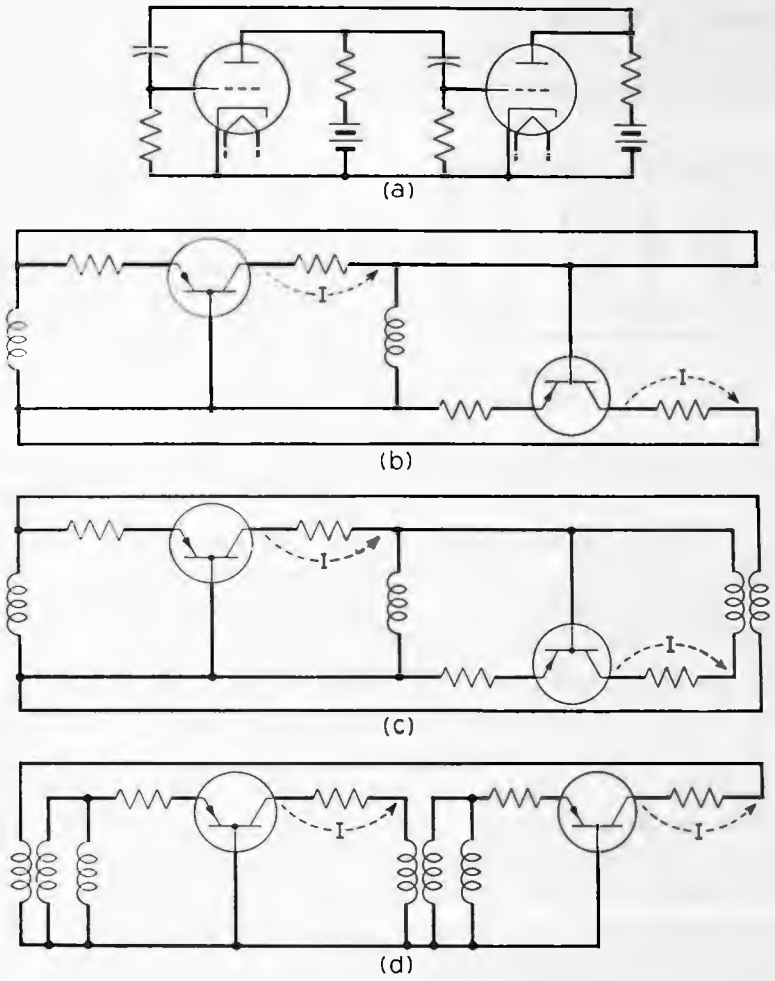


Fig. 17—Multivibrator and duals.

negative grid rises toward zero at a rate controlled by the grid resistor and the coupling condenser. When the grid potential rises above the cutoff

voltage, the plate potential falls, and the positive feedback accelerates the process so the first grid rises to a positive potential and the other grid falls to a large negative potential. The process is then repeated.

The dual behavior of the transistor multivibrator is as follows: Suppose that the emitter current of one transistor is very large, and of the other about zero. The large emitter current passes through the emitter, an inductance, and a small resistor, and will decay at a rate controlled by the inductance and the effective resistance of the resistor and emitter. As it decays, no effect will result in the collector circuit until the emitter current falls below the collector voltage cutoff point, after which the collector current will decrease. The emitter current of the other transistor will *increase*, as a consequence of the phase inversion built into the circuit, and the collector current of the second transistor will increase. As a consequence of positive feedback, the whole process will accelerate suddenly and proceed until the

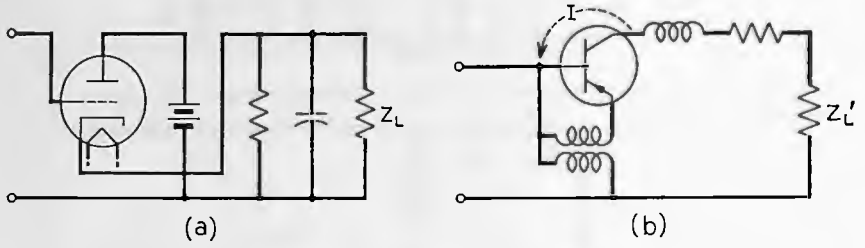


Fig. 18—Cathode follower and dual.

emitter current of the second transistor is large and that of the first is zero or negative. The process will then repeat.

Figure 18 shows a simple cathode follower and its dual. It has been explained in the text that, in circuits where the cathode current or the grid-to-plate voltage play an important part, the dual circuit will usually require a transformer. Alternatively it may be said that, in any circuit in which feedback in a single stage plays an important role, a transformer may be a necessity. In fact, we have found it impossible to avoid the use of a transformer in the dual of the cathode follower.

The transistor circuit shown will need another power supply for emitter bias, and a blocking condenser to prevent the bias current from flowing through one winding of the transformer. The power supply is required because the transformer will not transmit d-c. signals, and the condenser is necessary because the d-c. impedance of a transformer winding is nearly zero. Extra blocking condensers will appear in association with transformers in many circuits, especially in cases where the transformer is being used as the dual of another transformer.

A vacuum tube cathode follower ordinarily has a high input impedance

and a low output impedance, and has a voltage gain nearly equal to one. The dual circuit has a low input impedance and a high output impedance, and a current gain nearly equal to one. This comes about as follows: Suppose the collector circuit resistance and load are small. Then the collector is approximately at ground potential. Let the input current increase. This tends to increase the voltage drop from base to collector, and therefore the base potential tends to rise. This rise is transmitted to the emitter by the transformer, and therefore the emitter current rises. The rise in emitter current causes a drop in collector resistance, and counteracts the tendency for the collector-to-base voltage to rise. The result is that the input current passes through the collector circuit into the load without any corresponding rise in voltage between base and ground. This means that the input impedance is low. Of course, not all the input current passes to the load; some is passed to the emitter circuit. In a practical case tested, using a transistor whose

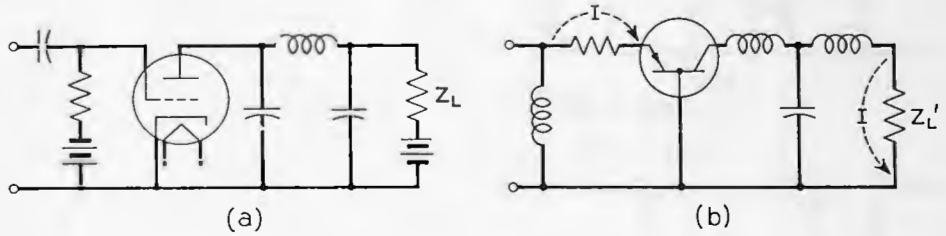


Fig. 19—Plate detector and dual.

base and emitter resistances were of the order of a few hundred ohms, with a load of 5000 ohms, the current gain was about .70 and the input impedance was about 40 ohms.

Figures 19, 20, and 21 show several amplitude modulation detectors and their duals. The purpose of all these circuits is to derive from an amplitude modulated wave a wave proportional to the envelope of the given wave.

Figure 19 shows a plate detector and its dual. The plate detector looks like a single-stage amplifier with a low-pass filter in its output circuit. It is biased approximately to plate current cutoff. As an amplifier it amplifies approximately the upper half of the input wave and does not pass the lower half. The filter smooths the succession of current pulses in the plate circuit and gives an output proportional to the average of the upper half of the input wave. If the input is a true amplitude modulated wave, this is also proportional to the envelope.

The dual circuit operates in the same way. The circuit looks like a one-stage amplifier with a low-pass filter in the collector circuit. It is biased to collector voltage cutoff. The negative part of the input signal is amplified, and the positive part is not. The collector voltage is a succession of pulses

of varying amplitudes. These are smoothed by the filter; and the output, as before, is proportional to the envelope of the input.

It is important for the proper operation of these circuits that the filter in the plate detector have low input impedance outside of the pass band, and that the filter in the dual circuit have low input admittance outside of the pass band. The exact form of the filter is immaterial.

Figure 20 shows a grid leak detector and its dual. The operation of the grid leak detector depends on the same principles as that of the grid leak bias circuit described before. The time constant of the bias circuit is chosen, however, so that the bias will be able to vary fast enough to follow the en-

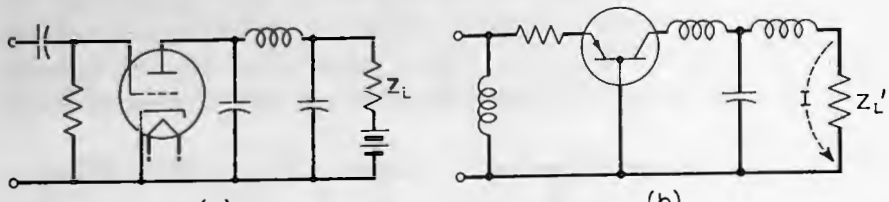


Fig. 20—Grid leak detector and dual.

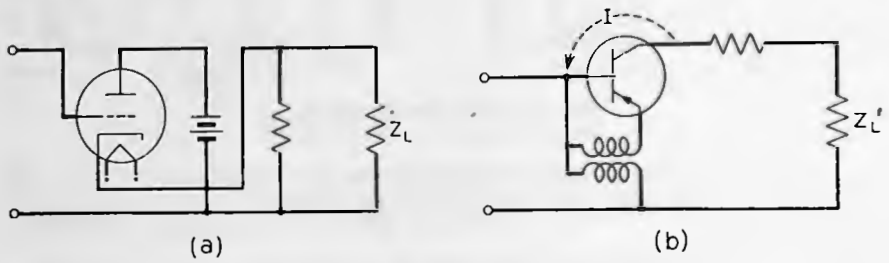


Fig. 21—Infinite impedance detector and dual.

velope of the input wave. The overall grid voltage or emitter current is then the input with a super-imposed wave proportional to the envelope of the input. These are amplified together, and the undesired high-frequency components are removed by a filter in the output circuit, leaving only the envelope wave. The filter must have approximately the same qualities as in the previous case.

Figure 21 shows an infinite impedance detector and its dual. This can be thought of as a cathode follower with a large capacitor across the cathode resistor. This capacitor charges through the comparatively low impedance of the tube when the signal is positive and reaches approximately the peak potential of the input wave. When the input voltage falls, the tube is cut off and the condenser must discharge through a comparatively high impedance. If the time constant of the discharge is properly chosen, the con-

denser will remain charged approximately to the instantaneous peak value of the input wave, which is the envelope of the input. The transistor circuit works in a similar way. When the input current is positive, the circuit behaves like a cathode follower, and a current is sent through the inductor L which is approximately equal to the input current. When the input current falls, the emitter current rises, and the collector presents a low impedance.

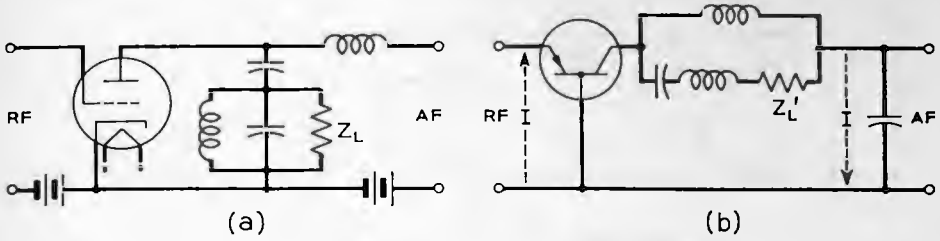


Fig. 22—Plate modulator and dual.

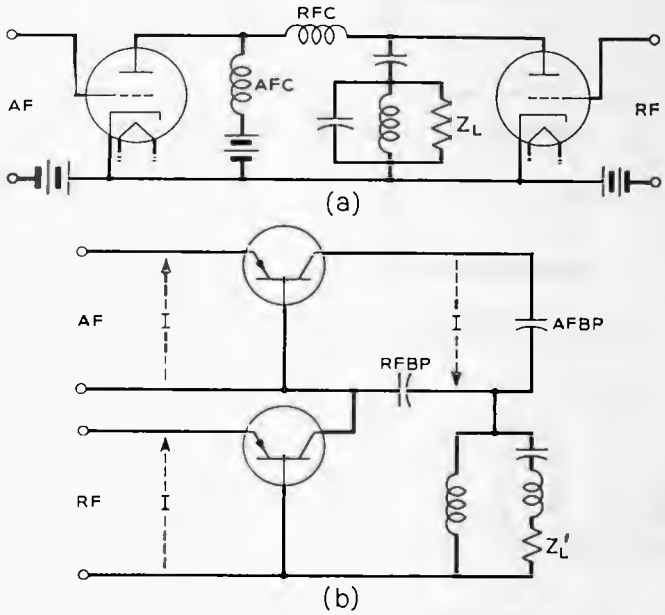


Fig. 23—Constant current modulator and dual.

The current through the inductor then decays slowly through the load and the low collector impedance. Each time the input signal has a positive peak, the effect is to draw a large current through the inductor, which persists during the rest of the cycle.

The dual of the infinite impedance detector is not a very attractive circuit, because the transformer must act for the carrier frequency as well as for the envelope frequencies.

Figures 22, 23, 24, 25, 26 and 27 show various amplitude modulator

circuits and their duals. These have as their object the production of a carrier frequency wave with a given envelope.

Figure 22 shows a plate modulator and its dual. This circuit makes use of the fact that the output of a class *C* amplifier is proportional approximately to the plate supply voltage. In 22(a) the vacuum tube acts as a class *C* amplifier amplifying the carrier frequency wave, and the supply voltage is varied

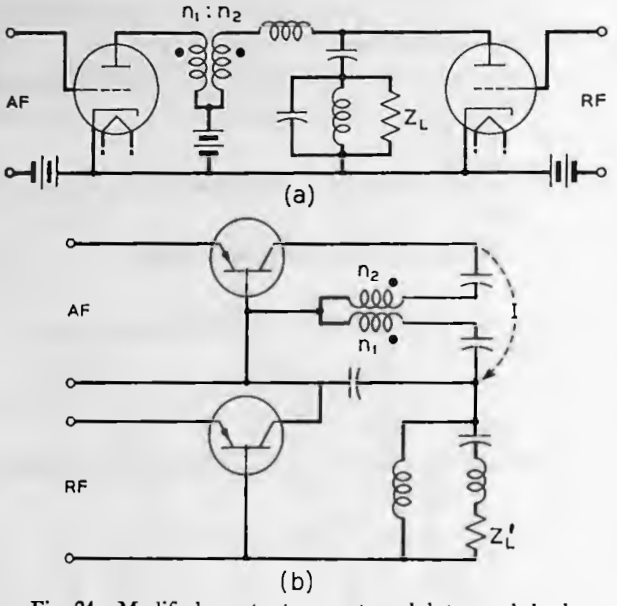


Fig. 24—Modified constant current modulator and dual.

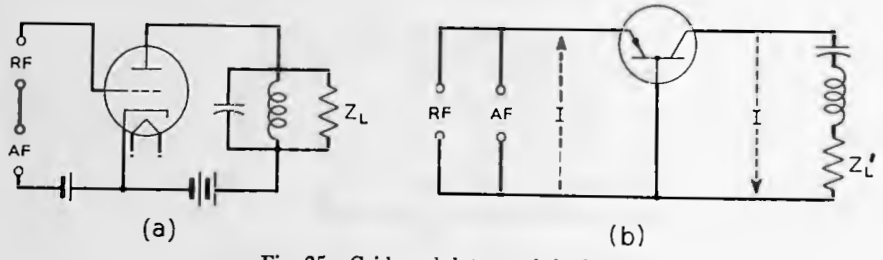


Fig. 25—Grid modulator and dual.

by adding to it the modulating voltage. The peak output voltage thus varies with the modulating voltage. The dual transistor circuit, 22(b), is also a class *C* amplifier, with its collector supply current varied by the addition of a modulating current. The output is proportional approximately to the total supply current, and hence varies with the modulating current.

Figure 23(a) shows a particular embodiment of the plate modulator which is called a constant current modulator, because the total supply current to the two tubes in the circuit is (approximately) constant. It is easier to ex-

plain this circuit by saying that the output of class C amplifier is proportional to the supply current. Two tubes, one a sort of audio amplifier and one a class C amplifier, are connected in parallel to a single constant current power supply consisting of a battery in series with a large inductor. The

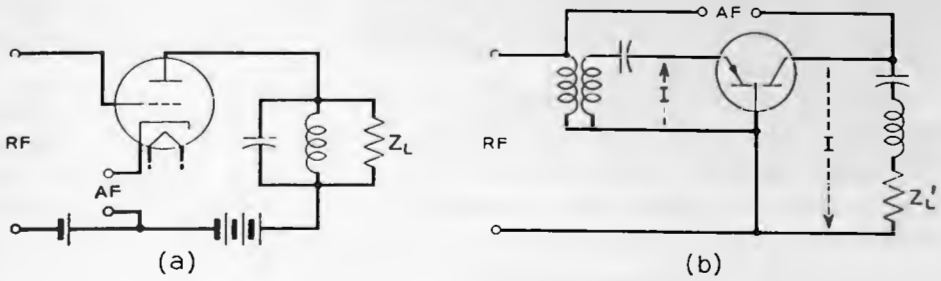


Fig. 26—Cathode modulator and dual.

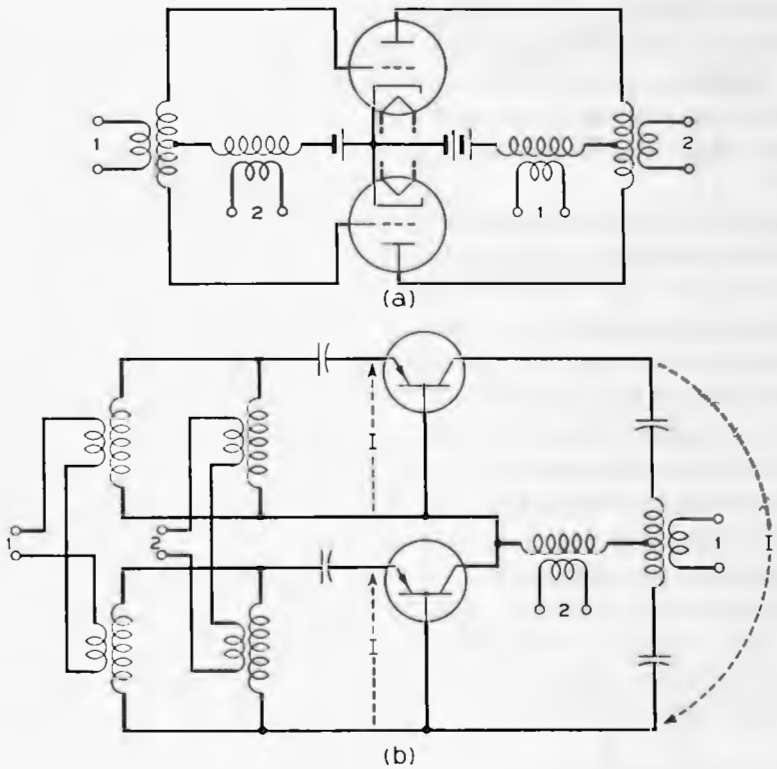


Fig. 27—General balanced modulator and dual.

class C amplifier can use only the current not used by the modulator tube, and hence its output varies inversely with the plate current of, and hence inversely with the grid voltage of, the modulator tube. The dual, Fig. 23(b), consists of a class C transistor amplifier in series with a class A modulator

transistor, connected to a source of constant voltage approximated here by a constant current supply in parallel with a large capacitor. The voltage available for the operation of the class *C* amplifier is the difference between the supply voltage and the collector voltage of the modulator transistor. Inasmuch as the output of the class *C* amplifier is proportional to its supply voltage, it is clear that the output will vary directly with the emitter current of the modulator transistor.

Both the vacuum tube and transistor circuit of Fig. 23 suffer because the modulator element can never take all of the available power supply voltage or current, and hence 100% modulation cannot be attained. The circuits of Fig. 24 correct this defect with a transformer, which amplifies slightly the variations in current or voltage in the modulator element. In both circuits $n_2 > n_1$, and the total supply current or voltage is no longer constant, but it is nearly so.

Figure 25 shows a grid modulator and its dual. Here the non-linearity of the transfer characteristics in the neighborhood of the cutoff point is made use of directly to produce modulation products. The tube is biased approximately to plate current cutoff, and the transistor approximately to collector voltage cutoff. The desired modulation products are selected by a tuned circuit.

Figure 26 shows a cathode modulator and its dual. These circuits combine some of the features of the grid modulator and of the plate modulator. Unfortunately the phase relationships are such in the transistor circuit that a transformer is required, and this transformer must be able to pass modulation frequencies as well as carrier frequencies.

The circuits of Fig. 25 and Fig. 26 can be operated as large-signal devices, using the gross non-linearities of the circuits to produce modulation products, or as small-signal devices, when they operate as 'square law' devices. They can, moreover, be combined to form various push-pull or balanced modulators. An example of such a circuit is shown in Fig. 27(a). This circuit has two inputs and two outputs. If it is operated as a square-law device the relations between the input and output frequencies will be as follows:

Input		*Output*	
1	2	1	2
a	—	2a	a
—	a	a, 2a	—
a	b	b, 2b, 2a	a, a+b, a−b
a, b	—	2a, 2b, a+b, a−b	a, b
—	a, b	a, b, 2a, 2b, a+b, a−b	—

The same relations hold for the dual circuit, Fig. 27(b). The action of the dual circuit is analogous to that of the vacuum tube circuit. It is a two-transistor circuit operated at the same time as a push-pull circuit and as two transistors in series, in phase. At various points in the circuit certain com-

ponents of the signal are zero because of the symmetries of the circuit. Notice that the dual of two vacuum tubes in parallel is two transistors in series.

Figure 28 shows a modulator which bears the same relation to the modulator of Reise and Skene (*U. S. Patent* 2,226,258) that the amplifier of Fig. 11 of the test bears to the Doherty amplifier. The carrier wave is fed into the tube and the transistor in the same way that the signal is fed into the amplifier, and the modulating signal is fed into the grid and the emitter through the transformers AF_1 and AF_2. The effect of the modulating signal is to vary the biases of the active elements. Inasmuch as both elements are used as class *B* or *C* amplifiers, their outputs are dependent on their biases.

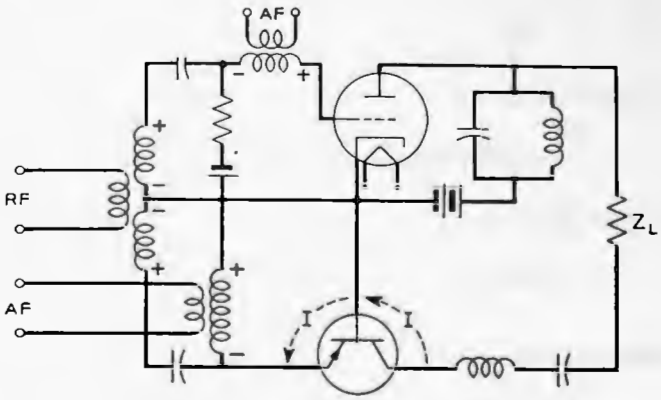


Fig. 28—High efficiency modulator.

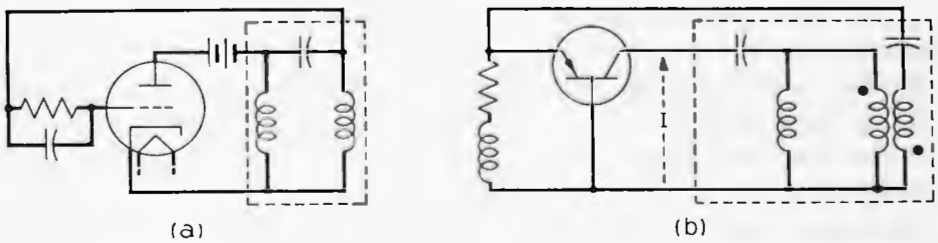


Fig. 29—Hartley oscillator and dual.

If the *RF* signal is large enough, and if the phases and turns ratios of the transformers are carefully chosen, the amplitude of the output will be nearly proportional to the modulating signal. A similar modulator can be based on the circuit of Fig. 10.

Figure 29 shows a Hartley oscillator and its dual. The configuration of elements in the Hartley oscillator may seem unfamiliar, but is chosen deliberately to emphasize the point of view that the Hartley oscillator is an amplifier with feedback through a coupling network. The part of the circuit enclosed in dotted lines is the coupling network. The dual circuit is also an amplifier with a coupling network, but because of the fact that the vacuum

tube has an inherent phase reversal and the transistor has not, an extra transformer is needed in the coupling network. It is conveniently placed as shown, where it can be combined with the inductor. The input circuits to the amplifiers have self bias circuits which have already been described.

The Colpitts oscillator is similar to the Hartley oscillator. The difference lies in the coupling circuit. Figure 30 shows the coupling circuit for the Colpitts oscillator and its dual. The tuned-plate tuned-grid oscillator can be treated in exactly the same way.

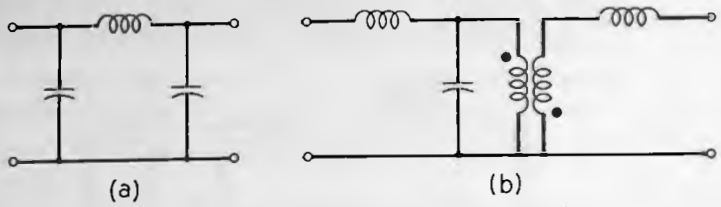


Fig. 30—Colpitts coupling network and dual.

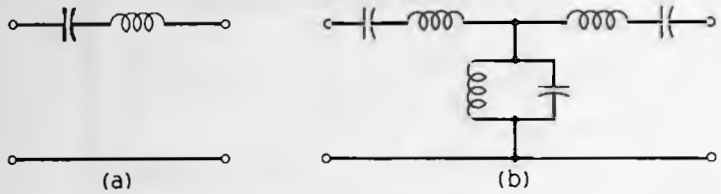


Fig. 31—Transistor oscillator coupling networks.

The coupling circuits used in successful vacuum tube oscillators are characterized by having a phase shift of 180°. This can be done with structures like low- or high-pass filter sections, by *R-C* networks, and in other ways. On the other hand, the coupling network required for a good transistor oscillator must have zero (or 360°) phase shift. It is therefore probably not most easily arrived at by designing a 180° phase shifting network and adding a phase inverting transformer. Two simple coupling networks which have zero phase shift are shown in Fig. 31. These are both band-pass structures which lead to oscillation in the pass band. Of the two, the network of Fig. 31(b) gives a more rapid change of phase with frequency and hence leads to a more stable oscillator. In addition, this circuit has the potential advantage of providing means for matching impedances.

SOME CIRCUIT ASPECTS OF THE TRANSISTOR

R. M. RYDER R. J. KIRCHER

BSTJ VOL.XXVIII pp.367-401 JULY 1949

Produced outside the work performed under this contract

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 Shockley[1] 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. Schockley, 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 transistor[3], 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,

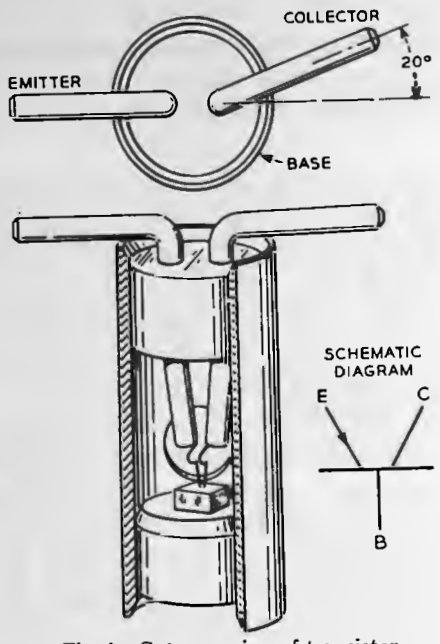


Fig. 1—Cutaway view of transistor.

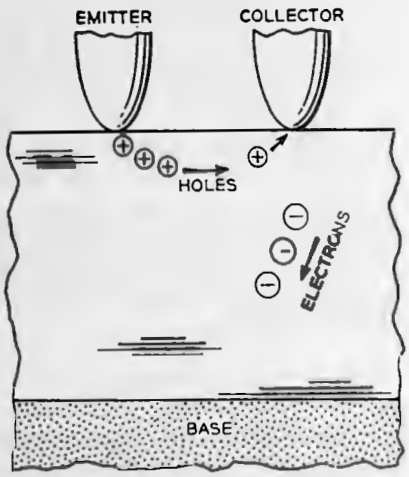


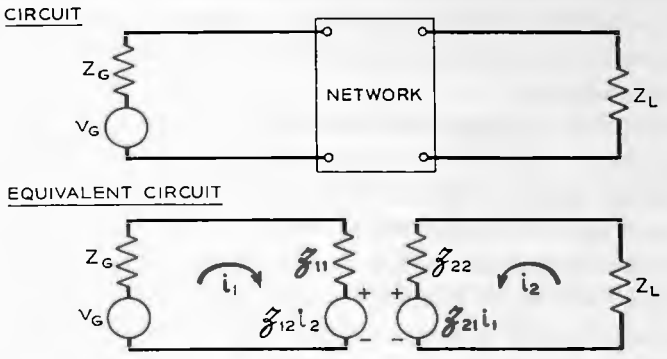
Fig. 2—Transistor mechanism.

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



Equations $\quad i_1(Z_g + \mathcal{Z}_{11}) + i_2\mathcal{Z}_{12} = v_G$
$\quad i_1\mathcal{Z}_{21} + i_2(\mathcal{Z}_{22} + Z_L) = 0$

Circuit determinant $\quad \Delta = (\mathcal{Z}_{11} + Z_g)(\mathcal{Z}_{22} + Z_L) - \mathcal{Z}_{12}\mathcal{Z}_{21}$

Input impedance $\quad Z_{11} = \mathcal{Z}_{11} - \dfrac{\mathcal{Z}_{12}\mathcal{Z}_{21}}{\mathcal{Z}_{22} + Z_L}$

Output impedance $\quad Z_{22} = \mathcal{Z}_{22} - \dfrac{\mathcal{Z}_{12}\mathcal{Z}_{21}}{\mathcal{Z}_{11} + Z_G}$

Operating power gain $\quad G_0 = 4R_G R_L \left|\dfrac{-\mathcal{Z}_{21}}{\Delta}\right|^2$

Insertion power gain $\quad G_I = \left|\dfrac{(Z_G + Z_L)\mathcal{Z}_{21}}{\Delta}\right|$

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.

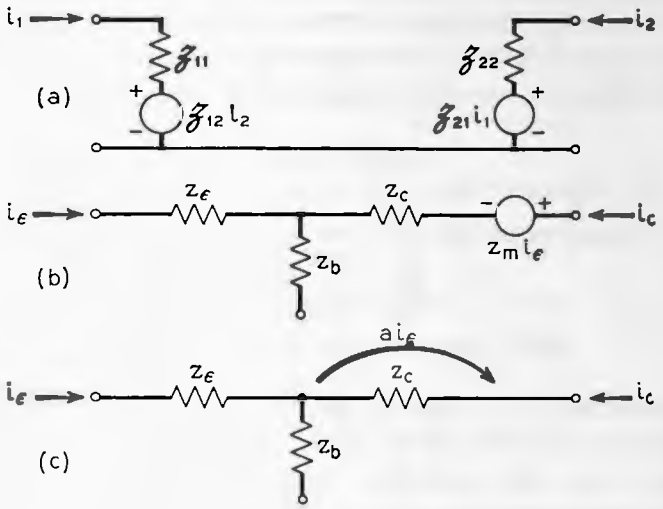
This description by open-circuit impedances happens to be a good one for many purposes, but there is nothing final or unique about it. In fact at high frequencies one of the other descriptions becomes more convenient.

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

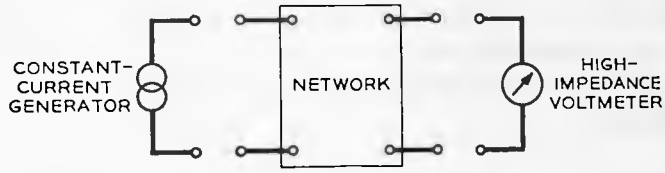


Fig. 5—Principle of measurement method.

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:	$I_e = 0.6$ ma	$V_e = 0.7V$
	$I_c = -2$ ma	$V_c = -40V$
Circuit Parameters:	$r_e = 240$ ohms	$r_b = 290$ ohms
	$r_c = 19000$ ohms	$r_m = 34000$ ohms
	$\mathcal{G}_{11} = 530$ ohms	$\mathcal{G}_{12} = 290$ ohms
	$\mathcal{G}_{21} = 34000$ ohms	$\mathcal{G}_{22} = 19000$ ohms

Fig. 6—Equivalent circuit parameter values.

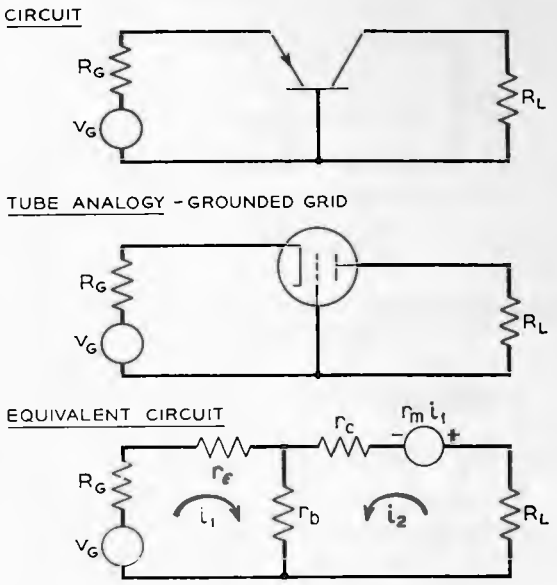
Some comments are in order on how this amplifier compares with an electron tube amplifier. First of all, the amplifying function and the manner of analyzing it from the circuit point of view are very similar, even though the internal mechanisms are markedly different. Secondly, there are qualitative 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:

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

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

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

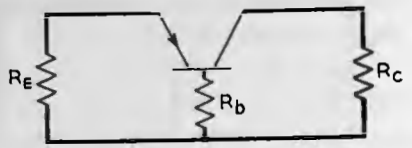
Operating power gain $G_0 = 4R_G R_L \left(\dfrac{-(r_b + r_m)}{\Delta}\right)^2$

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. Peterson[4] 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

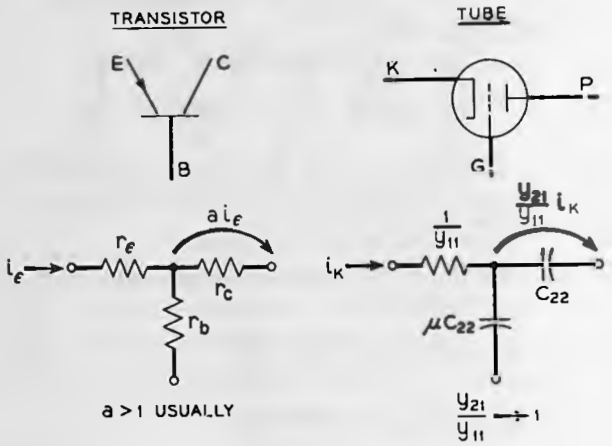


Fig. 9—Transistor-electron tube analogy.

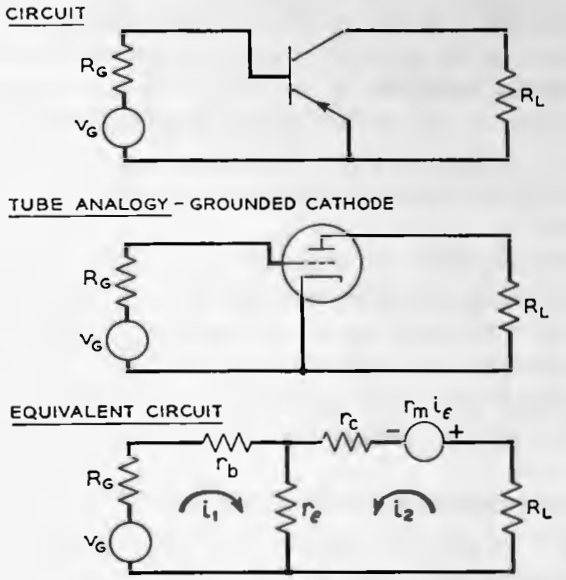
for the triode is close to unity for usual conditions. Another difference, of less importance, is the fact that the tube quantities analogous to r_c and r_b are capacitative reactances; their ratio, however, is like the ratio of r_c to r_b in magnitude.

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 for stability

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

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

Operating Gain $G_F = 4\, R_G R_L \left(\dfrac{r_m - r_e}{\Delta}\right)^2$

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

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.

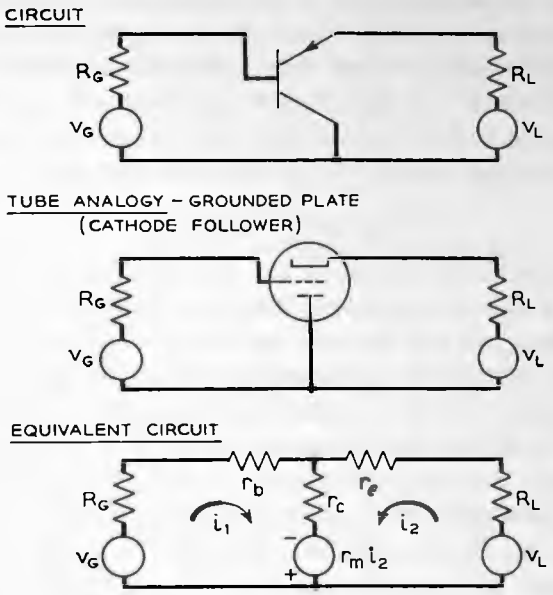
For numerical comparison we might work the grounded emitter amplifier between the same two terminations as the grounded base amplifier above, namely from 500 into 20,000 ohms. It would then have a gain of about 24 db, an improvement of 7 db over the grounded base, with about the same power output and 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_o = 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:

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

Circuit determinant

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

Input impedance

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

Output impedance

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

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:

For $R_G = 20000^{\omega}$, $R_L = 10000^{\omega}$
Then $R_{11} = -41000^{\omega}$
$R_{22} = -7600^{\omega}$
$G_F = 15^{db}$
$G_R = 13^{db}$

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,

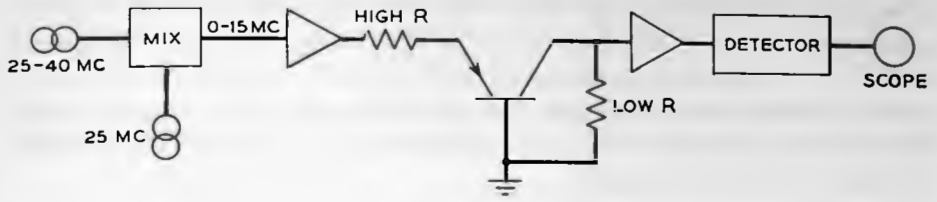


Fig. 12—Sweeper for measuring frequency response.

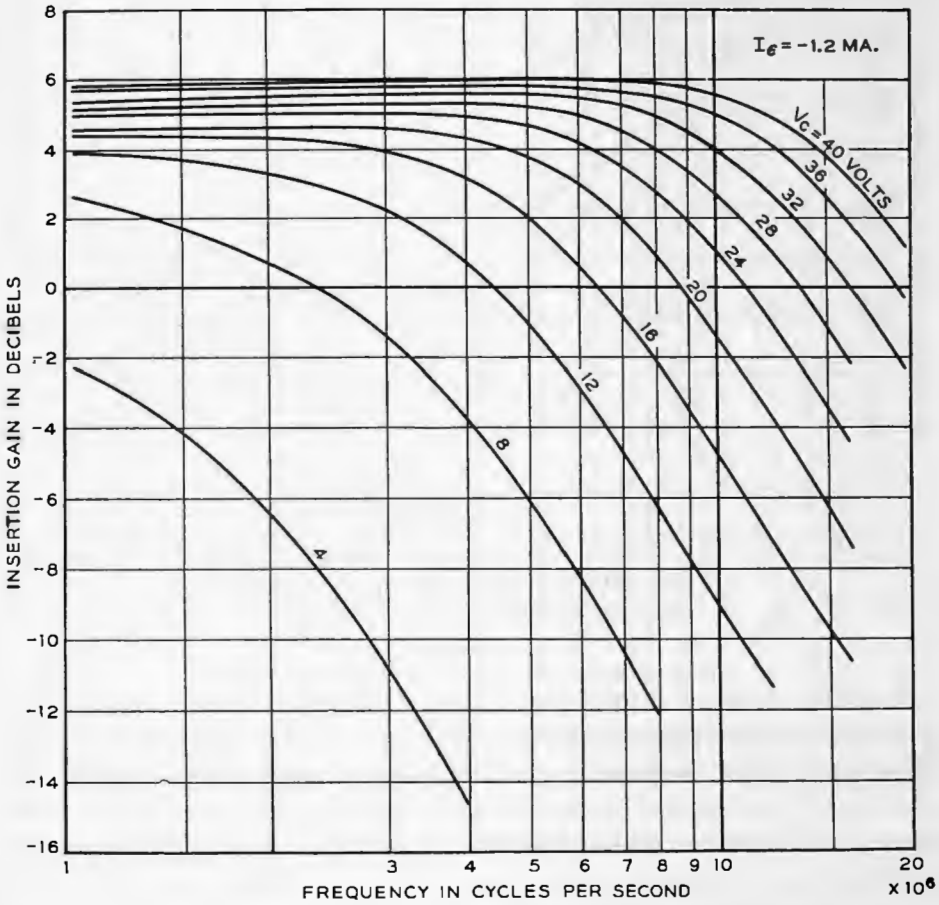


Fig. 13—Alpha versus frequency.

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.

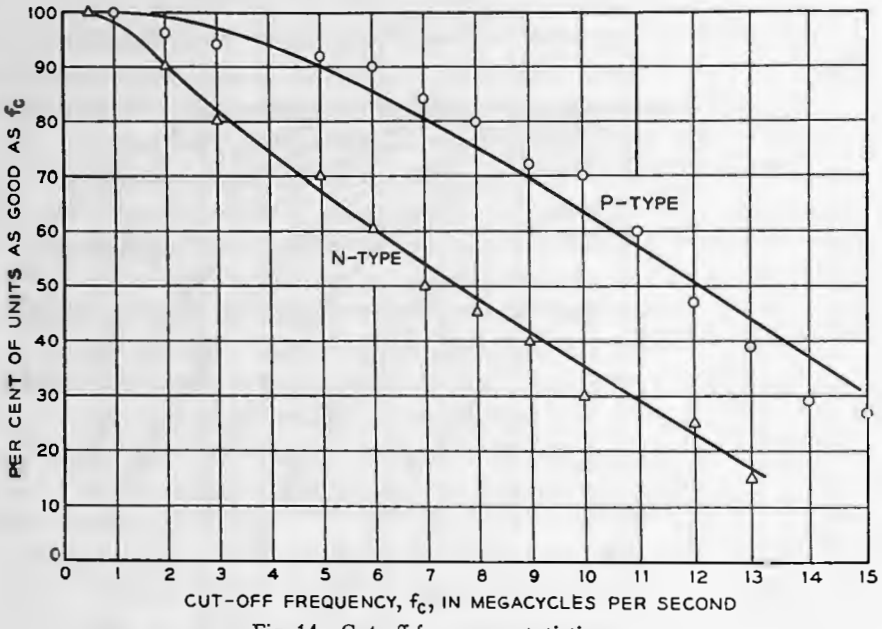


Fig. 14—Cut-off frequency statistics.

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, $\alpha = (\partial I_c/\partial I_e)_{V_c}$ 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.

mum phase" passive circuit.[6] 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

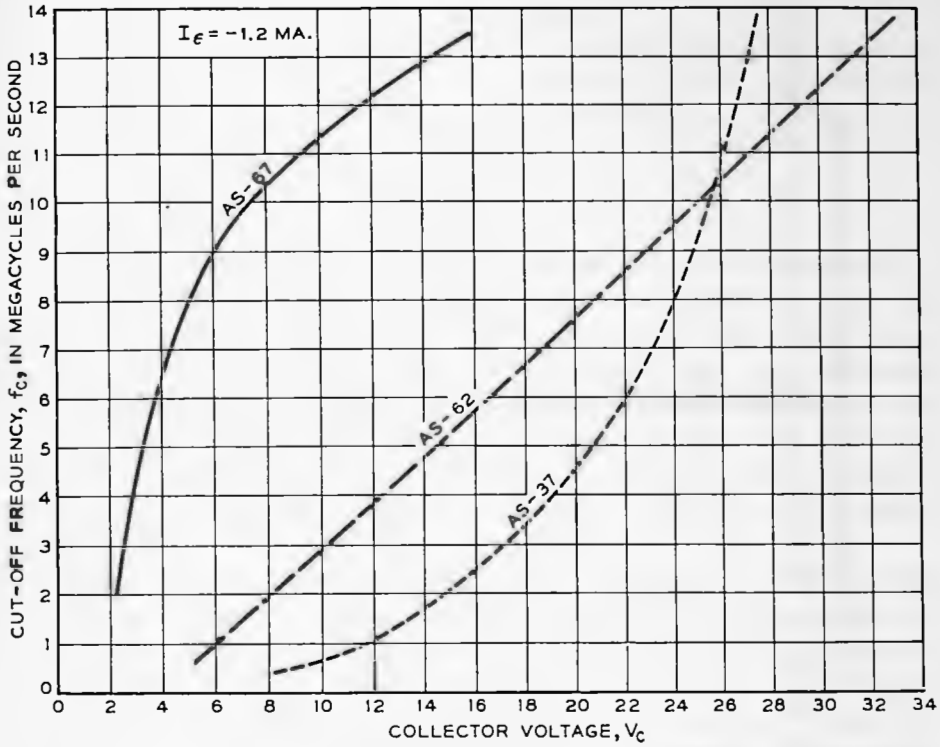


Fig. 15—Cut-off frequency versus collector voltage.

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.

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

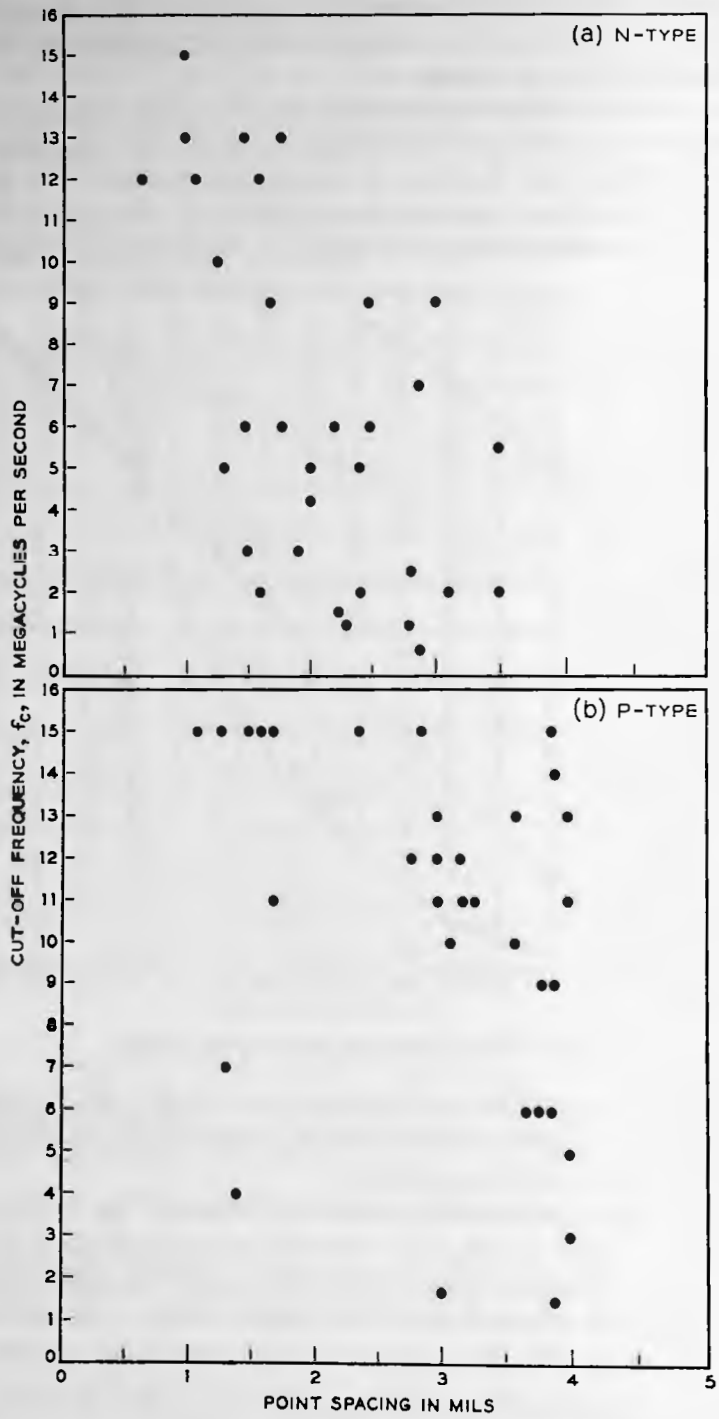


Fig. 16—Cut-off frequency versus point spacing.

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

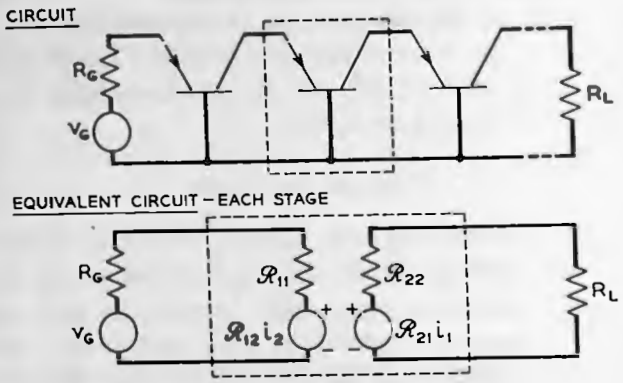
Many cascading possibilities exist, since any connection of the transistor might be used in combination with other connections, as well as involving all the parameter variations which might be made on each single stage. Some of the more elementary possibilities will be mentioned. Since feedback in each unit greatly complicates the situation, the essential features of the amplifiers may become clearer by discusing 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 $\Delta = (\mathcal{R}_{11} + \mathcal{R}_{22})^2$

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

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

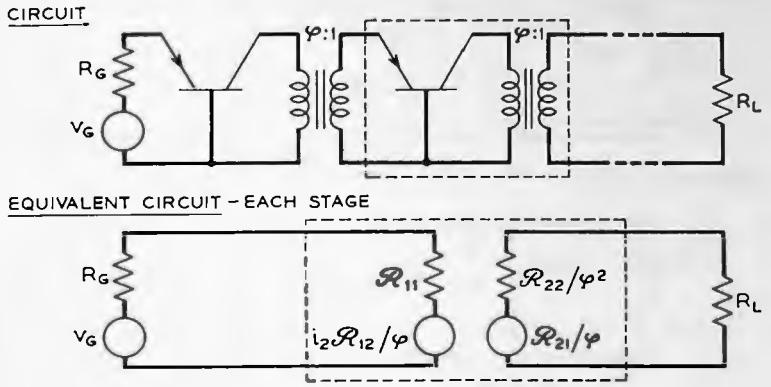
Nominal Type A Gain = 5^{db}

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 ($\mathcal{R}_{12} = 0$):

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

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

Insertion Power Gain $G_I = \left(\dfrac{-\mathcal{R}_{21}/\varphi}{\mathcal{R}_{11} + \mathcal{R}_{22}/\varphi^2}\right)^2$

Maximum when $\mathcal{R}_{11} = \mathcal{R}_{22}/\varphi^2$

$$G_{I\,\max.} = \frac{\mathcal{R}_{21}}{4\,\mathcal{R}_{11}\,R_{22}} = \frac{1}{4}\alpha^2\frac{\mathcal{R}_{22}}{\mathcal{R}_{11}}$$

Nominal Type A Gain: without feed back = 15^{db}
with $\mathcal{R}_{12}$ normal = 17^{db}

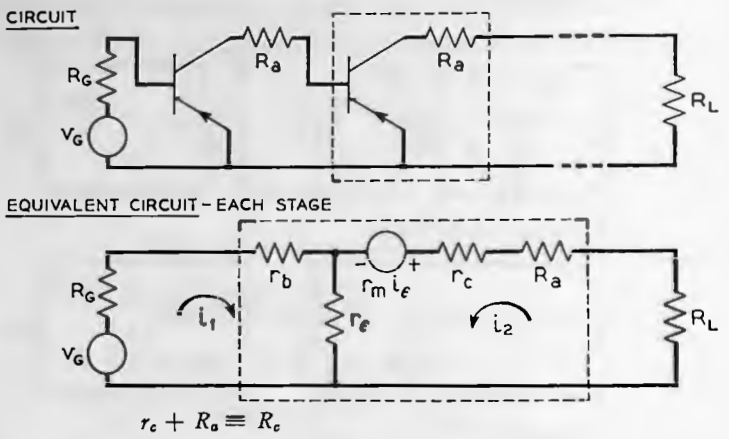
Fig. 18—Synopsis of grounded base cascade with transformers.

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



$r_c + R_a \equiv R_c$

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

Without feed back $(r_e = 0)$

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

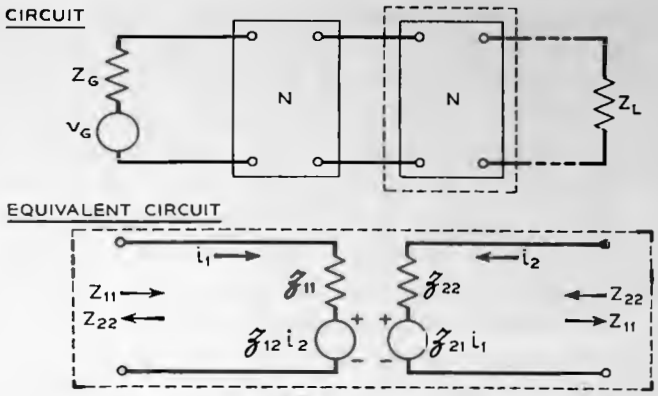
Circuit determinant $\Delta = (r_b + R_c - r_m)^2$

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.



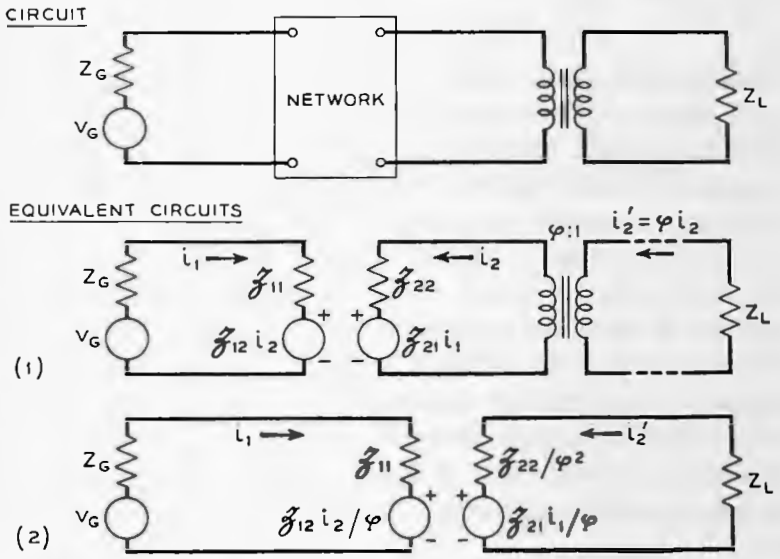
Equations: $i_1(\mathfrak{Z}_{11} + Z_{22}) + i_2\mathfrak{Z}_{12} = v_G$
$i_1\mathfrak{Z}_{21} + i_2(\mathfrak{Z}_{22} + Z_{11}) = 0$

Terminations: $Z_{11} = Z_L = -\mathfrak{Z}_{22} + \frac{1}{2}(\mathfrak{Z}_{11} + \mathfrak{Z}_{22})(1 + \sqrt{1-y})$
$Z_{22} = Z_G = -\mathfrak{Z}_{11} + \frac{1}{2}(\mathfrak{Z}_{11} + \mathfrak{Z}_{22})(1 + \sqrt{1-y})$
$y = 4\mathfrak{Z}_{12}\mathfrak{Z}_{21}/(\mathfrak{Z}_{11} + \mathfrak{Z}_{22})^2$

Circuit determinant $\Delta = \frac{1}{2}(\mathfrak{Z}_{11} + \mathfrak{Z}_{22})^2(1 - y + \sqrt{1-y})$

Insertion Power Gain $G_I = \left| \frac{\mathfrak{Z}_{21}}{\mathfrak{Z}_{11} + \mathfrak{Z}_{22}} \frac{2}{1 + \sqrt{1-y}} \right|^2$

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



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

$$i_1\mathfrak{Z}_{21}/\varphi + i_2'\left(\frac{\mathfrak{Z}_{22}}{\varphi_2} + Z_L\right) = 0$$

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

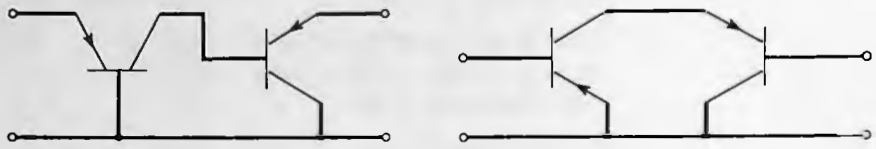


Fig. 22—Non-iterative cascade amplifiers.

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:

$$\begin{aligned} V_\epsilon &= f_1 (I_\epsilon, I_C) \\ V_C &= f_2 (I_\epsilon, I_C) \end{aligned} \tag{2}$$

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

$$\begin{aligned} v_\epsilon &= i_\epsilon \frac{\partial f_1}{\partial I_\epsilon} + i_C \frac{\partial f_1}{\partial I_C} \\ v_C &= i_\epsilon \frac{\partial f_2}{\partial I_\epsilon} + i_C \frac{\partial f_2}{\partial I_C} \end{aligned} \tag{3}$$

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.

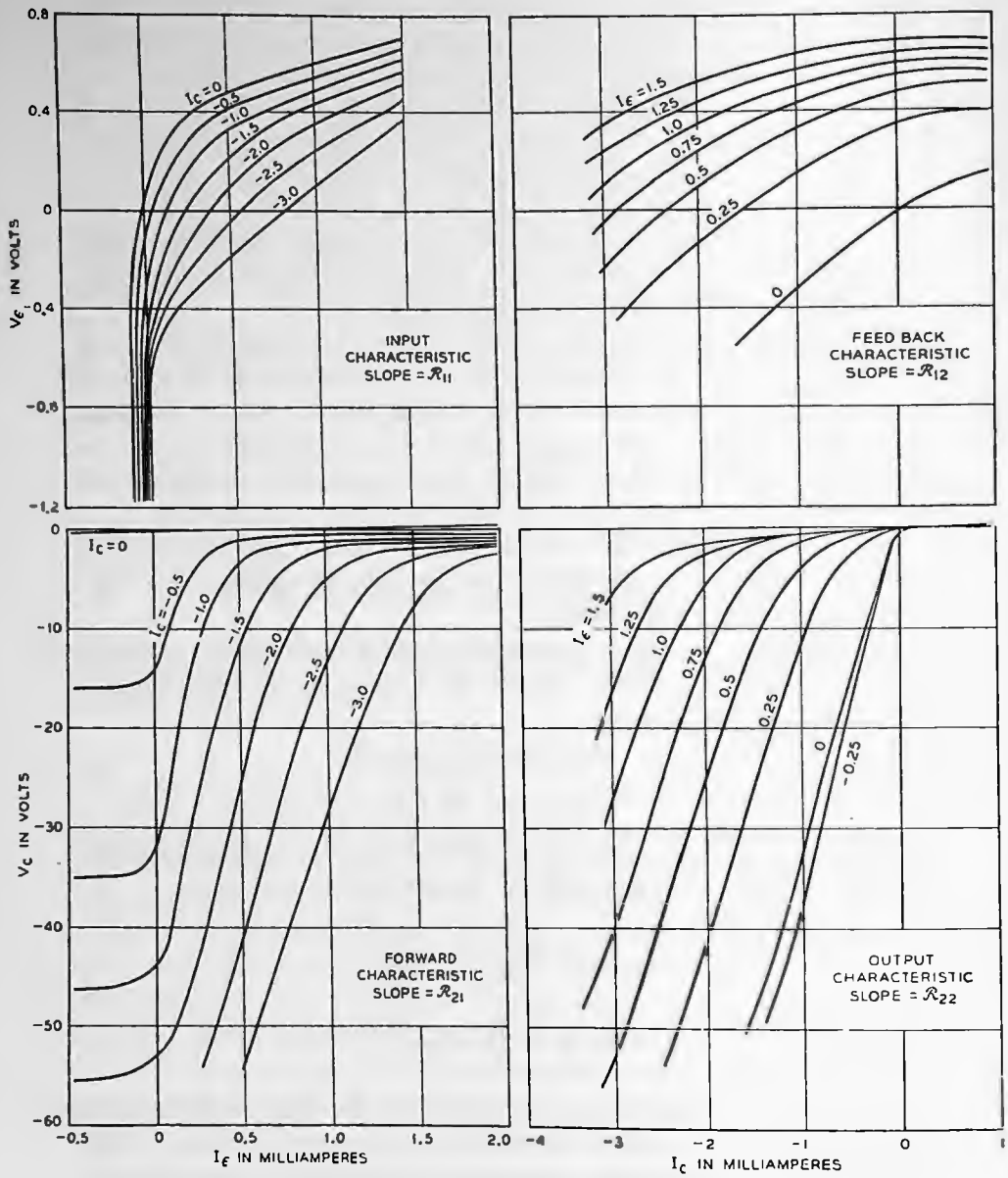


Fig. 23—Static characteristics.

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

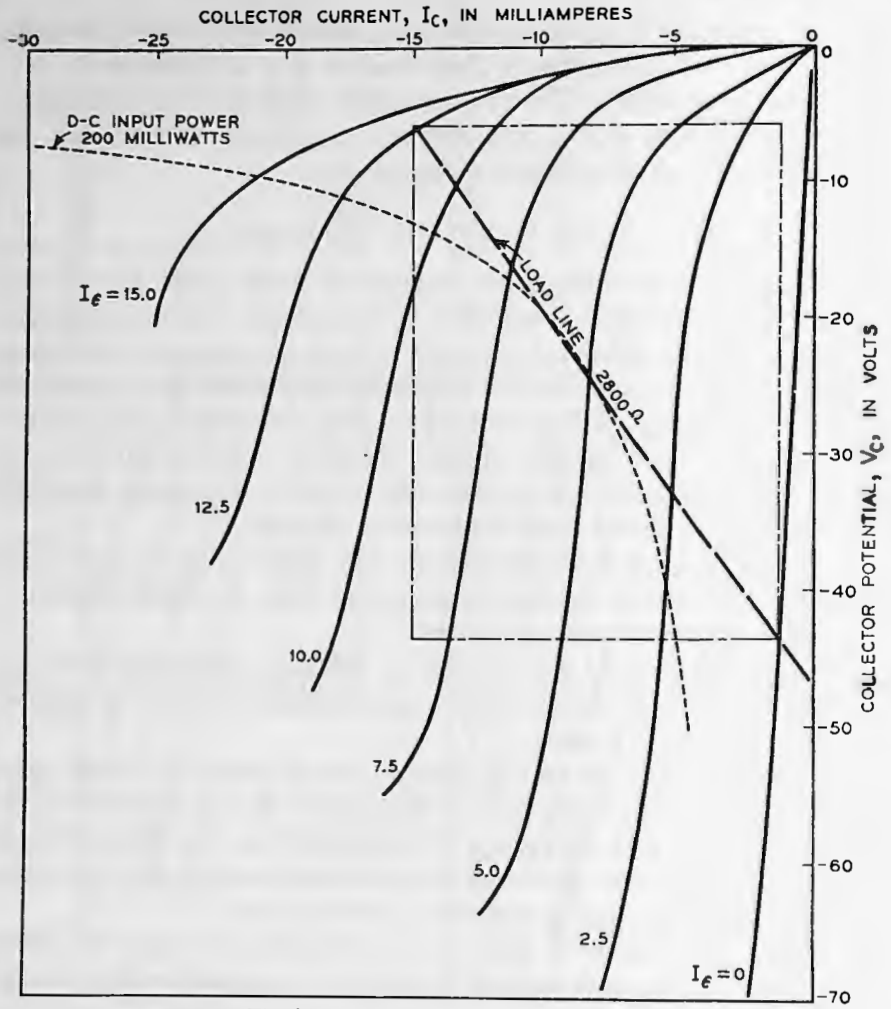


Fig. 24—Collector power output plot.

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 larger 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 transisto- include an oscillator with anti-resonant circuit in the base lead, or with a

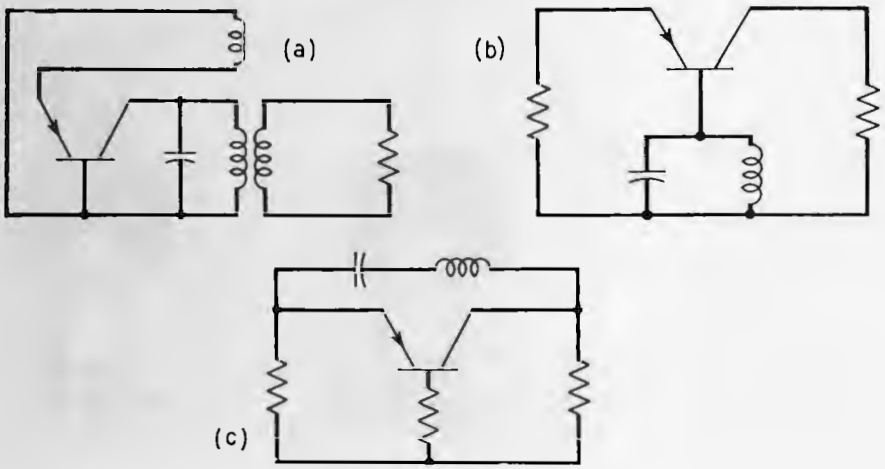


Fig. 26—Transistor oscillators.

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

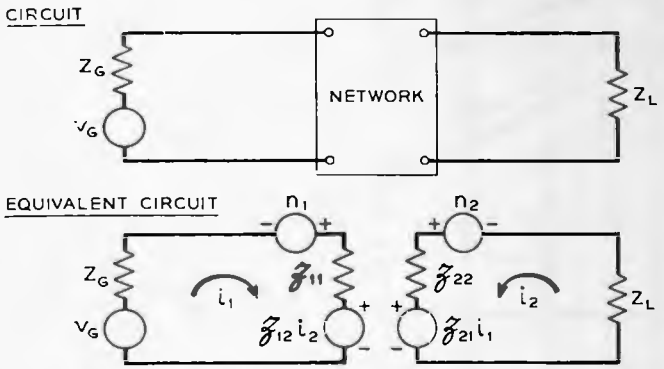
A discussion of small-signal amplifiers would be incomplete without some mention of the limiting factor of noise. The noise has been left to the last, however, because its discussion complicates the circuits slightly and perhaps because it is not well to present too early an aspect of performance which 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. 1264–1 272

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



Equations: $i_1(Z_G + \mathcal{Z}_{11}) + i_2\mathcal{Z}_{12} = v_G \oplus N_1$
$i_1\mathcal{Z}_{21} + i_2(\mathcal{Z}_{22} + Z_2) = \oplus N_2$

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

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

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 ooise 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

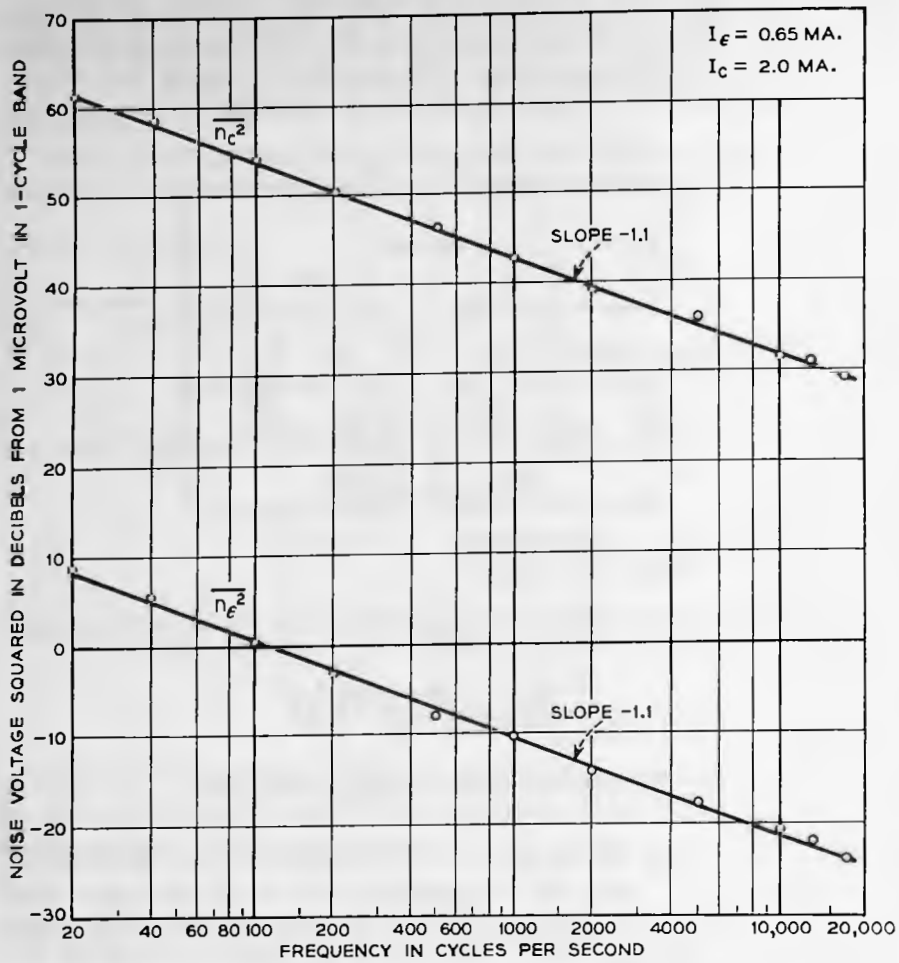


Fig. 28—Transistor noise versus frequency.

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.

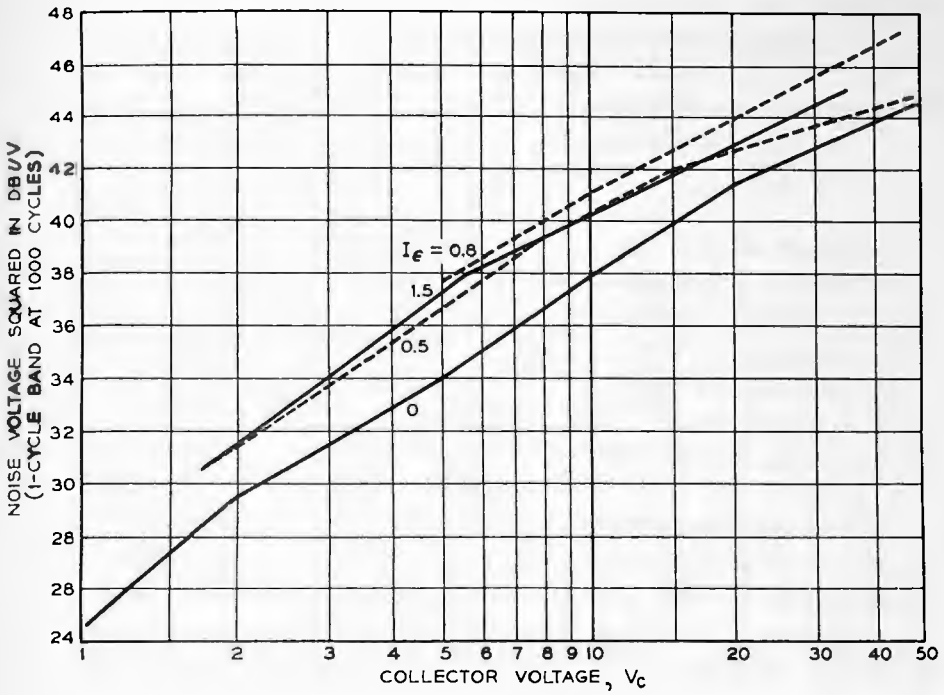


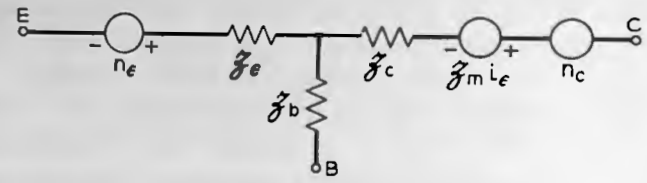
Fig. 29—Transistor noise versus operating point.

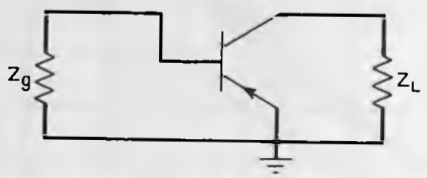
A representative noise figure for the Type A transistor at 1000 cycles is 60 db, with individual units ranging from 50 to 70 db.

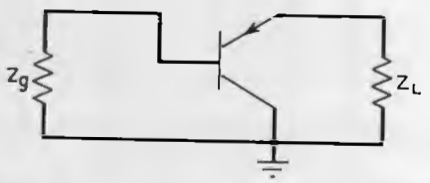
Noise figure formulas for the three single-stage connections are given in Fig. 30. The noise performance of the three connections would usually not be very different if it were not for stability considerations, which may render unusable the generator impedance which would give optimum performance. Mainly, 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

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

Grounded Emitter

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

Grounded Collector

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

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

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.

SOME CIRCUIT PROPERTIES AND APPLICATIONS OF n-p-n TRANSISTORS

R. L. WALLACE, JR. W. J. PIETENPOL

BSTJ VOL.XXX pp.530-563 JULY 1951
PROC. IRE VOL.39 NO.7 pp.753-767 JULY 1951

Produced outside the work performed under this contract

Some Circuit Properties and Applications of *n-p-n* Transistors

By R. L. WALLACE, JR. and W. J. PIETENPOL

Shockley, Sparks, and Teal have recently described the physical properties of a new kind of transistor. Preliminary studies of circuit performance show that it is a stable, high gain, quiet amplifier of considerable practical interest. This paper analyzes the performance of a few early experimental units.

Introduction

Almost two years ago, W. Shockley[1, 2] first published the theory of a transistor made from a single piece of germanium in which the conductivity type varies in such a way as to produce two rectifying junctions. Since that time, M. Sparks, G. K. Teal and others at the Bell Telephone Laboratories[3,4] have contributed notably to the physical realization of this device.

Recently Sparks has produced a number of *n-p-n* transistors and has found their behavior to be closely in accord with Shockley's theory.[4] Preliminary circuit studies on these devices have shown that in several respects their performance is remarkable. In view of this, our transistor development group has undertaken to produce small quantities of *n-p-n* transistors in a form suitable for incorporation in working circuits.

This paper will deal principally with the circuit aspects of the *n-p-n* transistor by presenting and analyzing performance data on a small number of experimental units. For a discussion of the solid state physics of its design and operation the reader is referred to the previously mentioned works of Shockley, Sparks, and Teal.

Outstanding Properties

Before getting lost in a maze of detail, it seems worthwhile to list and mention briefly the salient features of this new transistor. They are:

1. *Relatively low noise figure.* Most of the units measured so far have a noise figure between 10 and 20 db at 1000 cps.

2. *Complete freedom from short-circuit instability.* The input and output impedances are always positive whether the transistor is connected grounded-emitter, grounded-base, or grounded collector. This permits a great deal of freedom in circuit design and makes it possible, by choosing the appropriate connection, to obtain a considerable variety of input and output impedances.

3. *High gain.* Power gains of the order of 40 to 50 db per stage have been obtained

4. *Power handling capacity and efficiency.* The design can readily be varied to permit the required amount of power dissipation up to at least two watts. Furthermore the static characteristics are so nearly ideal that Class A efficiencies of 48 or 49 out of a possible 50% can be realized. The efficiencies for Class B and Class C operation are correspondingly high.

5. *Ruggedness and small size.* The germanium part of the transistor is enclosed in a hard plastic bead about $\frac{3}{16}$ inch in diameter. Inside the bead three connections are mechanically as well as electrically fastened to the germanium and are brought out as pigtails through the bead. This gives a very sturdy unit.

6. *Freedom from microphonics.* Vibration tests in the audio frequency range indicate that these devices are relatively free from microphonic noise.

7. *Limited frequency response.* Collector capacitance limits the frequency response at full gain to a few kilocycles. By using a suitable impedance

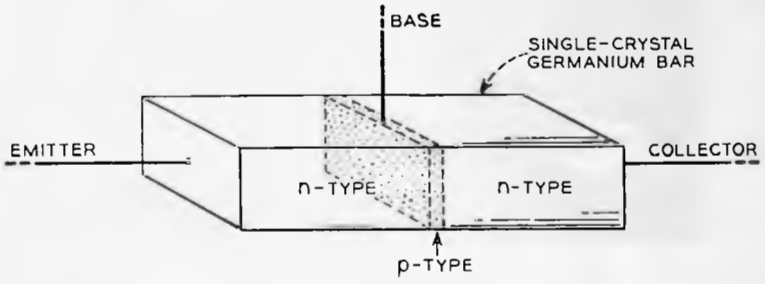


Fig. 1—The heart of an *n-p-n* transistor is a tiny bar of germanium to which three mechanically strong electrical connections are made.

mismatch it is possible to maintain the frequency response flat to at least one megacycle while still obtaining a useful amount of gain.

8. *Operation with exceedingly small power consumption.* Perhaps the most remarkable feature of these transistors is their ability to operate with exceedingly small power consumption. The best example of this to date is an audio oscillator which requires for a power supply only 6 microamperes at 0.1 volts. This represents 0.6 microwatts of power which contrasts sharply with the million or more microwatts required to heat the cathode of an ordinary receiving-type vacuum tube.

Physical Appearance and Construction

Figure 1 shows schematically the configuration of an *n-p-n* transistor. The small bar of single crystal germanium contains a thin layer of *p*-type interposed between regions of *n*-type. Mechanically strong ohmic connections are made to the three regions as indicated and brought out through a

hard plastic bead. A finished transistor is shown in the photograph of Fig. 2. It should be pointed out that Fig. 1 is not drawn to scale and that the *p*-layer may be less than a thousandth of an inch thick.

Static Characteristics

A great deal of information about the low frequency performance of a transistor can be obtained from a set of static characteristics such as those shown in Fig. 4. Curves of this sort are obtained simply by connecting suit-

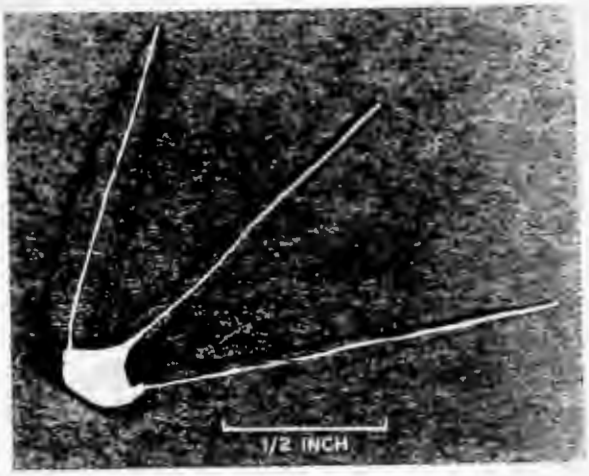


Fig. 2—A beaded *n-p-n* transistor.

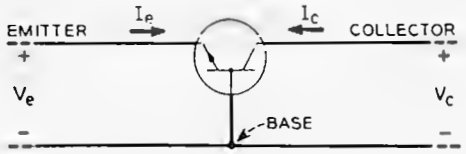


Fig. 3—The symbol for a *p*-type transistor on which the convention of signs for currents and voltages is indicated.

able current sources to the emitter and collector circuits of the transistor and measuring the resulting voltages. The currents are called positive when they flow into the emitter and collector as shown and the voltages are called positive when they have the signs shown in Fig. 3.

Let us first examine these curves with an eye to finding out what kind of voltage and current supplies are needed to bias the transistor into the range in which it can amplify. To make this easy, that part of the characteristics which lies within the normal operating range has been shown as solid lines and that part of the characteristics corresponding to cutoff has been shown as dotted lines.

Note from the upper set of curves that V_c is positive in the operating range. This means that the collector must be biased positive with respect to the

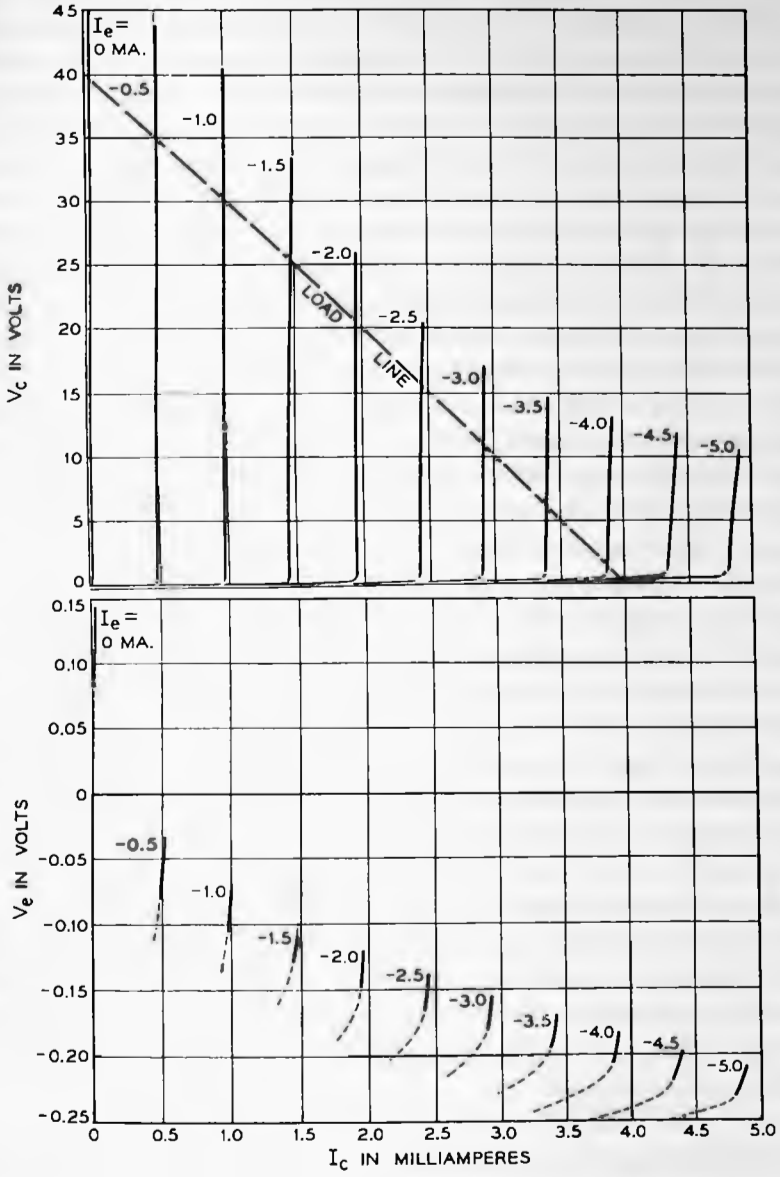


Fig. 4—Static characteristics of an *n-p-n* transistor.

base. For this particular transistor a bias voltage anywhere between about 0.1 volts and 35 volts is suitable. Note also that all the curves on this plot correspond to negative emitter currents. This means that the emitter must

be biased in such a way that current flows out of the emitter into a suitable current supply. Furthermore, the collector current corresponding to any given emitter current can be seen to be almost equal in magnitude to the emitter current. Since these two currents are opposite in sign, this means that most of the current which flows into the collector leaves by way of the emitter with the result that the current in the base circuit is very small.

Suppose that the collector is held at a constant positive voltage as, for example, by connecting a battery between collector and base (with a transformer winding in series, perhaps). Now if a negative current is forced into the emitter by a battery and resistance connected in series between emitter and base, the collector current can be controlled by varying the emitter current and will always be approximately equal in magnitude to the emitter current. Suitable collector currents for this particular transistor range from about 20 microamperes to about five milliamperes.

The exact choice of collector current and voltage within the ranges mentioned above will be dictated largely by the amount of power output required. The more power output required, the more current and voltage will be needed from the power supply. Since the collector circuit efficiency cannot exceed the theoretical limit of 50% in Class A operation, the signal power output cannot exceed half the power supplied by the battery. This means, for example, that if the collector is worked at 20 volts and 2 milliamperes the Class A power output cannot exceed 20 milliwatts.

From the lower plot of Fig. 4 it is possible to obtain information about the bias voltage required for the emitter. Note, first, that the entire emitter voltage plot corresponds to a very small range of emitter voltages near zero and, furthermore, that the part of the characteristics corresponding to the operating range covers only a few thousandths of a volt. This means that if the collector voltage is held constant very small changes in emitter voltage will produce fairly large changes in collector current, or if the collector current is held constant very small changes in emitter voltage will produce relatively enormous changes in collector voltage. This at once suggests the use of this transistor as a d-c. amplifier between a low impedance source and a high impedance load. In this application, voltage stepup of the order of 10,000 times is possible.

The very great sensitivity of the collector circuit to emitter voltage suggests, however, that for a-c. amplifiers one should use a current source as an emitter bias supply. This can be obtained from a battery and a large resistance in series. Furthermore, since the emitter voltage is always nearly zero, the emitter current can be calculated in advance by dividing the battery voltage by the value of the series resistance (provided, of course, that

the supply voltage is large compared to the few hundredths of a volt drop across the emitter circuit).

One can also draw some interesting conclusions from the static characteristics about the large signal operation of the transistor. If the load is resistive, the instantaneous operating point will swing up and down along a straight line such as the load line shown in the upper plot of Fig. 4. This particular load line corresponds to an a-c. load resistance of 10,000 ohms. Suppose that the steady collector biases are 20 volts and 2 milliamperes so that the drain from the power supply is 40 milliwatts. Now consider the permissible swings of collector voltage and current. Since the collector characteristics are quite straight and evenly spaced over a wide range of current and voltage values, the output signal can swing nearly down to zero collector volts and nearly up to zero collector current without distortion. The limit on the lower end is imposed by the fact that the collector characteristics begin to be curved when V_c is less than about 0.1 volts; and the limit on the upper end is imposed by the fact that the collector current does not drop completely to zero when I_e drops to zero. The lower limit of collector current is, in this case, about 50 microamperes and, since this amount of current in 10,000 ohms corresponds to 0.5 volts, this means that the instantaneous collector voltage is limited to swings between 39.5 volts and 0.1 volts. Starting from a quiescent value of 20 volts, the permissible positive swing is then 19.5 volts and the permissible negative swing is 19.9 volts. Reducing the quiescent voltage to 19.8 volts (and keeping the same load line) makes it possible to obtain a peak swing of 19.7 volts which corresponds to 19.45 milliwatts of signal delivered to the load. This gives a collector circuit efficiency of 48.5% out of a possible 50%. Some transistors take even less collector current when the emitter current is zero and hence permit even higher efficiencies.

These computations of efficiency have all been based on the assumption of sinusoidal *current* applied to the emitter. It will be shown in a later section that the emitter resistance varies with emitter current, however, and this means that to realize high efficiency with low distortion it is necessary to drive the emitter from a high impedance source.

Operation with Small Power Consumption

For small signal applications the transistor represented by the characteristics of Fig. 4 can deliver useful gain at very much lower voltages and currents than those used in the example above. In order to show this, the characteristics of Fig. 5 have been plotted for a range of collector voltage extending up to only 2 volts and for a range of collector currents extending

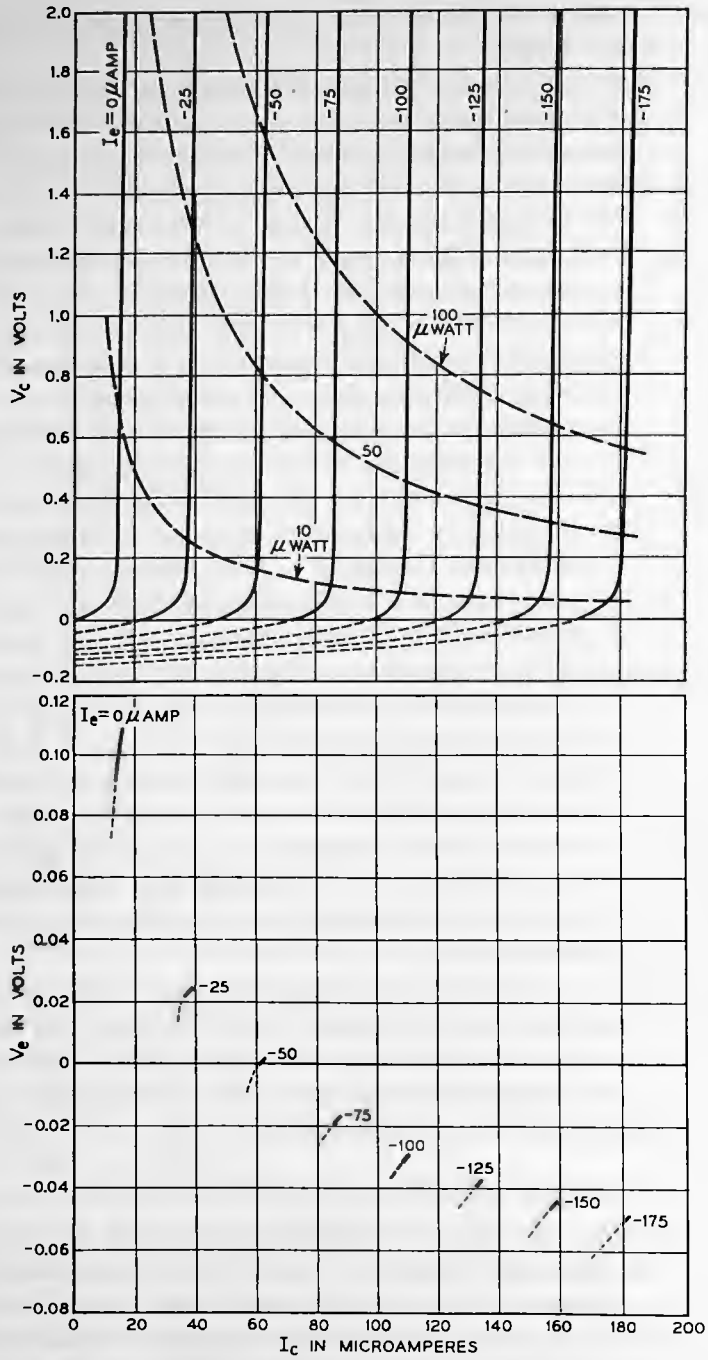


Fig. 5—Static characteristics showing behavior at very low applied voltages and currents.

up to only 200 microamperes. It can be seen from the upper plot that the collector circuit characteristics are still quite usefully straight and evenly spaced in this micro-power range. In fact, for small signal operation it is sufficient to use a collector voltage only a little in excess of 0.1 volts and a collector current a little in excess of 20 microamperes. This means that the power required to bias the collector into the operating range amounts to only a few microwatts. Contours are shown for 10, 50, and 100 microwatts of power supply.

This ability of the transistor to work with extremely small power consumption is one of its most striking and perhaps most important features

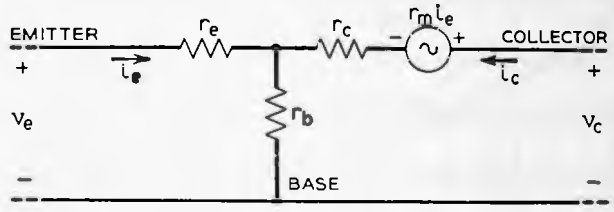


Fig. 6—The low-frequency equivalent circuit of a transistor.

When one considers that the total power consumption of a single transistor stage can be smaller by many thousands of times than the power required to heat the cathode in a vacuum tube, it is obvious that the advent of this device will make possible many new kinds of application.

Variation of Transistor Properties with Operating Point

Ryder and Kircher[5] have shown that it is convenient to analyze the small signal properties of a transistor at low frequencies in terms of the equivalent circuit of Fig. 6 where r_e is called the emitter resistance, r_b is called the base resistance, and r_c is called the collector resistance. The internal generator, $r_m i_e$, is the active part of the circuit and in this respect corresponds to the familiar μe_g of vacuum tube circuit theory. It is the purpose of this section to show what values these quantities have for a particular *n-p-n* transistor and to show how they vary with the applied biases. This will form a basis for the next section in which these quantities will be used to compute such things as the input and output impedances and the gains of various transistor connections.

Ryder and Kircher have shown that these four *r*'s can be obtained directly from static characteristics such as those shown in Fig. 4 and Fig. 5. In the case of *n-p-n* transistors, however, the magnitudes of these quantities are such that it is difficult to obtain satisfactory accuracy in this way and it has been more convenient to measure the 4-pole *r*'s by a-c. methods.

These measurements have shown that all of the r's are, to a first approximation, independent of collector voltage so long as the collector voltage is above a few tenths of a volt and so long as the total dissipation is small enough to prevent appreciable heating of the transistor.

In view of this fact it is perhaps sufficient to show how these quantities vary with emitter current for a moderate fixed value of collector voltage. Figures 7 and 8 show that r_c and r_m are very nearly equal and that they tend to decrease as I_e increases. Theoretically r_m and r_c should both be infinite. The fact that they reach values as low as 10 megohms in this case is a meas-

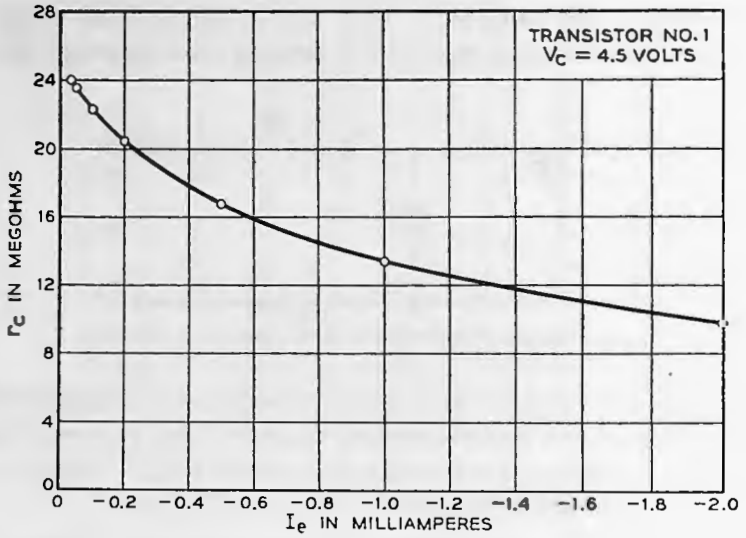


Fig. 7—The variation of collector resistance with emitter current at a fixed value of collector voltage.

ure of the imperfection in technique of fabricating the transistor. Values as high as 60 megohms have been achieved in the laboratory.

Figure 9 shows that r_b in this transistor is approximately 240 ohms and is independent of I_e.

Figure 10 shows that r_e decreases with increasing emitter current, ranging from about 500 ohms at 50 microamperes down to about 5 ohms at 5 milliamperes. Shockley[4] has shown that r_e should be given by

$$r_e = \frac{kT}{qI_e} \tag{1}$$

where q is the charge on an electron, k is Boltzman's constant, T is the Kelvin temperature and I_e is the emitter current. When the temperature

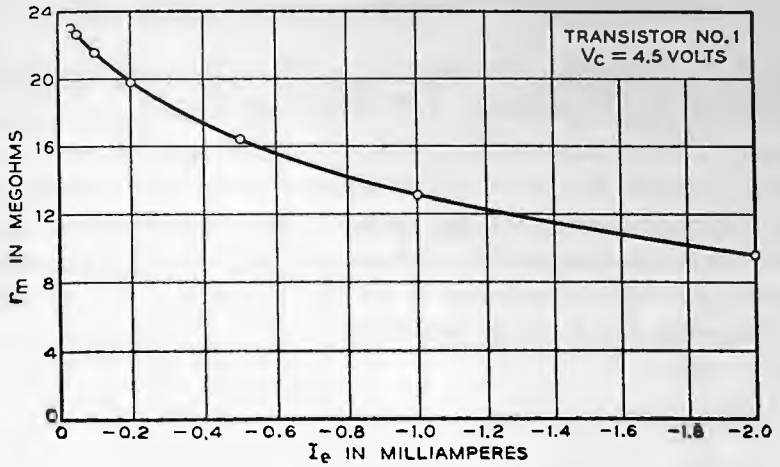


Fig. 8—Variation of r_m with emitter current at fixed collector voltage.

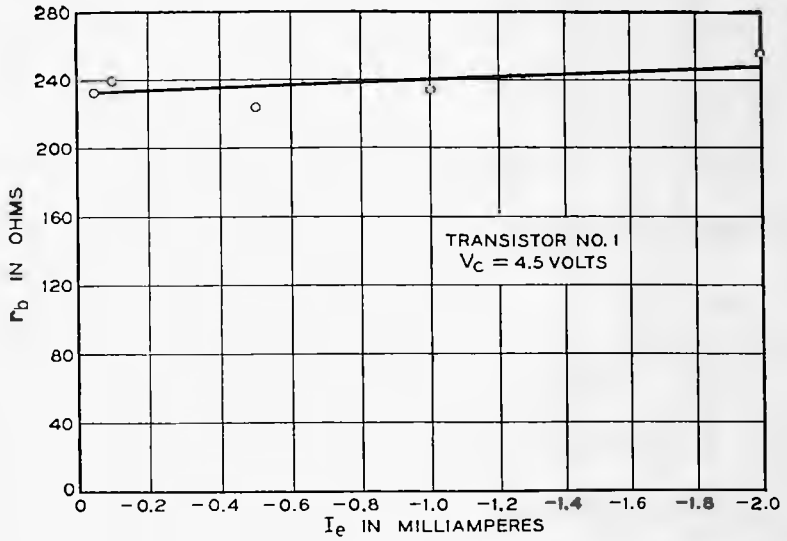


Fig. 9—Variation of base resistance with emitter current. Scatter of the data indicates that the measurements were not accurate.

is about 80° F., this reduces to

$$r_e = \frac{25.9}{I_e} \tag{2}$$

where I_e is measured in milliamperes. Within experiment error, values of r_e computed from this relation agree perfectly with the measured curve shown in Fig. 10.

Figure 11 introduces a new quantity, α, the current amplification factor of the transistor. This quantity is defined by the equation

$$\alpha = \frac{r_m + r_b}{r_c + r_b} \tag{3}$$

Since r_m and r_c are both very large compared to r_b, α is approximately equal to the ratio of r_m to r_c. It will be shown in a later section that this quantity is important in determining some of the circuit properties of the transistor and that many of the circuit properties become more desirable as α approaches unity.

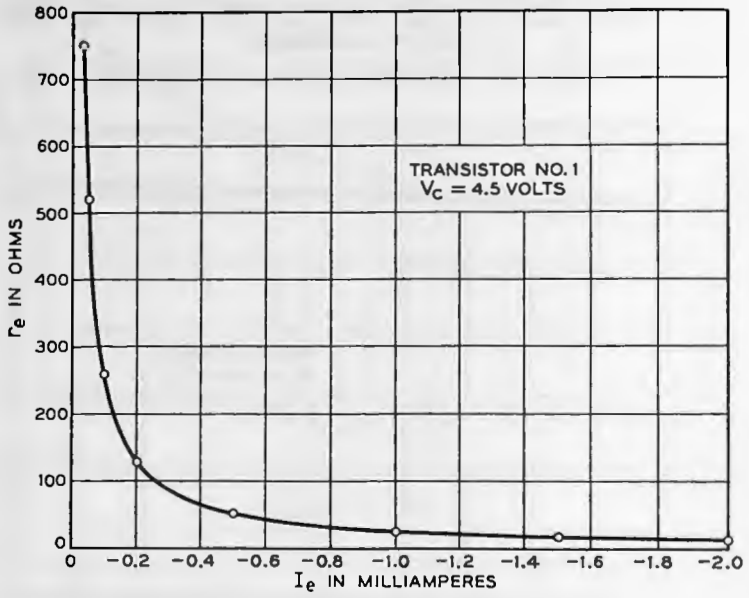


Fig. 10—The emitter resistance is inversely proportional to emitter current.

It can be seen from Fig. 11 that in this transistor α is approximately equal to 0.98 and that it increases slightly with increasing emitter current. The highest value of α so far achieved is 0.9965.

Those units which have been made in the laboratory so far show considerable variation in some of the properties, but this is partly due to the fact that changes have been made deliberately to test one aspect or another of Shockley's theory. The data in table I are presented to indicate what properties have been achieved to date. The collector capacitance C_c will be discussed in a later section.

General Considerations and Formulae

It is a consequence of the fact that α is always less than unity in this structure that these transistors are unconditionally stable with all termina-

tions. This means that stability considerations do not prevent working with matched terminations. Furthermore it is possible to obtain a variety of input and output impedances by connecting the transistor as a grounded-emitter, grounded-base, or grounded-collector stage. It is the purpose of this section to give some idea of the characteristics of these various stages and to show in each case at least one way of supplying the required biases and couplings to the stage.

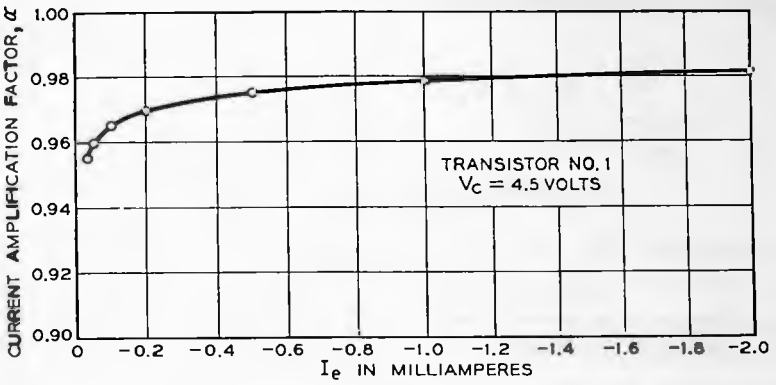


Fig. 11—The current amplification factor, α, increases slightly with increasing emitter current. Note the expanded scale for α.

TABLE I

CONSTANTS FOR VARIOUS TRANSISTORS MEASURED AT $V_c = 4.5$ v., $I_e = 1.0$ ma.

Transistor No.	I	II	III	IV	V
r_e (ohms)	25.9	31.6	33.1	30.2	38.8
r_b (ohms)	240	44	300	3070	180
r_c (megohms)	13.4	0.626	1.11	1.21	2.00
$r_c - r_m$ (megohms)	0.288	0.00387	0.0168	0.00422	0.0439
α	0.9785	0.9936	0.9848	0.9965	0.9780
C_c ($\mu\mu$f.)	7	7.7	18.9	27.9	21.2

It will be convenient to begin by writing down general relationships which will apply to all the possible connections. To this end let the transistor be represented by the box in Fig. 12. At low frequencies, the signal currents and voltages are related through the equations:

$$\begin{aligned} R_{11}i_1 + R_{12}i_2 &= v_1 \\ R_{21}i_1 + R_{22}i_2 &= v_2 \end{aligned} \tag{4}$$

If a generator of open circuit voltage v_g and internal resistance R_g is connected to the input terminals of the device as shown in Fig. 13, then

$$v_1 = v_g - i_1R_g \tag{5}$$

and if a load of resistance R_L is connected to the output terminals

$$v_2 = -R_L i_2 . \tag{6}$$

The equations for the circuit of Fig. 13 are, therefore,

$$\begin{aligned} (R_{11} + R_g) i_1 + R_{12} i_2 &= v_g \\ R_{21} i_1 + (R_{22} + R_L) i_2 &= 0 \end{aligned} \tag{7}$$

Solving for the voltage developed across the load ($= -R_L i_2$) gives

$$v_2 = \frac{R_L R_{21}}{(R_{11} + R_g)(R_{22} + R_L) - R_{12} R_{21}} v_g \tag{8}$$

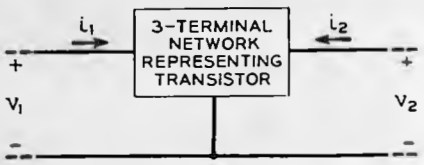


Fig. 12—A three-terminal network representing either grounded emitter, grounded base, or grounded collector connection of a transistor. Note the convention of signs.

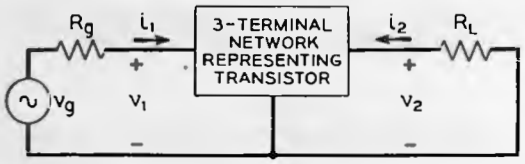


Fig. 13—The three-terminal network of Fig. 12 connected between a generator and a load.

The power gain in the circuit is the power delivered to the load (v_2^2/R_L) divided by the power available from the generator ($v_g^2/4R_g$). From equation (8), this gives

$$G = \frac{4R_g R_L R_{21}^2}{[(R_{11} + R_g)(R_{22} + R_L) - R_{12} R_{21}]^2} \tag{9}$$

The gain depends on R_g and R_L and will be maximum when these are chosen to match the input and output impedances of the transistor stage. But the input impedance depends on R_L and the output impedance depends on R_g in the following way:

$$\text{Input impedance} = R_i = R_{11} - \frac{R_{12} R_{21}}{R_{22} + R_L} \text{ and} \tag{10}$$

$$\text{Output impedance} = R_o = R_{22} - \frac{R_{12} R_{21}}{R_{11} + R_g} \tag{11}$$

If $R_i = R_g$ and $R_o = R_L$ then impedances are matched at the input and output terminals and the gain is a maximum. The conditions are:

Matched input impedance =

$$R_{im} = R_{11}\sqrt{1 - R_{12}R_{21}/R_{11}R_{22}}\,, \tag{12}$$

Matched output impedance =

$$R_{om} = R_{22}\sqrt{1 - R_{12}R_{21}/R_{11}R_{22}}\,, \tag{13}$$

Maximum available gain =

$$\text{M.A.G.} = \frac{R_{21}^2}{R_{11}R_{22}}\,\frac{1}{[1 + \sqrt{1 - R_{12}R_{21}/R_{11}R_{22}}]^2} \tag{14}$$

The Grounded Base Stage

In this and the following two sections we will put into equations (7) through (14) the appropriate 4-pole r's to obtain expressions for impedances and gains. As a numerical example we will substitute into the resulting equations the measured values of these r's for Transistor No. I working at V_c = 4.5v. and I_c = 1 ma. It must be understood that the numerical values may vary appreciably from transistor to transistor and that these numerical calculations are intended only for illustration and not as a basis for final circuit design. The numerical values to be used are

$$\begin{aligned} r_e &= 25.9 \text{ ohms} \\ r_b &= 240 \text{ ohms} \\ r_c &= 13.4\ (10)^6 \text{ ohms} \\ r_c - r_m &= 0.288\ (10)^6 \text{ ohms} \\ \alpha &= 0.9785 \end{aligned} \tag{15}$$

In this section it will be shown that the grounded base connection is suitable for working between a low impedance source and a high impedance load. The input impedance may be of the order of a hundred ohms and the output impedance of the order of one or more megohms. In this connection the current amplification is always less than unity but the voltage amplification may be very large indeed. Power gains of the order of 40 to 50 db can be obtained between matched impedances, and appreciable gains can still be obtained if the load resistance is reduced to a few thousand ohms (because the current gain is then almost equal to unity). In this case the gain of the stage is almost completely independent of those transistor properties

which tend to vary from unit to unit. This sort of stage does not produce a phase reversal.

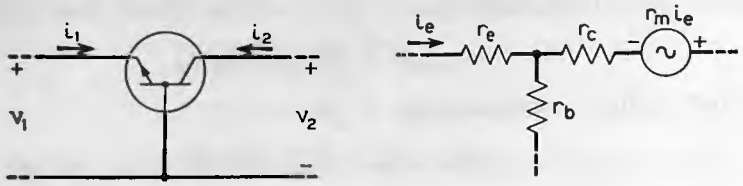


Fig. 14—The grounded base connection of a transistor.

For the grounded base stage shown in Fig. 14,

$$\begin{aligned} R_{11} &= r_e + r_b = 266 \text{ ohms} \\ R_{12} &= r_b = 240 \text{ ohms} \\ R_{21} &= r_m + r_b = 13.1\ (10)^6 \text{ ohms} \\ R_{22} &= r_c + r_b = 13.4\ (10)^6 \text{ ohms} \end{aligned} \tag{16}$$

$$\begin{aligned} \alpha &= \frac{r_m + r_b}{r_c + r_b} \\ &\doteq \frac{r_m}{r_c} \\ &= 0.9785 \end{aligned}$$

In this case if r_b is neglected by comparison with r_m and r_c, equation (8) leads to

$$v_2 = \frac{\alpha R_L v_g}{(r_e + r_b + R_g)(1 + R_L/r_c) - \alpha r_b}. \tag{17}$$

Since for these transistors $\frac{r_m}{r_c}$ $(\doteq \alpha)$ is always less than unity, the output voltage is in phase with the input voltage. Furthermore, if R_L is very high, the output voltage is enormous by comparison with the input voltage. For example, if $R_g = 0$ and R_L is infinite

$$v_2 = v_g \frac{r_m}{r_e + r_b} \tag{18}$$

and for the numerical example this is

$$v_2 = 4.93\ (10)^4\ v_g.$$

To achieve this step-up would require a load impedance very large compared to 13 megohms, but even with more modest values of load impedance the voltage step-up is large.

If R_L is small compared to r_c, the second of equations (7) leads to

$$\begin{aligned} i_2 &= -\frac{r_m}{r_c} i_1 \\ &\doteq -i_1 \end{aligned} \tag{19}$$

and the current delivered to the load is approximately equal to the current which the generator delivers to the transistor.

From equations (10) and (11), the input and output impedances are

$$R_i = r_e + r_b - \frac{r_b(r_m + r_b)}{r_c + R_L + r_b} \tag{20}$$

$$R_o = r_c + r_b - \frac{r_b(r_m + r_b)}{r_e + r_b + R_g} \tag{21}$$

As the load impedance varies from zero to infinity, the input impedance varies from

$$\begin{aligned} R_i &= r_e + r_b\left[1 - \frac{r_m + r_b}{r_c + r_b}\right] \quad \text{for } R_L = 0 \\ &\doteq r_e + r_b(1 - \alpha) \\ &= 31.1 \text{ ohm} \end{aligned} \tag{22}$$

to

$$R_i = r_e + r_b = 266 \text{ ohms} \quad \text{for } R_L = \infty. \tag{23}$$

When $R_g = 0$, the output impedance is

$$R_o = r_c - \frac{r_b}{r_e + r_b}(r_m - r_e) \tag{24}$$

$$\begin{aligned} &\doteq r_c - \frac{r_b}{r_e + r_b} r_m \\ &= 1.56\ (10)^6 \text{ ohms.} \end{aligned} \tag{25}$$

As R_g increases to infinity

$$R_o = r_c + r_b = 13.4\ (10)^6 \text{ ohms.} \tag{26}$$

From equations (12) and (13), the matched input and output impedances are approximately

$$R_{im} = (r_e + r_b)\sqrt{1 - \alpha r_b/(r_e + r_b)} = 91 \text{ ohms} \tag{27}$$

$$R_{om} = (r_c + r_b)\sqrt{1 - \alpha r_b/(r_e + r_b)} = 4.58(10)^6 \text{ ohms.} \tag{28}$$

With matched impedances, the maximum available gain is

$$\text{M.A.G.} = \frac{\alpha(r_m + r_b)}{r_e + r_b}\left[1 + \sqrt{1 - \alpha r_b/(r_e + r_b)}\right]^{-2} = 2.7\ (10)^4 \text{ or } 44.3 \text{ db.} \tag{29}$$

The matched output impedance of this stage is inconveniently high but a useful amount of gain can be maintained if R_L is reduced to a more reasonable value. For example, if $R_L = 200{,}000$ and $R_g = 25$, equation (9) gives

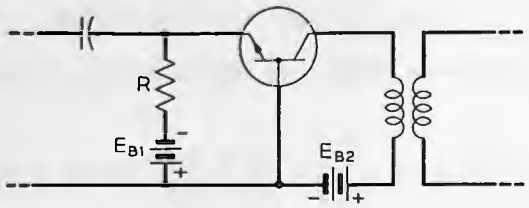


Fig. 15—One practical arrangement of a grounded base amplifier stage.

$$G = 5.3\ (10)^3 \text{ or } 37.2 \text{ db.}$$

If stages of this sort are to be cascaded, a step-down transformer must be used to couple each collector to the following emitter. Otherwise, since the current amplification factor of the transistor is slightly less than unity, the gain per stage will also be slightly less than unity.

One practical arrangement of a grounded base stage would be as shown in Fig. 15. The required value of R will be approximately

$$R = \frac{E_{B1}}{I_c} \tag{30}$$

where I_c is the desired collector current and E_{B1} is the voltage of the emitter-bias battery. For operating at $I_c = 1$ ma, for example, $E_{B1} = 1.5$ v and $R = 1500$ ohms would be suitable.

The Grounded Emitter Stage

For many applications the grounded emitter connection is more desirable than either of the other two. The power gains which can be obtained are high—of the order of 50 db—and the interstage coupling problem is simplified by the fact that the input impedance is somewhat higher than that of the grounded base stage while the output impedance is very much lower. The input impedance may be of the order of a few hundred ohms and the output impedance of the order of a few hundred thousand ohms. Both voltage and current amplification are produced (with a phase reversal) and gains of the order of 30 db or more per stage can be obtained without the

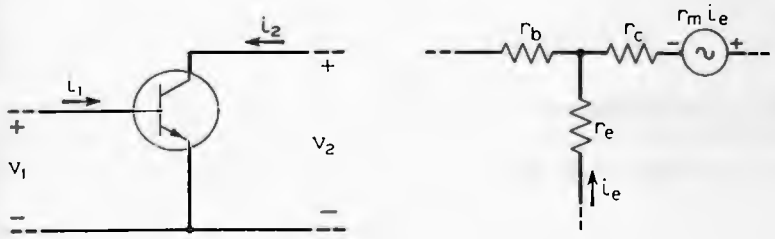


Fig. 16—The grounded emitter connection of a transistor and the equivalent circuit

use of interstage coupling transformers. The input and output impedance depend very critically on α and may vary appreciably from unit to unit

For this connection, which is indicated schematically in Fig. 16,

$$\begin{aligned} R_{11} &= r_e + r_b = 266 \text{ ohms} \\ R_{12} &= r_e = 25.9 \text{ ohms} \\ R_{21} &= r_c - r_m = -13.1\ (10)^6 \text{ ohms} \\ R_{22} &= r_e + r_c - r_m = 0.288\ (10)^6 \text{ ohms} \end{aligned} \tag{31}$$

Putting these values into equation (8) shows v_2 is always opposite in sign compared with v_g, that is, that the grounded emitter stage produces a phase reversal as does the grounded cathode vacuum tube.

If R_L is infinite and $R_g = 0$

$$\begin{aligned} v_2 &= v_g \frac{r_e - r_m}{r_e + r_b} \\ &= -4.93\ (10)^4\ v_g \end{aligned}$$

which is the same as for the grounded base stage. But if $R_L = 0$

$$i_2 = \frac{r_m - r_e}{r_e + r_c - r_m} i_1 \tag{32}$$

$$\doteq \frac{\alpha}{1 - \alpha} i_1 \tag{33}$$

$$= 45.5\, i_1.$$

Thus it is seen that the grounded emitter amplifier can produce quite appreciable current amplification—particularly so when α approaches unity.

The input impedance to the stage is

$$R_i = r_e + r_b + \frac{r_e(r_m - r_e)}{r_e + r_c - r_m + R_L}. \tag{34}$$

When $R_L = 0$ this reduces to

$$R_i = r_b + r_e \frac{1}{\frac{r_e}{r_c} + 1 - \frac{r_m}{r_c}} \tag{35}$$

$$\doteq r_b + r_e \frac{1}{1 - \alpha} \tag{36}$$

$$= 1440 \text{ ohms.}$$

As R_L increases to infinity, the input impedance decreases to $r_e + r_b$ which, for the numerical example, is 266 ohms.

The output impedance is

$$R_o = r_e + r_c - r_m + \frac{r_e(r_m - r_e)}{r_e + r_b + R_g} \tag{37}$$

When $R_g = 0$, this gives

$$R_o = r_c - \frac{r_b}{r_e + r_b}(r_m - r_e) \tag{38}$$

$$\doteq r_c \left[1 - \frac{r_b}{r_e + r_b} \alpha\right] \tag{39}$$

$$= 1.56\ (10)^6 \text{ ohms}$$

As R_g increases to infinity, R_o decreases to

$$R_o = r_e + r_c - r_m$$
$$= 0.288\ (10)^6 \text{ ohms.} \tag{40}$$

The matched input and output impedances are

$$R_{im} = (r_e + r_b)\sqrt{1 + r_e(r_m - r_e)/(r_c + r_b)(r_e + r_c - r_m)} \tag{41}$$

$$= 619 \text{ ohms and}$$

$$R_{om} = (r_e + r_c - r_m)\sqrt{1 + r_e(r_m - r_e)/(r_e + r_b)(r_e + r_c - r_m)} \tag{42}$$

$$= 0.671\ (10)^6 \text{ ohms}$$

As α increases toward unity the matched input impedance increases and the matched output impedance decreases. They approach the limits

$$R_{im} = \sqrt{(r_b + r_c)(r_e + r_b)} \tag{43}$$

$$R_{om} = r_c\sqrt{(r_b + r_e)/(r_e + r_b)} \tag{44}$$

$$\text{as } \alpha \to 1.$$

If r_m in the transistor of our numerical examples could be increased to exactly the value of r_c ($\alpha = 1$) then the matched impedances would be

$$R_{im} = 59{,}700 \text{ ohms}$$

$$R_{om} = 5{,}800 \text{ ohms}$$

From this example, it is seen that the impedances vary rapidly with α as α approaches unity.

With matched impedances, the maximum available gain from the grounded emitter stage is

$$\text{M.A.G.} = \frac{(r_e - r_m)^2}{r_e r_c}\left[\sqrt{\left(1 + \frac{r_b}{r_c}\right)\left(\frac{r_e}{r_c} + 1 - \frac{r_m}{r_c}\right)} + \sqrt{1 + \frac{r_b}{r_e}\left(\frac{r_e}{r_c} + 1 - \frac{r_m}{r_c}\right)}\right]^2 \tag{45}$$

$$= 2.02\ (10)^5 \text{ or } 53 \text{ db}$$

When α is exactly unity this expression reduces to r_e/r_c provided r_e and r_b are small compared with r_c. For values of α which are enough smaller than unity so that

$$\frac{r_e}{r_c} \ll 1 - \alpha$$

the expression for maximum available gain reduces to the approximate expression

$$\text{M.A.G.} = \alpha(r_m/r_e)\left[\sqrt{(1 - \alpha)\frac{r_b}{r_e}} + \sqrt{1 + (1 - \alpha)\frac{r_b}{r_e}}\right]^2 \tag{46}$$

From the above equations it can be seen that gain does not increase rapidly with α when α is sufficiently near unity. In the case of our numerical example, increasing α from 0.9785 to unity increases the gain by only 4.1 db. The gain of the stage is approximately proportional to r_c and inversely proportional to r_e and can therefore be increased by operating at higher emitter currents or by fabricating the transistor in such a way as to obtain higher values of r_c. In the latter case it would be desirable also to increase α in order to keep the output impedance from becoming unreasonably high.

It has been seen that, for the case of our numerical example, the matched output impedance is large compared to the input impedance (671,000 ohms compared to 619 ohms). This means that if the maximum available gain is to be obtained in cascaded stages, step-down interstage transformers must be used. But an appreciable amount of gain can be obtained without interstage impedance matching. This is because of the short-circuit current amplification previously mentioned which amounts approximately to

$$\frac{\alpha}{1-\alpha}$$

or 45.5 times in the case of our numerical example. For this transistor, then, the iterative gain per stage without impedance transformation would be 33.2 db. This gain increases very rapidly as α approaches unity, not only because the short-circuit current amplification increases, but also because the output impedance decreases and the input impedance increases so that a better interstage impedance match is obtained. From equations (41) and (42), it is seen that the matched input and output impedances are equal when

$$r_c - r_m = r_b$$

or when

$$1 - \alpha = \frac{r_b}{r_c + r_b}$$

In this case the gain per stage would be approximately r_c/r_e. This says that if r_m could be increased in the numerical example until

$$\alpha = 0.999821$$

the gain per stage (without impedance transformation) would be increased to 57.1 db. Values of α this near to unity have not even been approached in transistors made to date. This unrealistic example is included only to indicate one of the reasons for seeking to make α very near unity.

Consider next one possible way of supplying biases to a grounded emitter

stage. Suppose that the collector is connected through a transformer winding to a fixed voltage supply as indicated in Fig. 17. Since V_e is always a very small fraction of a volt (see Fig. 4 and Fig. 5) the collector voltage will be approximately equal to the supply voltage. If no d-c. connection is supplied to the base, it will float at a potential above ground equal (in magnitude) to V_e—i.e., a very small fraction of a volt—and the collector current will be exactly equal to the emitter current. To find out approximately what this value of the current will be, consider the upper set of static characteristics in Fig. 5. Note that, when the emitter current is zero, the collector current is of the order of 20 microamperes (the exact value varies from 1 to 30 microamperes in transistors tested so far). The collector current is then of the order of 20 microamperes greater than the emitter current. If the emitter current is now increased by ΔI_e, the collector current will increase by $\alpha \Delta I_e$. That is, the increments of collector current will be slightly smaller than the incre-

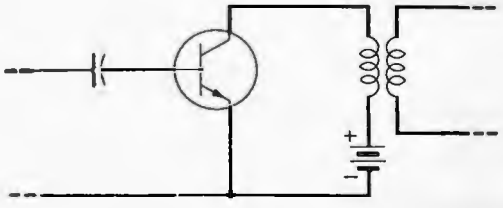


Fig. 17—One practical arrangement of a grounded emitter stage.

ments of emitter current and, as the emitter current is increased, the emitter and collector currents will become more nearly equal. If α were perfectly constant they would become exactly equal when

$$I_e = I_c = \frac{I_{co}}{1 - \alpha} \tag{47}$$

where I_{co} is the collector current which flows when the emitter current is zero. Equation (47), then, gives the value of emitter (and collector) current which will flow in a grounded emitter stage if no d-c. connection is made to the base. This current varies rapidly with α. For the transistor which has been considered numerically, this current would amount to about 465 microamperes which is certainly within the range of suitable values for the transistor. In certain low level applications, however, it might be desirable to work at a smaller current for the sake of decreasing battery power consumption. This requires that a small current be drawn out of the base. The required current is small because the collector current will decrease by $1/(1 - \alpha)$ microamperes for each microampere drawn from the base. One method of obtaining this base current is to provide a resistive path between

base and ground as shown, for example, in Fig. 18. Since the base floats at a positive potential with respect to ground, this circuit produces a base current of the right sign to decrease the collector current. As the value of the series resistance is decreased to zero, the collector current decreases to

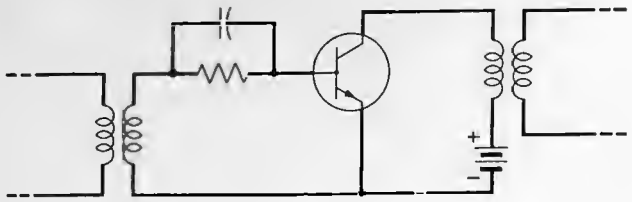


Fig. 18—Modification of Fig. 17 to obtain lower collector current.

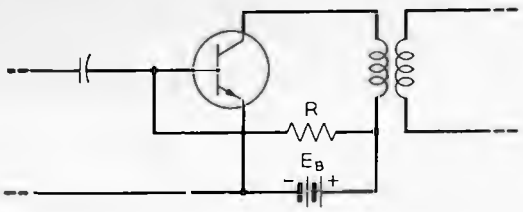


Fig. 19—Modification of Fig. 17 to obtain higher collector current.

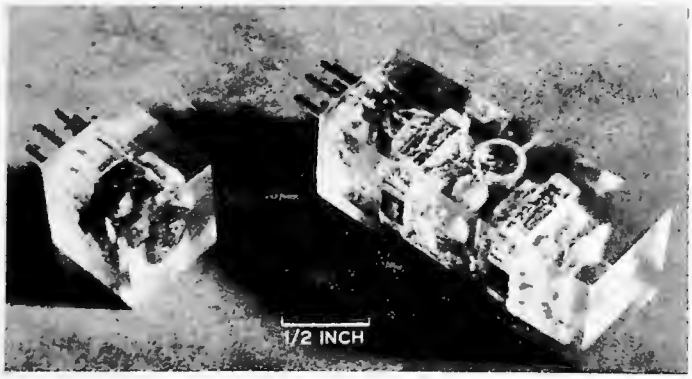


Fig. 20—A two-stage grounded emitter amplifier which produces approximately 90 db power gain is shown on the right and a micro-power audio oscillator is shown on the left.

a value corresponding to zero emitter voltage. A still further decrease in collector current can be obtained by inserting resistance between emitter and ground.

In order to increase the collector current to values higher than that corresponding to zero base current, a high resistance path between base and the positive supply voltage may be used as shown in Fig. 19. In this case the

collector current will increase by $1/(1 - \alpha)$ microamperes for each microampere which flows through the bias resistor. Since the current in the bias resistor will be approximately E_B/R, it is a simple matter to compute the required value of bias resistor once the desired collector current is known.

Figure 20 shows a two-stage audio amplifier which gives approximately 90 db gain. The circuit is shown in Fig. 21.

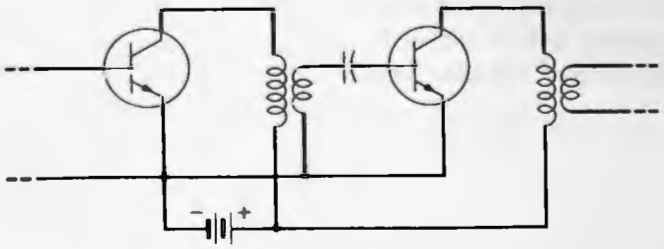


Fig. 21—Circuit of the amplifier shown in Fig. 20.

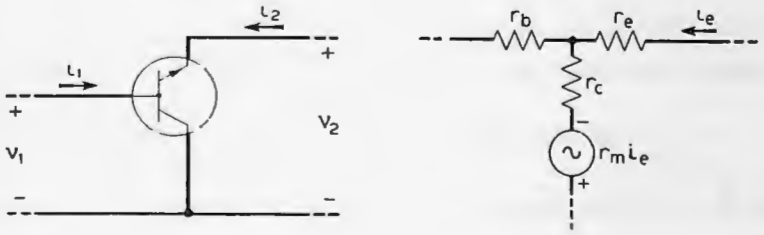


Fig. 22—The grounded collector connection of a transistor and the equivalent circuit.

The Grounded-Collector Stage

Although the power gain obtainable from this connection is relatively low—of the order of 15 or 20 db—it has very interesting possibilities in producing very high input impedances or very low output impedances. If it is worked into a fairly high load impedance, the input impedance may be several megohms, or if it is worked from a source of moderately low impedance (a few thousand ohms), the output impedance may be of the order of 25 ohms or lower.

For this type of stage, which is shown schematically in Fig. 22,

$$\begin{aligned} R_{11} &= r_b + r_c = 13.4\ (10)^6 \text{ ohms} \\ R_{12} &= r_c - r_m = 0.288\ (10)^6 \text{ ohms} \\ R_{21} &= r_c = 13.4\ (10)^6 \text{ ohms} \\ R_{22} &= r_e + r_c - r_m = 0.288\ (10)^6 \text{ ohms} \end{aligned} \tag{48}$$

If this stage is worked from a zero impedance generator into an infinite impedance load

$$v_2 = v_g \left(\frac{r_c}{r_b + r_c} \right) \tag{49}$$

$$\doteq v_g$$

and so, like a cathode follower, it gives an output voltage which is less than the input voltage, but in the same phase.

If the stage is operated into a short circuit

$$i_2 = -i_1 \frac{r_c}{r_b + r_c - r_m} \tag{50}$$

$$\doteq -i_1 \frac{1}{1 - \alpha} \tag{51}$$

$$= -46.5\, i_1$$

which indicates that the stage can give an appreciable current gain.

The input impedance is

$$R_i = r_b + r_e - \frac{r_c(r_c - r_m)}{r_c + r_c - r_m + R_L} \tag{52}$$

When $R_L = 0$, this reduces to

$$R_i = r_b + r_e \frac{1}{\frac{r_e}{r_c} + 1 - \frac{r_m}{r_c}} \tag{53}$$

$$\doteq r_b + r_e \frac{1}{1 - \alpha} \tag{54}$$

$$= 1445 \text{ ohms.}$$

When R_L is infinite

$$R_i = r_b + r_c \tag{55}$$

$$= 13.4(10)^6 \text{ ohms.}$$

With respect to input impedance, the grounded collector stage is again seen to be like a cathode follower in that the input impedance is high when the load impedance is high.

The output impedance is

$$R_o = r_e + r_c - r_m - \frac{r_c(r_c - r_m)}{r_b + r_c + R_g} \tag{56}$$

For $R_g = 0$, this reduces to

$$R_o = r_e + r_b \frac{r_c - r_m}{r_b + r_c} \tag{57}$$

$$\doteq r_c + r_b (1 - \alpha) \tag{58}$$

$$= 31.1 \text{ ohms.}$$

For R_g infinite

$$R_o = r_e + r_c - r_m \tag{59}$$

$$= 0.288(10)^6 \text{ ohms.}$$

The matched input impedance is

$$R_{im} = (r_b + r_c) \sqrt{\frac{r_b}{r_b + r_c} + \frac{r_c r_c}{(r_b + r_c)(r_e + r_c - r_m)}} \tag{60}$$

$$\doteq \sqrt{r_c[r_b + r_e/(1 - \alpha)]} \tag{61}$$

$$= 139{,}000 \text{ ohms.}$$

The matched output impedance is

$$R_{om} = (r_c + r_c - r_m) \sqrt{\frac{r_b}{r_b + r_c} + \frac{r_e r_c}{(r_b + r_c)(r_e + r_c - r_m)}} \tag{62}$$

$$\doteq (1 - \alpha) \sqrt{r_c[r_b + r_e/(1 - \alpha)]} \tag{63}$$

$$= 2990 \text{ ohms.}$$

With matched impedance, the maximum available gain of the grounded collector stage is

$$\text{M.A.G.} = \frac{\dfrac{r_c^2}{(r_b + r_c)(r_c + r_c - r_m)}}{\left[1 + \sqrt{\dfrac{r_b}{r_b + r_c} + \dfrac{r_c r_c}{(r_b + r_c)(r_c + r_c - r_m)}}\right]^2} \tag{64}$$

As α approaches unity, this approaches approximately

$$\text{M.A.G.} = r_c/4r_e \tag{65}$$

but so long as $r_e \ll r_c - r_m$, a good approximation is

$$\text{M.A.G.} = 1/(1 - \alpha) \tag{66}$$

$$= 46.5 \text{ or } 16.7 \text{ db.}$$

The considerations involved in supplying biases to a grounded collector stage are rather similar to those discussed already for the grounded emitter case. If the base is allowed to float, the collector current will be given approximately by equation (47) as discussed for the grounded emitter case. A resistance between base and the negative side of the supply battery in Fig. 24 will serve to decrease the collector current while a resistance between base and ground will serve to increase it. In applications where it is desired to make full use of the high input impedance which this stage can afford, it may be most desirable to let the base float as shown in Fig. 23.

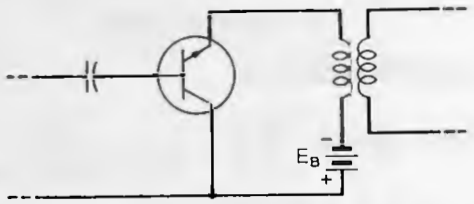


Fig. 23—One practical arrangement of a grounded collector stage.

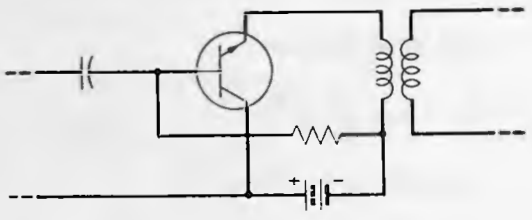


Fig. 24—Modification of Fig. 23 to obtain lower collector current. To raise collector current remove the resistance shown and connect a high resistance between base and ground.

Frequency Response—General Remarks

Shockley has shown that there are several different physical considerations which lead one to expect a high-frequency cutoff in the response of *n-p-n* transistors. The frequency at which cutoff occurs depends in a theoretically understandable way on such things as the geometry of the transistor and the physical properties of the germanium from which it is made. If these factors could all be controlled and varied at will, it would be possible to design a transistor to have a specified cutoff frequency.

One limitation comes about in the following way: In order to produce transistor action, the electrons which are injected into the *p* layer at the emitter junction must travel across this thin layer and arrive at the collector junction. They do this principally by a process of diffusion and require a finite (but small) amount of time to make the journey. If this time were

exactly the same for all electrons, the effect would be simply to delay the output signal with respect to the input and there would be no effect on frequency response. But there is a certain amount of dispersion in transit time which means that the electrons corresponding to a particlar part of the input signal wave do not all arrive simultaneously at the collector. When this difference in time of arrival amounts to an appreciable part of a cycle there is a tendency for some of the electrons to cancel the effect of others so that the frequency response begins to fall off. As the signal frequency increases beyond this point, the effect becomes more and more pronounced and the response continues to fall with increasing frequency.

In terms of the equivalent circuit, this dispersion in transit time means that beyond a certain frequency, r_m (and hence α) begins to decrease with increasing frequency and so the transistor may be said to have a certain α-cutoff which we will call $f_{c\alpha}$.

Shockley has shown that $f_{c\alpha}$ is inversely proportional to the square of the p-layer thickness and hence increases rapidly as the p layer is made thinner. For n-p-n transistors now available, this cutoff should occur at frequencies between five and twenty megacycles.

Another limitation on frequency response comes about from the fact that, at sufficiently high frequencies, the emitter junction fails to behave as a pure resistance and is, in effect, shunted by a capacitance. In terms of the equivalent circuit, this means that r_e is shunted by a capacitance.

The effect which this has on frequency response can be reduced by reducing the impedance of the source from which the emitter is driven. But so far as the emitter junction is concerned, r_b is always in series with the source impedance and so it is the value of r_b which ultimately determines the emitter cutoff frequency.

This capacitative reactance should begin to become appreciable with respect to emitter resistance at a frequency which may be of the same order as $f_{c\alpha}$. If r_b is high, the emitter cutoff frequency f_{ce} will then be of the same order of magnitude as $f_{c\alpha}$ and will increase as r_b is decreased.

A third cause for limited frequency response is the capacitance of the collector junction. The n-type germanium on one side of the junction behaves as one plate of a parallel-plate condenser and the p-type germanium on the other side behaves as the other plate. Since the transition from n to p type germanium may be made in an exceedingly small fraction of an inch, the plates of the condenser are very closely spaced and the capacitance may be appreciable.

Collector capacitance also depends on collector voltage, decreasing with increasing voltage. Theoretically, the capacitance should be in proportion to the negative one-third power of V_c.

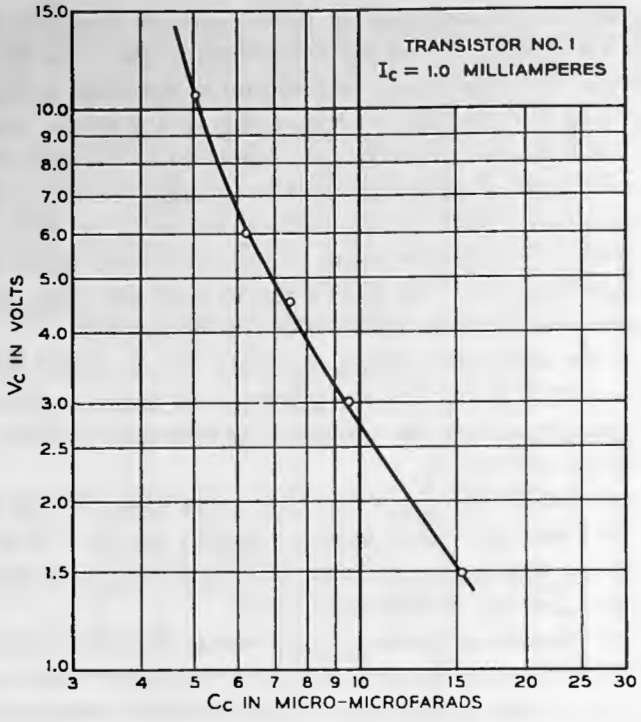


Fig. 25—Collector capacitance decreases as collector voltage is increased.

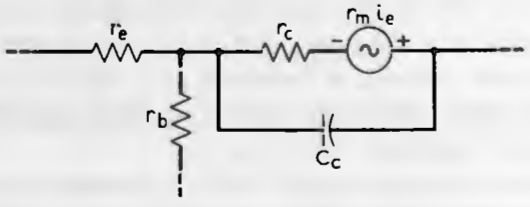


Fig. 26—The equivalent circuit of a transistor with collector capacitance shown.

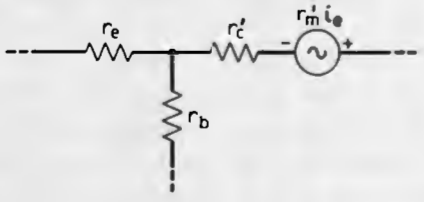


Fig. 27—The effect of collector capacitance is to change r_m and r_c to r_m' and r_c'. See equations (67) and (68).

Figure 25 shows measured values of C_c as a function of collector voltage. For reasons which are not understood at present, these data show a departure from the usual inverse one-third power variation. At $V_c = 4.5$ volts

the capacitance is seen to be approximately 7 micro-microfarads. In terms of the equivalent circuit, this capacitance is in shunt with the series combination of r_c and the generator, $r_m i_e$, as shown in Fig. 26. This can be shown to be equivalent to the circuit of Fig. 27 in which r_c has been replaced by

$$r_c' = r_c/(1 + jC_c r_c \omega) \tag{67}$$

and r_m has been replaced by

$$r_m' = r_m/(1 + jC_c r_c \omega). \tag{68}$$

The effect of collector capacitance can be computed by substituting r_c' and r_m' for the values of r_m and r_c (implicitly contained) in equation (8). In the sections which follow, this will be done for each of the three transistor connections and the resulting collector cutoff frequencies f_{cc} will be computed. It will be shown that at least for the transistor on which data are presented collector capacitance tends to produce a cutoff frequency well below those to be expected from emitter cutoff or alpha cutoff. For this reason, only collector cutoff will be considered.

Collector Cutoff in the Grounded Base Stage

If the values of r_c' and r_m' from equations (67) and (68) are substituted for r_c and r_m in equations (16) and the resulting values of the R's are substituted into equation (8), the result is

$$v_2/v_g = \frac{\alpha R_L}{(r_e + r_b + R_g)[1 + R_L(1 + j\omega C_c r_c)/r_c]} \tag{69}$$

The cutoff frequency f_{cc} is defined as the frequency at which the voltage across the load has dropped 3 db compared to its low-frequency value. This is the frequency at which the imaginary part of the denominator of (69) is equal to the real part. Solving for f_{cc} gives

$$f_{cc} = \frac{1}{2\pi C_c}\left[\frac{1}{R_L} + \frac{1}{r_c} - \frac{\alpha r_b}{R_L(r_e + r_b + R_g)}\right] \tag{70}$$

Substituting into this equation $C_c = 7(10)^{-12}$ farad, numerical values of the r's from (16), and the values $R_g = 91$ and $R_L = 4.58\ (10)^6$ ohms, corresponding to maximum available gain gives

$$f_{cc} = 3390 \text{ cps.}$$

With these terminations, the low-frequency gain is 44.3 db. If R_g and R_L are reduced to 25 and 200,000 ohms, respectively, f_{cc} is raised to 23,500 cps and the gain is lowered to 37.2 db. A further reduction of R_L to 20,000 ohms increases f_{cc} to 0.22 megacycles and reduces the gain to 27.8 db. This corre-

sponds to a gain-bandwidth product of $1.2(10)^8$ cps and shows that useful gain could be obtained at frequencies well above a megacycle, *provided* alpha and emitter cutoffs did not interfere.

Collector Cutoff in the Grounded-Emitter Stage

The procedure described in the last section leads, in this case, to

$$v_2/v_1 = \frac{-R_L r_m/r_c + (R_L r_c/r_c)(1 + jr_c C_c \omega)}{r_e + (r_b + R_g)(1 - r_m/r_c) + [r_e R_L/r_c + (r_b + R_g)(r_c + R_L)/r_c](1 + jr_c C_c \omega)} \tag{71}$$

For the transistor of our numerical example, the imaginary term in the numerator is completely negligible at frequencies below $(10)^9$ cps. Neglecting it leads to

$$f_{cc} = \frac{1}{2\pi C_c} \frac{1 + R_L/r_c + [(r_b + R_g)/r_c][1 - (r_m - r_e - R_L)/r_c]}{R_L + [(r_b + R_g)/r_c](r_e + R_L)} \tag{72}$$

In this case the values of R_g and R_L (619 ohms and 671,000 ohms respectively) which correspond to maximum available gain give

$$f_{cc} = 3740 \text{ cps and}$$

$$\text{M.A.G.} = 53 \text{ db.}$$

Reducing R_L to 100,000 and increasing R_g to 1000 ohms gives

$$f_{cc} = 11{,}120 \text{ cps}$$

$$G = 50 \text{ db.}$$

For $R_g = 1000$ and $R_L = 10{,}000$,

$$f_{cc} = 97{,}900 \text{ cps}$$

$$G = 41.3 \text{ db.}$$

and for $R_g = R_L = 1000$ ohms,

$$f_{cc} = 943{,}000 \text{ cps}$$

$$G = 31.4 \text{ db.}$$

The gain-bandwidth product for this stage is $1.3(10)^9$ cps as compared to $1.2(10)^8$ cps for the same transistor connected as a grounded base amplifier. It should be pointed out, however, that this stage is particularly sensitive to change in α and on this account alpha cutoff may influence the response at fairly low frequencies. For example, when the terminating resistances are both 1000 ohms, reducing α from 0.9785 to 0.900 reduces the gain from 31.4 db to 0.2 db.

Collector Cutoff in the Grounded Collector Stage

In this case

$$v_2/v_g = \frac{R_L}{[r_e + R_L + (r_b + R_g)(1 - r_m/r_c)] + (1/r_c)(r_b + R_g)(r_e + R_L)(1 + j\omega C_c r_c)} \quad (73)$$

and

$$f_{cc} = \frac{1}{2\pi C_c}\left[\frac{1}{r_c} + \frac{1}{r_b + R_g} + \frac{1 - r_m/r_c}{r_e + R_L}\right] \quad (74)$$

For matched impedances (R_g = 139,000 ohms and R_L = 2990 ohms),

$$f_{cc} = 320{,}000 \text{ cps}$$

$$G = 16.7 \text{ db.}$$

The cutoff frequency can be raised by decreasing either R_g or R_L. With R_g = 139,000 and R_L = 25 ohms

$$f_{cc} = 9.77 \text{ megacycles}$$

$$G = 1.8 \text{ db}$$

The gain-bandwidth product in this case is $1.5(10)^7$.

Noise

The data now available on noise are insufficient to give an adequate picture of the performance of n-p-n transistors in this respect. Such measurements as have been made, however, make it clear that these devices are very much quieter than early point-contact transistors reported on by Ryder and Kircher.

Transistor noise seems still to decrease with increasing frequency at a rate of something like 11 db per decade. It also decreases as the thickness of the p layer is decreased.

Of the order of half a dozen units of various dimensions have been measured at 1000 cps and have shown noise figures as low as 8 db and as high as 25 db.

The dependence of noise figure on operating point has been measured for only one transistor. As indicated in Fig. 28 and Fig. 29, these data show that the noise figure improves as V_c is reduced and that it may be roughly independent of collector current. These data were taken on a grounded emitter stage with impedance match at the input terminals. Noise figure for this connection varies slightly with source impedance and has been found

to be a minimum when the source impedance is roughly equal to the input impedance of the stage.

It must be emphasized that this functional dependence of noise figure on operating point and source impedance has been measured for only one transistor. Further measurements may show that these results are not typical.

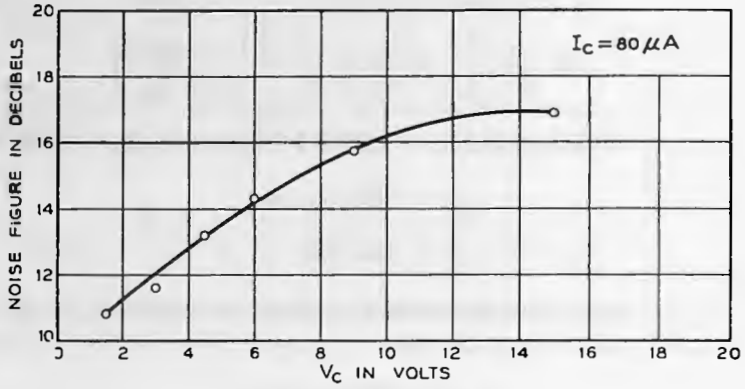


Fig. 28—Noise figure increases with increasing collector voltage.

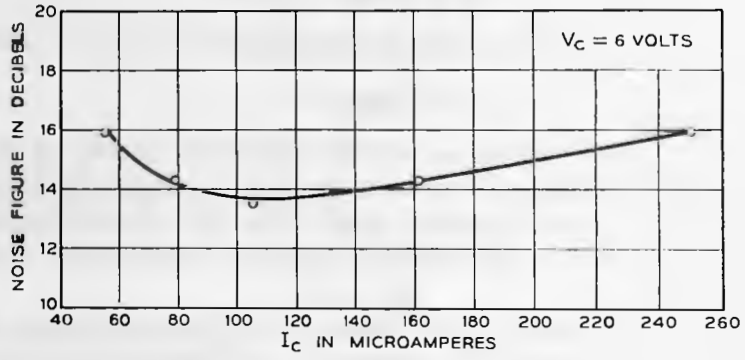


Fig. 29—Noise figure does not vary much with collector current.

Final Comments

In this paper we have attempted to present what is known about the circuit performance of *n-p-n* transistors. Since these devices are still undergoing exploratory development and since only a limited number has been produced, it is obviously impossible to give statistical data on reproducibility or on such reliability factors as the effect of ambient temperature.

It is much too soon to know what properties may be achieved after further development, but the results obtained to date seem encouraging and worth reporting.

Acknowledgment

The authors are happy to acknowledge their indebtedness to W. Shockley who was first to conceive the *n-p-n* transistor and has provided much of the inspiration and guidance which has made its physical realization possible. His comments have been of great help in the preparation of this paper.

We are also much indebted to J. A. Morton for his encouragement and helpful guidance and to M. Sparks for providing most of the transistors which have been studied. We wish to thank L. O. Schott, L. C. Geiger, and K. D. Smith for taking some of the data presented and to thank G. Raisbeck and L. G. Schimpf for proofreading and correcting the manuscript.

References

1. W. Shockley, "The Theory of *p-n* Junctions in Semiconductors and *p-n* Junction Transistors," *B. S. T. J.*, XXVIII, 435 (1949).
2. W. Shockley, "Electrons and Holes in Semiconductors," Van Nostrand (1950).
3. F. S. Goucher, G. L. Pearson, M. Sparks, G. K. Teal, W. Shockley, "Theory and Experiment for a Germanium *p-n* Junction," *Phys. Rev.*, *81*, 637 (1951).
4. W. Shockley, M. Sparks, G. K. Teal, "*p-n* Transistors," *Phys. Rev.*, *83*, 151 (1951).
5. R. M. Ryder, R. J. Kircher, "Some Circuit Aspects of the Transistor," *B. S. T. J.*, XXVIII, 367 (1949).

GENERAL CONSIDERATIONS CONCERNING NON-LINEAR CIRCUITS AND NEGATIVE RESISTANCE

B. G. FARLEY

Produced under contract W 36-039 sc-44497

GENERAL CONSIDERATIONS CONCERNING NON-LINEAR CIRCUITS AND NEGATIVE RESISTANCE

In general, it can be said that switching circuits make use of non-linear characteristics of particular devices. Numerous instances of such devices and circuits have been invented in the past. Examples are the non-linear characteristic of the armature motion in a relay, cutoff and saturation in a vacuum tube, sudden change in resistance of a gas tube upon ionization, and collector "cutoff" and saturation in a transistor.

Utilizing these non-linearities, many types of circuits can be built having useful properties which are not possible to obtain using only linear elements. Oscillators, multivibrators, counters, and trigger circuits are examples.

These circuits have generally been invented and designed "by themselves", for particular applications, and any mutual relationship not recognized until much later, or not at all. In other words, there exists no point of view generally available by which all, or many, of such non-linear applications can be seen in "perspective".

This situation is in contrast to that prevailing in linear circuit theory where there exists a theory giving both qualitative and quantitative design criteria on a wide variety of devices. For example, figures of merit are available to compare linear circuits in a quantitative way, criteria for cascading amplifiers are well known, and response to varying types of input can be accurately calculated. None of these advantages is available for non-linear circuits. Of course, there is a good reason for this. The non-linear differential equations governing the devices of interest are considerably more difficult to solve than linear ones. However, in developing a new field, that of transistor switching circuits, it would be very advantageous to have a design theory which would incorporate at least some of the good points of linear design theory, and allow comparison of transistor switching circuits with each other and with previously known circuits. It is believed that the common property of a broad class of switching functions, that of the oscillators or "trigger circuits", is that of negative resistance. While it may be difficult to prove this exclusively and a priori, the evidence seems to be preponderately in this direction as all such circuits, upon analysis, do have negative resistance. It is then proposed to

use the mechanism of negative resistance as a means to codify switching functions so as to delineate the design theory in such a fashion as to be applicable to all modes of manifestation be they negative grid tubes, gas tubes, transistors, thermistors, mechanical devices or what not. A goodly part of such a formulation is the appreciation of points which lead to satisfactory and practical devices which suggests a study of existing devices and circuits for an analytical determination of such features. However, a basic understanding of the first order, basic non-linear behavior is necessary before either analysis or synthesis can be accomplished. It is expected, and in part realized, that such a study will reveal controlling factors which relate to stability conditions, wave shape, speeds, sensitivity, gain, etc. When basic relationships are once established in terms of the common medium of a negative resistance characteristic it remains: (1) to interpret these factors to device in this case transistor, parameters, and (2) to devise circuits to yield such requirements. The actual devisement may require ingenuity and invention, but would have the decided advantages of an ordered directed procedure, definite objectives and a mechanism for quantitative comparison.

Accordingly, work has been started which it is hoped may lead toward such a goal. How far it is possible to go without running into unrewarding complexity, is of course not yet known. Complete quantitative results will probably always depend on computing machines. Nevertheless, it is believed that qualitative results can be achieved which will be highly useful, and also a qualitative understanding is important even when depending on machines, if effective use is to be made of them.

In part, lack of progress in this field is due to lack of dissemination of knowledge. Non-linear theory of differential equations has made considerable progress within the past 50 years, but applications to circuits have not become widely known because of the relative inaccessibility of the information, due to excessively mathematical formulation, or foreign publication. Only within the last year or two have any but specialized works on non-linear circuit applications been published in English, with the possible exception of Van der Pol's papers published in England in the late 1920's. Preliminary references are given in the bibliography at the end of this report.

In this report some preliminary, qualitative results on a negative resistance circuit will be discussed.

The circuit to be considered is that of Fig. 1. We assume that the rectangular box contains an element with an

"N"-type negative resistance, that is, it has the non-linear v'-i characteristic shown which is assumed to be independent of frequency. Straight lines have also been assumed for simplicity, although this will not affect the qualitative nature of the results which will apply to any single-valued function of i. It is an observed fact that all negative resistance devices have an internal parameter which is frequency sensitive and to the extent that the delay in a transistor is representable by an inductance (which is believed good for first approximation), this circuit is that often used with the emitter characteristic when there is a resistance in the base lead. (See Fourth Quarter Progress Report). Care should be taken to avoid confusion between v' and v.

We wish to understand the behavior of this circuit qualitatively as the parameters r, R, L and C vary.

In this and subsequent treatments mathematical manipulation will be kept to a minimum, only the final results being given in the text. Any further mathematics thought necessary will be relegated to an appendix.

To simplify matters, the rectangular box will be considered at first to contain a linear resistance of value r, which may be either positive or negative. It may also contain a dc term equivalent to a battery, (i.e. its v'-i line need not go through the origin), but the measurement of v and i (across the condenser and through the inductance respectively) will be chosen in such a way as to eliminate any constant term of r as well as that due to the battery voltage V. In other words, the origin of the v-i coordinates will be chosen at the point of intersection of the v-i lines of R and r.

With this convention, the basic differential equations of the circuit become: (see Appendix)

$$\frac{dv}{dt} = -\frac{1}{RC}(v + iR) \tag{1}$$

$$\frac{di}{dt} = \frac{1}{L}(v - ir) \tag{2}$$

$$\frac{dv}{di} = -\frac{L}{RC} \cdot \frac{\frac{v}{i} + R}{\frac{v}{i} - r} \tag{3}$$

$$\frac{d^2v}{dt^2} + \left(\frac{1}{RC} + \frac{r}{L}\right)\frac{dv}{dt} + \frac{1}{LC}\left(1 + \frac{r}{R}\right)v = 0 \tag{4}$$

The first two come directly from the circuit after eliminating V and any constant term in r as described above. Equation (3) results from dividing (1) and (2), and (4) results from eliminating i between (1) and (2). All the information we need is contained in these equations.

To find the circuit behavior, we make use of the v-i plane and find what happens to v and i if they are initially at v_o and i_o. Equation (3) gives this information, because it is the differential equation of the paths in the v-i plane, and it is a first order one, so one initial condition is enough to determine a path for all time "before" and after. Also, only one path goes through any ordinary point of the plane. (Ordinary here means that $\frac{dv}{di} \neq \infty$. In this case only the origin is not ordinary; it is called singular.) Now it is possible to solve equation (3) by standard methods, since it is a homogeneous equation of the first degree, however the solution is not what we want, but a general method. The method to be presented is known as the method of isoclines (curves of constant slope). These are the lines given by putting equation (3) equal to a constant ($\frac{dv}{di}$ = const.). In this case, we have a particularly simple situation. (See Fig. 31) It can be seen from (3) that if v/i = const. so is $\frac{dv}{di}$. Therefore, straight lines through the origin are isoclines. In particular, the slope of the solutions (paths) is zero when v = -Ri and infinite when v = ri as seen from (3). We can therefore take as many lines through the origin as we please, compute the slope and draw short segments of this slope at intervals along the appropriate line. The direction of motion is thus determined at every point of the plane. The sense at any point remains to be found. It follows directly from (1) and (2). If, for example, at a given point (v,i), $v + iR > 0$ and $v - iR > 0$ also, then (1) shows that v is decreasing with time while i is increasing. Thus arrows can be placed on the direction lines to show sense.

In this way a vector field diagram can be drawn, and if the isoclines are close enough, the nature of the motion can be seen directly. For our case, a simplified drawing is shown in Figures 2 and 3. Fig. 2 shows the case when r is negative; R is the "load" line. In this case, any (v,i) point moves on an outward spiral from the point of intersection of r and the load line, that is, the origin is unstable. A typical spiral is shown corresponding to particular initial conditions. Of course, if r remained linear, it would spiral out forever, however, it will eventually run out into a region like that shown in Fig. 3. In this case, r is positive, and the (v,i) point now moves in a decreasing spiral toward the origin, that is, the origin is stable

A point which is the center of spirals like Figs. 2 or 3 is known as an unstable or stable focus, as the case may be. This means that the characteristic roots of (4) are complex, corresponding to exponentially increasing or decreasing oscillations. There are other types of centers, depending of course on the magnitudes of r, R, L and C, which determine the nature of the solutions of (4). However, this point will not be pursued further at this time, except to say that the speed of response of the circuit will depend on the nature of the center.

Now in Fig. 4, the two types of centers shown in Figs. 2 and 3 are combined to show the behavior of the circuit with N-type negative resistance. The v-i plane is divided into 3 regions by the vertical dotted lines. In the middle region, $r > 0$, so Fig. 2 holds with the middle heavy point as unstable center, (pt. #1), whereas in the outer regions Fig. 3 applies, with the outside heavy points (where "R" meets "r" extended in the Fig.) as stable centers (pts. #2 and 3).

The beginning of a path naturally depends on the initial switching arrangement. If the system is initially near pt. 1, it spirals outward until it crosses a dotted vertical line. It then tends to spiral in toward pt. 2 or 3, but never reaches it, because it crosses a dotted line first when it then again has pt. #1 as unstable center.

However, if it starts at a great distance from pt. 1, the general trend is inward because it spends so much time trying to get near pts. 2 and 3. Therefore, there is some path which is closed, toward which both outward and inward spirals tend. This is called a limit cycle, and is the path of steady oscillations reached by this type of circuit. (See Appendix) It can be considered as that path for which the outward and inward spiralling tendencies just cancel. Note that this applies equally well to any oscillator depending on this type of negative resistance, and indeed is characteristic of many types of oscillators.

It is to be noted that as L approaches zero, (3) shows that the slope of the paths approaches zero everywhere in the plane except where $\frac{v}{I} - r = 0$, that is, along the "r" line. It can be shown that the point then tends to move along the "r" line, so that the oscillation tends to snap horizontally from the "peak" to the far "r" line, down it to the "valley" point, horizontally back to the original "r" line, etc. Thus, when L is small, the oscillation is a "relaxation" oscillation whereas when L is larger, the limit cycle rounds out until when it is oval, the oscillation is nearly sinusoidal. Thus, there is a continuous transition between relaxation and "sinusoidal" oscillations.

If the "R" line were such that pts. 2 and 3 were "real" instead of "virtual", that is the "r" line proper crossing "R", the system would come to rest at pt. 2 or 3 and would remain there until "triggered".

Preliminary Bibliography

I. Books

A-1 Andronow, A. A. and Chaikin, C. E.

Theory of Oscillations. Princeton 1949

M-1 Minorsky, N. Introduction to Non-Linear

Machanics. Edwards 1947

II. Papers

R-1 Reichardt, W. Uniform Relationship between Sinusoids, Relaxation Vibrations, and Relaxation Discontinuities (in German)

Elektrische Nachrichten - Technik.

V. 20, No. 9, Sept. 1943, and V. 20, No. 3, March 1943

R-2 Reichardt, W. "Negative Resistances, Their Characteristics and Effects. Sinusoids, Relaxation Oscillations and Relaxation Discontinuities." (In German)

Elektrische Nachrichten - Technik.

V. 20, No. 3, March 1943

V-1 Van der Pol, B. "Relaxation Oscillations".

Phil. Mag. Vol. 2, p. 978 (1926)

APPENDIX

Derivation of equations (1) and (2)

From Fig. 30

$$V = v + IR, \text{ I is current in R} \tag{A-1}$$

$$v = \Psi(i) + L\frac{di}{dt}, \text{ where } v' = \Psi(i) \tag{A-2}$$

$$\text{Also, since } v = \frac{q}{c}, \text{ I condenser} = C\frac{dv}{dt} \tag{A-3}$$

and,
$$I = C\frac{dv}{dt} + i \tag{A-4}$$

by (A-4) and (A-1)

$$V = v + \left(C\frac{dv}{dt} + i\right) R$$

or
$$RC\frac{dv}{dt} = V - v - iR \tag{A-5}$$

$$L\frac{di}{dt} = v - \Psi(i) \text{ by (A-2)} \tag{A-6}$$

Now, assume $\Psi = ri + b$ (A-7)

and make the transformation

$$v = \bar{v} + \frac{Vr + Rb}{r + R} \tag{A-8}$$

$$i = \bar{I} + \frac{V - b}{r + R} \tag{A-9}$$

then (5) becomes after a short calculation

$$\frac{d\bar{v}}{dt} = -\frac{1}{RC}\left(v + \bar{I}R\right)$$

while (6) becomes

$$\frac{d\bar{I}}{dt} = \frac{1}{L}\left(\bar{v} - \bar{I}r\right)$$

which are the desired equations. The bars are omitted in the text for convenience.

Note in regard to limit cycles. Other types of limit cycles are possible in more complicated systems. In fact, it is possible to have more than one limit cycle in the same system, or limit cycles plus stable points. Mathematically, a discussion of existence and position of limit cycles in general is quite complex and will not be discussed here, since the existence of a limit cycle in the simple case discussed is evident from physical considerations.

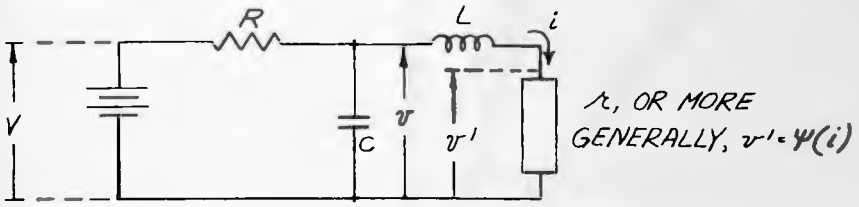


Fig. 1 RLC Circuit with N-Type Negative Resistance

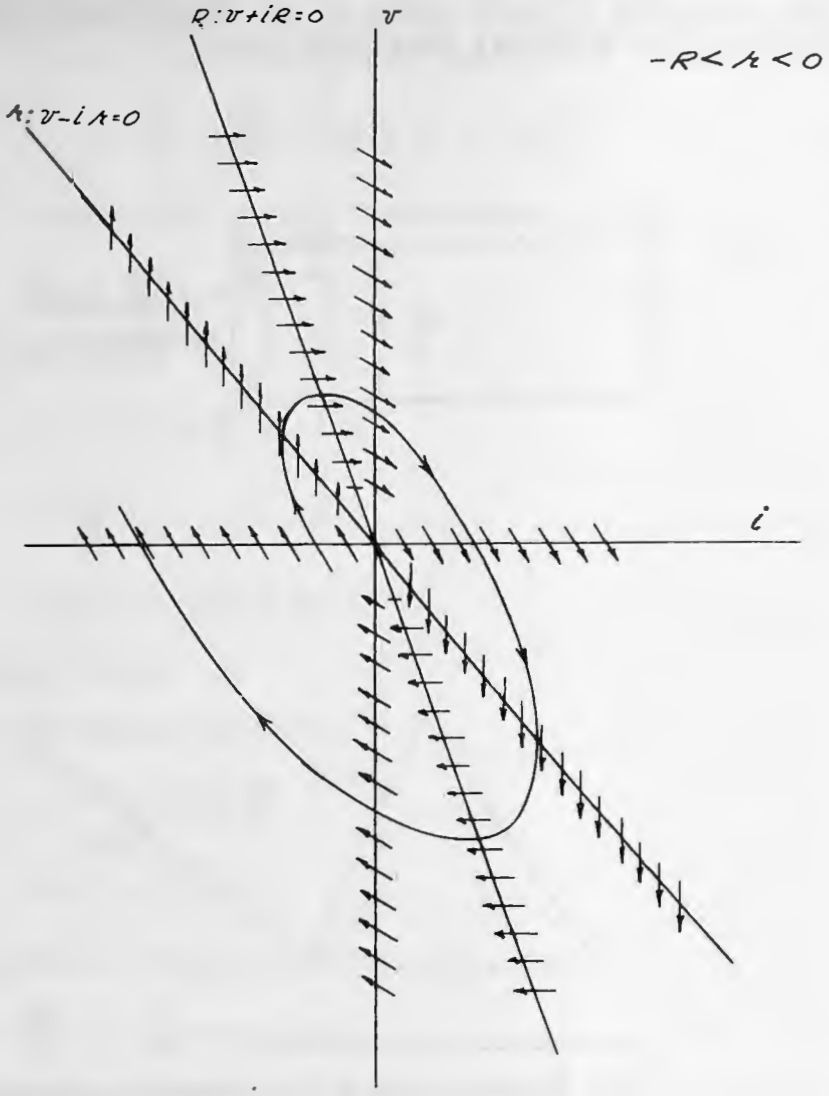


Fig. 2 Unstable Focus with r Negative

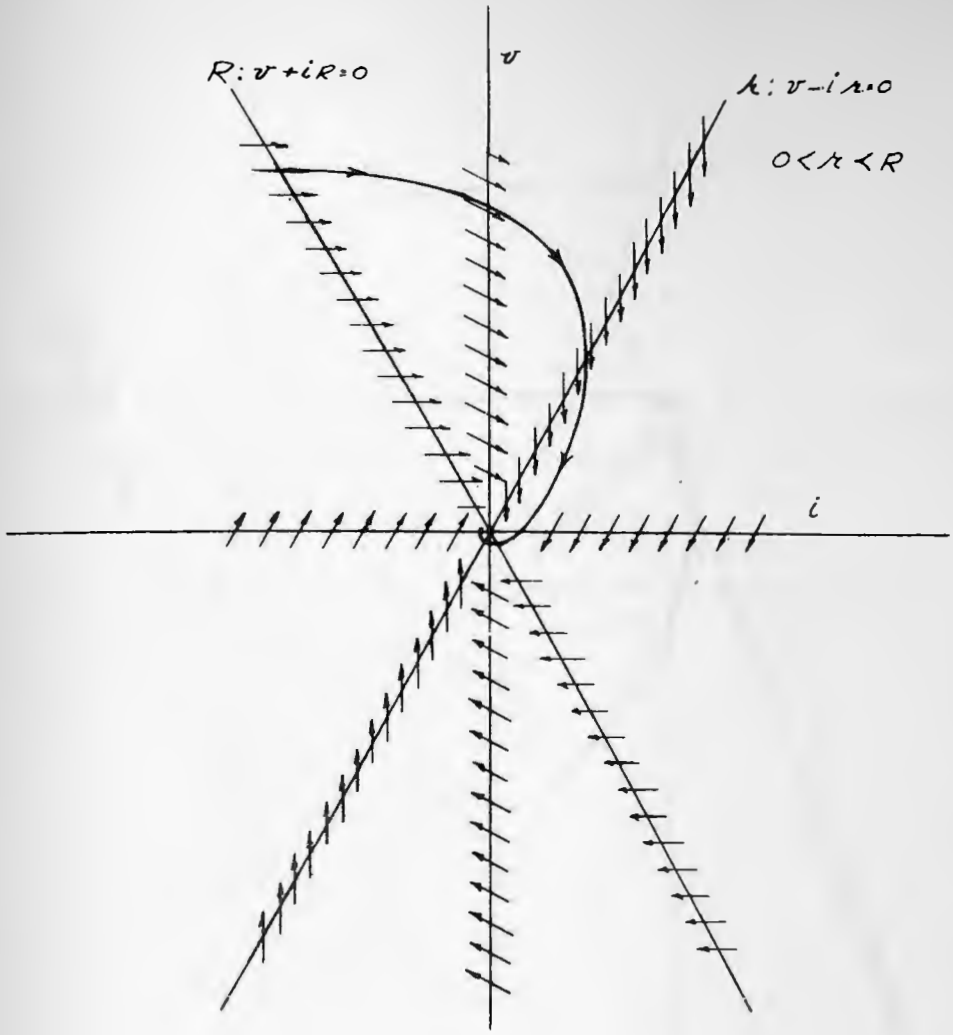


Fig. 3 Stable Focus with r Positive

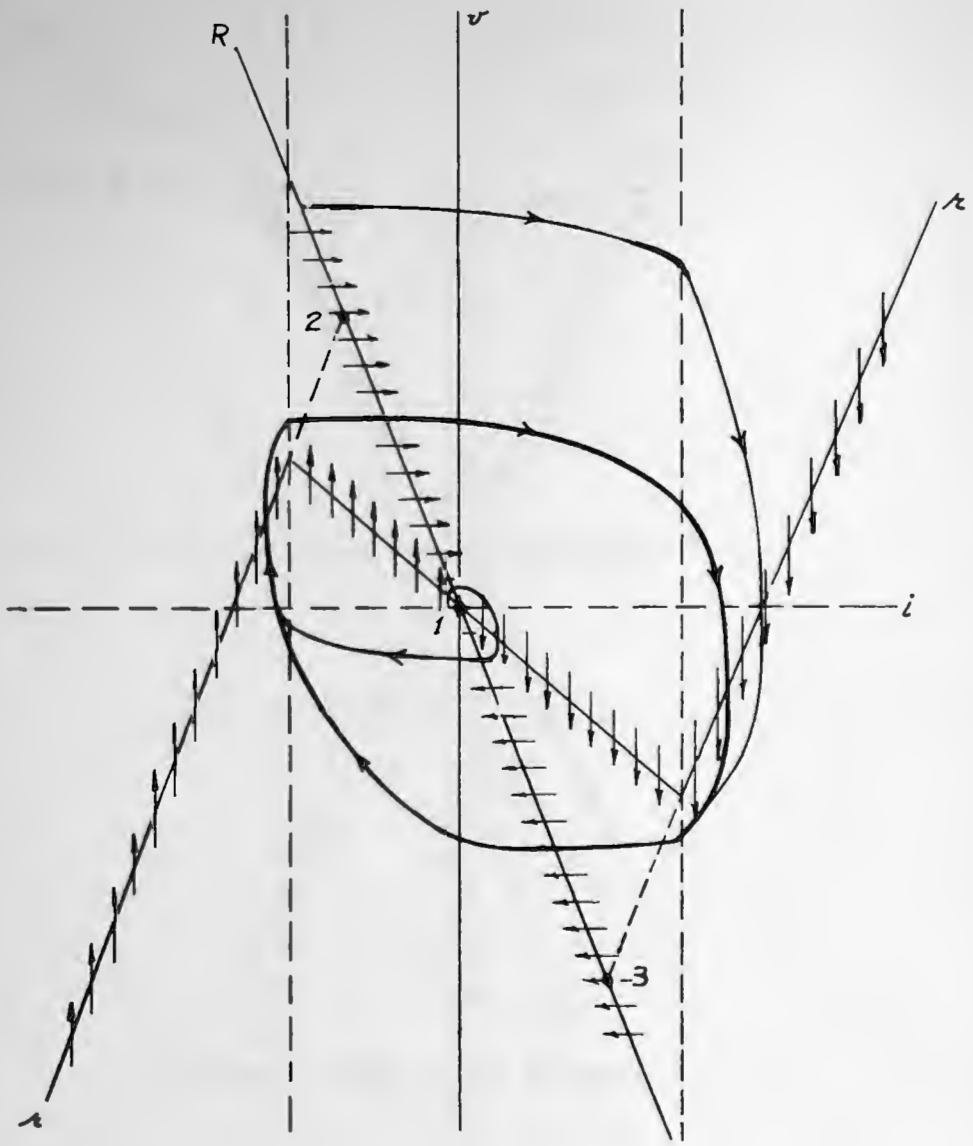


Fig. 4 N-Type Negative Resistance Oscillation

IDEALIZED NEGATIVE RESISTANCE CHARACTERISTICS OF THE TRANSISTOR

R. L. TRENT

Produced under contract W 36-039 SC-44497

IDEALIZED NEGATIVE RESISTANCE CHARACTERISTICS OF THE TRANSISTOR

Abstract

Circuital methods by which the negative resistance characteristics of the transistor may be modified to approach "idealized" N or S type (open circuit or short circuit stable) characteristics are discussed. Stabilization of the turning point in the characteristic is also covered. In various large signal switching applications, such as binary counters, single-shot multivibrators, etc., the triggering action from one state to another may be explained in terms of a negative resistance characteristic. For many of these applications, the normal asymmetric characteristic may place undesirable triggering requirements upon the circuit designer, and the purpose of this memorandum is to illustrate circuital methods by which the negative resistance curve for the transistor may be modified to approach "idealized" N or S type (open circuit or short circuit stable) characteristics, as illustrated in Figure 1.

Each of the characteristics, emitter, collector and base will be analyzed, and the methods illustrated for each case, together with experimental verification.

General

The input impedance of any linear active four-pole is given by:

$$Z_1' = Z_{11} - \frac{Z_{12}Z_{21}}{Z_{22}} \qquad (1)$$

where the Z's are the usual open circuit loop impedances, valid for linear characteristics only. For "large signal" switching application analysis we may suitably sectionalize the device characteristic such that each section of the curve may be considered approximately linear[3]. Figure 2 illustrates such a division

(3)This treatment is due to B. G. Farley

ffor a typical emitter (V'_ε vs I_ε) characteristic into Regions I, III and III, with each region having its own set of device parameters as listed in Table A. In all of the following analysis, lower case impedances or resistances refer to device parameters and capitals to external circuit elements.

II.0 Emitter Characteristic (V'_ε vs I_ε)

1.1 Impedance Relationships

Region I

From equation (1) the impedance or slope may be written for the typical circuit configuration shown in Figure 2:

$$R'_1 = r_{11} + R_B - \frac{(r_{12}+R_B)(r_{21}+R_B)}{r_{22} + R_L + R_B} = r_{11} + \frac{R_B(r_{22}+R_L)}{r_{22}+R_L+R_B} \qquad (2)$$

For typical values of R_B = 10K, R_L = 2K, the input resistance may be calculated as:

$$R'_1 \doteq 107{,}000 \text{ ohm}$$

By inspection it may be seen that the slope is determined principally by r_{11} in this region, and that as long as $r_{11} > 70{,}000$ ohms, the other factors will amount to less than 10%.

Region II

Using device parameter values in accordance with Table A, the impedance relationship may be simplified to:

$$R'_1 \doteq \frac{R_B(r_{22}+R_L-r_{21})}{r_{22}+R_L+R_B} \qquad (3)$$

For the same typical values of external circuit elements:

$$R'_1 \doteq -\ 8700 \text{ ohms.}$$

This is the negative resistance portion of the characteristic.

Region III

$$R_1' \approx \frac{R_B(r_{22}+R_L)}{r_{22}+R_L+R_B} \qquad (4)$$

$$R_1' \approx + 2300 \text{ ohms}$$

An examination of the composite characteristic in view of the input resistances obtained by the above analysis shows that it approaches the "ideal" unsymmetrical voltage to current negative resistance of Figure 1. It is proposed to modify the characteristic so that it approaches the voltage to voltage type of negative resistance. Three circuit modifications by means of which the resistance of Region III may be increased will now be illustrated and discussed.

1.20 Circuit and Modifications

1.21 Method 1

The first method is to attempt to limit the collector current when the desired turnover point is attained, equivalent to raising R_L to a high value. If a diode, properly biased so that it is normally conducting except when the collector current exceeds a certain value, is inserted in series with the load resistor R_L as shown in Figure 3, the characteristic is modified as illustrated. The turning points x', x'', x''' may be chosen at will by suitably varying the biasing current I_{RV_1}, and the slope of the line x' - x', etc., is given by the expression:

$$R_1' \approx \frac{R_B(r_{22}+R_L'-r_{21})}{R_B+R'_L+r_{22}} \qquad (5)$$

where R'_L represents the entire external collector circuit resistance. Since the diode will exhibit a very high resistance in the order of 0.5 megohm, the effective value of the resistance in series with R_L is determined largely by the amount of voltage available for biasing, setting R_{R_1}. Calculating the new slope for R_{R_1} = 60K:

$$R_1' = 8600 \text{ ohms}$$

indicating an increase by a factor of 3 to 4. The impedance is of course greatly increased in the region near the turning point, since the normal characteristic exhibits a considerable region of zero slope at this point.

1.22 Method 2

The next step is to make use of the fact that we have limited the collector current by the above device and that, referring to Figure 4:

$$I_b = I_c - I_\varepsilon \tag{6}$$

Thus when the varistor RV_1 operates to limit the collector current at the turning point x', the decreased collector current with respect to the emitter current allows the base varistor RV_2 to operate in its reverse direction, and to further increase the slope. The impedance is now determined by the expression:

$$R_1' = \frac{R'_B(r_{22}+R'_L-r_{21})}{r_{22}+R'_B+R'_L} \tag{7}$$

where the R'_B is now the entire external resistance in the base circuit. Again the value for R_{R_2} is determined by the voltage supply available for biasing, and

$$R'_B \doteq R_B + R_{R_2} \tag{8}$$

The shunting resistance of the diode is neglected here since it is an order of magnitude greater than R_{R_2}. Calculating the new slope for R_{R_2} = 56,000 ohms:

$$R_1' \doteq 32{,}000 \text{ ohms.}$$

Experimental verification of methods 1 and 2 has been made and the slopes obtained agreed with the analysis within experimental accuracy of slope determination.

1.23 Method 3

In many cases, the impedances obtained by using the foregoing methods will be sufficiently high to match the impedance obtained in Region I. However, it is possible that still higher values would be desired. Figure 5 illustrates the circuit arrangement for placing a diode in the emitter circuit, suitably biased so that it is normally conducting until the desired turnover point is attained, at which time the diode becomes a high resistance and limits the input current to the device. An analytical derivation of the impedance relationships in the three regions of interest for this circuit is made in Appendix A of this memorandum.

Region I

$$R_1' \approx \frac{R_1 R_2 (r_{11}+R_B)}{R_1 R_2 + (R_1+R_2)(r_{11}+R_B)} \qquad (9)$$

The following values for R_1 and R_2 were used in a test setup:

$$R_1 = 0.2M$$

$$R_2 = 0.36M$$

Substituting:

$$R_1' = 59{,}000 \text{ ohms.}$$

Region II

$$R_1' \approx \frac{R_1 R_2 R_B}{R_1 R_B + R_2 R_B + R_1 R_2 \left(\frac{r_{22}+R_L+R_B}{r_{22}+R_L+r_{21}}\right)} \qquad (10)$$

Substituting:

$$R_1' \approx -9400 \text{ ohms}$$

Region III

$$R_1' \approx \frac{R_1R_2R_4}{R_1R_2+R_2R_4} \qquad (11)$$

Substituting:

$$R_1' \approx 67{,}000 \text{ ohms.}$$

It is obvious that values for the circuit components may be so chosen that equal slopes in the positive resistance regions may be achieved; also that control over the slope in Region III may be effected by a suitable shunting resistance across the diode. It may therefore, be concluded that by the general method of employing biased diodes to modify the V_ε'-I_ε negative resistance characteristic of the transistor, a range of impedance values in Region III may be obtained. The exact method, or combination of methods, will depend upon the circuit configuration under consideration.

2.0 Collector Characteristic (V_c' vs I_c)

2.1 Impedance Relationships

Region I

From equation (1), the impedance or slope may be written for the typical circuit configuration as shown in Figure 6.

$$k_1' \approx r_{22}+R_B - \frac{(r_{12}+R_B)(r_{21}+R_B)}{r_{11}+R_L+R_B} \approx r_{22} + \frac{R_B(r_{11}+R_L-r_{21})}{r_{11}+R_L+R_B} \qquad (12)$$

Substituting

$$R_1' \approx 29{,}100 \text{ ohms}$$

It may be seen that the slope is determined principally by r_{22} and R_B in this region, and that as long as r_{11} is large, the other factors will amount to less than 3%.

Region II

Using device parameter values in accordance with Table A of Figure 2, the impedance relationship may be simplified to:

$$R_1' \approx r_{22} + \frac{R_B(R_L-r_{21})}{R_L+R_B} \qquad (13)$$

Substituting:

$$R_1' \doteq -20{,}000 \text{ ohms}$$

This is the negative resistance region of the characteristic.

Region III

$$R_1' \doteq r_{22} + \frac{R_B R_L}{R_B + R_L} \qquad (14)$$

Substituting:

$$R_1' \doteq 2800 \text{ ohms}$$

Therefore, the composite characteristic approaches the ideal current-to-voltage negative resistance of Figure 1. It is proposed to modify the characteristic so that it approaches the voltage-to-voltage type of negative resistance. Circuit modification by means of which the resistance in Region III may be increased will now be illustrated and discussed.

2.2 Circuit and Modifications

2.21 Method 1

As for the emitter characteristic, we will attempt to limit the load current (in this case the emitter current) when the desired turnover point is attained, equivalent to raising R_L to a high value. Inserting a properly biased diode in the emitter lead, so that it is normally conducting except when the emitter current exceeds a certain value, as shown in Figure 7, modifies the characteristic as illustrated. The slope of the lines x'x', etc. is given by the following expression and the intercept point may be chosen at will by varying the biasing current I_{RV_1}:

$$R_1' \doteq r_{22} + \frac{R_B R'_L}{R_B + R'_L} \qquad (15)$$

where R'_L represents the entire emitter circuit resistance. Again we may neglect the shunting effect of the diode due to its high back resistance. Calculating the new slope for R_{R_1} = 60K:

$$R_1' \doteq 9600 \text{ ohms}$$

indicating an increase by a factor of 3 to 4.

2.22 Method 2

We again may make use of the fact that the load current is limited by the above device, and introduce a biased diode in the base lead to further increase the slope. The impedance is now determined by the following expression:

$$R_L' \approx r_{22} + \frac{R_B' R_L'}{R_L' + R_L'} \qquad (16)$$

where R_B' is now the entire external resistance in the base circuit. For R_{R_2} = 56000 ohm:

$$R_1' \approx 32000 \text{ ohms}$$

neglecting the shunting effect of the diode resistance. Experimental verification of methods 1 and 2 have been made and the slopes obtained agreed with the analysis within experimental accuracy.

2.23 Method 3

In a manner analogous to the emitter case, Figure 9 illustrates the circuit arrangement for placing a diode in the collector circuit suitably biased so that it is normally conducting until the desired turnover-point is attained, at which time the diode becomes a high resistance and limits the input current to the device. An analytical derivation of the impedance relationships in each of the three regions of interest for this circuit is made in Appendix B of this memorandum.

Region I

$$R_1' \approx \frac{R_1 R_2 (r_{22}+R_B)}{(R_1+R_2)(r_{22}+R_B)+R_1 R_2} \qquad (17)$$

The following values for R_1 and R_2 were used in a test setup:

$$R_1 = 0.16M$$

$$R_2 = 0.1M$$

Substituting:

$$R_1' \doteq + 20{,}000 \text{ ohms}$$

Region II

$$R_1' \doteq \frac{R_1 R_2}{R_1+R_2+R_1R_2\left(\frac{R_B+R_L}{r_{22}(R_B+R_L)-R_B r_{21}}\right)} \qquad (18)$$

$$R_1' \doteq - 33{,}000 \text{ ohms}$$

Region III

$$R_1' \doteq \frac{R_1 R_2 R_4}{R_1 R_2 + R_2 R_4} \qquad (19)$$

$$R_1' \doteq + 62{,}000 \text{ ohms}$$

It is obvious that values for the circuit components may be chosen to obtain equal slopes in the positive resistance regions if desired, and that control over the slope in Region III may be achieved by a suitable shunting resistance across the diode.

It may be concluded that by the general method of employing biased diodes to modify the V_c' vs I_c negative resistance characteristic of the transistor, a range of impedance values in Region III may be obtained, and that the negative resistance may also be varied, if desired.

3.0 Base Characteristic ($-V_\varepsilon'$ vs I_b)(Grounded Emitter)

3.1 Impedance Relationships

Region I

From equation (1), the general impedance expression for the circuit as shown in Figure 10 may be written:

$$R_1' = r_{11}+R_1+R_g + \left[\frac{(r_{21}-r_{11}-R_1)(r_{11}-r_{12}+R_1)}{r_{22}-r_{21}-r_{12}+r_{11}+R_L+R_1}\right] \qquad (20)$$

This may be simplified from a consideration of device parameter values for Region I:

$$R_1' = r_{11}\left[1 - \frac{r_{11}}{r_{22}+r_{11}+R_L+R_1}\right] \qquad (20A)$$

For typical values of R_L = 1200 ohms, R_1 = 220 ohms, the input resistance may be calculated as

$$h_1' \approx 17{,}500 \text{ ohms}$$

Region II

Using device parameter values, the impedance expression may be simplified to:

$$R_1' = r_{11}+R_1+R_g + \left[\frac{r_{21}(r_{11}-r_{12}+R_1)}{r_{22}-r_{21}+R_L}\right] \qquad (21)$$

Substituting:

$$R_1' = -127 \text{ ohms.}$$

By reducing R_1, the value of the negative resistance may be increased to the order of - 300 ohms. A small value of R_1 may be advisable for current limiting.

Region III

$$R_1' = r_{11}+R_1+R_g + \frac{(r_{21}-r_{11}-R_1)(r_{11}-r_{12}+R_1)}{r_{22}-r_{21}-r_{12}+r_{11}+R_L+R_1} \qquad (20)$$

$$R_1' = 390 \text{ ohms}$$

An examination of the composite characteristic in view of the impedances obtained by the above analysis shows that it approaches the "ideal" short circuit stable voltage-to-current negative resistances of Figure 1. It is proposed to outline a circuit modification by means of which the resistance in Region I may be reduced.

3.2 Circuit and Modification

3.21 Method

A diode is inserted between base and ground, as illustrated in Figure 11. This results in the "S" type characteristic depicted, current-to-current. An analysis of this circuit (Appendix C) shows that in Region I the following expression may be derived:

$$R_1' = R_g + R_{RV_1} \qquad (22)$$

Calculating:

$$R_1' = 210 \text{ ohms}$$

This modification, therefore, results in a nearly perfect current-to-current short circuit stable negative resistance.

4.0 Stabilization of Characteristics

4.10 The foregoing modifications illustrated methods by which the slopes or impedances of the negative resistance characteristics may be controlled. In addition to the slopes, the turning points of the characteristic are of interest to the circuit designer; in fact, the triggering impulse required will depend upon how closely the turnover point may be approached by the load line. Referring to Figure 12, which depicts a normal emitter open circuit stable negative resistance characteristic, on which a biased load line R_1 has been drawn, it can readily be appreciated that if the $V'_{\varepsilon o}$* of the transistor varies over an appreciable range, that a circuit which is designed to have two stable conditions at A & B may become monostable at B, or require a larger triggering impulse v, unless the bias potential is varied, or circuit elements changed in value. It is proposed to describe a circuital means of mitigating this disadvantageous variation in $V'_{\varepsilon o}$.

---

*The $V'_{\varepsilon o}$ may be defined as the voltage developed from emitter to ground with emitter current $I_\varepsilon = o$.

4.11 Figure 12 has been derived from the "broken line" analysis due to B. G. Farley. The following equations based upon this analysis have been derived:

$$I_{\varepsilon t} = \frac{V_{cc}}{R_B - \alpha(R_B+R_L)} \qquad (23)$$

(23)

$$V'_{\varepsilon t} = \frac{V_{cc}}{1 - \frac{\alpha}{1-\alpha}\left(\frac{R_L}{R_B}\right)}$$

These equations define the turnover point corresponding to the onset of current saturation, or the end of the negative resistance region, and show that this turnover point is reasonably fixed, since the α of the A1698 switching transistor is fairly consistently in the order of 2 in this region. The equation defining the turnover point at $I_\varepsilon = o$, the beginning of the negative resistance region is as follows:

$$V'_{\varepsilon o} = \frac{R_B V_{cc}}{R_B+R_L - \frac{V_{co}}{I_{co}}} = \frac{R_B V_{cc}}{R_B+R_L-500\,V_{co}} \qquad (24)$$

Using this relationship, we may calculate the effect of V_{co}* variation. V_{co}'s greater than 40 are proposed as standard acceptance values for the A1698 switching transistor; allowing for a degredation down to 30 during operating life and an upper limit of approximately 120 would result in the spread illustrated as the shaded portion of the curve of Figure 12. The disadvantages of biasing the emitter load resistance R_1 so as to insure stability at point A for all V_{co}'s down to 30 are:

(1) Large triggering impulses required for high V_{co} units.

(2) Narrow range of variation allowable in resultant bias potential, since if the load line falls below the saturation turning point, the unit becomes monostable at A.

---

* The V_{co} may be defined as the voltage developed from base to collector with emitter open-circuited and 2MA Collector current.

4.12 A stabilization means suggested by A. J. Rack is illustrated in Figure 13. It involves the introduction of a biased diode in the base lead, so poled that it is normally conducting for $I_\varepsilon = 0$. The effective resistance in the base lead is then the forward resistance of the diode (~250 ohms), and the changes in the $V'_{\varepsilon o}$ due to variations in V_{co} from unit to unit is reduced from about 13 volts as shown on Figure 12 to approximately 0.4 volt, although the percentage variation remains the same. The proportion of voltage V_2 which appears across R'_B when added, by the principal of superposition, to the $V'_{\varepsilon o}$ acts to center this variable voltage about the origin. Then a very small biasing potential V_1, in the order of -1 volt, will insure stability at point A for all units having V_{co}'s between 30 and 120. As soon as the combined effect of emitter and collector currents exceed the biasing current through the base varistor, the varistor assumes its high resistance condition; the resistance in the base lead changes from R'_B, the forward resistance of the diode, to R_B; the voltage impressed upon the device becomes $V_{cc} + V_2$, and the emitter resistance R_1 is biased a total of $V_1 + V_2$ volts. This change is illustrated in Figure 13 as the vertical axis shift in R_1, the load line. Variations in the biasing voltage V_2 will result in a flat topped pedestal which by design, may have a slight negative resistance as shown. It should also be noted that if $+V_2$ be increased so that the R_1 load line, after the shift in axis, falls below the current saturation turning point, the circuit is monostable, with a stabilized triggering point at A.

This circuit modification has the following advantages:

(1) Stability from unit to unit is achieved.

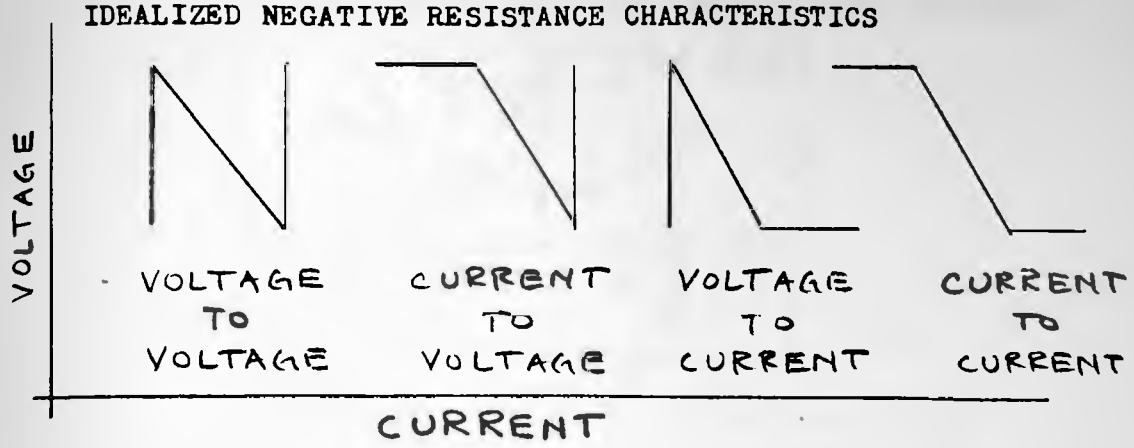
(2) Dissipation in the unit is decreased in the "off" condition. This is the condition of greatest dissipation, and for similar output conditions, is decreased by the modification due to the fact that V_2 is, in effect, not impressed across the unit when the diode is in its low resistance state. The modification has the disadvantage that if more than the V_2 voltage required for correct turnover is applied, the pedestal which results introduces a delay in triggering. As stated heretofore, monostability will result if V_2 is increased to the point where the load line does not intersect the characteristic.

5. Conclusions

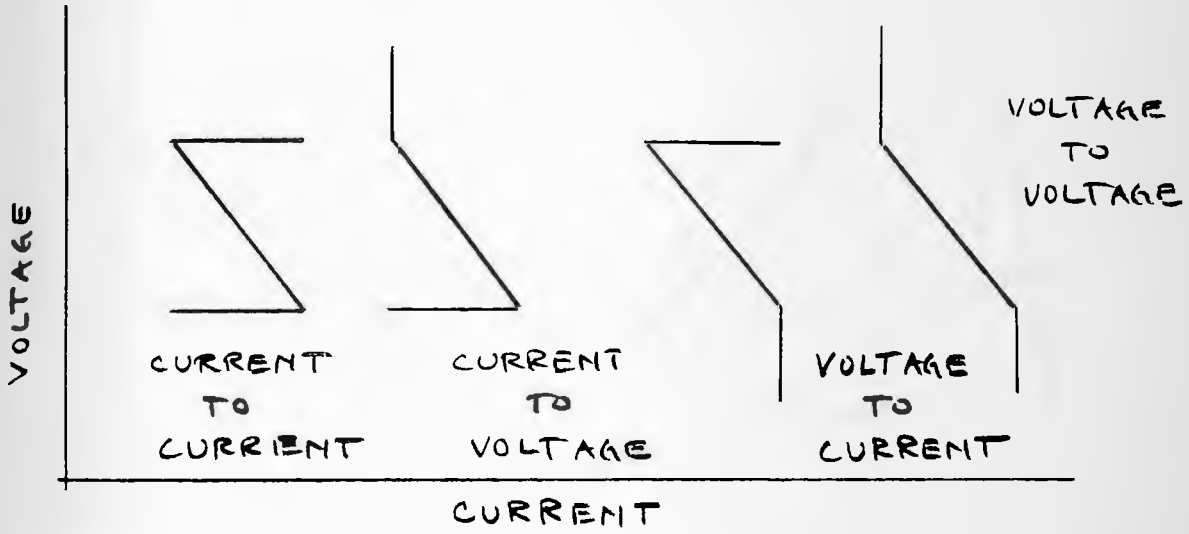
5.1 The slopes of the negative resistance characteristics of transistors, both open circuit stable and short circuit stable types, may be modified by means of biased diodes to obtain symmetrical characteristics which may prove to facilitate the design of switching circuits.

5.2 Stabilization of the $V'_{\varepsilon o}$ turning point of the emitter or collector characteristics may be achieved by means of a biased diode, effecting a high degree of interchangeability and independence from transistor variations.

IDEALIZED NEGATIVE RESISTANCE CHARACTERISTICS



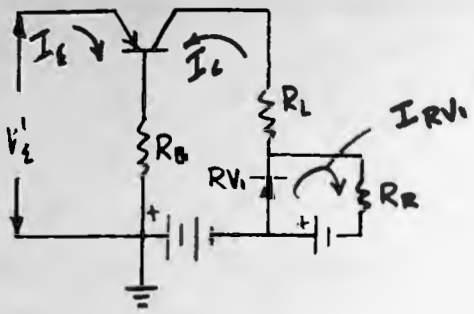
OPEN CIRCUIT STABLE NEGATIVE RESISTANCE TYPES



SHORT CIRCUIT STABLE NEGATIVE RESISTANCE TYPES

FIG. 1

EMITTER CHARACTERISTIC MODIFIED PER METHOD 1



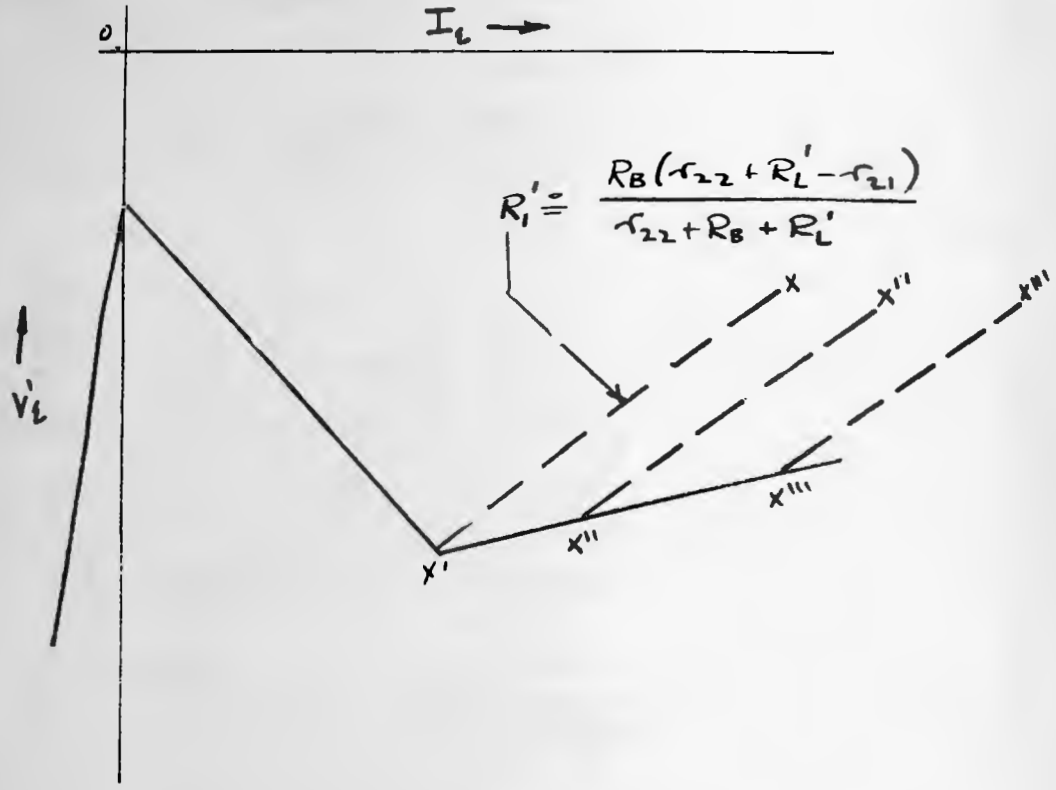
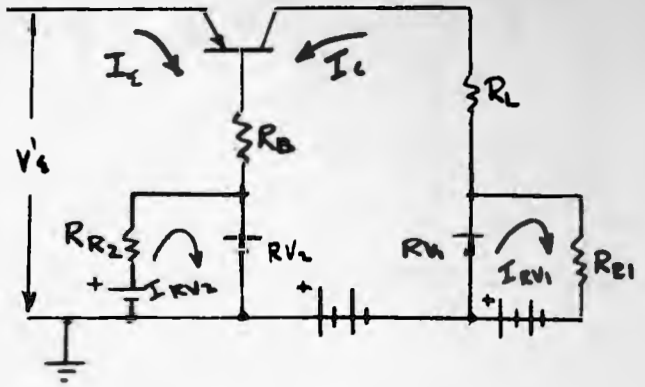


FIG 3

EMITTER CHARACTERISTIC MODIFIED PER METHOD 2



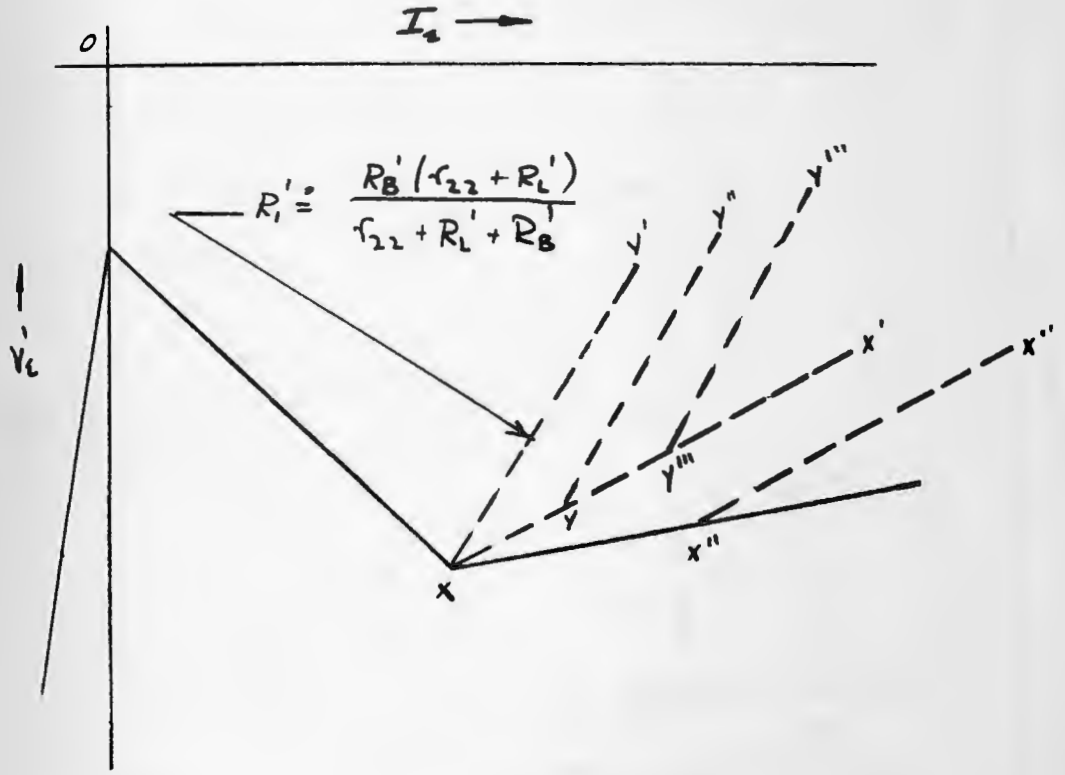
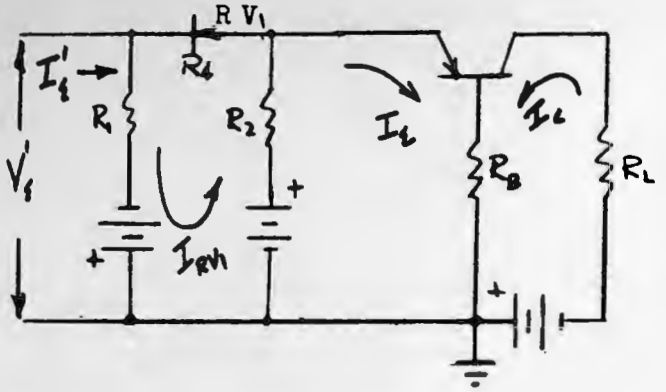


FIG 4

EMITTER CHARACTERISTIC MODIFIED PER METHOD 3



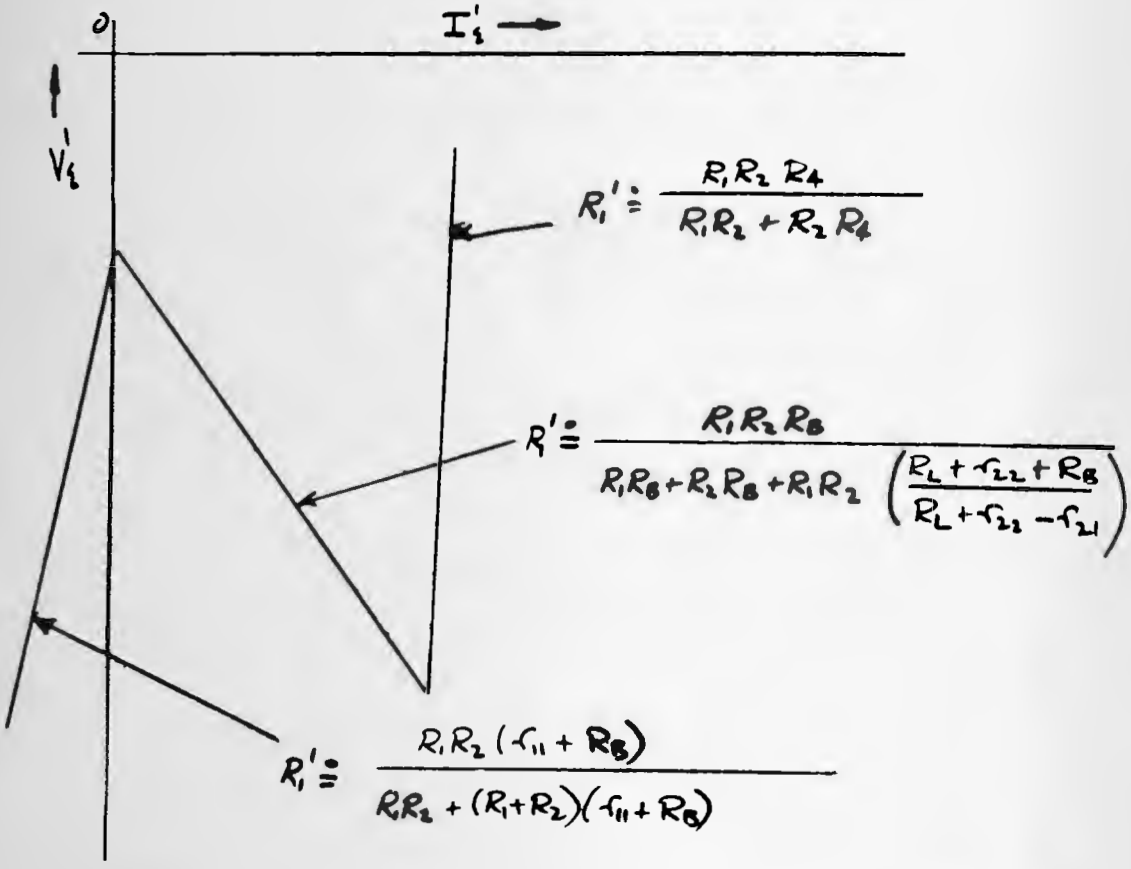
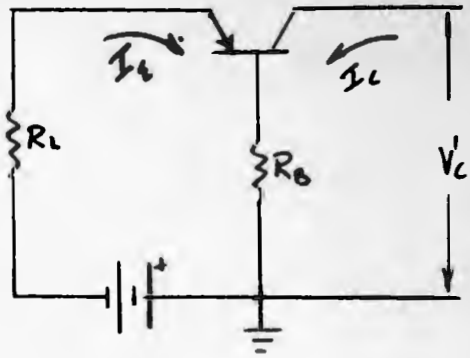
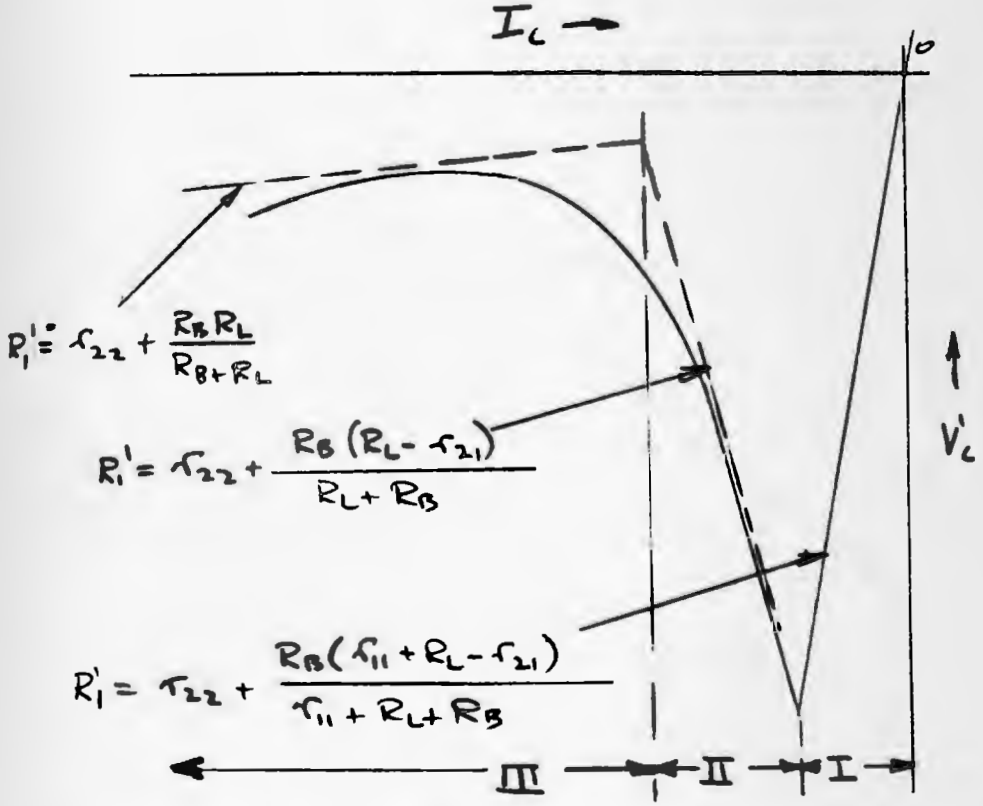


FIG 5

COLLECTOR CHARACTERISTIC LINEAR APPROXIMATION

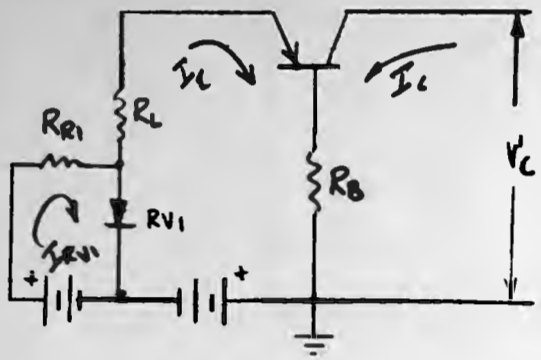




Typical Collector Characteristic (V'_c vs I_c)
Linear Approximation into Three Regions

FIG 6

COLLECTOR CHARACTERISTIC MODIFIED PER METHOD 1



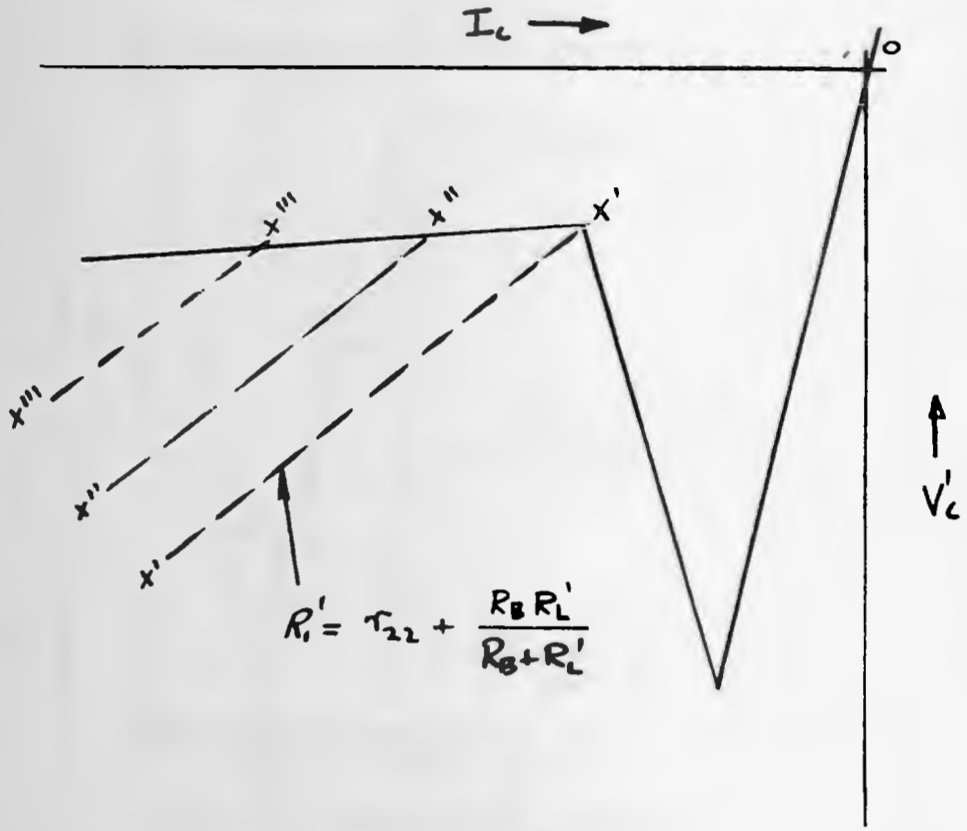
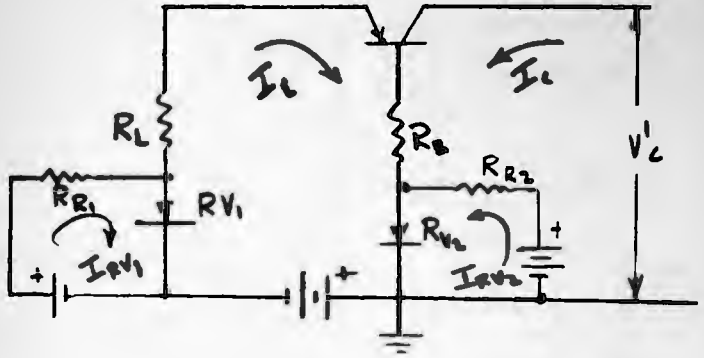


FIG 7

COLLECTOR CHARACTERISTIC MODIFIED PER METHOD 2



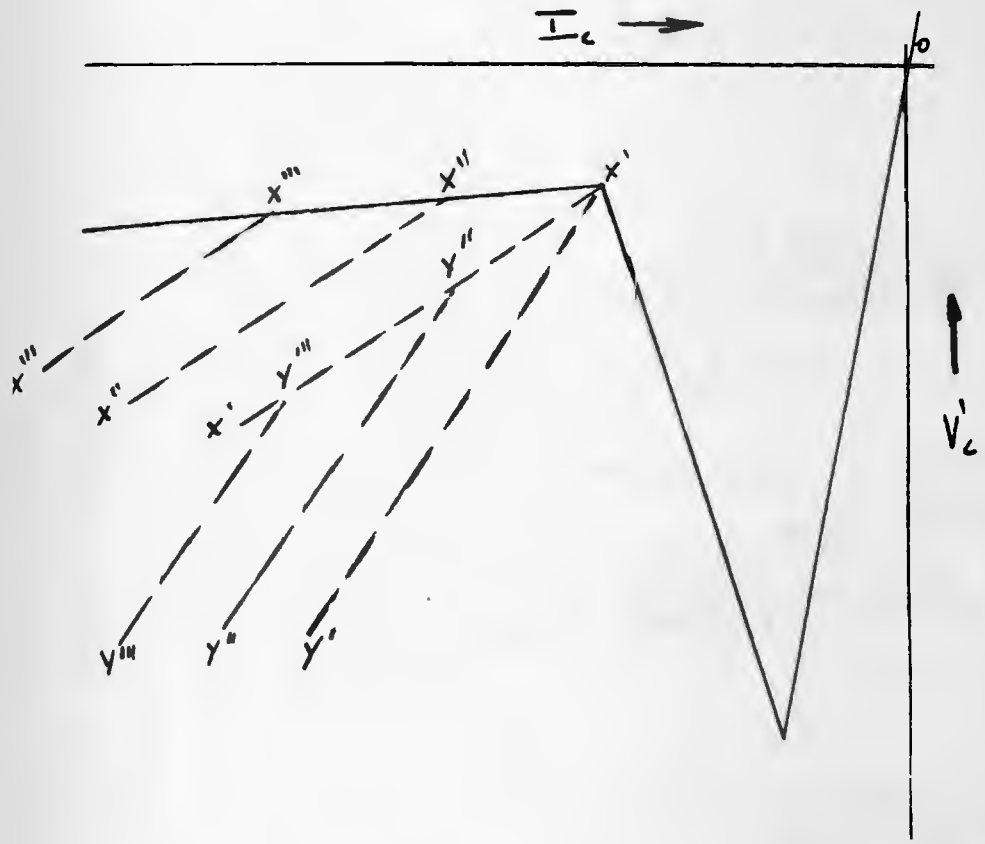
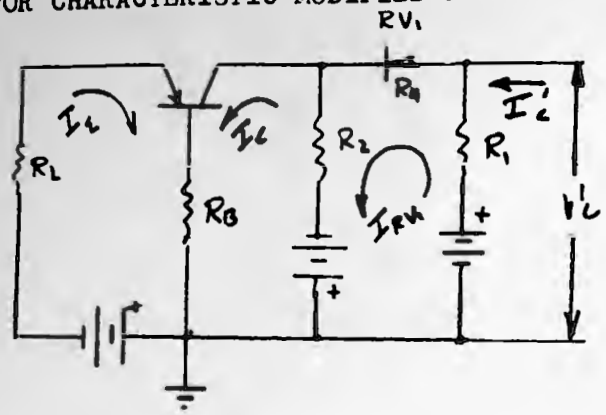


FIG 8

COLLECTOR CHARACTERISTIC MODIFIED PER METHOD 3



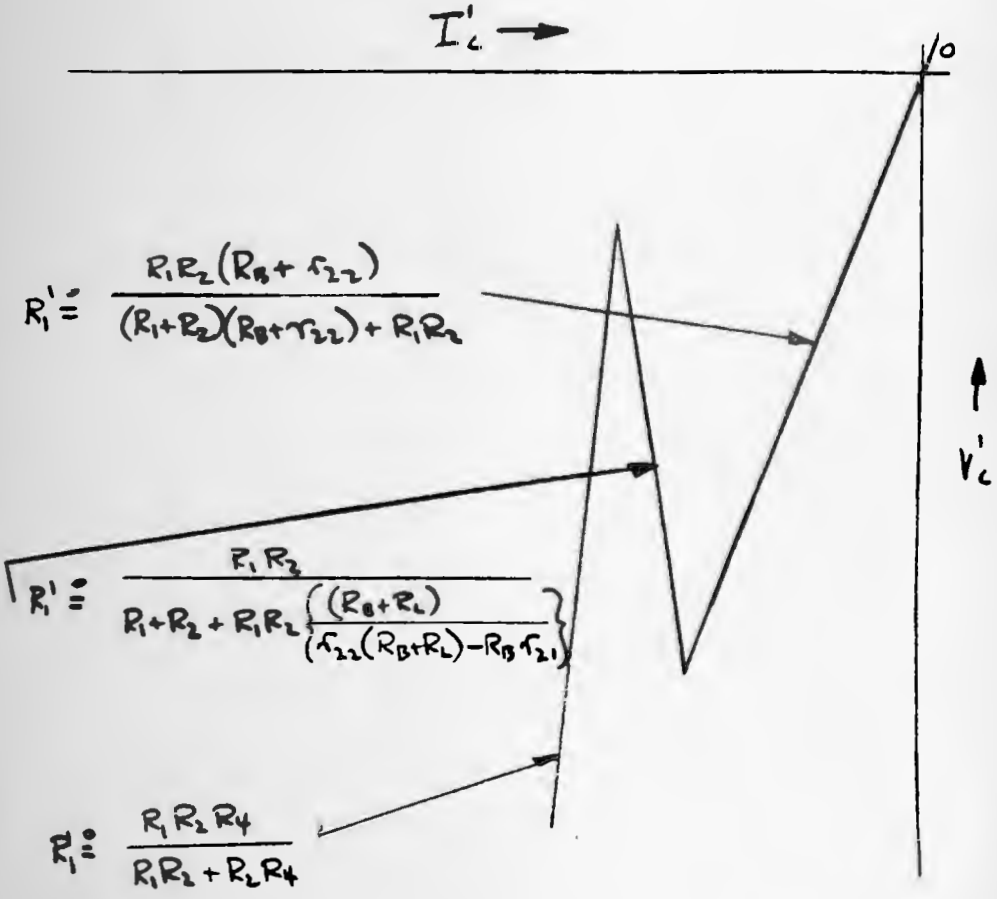
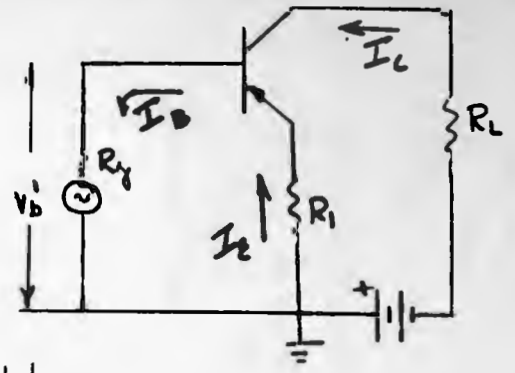


FIG 9

BASE CHARACTERISTIC LINEAR APPROXIMATION



$$R_1' = r_{11} - \frac{r_{11}^2}{r_{11} + r_{22} + R_1 + R_L}$$

V_b'

I

I_b

II

II

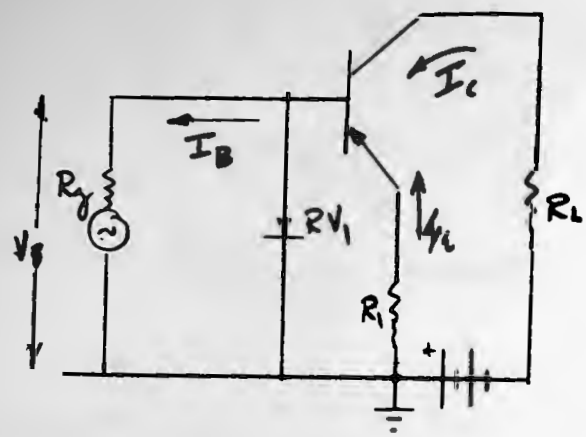
III

$$R_1' = r_{11} + R_1 + R_y + \frac{r_{21}(r_{11} - r_{12} + R_1)}{r_{22} + R_L - r_{21}}$$

$$R_1' = r_{11} + R_1 + R_y + \frac{(r_{21} - r_{11} - R_1)(r_{11} - r_{12} + R_1)}{r_{22} - r_{21} - r_{12} + r_{11} + R_L + R_1}$$

FIG10

BASE CHARACTERISTIC MODIFIED PER METHOD 1



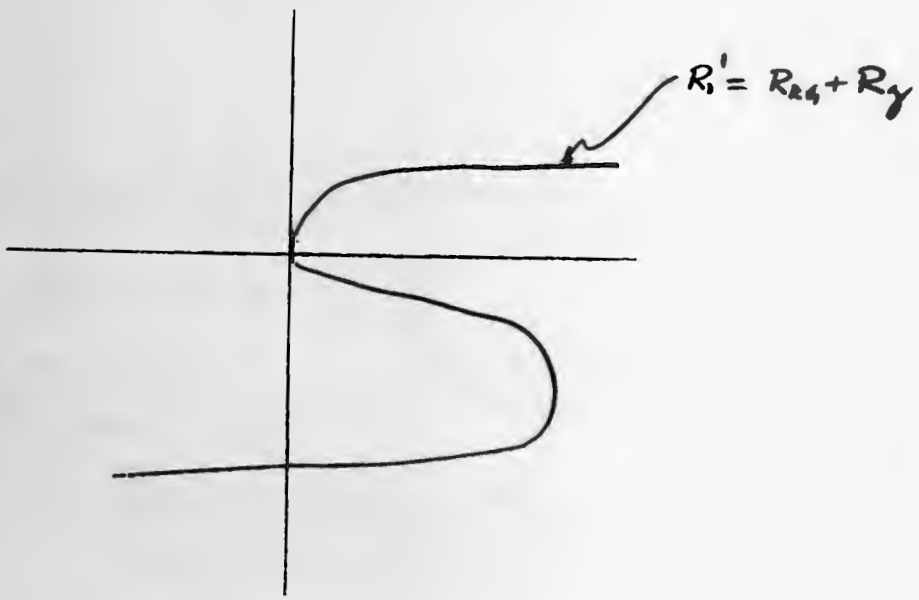
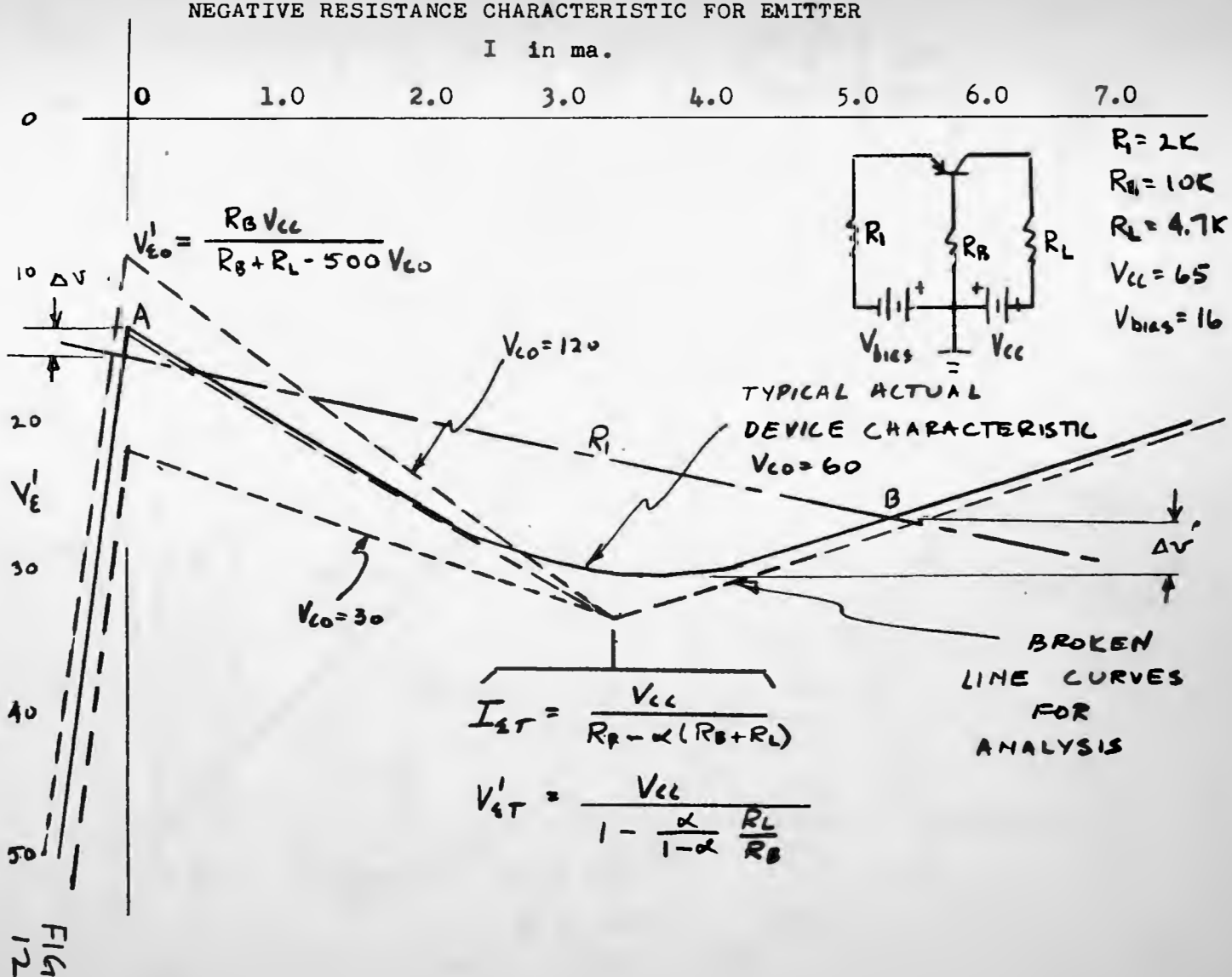
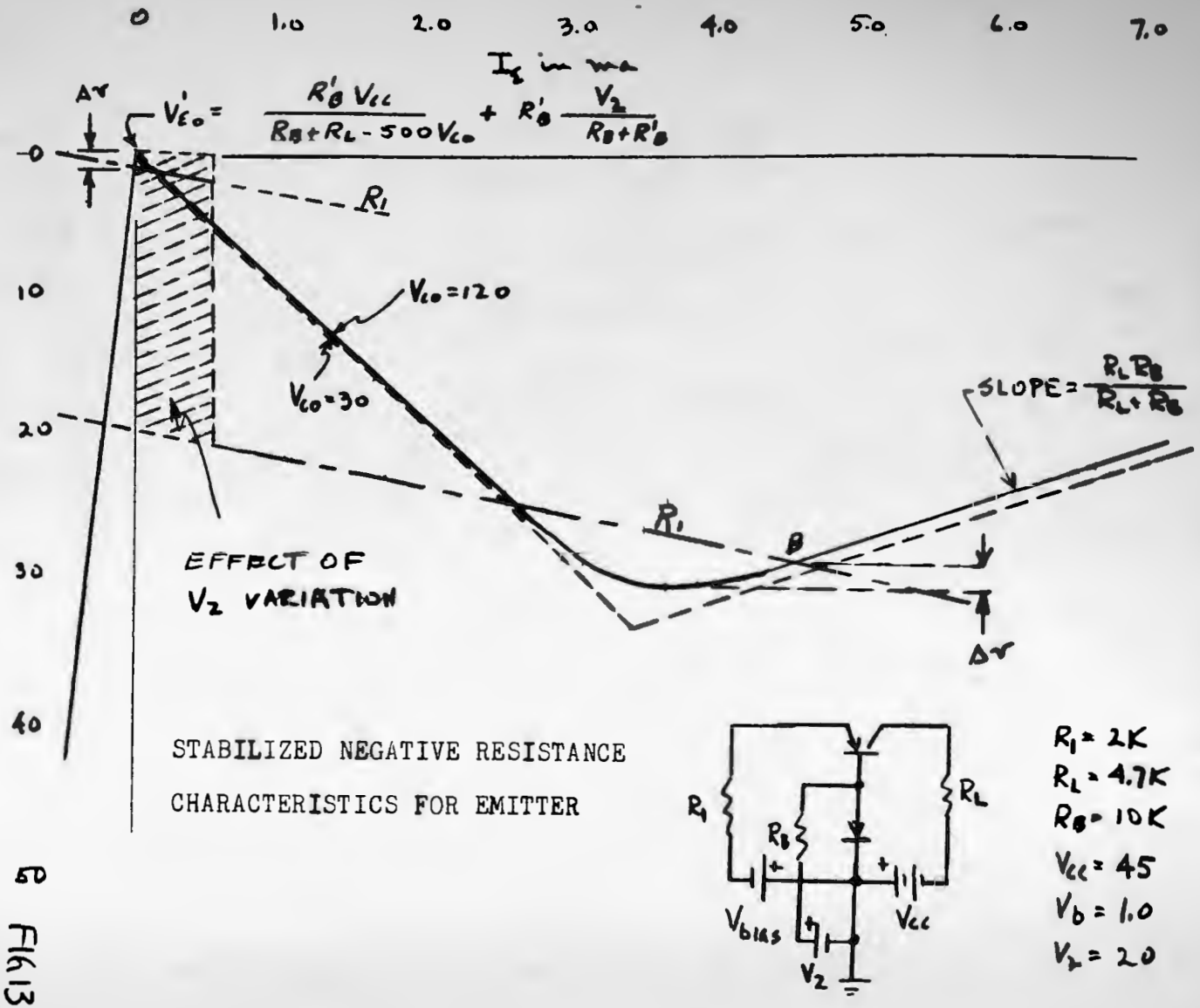


FIG 11

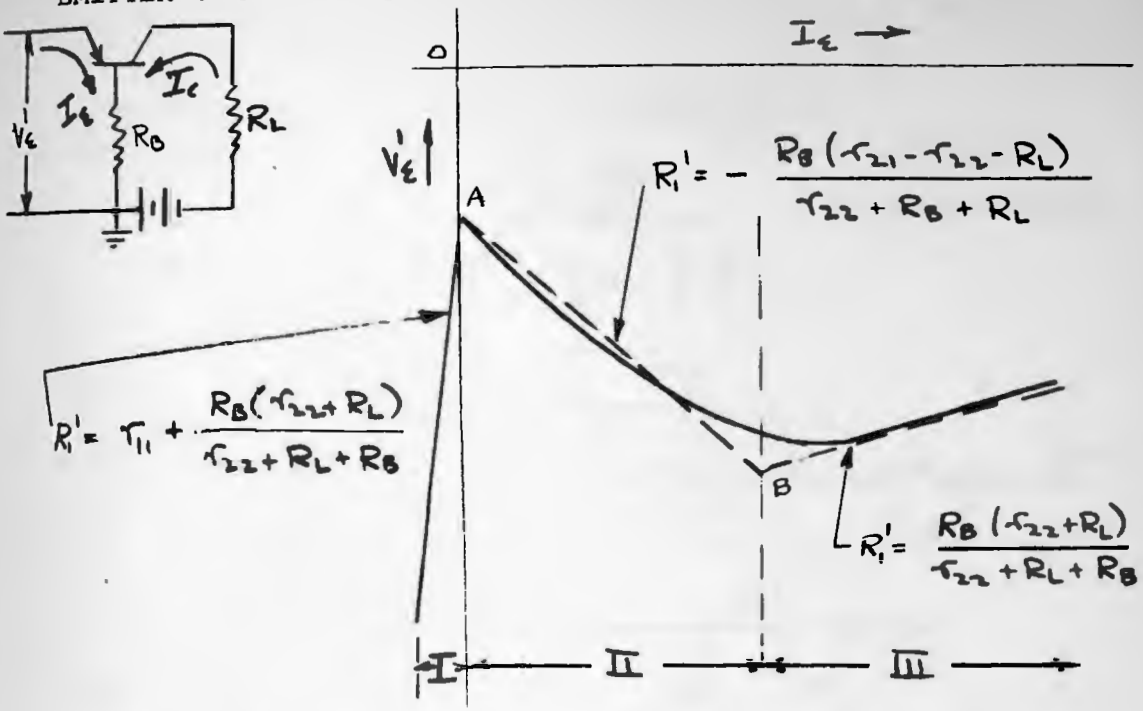

NEGATIVE RESISTANCE CHARACTERISTIC FOR EMITTER
I in ma.
0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
$V'_{EO} = \frac{R_B V_{CC}}{R_B + R_L - 500 V_{CO}}$
$R_1 = 2K$
$R_B = 10K$
$R_L = 4.7K$
$V_{CC} = 65$
$V_{bias} = 16$
R_1
R_B
R_L
V_{bias}
V_{CC}
10
ΔV
A
$V_{CO} = 120$
TYPICAL ACTUAL
DEVICE CHARACTERISTIC
$V_{CO} = 60$
20
V'_E
R_1
B
ΔV
30
$V_{CO} = 30$
BROKEN
LINE CURVES
FOR
ANALYSIS
40
$I_{ST} = \frac{V_{CC}}{R_B - \alpha(R_B + R_L)}$
$V'_{ST} = \frac{V_{CC}}{1 - \frac{\alpha}{1-\alpha}\frac{R_L}{R_B}}$
50
FIG 12



STABILIZED NEGATIVE RESISTANCE CHARACTERISTICS FOR EMITTER

Fig 13

EMITTER CHARACTERISTIC LINEAR APPROXIMATION AND TABLE A



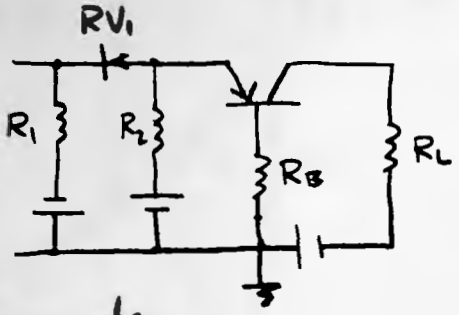
Typical Emitter Characteristic (V_E' vs I_E)
Linear Approximation into Three Regions

TABLE A

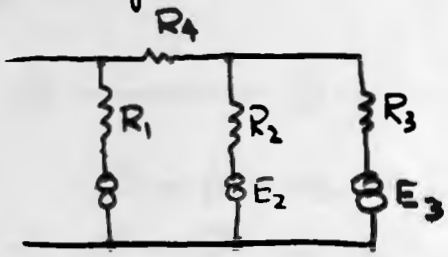
Device Parameter	Approximate Value in Ohms Region I	II	III
r_{11}	0.1M	400	200
r_{22}	20K	20K	~1000
r_{12}	100	100	100
r_{21}	0	50K	~100

FIG 2

For the circuit:



The input circuit may be represented by:



Since the device may be characterized by:

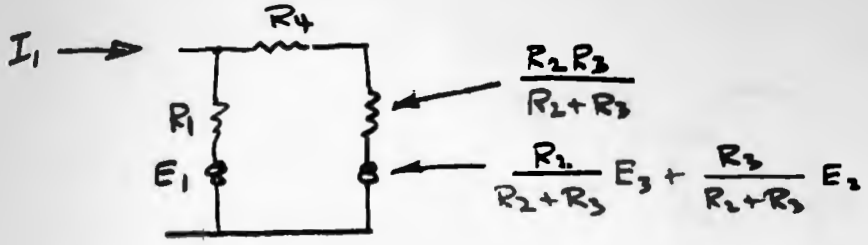
$$V_{E_2} + i_{E_2}\left[r_{12} - r_{11} + \frac{(R_B + r_{12})(r_{21} - r_{22} - R_L)}{R_B + R_L + r_{22}}\right] = V_{CC}\left[\frac{R_B + r_{12}}{R_B + R_L + r_{22}}\right]$$

Then:

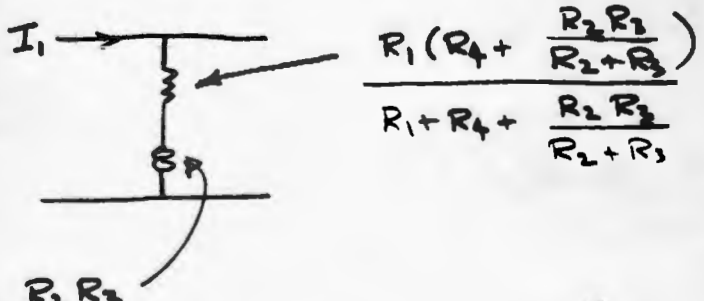
$$R_3 = r_{12} - r_{11} + \frac{(R_B + r_{12})(r_{21} - r_{22} - R_L)}{R_B + R_L + r_{22}}$$

$$E_3 = V_{CC}\left[\frac{R_B + r_{12}}{R_B + R_L + r_{22}}\right]$$

By Thevenin's Theorem, reflecting equivalent generators:



and:



$$\frac{R_4+\frac{R_2R_3}{R_2+R_3}}{R_1+R_4+\frac{R_2R_3}{R_2+R_3}}E_1+\frac{R_1}{R_1+R_4+\frac{R_2R_3}{R_2+R_3}}\left[\frac{R_2}{R_2+R_3}E_3+\frac{R_3}{R_2+R_3}E_2\right]$$

Then:

$$R_1'=R_1\left[\frac{R_3(R_2+R_4)+R_2R_4}{R_3(R_1+R_2+R_4)+R_1R_2+R_2R_4}\right]$$

In the regional interest, the general expression for R_3 given on page 1 may be simplified:

$$R_3=r_{11}+R_B-\frac{(R_B+r_{12})(R_B+r_{21})}{r_{22}+R_L+R_B}$$

Substituting this expression and simplifying:

Region I:

$$R_1' \doteq \frac{R_1 R_2 (r_{11} + R_B)}{(R_1 + R_2)(r_{11} + R_B) + R_1 R_2}$$

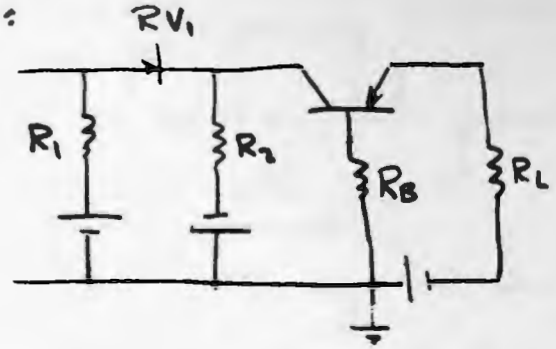
Region II:

$$R_1' \doteq \frac{R_1 R_2 R_B}{R_B(R_1 + R_2) + R_1 R_2 \left[\frac{r_{22} + R_L + R_B}{r_{22} + R_L - r_{21}} \right]}$$

Region III

$$R_1' \doteq \frac{R_1 R_2 R_4}{R_1 R_2 + R_2 R_4}$$

For the circuit:



From appendix 1:

$$R_1' = R_1\left[\frac{R_3(R_2+R_4)+R_2R_4}{R_3(R_1+R_2+R_4)+R_1R_2+R_2R_4}\right]$$

Substituting for R_3 and simplifying:

$$R_3 = r_{22} + \frac{R_B(r_{11}+R_L-r_{21})}{r_{11}+R_L+R_B}$$

Region I:

$$R_1' \doteq \frac{R_1R_2(r_{22}+R_B)}{(R_1+R_2)(r_{22}+R_B)+R_1R_2}$$

Region II:

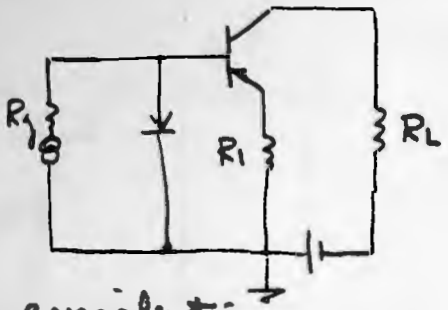
$$R_1' \doteq \frac{R_1R_2}{R_1+R_2+R_1R_2\left[\dfrac{R_B+R_L}{r_{22}(R_B+R_L)-R_Br_{21}}\right]}$$

Region III:

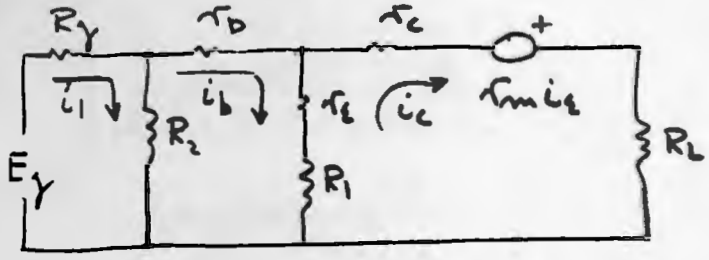
$$R_1' \doteq \frac{R_1R_2R_4}{R_1R_2+R_2R_4}$$

Appendix 3

For the circuit:



We may draw the equivalent:



Setting up mesh equations:

$$\textcircled{1}\quad (i_b + i_1)R_\gamma + i_b(r_b + r_\epsilon + R_1) - i_c(R_1 + r_\epsilon) = E_\gamma$$

$$\textcircled{2}\quad (i_b + i_1)R_\gamma + i_1 R_2 = E_\gamma$$

$$\textcircled{3}\quad i_b(r_m - r_\epsilon - R_1) + i_c(r_c + R_L + r_\epsilon - r_m + R_1) = 0$$

Solving for input R_1':

$$R_1' = \frac{(R_2 + R_\gamma)\left[r_b + r_\epsilon + R_1 + R_\gamma - \frac{R_\gamma^2}{R_2 + R_\gamma} + \frac{(r_m - r_\epsilon - R_1)(R_1 + r_\epsilon)}{r_c + R_L + r_\epsilon - r_m + R_1}\right]}{\left[r_b + r_\epsilon + r_1 + R_\gamma - \frac{R_\gamma^2}{R_2 + R_\gamma} + R_2 - \frac{R_\gamma R_2}{R_\gamma + R_2} + \frac{(r_m - r_\epsilon - R_1)(R_1 + r_\epsilon)}{r_c + R_L + r_\epsilon - r_m + R_1}\right]}$$

since the numerator & denominator functional expressions are almost exactly equal.

$R_1' = R_2 + R_y$ in Region I.

The characteristic expressions in Regions II and III are the same as for the analysis without RVI.

Summarizing:

Region I:

$$R_1' \doteq R_y + R_2$$

Region II:

$$R_1' \doteq r_{11} + R_1 + R_y + \frac{r_{21}(r_{11} - r_{12} + R_1)}{r_{22} - r_{21} + R_L}$$

Region III:

$$R_1' \doteq r_{11} + R_1 + R_y + \frac{(r_{21} - r_{11} - R_1)(r_{11} - r_{12} + R_1)}{r_{22} - r_{21} - r_{12} + r_{11} + R_L + R_1}$$

SOME SWITCHING ASPECTS OF TRANSISTORS

A. E. ANDERSON

Produced under contract W 36-039 sc-44497

SOME SWITCHING ASPECTS OF TRANSISTORS

INTRODUCTION

It is proposed to discuss some of the properties of transistors which are applicable to switching, to develop elementary analysis methods and to describe a few circuits.

The bounds or limits of the field of switching are difficult to define. The common thread usually involves definite states of being as "open or closed", "off or on", "0 or 1" and so on, rather than a continuum of conditions. Even when consideration is given to more than two states, the thought involves distinct recognition of each state. The field is termed to be non-linear in distinction to linear manipulation of information. Any number of anomalies in definition may be raised.

Without attempting to either define or to limit the field some of the functions which are often employed are: wave form generation, as rectangular pulses, sawtooth waves, etc; memory or storage which may be for short, intermediate or long periods and involves the retention of information for subsequent use; operations involving addition, subtraction, multiplication and division; translation of information from one form or code to another; gating, involving the routing of signals according to a predetermined pattern or set of conditions; regeneration of signals in amplitude and wave form; delay, which may be thought of as a form of storage; and timing. Some of these functions are simple; others result from fairly complex structures of simpler functions.

Present trends in switching systems are toward complicated automata as exemplified by digital computers.[1] The reliability, power consumption and physical size of the electron devices employed largely determines the degree of realizability of such systems. It is believed that the transistor will find a significant application in this field.

The transistor can reduce power consumption by the elimination of heater or filament power. In addition, particularly in broadband applications as in high speed pulse systems, the "B" power may be reduced by the order of decades if not more.

- -

1 "High Speed Digital Computers", L. N. Ridenour, Journal of Applied Physics, vol. 21, no.4, April, 1950, pp.263-270.

Transistor circuits with 0.02μs rise time have been made to operate with an input power of 20 milliwatts which compares with approximately 2.5 watts (1 watt heater, 1.5 watt plate) for an equivalent tube circuit. Transistors have operated with less than one microwatt input power.[2]

Such power reductions result from the low operating voltages, low internal resistances and low capacitances of transistors. Low internal impedances greatly reduce the importance of stray wiring capacitances thereby making mechanical design much simpler and often eliminating the need for the isolating or buffer amplifiers.

The transistor can contribute difinite reduction in size directly. Fig. 1 is a photograph of a "bead" transistor which has a volume of approximately one-thousandth of a cubic inch and a weight of 5/1000 ounce. Indirectly the transistor can contribute to size reduction through the use of smaller, lower voltage, lower dissipation components. The reduction of power supply requirements in terms of size, regulation and capacity is also quite appreciable.

Transistors have been subjected to shocks in excess of 20,000 G without change in characteristics. Vibration tests have shown no resonances in the transistor shown in Fig. 1 to several thousand cycles. Harmonic accelerations of 100G at 1000 cycles have produced no detectable current modulation. Life reliability and expectancy is difficult to determine due to the relative infancy of transistors, the definite finiteness of time, the many variables involved and the rate of development progress. Average life is presently estimated to be in excess of 70,000 hours. Life is a function of the operating conditions and may be materially reduced accordingly.

At present transistors also have limitations. Noise is high as compared to electron tubes; input impedances are low, which may be either advantageous or disadvantageous; power output is limited; frequency response is relatively low; circuit instability may cause design difficulties; and the devices are sensitive to temperature changes. There is also an absence of a long practical experience with a consequent art background in both devices and circuits.

While it is difficult to define the switching field, it is no less difficult to discuss circuit and device properties

- -

2 "Some Circuit Properties and Applications of n-p-n Transistors" R. L. Wallace, Jr. and W. J. Peitenpol, Bell Syst. Tech. Jl, vol.XXX, no.3, July, 1951, pp.530-563.

on a general basis. This is related to the non-linear nature of the circuits and devices in distinction to the virtually classical linear small signal field. The lack of a classical method of analysis is a serious handicap in the synthesis of contemporary circuits and devices. When new devices, as the transistor, are to be considered, the problem is multiplied due to the lack of a long practical background of experience.

It has been assumed that negative resistance is a common thread between "trigger circuits" and oscillators regardless of the device employed-electron tube, gas tube, transistor, mechanical structures, etc. This is not a new or novel idea and there is no intent to present it as such.[3] Rather, it is used as a pattern upon which a certain degree of transistor analysis may be based, leading to simple understanding. The analysis assumes that the negative resistance characteristic can be broken into three regions, in which each region can be considered on a linear basis.

Section I will deal with simple properties; Section II with analysis and device properties; and Section III with circuit design.

I - SIMPLE PROPERTIES

The common property, which is ascribable to swtiching functions is that of definiteness of state. The condition of the function is either "off" or "on". Switches are either open or closed; relays operated or not; tubes are in cutoff or saturation; doors are open or closed and so on. This is common regardless of the phenomena being exploited.

There is an intermediate region between these two conditions usually characterized by a time which is related to how fast the function may go from one state to another. Functionally these times of closing and opening are taken to be zero; practically, they are of determining importance. Relays replace hand-operated switches and electronic devices replace relays as speed becomes important. Obviously, no function or system can be faster than can its state-devices operate.

- -

3 See for example "Negative Resistances, their Characteristics and Effects. Sinusoids, Relaxation Oscillations and Relaxation Discontinuities", Walter Reichardt, Elektrische Nachtrichten-Technik, v.20, no.9, March, 1943, pp.76-87. "Uniform Relationship Between Sinusoids, Relaxation Vibrations and Discontinuites", Walter Reichardt. Elektrische Nachtrichten-Technik, v.20, no.9, Sept., 1943, pp.213-225. For transistors: "Counter Circuits Using Transistors", E. Eberhard, R. O. Endres and R. P. Moore, RCA Review, Dec., 1949, pp.459-476 and "A Transistor Trigger Circuit", H. J. Reich and Wugvary, Rev. Sci. Inst., v.20, p.8, Aug., 1949, p.586.

All such state-devices will have separate attendant properties such as the degree of reverse coupling between the controlling signal and the controlled signal. Separated into families, however, there are those which are passive and those which are active. The latter are threshold devices in which a small amount of signal or control energy results in the translation of a relatively larger amount of stored energy into dynamic energy which consummates the change in state. As long as the control signal is below the inital threshold there is no response and any change is directly related to the control signal itself. When the signal exceeds the threshold the second state is assumed. Watch escapements, thyratrons, and the whole family of oscillators fall into this category. When such functions are analyzed they are found to involve in one way or another two states, which might be termed cutoff and saturation, separated by a region in which there is positive feedback with resultant negative resistance. Taking the proposition that a negative resistance characteristic is common to trigger or threshold switching circuits, the next step is to examine the transistor for such behavior and classify the properties.

That the transistor* can exhibit negative resistance has been demonstrated analytically[4] and experimentally. The resistances seen looking into the emitter and collector of the transistor with grounded base are shown in Figs. 2A and 2B. Capital letters as R_{11} etc., refer to circuit parameters; lower cases to device parameters as r_c, r_ε; $r_{11} = r_\varepsilon + r_b$; $r_{22} = r_c + r_b$, etc. This nomenclature has been covered thoroughly in the literature and will not be discussed further. See, for example, reference 4.

Taking the collector characteristic, Fig. 2B, for example,

$$R_{in} = (r_c + r_b) - \frac{r_b (r_b + r_m)}{r_\varepsilon + r_b + R_\varepsilon} \qquad (1)$$

- -

* Discussion is limited primarily to point contact transitors with α's or current gains greater than unity.

4 "Some Circuit Aspects of the Transistor", R. M. Ryder and R. J. Kircher, Bell Syst. Tech. Jl., vol. XXVIII, no.3, July, 1949, pp.367-400

R_{in} can be negative or positive depending upon the relative magnitudes of the two terms. Actually, of course, r_m has a phase factor and so is frequency dependent. Frequencies wherein r_m is essentially resistive will be assumed. For negative resistance r_m must be large, R_ε small and r_b not too small or augmented by external resistance. Negative resistance is thus predicted on a small signal basis. The large-signal behavior may be studied experimentally by adding sufficient resistance as R_c to the first or positive term to insure stability. This is shown in Fig. 3 with the resultant characteristic. External base resistance R_b has been added.

Fig. 3B, shows the pattern of cutoff and saturation, separated by a negative resistance. The negative resistance is obvious in the b region, c is equally obvious as a saturation wherein the current increases with very little change in voltage. Region a as a cutoff is not so clearly obvious since the slope in finite. Thus it is a sort of "leaky" cutoff.

An evaluation of the emitter input characteristic leads to similar results, using the circuit of Fig. 4A. R_b has been added here also. The general pattern is again present. Region a corresponds to cutoff, b to the negative resistance region and c to saturation. Here, however, the cutoff region more nearly approaches infinite slope. The emitter has diode properties so that when I_ε is made negative the emitter resistance is very high, as in a diode. This fact will be employed later on in the analysis where recognition is made of the non-linear nature of the emitter resistances.

Both of the negative resistances discussed are of the so-called open-circuit stable type. If loads are applied to the "in" terminals of Fig. 2 which are larger in magnitude than the negative resistances, the circuits will be stable; that is, there will be singular operating points. This is shown in Fig. 4B by the dotted lines marked, R_ε' , R_ε'' , R_ε'''. A load resistance smaller in magnitude than the negative resistance may intersect the characteristic in three positions as shown by the load line R_ε. The load line R_ε can be made to have singular or multiple intersections by biasing properly as shown in Fig. 5, where the three possibilities are shown as R_ε, R_ε', R''. Singular intersection in the cutoff or overload region, or multiple intersections result in accordance with the choice of emitter bias, $V_{\varepsilon\varepsilon}$, as shown. It can be shown that the intersection of load line R_ε with the characteristic at b as shown in Fig. 5 is unstable whereas those at a and c are stable. Transient analysis and experiment show also in the multiple intersection case

that as $V_{\varepsilon\varepsilon}$ is slowly increased (decreased in absolute magnitude) the load line moves upward and that the assumed operating point a moves up along the cutoff portion of the characteristic. At the turning point shown on the current axis, the operating point suddenly flips to the saturation region, returning along the curve to c as $V_{\varepsilon\varepsilon}$ is returned to the original value.

A decrease in $V_{\varepsilon\varepsilon}$ toward $V''_{\varepsilon\varepsilon}$ moves the operating point at c downward along the characteristic until it "escapes" past the lower turning point and flips to the cutoff portion, returning to a as $V''_{\varepsilon\varepsilon}$ is returned to the original value. This then is an elementary switching circuit, a bistable trigger circuit or "flip-flop". A positive emitter pulse will cause the circuit to flip to high current, a negative pulse to low current. The polarities may be reversed or used in combination by application of the trigger to the base. Trigger sensitivities are shown in Fig. 6. Such a circuit is often used for register or storage purposes. It can store one bit of information for a potentially infinite period, be sampled for the presence of such information, and be cleared or restored to the original condition for reuse when the stored information is no longer useful. With the addition of suitable steering circuits it can be made to count by a scale of two.

The addition of a capacitor to the circuit as in Fig. 7 leads to either monostable or astable operation. In Fig. 7A the normal operating point is stable at a as discussed previously by virtue of the bias $V_{\varepsilon\varepsilon}$. As $V_{\varepsilon\varepsilon}$ is increased as by a trigger, the load line is moved up and over the turning point. Without capacitor C in the circuit, the operating point would move to b with the resultant rapid change in voltage and current. However, a capacitor has in effect voltage inertia; this is equivalent to saying that a capacitor is a short circuit to a voltage change, the degree of short-circuitedness being related to both the capacitance and the rate of change of voltage, both of which are assumed high. Thus at the turning point the capacitor effectively short circuits the emitter and the operating point moves along dotted line (1) to intersect the characteristic. This point is quasi-stable as the capacitor discharges rapidly along line (2) to the second turning point where the emitter is again effectively short circuited and the operating point moves along (3) to intersect the cutoff portion of the characteristic. This point is again quasi-stable and the operating point moves slowly up to the initial or dc stable operpoint. A single trigger thus causes a complete cycle of operating tion. The emitter current shifts rapidly to a high value of

current, falls relatively slowly to an intermediate value, then shifts rapidly to a small negative value and finally slowly returns to the original value. The emitter current and voltage are sketched in Fig. 8. It is a so-called "single-shot" circuit.

This primitive behavior has the basic properties of several switching functions. It can be used as a regenerative amplifier to restore level and wave shape, as a delay and as a wave form generator. The emitter voltage (and base voltage) has a saw-toothed shape; design factors leading to linearity, etc., apply. The current wave has the shape of rectangular pulses without the desirable features of flat top and base. The base line may be flattened by taking the signal from the collector which follows the emitter except when the emitter is in the negative current region. The dc stable point may be in the saturation region as in Fig. 5 with suitably chosen positive emitter bias.

When the load line and bias are chosen to result in intersection in the negative resistance portion, astable operation or continuous oscillation results. This mode is illustrated in Fig. 7C. Proper bias and $R_\varepsilon > |-R_{in}|$ are required. There are at least two physical explanations of this mode of behavior. The operating point formed by the intersection of the load line on the negative resistance portion of the characteristic would normally be stable. However, the capacitor provides a virtual short circuit in parallel with R_ε causing the path (1), (2), (3), (4) to be followed continuously. Another form of physical explanation of this relaxation oscillation, usually applied to gas tubes, is that the capacitor C is charged slowly through R_ε to a critical or breakdown value with a resultant discharge of the capacitor. When the capacitor charge is dissipated, the discharge can no longer be maintained due to the IR drop in R_ε and the tube or device goes off and the capacitor is recharged.

The above suggests a strong simularity to gas tube behavior and this is indeed so. In fact, the modes described above are common to all open-circuit stable devices; only the parameters involved are different.

Since the collector characteristic shown in Fig. 3 is also open-circuit stable, the same sort of circuits can be constructed using this characteristic. Bistable, monostable and astable circuits are shown in Fig. 9.

The resistances seen looking into the base are given in Fig. 10. These circuits are short-circuit stable. That is, high values of R_b result in instability. Bistable, monostable and astable circuits can be constructed also, but use is made of an inductor instead of a capacitor. The reactance of the inductor affords a quasi-open circuit in the same manner as the capacitor afforded a quasi-short circuit in the previous cases. Circuits are shown in Fig. 11.

These simple circuits by no means exhaust the switching circuit possibilities; they are often useful if required properties can be obtained. A study and analysis of these basic circuits will involve a detailed understanding of transistor behavior which is necessary to understand more complex circuits in any event.

II ANALYSIS AND DEVICE PROPERTIES

In order to deal analytically with circuits and devices it is necessary to have analytic expressions for the device characteristics. For small signal analysis this is relatively easy. In large signal applications, as in switching, the situation is not so simple. The problems arise because of the high degree of non-linearity wherein the simplifying assumptions employed in small signal analysis are by no means valid. Further it is desirable to retain dc terms in many cases.

The method to be employed here is the so-called broken-line method and involves breaking the characteristic into three regions corresponding to: I Cutoff, II Active, and III Saturation. It is assumed that linear relationships hold separately in each region with proper choice of parameters. This is illustrated in Fig. 12A in terms of a negative resistance characteristic and in Fig. 12B in terms of an idealized transistor output characteristic ($r_{22} = r_c + r_b$). The assumption that the negative resistance characteristic can be simulated by three straight lines is reasonably valid for gross considerations; for fine detail near the turning points it is by no means accurate although affording zero order information.

The idealization of the transistor collector characteristics does not depart too markedly from actual curves as may be seen by comparison to actual characteristics in Fig. 20. The saturation line in Fig. 12B is shown as common for all values of I_e. This is factually not true, although in practice the several lines sensibly overlap.

The curves of Fig. 12B have the equation, accounting for the signs of I_ε and I_c:

$$V_c = r_{22} (I_c + \alpha I_\varepsilon - I_{co}) \tag{2}$$

where V_c is the collector voltage and $r_{22} = r_c + r_b$. I_{co} is a small dc residual current and is usually of concern only when r_c is very high or when investigations are being carried out very near the origin. α is the current multiplication factor and I_ε is the emitter current. Since $\alpha = \frac{r_{21}}{r_{22}}$

$$V_c = I_\varepsilon r_{21} + (I_c - I_{co}) r_{22} \tag{3}$$

or,

$$V_c = I_\varepsilon (r_m + r_b) + (I_c - I_{co})(r_c + r_b) \tag{4}$$

It remains then only to evaluate Equation 4 in each of the three regions. In cutoff, Region I, there are no transfer properties; hence r_m is zero. In Region II, all parameters have normal values. In Region III, saturation, r_m and r_c are both zero. Hence,

$$\text{Region I} \quad V_c = I_\varepsilon r_b + (I_c - I_{co})(r_c + r_b) \tag{5}$$

$$\text{Region II} \quad V_c = I_\varepsilon (r_m + r_b) + (I_c - I_{co})(r_c + r_b) \tag{6}$$

$$\text{Region III} \quad V_c = I_\varepsilon r_b + (I_c - I_{co}) r_b \tag{7}$$

From (7) it is clear that the saturation curve slope shown in Fig. 12 is r_b when $I_\varepsilon = 0$.

The effect of load resistance R_c can be introduced simply since

$$V_{cc} - I_c R_c = V_c \tag{8}$$

where V_{cc} is the supply voltage.

Thus Equation 4 can be rewritten as:

$$V_{cc} = I_\varepsilon(r_m + r_b) + I_c(r_c + r_b + R_c) - I_{co}(r_c + r_b + R_c) \quad (9)$$

Equation (9) has the general form:

$$V_2 = I_1 R_{21} + I_2 R_{22} \quad (10)$$

where $V_2 = V_{cc} + I_{co}(r_c + r_b + R_c)$, $I_1 = I_\varepsilon$, $I_2 = I_c$, and R_{21} and R_{22} are circuit parameters.

The idealized emitter family is shown in Fig. 13 wherein the double valuedness of r_ε is vividly illustrated. This comes about through the normal emitter diode non-linearity; that is when the emitter is positive with respect to the base, the emitter has low resistance and all the normal diode forward properties with $I_\varepsilon > 0$. In the reverse direction the emitter has very high resistance of the order of 50,000 to 500,000 ohms. In terms of the usual equivalent tee, $r_{11} = r_\varepsilon + r_b$. r_ε is assumed to have two values, one for Regions II and III and one for Region I.

The equation for this family of curves is:

$$V_\varepsilon = I_\varepsilon r_{11} + I_c r_{12} \quad (11)$$

or,

$$V_\varepsilon = I_\varepsilon(r_\varepsilon + r_b) + I_c r_b \quad (12)$$

An emitter load may be introduced as with the collector family,

$$V_{\varepsilon\varepsilon} - I_\varepsilon R_\varepsilon = V_\varepsilon \quad (13)$$

or

$$V_{\varepsilon\varepsilon} = I_\varepsilon(r_\varepsilon + r_b + R_\varepsilon) + I_c r_b$$

which is of the general form

$$V_1 = I_1 R_{11} + I_2 R_{12} \quad (15)$$

where R_{11} and R_{12} are circuit parameters.

Summarized in tabular form, the two two-terminal equations and parameter values for the three regions are shown in Table 1. Typical values for a point contact transistor are also given.

TABLE 1

$$V_{\varepsilon\varepsilon} = I_\varepsilon(r_\varepsilon + r_b + R_\varepsilon) + I_c r_b$$

$$V_{cc} = I_\varepsilon(r_m + r_b) + I_c(r_c + r_b + R_c) - I_{co}(r_c + r_b + R_c)$$

Parameter	Region I		Region II		Region III	
	Symbol	Typical	Symbol	Typical	Symbol	Typical
r_ε	r'_ε	100,000Ω	r_ε	130Ω	r_ε	130Ω
r_b	r_b	150Ω	r_b	150Ω	r_b	150Ω
r_c	r_c	18,000Ω	r_c	18,000Ω	0	--
r_m	0	--	r_m	40,000Ω	0	--

Summary of terminal voltage and current relationships with assumed approximations for the three regions of the broken-line method.

In order to derive the negative resistance characteristics of Figs. 3B, 4B and 11 it is only necessary to solve Equations 9 and 14 and substitute applicable parameter values from Table 1 for each region.

Taking the emitter characteristic first, it is desired to know $V_{\varepsilon\varepsilon}$ in terms of I_ε.

$$V_{\varepsilon\varepsilon} = I_\varepsilon(r_\varepsilon + r_b + R_\varepsilon) + I_c r_b \qquad (14)$$

$$V_{cc} = I_\varepsilon(r_b + r_m) + I_c(r_c + r_b + R_c) - I_{co}(r_c + r_b + R_c) \qquad (9)$$

A great deal of manipulative effort is saved if we solve and manipulate the more functional forms of 9 and 14 given by 10 and 15 as:

$$V_1 = I_1 R_{11} + I_2 R_{12} \tag{15}$$

$$V_2 = I_1 R_{21} + I_2 R_{22} \tag{10}$$

where $V_1 = V_{\varepsilon\varepsilon}$, $V_2 = V_{cc} - I_{co}(r_c + R_b + r_b R_c)$, $I_1 = I_\varepsilon$, $I_2 = I_c$, and generally:

$$R_{11} = r_\varepsilon + R_\varepsilon + r_b + R_b,\ R_{12} = r_b + R_b,$$

$R_{21} = r_b + R_b + r_m$, $R_{22} = r_c + R_c + r_b + R_b$

Thus:

$$V_1 = I_1\left(R_{11} - \frac{R_{12}\,R_{21}}{R_{22}}\right) + \frac{V_2\,R_{12}}{R_{22}} \tag{16}$$

which becomes generally,

$$V_{\varepsilon\varepsilon} = I_\varepsilon\left(r_\varepsilon + R_\varepsilon + r_b + R_b - \frac{(r_b + R_b)(r_b + R_b + r_m)}{r_c + R_c + r_b + R_b}\right) + \frac{\left[V_{cc} - I_{co}(r_c + R_c + r_b + R_b)\right](r_b + R_b)}{r_c + R_c + r_b + R_b} \tag{17}$$

Taking the simplest case as in Fig. 14 and assuming that $I_{co} \ll I_c$, $R_\varepsilon = 0$, $R_c = 0$, r_ε and $r_b \ll R_b$, (17) reduces to:

$$V_\varepsilon = I_\varepsilon\left[r_\varepsilon + R_b - \frac{R_b(R_b + r_m)}{r_c + r_b}\right] + \frac{V_c R_b}{r_c + R_b} \tag{18}$$

(18) becomes for the three regions:

Region I

$$V_\varepsilon = I_\varepsilon\, r'_\varepsilon + R_b\left(1 - \frac{R_b}{r_c + R_b}\right) + \frac{V_c R_b}{r_c + R_b} \tag{19}$$

If $R_b \ll r_\varepsilon'$, as is usual,

$$V_\varepsilon = I_\varepsilon r_\varepsilon' + \frac{V_c R_b}{r_c + R_b} \tag{20}$$

Region II

$$V_\varepsilon = I_\varepsilon \left[R_b + r_\varepsilon - \frac{R_b(R_b + r_m)}{r_c + R_b} + \frac{V_c R_b}{r_c + R_b} \right] \tag{21}$$

or

$$V_\varepsilon \approx I_\varepsilon R_b \left(1 - \frac{R_b + r_m}{r_c + R_b}\right) + \frac{V_c R_b}{r_c + R_b} \tag{22}$$

If $R_b \ll r_c$ and r_m, then (22) is approximately:

$$V_\varepsilon = I_\varepsilon R_b (1-\alpha) + \frac{V_c R_b}{r_c + R_b} \tag{23}$$

Region III

$$V_\varepsilon = I_\varepsilon \left[r_\varepsilon + R_b - \frac{R_b(R_b)}{R_b} \right] + \frac{V_c R_b}{R_b} \tag{24}$$

$$V_\varepsilon = I_\varepsilon r_\varepsilon + V_c \tag{25}$$

Equations (20), (22) and (25) each contain two terms. The first concerns the input resistance; the second relates to dc currents and voltages. Both terms are important, the first gives the slopes of the characteristic; the second determines the turning or trigger point voltages and currents.

In Region I the input resistance is essentially that of the reverse resistance of the emitter, r_ε. Region II resistance is negative with choice of parameters. The abosolute value is determined primarily by R_b and α, the current gain, as shown in (23). The trigger point formed by the intersection of the curves of Regions I and II is of great importance as it determines the trigger sensitivity. That is, with normally fixed biases, should the trigger point voltage given as $V_{\varepsilon x}$

vary, the trigger sensitivity will also vary. To determine $V_{\varepsilon x}$, 21 and 23 are solved simultaneously. The only solution possible is that for $I_\varepsilon = 0$. Thus,

$$V_{\varepsilon x} = \frac{V_c R_b}{R_b + r_c} \tag{26}$$

That the turning point falls on the emitter current axis, i.e. $I_{\varepsilon x} = 0$, is a consequence of the original idealization which assumed that $r_m = 0$ for $I_\varepsilon < 0$ and $r_m > 0$ for $I_\varepsilon > 0$. This is not a precise assumption and the turning point will actually lie slightly in the positive emitter current region. For very small triggers, consideration must be given to closer approximations of $r_m = f_1(I_\varepsilon)$ and $r_{11} = f_2(I_\varepsilon)$.

The consequences of (26) can be quite serious. V_c and R_b are of course fixed, but r_c is variable, both from unit to unit and with temperature, decreasing with ambients much above room temperatures. The variability of $V_{\varepsilon x}$ can result in failure to trigger with signals or, at the other extreme, lock-up at high current. This subject will be discussed later in terms of both device and circuit mitigation means.

Equation 25 for Region III is quite interesting and reveals that the saturation current is limited by r_ε, the emitter resistance. Consequently saturation currents may be quite high under short circuit conditions; that is $V_\varepsilon = 0$. For example, suppose V_c is -22 volts and using the typical parameters given in Table I,

$$I_{\varepsilon III} = \frac{V_c}{r_\varepsilon} = \frac{-22}{130} \approx -170 \text{ mA}$$

$$(V_\varepsilon = 0)$$

Such currents are likely to damage the transistor and limiting means should normally be employed. The possibility of such currents may help explain sudden decease of both transistors and meters sometimes reported.

The high-current turning or trigger point is determined by a simultaneous solution of Equation (21) and (25) which leads to:

$$I_{\varepsilon s} = \frac{V_c - V_{\varepsilon x}}{R_b\left(1 - \frac{R_b + r_m}{R_b + r_c}\right)} \tag{27}$$

or,

$$I_{\varepsilon s} \approx \frac{V_c}{R_b(1-\alpha)} \tag{28}$$

Substituting (27) into (25):

$$V_{\varepsilon s} = V_c \left[1 + \frac{r_\varepsilon}{R_b\ (1-\alpha)}\right] \tag{29}$$

The current and voltage coordinates of the lower trigger point as given by (28) and (29) are fairly fixed in terms of transistor parameters, varying only with α. It is interesting to note that the turning point will move along the Region III curve, given by (25), since r_ε is fairly constant in Regions II and III. The locus of points for $I_{\varepsilon s}$ and $V_{\varepsilon s}$ is thus along the characteristic for Region III.

The ratio of $V_{\varepsilon s}$ to $V_{\varepsilon x}$ is often of interest to estimate voltage swings or perhaps as a design objective in some switching circuits. Thus from (29) and (26),

$$\frac{V_{\varepsilon s}}{V_{\varepsilon x}} = \frac{[Rb(1-\alpha) + r_\varepsilon]\ (Rb + r_c)}{Rb^2\ (1-\alpha)} \tag{30}$$

If $r_\varepsilon \ll Rb(1- \ )$ then (30) becomes:

$$\frac{V_{\varepsilon s}}{V_{\varepsilon x}} = \frac{Rb + r_c}{R_b} \tag{31}$$

If Rb is very large, the ratio approaches unity with the implication of the existence of only two regions. This is equivalent to saying that the negative resistance becomes infinite over an infinitely short range. The proper choice of Rb in terms of (31) may well be a design problem where it is desired to have a high ratio of $V_{\varepsilon s}$ to $V_{\varepsilon x}$, as in lockout circuits.

The collector negative resistance characteristic shown in Fig. 15 may be determined in like fashion, using Equations (14) and (9), but solving for V_{cc} in terms of I_c.

$$V_{\varepsilon\varepsilon} = I_\varepsilon(r_\varepsilon + r_b + R_\varepsilon) + I_c r_b \tag{14}$$

$$V_{cc} = I_\varepsilon(r_b + r_m) + I_c(r_c + r_b + R_c) - I_{co}(r_c + R_c + r_b) \tag{9}$$

Adding R_b as before and neglecting the I_{co} term which may be arbitrarily subtracted from V_{cc}:

$$V_{cc} = I_c\left[r_c + R_b + r_b + R_b - \frac{(r_b + R_b)(r_b + R_b + r_m)}{r_\varepsilon + R_\varepsilon + r_b + R_b}\right] + \frac{V_{\varepsilon\varepsilon}(r_b + R_b + r_m)}{r_\varepsilon + R_\varepsilon + r_b + R_b} \tag{32}$$

Taking the simplest case as before with $R_\varepsilon = R_c = 0$, $r_b \ll R_b$, (31) becomes:

$$V_c = I_c\left[r_c + R_b - \frac{R_b(R_b + r_m)}{r_\varepsilon + R_b}\right] + \frac{V_{\varepsilon\varepsilon}(R_o + R_m)}{r_\varepsilon + R_b} \tag{33}$$

The equations for the three regions in terms of the assumptions of Table 1 are:

Region I

$$V_c = I_c\left[r_c + R_b - \frac{R_b 2}{r'_\varepsilon + R_b}\right] + \frac{V_{\varepsilon\varepsilon} R_b}{r'_\varepsilon + R_b} \tag{34}$$

or since, $r'_\varepsilon \gg R_b$,

$$V_c = I_c(r_c + R_b) \tag{35}$$

Region II

$$V_c = I_c\left[r_c + R_b - \frac{R_b(R_b + r_m)}{r_\varepsilon + R_b}\right] + \frac{V_{\varepsilon\varepsilon}(R_b + r_m)}{r_\varepsilon + R_b} \tag{36}$$

Approximately, where $r_\varepsilon \ll R_b$,

$$V_c = I_c[r_c(1-\alpha)] + \frac{V_{\varepsilon\varepsilon}(r_m + R_b)}{R_b} \tag{37}$$

Region III

$$V_c = I_c\left(R_b - \frac{R_b 2}{r_\varepsilon + R_b}\right) + \frac{V_{\varepsilon\varepsilon}\ R_b}{r_\varepsilon + R_b} \tag{38}$$

or:

$$V_c = I_c R_b\left(\frac{r_\varepsilon}{r_\varepsilon + R_b}\right) + \frac{V_{\varepsilon\varepsilon}\ R_b}{r_\varepsilon + R_b} \tag{39}$$

Approximately, where $r_\varepsilon \ll R_b$,

$$V_c = I_c\ r_\varepsilon + V_{\varepsilon\varepsilon} \tag{40}$$

Equation (34) and its approximation, (35) show that in Region I, the slope is, to a very good order of approximation, given by $r_c + r_b$. r_c is variable with temperature and between units and is therefore of concern. Equations (36) and (37) show the dependence of the negative resistance portion, Region II, upon the alpha characteristic. Negative resistance values will in general be higher than in the emitter characteristic. The saturation characteristic is quite similar to that of the emitter, the slope being the same and given by the emitter resistance. Comments relating to high currents apply here also.

The turning or trigger points are determined in the same fashion. Solving for I_{cx} using Equations (35) and (37):

$$I_{cx} = \frac{V_{\varepsilon\varepsilon}}{R_b} \tag{41}$$

and hence,

$$V_{cx} = V_{\varepsilon\varepsilon}\left(1 + \frac{r_c}{R_b}\right) \tag{42}$$

A physical interpretation of (41) is that there will be no positive emitter current and hence no active properties until the negative emitter bias, $V_{\varepsilon\varepsilon}$, is overcome. With $I_\varepsilon = 0$, this will occur just as $R_b I_c = V_{\varepsilon\varepsilon}$. It is to be appreciated that both I_{cx} and V_{cx} will depart directly from the relations of (41) and (42) in accordance with the degree of departure of r_m or α from a true discontinuity at $I_\varepsilon = 0$. These two equations also show that with constant α, a reasonable zero order assumption, the locus of the trigger point will be along a constant current line, I_{cx}. The variation in V_{cx} will be due to the term in (42) involving r_c. Note that if $V_{\varepsilon\varepsilon}$ is zero, that is, no emitter bias, both V_{cx} and I_{cx} will be zero. In short, the circuit will enter saturation immediately.

The high-current trigger point is given by a solution of (36) and (39). Using the same approximations,

$$I_{cs} \approx \frac{V_{\varepsilon\varepsilon}}{R_b}\left[\frac{\alpha}{(\alpha-1)}\right] \tag{43}$$

$$V_{cs} \approx V_{\varepsilon\varepsilon}\left(1 + \frac{r_{\varepsilon 1}}{R_b(\alpha-1)}\right) \tag{44}$$

The high-current point will be fairly constant since only r_ε and α are involved. Bothe parameters are reasonably constant, particularly in this high current range.

It is interesting to note the ratio of V_{cx} to V_{cs}, Equations (42) and (44), as was done before for the emitter values.

$$\frac{V_{cx}}{V_{cs}} = \frac{(Rb + r_c)(\alpha-1)}{Rb(\alpha-1) + r_\varepsilon\alpha} \tag{45}$$

If it is assumed that $r_\varepsilon\alpha \ll Rb(\alpha-1)$ then (45) becomes,

$$\frac{V_{cx}}{V_{cs}} = \frac{Rb + r_c}{R_b} \tag{46}$$

which is the same ratio obtained for the emitter characteristic case.

The equations for the current-voltage relationships looking into the base, in terms of Fig. 16, are:

$$V_b = I_b(r_b + r_\varepsilon + R_\varepsilon) + I_c(r_\varepsilon + R_\varepsilon) \tag{47}$$

$$V_{cc} = I_b(r_\varepsilon + R_\varepsilon - r_m) + I_c(r_\varepsilon + R_\varepsilon + r_c + R_c - r_m) \tag{48}$$

It is to be noted that $I_\varepsilon = -(I_b + I_c)$. Solving (47) and (48) for V_b in terms of I_b:

$$V_b = I_b \left[r_b + r_\varepsilon + R_\varepsilon - \frac{(r_\varepsilon + R_\varepsilon)(r_\varepsilon + R_\varepsilon - r_m)}{r_\varepsilon + R_\varepsilon + r_c + R_c - r_m} \right] + \frac{V_{cc}(r_\varepsilon + R_\varepsilon)}{r_\varepsilon + R_\varepsilon + r_c + R_c - r_m} \tag{49}$$

Using the simplified circuit ($R_c = 0$), the parameter values for the three regions given in Table 1 and making the the approximations that $R_\varepsilon \ll r_c$, $\left|R_\varepsilon\right| \ll \left[r_c(1-\alpha)\right]$, the equations for the three regions become:

Region I

$$V_b = I_b \left[\frac{r_c r'_\varepsilon}{r_c + r_\varepsilon} \right] + \frac{V_{cc} r'_\varepsilon}{r'_\varepsilon + r_c} \tag{50}$$

Region II

$$V_b = I_b\left[r_b + \frac{r_\varepsilon + R_\varepsilon}{1-\alpha}\right] + \frac{V_{cc}(R_\varepsilon + r_\varepsilon)}{r_c(1-\alpha)} \tag{51}$$

Region III

$$V_b = I_b r_b + V_{cc} \tag{52}$$

As mentioned previously the base characteristic is short-circuit stable in distinction to the emitter and collector characteristics which are open-circuit stable. It would have been more appropriate to solve the base circuit in terms of conductances rather than resistances. One result is that the negative resistance obtained in this connection is quite low in magnitude; the negative conductance is quite high, however.

Solutions for the trigger or turning points are determined as before by solving (50) and (51) for V_{bx} and I_{bx} and (51) and (52) for I_{bs} and V_{bs}. Thus, approximately,

$$I_{bx} = -\frac{V_{cc}}{r_c} \tag{53}$$

$$V_{bx} = 0 \tag{54}$$

$$I_{bs} = V_{cc}\left[\frac{1-\alpha}{R_\varepsilon + r_\varepsilon}\right] \tag{55}$$

$$V_{bs} = V_{cc}\left[1 - \frac{(\alpha-1)r_b}{R_\varepsilon + r_\varepsilon}\right] \tag{56}$$

As shown in Fig. 15, the low-current turning point lies on the voltage axis. This is similar to the emitter case but opposite in sense. The reasons are the same: the assumption that $r_m = 0$ for $I_\varepsilon < 0$ and $r_m > 0$ for $I_\varepsilon > 0$. Shifting of the low-current trigger point along the voltage axis to change trigger conditions is directly controlled by r_c as in the emitter case.

An example to illustrate the foregoing analysis is shown in Fig. 17. A circuit including a collector load resistor has been chosen for more complete illustration. The heavy line is an experimental curve and the light line is the approximated broken-line characteristic.

The low-current trigger point, $V_{\varepsilon x}$, $I_{\varepsilon x}$, is easily derived from the Region II calculation since $I_{\varepsilon x} = 0$. From (17) assuming $I_{co} = 0$ and using the small signal values of Fig. 17,

$$V_\varepsilon = I_\varepsilon \left[120 + 160 + 6.8k - \frac{(160 + 6.8k)(160 + 6.8k + 50k)}{19k + 2.2k + 160 + 6.8k} \right] - \frac{45(160 + 6.8k)}{19k + 2.2k + 160 + 6.8k} \qquad (57)$$

All parameters have been included in order to illustrate the degree of approximation made in the analysis. $V_{\varepsilon x}$ is given by (57) when $I_\varepsilon = 0$; hence,

$$V_{\varepsilon x} \approx - \frac{45(6.8k)}{28k} = -10.9 \text{ volts} \qquad (58)$$

The calculated $V_{\varepsilon x} = -10.9$ corresponds quite well to the measured -11 volts. The approximation was of course that $r_b \ll R_b$.

The first term of (57) is the slope of Region II and is given by:

$$V_\varepsilon = I_\varepsilon \left[6.8k - \frac{(6.8k)(64.8k)}{28k} \right] = I_\varepsilon (6.8k - 15.7) = I_\varepsilon 8.9k \qquad (57)$$

The calculated slope of Region II is -8900 ohms whereas the measured value is about -9200 ohms. Proper choice of operating

point for the measurement of small signal parameters, principally r_c and r_m or α, yield these close checks.

Using (17) again the equation for Region III is

$$V_b = I_\varepsilon \left[r_\varepsilon + \frac{R_b R_c}{R_b + R_c} \right] + \frac{V_{cc} R_b}{R_b + R_c} \qquad (59)$$

or:

$$V_b = I_\varepsilon \left[120 + \frac{(6.8k)(2.2k)}{6.8k + 2.2k} \right] - \frac{45\ (6.8k)}{6.8k + 2.2k}$$

$$V_b = I_\varepsilon (1785) - 34 \qquad (60)$$

Plotting (60) results in an error in the dc value but quite good agreement in slope. If r_b is determined from the output saturation characteristic and is added to R_b in (59) the check is surprisingly good.

The Region I characteristic yields a perfect check due to the high value of r'_ε. The approximation given by (20) is quite good.

The photo properties of transistor devices may be employed in switching also. That photo-diodes can be coupled to trigger circuits is obvious and will not be discussed. The example chosen is that of direct triggering of a trigger circuit by means of light. The photo diode is in simple essence just the collector portion of a transistor with the emitter replaced by light. Thus there is only a collector family since r_{11} and r_{12} are zero. Neglecting I_{co}, the equation for the collector family is:

$$V_c = kL\ r_{22} + Ic\ r_{22} \qquad (61)$$

where k is the quantum efficiency and has the dimensions of amperes/Lumen and L is the light in Lumens. The equation for an

output family where there are transfer properties due to both an emitter and to light is, accounting for signs and neglecting I_{co},

$$V_c = I_\varepsilon r_{21} + I_c r_{22} + kL r_{22} \tag{62}$$

which is equivalent to:

$$V_{c'} = I_\varepsilon r_{21} + I_c r_{22} \tag{63}$$

where:

$$V_{c'} = V_c - kL r_{22} \tag{64}$$

Thus the photo effects may be introduced by replacing V_c with $V_{c'}$ in the analysis of Section II and taking r_{22} into proper account for each region.

The emitter negative resistance characteristic as an example is shown on Fig. 22. The load line has been chosen to result in a stable operating point as at a. The voltage of point a in the simple circuit case is given by:

$$V_{\varepsilon x} = \frac{V_{cc} R_b}{R_b + r_c} \tag{26}$$

Since $V_{cc} = V_c$ in the circuit of Fig. 22, an increment of light will increase I_c and change $V_{\varepsilon x}$ to $V_{\varepsilon x'}$ as :

$$V'_{\varepsilon x} = \frac{(V_{cc} - kLr_{22})R_b}{R_b + r_c} \tag{65}$$

and thus:

$$V'_{\varepsilon x} = V_{\varepsilon x} - kLR_b \tag{66}$$

If the change in $V_{\varepsilon x}$ exceeds the trigger sensitivity as is shown in Fig. 21, the upper trigger point will fall below the load line and hence the only stable point will be at b in

the saturation region. The point b will be maintained as the stable operating point as long as the light remains. Removal of the light will result in return to the normal operating point a. If $R_\epsilon < |-R_{in}|$ triggering will result in lock-up in the saturation region, and would require either a reduction in light from a given nominal value or the use of an electrical trigger. Light may be employed to trigger astable circuits as well.

From analysis of this sort it can be deduced that certain transistor characteristics must be controlled in a reproducible manner and must be reasonably constant under operating conditions. r_c is of particular importance since it largely determines the low-current trigger point. In particular the collector resistance at $I_\epsilon = 0$ is of concern. α is of importance both for gain and step properties at $I_\epsilon = 0$. Good saturation characteristics are also necessary. The reverse resistance of the emitter is an important factor which does not appear in small signal applications.

The broken line method can thus lead to objective device design requirements and properly employed will reasonably predict circuit behavior. The analysis will be accurate to the degree to which the original idealized characteristics are approximated. Other idealizations are necessary where family characteristics are different from those employed here. Extended, it will predict fine detail to the degree to which device parameters are known and to which patience will permit. Transient behaviors and analysis have not been discussed and are needed for a more complete understanding.

III DEVICE PROPERTIES

The foregoing Section has brought out that all of the equivalent transistor parameters are of importance to switching applications. The collector resistance should be high; the emitter resistance low; α or r_m high and incorporating a step characteristic, conveniently at $I_\epsilon = 0$. Although it has not yet been brought out too clearly, r_b should be low. Reverse emitter resistance should be high. That all parameters should be constant with operating conditions and uniform from unit to unit goes virtually without saying.

The choice of parameters for measurement to define characteristics for large signal applications must be a subtle compromise between the almost infinite sets that could be determined and those which are practical. Further it is desirable to measure the parameters involved under conditions which result in the most useful parameter values. For example, r_c is an important parameter. In most instances the parameter actually concerned is the dc large-signal collector resistance at $I_\varepsilon = 0$ and not the small-signal r_c at a mid-range operating point. If consideration is given to I_{co} these two values will agree quite closely. A present compromise has been to measure the collector current at constant dc collector voltage with $I_\varepsilon = 0$ and term this value r_{co}. Thus $I_c r_{co} = I_c r_c - I_{co} r_c$. The discrepancy between r_{co} and r_c is not nearly so noticeable in point contact devices, where the two are nearly equal, as in junction devices where r_c may be many times higher than r_{co} at low voltages. r_{co} and r_c are illustrated in Fig. 18. Fig. 18 has been exaggerated in order to bring out this point. The assumptions that $r_c \gg r_b$ and hence $r_{22} \approx r_c$ are made.

Since $r_{22} = r_c + r_b$, the saturation line in Fig. 18 affords a good measure of r_b. Hence a voltage noted as V_{c3} in Fig. 18 has been chosen as a limit with constant I_c and sufficient I_ε to insure saturation. V_{c3} is measured at $I_\varepsilon = +3$ mA and $I_c = 5.5$ mA. As a limit such a test assures a good saturation characteristic. Small signal measurements of r_b at a mid-range operating point have also been made.

A gross evaluation of α can be made from the above data, but it will be a lower value than actual and will not afford a great deal of information concerning the nature of α as a step function. Therefore, and since alpha is relatively easy to measure, three small-signal tests have been presently employed. These are illustrated in Fig. 19 α_3 is in the negative emitter current range and is a maximum limit intended to insure cutoff. α_2 is in the low positive current range, is a minimum limit, and is intended to insure an α step function in conjunction with α_3. α_1 is a minimum limit in the mid-range.

The reverse emitter resistance, r_ε', is conveniently measured by a $-I_\varepsilon$ maximum limit at a nominal negative emitter voltage.

Small signal parameter values of r_ε and r_m are also measured at the same time as r_c and r_b.

A typical family of characteristics for the open circuit parameters, $r_{11} = r_\epsilon + r_b$, $r_{12} = r_b$, $r_{21} = r_m + r_b$ and $r_{22} = r_c + r_b$ are given in Fig. 20. Distribution in parameter values are shown in Fig. 20. The unit-to-unit variations involved compare favorably to those expected in electron tubes.

Transistor parameters are quite constant with temperatures below room temperature to at least - 70°C. Above room temperatures some of the parameters are variable. r_ϵ and r_b are fairly constant, changing very little to 70°C. r_c and r_m decrease fairly rapidly, maintaining a ratio such that alpha rises slightly. r'_ϵ and r_{co} change most rapidly and while both of these parameters are of little consequence in small-signal applications, are quite important in switching. This is particularly true of r_{co} which controls trigger or turning points of the negative resistance characteristic.

Early transistors might exhibit a change in r_{co} at 60°C of 3 to 1 or more from room temperature values. The transistors of which the data in Figs. 20 and 21 are typical have an r_{co} temperature coefficient of about 3/4%/°C. That is the room temperature value of r_{co} might be reduced to 70% at 70°C. This implies a corresponding reduction in variation in trigger sensitivity.

Variation in characteristics will arise from self-engendered heat, that is, dissipation. Transistors may be thermally unstable under constant voltage conditions. Since the switching properties are exhibited under short circuit or constant voltage terminations thermal properties are of concern. The limitations involved are similar to those of any positive feedback circuit. If the thermal loss through radiation and conduction exceeds the heat input, the system will be stable. The practical significance is to place limitations on dissipation and to employ designs which result in rapid heat loss. Other design criteria such as miniaturization may limit the latter.

If perfect switching characteristics were obtainable dissipation would be of little consequence in switching. This is akin to saying that neither a short circuit nor an open circuit dissipates any energy. Further the perfect device has zero transitional time and therefore involves no loss. The transistor has finite resistance both open and closed and a finite although rapid transition time. There is some advantage however. A constant dissipation curve shown as a dotted line

has been included in Fig. 18. Small-signal operation at mid-range currents and voltages results in fairly low limitations on both current and voltage. The intersection with the r_{22} cut-off line ($I_\varepsilon = 0$) is at fairly high voltage. Similarly, the intersection with the saturation curve is at high current. Thus for constant dissipation, approximately,

$$\text{Cutoff:} \quad P_d \approx \frac{V_c^2}{r_{co}} \tag{67}$$

$$\text{Saturation:} \quad P_d \approx I_c^2 \, r_\varepsilon \tag{68}$$

Depending upon the circuit the assumed dissipation limit may or may not be exceeded during the transitions. Should the limit be exceeded, and substantially so, there are normally no serious consequences due to the very rapid transitions and consequent low thermal energy generated.

Transistors may not be able to tolerate excess dissipation on this basis if the circuits are slow, that is with transitional times in excess of perhaps a few tenths of a microsecond. Such conditions may arise, for example, if loads are inductive. In many such cases, shunting capacitor networks will often permit a rapid transition with consequent transfer of current to the inductive load.

Practically, more dissipation can be tolerated in the saturated condition than in cutoff due to the high sensitivity of r_{co} to temperature. r_ε and r_b are much less sensitive to temperature and in any event any decrease is of advantage. Decrease in r_b and r_ε with dissipation may result in higher saturation currents on a dc basis than with transients. When full account is taken of the non-linearity of r_ε, the saturation currents may be surprisingly high. For example, currents of 100 mA at 2 volts drop, emitter to collector, have been employed on a dc basis. There may be hysteresis in such operation, however. That is, r_{co} may be temporarily lowered on returning to a low dissipation cutoff from a high dissipation saturation until a new thermal equilibrium is established.

The frequency response of point contact transistors is sufficiently good to insure switching type operation with

rise times of the order of 0.1 to 0.01μS. Fall times are somewhat longer due to the hole storage effect. In regenerative circuits operating speeds are faster than might be imagined from small-signal low frequency cutoffs. Reliable operation with rise times of 0.1μS are obtained with only nominal attention to frequency cutoff. 0.02μS requires a 10 Mc lower limit. Junction transistors are substantially slower.

Accurate life estimates are difficult due to the rapid rate of development, the relative age of the transistor and the number of parameters involved. A given device is quite likely to be obsolete and forced to give way to an improved version before sufficient models can be obtained for life tests. A small quantity of transistors having properties similar to those of Fig. 20 and 21 have been operated for over 6000 hours with an indicated life of 30,000 hours. Other somewhat similar transistors with longer life histories have indicated lives of better than 70,000 hours. The pattern appears to be similar to that of electron tubes - an early failure and change rate followed by a very slow exponential rate. It is believed that life is extended by low power operation and is decreased by high temperature operation.

The relatively high noise level of transistors does not appear to be a significant problem at present when considered in terms of automata. Systems employing switching type circuits in pulse communication will of course be concerned. It is mildly suggested that the non-concern for noise is largely a reflection of the ease with which high magnitudes of state changes are obtained. With objective design toward low power and low operating levels, noise will undoubtedly set a lower limit as in small signal transmission.

The extreme resistance of the transistor to shock and vibration with a consequent absence of microphonism may in some applications, result in effective lower noise. Shocks in excess of 20,000G have resulted in no damage. No evidences of current modulation in excess of noise have been detected with vibrational forces of the order of 100G at frequencies as high as 1000 cycles in tests on the transistor of Fig. 1. Transistors may be included in plastic embedded circuits without change of characteristics.

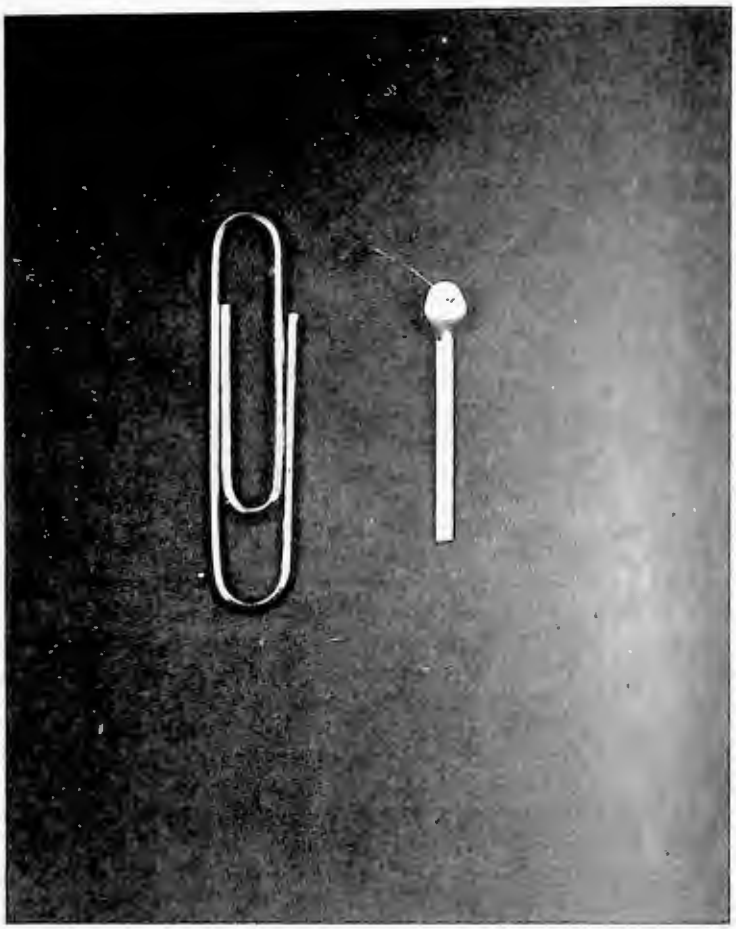
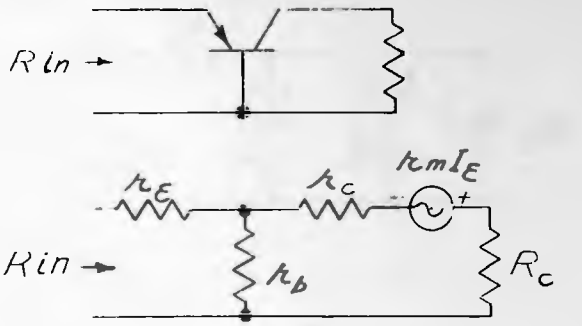
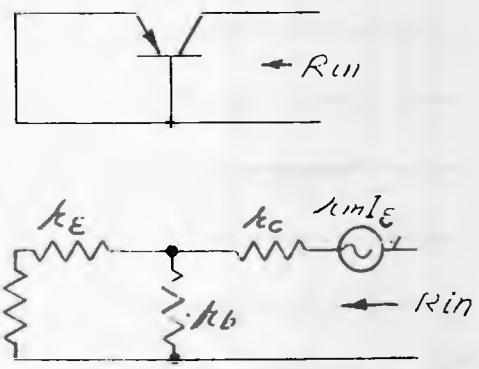


FIG. 1 POINT-CONTACT "BEAD" TRANSISTOR



$$R_{in} = R_{11} - \frac{R_{12} h_{21}}{R_{22}} = h_E + h_b - \frac{h_b(h_b + h_m)}{h_c + h_b + h_c}$$

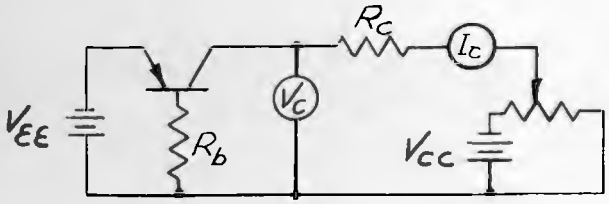
FIG. 2A

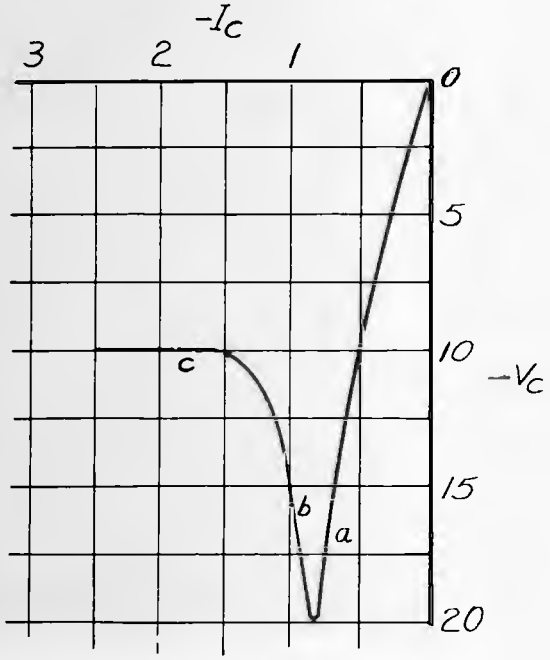


$$R_{in} = R_{22} - \frac{K_{12} K_{21}}{R_{11}} = h_c + h_b - \frac{h_b(h_b + h_{in})}{h_E + h_b + R_E}$$

FIG. 2B

TRANSISTOR EMITTER AND COLLECTOR INPUT RESISTANCES.





COLLECTOR NEGATIVE RESISTANCE CHARACTERISTIC.

FIG. 3

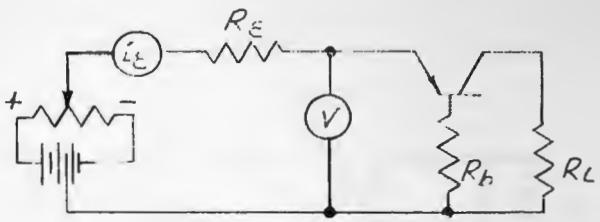


FIG. 4A

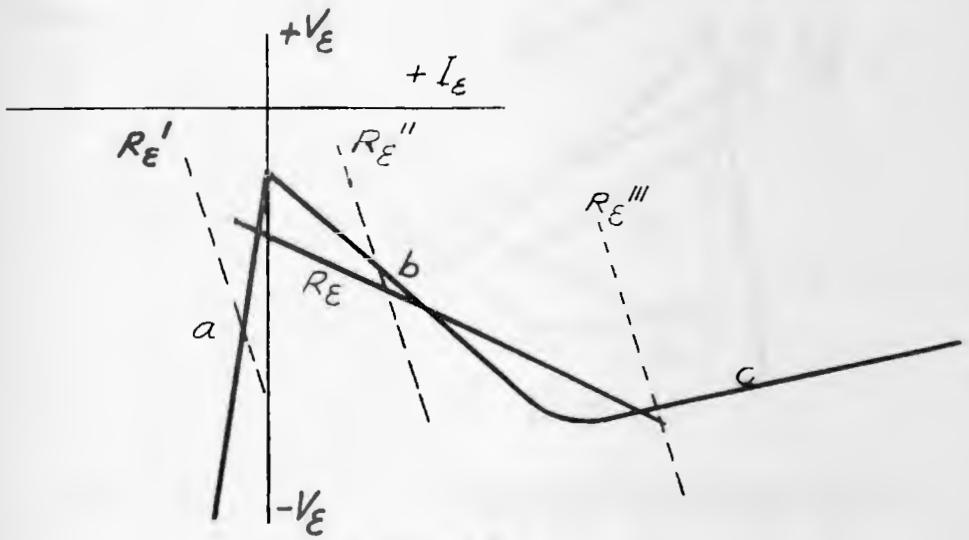


FIG. 4B

EMITTER NEGATIVE RESISTANCE CHARACTERISTIC

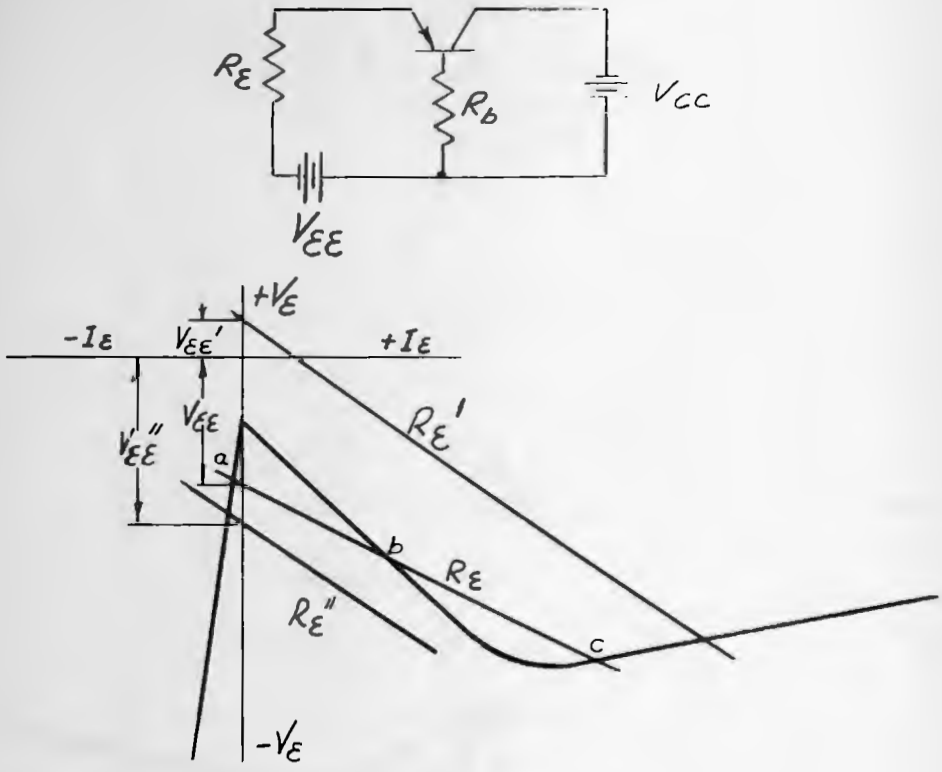
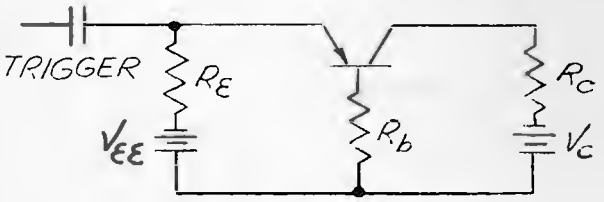


FIG. 5

EMITTER NEGATIVE RESISTANCE CHARACTERISTIC



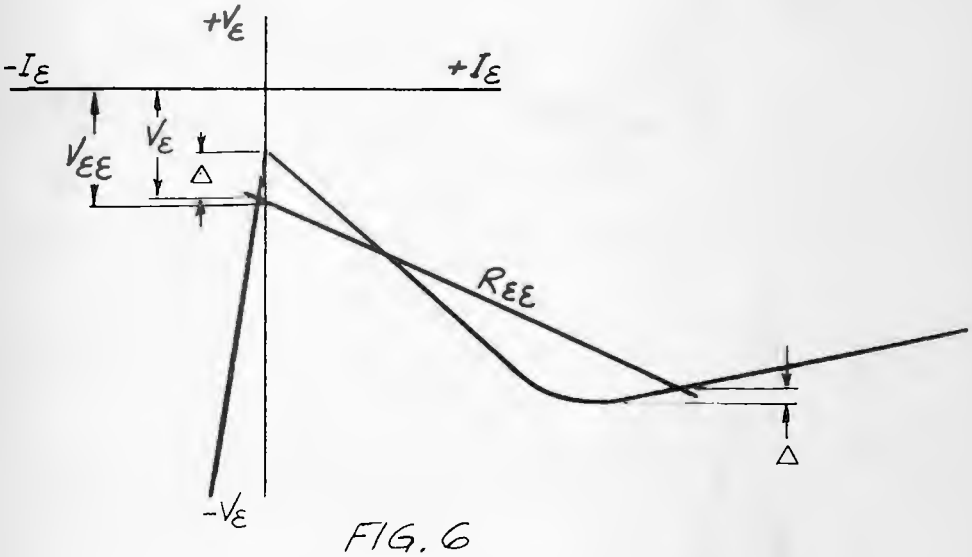


FIG. 6

BISTABLE CIRCUIT SHOWING TRIGGER SENSITIVE, Δ

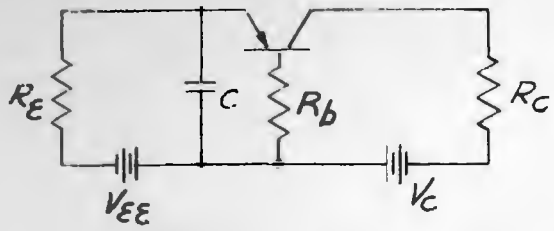
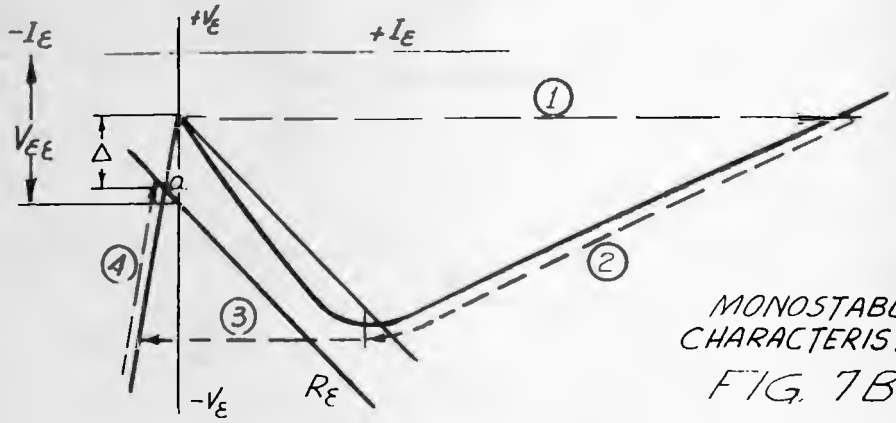
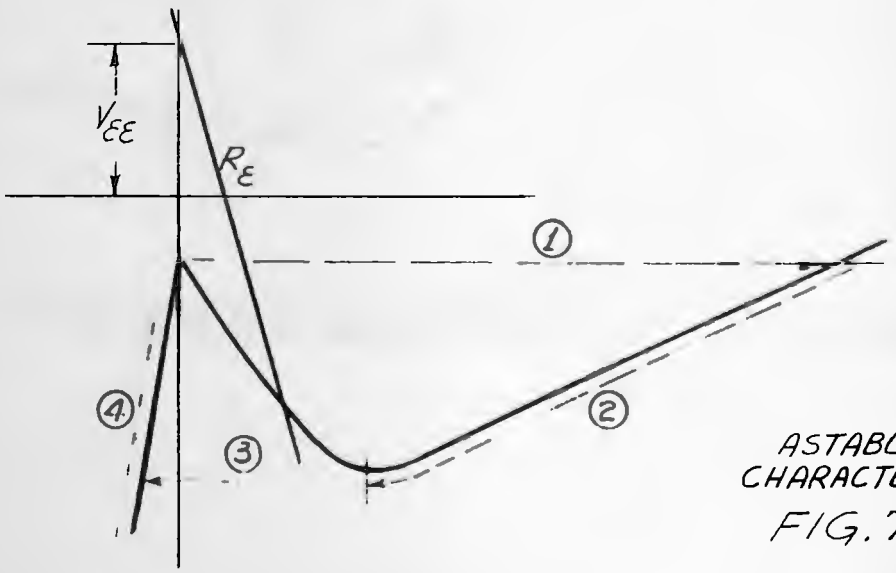


FIG. 7A



MONOSTABLE CHARACTERISTIC

FIG. 7B



ASTABLE CHARACTERISTIC

FIG. 7C

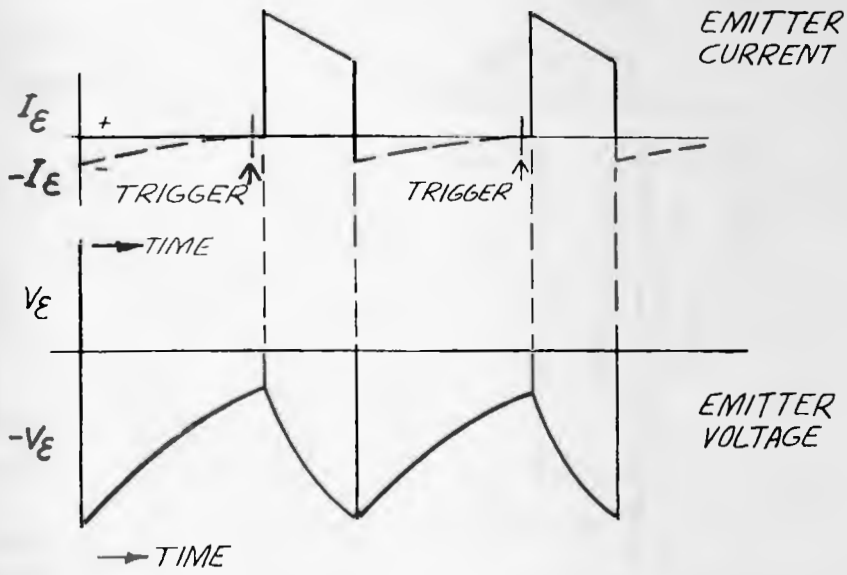
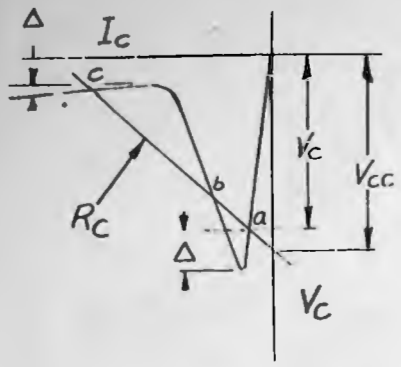


FIG. 8

MONOSTABLE RELAXATION OSCILLATOR
WAVE FORM



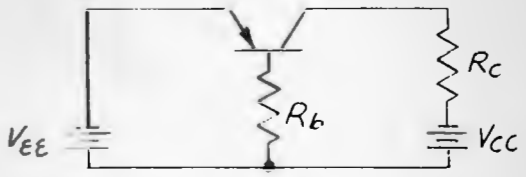
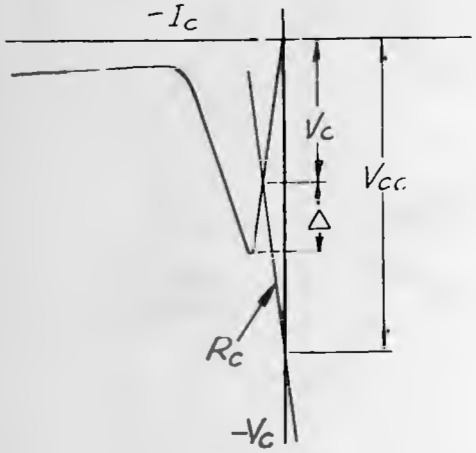


FIG. 9A

COLLECTOR
BISTABLE CHARACTERISTIC



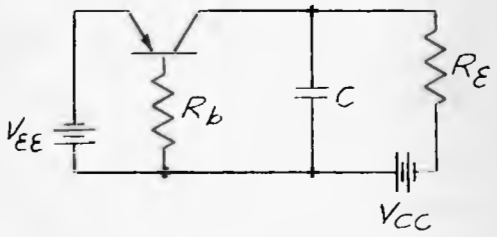
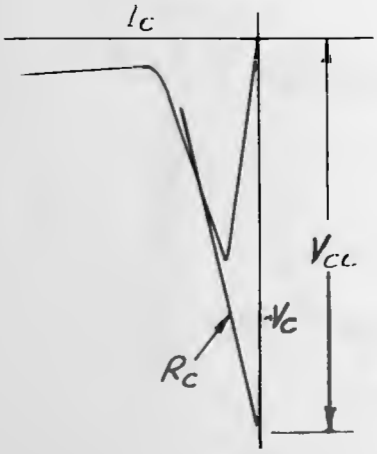


FIG. 9B

COLLECTOR
MONOSTABLE CHARACTERISTIC



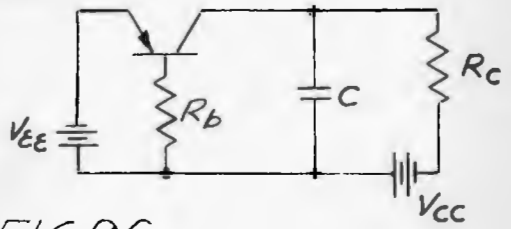
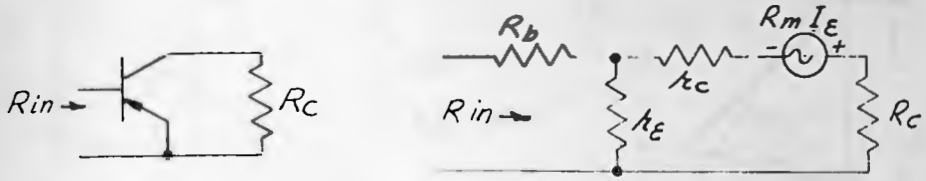


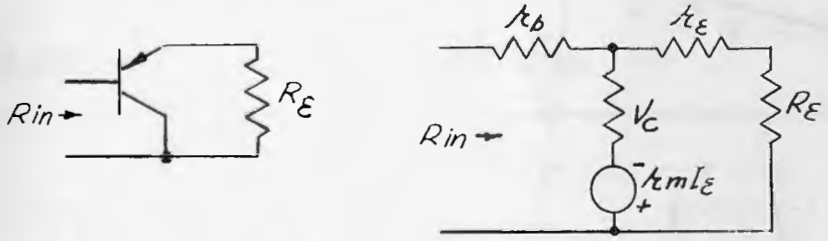
FIG. 9C

COLLECTOR
ASTABLE CHARACTERISTIC



$$Rin = r_b + r_\varepsilon + \frac{r_\varepsilon (r_m - r_\varepsilon)}{R_C + r_\varepsilon + r_c - r_m}$$

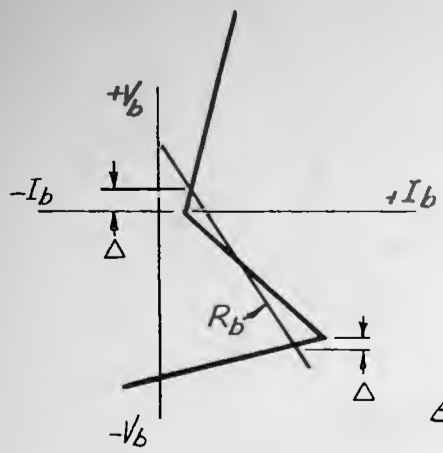
FIG. 10A

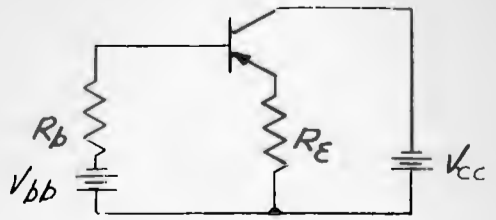


$$Rin = r_b + r_c - r_m + \frac{r_c (r_m - r_c)}{R_L + r_\varepsilon + r_c - r_m}$$

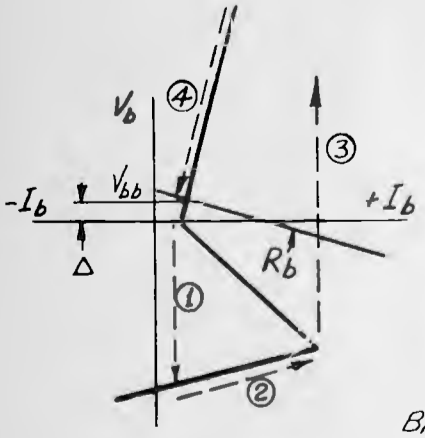
FIG. 10B

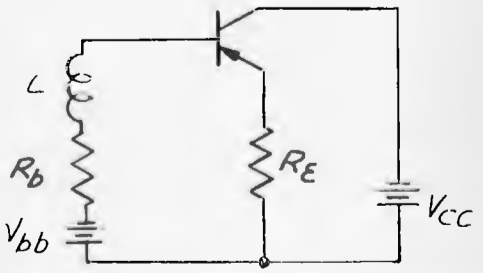
BASE INPUT RESISTANCE



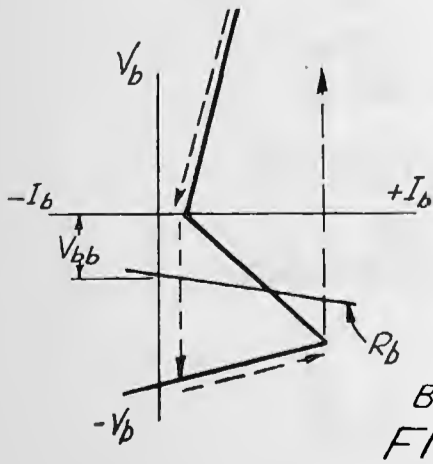


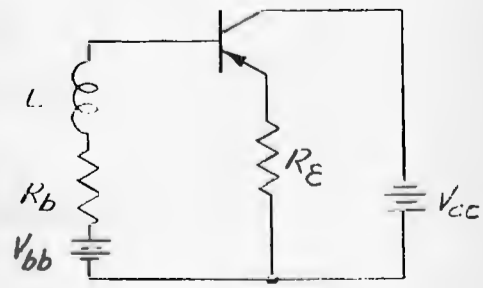
BASE BISTABLE CHARACTERISTIC





BASE MONOSTABLE CHARACTERISTIC





BASE ASTABLE CHARACTERISTICS

FIG. 11

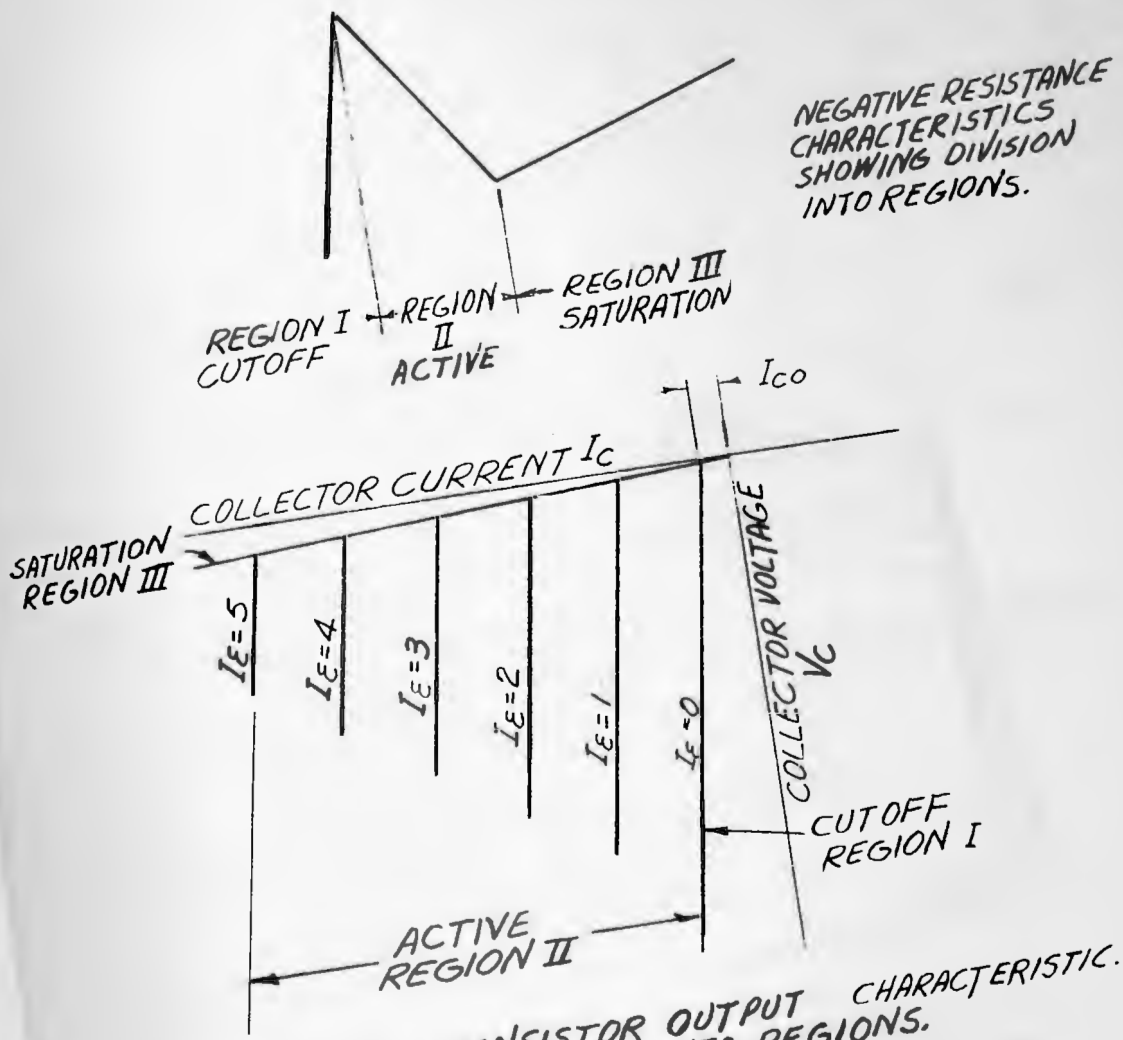
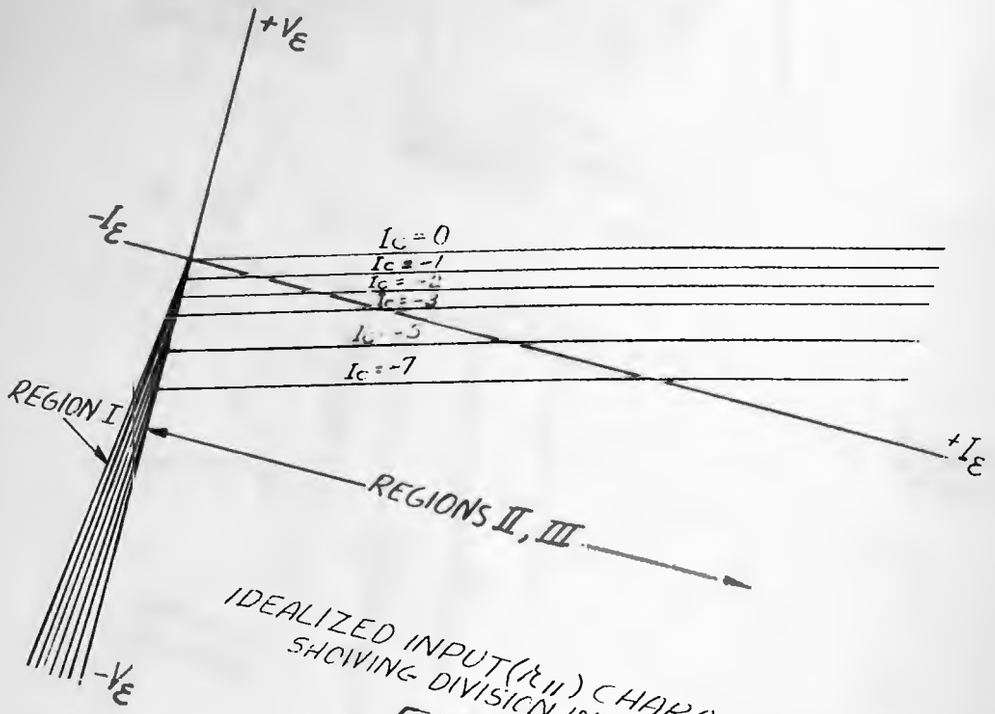
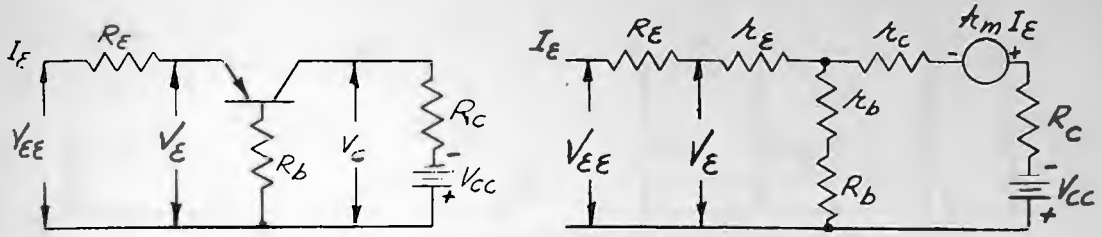

NEGATIVE RESISTANCE CHARACTERISTICS SHOWING DIVISION INTO REGIONS.
REGION I CUTOFF
REGION II ACTIVE
REGION III SATURATION
I_{CO}
COLLECTOR CURRENT I_C
SATURATION REGION III
$I_E=5$
$I_E=4$
$I_E=3$
$I_E=2$
$I_E=1$
$I_E=0$
COLLECTOR VOLTAGE V_C
CUT OFF REGION I
ACTIVE REGION II
IDEALIZED TRANSISTOR OUTPUT CHARACTERISTIC. SHOWING DIVISIONS INTO REGIONS.

FIG. 12

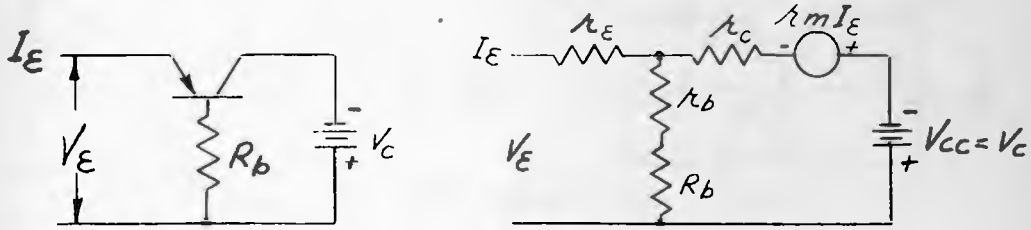


IDEALIZED INPUT(R_{11}) CHARACTERISTIC SHOWING DIVISION INTO REGIONS.

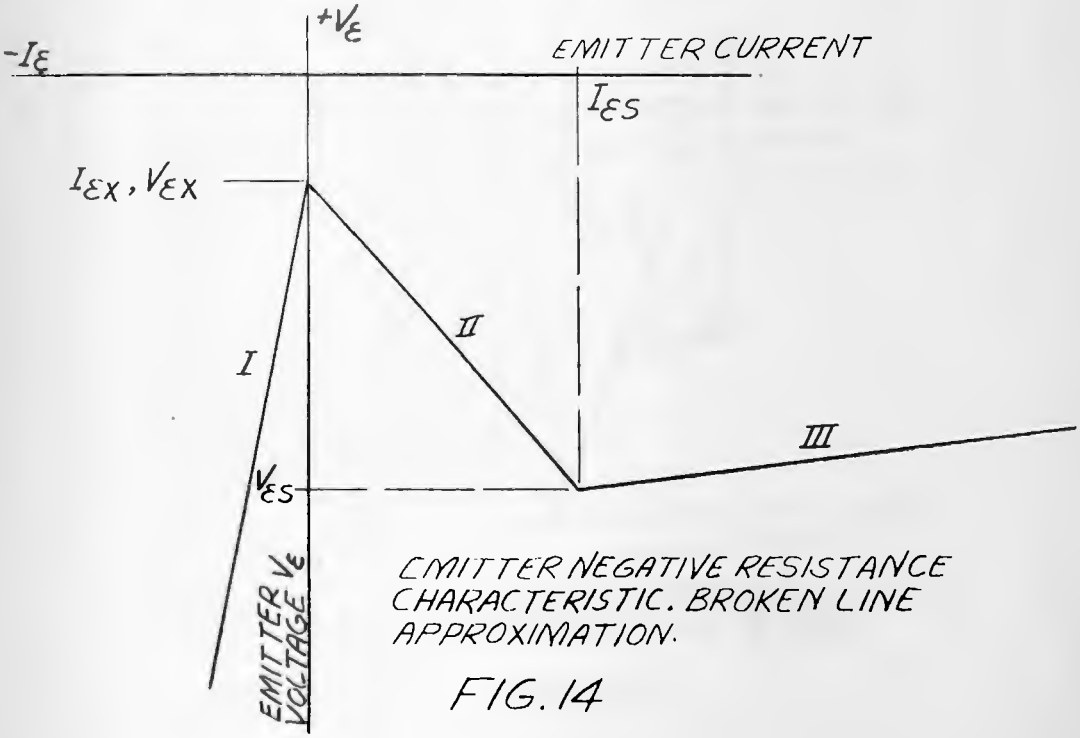
FIG. 13



GENERALIZED EMITTER CIRCUIT

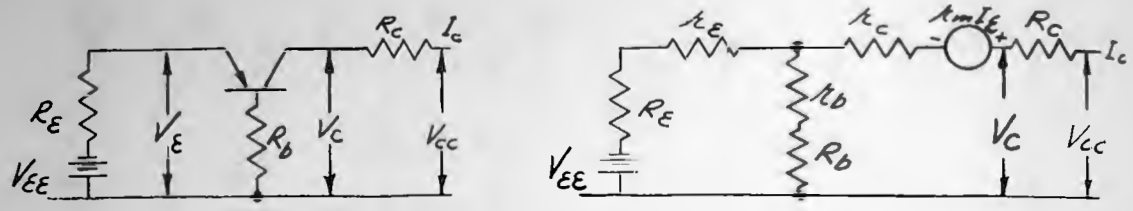


SIMPLIFIED EMITTER CIRCUIT

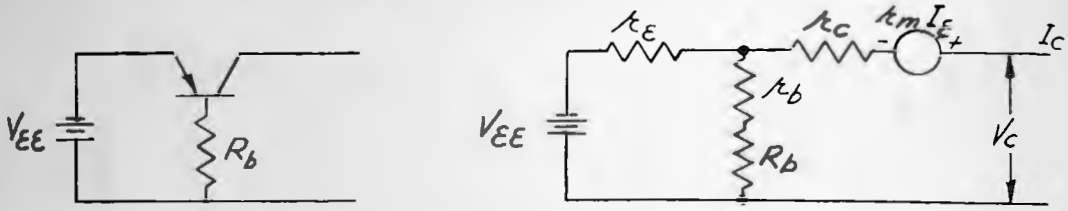


EMITTER NEGATIVE RESISTANCE CHARACTERISTIC. BROKEN LINE APPROXIMATION.

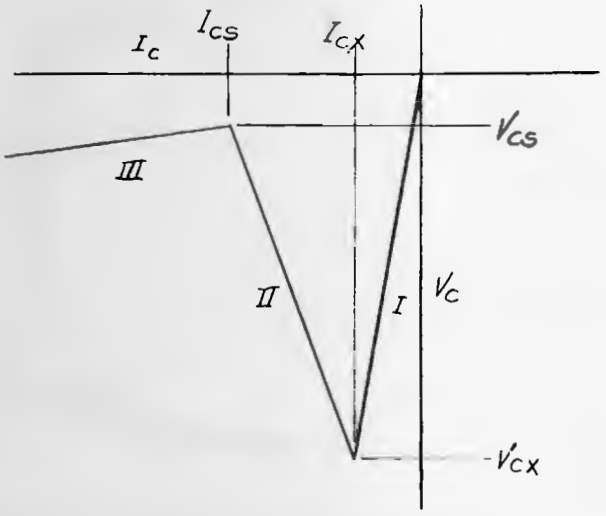
FIG. 14



GENERALIZED COLLECTOR CIRCUIT

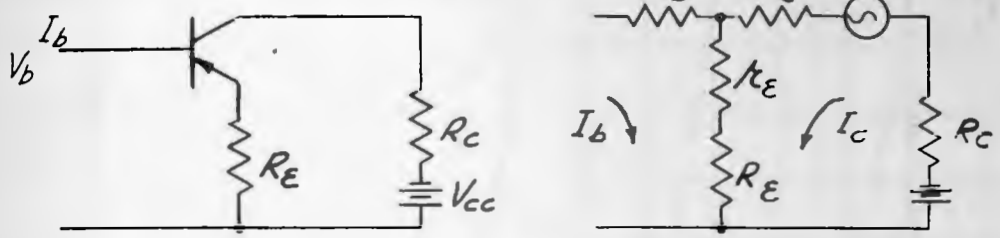


SIMPLIFIED COLLECTOR CIRCUIT

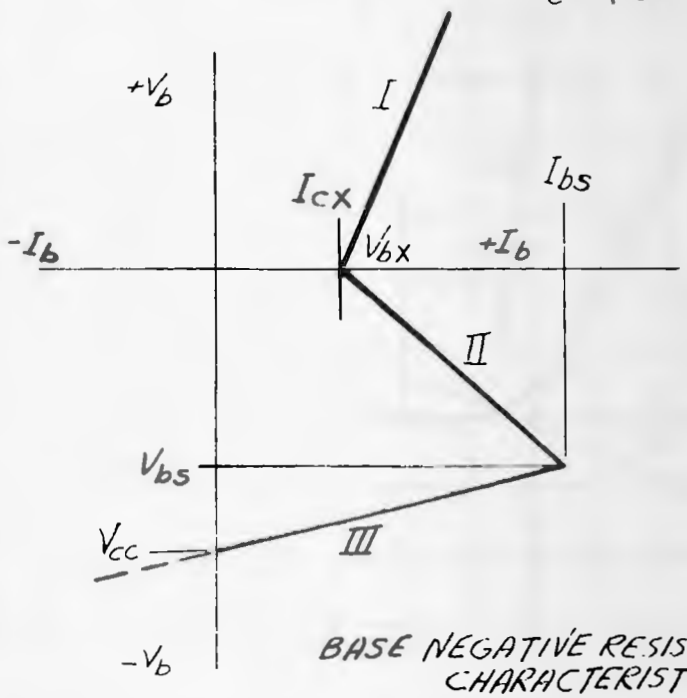


COLLECTOR NEGATIVE RESISTANCE CHARACTERISTIC
BROKE LINE APPROXIMATION

FIG. 15

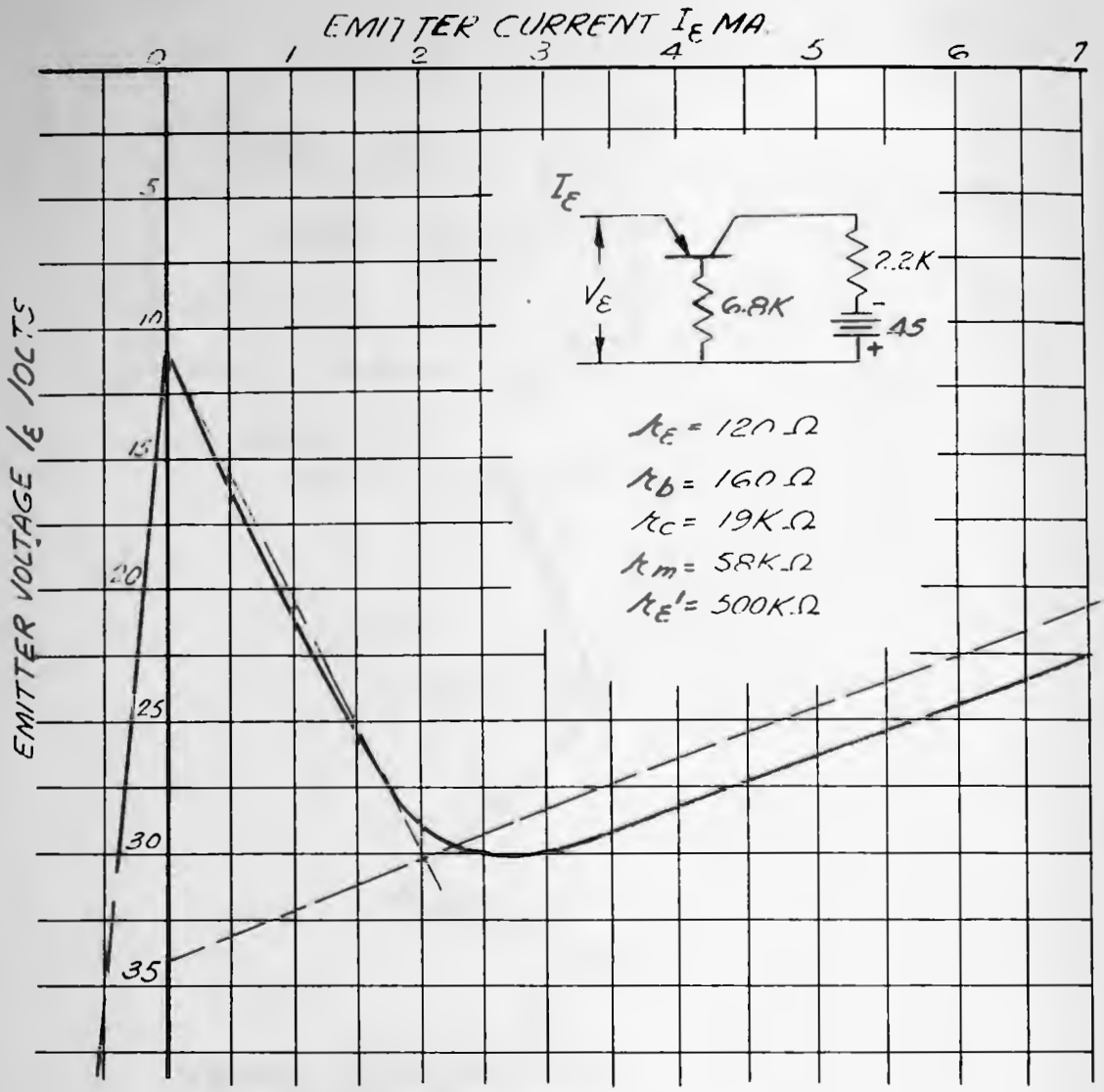


$$I_E = -(I_b + I_c)$$



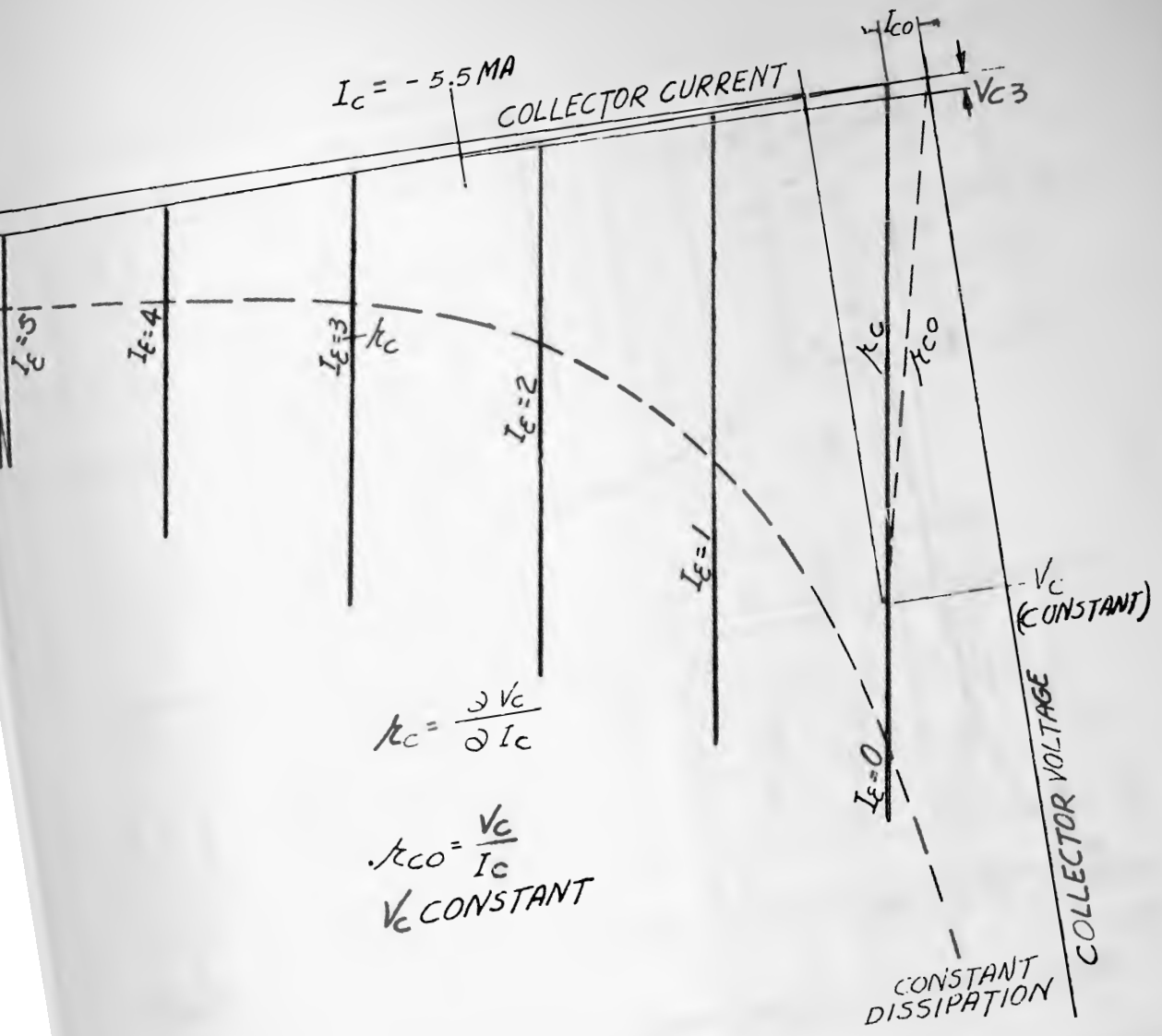
BASE NEGATIVE RESISTANCE CHARACTERISTIC

FIG. 16



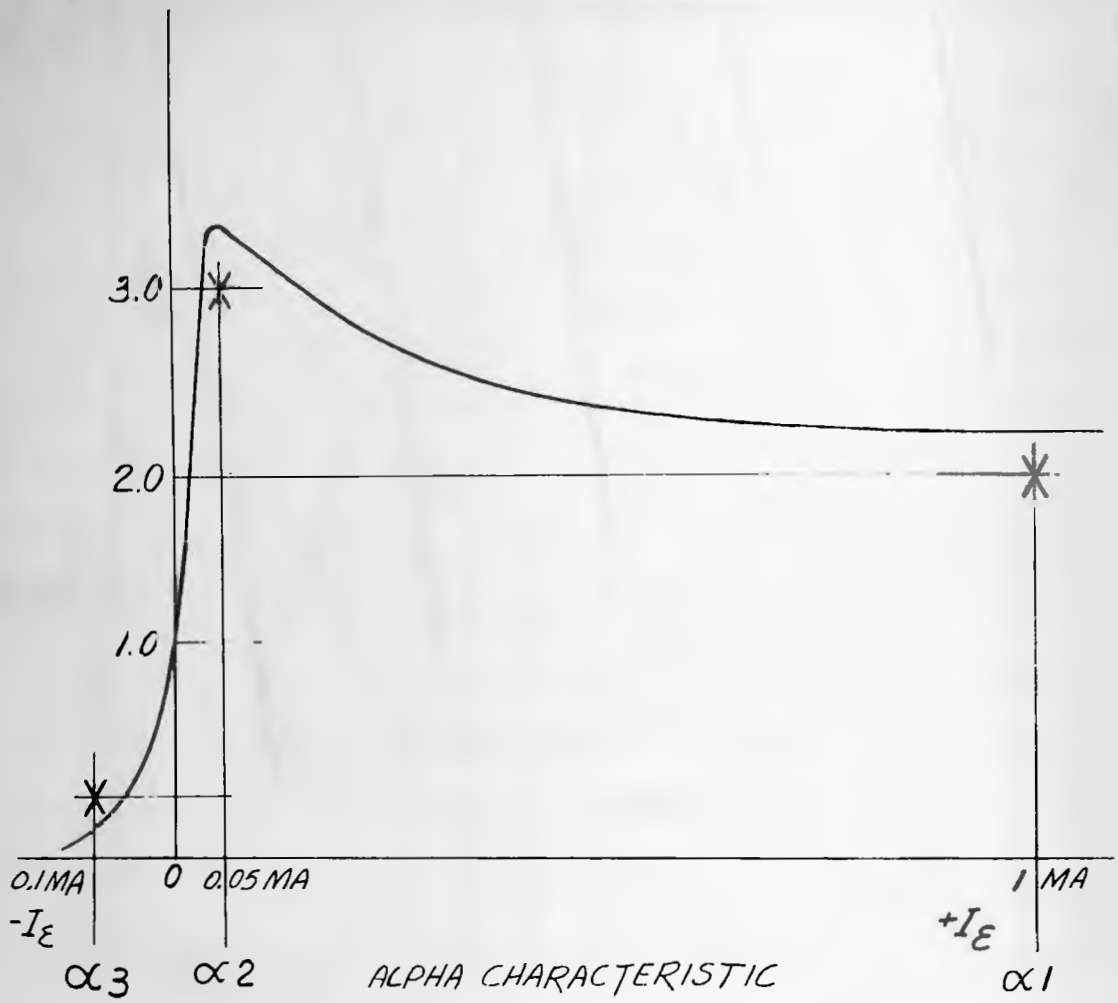
MEASURED AND CALCULATED EMITTER NEGATIVE RESISTANCE CHARACTERISTIC.

FIG. 17



IDEALIZED TRANSISTOR OUTPUT CHARACTERISTIC SHOWING PARAMETERS

FIG. 18



ALPHA CHARACTERISTIC

FIG. 19

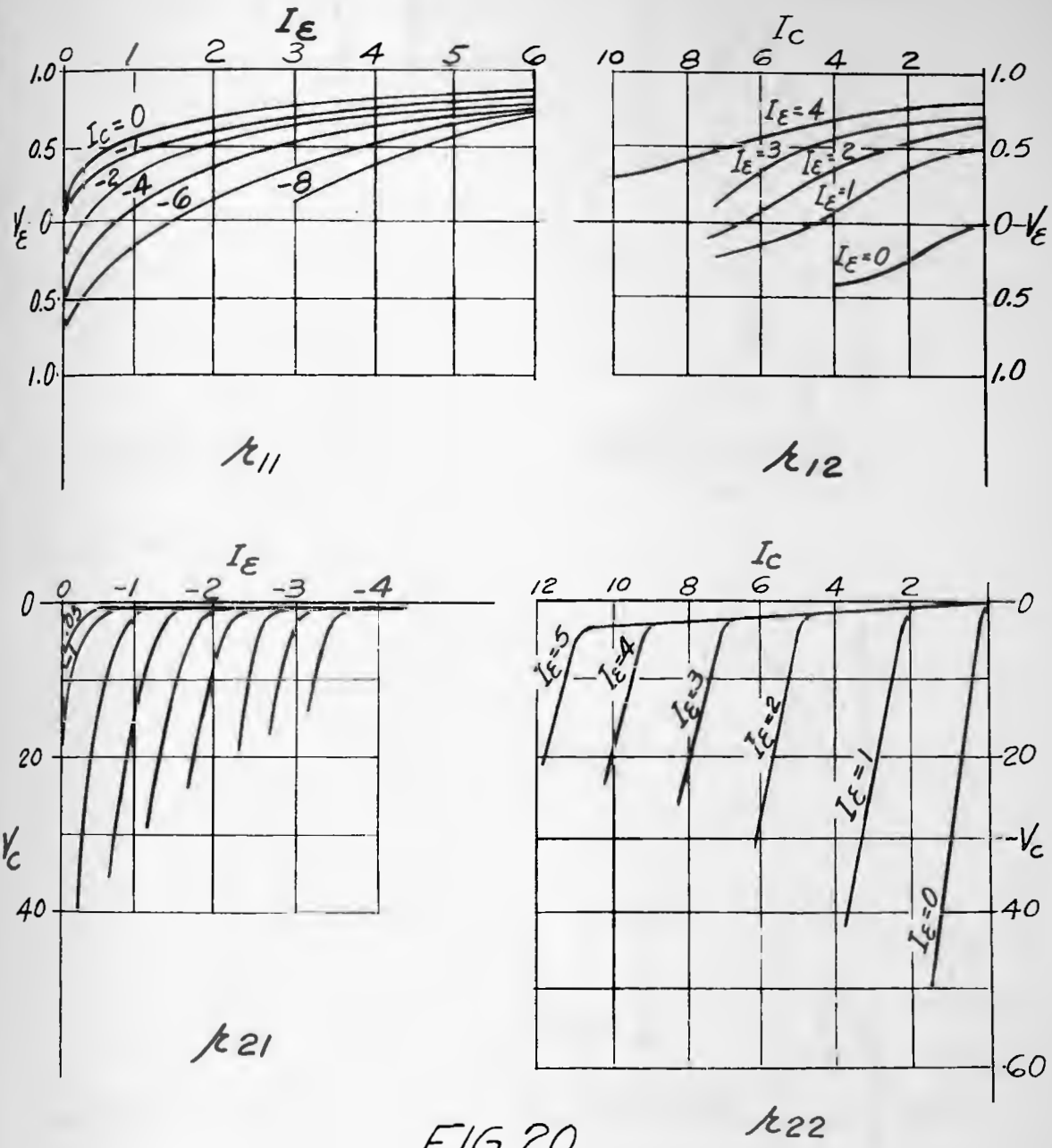


FIG. 20
STATIC CHARACTERISTICS

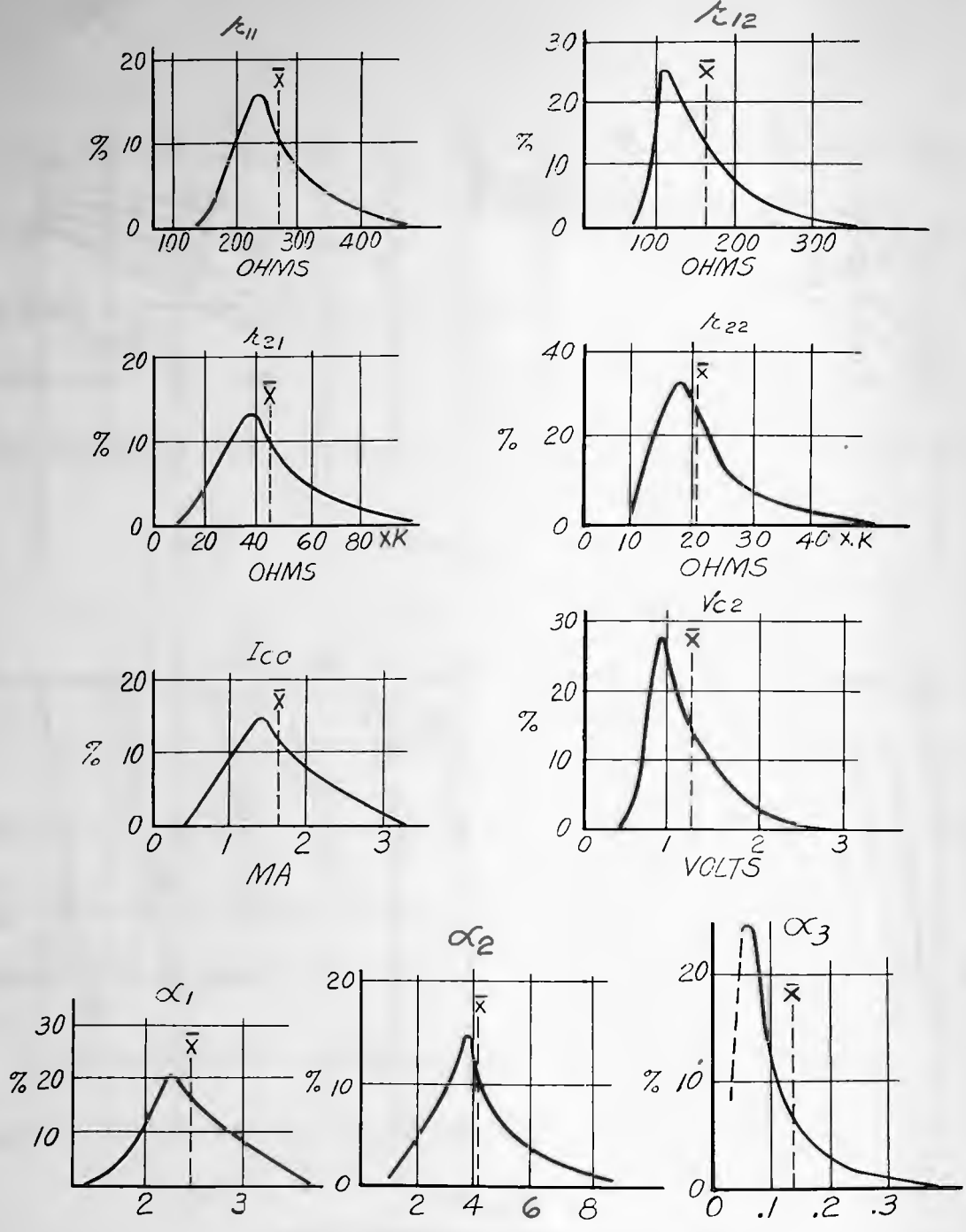


FIG. 21

TYPICAL SWITCHING TRANSISTOR PARAMETER DISTRIBUTION

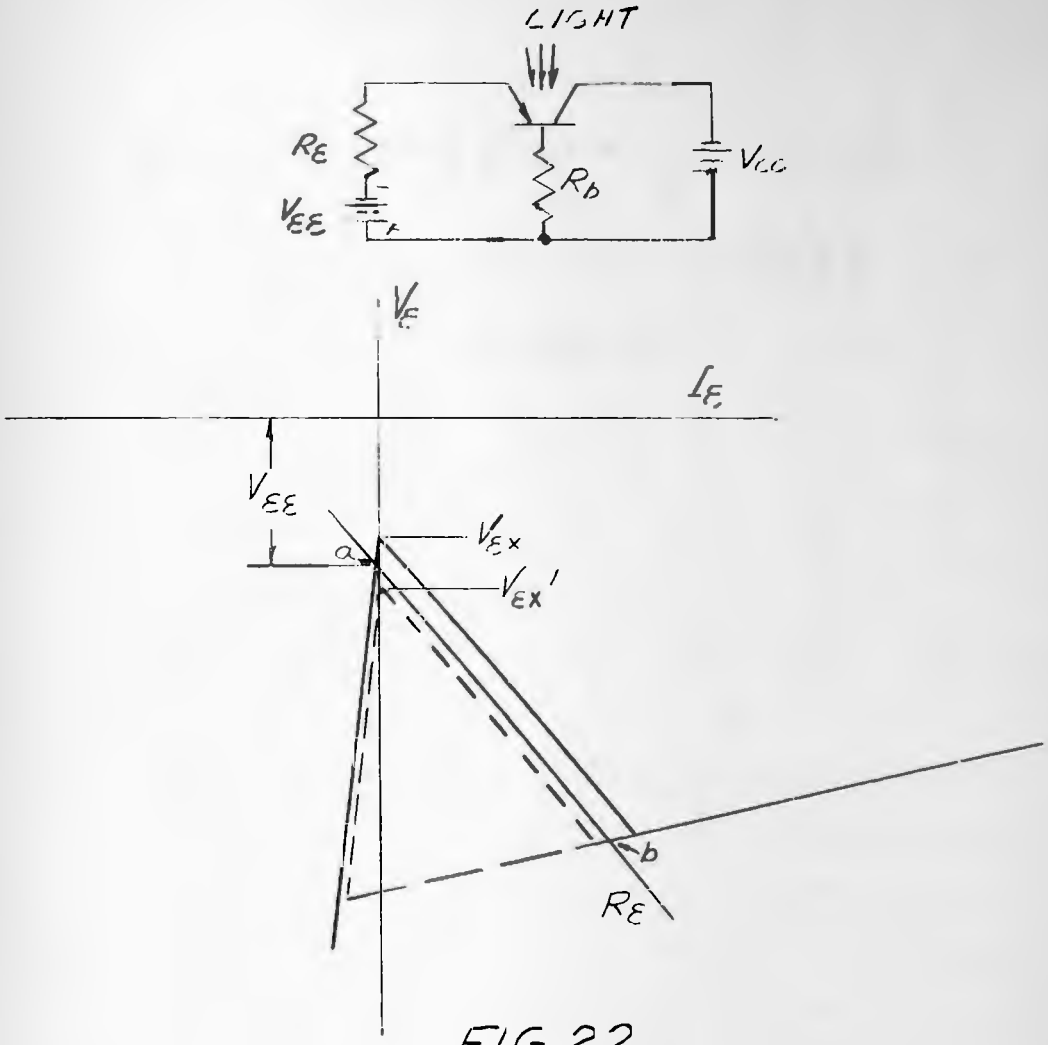


FIG. 22

PHOTO TRIGGER TRANSISTOR CIRCUIT

SOME SYSTEM APPLICATIONS OF TRANSISTOR AMPLIFIERS

R. S. CARUTHERS

Produced outside the work performed under this contract

SOME SYSTEM APPLICATIONS OF TRANSISTOR AMPLIFIERS

While use of vacuum tubes is small in the Bell System compared to use in radio and television business, concentrated use in limited space is only rivaled by the high concentrations of vacuum tubes found in present day computer equipment. More and more equipment has been miniaturized crowding ever increasing numbers of circuits into the same office space. Our circuits have demanded greatly expanded use of vacuum tube functions and the extensive use of miniature pentodes of the 6AK5 type and double triodes of the 396A type has increased the crowding. Now a standard telephone office 19" relay rack may contain more than 500 tubes with more than a 1000 watts of heat to dissipate. Back-to-back use of these bays vitually doubles the tube and heat concentration.

It is in this field of compacted vacuum tube use that the transistor has a particularly useful application. It is to be expected that more and more the transistor will help the heat problem. Perhaps off-setting this would be the ability to crowd even more transistors into a given space. An adding stimulus would also be given to their use in more varied circuit functions because of their small size and permanence.

Before discussion the ways in which transistors have been arranged in some of our experimental circuits, a description will be given of performance objectives in certain telephone vacuum tube circuits. Typical of the circuit using the

miniature pentode is the V3 amplifier shown in Fig. 1. This amplifier represents an advance in the art of small telephone repeaters of which the Bell System has been quite proud. It was introduced in 1948 largely in the toll plant on trunks running between cities 20-200 miles apart. This amplifier, including transformers, tube, other components and mounting, occupies a space little larger than the vacuum tube in the prior amplifier replaced by the V3. As a result some 600 amplifiers can be mounted in a standard telephone bay that formerly would hold less than 1/10 that number of the earlier V1 amplifiers. Ten amplifiers plug into octal sockets in a 1-3/4" mounting strip across the width of a 19" relay rack. Gain control and test points on the top surface eliminate the need for fields of space consuming test jacks normally used for maintenance purposes. A plug-in amplifier can be removed from its socket, inserted in a test panel on the bay and restored to operating conditions by a test cord between the vacated socket and the test panel.

The circuit schematic of the V3 amplifier is shown in Fig. 2. Transformers at input and output provide an amplifier with 50-55 db non-regenerative gain. Circuit application requires input and output impedances held rather closely to 600 ohms resistance over the speech transmission band from 200 cycles to about 3500 cycles. Resistive terminations on the transformer provide the 600 ohm input impedance, but on the output where none of the output power can be wasted a combination of transformer and cathode feedback is used to provide the required impedance termination.

In Fig. 3 typical gain-frequency, load, and harmonic distortion curves are plotted for the V3 amplifier. Aside from the low frequency end of the band, flatness with frequency is easily gotten within a few tenths of a db. The smaller the transformers are made using finer and finer wire the higher the transmission band extends in frequency. Flatness to 5 or 6 kc as shown was obtained without effort in that direction as the useful voice band required was not that high. On the low end d-c through the output transformer impaired the transformer and amplifier frequency characteristic. Such transmission would have been even more impaired without feedback including the output transformer in the mu path.

Gain of 36 to 38 db and maximum output of about 200 milliwatts (23 db above 1 milliwatt) are found necessary in the applications where this type of amplifier is most valuable. Feedback of 14 db or more db reduces second and third harmonic distortions to the levels shown. Because of the large numbers of these amplifiers in tandem in a speech channel of a telephone call, distortion must be at least as small as the magnitudes shown to prevent speech impairment.

No transistor type exists that will reliably provide the 50-55 db non-regenerative gain of the V3 amplifier. In the NPN transistor nearest meeting the gain requirements, maximum output power is about 10 db short of the 200 mw output desired. The point contact transistor gain of 17 or 18

db is of little interest in getting a 50 db gain amplifier. The three stages required would need to be transformer coupled to provide the full gain. The circuit complexity of such an amplifier would make it quite unattractive compared to the V3 vacuum tube amplifier. In output power a point contact type with radiating fins for cooling has come closest to meeting the V3 200 milliwatt output requirement.

Fig. 4 is a photograph of a two stage point contact transistor amplifier. The output stage transistor is provided with cooling fins which allow a maximum collector dissipation of about 400 mw, about four times that of the regular point contact 1729 type. Although most of the inactive components, the condensers, resistances, and interstage transformers have been reduced to sizes comparable with the transistors, the output transformer is about half the volume of the entire amplifier. This large size results from the requirement that in this circuit the transformer be capable of carrying the d-c collector current of the output stage. Means for avoiding this d-c current requirement to allow a reduction in transformer size will be discussed later.

Fig. 5 shows the circuit configuration used in the two stage point contact transistor amplifier. The circuit is essentially a straightforward application of grounded base techniques, the operating points of the transistors being controlled by constant emitter current supplied from +130

volt battery. Other circuit configurations were tried in which the self-bias for the emitter was derived by adding resistance in the base lead. The main advantage of such a circuit would be, of course, the need of only a single battery supply. It was found that the positive feedback caused by the base resistance reduced the stability of the operating point. Variation from unit to unit and with temperature and operating voltages were annoyingly large.

As the first stage collector current was low, this transistor was resistance shunt fed, thus keeping d-c from the interstage transformer and resulting in a large reduction in its size. The input and interstage transformer ratios were chosen as a compromise between the maximum gain of impedance matching and the minimum distortion and low gain of constant emitter current drive. The output transformer impedance was determined by the optimum load line for the transistor used.

All condensers are of the new miniature tantalum type. The 8V d-c maximum limit on the 25 mf emitter circuit condensers made it necessary to provide the safety resistances from emitter to ground. Thus, in the event a transistor were removed from its socket while bias voltage was supplied to the amplifier, only a part of the battery voltage would appear across the condenser instead of the full +130 volts.

The total battery power required for the transistor amplifier of Fig. 5 is about 1.33 watts, about 2/3 that required

for the V3 vacuum tube amplifier. A more marked reduction in power might have been obtained, however, if we had not restricted ourselves to the use of standard telephone office battery supplies.

The performance characteristics for this two stage point contact transistor amplifier are shown by the curves of Fig. 6. Here it can be seen that the maximum output power of this amplifier is short by 1 to 2 db of the power of the V3 amplifier. The harmonic distortion curves show a more marked difference, however. At +16 dbm the second harmonic distortion is only 25 db below the fundamental where at the same point on the V3, second harmonic distortion is 35 db below fundamental. The third harmonic distortion is degraded similarly. The over-all gain of the amplifier is about 30 db, 6 db short of the V3. A few transistors have been tested, however, that give an over-all gain of about 35 db in this amplifier. The frequency characteristic curve shows that, although the response at the high end of the voice band is satisfactory, the drop off at the low end due to the output transformer characteristic would not be acceptable in telephone speech channel use. The lack of adequate non-regenerative gain prevents us from applying feedback to this amplifier to improve both this frequency characteristic and the harmonic distortion products. The conclusion was that an additional transistor gain stage would be required to meet the desired performance characteristics.

The development of the NPN transistor offered new

possibilities for voice frequency amplifier design. Although these transistor units were somewhat limited in availability in the early stage of development, a fairly large number of NPN junction units were obtained as by product rejects in constructing high quality units. These units were rejected for low alpha characteristics, the alpha being about .8 as compared with the desired value of .95 or greater. Preliminary work with a two stage amplifier of these low alpha units showed some interesting results. A model of such an amplifier is shown in Fig. 7. This amplifier was constructed of miniature components and assembled using printed wiring techniques. Its size is shown compared to the size of the 408A vacuum tube of the V3 amplifier.

It should be understood that at the outset it was realized that the available NPN transistors were not capable of Class A power outputs equivalent to the V3 amplifier, these units being limited to a maximum dissipation of 50 mw. It was believed, however, that higher output performance might be extrapolated from the initial performance characteristics derived from the lower power amplifier. It is expected that such higher power units will be constructed soon, perhaps using cooling fins.

The circuit of the two stage NPN amplifier is shown in Fig. 8. Each stage is connected in grounded emitter configuration for audio frequency and biased in a grounded base

configuration for d-c. This type of biasing holds a constant emitter current in contrast to other systems which control bias by maintaining a constant base current. By this means a more constant collector operating point is held, an important feature in an amplifier operating at close to 50% efficiency. A more detailed discussion of biasing will be given in a later section.

The output stage collector in Fig. 8 is resistance shunt fed, permitting use of a sub-miniature transformer for transistor to line impedance matching. This was done, however, at the expense of a considerable reduction in overall battery efficiency and an output shunt power loss of about 1.1 db. This loss in battery efficiency is evidenced by the fact that, although the amplifier power output is over 10 db less than the V3, its battery power input is only 1/4 or about 6 db less. It should be pointed out, however, that the output transistor itself is operating at remarkably high efficiency. In this circuit it is biased at an operating point of 22.5 volts at 2.2 ma, or at 50 mw. The maximum sine-wave output is +11.1 dbm (db above 1 mw). Including the output transformer loss of 1.6 db and the shunt loss of 1.1 db, the transistor is delivering +13.8 dbm, only 0.2 db short of the theoretical maximum of 50% Class A efficiency, or +14 dbm.

The non-regenerative gain of the two stage NPN amplifier is about 55 db. This gain gives at least 15 db of negative feedback while still meeting regenerative gain requirements. Feedback was added in the form of local series feedback by use of unby-passed emitter resistance and in the form of overall shunt feedback by the coupling of the output collector to the input base.

A detailed study of the $\mu\beta$ characteristics of this amplifier has not yet been made but it is expected that such an analysis would be made in the same manner as with multiple stage vacuum tube circuits. The general performance characteristics of the amplifier are shown on Fig. 9. Here it is seen that the maximum square wave output is +13 dbm. Although this is 10 db less than that of the V3, it should be noted that the second harmonic distortion is down 35 db at +10 dbm, only 3 db back from maximum output. This distortion figure compares with a similar point on the V3 performance characteristic at +16 dbm, or 7 db back from maximum output. It should be understood that the output performance of this NPN amplifier holds only when the operating point is maintained. Any large deviation from this operating point by a variation in the bias would appreciably reduce the maximum output.

The 40 db of gain in the two stage NPN amplifier is constant within a few tenths of a db over the range of

temperature encountered in the laboratory. No extensive temperature stability tests have yet been made, however. The frequency response is flat over the voice frequency band, extending well beyond the high end of the audio range.

Some investigation of a single stage amplifier using an NPN junction with an alpha greater than .95 was made using the circuit given in Fig. 10. The transistor was operated from a single battery source. The means for biasing for single battery operation will be discussed in a later section. Measurements were made with collector current fed through the output transformer. The transistor was biased at an operating point of 25 volts at 2 ma, or 50 mw dissipation. The remaining 150 mw of battery power is used in the emitter bias resistance and in the voltage divider for base voltage. Some difficulty was experienced in determining the optimum ratio for the input transformer. Large changes of input impedance were experienced with changes in temperature, the impedance being in the order of a few thousand ohms for the unit at room temperature and increasing to about 80,000 ohms with increase of temperature due to internal dissipation. A transformer ratio of 600 to 20,000 ohms was chosen and all measurements were made on the amplifier after a warm-up time of about 20 minutes.

The results of performance tests on this single stage NPN circuit are shown in Fig. 11. As with the previous

amplifier the output power was limited by the maximum allowable dissipation on the unit of 50 mw. The non-regenerative gain was about 38 db but showed a tendency to drift up or down about 2 to 3 db from this value. This drift was probably due to slow changes in input impedance. The harmonic distortion was observed to be quite high, indicating that feedback would be needed to meet our requirements. The poor frequency response at the low end is due primarily to the fact that the 4 mf of by-pass condenser on the emitter bias resistance was not sufficient. Increasing the value of this condenser would improve the low end response to a characteristic limited by the output transformer. Feedback would be still required to overcome this limitation.

Although tests on the single stage NPN amplifier were made on only one transistor, others not being available at the time, it is felt that these observations show the improbability of making a single stage amplifier with all of the desired characteristics of the V3 amplifier even with high α units. It is fundamental that the higher the gain due to the increasing value of $\frac{1}{1-\alpha}$, the higher will be the large signal harmonic distortion due to small α curvature. It appears, therefore, that with sufficient feedback to reduce distortion a maximum regenerative large signal gain would be about 10 db short of requirements. As with the lower α units a two stage amplifier could be used. A

possible advantage of the higher α units in such an amplifier would be direct coupling rather than transformer coupling of the interstage.

The use of the grounded base connection was investigated but dropped in favor of the higher gain grounded emitter configuration. For high alpha units where the current gain in a grounded base circuit is essentially unity, the power gain must be obtained entirely from the output to input impedance ratio. For maximum power output for a reasonable operating point of 25 volts at 2 ma the optimum output impedance is about 12,000 ohms. For acceptable distortion a minimum generator input impedance is not less than 600 ohms. This represents a maximum power gain of 13 db. A regenerated gain of at least 20 db can be expected from a grounded emitter stage, even with low α units.

It can be appreciated that optimum large signal performance with a fixed load can be attained only by the maintaining of constant operating point. Where the operating voltage is held constant by transformer or choke feed a large increase in collector current could dangerously increase the dissipation of the unit; a decrease would reduce the available power output. For resistance coupled circuits, variation in collector current degrades the performance even greater as the current change is accompanied with an operating voltage change.

In the preliminary work with grounded emitter NPN junction circuits for small signal, low power application, three methods of biasing were proposed[1] as shown in Fig. 12. These methods were based on maintaining a constant base current of some value, the amount and direction depending on the operating region. For our work at maximum dissipation, these systems were not satisfactory, allowing large changes in collector current from unit to unit. Similar changes occurred in individual transistor units with operating temperature variation. Such changes from unit to unit appear to be due to the variation of α and I_{co}. The collector current change due to temperature in individual units is caused by the variation of I_{co} with temperature.

The reasons for changes due to I_{co} variation in grounded emitter NPN constant base current operation may best be seen qualitatively from the NPN static characteristics shown in Fig. 12. Although these characteristics are not strictly correct, the r_c slope being exaggerated for reasons of clarity, an exact examination of a more accurate set of curves would show the same results. An α of .9 and r_c equal to 200K have been assumed. The solid line characteristics are curves of constant emitter current plotted

- -

(1) Wallace and Pietenpol - BSTJ - July, 1951

for collector voltage vs. collector current and their slopes represent the r_c of the transistor. The slopes in the operating region are constant for all values of emitter current, an ideal condition for an essentially constant value of α. The collector current for any particular value of collector voltage for zero emitter current is called the I_{co}. It is this current which varies due to temperature changes in the transistor unit. To see how the I_{co} change would effect the operating point for constant base current bias a new set of characteristics is drawn in Fig. 12, a set of constant base current lines. These lines are derived using the fundamental criterion that the algebraic sum of base, emitter, and collector currents must equal zero. The first is a simple one, a curve through those points at which the emitter current and collector current are equal; this is shown by the dotted line $I_b = 0$. Another is shown as a curve through points in which the emitter current is 100 μa less than the collector current, represented by $I_b = +100$ μa. The other I_b curves can be plotted in the same manner.

If in constant base current operation a desired operating point of 20 volts at 1 ma is assumed it is readily seen from Fig. 12 that a base current $I_b = 0$ is required. At this voltage the I_{co} is equal to 100 μa. Assume now a temperature increase causes a change in I_{co} from 100 μa to 200 μa. The entire family of I_ε curves moves 100 μa to

the right, the I_e = 2 ma curve passing through I_c = 2 ma instead of I_c = 1.9 ma at 20 volts. A new family of constant I_b curves would result, the I_b = 0 curve passing through I_c = 2 ma instead of I_c = 1 ma. The result is, of course, that the operating point has changed from 20 volts at 1 ma to 20 volts at 2 ma. This assumes constant voltage was maintained by transformer or choke feed. Had resistance shunt feed been used, a corresponding reduction of operating voltage would have resulted with a shift along the d-c load line.

The formulae associated with each of the biasing figures on Fig. 12 show quantitatively the I_c variation with I_{co}. In the first case where I_b = 0, I_c varies directly with I_{co}. In the second and third cases, where I_b is likely to be about the same or less than I_{co}, the variation in I_{co} still has a large effect on I_c. It should be understood that changes in I_{co} of as much as 10 to 1 may be met in the operational temperature range.

In Fig. 13 biasing methods are shown in which close control of collector current is maintained for wide changes in I_{co} with temperature and from unit to unit. In the first arrangement in Fig. 13 the system for biasing for NPN grounded base operation is shown[1], this method being similar to the biasing of point contact transistors. If the voltage between

- -

(1) Wallace and Pietenpol - BSTJ - July, 1951

emitter and base is small compared with the emitter bias battery a constant emitter current through the emitter resistance R is maintained. For collector operation currents in the order of 1 to 2 ma the term αI_ε as shown in the formula is very much larger than I_{co} (10 to 100 μa). Therefore a large change in I_{co} will have very little effect on the collector current, and thus a constant operating point can be maintained. A method of biasing for grounded-emitter operation is shown in Fig. 13. In effect the NPN transistor is connected ground base for d-c, and grounded emitter for a-c. The same control of a constant operating point is maintained. From this figure it can be seen that the emitter current tends to be self-regulating. If I_ε tends to increase over the value of $\frac{E_1}{R}$ the emitter voltage becomes more positive with respect to the base, tending to reduce the current flow into the base. This change in turn tends to reduce to emitter current by the relationship $\Delta I_\varepsilon = \frac{1}{1-\alpha} \Delta I_b$. Thus, whether a negative, positive or zero base current is required, the emitter current is held essentially constant.

The third circuit in Fig. 13 shows a method for biasing in which a single battery is used. If the base current is small compared with the current in the divider R_2 and R_3, the base voltage is held constant by the divider at a value

$E_{base} = E \frac{R_2}{R_2 + R_3}$. This voltage tends to drive a base current and hence an emitter current until the base and emitter voltage are almost equal. At this point the emitter current is determined by $I_e = \frac{E_{base}}{R_1} = E \frac{R_2}{R_1(R_2 + R_3)}$. As with the other circuits of Fig. 13 the emitter current is held constant, the result being that the collector operating point is only slightly dependent on the temperature or unit to unit variation of I_{co}.

Fig. 14 shows in pictorial form the comparison of constant I_b and constant I_e methods of biasing. Here collector current is shown as a function of I_{co}. A constant for α of .97 was used. For higher values of α the difference is more marked in the two methods of biasing.

In a previous discussion of the two stage point contact transistor amplifier, it was observed that the output transformer carrying the d-c collector current was not commensurate in size with the other amplifier components. Although these other components have been reduced in size to as much as 1/10 their equivalents in the V3 repeater the size of the output transformer would limit the overall reduction in size of a transistor amplifier to only one-half the V3. With present core materials for voice frequency transformers, further reductions in size can be attained only by a sacrifice of performance with a degradation of low frequency response and an increase in mid-band loss or by the elimination of the d-c component by circuit means. As the performance requirements for these transformers have already been reduced to a minimum,

the only alternative for Class A operation is to investigate other means for collector current feed. In Fig. 15 two such alternatives are shown. The first is the standard resistance shunt feed commonly used in resistance coupled amplifier circuits. Though quite simple from the equipment standpoint, this method greatly reduces the overall operating efficiency. On Fig. 15 a curve is shown for the theoretical maximum amplifier efficiency versus the shunt loss to audio output of the collector feed resistance, this loss being a function of the ratio of output load impedance to the shunt feed resistance. The maximum efficiency is about 8.5% with a shunt loss for the a-c output of 2.3 db. This point corresponds to a load to shunt-feed resistance ratio of $\sqrt{2}/2$. Where a-c power output is limited by maximum transistor dissipation, it is important to keep this shunt power loss to a minimum. This loss may be reduced at the expense of a further loss in efficiency. In an effort to reduce the size of an amplifier commensurate with the transistor, it is apparent that the loss in efficiency has increased power input and hence greatly increased dissipation per unit volume.

A second method of collector current feed is shown on the lower left of Fig. 15. In this arrangement an additional NPN transistor is connected from the battery supply to the collector of the amplifying transistor. This additional transistor is biased in such a way that its emitter to collector

d-c voltage is equal to half the supply battery, the remaining half of the battery voltage appearing across the active transistor. Because of the condenser coupling from emitter to base the shunt feed transistor is held at constant current for a-c. In effect, this transistor presents a low impedance for the d-c supply and very high impedance shunt to the audio frequency load. With practically no loss to the output power but half the battery power being dissipated in the extra transistor, a maximum theoretical efficiency of 25% is possible.

The transformer coupled circuit is shown on Fig. 15 for comparison. For an NPN transistor biased at the proper operating point, a maximum Class A efficiency of almost 50% is possible. An alternative to the transformer feed would, of course, be choke feed. No improvement in size would be attained, however, as the added volume of a non-saturable choke would be equal to or greater than the volume reduction of the transformer.

A possible means for avoiding the conflict between power dissipation and size would be some form of push-pull operation. By use of the duality principle, Mr. Raisbeck has shown that optimum push-pull Class B performance on point contact transistors is attained by constant collector current operation. As constant current power supplies are not available in the telephone office, this method of operation does not seem practical for our use. The characteristics of NPN

junction transistors, however, appear to be highly suitable to constant voltage operation. In addition, the matching of pairs of NPN transistors by production control or selection and the expected inherent stability of their characteristics with little aging may avoid the usual unbalance problems met in push-pull vacuum tube circuits.

Two possible circuit arrangements of push-pull NPN transistors are shown in Fig. 16. In the first, each transistor is biased by constant emitter current control. Although a two battery arrangement is shown, the single battery arrangement previously described would be satisfactory. By proper adjustment of bias conditions and load line for Class A operation, a maximum power output of almost 50 mw or half the total allowable dissipation of the two transistors could be attained. If the two transistor collector currents are well balanced, a center tapped transformer of the sub-miniature size could be used. In telephone speech transmission, where maximum power is required only at syllabic rates, an adjustment of the load and bias for Class AB operation might improve the overload characteristics of such an amplifier. Under this condition the storage of the emitter by-pass condensers must be sufficient to maintain the quiescent operating point as rectification takes place when the maximum Class A signal is exceeded.

The second circuit in Fig. 16 is a possible arrangement for Class B operation. The transistors are biased at almost zero current with full battery voltage. The desirability of such an arrangement is, of course, that practically no power would be required during the idle condition and that maximum office battery load would occur only at speech syllabic rates.

For the circuits described above, an additional driver stage, not shown in the figures, would be needed to give the required non-regenerative gain. The reduction in size and power drain of such a transistor amplifier would make this type of circuit attractive, although the amplifier would require a total of three transistors to replace the single tube of the V3.

Fig. 17 is a picture of the type of components used in voice frequency amplifiers and their comparison in size with transistors. From left to right components are as follows:

1. 437C condenser 2-2 μf, 200V (used in voice frequency telephone repeater).
2. Miniature output transformer, 10 ma d-c.
3. Hearing-aid output transformer, 2 ma d-c.
4. 1 μf, 150V paper condenser (V3 repeater).
5. 408A vacuum tube and socket (V3 repeater).
6. 1 μf, 150V tantalum condenser.

7. 1763 transistor and socket; point contact, power type.
8. Sub-miniature output or interstage transformer, 50 μa d-c.
9. 4 μf, 60V tantalum condenser.
10. Sub-miniature interstage transformer, 0 d-c.
11. 1729 transistor, point contact type.
12. 1/2 watt Allen Bradley resistance.
13. Miniature turret type terminal.
14. NPN junction transistor.
15. 1/4 watt carbon resistance.

One contrast becomes obvious when examining this photograph. This is the comparison in size of the miniature turret terminal, commonly used in equipment design for the mounting of pigtail apparatus components, and the NPN transistor. From this it is apparent that, in the design of new miniature equipment using transistors, new production methods will be necessary probably including printed wiring and circuit packaging techniques.

Circuits for a single tuned and a double tuned 22 mc IF amplifier using point contact transistors are shown in Figs. 18 and 19, respectively. The performances are given in the associated curves. In each case the external collector inductances are resonated with the internal and stray collector to base capacity. Over-coupling in the double-tuned circuit produces the double humped characteristic.

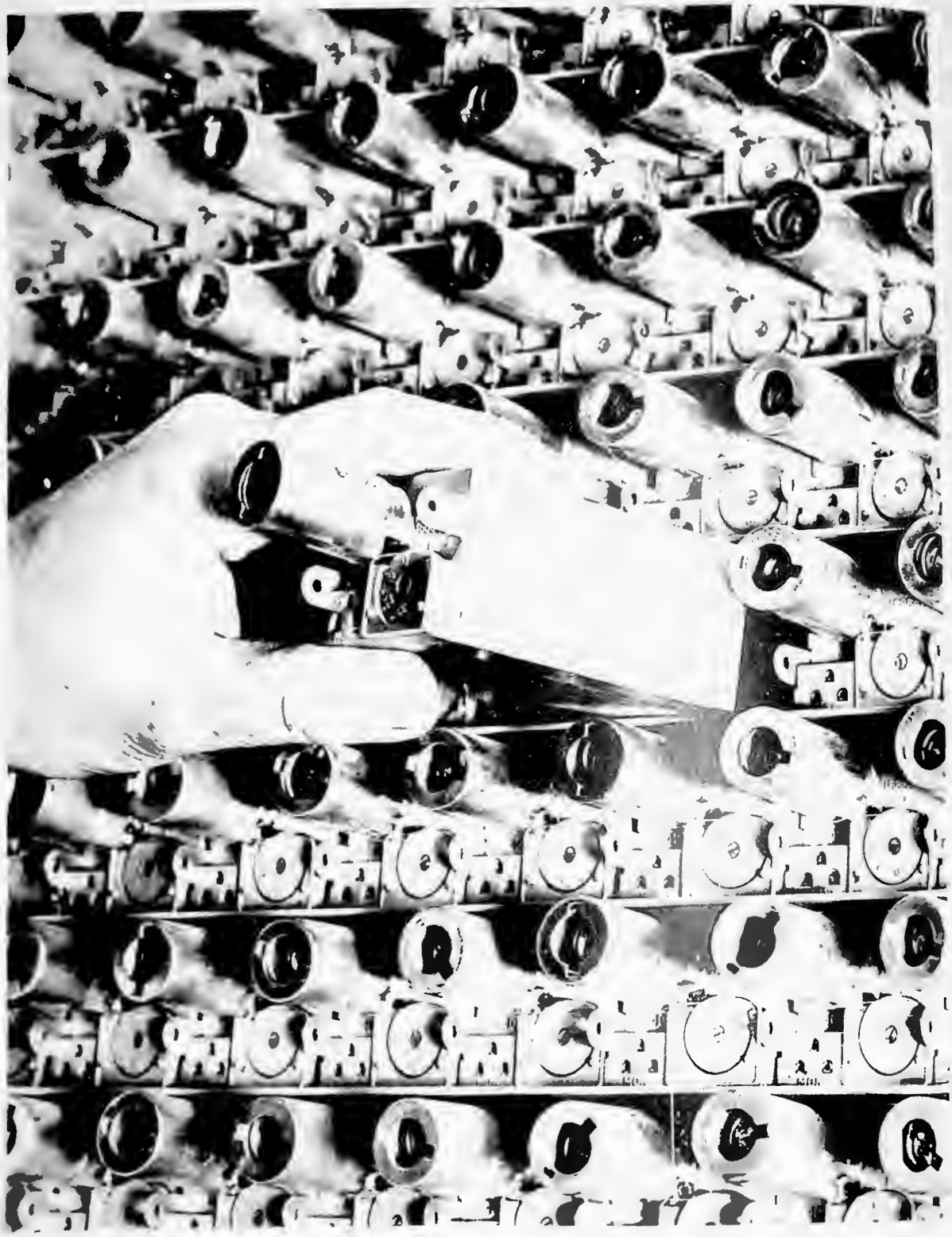


FIG. 1 - V3 REPEATER FOR VOICE FREQUENCY TELEPHONE USE

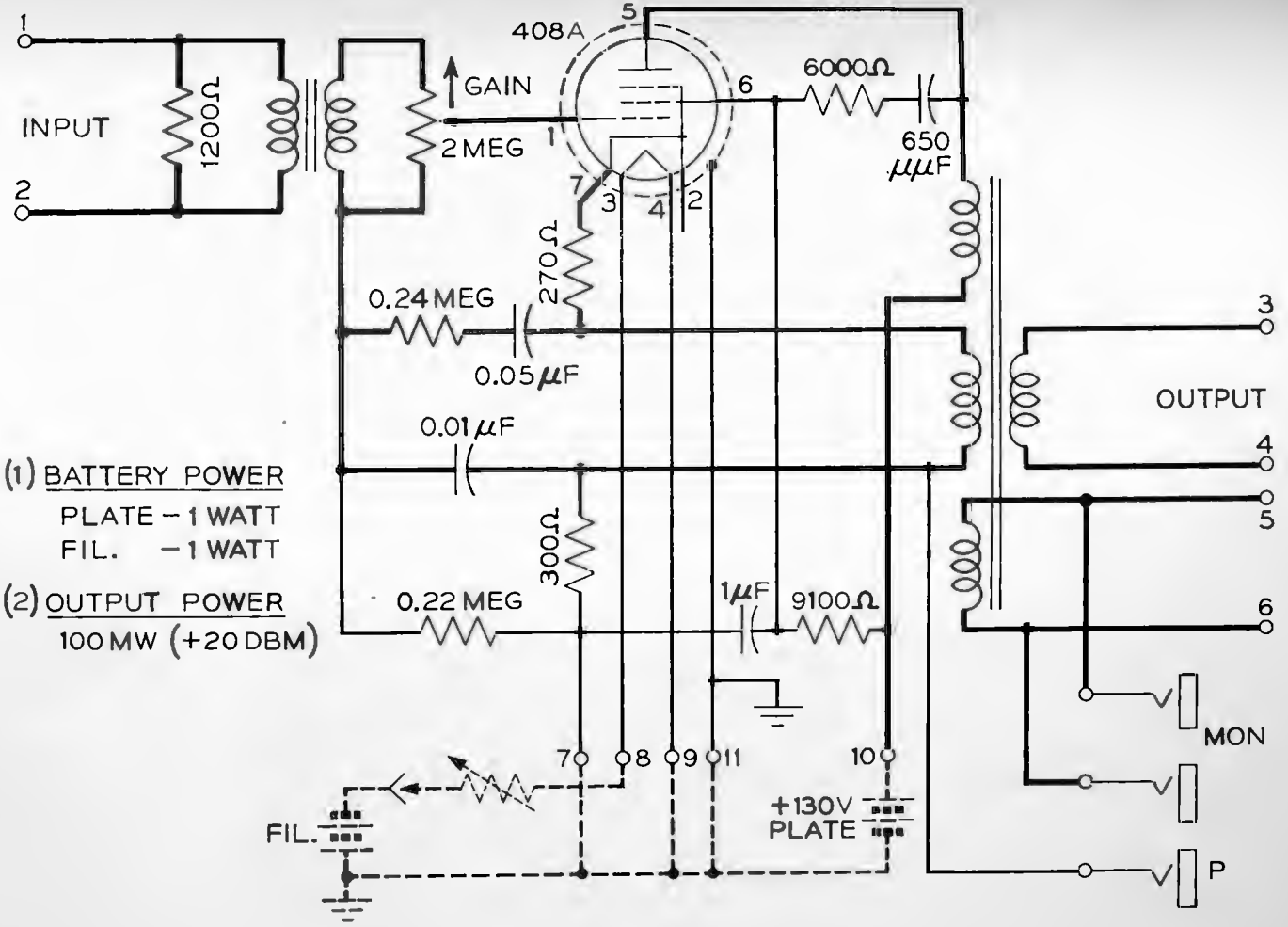
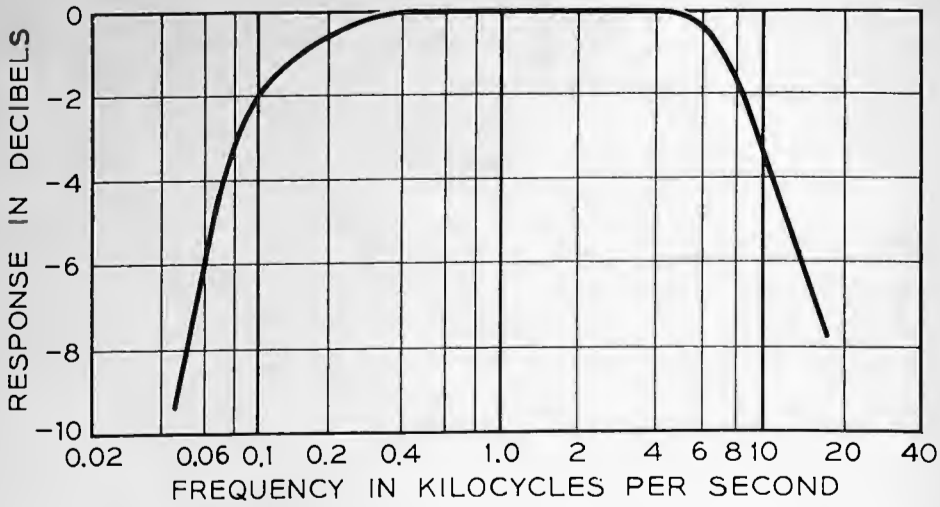


FIG.2 SCHEMATIC OF V3 REPEATER



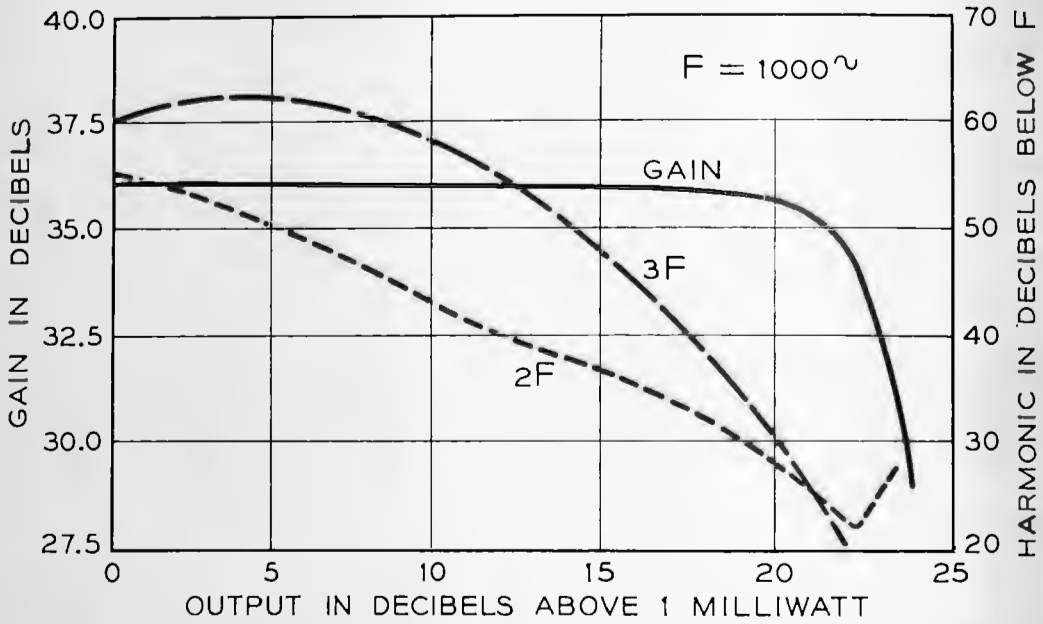


FIG. 3 PERFORMANCE CHARACTERISTICS OF V3 REPEATER

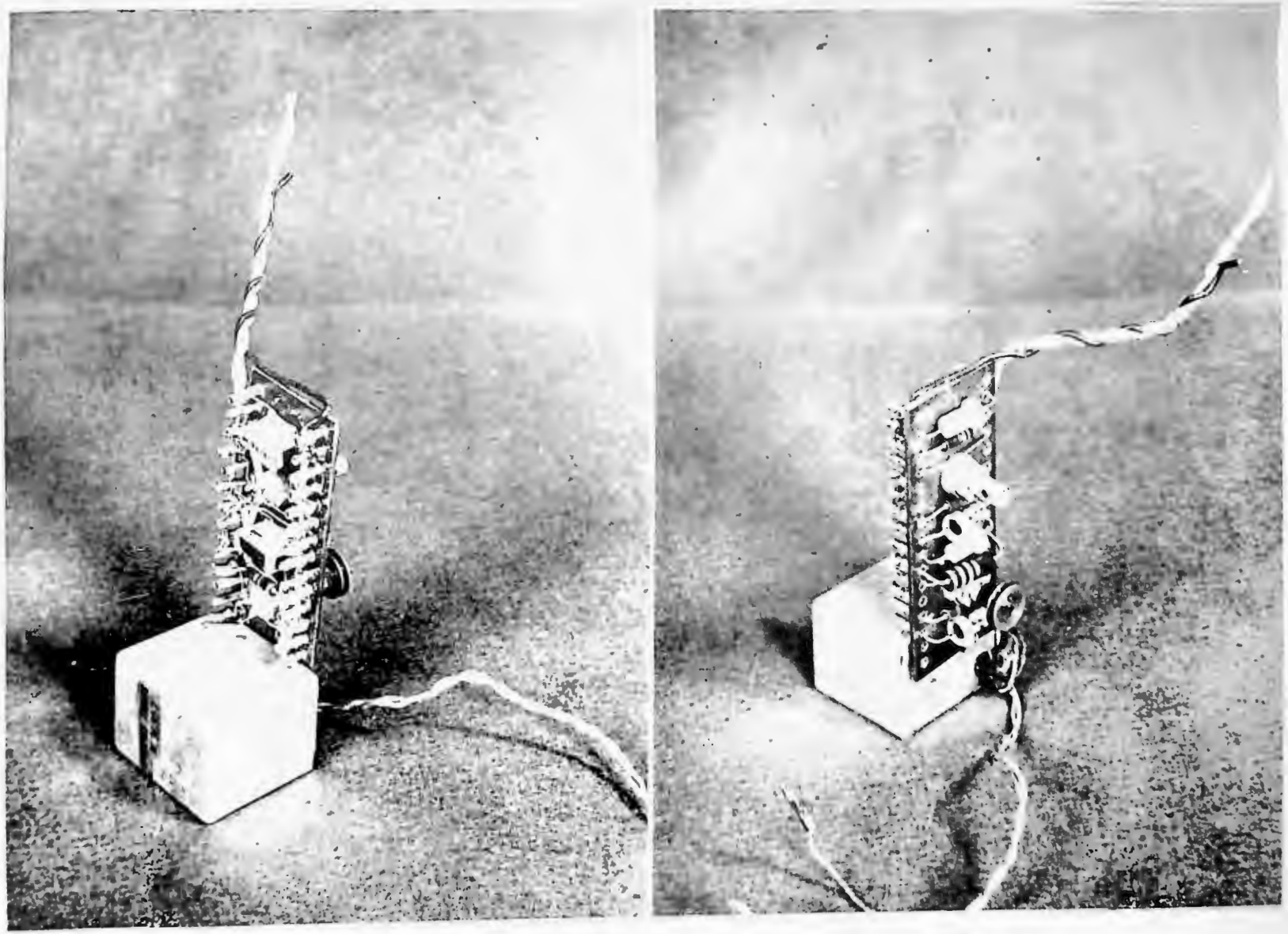
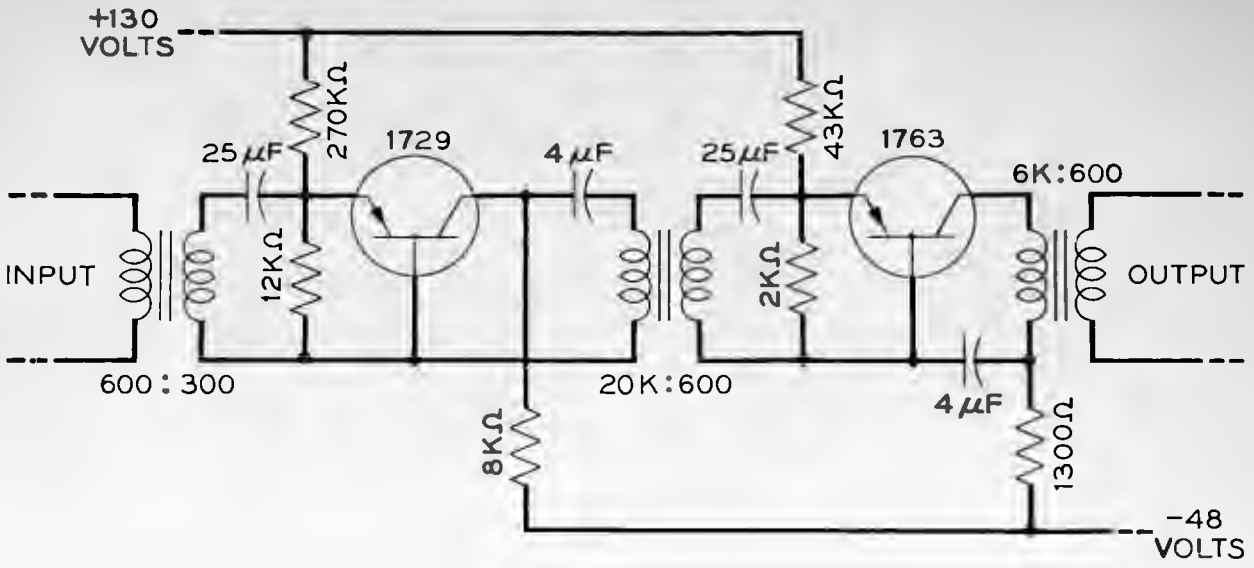


FIG. 4 - TWO STAGE POINT CONTACT TRANSISTOR AMPLIFIER



BATTERY POWER

1ST STAGE	0.23 WATTS
2ND STAGE	1.10 WATTS
TOTAL	1.33 WATTS

OUTPUT

63 MILLIWATTS (+18 DBM)

FIG. 5 SCHEMATIC OF TWO STAGE POINT CONTACT TRANSISTOR AMPLIFIER

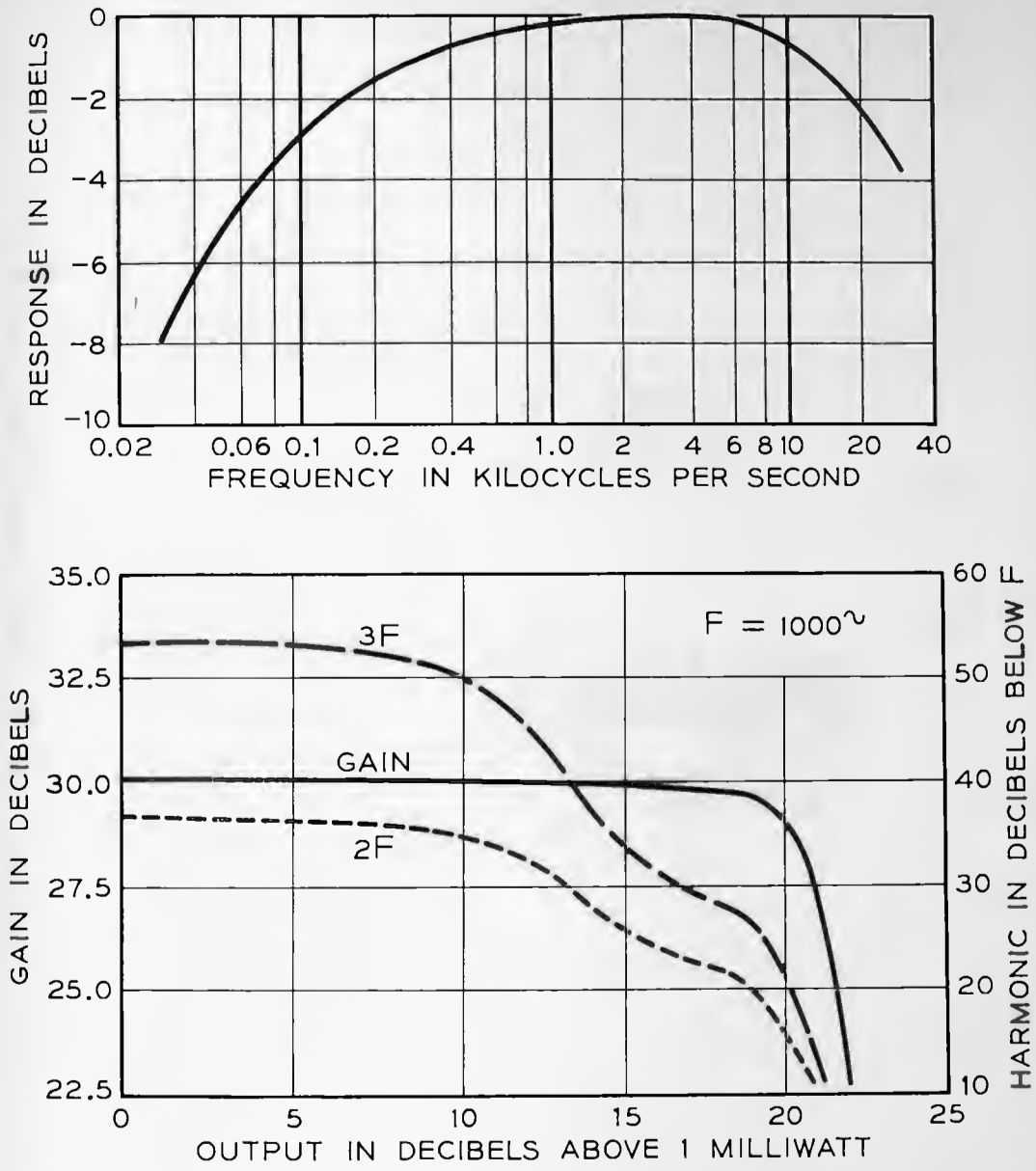


FIG. 6 PERFORMANCE CHARACTERISTICS OF TWO STAGE POINT CONTACT TRANSISTOR AMPLIFIER.

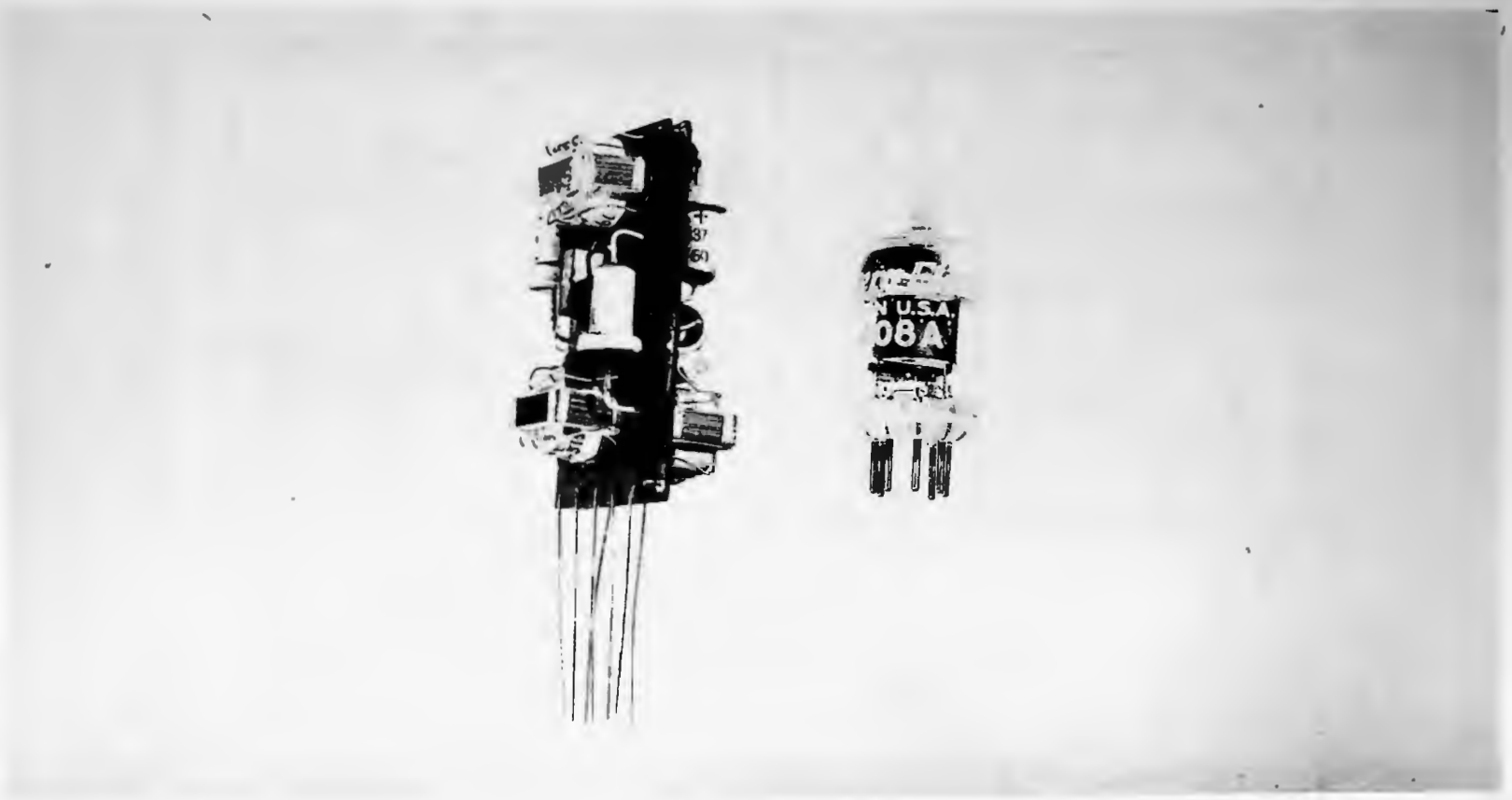
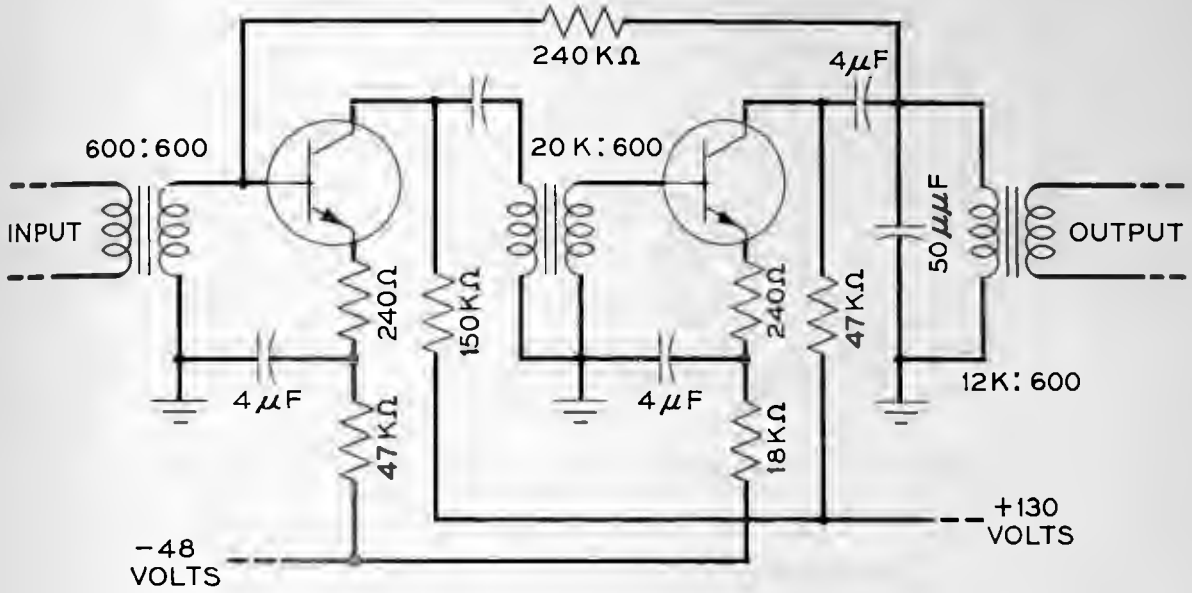


FIG. 7 - TWO STAGE NPN TRANSISTOR AMPLIFIER



BATTERY POWER

1ST STAGE	0.14 WATTS
2ND STAGE	0.33 WATTS
TOTAL	0.47 WATTS

OUTPUT

12.5 MILLIWATTS (+11 DBM)

FIG. 8 SCHEMATIC OF TWO STAGE NPN TRANSISTOR AMPLIFIER

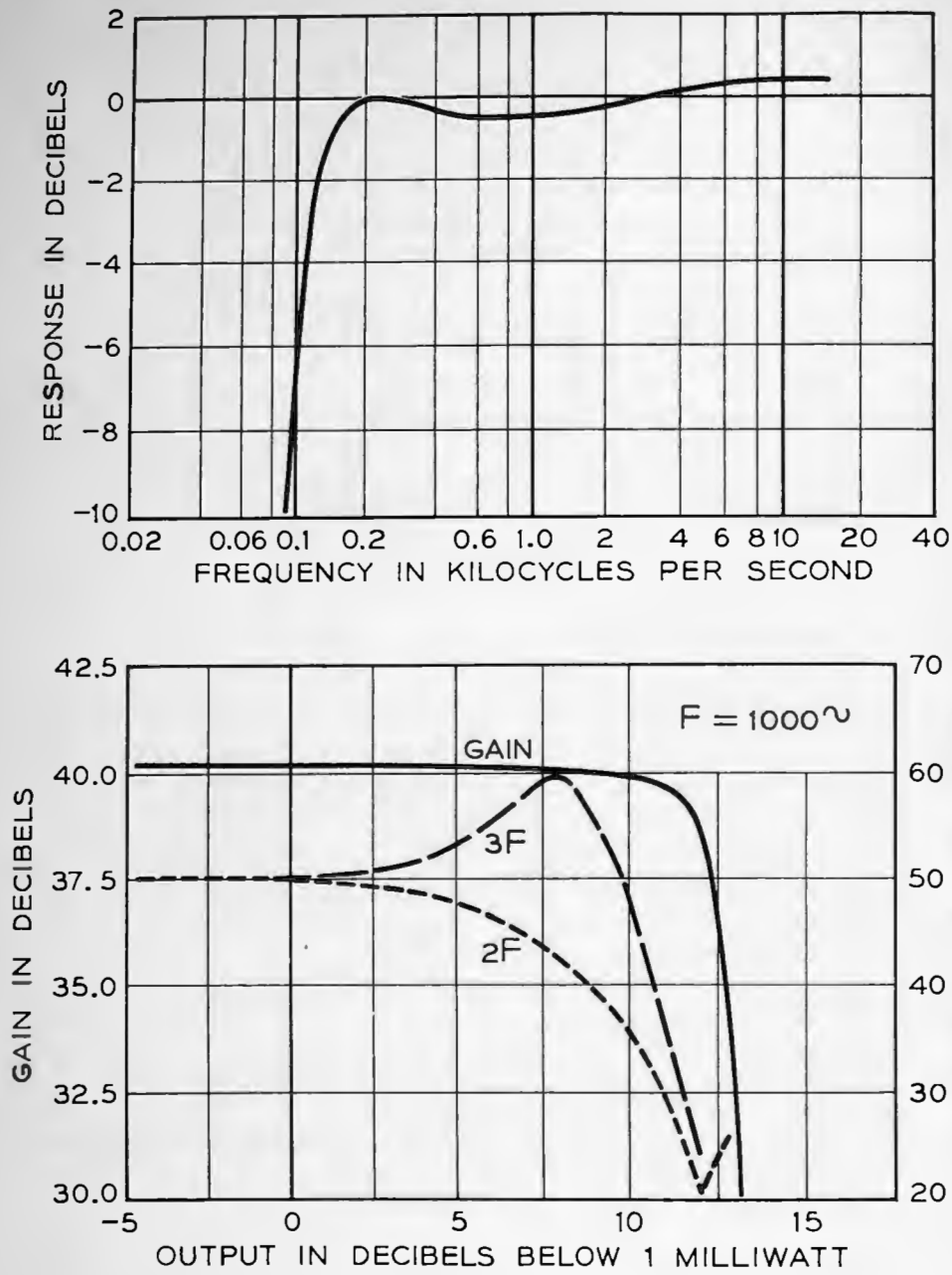
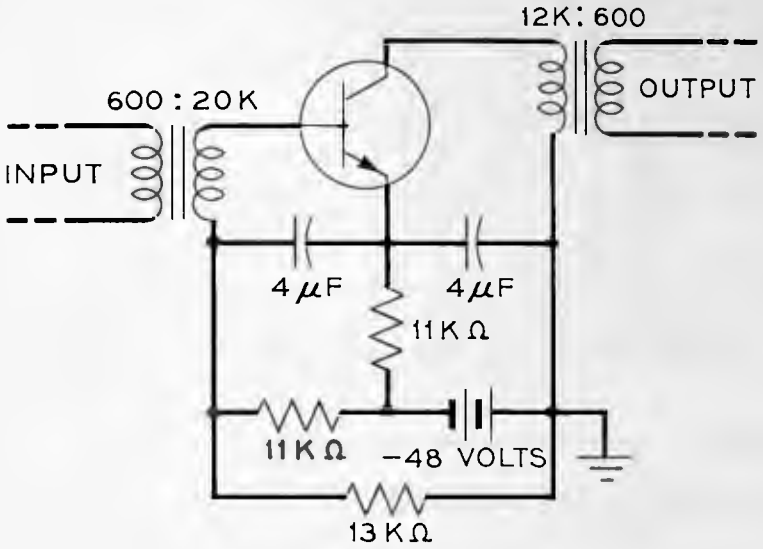


FIG. 9 PERFORMANCE CHARACTERISTICS OF TWO STAGE NPN TRANSISTOR AMPLIFIER.



$\alpha > 0.95$

$E_C = 24$ VOLTS, $I_C = 2$ MILLIAMPERES

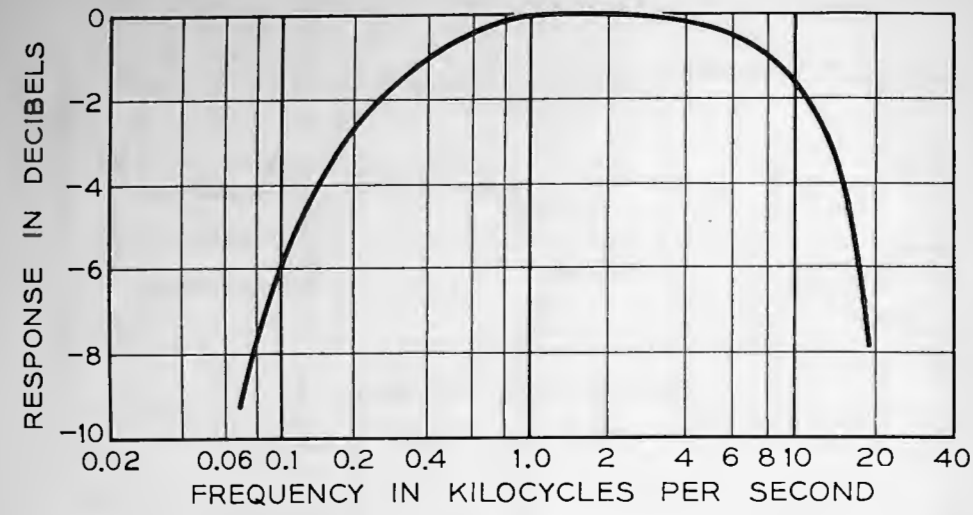
(1) BATTERY POWER

48 VOLTS – 0.2 WATT

(2) OUTPUT POWER

10 MILLIWATTS (+10 DBM)

FIG. 10 SCHEMATIC OF SINGLE STAGE NPN TRANSISTOR AMPLIFIER



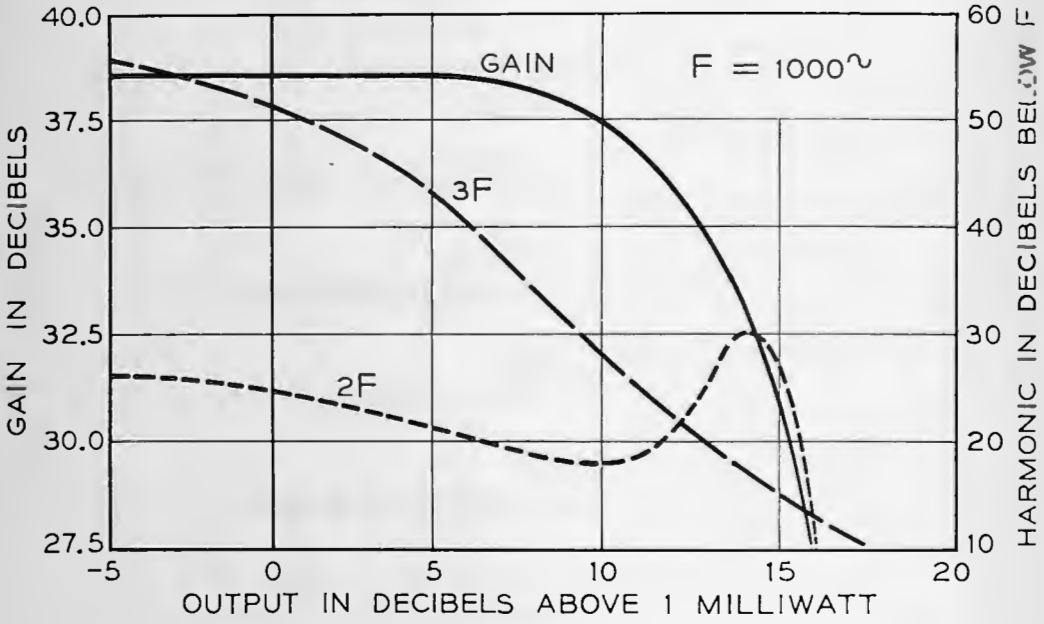
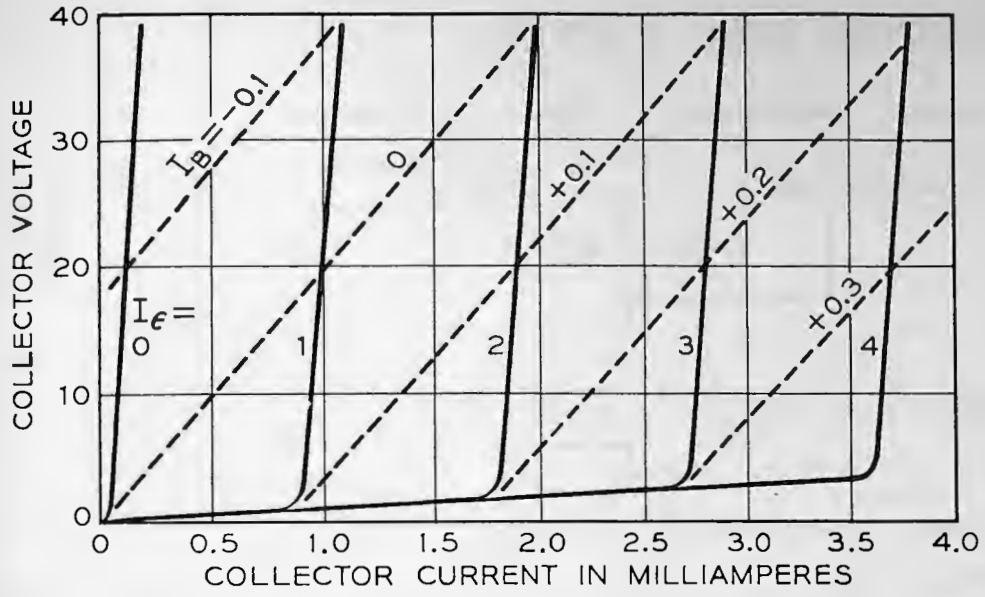
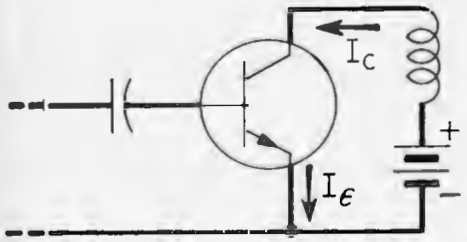
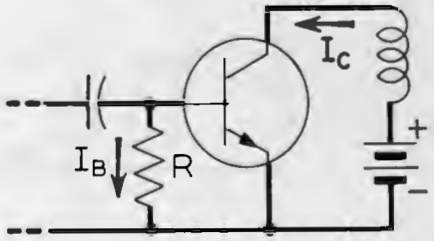


FIG. 11 PERFORMANCE CHARACTERISTIC OF SINGLE STAGE NPN TRANSISTOR AMPLIFIER.

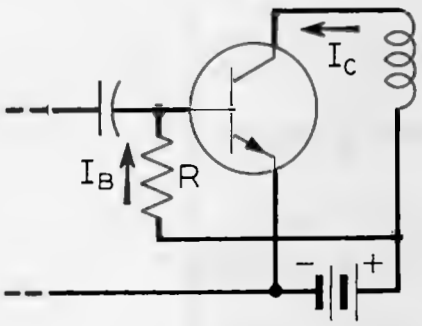




$$I_C = I_\epsilon = \frac{I_{CO}}{1-\alpha}$$



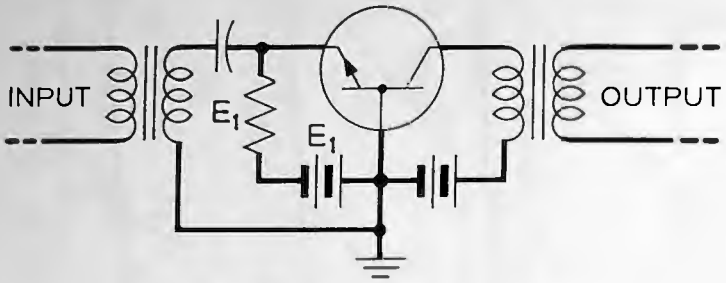
$$I_C = \frac{I_{CO}-\alpha I_B}{1-\alpha}$$



$$I_C = \frac{I_{CO}+\alpha I_B}{1-\alpha}$$

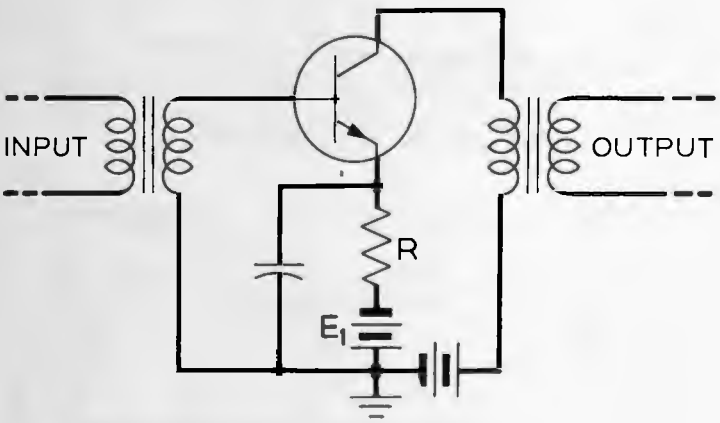
FIG. 12 OPERATION OF GROUNDED EMITTER NPN AMPLIFIER WITH CONSTANT BASE CURRENT FOR BIAS

GROUNDED BASE - 2 BATTERY



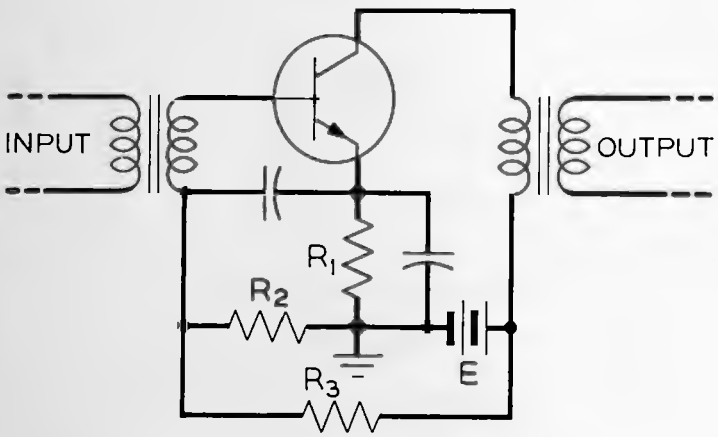
$$I_C = I_{CO} + \alpha I_\epsilon \qquad \left(I_\epsilon = \frac{E_1}{R}\right)$$

GROUNDED EMITTER - 2 BATTERY



$$I_C = I_{CO} + \alpha I_\epsilon \qquad \left(I_\epsilon = \frac{E_1}{R}\right)$$

GROUNDED EMITTER - SINGLE BATTERY



$$I_C = I_{CO} + \alpha I_\epsilon \qquad \left(I_\epsilon = \frac{E\dfrac{R_2}{R_2+R_3}}{R_1}\right)$$

FIG. 13 OPERATION OF NPN AMPLIFIER WITH CONSTANT EMITTER CURRENT FOR BIAS

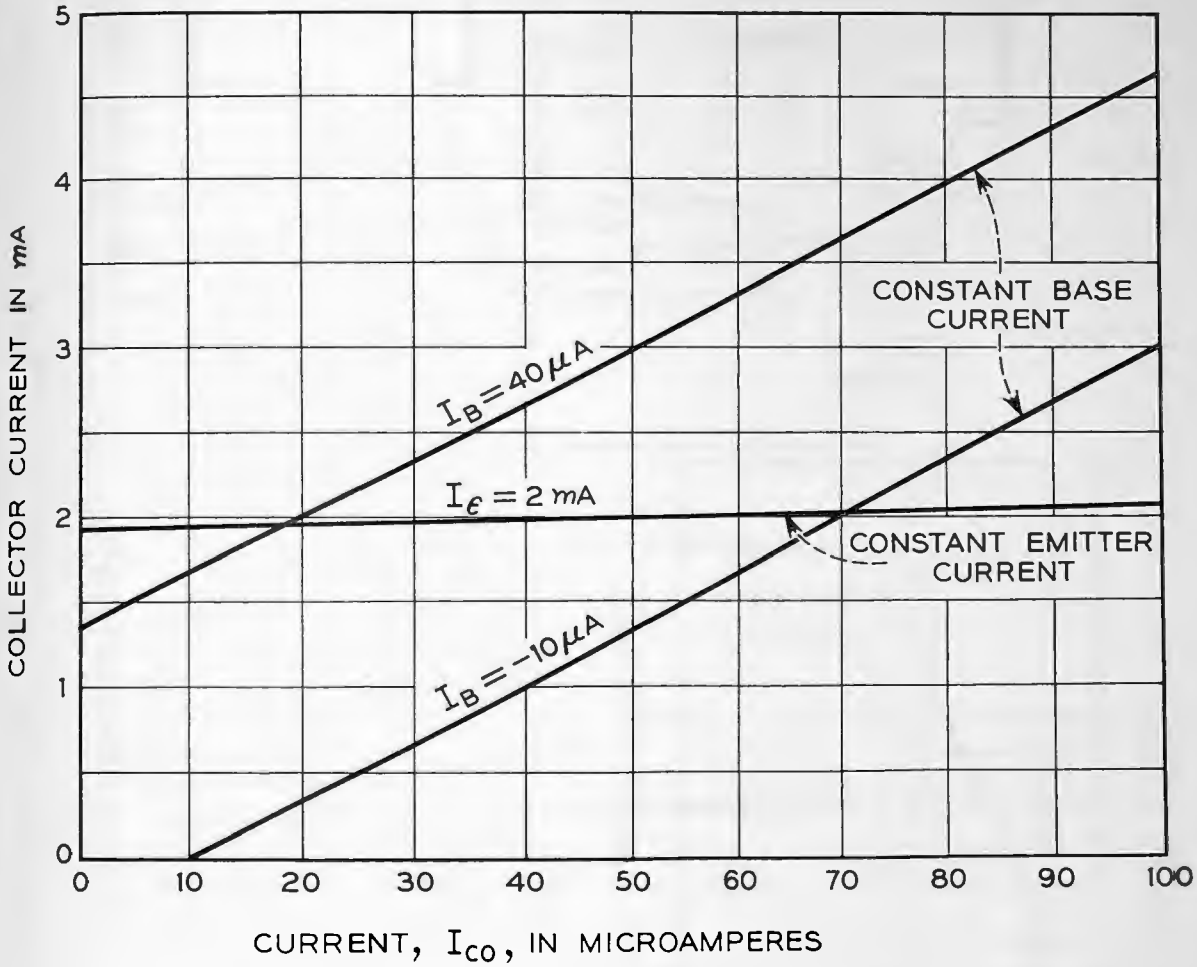
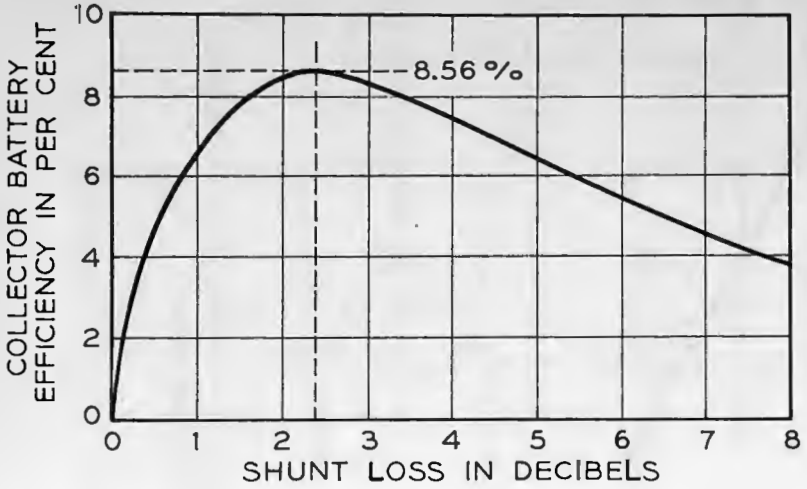
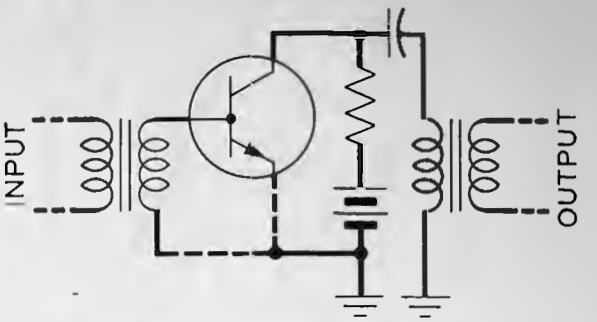


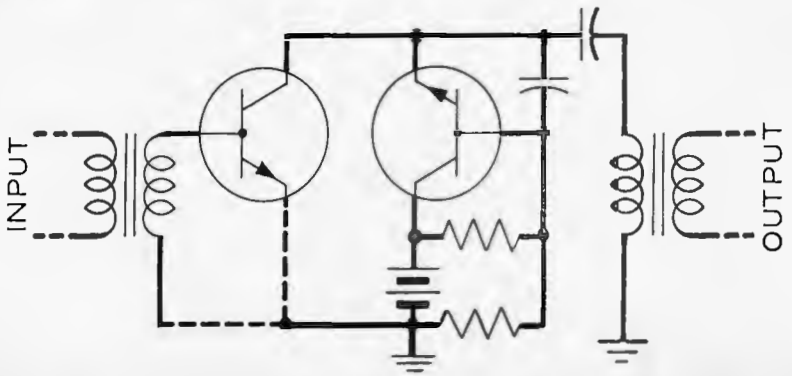
FIG. 14 COMPARISON OF CONSTANT BASE AND CONSTANT EMITTER CURRENT BIASING.

(1) RESISTANCE SHUNT FEED



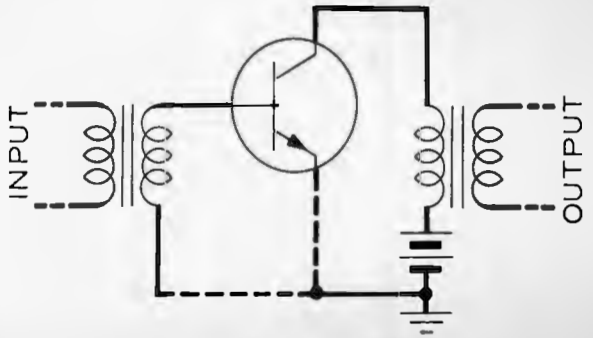


(2) TRANSISTOR SHUNT FEED



THEORETICAL MAXIMUM EFFICIENCY, 25%

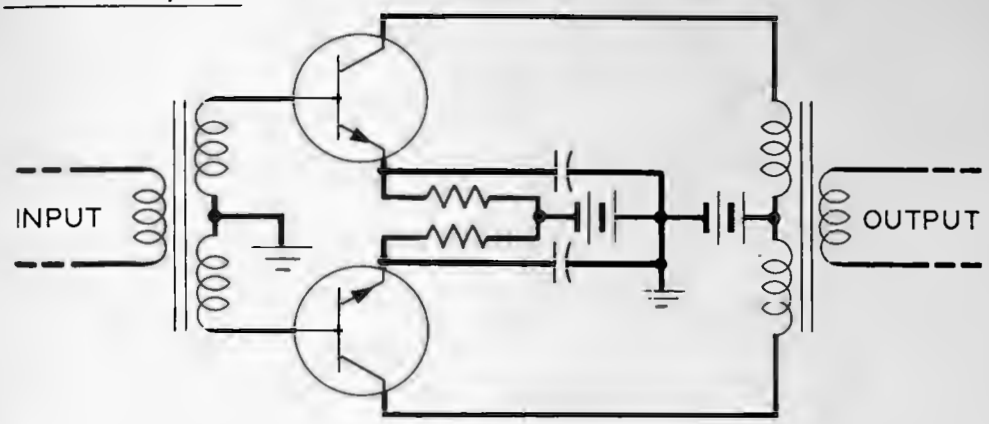
(3) TRANSFORMER FEED



THEORETICAL MAXIMUM EFFICIENCY, 50%

FIG. 15 METHODS FOR SUPPLYING DC CURRENT TO COLLECTOR

CLASS A, AB



CLASS B (CONSTANT VOLTAGE OPERATION)

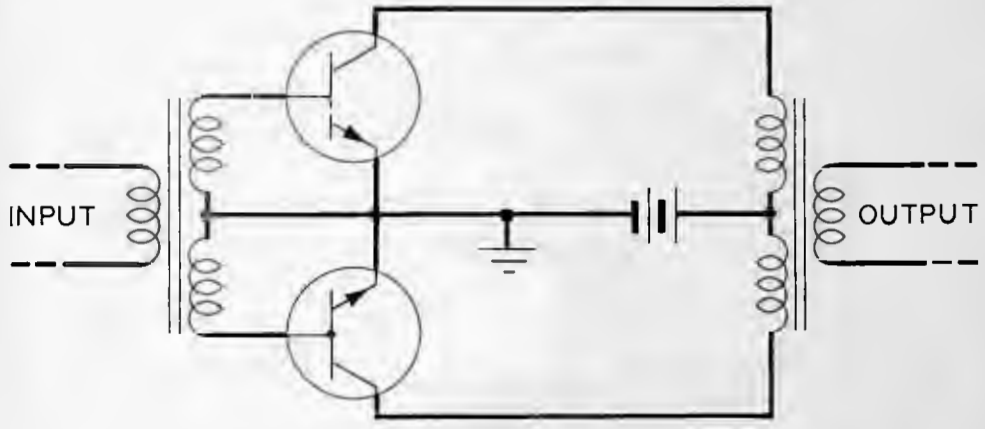


FIG. 16 PUSH-PULL CONNECTIONS FOR NPN TRANSISTOR AMPLIFIERS

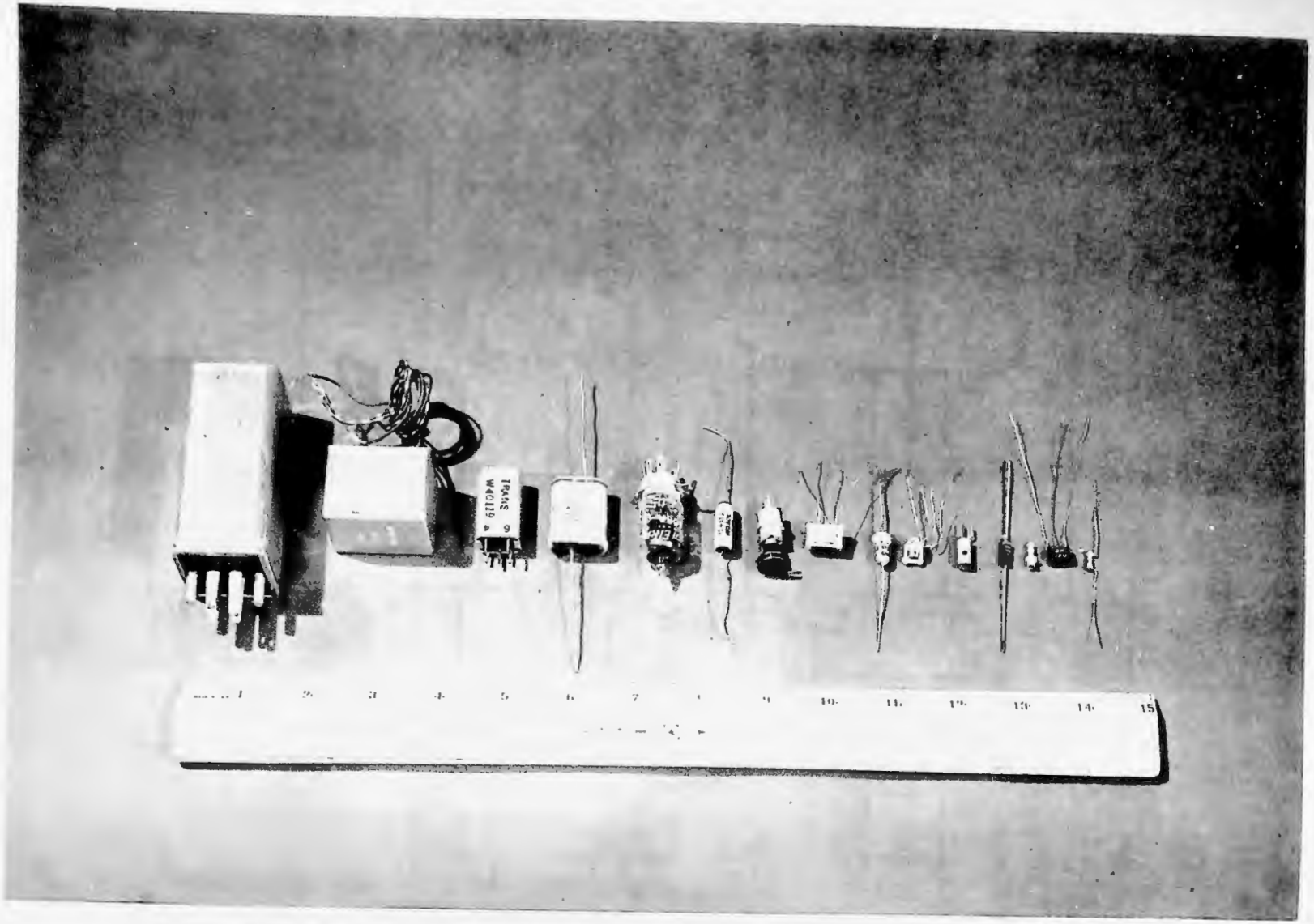
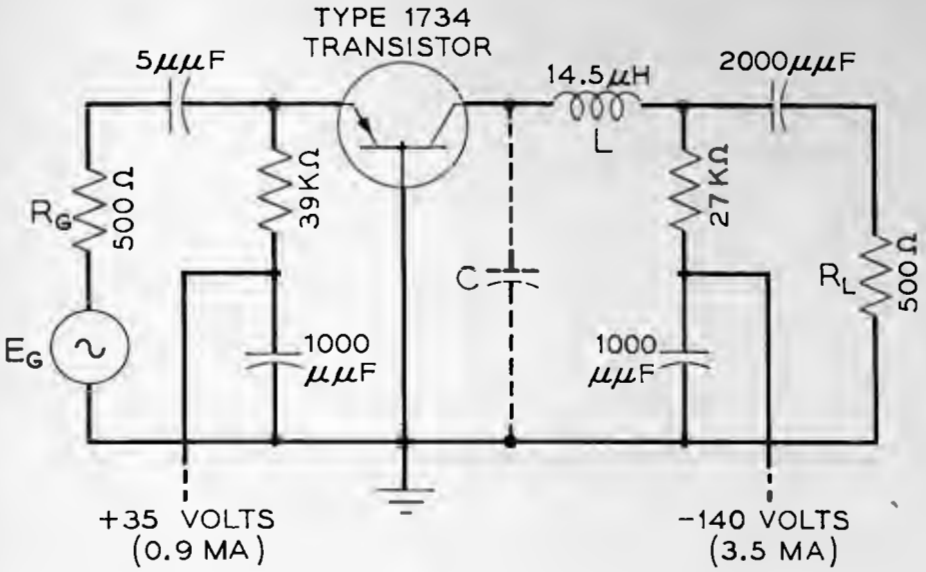


FIG. 17 - COMPARISON OF SIZE OF COMPONENTS USED IN VOICE FREQUENCY AMPLIFIERS



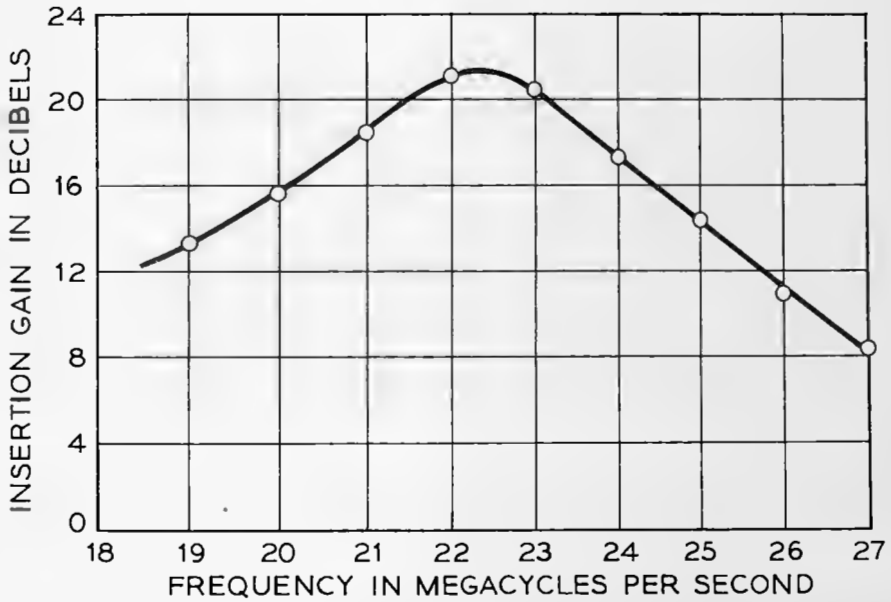
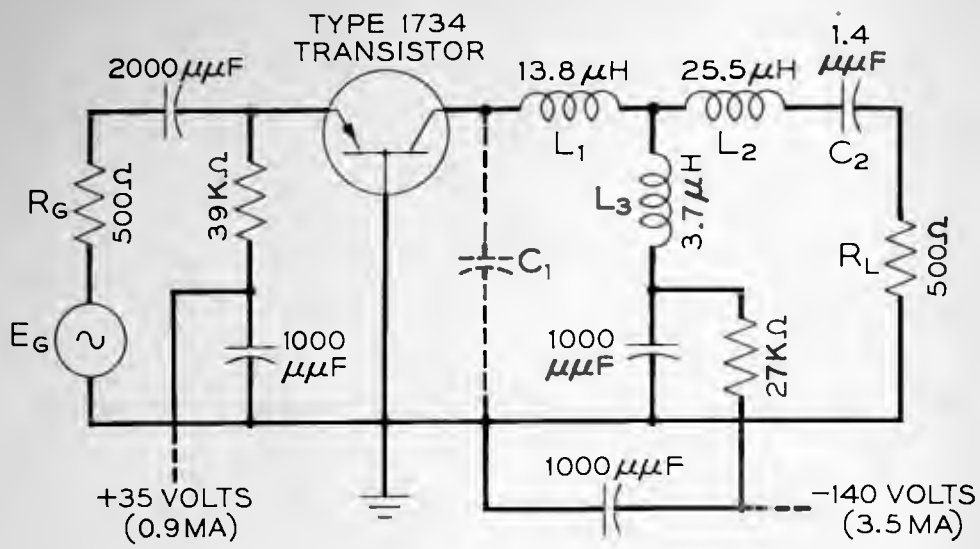


FIG.18 SINGLE TUNED 22 MC IF AMPLIFIER USING POINT CONTACT TRANSISTOR



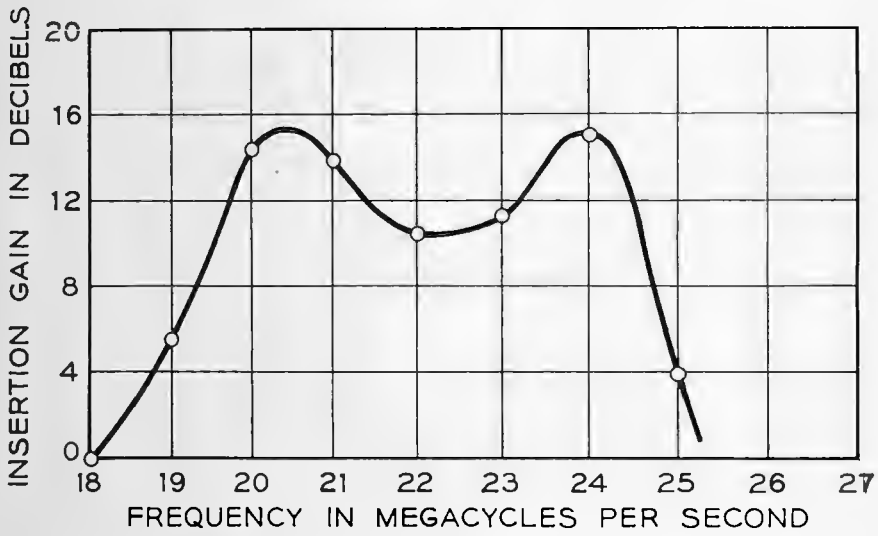


FIG. 19 DOUBLE TUNED 22 MC IF AMPLIFIER USING POINT CONTACT TRANSISTOR

A NEW HIGH-EFFICIENCY LINEAR AMPLIFIER USING A TRANSISTOR AND A VACUUM TUBE

G. RAISBECK

Produced outside the work performed under this contract.

A NEW HIGH-EFFICIENCY LINEAR AMPLIFIER USING A TRANSISTOR AND A VACUUM TUBE

A new linear amplifier is derived from the Doherty amplifier by transforming part of the circuit according to the duality principles explained by Wallace. The new circuit preserves many of the features of the Doherty amplifier, including linearity and high efficiency. A modulator designed according to the same principles is also described.

In 1936 W. H. Doherty described[1] a high efficiency linear amplifier. Doherty's amplifier uses two tubes in an unbalanced circuit, and attains high efficiency as an amplifier for modulated waves because during small-signal conditions one tube is operating as a class B amplifier delivering maximum voltage to the load and the other tube is almost not conducting at all. Further increase in power output is attained very ingeniously by operating the second tube as a class C amplifier in such a way that it lowers the effective load seen by the first tube. In this way the first tube may double its power output while always maintaining high class B efficiency. The second tube, which operates as a class C amplifier, always has a reasonably high efficiency. Doherty shows both theoretically and experimentally that his amplifier can have a no-signal efficiency around sixty-three per cent, and that the efficiency for modulated waves is nearly equal to this for all degrees of modulation.

The most striking feature of the Doherty amplifier is a pair of impedance inverting networks. These have the property that the impedance seen at one pair of terminals is the reciprocal, apart from a constant factor, of the impedance connected to the other two terminals. In other words, an impedance inverting network is a four-pole whose input impedance is the dual of the impedance seen by the other two terminals. In view of the work[2] done by Wallace and Raisbeck on duality and the transistor, it might be suspected that these networks could be eliminated from

- -

1. "A New High-Efficiency Power Amplifier for Modulated Waves, W. H. Doherty, Proc. IRE, Sept. 1936
2. "Duality as a Guide to Transistor Circuit Design, R. L. Wallace and G. Raisbeck, B.S.T.J., XXX, 381 (1951)

the circuit if one of the tubes were replaced by a transistor. This proves in fact to be the case.

Figure 1 shows the output circuit of a Doherty amplifier. It is seen that in this circuit the inverting network transforms the impedance of the plate circuit of a vacuum tube. Using Wallace's duality methods it is easy to find a circuit containing a transistor whose impedance is the dual of the impedance which is connected to the inverting network. If this is done, the inverting network can be eliminated, and the result is the circuit of Figure 2. This is the fundamental circuit of the new linear amplifier.

At this point all that remains is to complete the input circuit and add power supplies. The inverting network in the input circuit of Doherty's amplifier disappears also, for its only function was to introduce a phase change to compensate for the phase change in the other impedance inverting network. A possible circuit is shown in Figure 3.

Alternatively, we can take the form of the Doherty amplifier in which the two inverting networks are in the input and output circuits of the same tube. Then we can simultaneously perform on both the input and output circuits the process illustrated in Figure 1 for the output circuit alone. The result is the same as before.

A simplified explanation of the action of the new amplifier is as follows. The vacuum tube is biased at cutoff, so that it will operate as a linear amplifier. The transistor is biased with a large emitter current, so that for small carrier amplitudes the collector impedance is nearly zero. The two tanks are tuned to the carrier frequency. T_2 therefore presents a very low impedance, and for small signals the circuit is equivalent to the circuit of Figure 4. This is an ordinary linear amplifier. As the carrier amplitude increases, the alternating plate voltage increases until finally it approaches a maximum. At this point, the transistor begins to amplify, and its phase is such that it acts like a negative resistance in the load circuit as in Figure 5. Under these conditions the tube delivers more power without increasing its alternating plate voltage, and the transistor, behaving like a negative resistance, also delivers power. When the carrier voltage reaches its maximum, the negative resistance presented by the transistor is numerically half the load resistance, and the power delivered to the load is four times the maximum power which it can draw from the tube alone. This power is supplied half by the tube and half by the transistor.

Figure 6 shows an experimental model using a CK5703 sub-miniature triode and a type A transistor. The input circuits

use resistive networks and capacitive coupling instead of transformers in order to permit easy adjustments of relative input amplitudes and biases. The transistor is operated with one power supply simply by letting the collector current equal the emitter current. For many transistors this is a satisfactory point of operation for class C amplification. Figure 7 shows the relation of output voltage and efficiency to input voltage for the amplifier of Figure 6. It is clear that the linearity of the amplifier is not perfect, and that the maximum efficiency falls somewhat short of the sixty-three or more per cent predicted by Doherty. As for the linearity, this can be corrected by feedback, as Doherty has already suggested; and as for the efficiency, this is excused by saying that the voltage used is far too low to attain maximum efficiency with the particular tube used, and that in practice transistors have not attained quite the efficiency which is attained in the best class C amplifiers using vacuum tubes.

If one proceeds along the ideas of Reise and Skene[3], one can derive a circuit for a modulator. This bears the same relation to the new linear amplifier that the standard grid modulated class C amplifier bears to the standard one tube linear amplifier. A possible circuit is shown in Figure 8. The carrier and audio signals are now introduced separately, but each appears on both grids. The biases are arranged so that on the lower part of the modulation cycle the transistor is a short circuit and the tube behaves like a conventional grid modulated amplifier, while on the upper half of the modulation cycle the transistor operates as a negative resistance in series with the load, very much as in the linear amplifier. The chief difference between the two is that in the linear amplifier the positive peaks of the carrier input are raised by raising the carrier amplitude, while in the modulator they are raised by momentarily changing the bias of grid or emitter, as the case may be.

In both the amplifier and the modulator, the peak power delivered by the transistor equals the peak power delivered by the vacuum tube. This does not mean, however, that the average powers delivered by the two are comparable. If the modulation envelope is sinusoidal and the per cent modulation is 100, then the tube delivers on the average about twice as much power to the load as the transistor, and ordinarily dissipates more than twice as much energy in the form of heat. At first sight this would seem to indicate that the transistor does not help very much, but the fallacy of that view can be shown by several examples.

First, suppose that there is a limit on the amount of power which may be consumed. Then the increase in efficiency from about 33 per cent to over 50 per cent will permit the power output to increase by half without further increase in power input.

- -

3. U.S. Patent 2,226,258, H. A. Reise and A. A. Skene

Second, suppose that a tube of limited dissipation capacity is being used. For the sake of argument, suppose a tube is used which can dissipate one half watt. Used as a linear amplifier at 33 per cent efficiency for 100 per cent modulated waves, it will deliver about 250 milliwatts of usable power. Two such tubes in parallel or in push pull operated as a conventional linear amplifier will still deliver only 500 milliwatts. However, if the tube is used with a transistor whose power dissipating capacity is only 200 milliwatts, the two can deliver over 700 milliwatts of usable power.

Third, suppose that in a given application, the average degree of modulation is very low, but there are occasional modulation peaks which must be transmitted. Such a situation occurs, for example, in television, where the synchronizing pulses are large in amplitude but short in time. In such a case the duty cycle of the transistor would be very low, and it could be teamed with a tube of much greater power capacity so that the efficiency of the latter was raised considerably.

Other applications will suggest themselves to those who are familiar with the Doherty amplifier. Of course, the present amplifier is not suitable, as yet, for high power applications, because of the limitations imposed by the transistor. It is hoped, however, that it may be of use in cases where the Doherty amplifier might be undesirable on account of the large number of tunable elements and the complexity and size of the connecting networks.

I am indebted to Messrs. R. L. Wallace and W. H. Doherty for advice and criticism.

FIGURES 1 & 2 – OUTPUT CIRCUIT OF DOHERTY AMPLIFIER COMPARED WITH OUTPUT CIRCUIT OF NEW AMPLIFIER

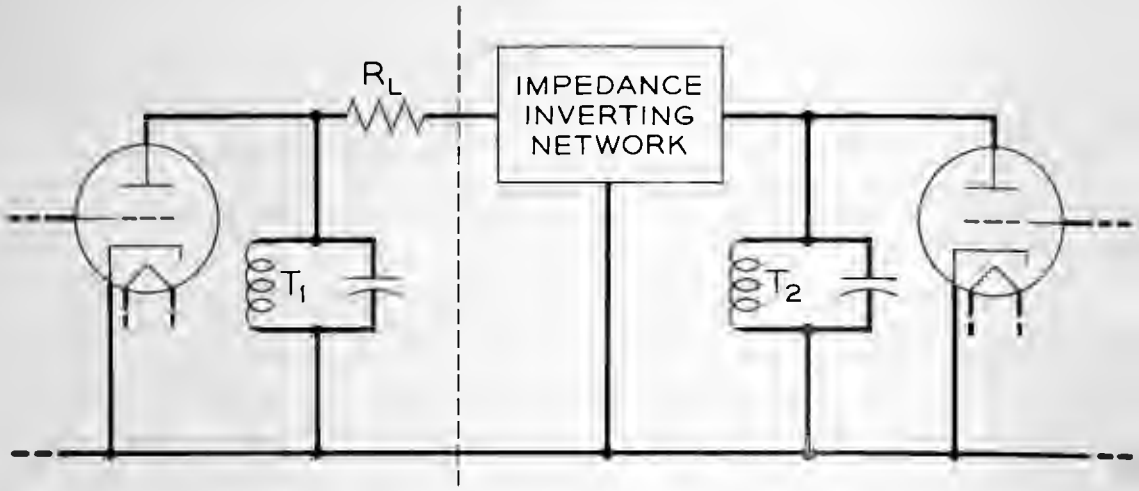


FIGURE 1

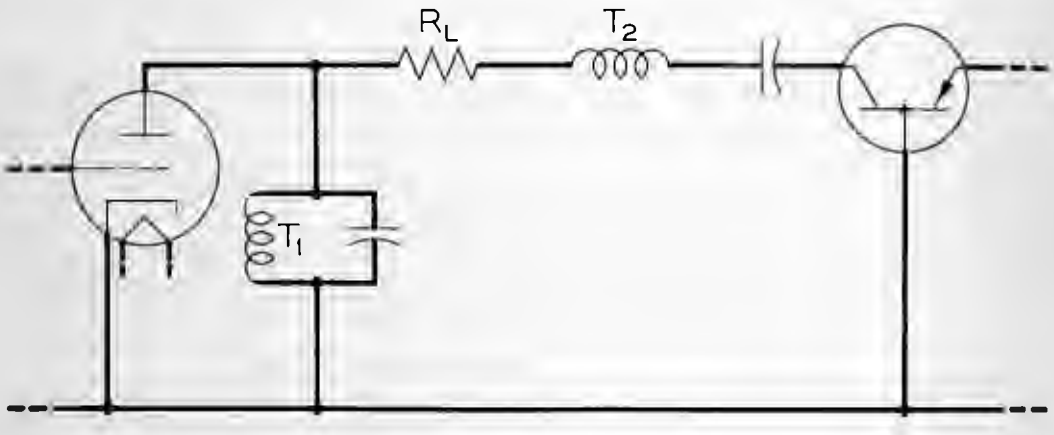
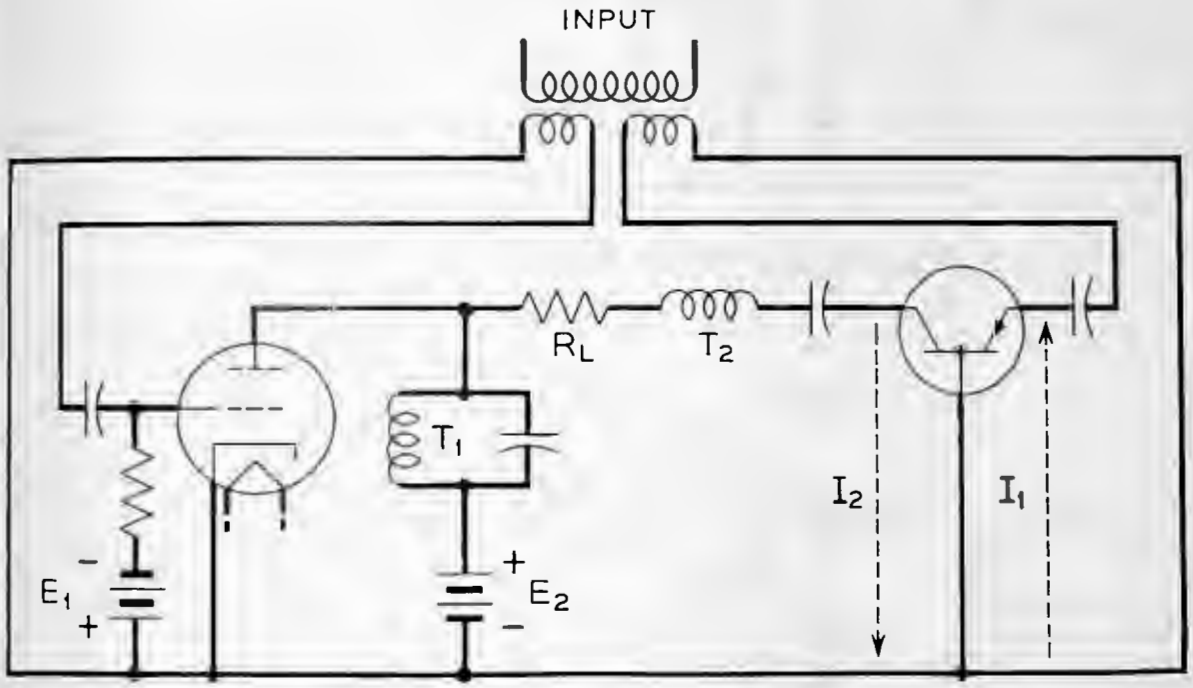


FIGURE 2

FIGURE 3 — LINEAR AMPLIFIER SHOWING INPUT CIRCUIT AND POWER SUPPLIES



FIGURES 4 & 5 – EQUIVALENT CIRCUITS

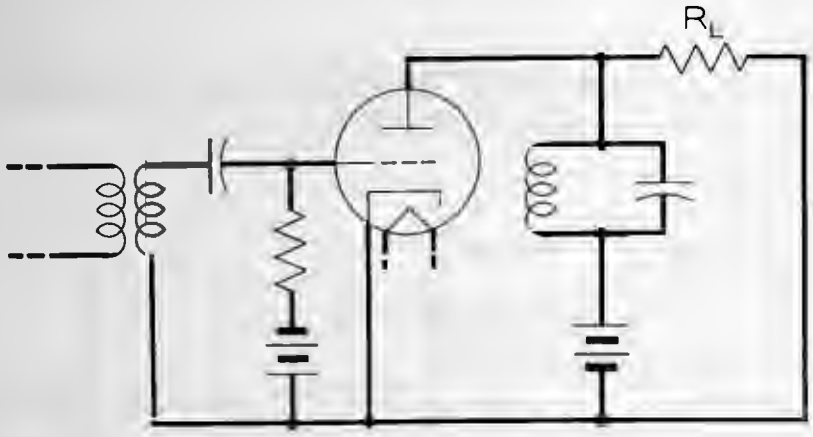


FIGURE 4 – SMALL SIGNALS

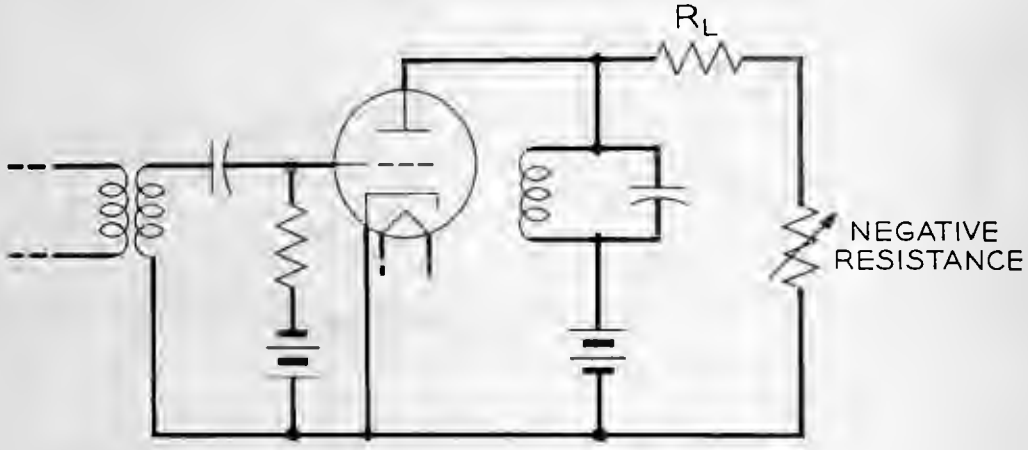


FIGURE 5 – LARGE SIGNALS

FIGURE 6- EXPERIMENTAL LINEAR AMPLIFIER

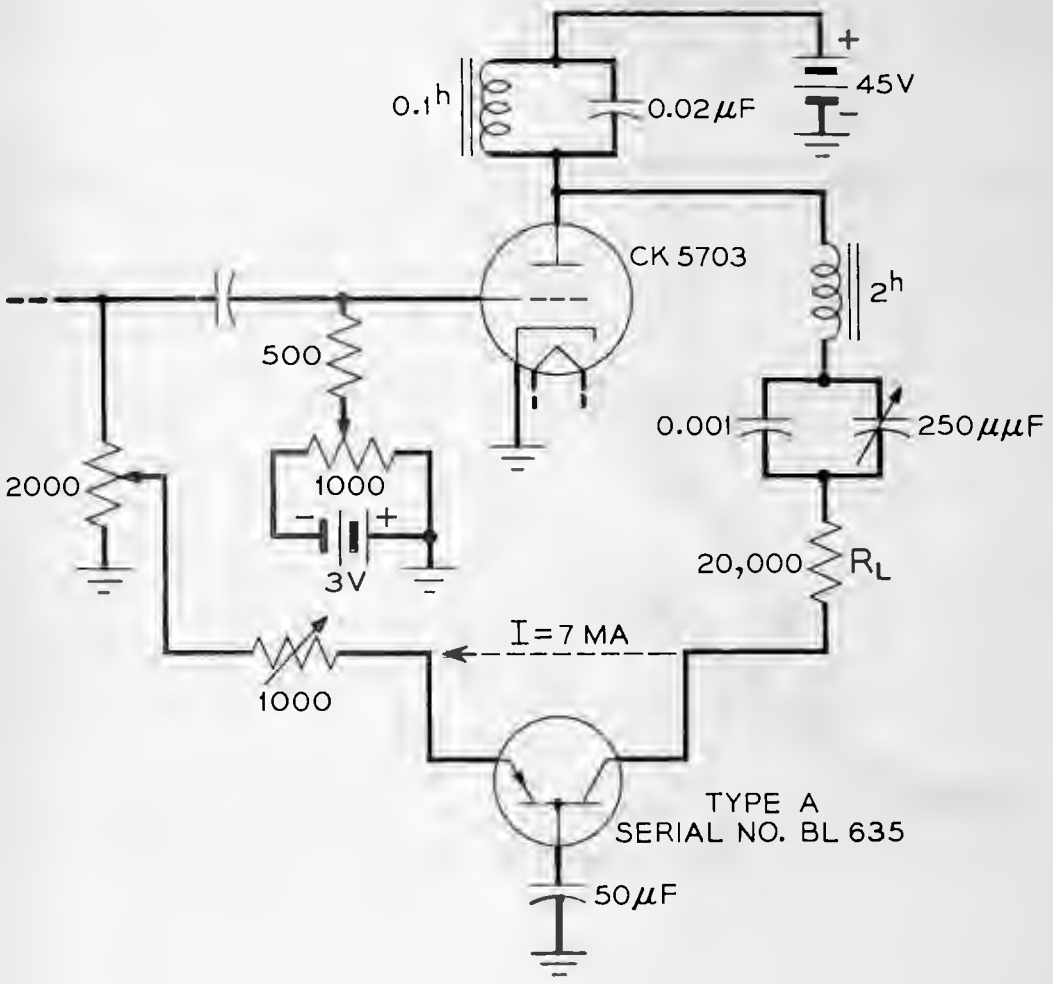


FIGURE 7 – LINEAR AMPLIFIER: OUTPUT AND EFFICIENCY AS FUNCTION OF INPUT

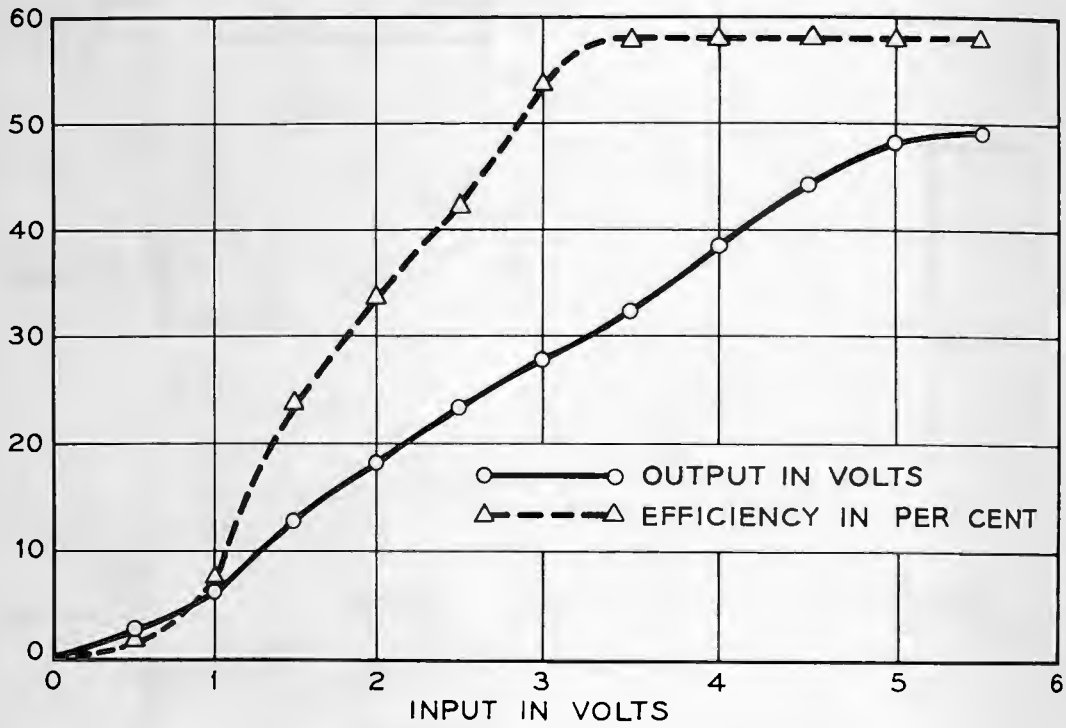
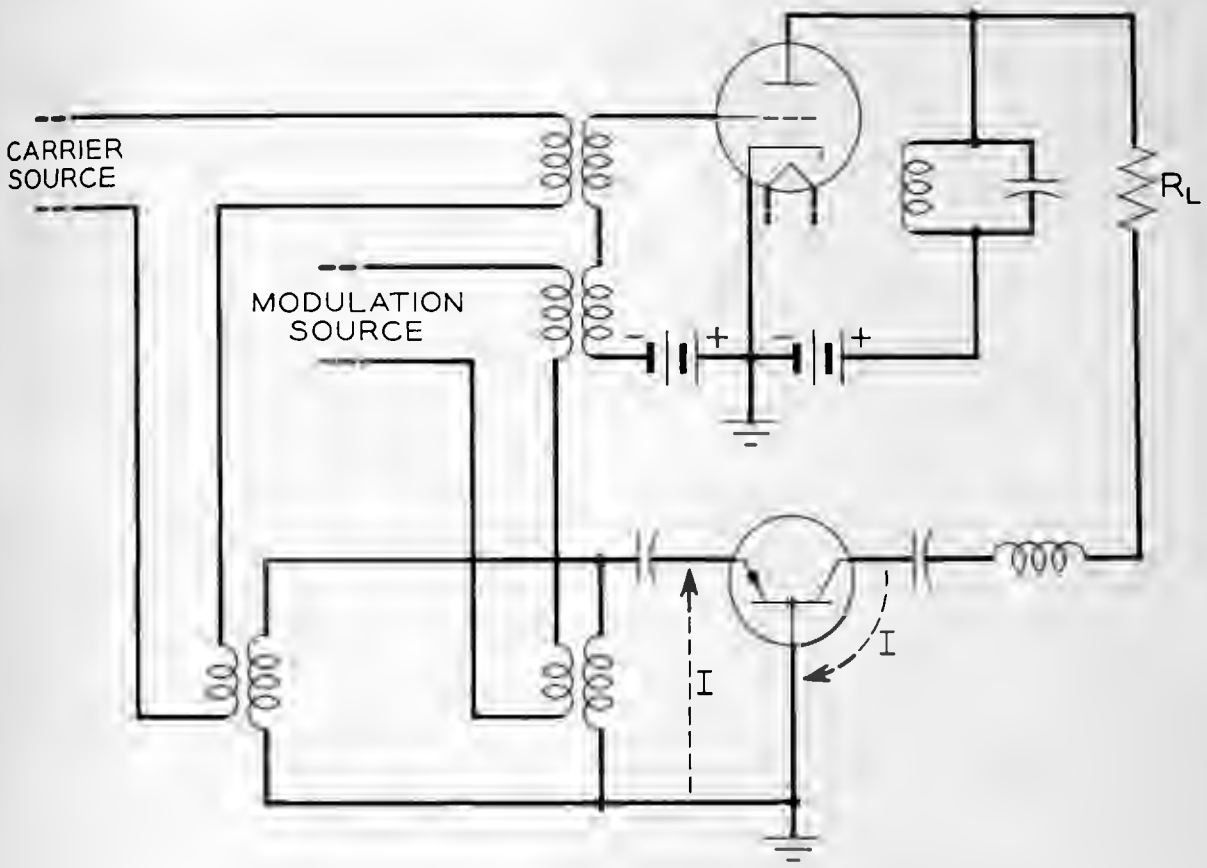


FIGURE 8 - MODULATED LINEAR AMPLIFIER



NO.

A HIGH-EFFICIENCY UNTUNED AMPLIFIER

G. RAISBECK

Produced outside the work performed under this contract.

A HIGH EFFICIENCY UNTUNED AMPLIFIER

A high efficiency amplifier having many characteristics of the Doherty amplifier, but having no tuned circuits, is described. It uses two transistors and two vacuum tubes in two push-pull circuits, connected in series to the load with proper biases.

A recent memorandum* by this author describes a high efficiency linear amplifier using a triode and a transistor which is derived from the well-known Doherty amplifier by replacing one of the power tubes by a transistor. An application of Wallace's duality principles shows that in this manner the impedance inverting networks so prominent in the Doherty amplifier can be eliminated.

The Doherty amplifier is a tuned amplifier. The requirements on the impedance inverting networks are such that they can never function properly over more than a narrow band if they are to be relatively free from loss. However, in the linear amplifier described by this author there are no tuned circuits except the two tank circuits. This suggests at once an attempt to modify the circuit to make an untuned amplifier out of it.

The first requirement is that the circuit must be a push-pull circuit, in order that it will amplify negative as well as positive signals. The second is that according to the manner of the Doherty amplifier, for small signals the vacuum tube amplifier should work alone. The third requirement is that for larger signals, the voltage drop across the vacuum tube should fall to its minimum value when the signal has reached about half the maximum value that the amplifier as a whole is designed to handle without distortion. The fifth requirement is that when the transistor shall begin to act in such a manner that 1) the transistor will deliver power to the load, and 2) the current through the vacuum tube shall increase.

These five requirements together suggest a circuit similar to the one in Figure 1. Here we have an ordinary class B

- -

* "A New High-Efficiency Amplifier using a Transistor and a Vacuum Tube"

triode amplifier and a class B transistor amplifier using the circuit of Wallace, connected by separate transformers to the load. The transistor amplifier is not biased, however, to behave as a class B amplifier, but the emitter bias is made large enough so that for small signals on the emitter both collectors will present very low impedances.

For small signals the vacuum tube amplifier behaves as an ordinary class B amplifier, because the impedance presented in the load circuit by T_2 is small. Let us imagine a signal which increases steadily from a small value. The plate voltage of one tube (say V_1) will fall steadily and the plate current will rise steadily. The current through the load is, apart from a constant factor, the same as the plate current of V_1. Presently, however, the plate voltage of V_1 falls to zero, and its plate current has risen to what we might suppose is a maximum value. At this point, however, the emitter current of Tr_1 falls to such a value that the collector circuit no longer presents zero impedance. The collector voltage falls to a negative value, and part of the current which was passing through the collector to the base is now driven through the primary of T_2 through the other transistor, which still presents a low impedance, to the common base. This induces further current in the load circuit, and incidentally causes more current to flow through the tube. This is possible now because the tube presents a low impedance to the secondary of T_1. The total effective driving voltage is now the sum of the voltage drop across the tube and the voltage drop across the transistor. These relations are shown in Figure 2.

Figure 3 shows the relations of voltage and current when the input is a sinusoidal wave of large amplitude. From the third and fourth lines one can see how the transistor supplies the missing voltage peaks which are cut off of the plate voltage wave when the tube limits. The sixth line shows the current and voltage across the secondary of T_1. The voltage wave shows the typical evidence of limiting. The next line shows the voltage and current for the secondary of T_2. The currents are of course the same. From these two graphs it is clear that the total effective exciting voltage is sinusoidal in spite of the fact that all four of the individual wave forms are far from sinusoidal.

It is possible to compute the efficiency of the device on the assumption that the collector voltage of the transistor and the plate voltage of the tube are strictly zero when they fall to their minimum values, i.e., that the tube and the transistor are perfect short circuits when they are supposed to present low impedances. The results of the computations are

shown in Figure 4, which shows the plate- and collector-circuit efficiencies for sine wave inputs of various magnitudes. The efficiencies predicted are obviously greater than those to be expected from a class B amplifier of the ordinary type. Of course, any actual circuit would fall short of this efficiency to the same degree that an ordinary class B amplifier falls short of its predicted maximum efficiency.

This amplifier might be useful in any low or medium power application where low power drain and low dissipation are important. It is true, furthermore, that as in the Doherty amplifier and the tuned linear amplifier described by the author, one of the active elements is worked much harder than the other. This allows us to use transistors of relatively low capacity to improve the efficiency of considerably larger vacuum tubes.

Up to now no example of an amplifier of the type described has been built. It is hoped that an experimental model will be built in the near future.

FIGURE 1 – HIGH EFFICIENCY UNTUNED AMPLIFIER

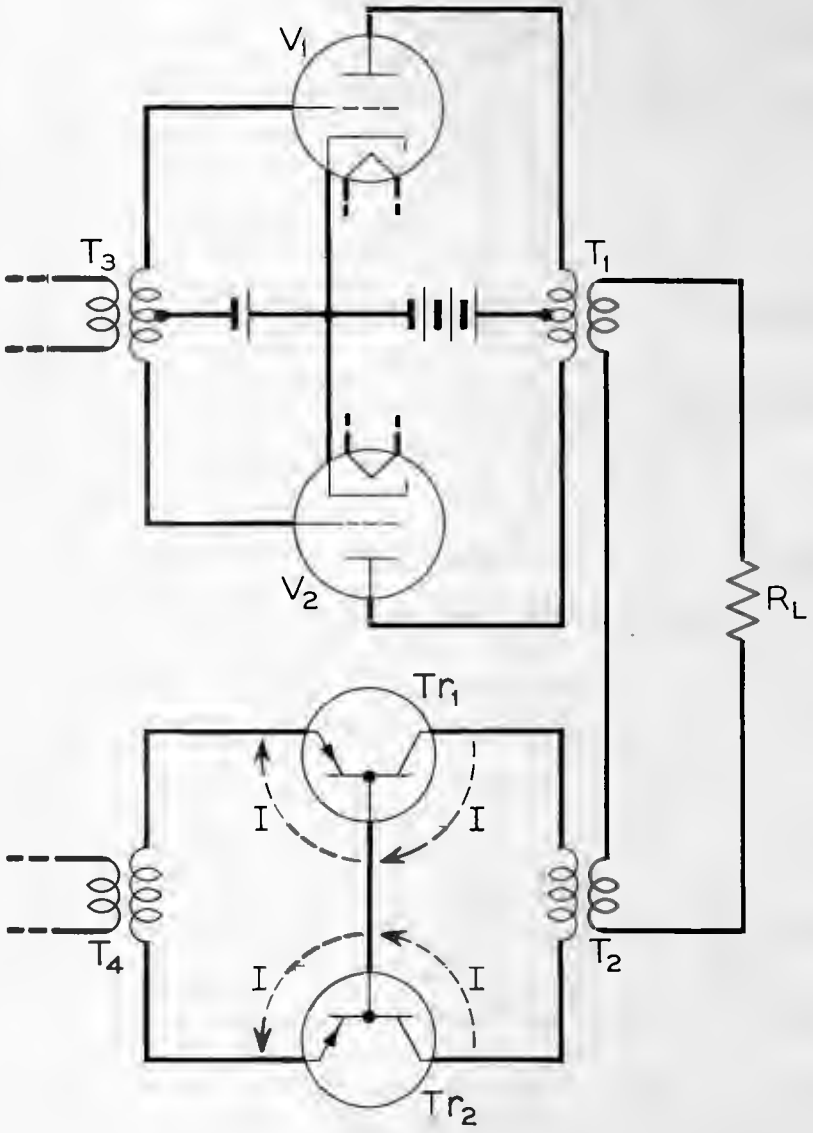


FIGURE 2 – VOLTAGE AND CURRENT RELATIONS, MONOTONIC LINEAR INPUT SIGNAL

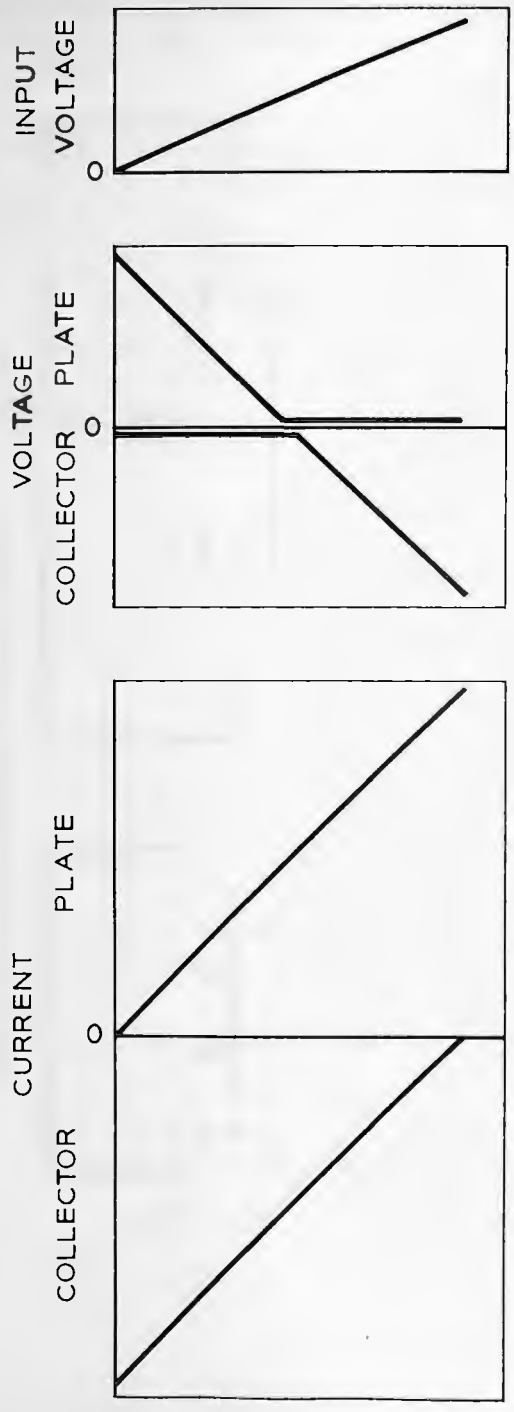


FIGURE 3 — VOLTAGE AND CURRENT RELATIONS FOR SINE WAVE INPUT

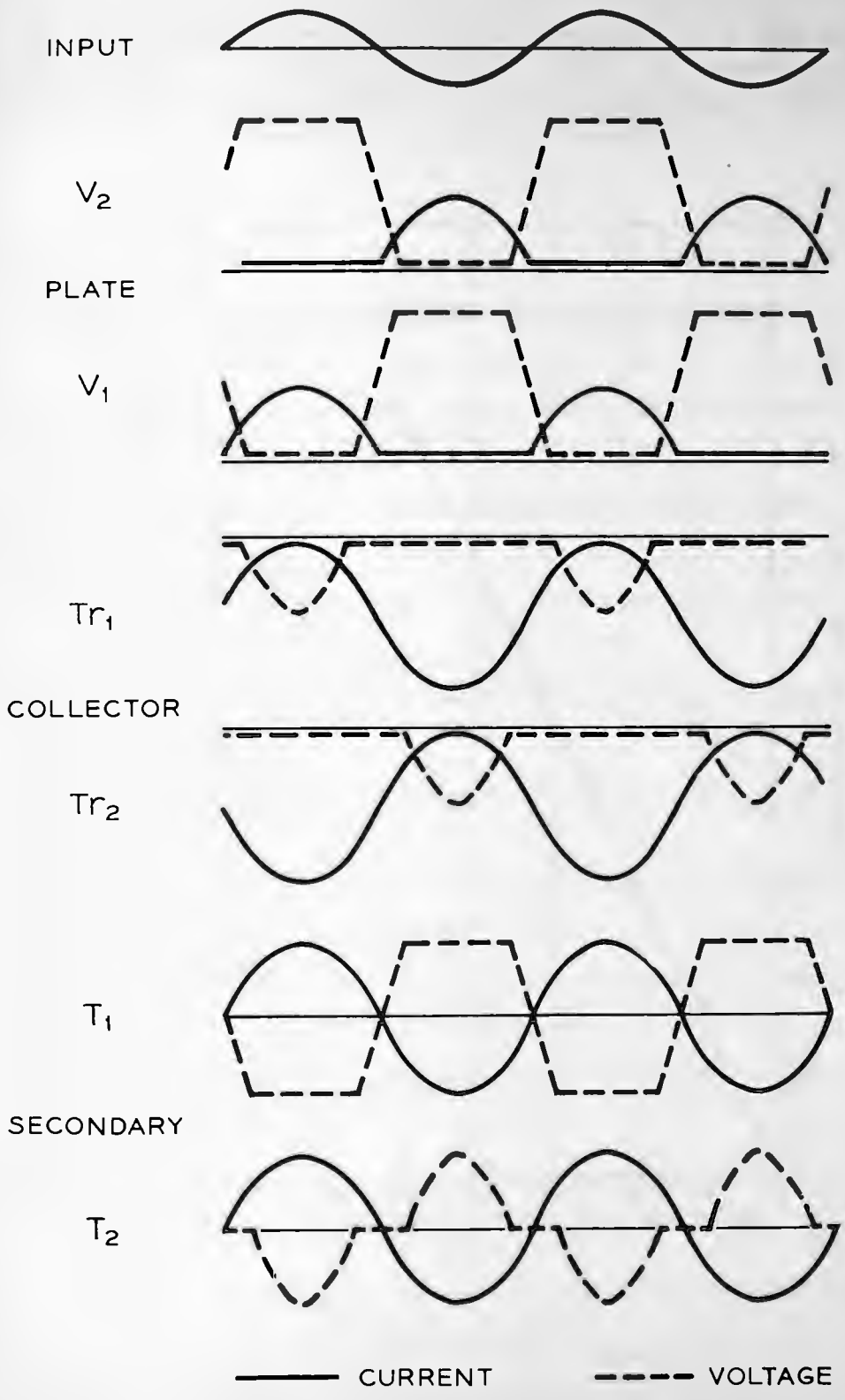
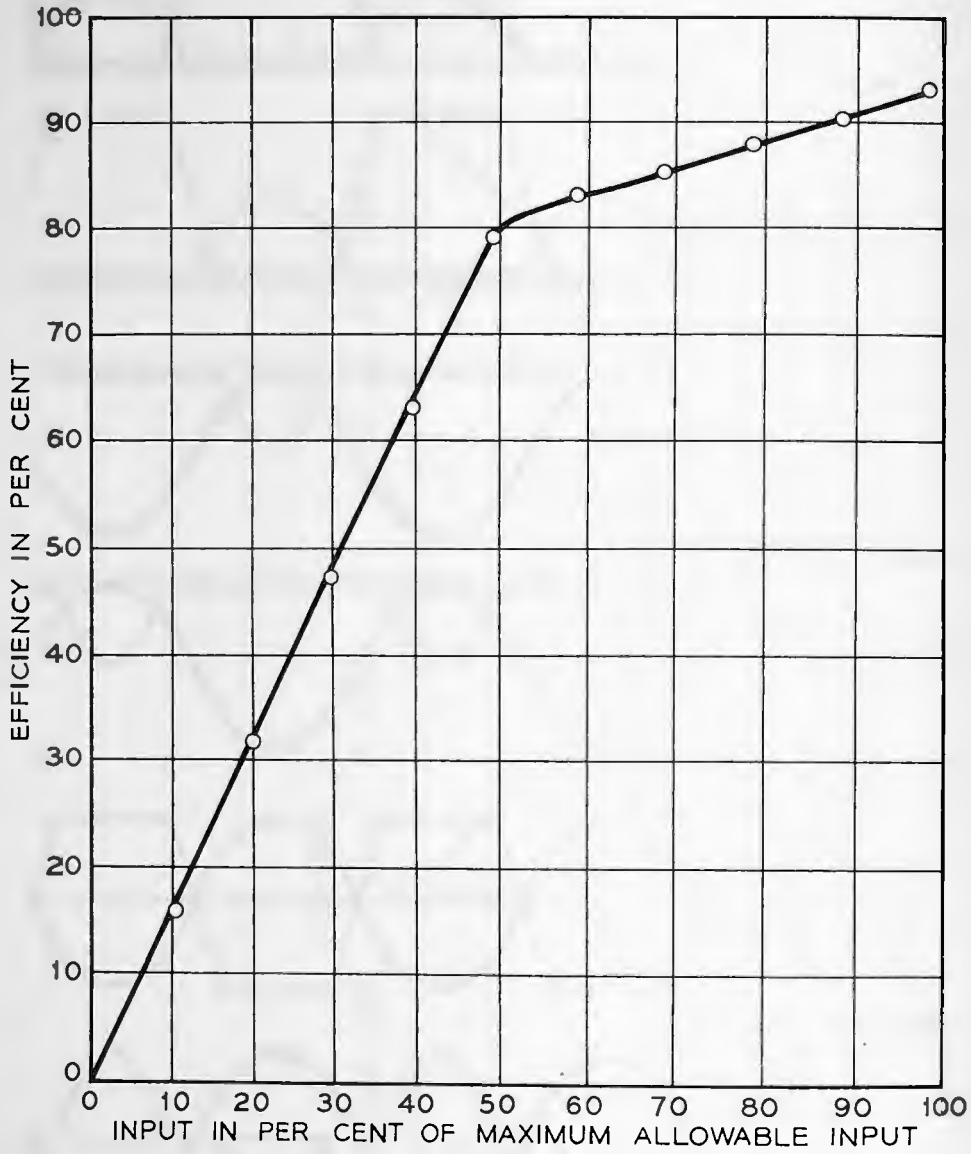


FIGURE 4 – THEORETICAL CIRCUIT EFFICIENCY AS A FUNCTION OF INPUT VOLTAGE FOR A SINUSOIDAL INPUT. SKETCHED FROM A FEW COMPUTED POINTS.



SOME EXPERIMENTAL AND PRACTICAL APPLICATIONS OF TRANSISTOR OSCILLATORS

R. S. CARUTHERS

Produced outside the work performed under this contract

SOME EXPERIMENTAL AND PRACTICAL APPLICATIONS OF TRANSISTOR OSCILLATORS

A small number of oscillator circuits built at BTL at frequencies ranging from audio to broadcast range are described in this section. Mostly these circuits were of an experimental nature but some were built for specific system uses and are either in the production stage or nearly so. All of the circuits described use point contact transistors, largely because NPN types were not available at the time of this work.

In general, the oscillators do not represent types where precision performance was required. However, it seems reasonable that transistors would have introduced no particularly difficult problems if had been necessary to obtain extremely high stability of output or frequency, very pure sine wave and uniformity with different units.

An example of an audio frequency oscillator using the early type A transistor is shown in Fig. 1. This unit has been widely used during the last four years for demonstrations throughout the country. Feedback from collector to emitter is by means of a transformer whose collector winding is tuned. A single battery is used and emitter bias is obtained from the drop in a resistor in the base lead.

Fig. 2 shows the schematic of two more early oscillators. These circuits are those of the local oscillators in

two experimental broadcast-band superhetrodyne receivers. In Fig. 2(a), feedback is by means of inductive coupling similar to that in the previous audio frequency oscillator. A single power supply is used with the emitter bias again taken from a resistor in the base circuit. In this case, however, a variable bleeder resistance is also provided so the emitter bias may be adjusted.

In the oscillator in Fig. 2(b), feedback from collector to emitter circuit is by means of a tuned circuit in the base lead which acts as a common impedance between these circuits. Here the emitter bias is furnished by a second voltage supply.

Both of the oscillators in Fig. 2 cover the frequency range of approximately 1 to 2 megacycles. Since the equipment was only experimental, selected type A transistors were used in each circuit and no attempt was made to attain interchangeability of the transistors.

Fig. 3(a) shows the schematic of an audio frequency transistor oscillator. This oscillator supplies a ringing tone of 1600 cycles in a lineman's test set used in maintaining the repeaters of a carrier telephone system. While batteries could be used to power the oscillator, the failure

of the batteries would leave the linesman without means of calling the test board. The oscillator is, therefore, operated from 20 to 30 volts d-c produced by a rectifier consisting of four germanium varistors. The rectifier is energized from a hand cranked magneto generator normally used by the linesman for ringing on voice frequency lines. This magneto generator has a nominal frequency of 16 cycles, but may vary with speed of cranking from 12 to 36 cycles and often has a very poor wave shape.

In the oscillator of Fig. 3, feedback from emitter to collector is through a transformer having a tuned collector winding. Additional feedback is obtained by means of the 100 ohm resistor in series with the base of the transistor. The 1200 ohm resistor in series with the emitter affords protection against surges originating both within the set and on the telephone line to which the set is connected. Any transistor of the A1723 type may be used without readjustment of the circuit.

The test set oscillator supplies approximately 1 milliwatt into a 600 ohm load and has a frequency of 1600 cycles. Frequency is well within the required tolerance of ± 20 cycles under all conditions of temperature, humidity, changes in transistors and in supply voltage due to variation in speed of cranking the magneto generator, etc. The wave shape is substantially sinusoidal with the total harmonic content approximately 40 decibels below the fundamental output.

In field equipment of the type being discussed, ability to operate under extreme temperature conditions is important. Therefore, the test set is designed for operation between the limits of -67°F to +150°F. The curve in Fig. 3(b) shows that the total variation in output over this range is between one and two decibels. Changes in temperature within this range changes the frequency only in the order of two to three cycles.

Fig. 4 shows the schematic of a low audio frequency (80 cycles) oscillator on which circuit development has now been completed. This oscillator features operation from a single 6 volt power supply, low power drain and constant output with a wide range of transistor characteristics.

As in the first and fourth oscillators emitter bias is supplied from the drop in an impedance in the base lead. However, in this oscillator the power available was so low that the drain through bleeders had to be restricted. By using two 3/16" disc copper oxide varistors in the base an impedance was obtained which has high d-c resistance at the moment power is applied. As oscillations build up, the additional collector current flowing through these varistors decreases their d-c resistance.

Constant output is achieved by an automatic-volume-control circuit. A portion of the output is rectified and the resulting voltage, of approximately 1/4 volt across the 250 ohm resistor between the center tap of the output transformer and ground, is applied to the emitter. The polarity of the AVC voltage opposes the emitter bias obtained from the drop across the varistors in the base circuit, giving a net negative emitter bias. Variation in this effective bias varies the operating angle thus controlling the output. This Class C operation also increases the output efficiency.

Ten milliwatts of output power is obtained with an input of 28.9 milliwatts when operating from a single six volt battery. This is an over-all efficiency of 34%. Actually, a large portion of this power is lost in bleeder current, transformer losses, etc. so the actual efficiency in collector circuit alone is of the order of 60%. At this output the total harmonic content is approximately 21 db down on the fundamental.

Fig. 5 is a statistical plot showing how the output of the oscillator of Fig. 4 changes due to variations in the characteristics of transistors. Curve B is the result of tests made using 61 transistors of the M1768 type all of which met the manufacturing specification. The degree of uniformity in output is exceptional. For 90% of the units tested, the variation was less than .4 db from the mean. Curve A shows the result of tests on 135 units including both the 61 units of Curve B and 74 additional

units which had been rejected for not meeting the M1768 limits. Even for this spread in varistor characteristics the variations in output were extremely small.

Fig. 6 is the schematic of a piezo-electric transistor oscillator which has been studied as a possible substitute for a vacuum tube oscillator used to supply carrier to a modulator in a carrier telephone system. In one experimental circuit, a grounded base point contact transistor amplifier driven by this oscillator delivered 30 milliwatts at 184 kilocycles. This is somewhat less than the 200 milliwatts that can be obtained from the electron coupled vacuum tube oscillator currently used but is nevertheless sufficient for the use contemplated.

An interesting test result in using the crystal transistor circuit is the fact that the total power required by the transistor oscillator and amplifier is .8 watts less than that required by the vacuum tube circuit. Both circuits were operated from normal +130 and -48 volt telephone office batteries. If more suitable power supply voltages can be made available, this power economy can be increased to over 1-1/4 watts. As most present day use of miniature tube circuits crowds a maximum number into small space, the saving of even such small amounts of power is important in reducing heating.

The simplicity of the circuit in Fig. 6 is due to the fact that in transistors the a-c collector current is in phase with the a-c emitter current. Hence to obtain oscillations it

is only necessary to feed the collector current back to the emitter through a series resonate circuit or its equivalent a piezo-electric plate. The fed back collector current must, of course, exceed the emitter current.

The operating point Q in Fig. 7 is initially determined by the supply voltages and the resistors in series with them. When the circuit oscillates the collector current and voltage follow the a-c load line shown in the lower portion of Fig. 7. Limiting action takes place on the peaks of the cycle due to decrease in alpha at both low and high values of emitter current. This is shown in the upper portion of Fig. 7.

A mathematical analysis shows that the approximate condition for oscillations is

$$\alpha \geq 1 + \frac{R_o}{R_L} + \frac{R_o}{r_c} \qquad (1)$$

in which R_o is the equivalent resistance of the piezo-electric plate, R_L the collector load resistor and r_c the usual parameter of the transistor. In point contact transistors r_c is from 15000 to 20000 ohms so the third term of (1) is of second order importance. R_o is from 2000 to 6000 ohms for small size piezo-electric plates thus restricting R_L to values of the same order of magnitude, as the alpha of most point contact transistors does not exceed 2 in the middle of their operating range.

Equation (1) indicates one important difference between a direct coupled collector-emitter circuit oscillator using point contact transistors and the usual vacuum tube oscillator. While

in most vacuum tube oscillators, there is considerable "excess gain", in this transistor oscillator there is practically no reserve margin in the current amplification. This would lead one to suspect considerable variations in operation in this type of transistor oscillator circuit.

The effect of changing transistors is shown in Fig. 8. Here curves of output versus temperature are shown for six different transistors of the 1729 type. There is considerable variation in output and the maximum operating temperatures vary widely. The failure to oscillate at higher temperatures is due to a large decrease in the value of r_c, a phenomena quite similar to the increase in reverse current in germanium diodes at high temperatures.

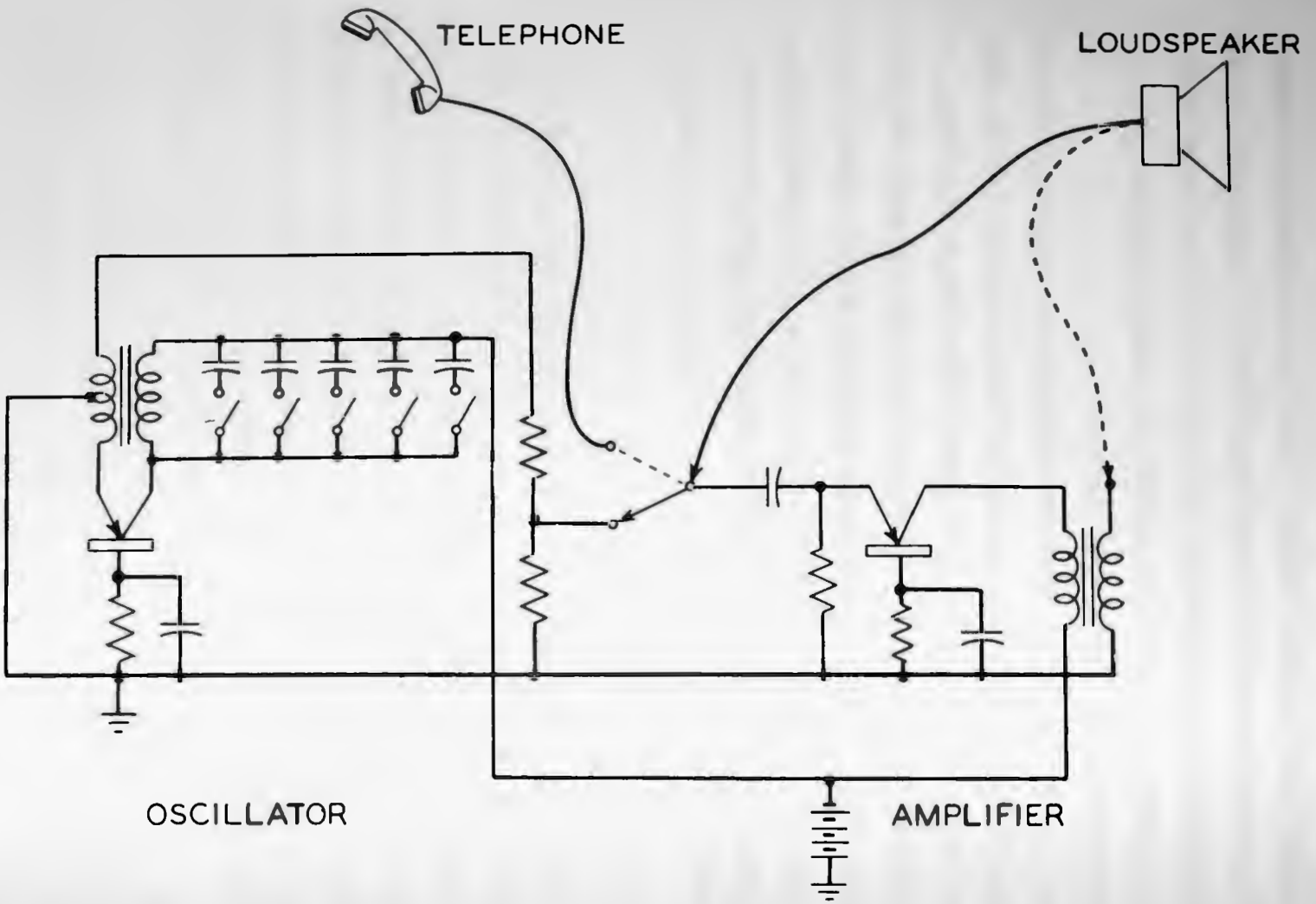


FIG. 1 - AN EARLY AUDIO-FREQUENCY TRANSISTOR OSCILLATOR

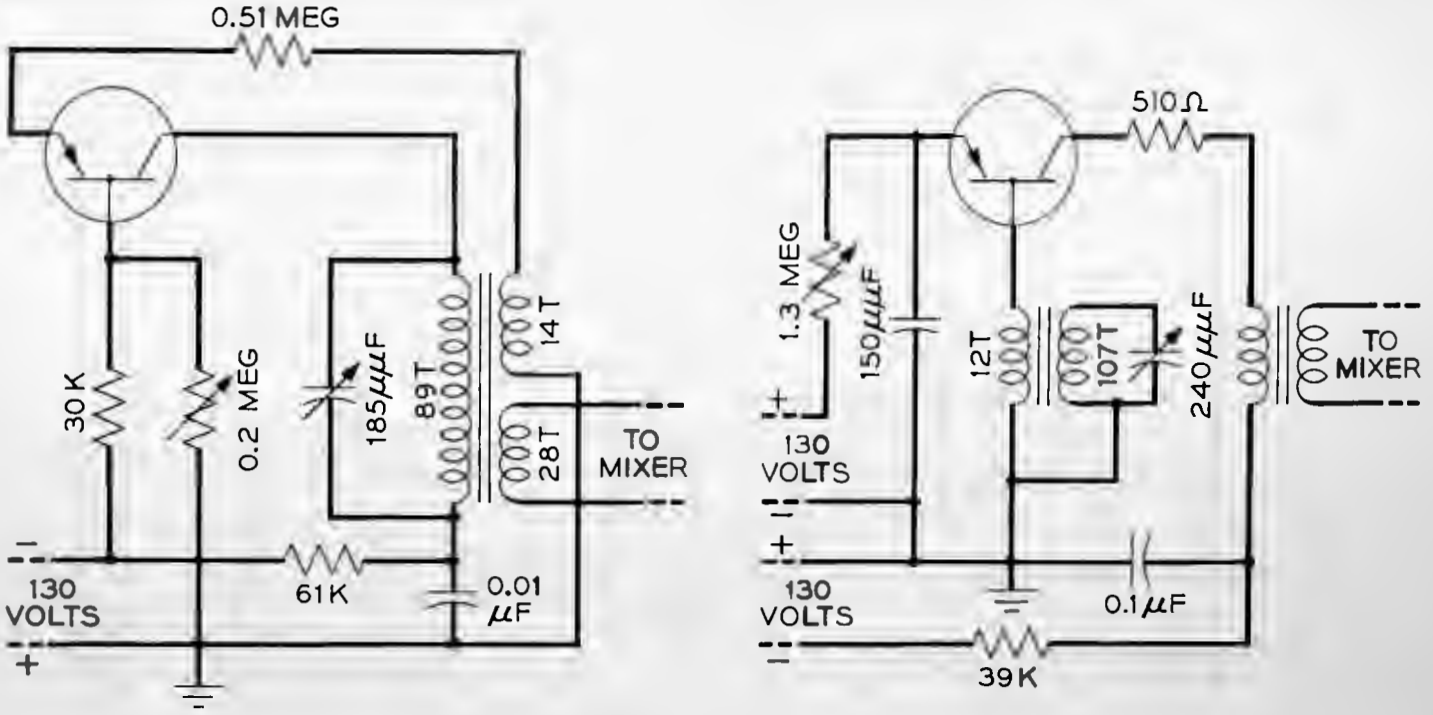
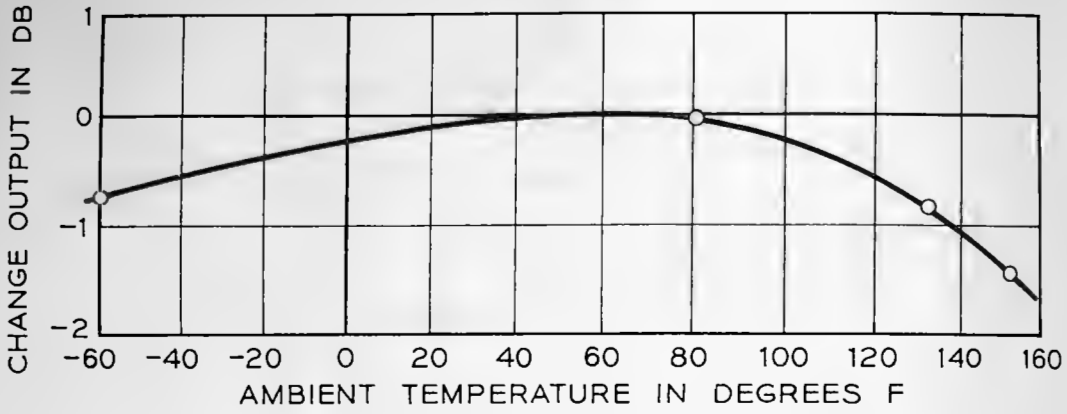


FIG. 2 EXPERIMENTAL OSCILLATORS FOR SUPER-HETRODYNE BROADCAST RECEIVERS



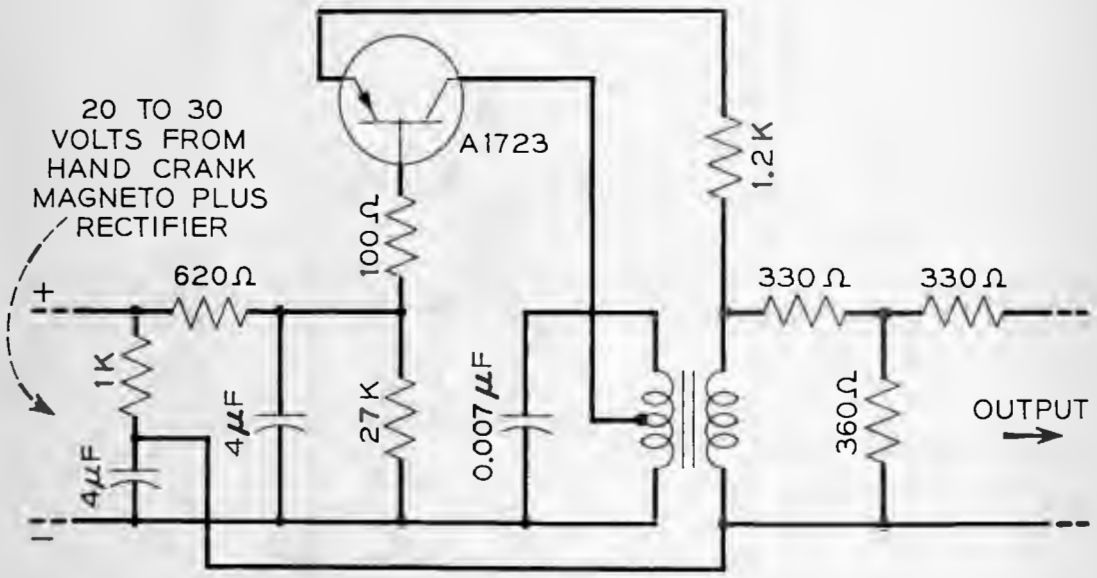


FIG. 3 AUDIO FREQUENCY OSCILLATOR FOR USE IN LINEMAN'S TEST SET.

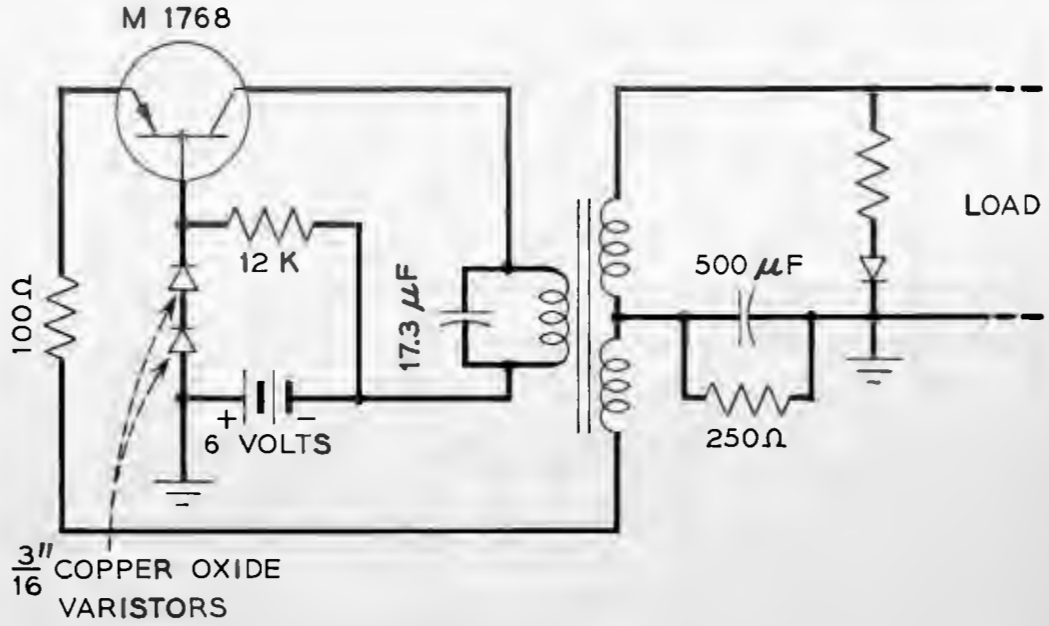


FIG. 4 HIGH EFFICIENCY AUDIO OSCILLATOR

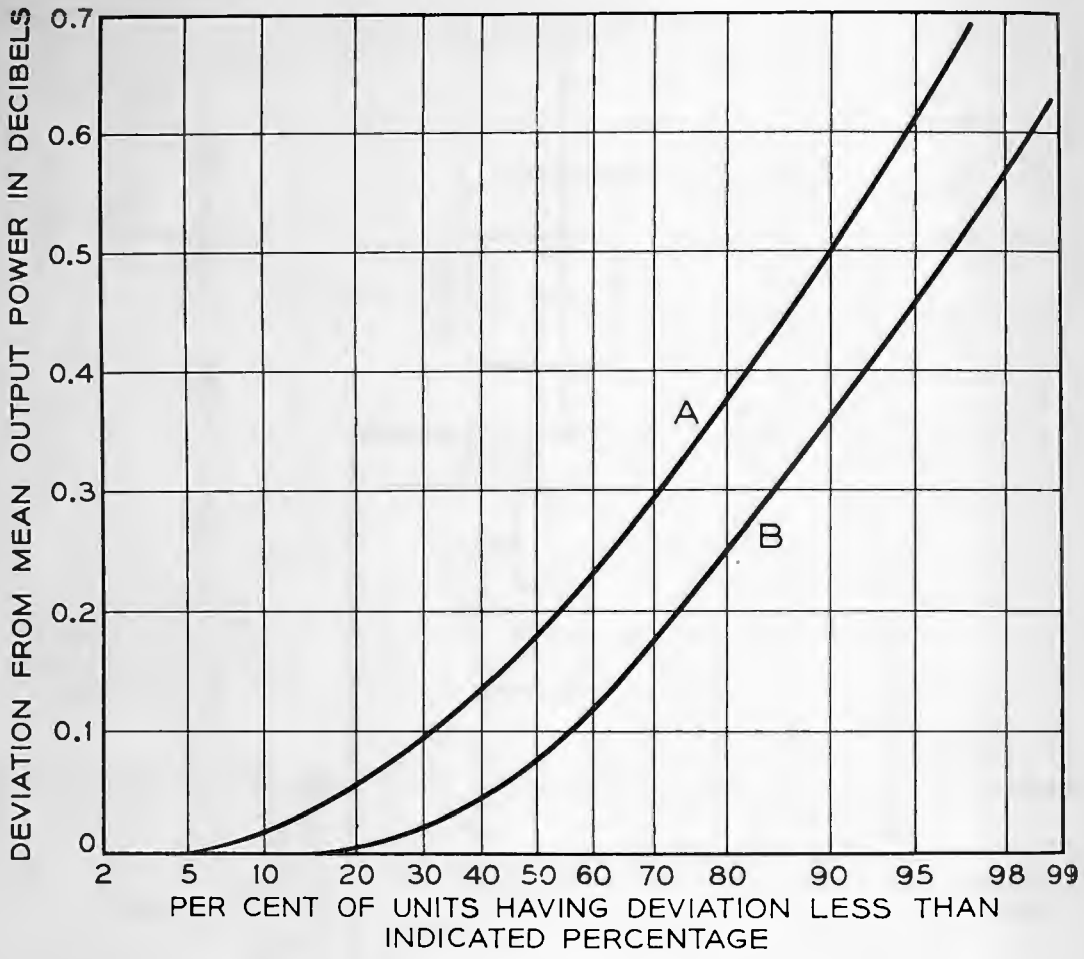


FIG. 5 DISTRIBUTION CURVES OF POWER OUTPUT FROM OSCILLATOR IN FIGURE 4

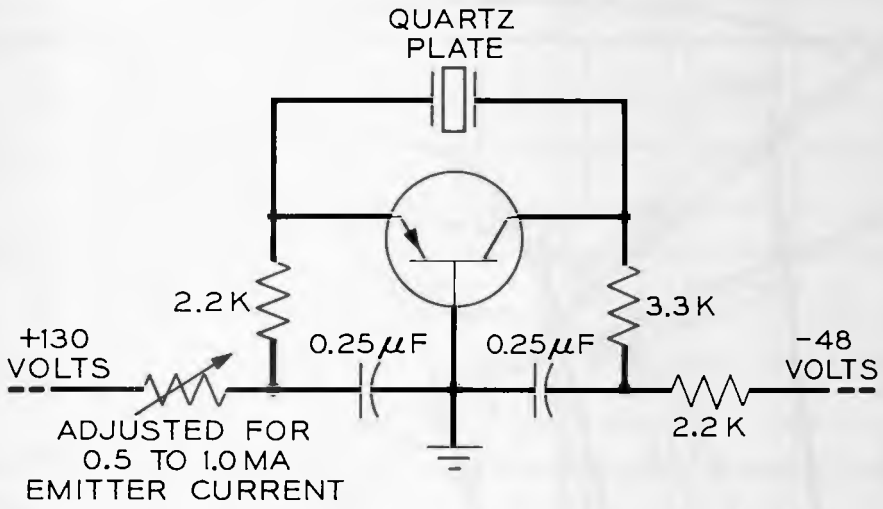
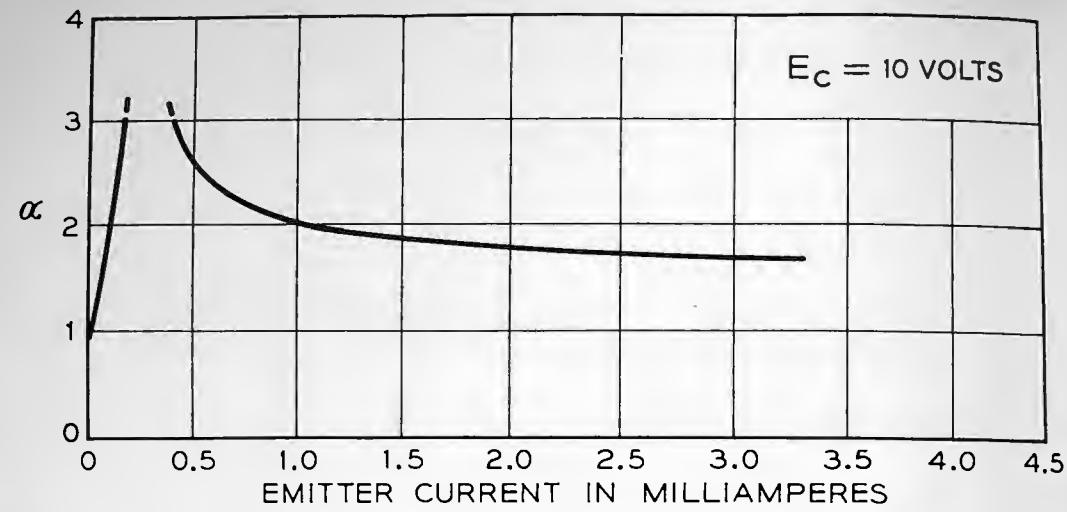


FIG. 6 PIEZO-ELECTRIC CARRIER FREQUENCY OSCILLATOR



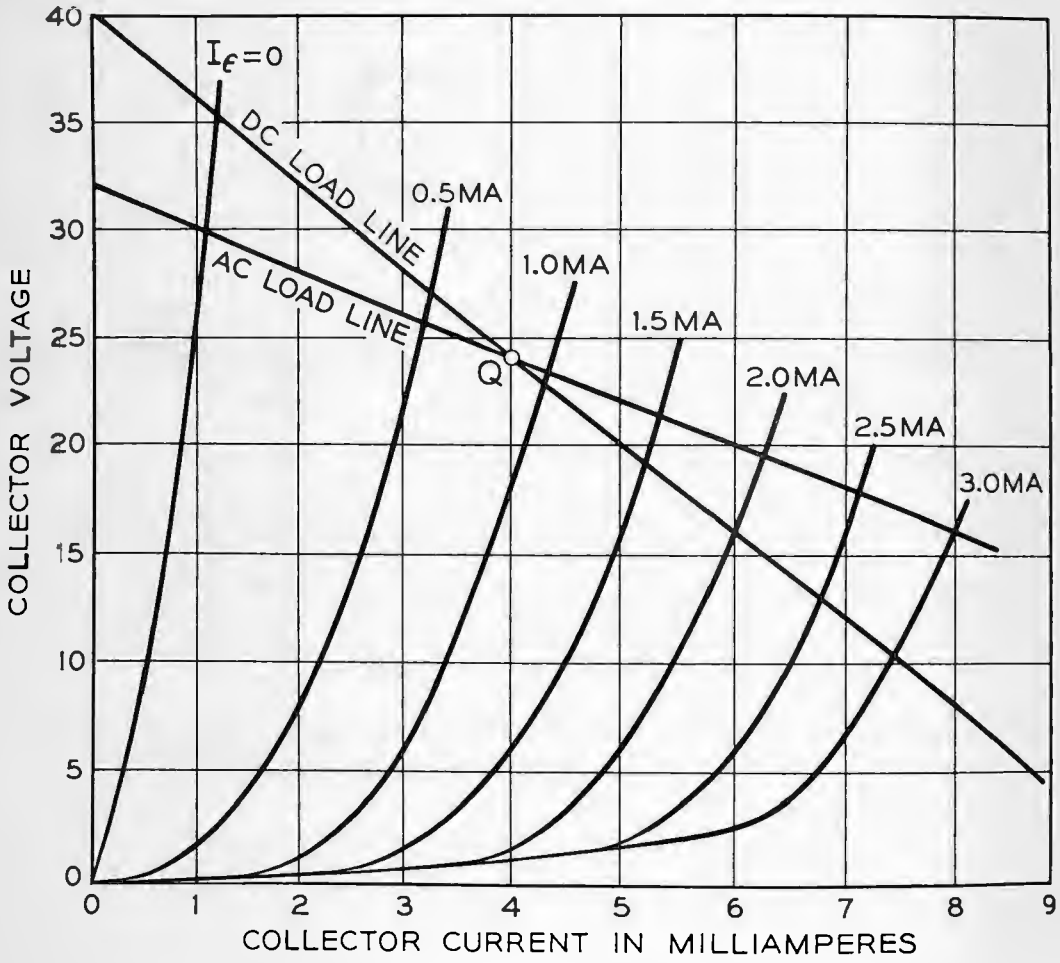


FIG. 7 COLLECTOR CHARACTERISTICS OF POINT CONTACT TRANSISTOR SHOWING OPERATING PATH.

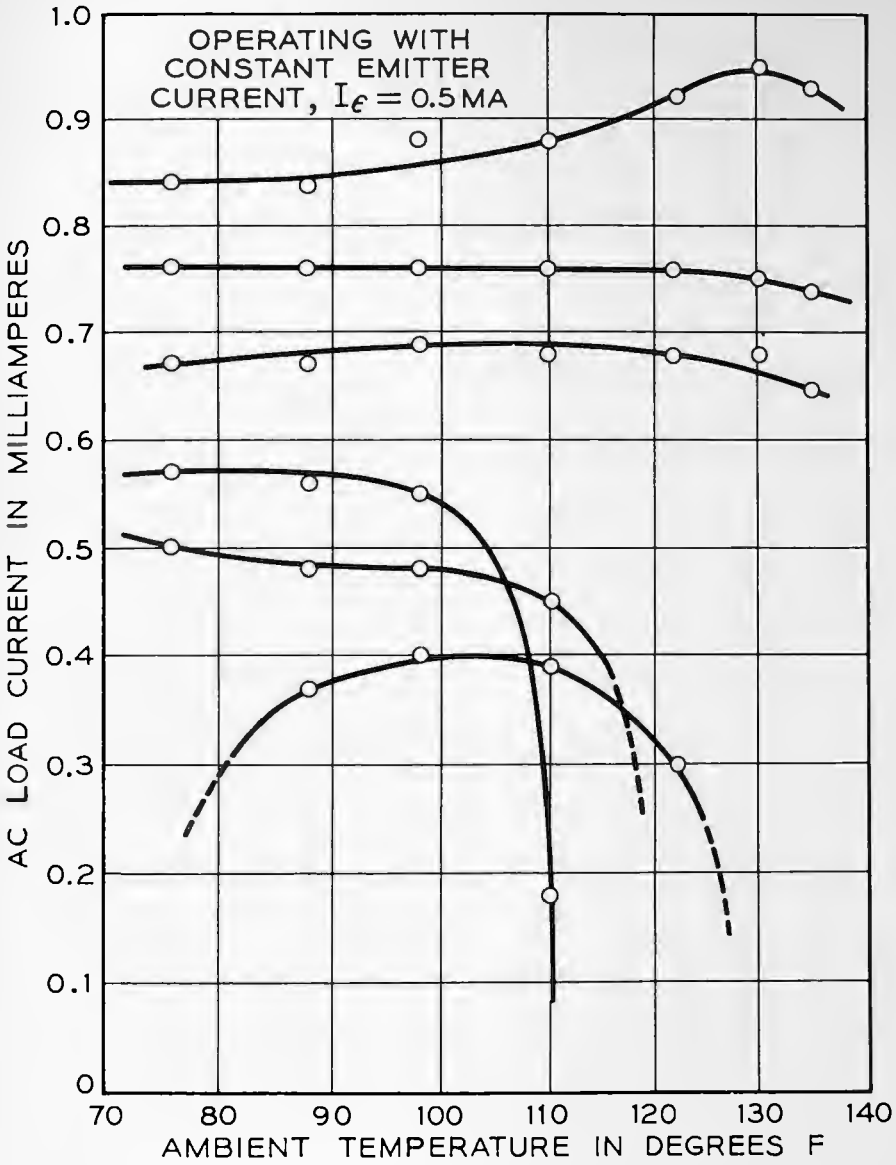


FIG. 8 EFFECT OF TEMPERATURE ON VARIOUS TRANSISTORS IN OSCILLATOR IN FIGURE 6.

MODULATORS IN CARRIER TELEPHONE SYSTEMS

R. S. CARUTHERS

Produced outside the work performed under this contract

MODULATORS IN CARRIER TELEPHONE SYSTEMS

Modulators are used in frequency division arrangements in the Bell System on either a single channel or group basis. As channel modulators they modulate or demodulate between the speech band frequencies and carrier frequencies seldom higher than a few hundred kilocycles. Most transmission is on a single sideband basis, and the carrier frequency modulated to, is set either by limitations of usable line frequencies or by desired bands where it is most practical to construct the single sideband separating band filters.

In nearly all present day systems of wide use, channels are first translated to a primary group allocation which may consist of 4 to 12 channels. These groups are then modulated to the line allocation desired, or they may be further modulated into super-group allocations before transmission as for example in the transcontinental microwave radio relay system. In the coaxial cable system as many as 10 of these super groups are sometimes used for transmission of as many as 600 channels.

In certain applications, group modulators are used at every repeater point. This serves to prevent sing and crosstalk coupling through use of different frequency bands at repeater input and output. Also in newer wire line systems, through inversion of the frequency allocations in the repeater modulators channels alternately traverse high and low frequency positions in successive line sections and thus encounter nearly constant attenuation. Although experimentally we have operated satis-

factorily through hundreds of such modulating repeaters in tandem, we do not expect such severe conditions in practical experience.

Low carrier leak and low interchannel modulation interference have been important factors in carrier telephone modulator design since the early 1920's. These conditions are reached through balanced bridge arrangements using high level carrier and low level signals. As a result a single sideband allocation is derived which can be sent through line repeaters at greatest signal/noise advantage. Fig. 1 shows a vacuum tube bridge arrangement used in all carrier telephone modulators until the late 1930's. Although the carrier was balanced from the output, frequent manual balancing was required using condensers to compensate for variations imposed by aging of the vacuum tubes. Such modulators were classed as single balanced, the prime objective being to balance the carrier from the output, even though it would have eased suppression requirements on associated output filters to balance the input signal and certain classes of unwanted modulation products from the output. Double balanced vacuum arrangements also shown in Fig. 1 were contemplated in the early 1930's to get further reduction in unwanted products. Both carrier and the input signal are balanced from the output and if we examine the circuit closely it is apparent that we are so poling tubes and transformers as to get a double pole double throwing reversing switch connection at the carrier rate

between the input and output signal branches. Neither the single or double balanced tube arrangement was ever used in a carrier telephone group modulator. This was because of the general instability in suppression of unwanted modulation bands, the circuit complexity and the greatly improved performance obtained with copper oxide varistor diode modulators beginning in the mid 30's.

In Fig. 2 are shown early versions of double balanced copper oxide varistor modulators. The bottom circuit shows the advantages of varistors in simplifying circuitry, as only two transformers need be used, in contrast to six transformers and four tubes in the vacuum tube arrangement.

The clarity with which the double balanced varistor arrangements showed that the problem was to get the reversing switch connection between input and output, contributed to devising a number of new arrangements. Third order arrangements where the connections were reversed at double the carrier rate were considered attractive. Desired output sidebands were utilized on the 2nd harmonic of the carrier frequency. With closely matched varistor diodes oppositely poled or with thyrite, carrier unbalance leak at the fundamental carrier frequency could easily be eliminated because of the wide frequency separation. In Fig. 3 a third order double balanced arranging is shown using thyrite. This was more speculative than useful. Three dimensional drawings

of the circuit aided in visualizing where the modulation product currents flowed.

In Fig. 4 series, shunt and lattice type varistor modulator circuits are shown. The lattice type is also called a double balanced modulator or a ring modulator. Most present day modulators are one of these types. Considerable stress has been placed on minimizing the number of transformers. They are completely eliminated in the series and shunt circuits. In the ring modulator a useful application is to take off dc or very low frequency currents through the normal carrier branch of the circuit, which does not include a transformer with two nearly equal input frequencies in the two signal branches.

In the circuits of Fig. 4, development has gone about as far as can be expected in the direction of circuit simplicity. Refinements in the direction of better reduction of modulation products is obtained through close selection of varistor diodes in the bridge arms. In some cases as many as 64 discs have been stacked in the bridge and held better than 50 db balances without adjustment in manufacture. Benefits from balance can, however, can be obtained only in sorting the modulation products into different classes. In the single balanced series and shunt modulators with only two circuit branches, all modulation products are divided into one branch or the other. In the signal branch, fundamental and odd harmonics of the signal frequencies are combined with fundamental and both even and odd harmonics of the carrier frequency.

In the carrier branch the even harmonics of the signal are combined with even and odd harmonics of the carrier. In the ring modulator, the four combinations of odd and even harmonics of signal and carrier are separated into four circuit branches: signal input, signal output, carrier, and varistor ring. Fig. 5 portrays these four classes of modulation products.

Until a few years ago, the only balanced modulators in telephone circuits used copper oxide varistor diodes. Recently an almost complete shift has taken place to use of germanium diodes in new circuit design. This can be attributed to finding in germanium a more stable, less temperature sensitive, less frequency conscious diode than copper oxide. In certain speech modulators modulated by carriers at syllabic frequencies, very close control and uniformity of the diode e-i characteristic has been required. In germanium the characteristics obtained in this use have shown extremely small change over a period of two years of operation. Consistently any change observed has been of the orders of magnitude that might have been expected from the few degrees of temperature variation accompanying the measurements.

It is not always preferable to use germanium, however, as the noise performance is at least 10 db poorer than copper oxide. Another factor favoring copper oxide, too, is the ability to stack large numbers of copper oxide diode varistors in the bridge arms for better balancing. The mechanical construction of the germanium varistor has not permitted such arrangements.

In examining transistor application to the circuits that have been under discussion, there is little advantage except for the gain of the transistors. The ultimate in modulator simplicity with varistor modulators has perhaps been reached. Transistor modulators tend toward the complexity of the vacuum tube modulator, requiring transformers for efficient use. Under these circumstances, it is hard to imagine even further complications through double balanced operation or through use of many transistors in the bridge arms to get closely matched characteristics and large reduction in undesired modulation. Progress in varistor modulators has advanced far beyond vacuum tubes and in using transistors only slight regression can be tolerated, to derive gain benefits. Under these circumstances it seems preferable to continue with varistor modulators and where gain is required, to follow the modulator with a transistor amplifier.

Fig. 6 shows how well varistor modulators have been performing in the Bell System. When first put in, a circuit element was hoped for which, once installed, would require no more attention than a soldered in transformer or resistance. Replacement, due to aging, was contemplated within five years. In the chart, a carrier balance distribution is shown for a large number of units installed in 1938. More than ten years later these same units show the average unit with bridge balance better than 40 db, and almost no change in the distribution curve. With transistors in our circuits whether in modulators, in oscillators, or in amplifiers, the same permanence in performance can be expected.

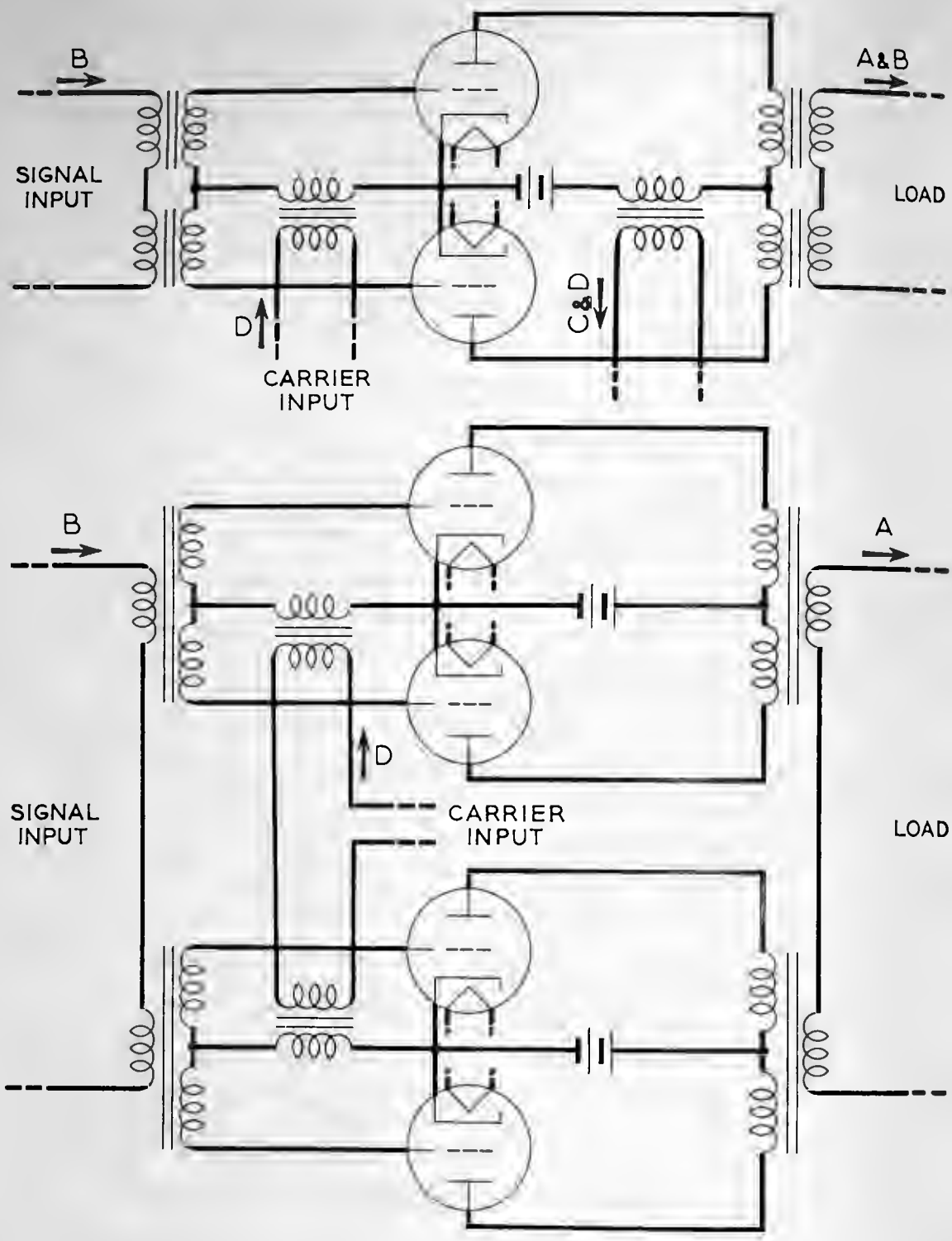


FIG. 1 EARLY SINGLE AND DOUBLE BALANCED VACUUM TUBE MODULATORS

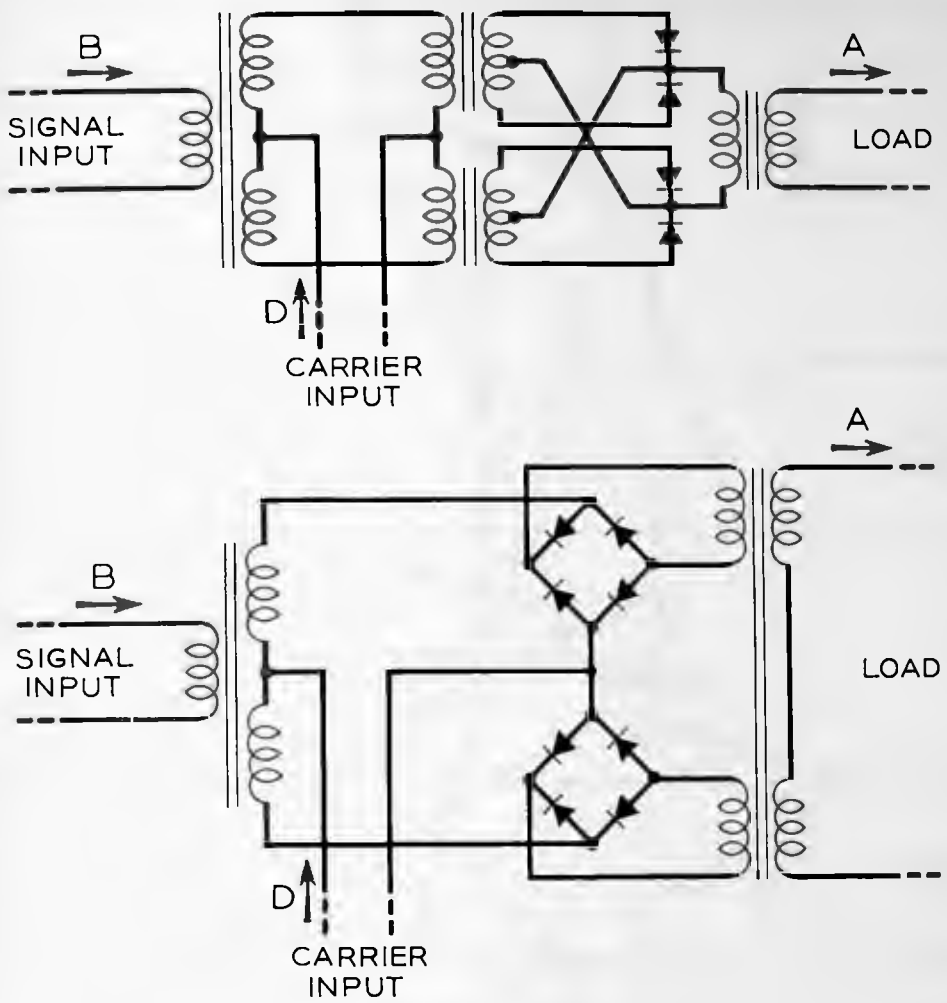


FIG. 2 EARLY DOUBLE BALANCED MODULATORS USING COPPER OXIDE VARISTORS.

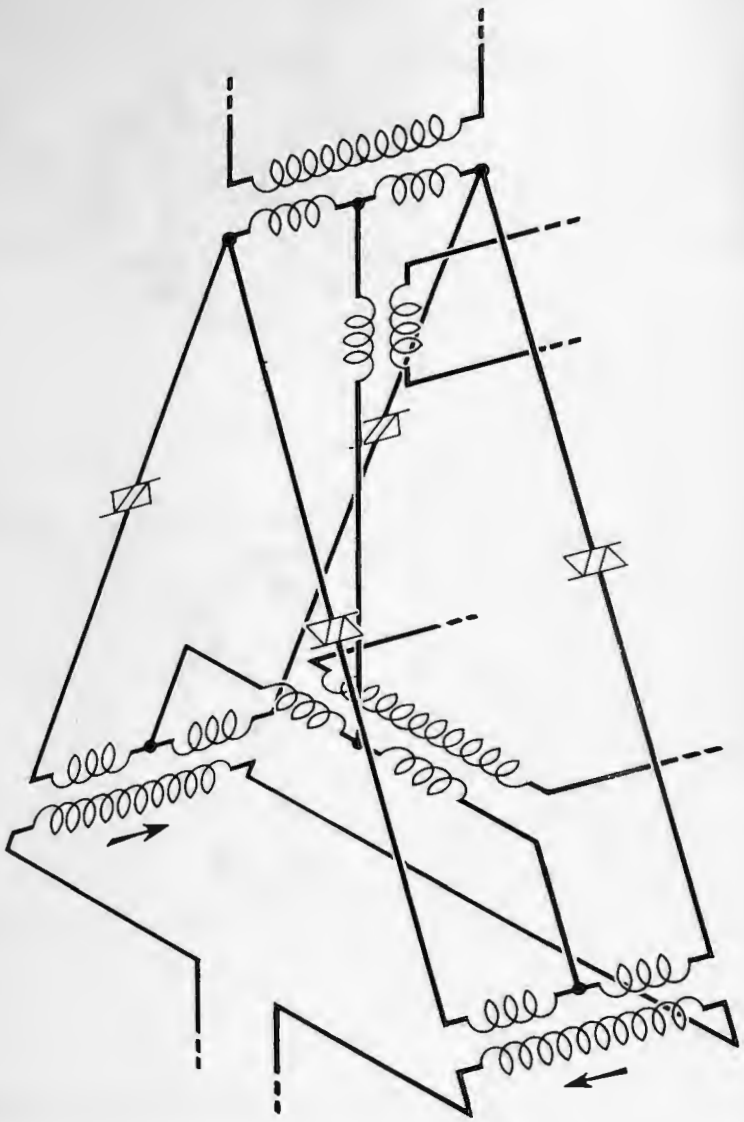


FIG. 3 THIRD ORDER DOUBLE BALANCED MODULATOR USING SYMMETRICAL VARISTORS

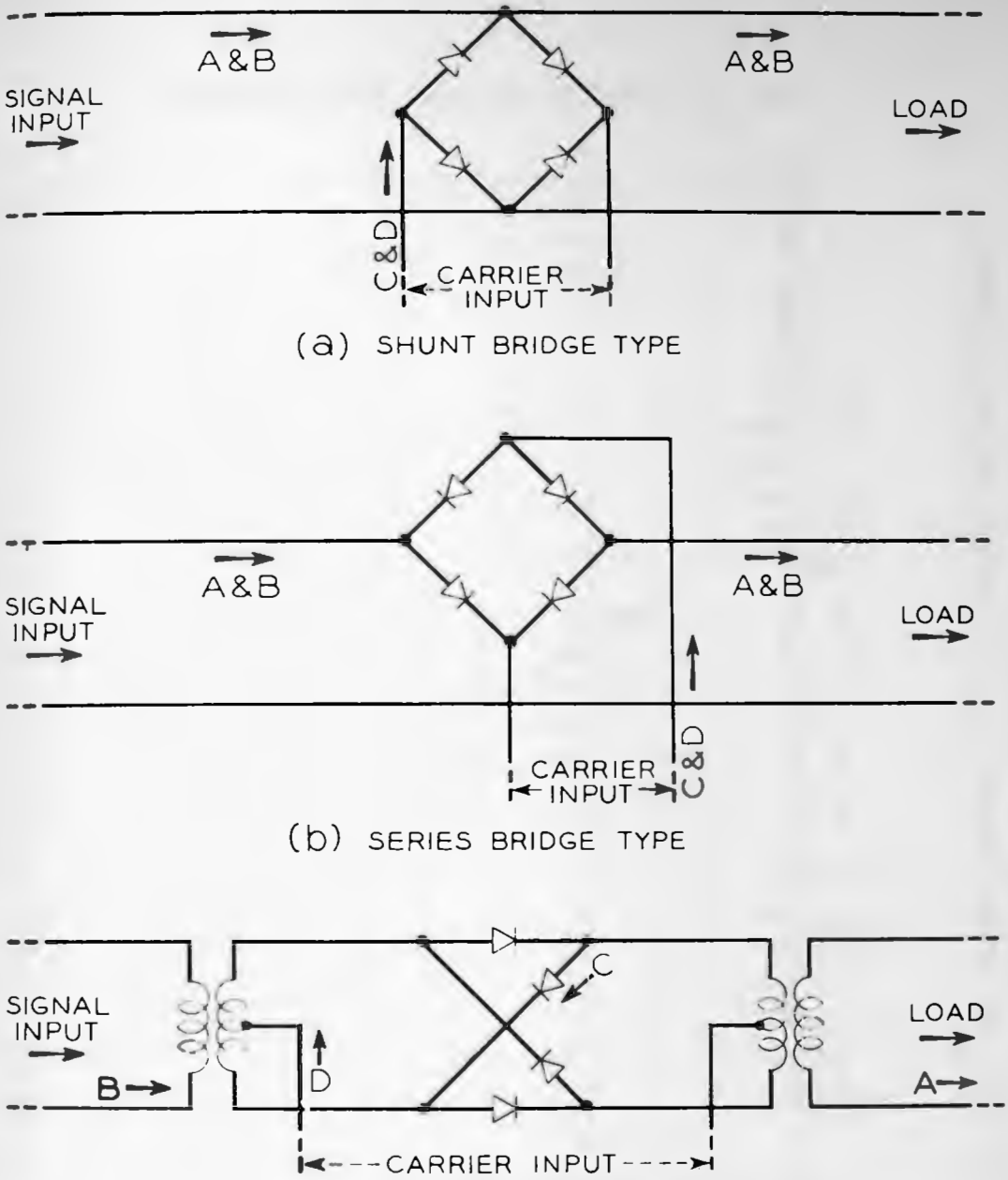
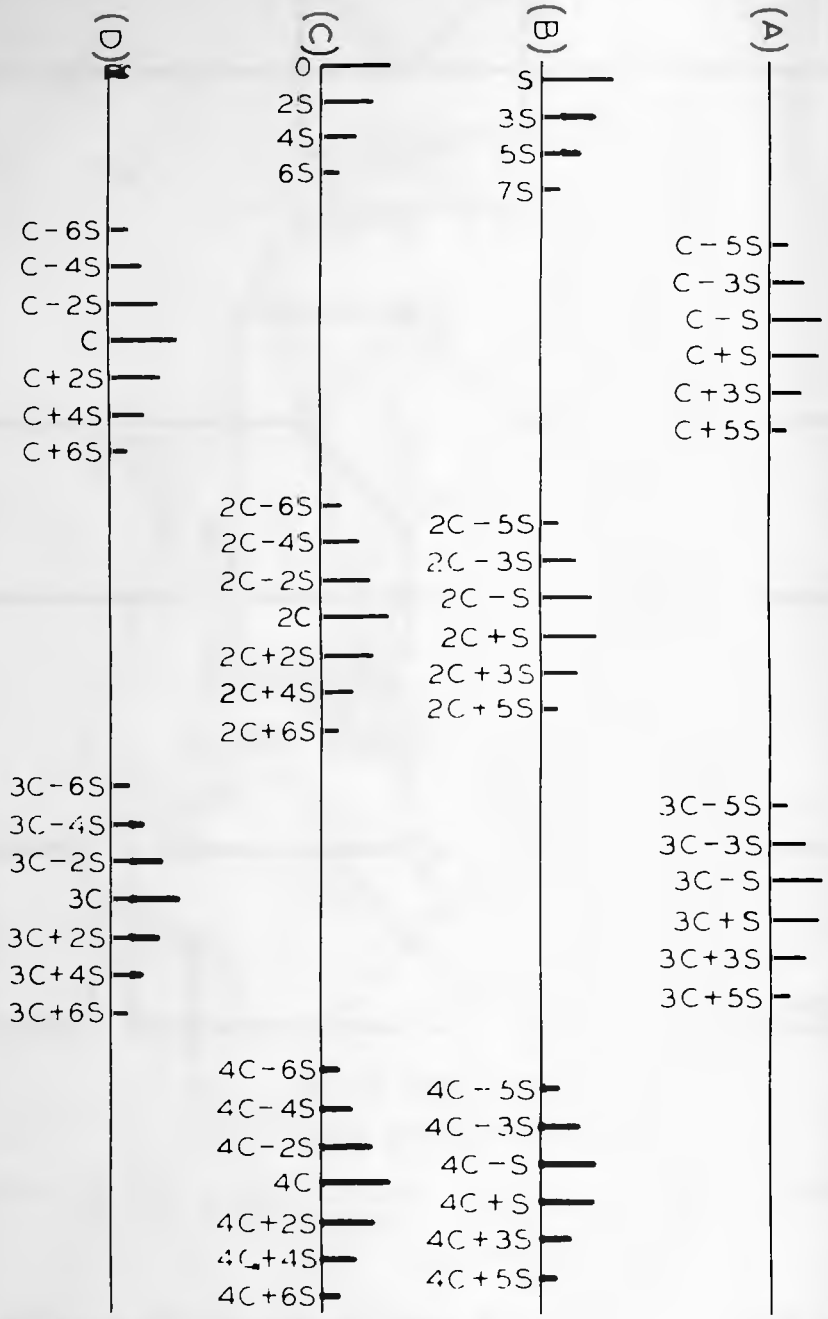


FIG. 4 - VARISTOR MODULATOR USED IN PRESENT DAY CARRIER TELEPHONE CIRCUITS

FIG. 5 - SPECTRA OF MODULATION PRODUCTS



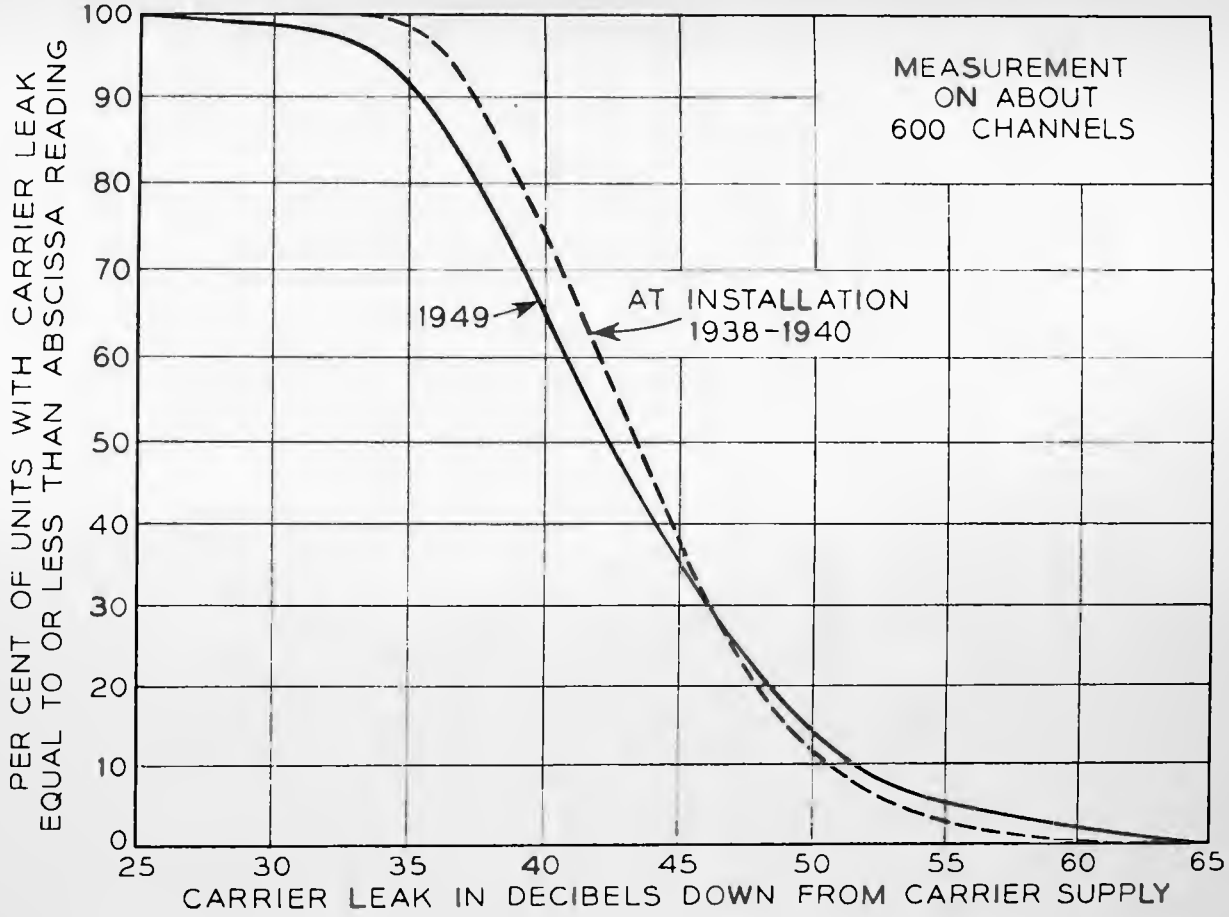


FIG. 6 - TEN YEARS EXPERIENCE IN STABILITY OF COPPER OXIDE MODULATORS IN BELL SYSTEM USAGE

A STABILIZED TRANSISTOR DELAY AND SWITCHING CIRCUIT

A. E. ANDERSON

Produced under contract W 36-039 SC-44497

A STABILIZED TRANSISTOR DELAY AND SWITCHING CIRCUIT

In most switching systems small delays, usually in the trigger or control circuit, are often necessary. The sense is not that of information storage, but that of delaying a trigger action until other actions can take place with certainty. Thus there is no requirement on linearity or even upon amplitude beyond a minimum limit. The single-shot or monostable multivibrator has been, in the past, the work horse for this application. The device sits in normal quiescent condition and is triggered off by the main trigger to execute one cycle - a sensibly rectangular pulse. The delayed trigger is derived from the trailing edge of this pulse.

It was though that a very simple transistor circuit could be used as a monostable relaxation oscillator in which the timing is by an R, C and device combination. Unfortunately, a large number of the delays are required to recycle very rapidly. That is to say that the duty cycle is 0.5 as contrasted to the normally expected duty cycle of perhaps 0.005 or 0.01 and seems to be, therefore, impossible for the simple relaxation oscillator which has a fairly long recovery period. This does not mean that it is impossible to use a transistor package for these rapid cycling delays; rather, a dual device sort of arrangement must be used wherein one unit is charging while the other is discharging. The simple relaxation oscillator can be used in other places where the recycle time is not so critical, in counters and, with slight modification, as a stabilized free-running oscillator which is also needed.

Fig. 1 shows the basic two-terminal relaxation oscillator which uses the negative resistance characteristic of the emitter to ground. R_b is an added base resistor to augment the feedback and is of the order of 10K. R_c is a load resistor in the collector circuit of 500 to 4K in value (the collector is used for output in order to obtain a flat base line and some degree of saturation). C is the storage or timing capacitor.

Fig. 2 represents the emitter characteristic which is employed. Certain parts of the characteristic have been exaggerated for clarity. R is the load line and the bias V_{ee} has been chosen to permit a singular intersection only, resulting in the monostable behavior. The application of a trigger to base or emitter momentarily will raise the load line over the trigger point A with the resultant completion of a single cycle.

The necessary trigger voltage is given by the difference between the bias, V_{ee} and the voltage of A. The voltage at A is very nearly the voltage developed across R_b due to the collector current when the emitter is open circuited. This is not quite true since there is internal r_b and A actually lines minutely to the right of the axis in the positive emitter current region, but is close enough except for extreme accuracy since $R_b >> r_b$ and $I_{eA} \approx 0$.

It follows very simply that for maximum sensitivity $V_e - R_b I_{co}$ should be as small as possible. However, V_{ee} is fixed and I_{co} is variable from unit to unit and with temperature having a range of perhaps 10:1. Thus if one were to choose V_{ee} small, the circuit might self-trigger under high I_{co} condition and lock-up in the high current region. Conversely, choosing V_{ee} so large as to completely negate this condition results in very low sensitivity. In either case there is a high degree of variability.

The proposed circuit of Figure 3 effectively ties the trigger voltage $R_b I_{co}$ and the quiescent point together so that as one varies, the other, to a large extent, moves with it so as to sensibly retain a constant small difference. This is accomplished by the cross resistor R_{eb} which holds the capacitor voltage to the base voltage, $R_b I_{co}$. Looking at the portion of the circuit within the box, Fig. 4 it is quite obvious that if R_{eb} is large enough that this portion of the circuit is dc stable and in the absence of any voltage perturbation the entire circuit is stable. If the voltage across C is varied, however, C suddenly appears to be a low impedance, approaching zero, and is quite small in comparison to R_{eb} so that R_{eb} effectively disappears from the picture and the circuit is thrown into the dynamic state. Such a circuit has been made so sensitive as to fire with random noise and pickup static.

Large values of R_{eb} are usually not too satisfactory due to the long recovery time and some variability in the trigger voltage; low values of R_{eb} may result in instability with consequent free running. In this latter case the circuit may be made monostable by the insertion of a battery bias in series with R_{eb}, suitably poled. More convenient, however, is the constant current bias afforded by the high valued shunt R_e in Fig. 5.

The interpretation and calculation of the quiescent or dc condition may be done graphically, but more easily with the equivalent circuit, Fig. 6. The assumptions are that I_e must be +50 uA before alpha exceeds unity (the trigger point A), that I_c is much greater than I_e, and R_e is sufficiently large as to afford constant current. r_e is an unknown and although we know that it has values of 300 ohms or so at positive currents, the region in which we are interested $I_e \rightarrow 0$ and therefore r_e

may have fairly high values. For first approximations consider r_e to be about 10K. If r_e is large the deviation will be in the direction of increased stability.

Knowing V_c, R_b, R_c and the range of r_c one may calculate I_{co} and its range. Since $I_c \gg I_e$,

$$I_e \approx \frac{I_{co}\, r_b}{R_{eb(+r_e)}} \quad \ldots\ldots \qquad (1)$$

If this value of I_e is less than the critical value of I_e for $\alpha > 1.0$, the circuit is stable and the trigger current is given by the difference. Experimentally, critical I_e is of the order of 50 ua and I_e from equation (1) may be of the order of +30 or 40 uA depending on R_{eb} and r_b. If now R_e is introduced and a constant current bias afforded, I_e from (1) will be reduced about one-half the bias current. The other half of the bias current flows directly through R_{eb} and R_b. Since the bias current is small compared to I_{co}, the deviation to $R_b I_{co}$ is small.

The net effect is a stable circuit in which variations are largely controlled by passive parameters. Variations which do exist and even though they may be percentage-wise large are small in terms of absolute values.

Since R_{eb} will in general be much smaller than the average emitter back resistance, variations in the recovery time or repetition rate are also considerably reduced. There is the disadvantage that the introduction of R_{eb} results in smaller entrance impedance making the circuit harder to drive; the smaller trigger possible with the circuit over weighs this disadvantage, however. Too small values of R_{eb} can result in poor dynamic properties.

The repetition rate can be speeded up by replacing R_{eb} with a diode as in Fig. 7. As long as the emitter is more negative than the base the diode is a near short circuit, but as soon as the emitter becomes more positive, as with trigger, the diode unlatches and drops out of the circuit.

The circuit may also be made into a stabilized free-running circuit by returning R_e to a positive voltage so as to afford a constant positive current bias.

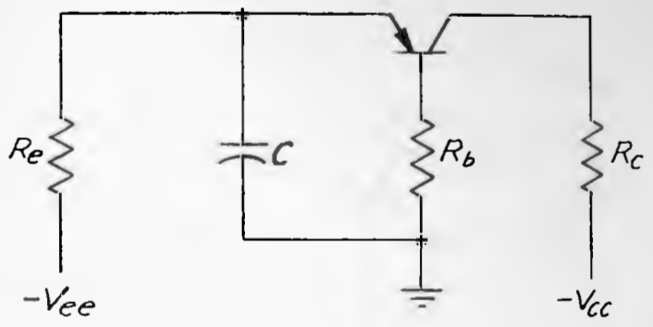


FIG. 1
SIMPLE TRANSISTOR
RELAXATION OSCILLATOR

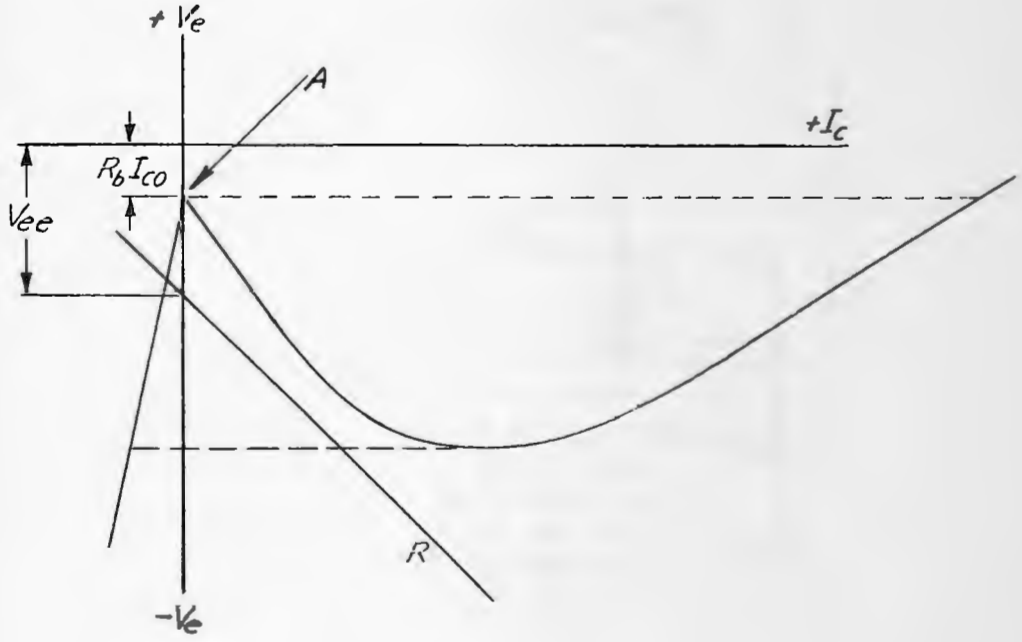


FIG. 2
EMITTER CHARACTERISTIC
WITH BASE FEEDBACK

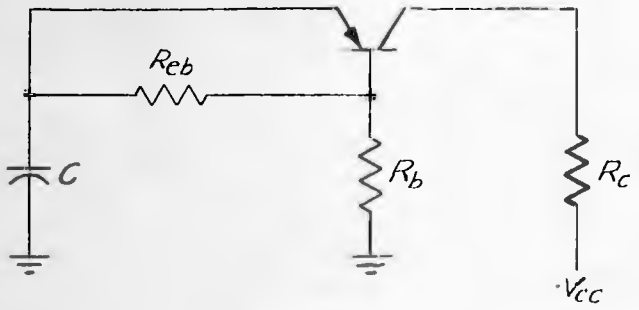


FIG. 3
STABILIZED OSCILLATOR
CIRCUIT

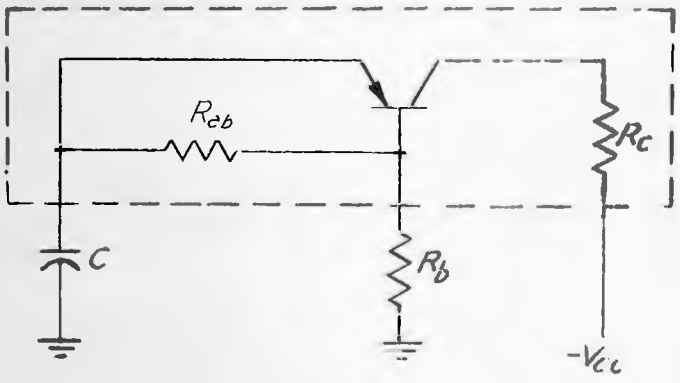


FIG. 4

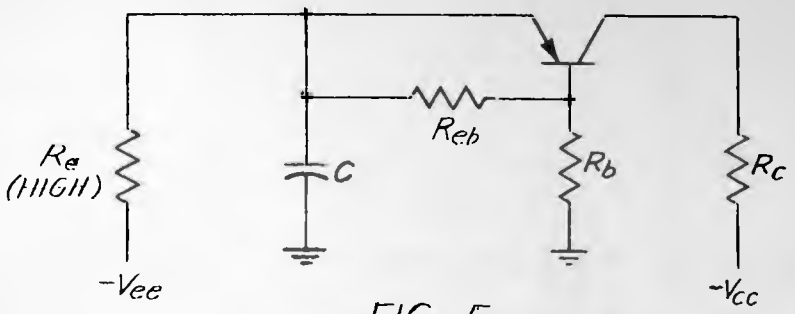
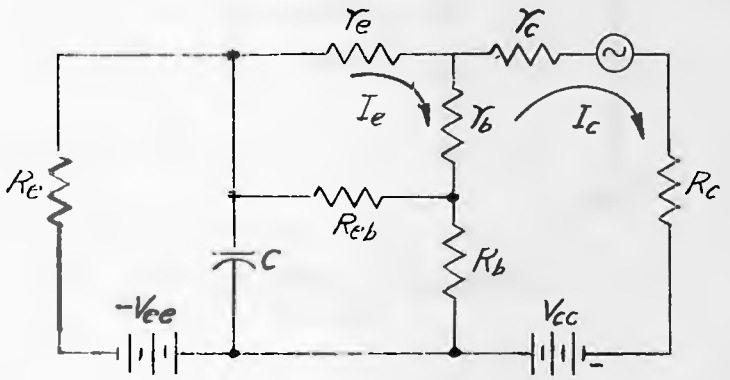


FIG. 5
METHOD OF BIAS



$\alpha > 1.0$ AT $I_e > 50\mu A$
$r_c \approx 20K$
$R_b \approx 10K >> r_b$
$R_e \rightarrow \infty$
I_c & I_c ARE DC, $I_c >> I_e$

FIG. 6
EQUIVALENT CIRCUIT

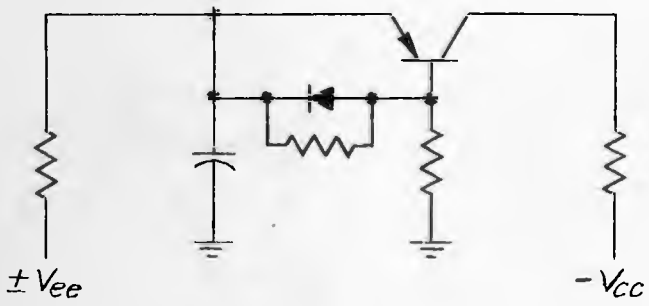


FIG. 7
STABILIZED RAPID RECYCLE
CIRCUIT

A TRANSISTOR PACKAGED CIRCUIT FOR MULTIFUNCTIONAL SWITCHING APPLICATIONS

R. L. TRENT

Produced under contract W 36-039 sc-44497

A TRANSISTOR PACKAGED CIRCUIT FOR MULTIFUNCTIONAL SWITCHING APPLICATIONS

In the design of switching systems, there are many applications for a threshold two gate, this is to say, a circuit which requires the simultaneous presence of two input signals to produce an output. When dealing with digital information or control signals with pulse techniques, the output often need not be a replica of any input. Further, gain or level restoration may also be desired. Other practical considerations often call for a device which has high input impedance and which is capable of being d-c coupled if necessary.

Past practice is exemplified by diode configurations involving transmission loss and having impedance matching requirements at input and output*. Such configurations are to a large degree independent of signal amplitudes after minimum levels have been exceeded. Various other diode configurations have also been developed which may preserve the waveform of the signal information, if desired.

In the considerations for the design of a threshold two gate employing a transistor, it was deemed feasible to utilize the negative resistance characteristics of a triggered regenerative amplifier. These characteristics have been investigated and successfully employed in the design of a threshold two gate. Further investigation of the basic circuit evolved has brought out the possibility of application of the basic circuit as a packaged device capable of performing additional useful switching functions.

1.0 General

The circuit to be described consists essentially of a regenerative amplifier, employing a transistor as an active element. Additional passive circuit elements and biasing voltages may be applied to this basic circuit to permit multifunctional usage. Some of the functions which may be performed are:

- - - -

1. Threshold two gating

   a. AND gating with two signals.

   b. BUT NOT gating with two coincident signals, one positive and one negative.

2. Pulse regeneration and delay

1.10 Basic Circuit Configuration

The elementary regenerative amplifier circuit configuration is illustrated in Figure 1, and a typical input (emitter) characteristic is shown on Figure 2. The circuit incorporates base stabilization** in order to clamp the turning point of the characteristic at a definite potential, in this case zero volts. In addition, a small resistance has been added in the base circuit to insure obtaining a negative resistance as soon as the emitter current becomes positive. This is to obtain ease of triggering at a small cost in stability of the turning point with transistor variations. The circuit constants as illustrated have been so adjusted that variations in the I_{CO} of the transistor from approximately 3 ma down to approximately 0.8 ma may be tolerated. A diode is inserted in series with the emitter to insure a slope in the negative emitter current region greater than 0.5 megohm for additional discrimination.

2.0 Applications of Basic Characteristic

2.10 Threshold Two Gating

2.11 AND Gating - Positive Pulse and Voltage Step Function

Figure 3 illustrates a circuit, using the basic regenerative amplifier, which is capable of performing the function of AND gating using a step voltage and a coincident positive pulse. The characteristic for this configuration is shown on Figure 4. An emitter load line, biased at -10 volts, is superimposed upon the input characteristic of the device, resulting in a d-c equilibrium point at A. We may assume that the pulse to be gated has an amplitude of 15 ±5 volts. If such a pulse be impressed at the input terminal, the circuit will be triggered and a positive pulse in the

- -

**"Idealized Negative Resistance Characteristics of the Transistor", by R. L. Trent.

order of 30 volts, 2 to 5 microseconds long, will appear at the output. If the biasing potential is changed to -30 volts, however, the load line will intersect the characteristic at point B and the input pulse will be incapable of triggering the circuit. Obviously the margins of operation depend primarily upon the amplitude variations of the trigger pulse, the value of emitter load resistance, and the slope of the characteristic. The circuit is characterized by a high input impedance to the triggering pulse source when in the OFF condition, determined by the emitter load resistance shunted by the back resistance of the series diode and the emitter back resistance. However, as soon as the positive emitter current region of the characteristic is entered, due to a triggering pulse, the pulse source sees a much lower input impedance determined by R_B, R_L and transistor characteristics. Therefore the impedance of the triggering pulse generator may be expected to affect the rise time and duration of the pulse, but is fairly noncritical up to a practical limitation in the order of 10,000 to 20,000 ohms, for the circuit constants as shown. The biasing potentials of -10 and -30 volts used in the discussion are for illustrative purposes only, but represent to a fair degree the potentials present at the collector of a transistor bistable flip-flop when in ON and OFF conditions, respectively.

The advantages of this gated amplifier circuit over a simple diode switch are as follows:

a. The source of emitter biasing potential always sees a very high impedance of at least the emitter load resistance value, for the illustrated plot, 0.1 megohm. During the intervals when no triggering pulse is impressed upon the input terminal, the source sees approximately 0.7 megohm. This permits many of these gated amplifiers to be operated in parallel by the same low impedance controlling potential source.

b. The pulse source also sees a high impedance except during a triggering interval, at which time the source sees approximately 3000 to 5000 ohms, depending upon the gate circuit constants and load impedance.

c. Pulse regeneration occurs, resulting in a positive pulse with a very steep leading edge. The duration and trail-off may be adjusted, within limits, by proper choice of circuit constants and triggering pulse. Output pulses of 2 to 5 microseconds duration, with amplitudes of 30 to 40 volts have been obtained

from this circuit. Inherently, the trailing edge of the pulse is much less steep than the leading edge, but with a properly chosen differentiating circuit in the output, a delayed negative pulse of about 6 to 10 volts amplitude may be obtained.

A discussion of the effects of experimental circuit parameter variations on the output pulse waveform is contained in Appendix A. Analytical relationships have not as yet been completed.

Several interesting circuit variations may be shown, in order to illustrate the flexibility of circuit package application to gating functions. These variations are discussed in Appendix B.

2.12 AND Gating - Two Coincident Positive Pulses

The circuit configuration illustrated on Figure 5 will perform the function of AND gating using two coincident positive pulses. The emitter load potentiometer with resistive arms R_1 and R_2 is connected to a source of negative biasing potential V, establishing a d-c stable equilibrium point at C, as shown on Figure 6, in the OFF negative emitter current region.

Input pulse amplitude limitations for use in this circuit are fairly stringent, since they must be so controlled that one of the pulses appearing at either A or B is not sufficient to drive the amplifier into the triggered stage, but that the summed effect of the two input pulses is sufficient to do so. Assuming equal pulse amplitudes at A and B, the following elementary requirements may be formulated:

$$V\left[\frac{r_{11}}{r_{11} + R_1 + R_2}\right] > A = B > \frac{V}{2}\left[\frac{r_{11}}{r_{11} + R_1}\right] \qquad (1)$$

where r_{11} is the slope of the emitter characteristic, including series diode, in the negative emitter current region. Assuming values as shown in Figure 5, the amplitude limits for pulses at A and B are:

$$17 > A = B > 9 \text{ volts peak approx.}$$

A diode pulse limiting circuit may be interposed in each input circuit to insure that the maximum pulse amplitudes are not exceeded.

Then if a pulse is impressed at either input A or input B alone, no output pulse will be obtained. If a pulse is impressed at input B, the load line is rapidly displaced along the vertical axis to achieve intersection with the characteristic at point D. The slope of the load line is halved if we assume that R_1 and R_2 are equal. If, simultaneously, a pulse is impressed at input A, the amplifier will be triggered and a positive pulse will be obtained at the output.

2.13 BUT NOT Gating - Two Coincident Pulses, One Negative and One Positive

The system function of BUT NOT pulse gating may be accomplished by a gate having two pulse inputs and a pulse output so arranged that if positive input A and negative input B are impressed simultaneously, no pulse appears at the output. However, when input A but not input B is impressed, a pulse output will be obtained. Also, input B alone produces no output. The circuit arrangement shown on Figure 5 will perform this function, as well as AND pulse gating. Consider the characteristic shown on Figure 7. The entire emitter load potentiometer is connected to a small negative biasing potential V. This establishes a d-c stable equilibrium point at C. With no pulse impressed at B, a positive pulse impressed at A of amplitude greater than "a" will trigger the circuit as previously described. However, if a negative pulse appears at input B simultaneously, causing the load line to be rapidly displaced along the vertical axis, and the equilibrium point to be at point D instead of point C, then if the positive pulse input at A is less than the amplitude "b" required for triggering, no pulse output will be obtained. If we assume that the source impedance of the B input is low with respect to R_2, and that R_1 and R_2 are equal, the slope of the load resistance at point D will be half that of point C.

From the above considerations, it is also evident that the input signal at B may equally well take the form of a negative step function. The form of coupling utilized depends upon the source impedance of the input step function.

3.0 Pulse Regeneration and Delay

The basic circuit may be modified in various ways to obtain pulse regeneration depending upon desired output waveshape and triggering pulse polarity.

3.10 Triggering with Positive Pulses

3.11 The circuit arrangement shown in Figure 4 may be utilized to obtain simple positive pulse regeneration by considering the bias potential V as fixed. For a given fixed bias potential a single input pulse of sufficient amplitude to trigger the device will result in a single output pulse, as discussed previously. Appendix A covers the effects of circuit parameter variations on pulse output waveform and duration. Delay may be achieved as previously described, by differentiation of the output to obtain a negative pulse delayed from the initiating triggering pulse by the duration of the output pulse.

3.12 The circuit arrangement shown in Figure 8 may be used to obtain pulse regeneration. This circuit has the advantage of high input impedance, and greater stability from voltage and transistor variations. In addition, the triggering sensitivity is very great, and it may be adjusted within limits by proper proportioning of the bias resistor R_2 and the current flowing through it during the OFF condition. These features make the circuit capable of being triggered from a high impedance trigger source, if desired.

3.20 Triggering with Negative Pulses

The circuit configuration and characteristic shown on Figure 9 may be utilized for pulse regeneration when it is desired to regenerate a negative pulse. A characteristic output pulse shape as shown illustrates that the leading edge of the negative pulse is poor but that the trailing edge is very steep, indicating that differentiation of the output would result in a delayed positive pulse of high amplitude. Considering the circuit, a fixed positive bias potential is applied to the emitter load resistance of such magnitude as to insure intersection of the load line with the emitter characteristic in the saturation or ON region. Then a negative pulse applied to the emitter from a source impedance Z will cause the unit to snap to the OFF condition. Since there is no d-c stable intersection of emitter load line with the characteristic in the negative emitter current region, the unit will return to the ON condition after a delay period. There are many complex factors affecting this delay period, such as source impedance, pulse amplitude and duration, and emitter load resistance, besides the usual transistor parameters. However, for a specific application it is usually possible to arrive at a suitable compromise among these factors to arrive at a delay period of from 2 to 5 microseconds. A typical output waveform is illustrated together with circuit constants.

4.0 Conclusions

A transistor gated amplifier circuit incorporating AND type switching characteristics and pulse regeneration may be designed suitable for switching system application. Modifications may be made to the basic circuit to obtain multifunctional usage, such as pulse delay, pulse regeneration and BUT NOT gating involving positive and negative coincidental pulses.

Appendix A

The purpose of this appendix is to describe the effects of additional or modified circuit elements, and variations in triggering pulse amplitude and duration, upon the output waveform of a gated amplifier circuit.

Since the d-c parameters are usually dictated by other factors, such as available voltages, allowable dissipations and loading conditions, and transistor variations in I_{CO} which the circuit must override, we may assume that the values of the resistors have been so chosen that these requirements have been met. Therefore, this discussion will be limited to the effects of those components which affect the a-c response of the circuit.

Considering the specific circuit configuration shown in Figure 10, an attempt will be made to illustrate the effects on the output waveform of triggering pulse amplitude and duration, and the effects of varying the absolute values of capacitors C_1, C_2 and C_3.

Effect of Triggering Pulse Amplitude and Duration

Figure 11 illustrates typical output waveforms under various conditions of triggering pulse amplitude and duration. A summary of these effects is as follows:

(1) For input pulses of short duration (0.2 microsecond) the major effect of increasing the amplitude is sharpening the leading edge. The trailing edge shape is unchanged, and the duration of the output pulse is substantially the same.

(2) For input pulses of duration longer than the ON condition of the curve (flat top position) the major effect of increasing the amplitude is sharpening the wavefront as before. A secondary point is the effect on the duration of the pulse. After a certain minimum triggering voltage is exceeded, the duration of the output pulse is roughly proportional to the input pulse amplitude.

Effect of Capacitor C_1

(1) As illustrated in Figure 12, after a certain minimum value of capacitor C_1 is exceeded, the duration of the ON condition of the curve (flat-topped portion) is

proportional to the size of the capacitor C_1. Effects of triggering pulse amplitude and duration are the same as previously described.

(2) The minimum value of triggering pulse amplitude is decreased for increased values of C_1, since a larger proportion of the triggering pulse appears between input terminal of the device and ground due to the voltage divider action between C_1 and stray capacitance to ground from the input terminal. (Shown as C_2)

Effect of Capacitor C_2

(1) As shown in Figure 13, increasing the capacitance to ground steepens the leading edge and introduces an oscillatory condition at the resonance frequency of the R-C combination at the input terminal of the device. Heavy damping of the resonance circuit is inherent due to the shorting effects of the lowered emitter circuit impedance, however, and a small additional capacitance of the order of 60 to 100 microfarads may be used to advantage if differentiation of the output is to be employed to obtain a positive output pulse.

(2) With low amplitude trigger pulses, a small delay of the order of 0.2-0.3 microsecond is introduced in the wavefront. This is due to the fact that the additional input capacitance to ground changes the ratio of the capacitance potentiometer formed by C_1 and C_2, and a smaller fraction of the pulse voltage appears at the input terminal of the device for triggering. Therefore lower sensitivity to triggering pulses results.

Effect of Capacitor C_3

(1) As shown in Figure 14, the major effect of increasing the series output capacitor C_3 is to alter the trailing edge of the wave by increasing the time constant of the collector circuit.

(2) The adverse loading effects of a larger output capacitor result in decreased peak amplitudes. However, this degradation of the waveshape in the trailing edge may be advantageous if differentiation of the pulse to obtain a positive pip with no following negative pip is desired for triggering a following stage.

Summary

The output wave from a transistor gated amplifier may be modified by changes in circuit elements to obtain various forms as desired. Differentiation may be applied to obtain a delayed negative pulse: if such a delayed pulse may adversely affect following circuits, the amplitude of the negative pulse may be minimized by proper choice of constants.

Appendix B

The purpose of this appendix is to illustrate several variations in circuital applicability in order to point out some of the possibilities of multi-functional usage of the basic circuit.

(1) Because of the clamping action of the base stabilization feature, it is possible to displace the characteristic so as to obtain a gated-amplifier configuration between any two voltage steps on the vertical axis, provided the usual margins between the voltage steps and triggering pulse amplitude are maintained. For instance, a very simple displacement results from modifying the applied potentials as shown in Figure 15. This results in a characteristic as shown on Figure 16. Gating may then be accomplished using a voltage step function between -40 and -60 volts.

(2) It is also evident that greatly increased triggering sensitivity may be achieved either by adjustment of circuit constants or by displacement of the characteristic so that it is possible to achieve larger ratios of output to input pulse amplitudes. For instance, in the circuit shown in Figure 4, a 9-volt triggering pulse is required with the load line intersecting the characteristic at point A. If we obtain a nominal 30-volt output pulse, a voltage gain of about 10 db is provided by the gated amplifier. However, if the biasing potential is decreased so that intersection of load line and characteristic occurs only 3 volts below the turnover point, a similar result may be obtained by changing the base resistance or bias voltage so that the turnover point of the emitter characteristic comes at -6 volts. Similar increases may also be obtained either by increasing the emitter load resistance or by shaping the characteristic to obtain intersection of load line and characteristic at lower voltages with respect to the turning point.

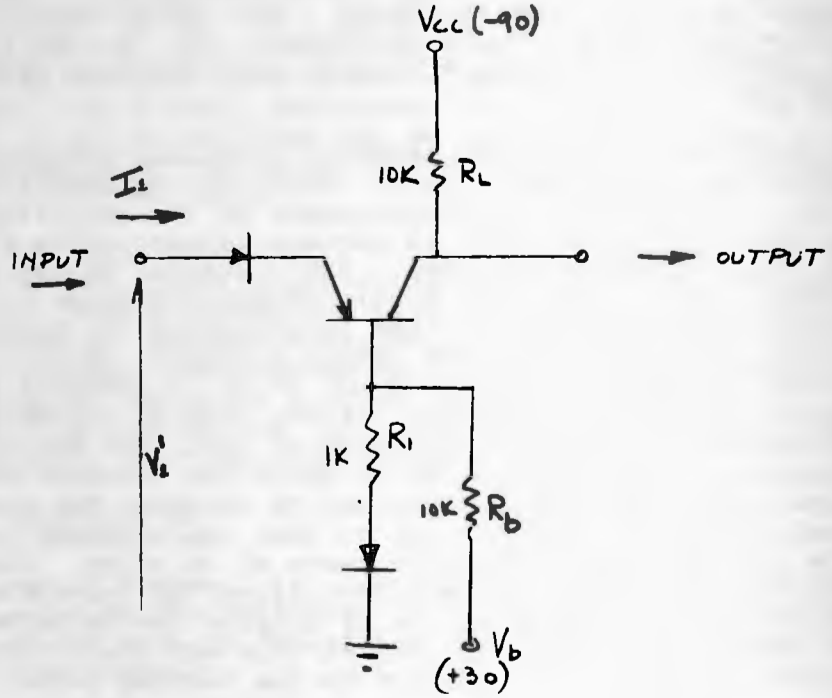


Fig. 1. Basic Circuit for Transistor Gated Amplifier

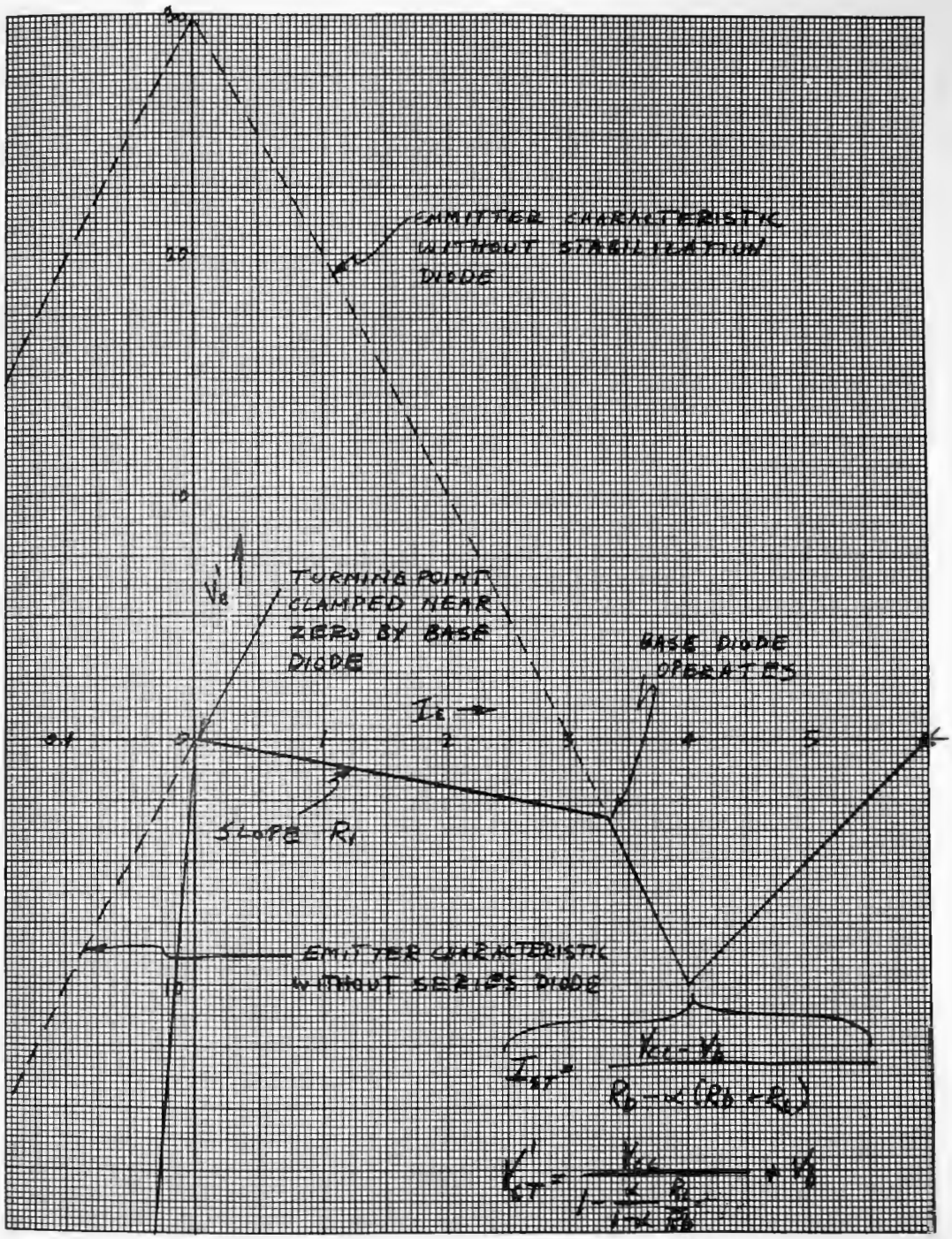


Fig. 2. Characteristics of Transistor Gated Amplifier

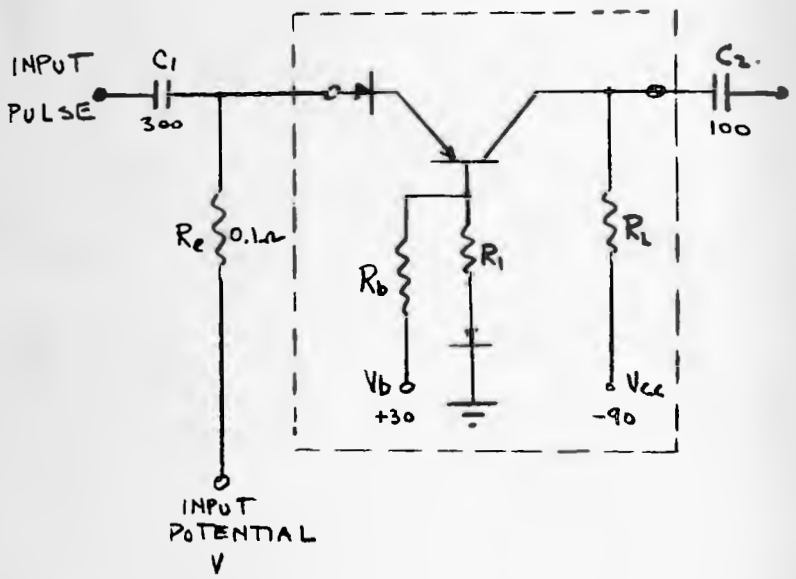


Fig. 3. Circuit for Transistor Gated Amplifier with Capacitor-Coupled Input and Output

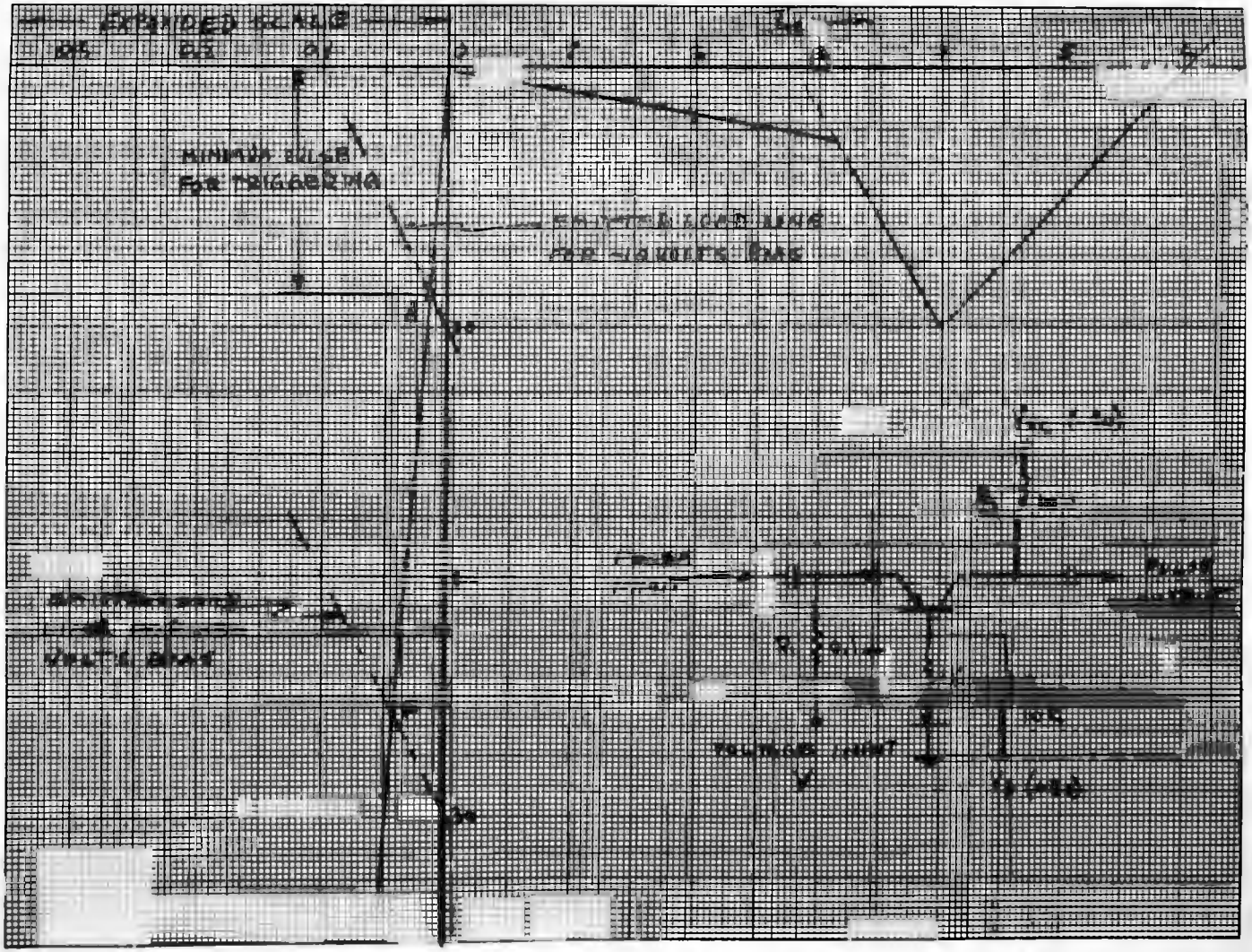


Fig. 4. Characteristics of Transistor Gated Amplifier with Capacitor-Coupled Input and Output

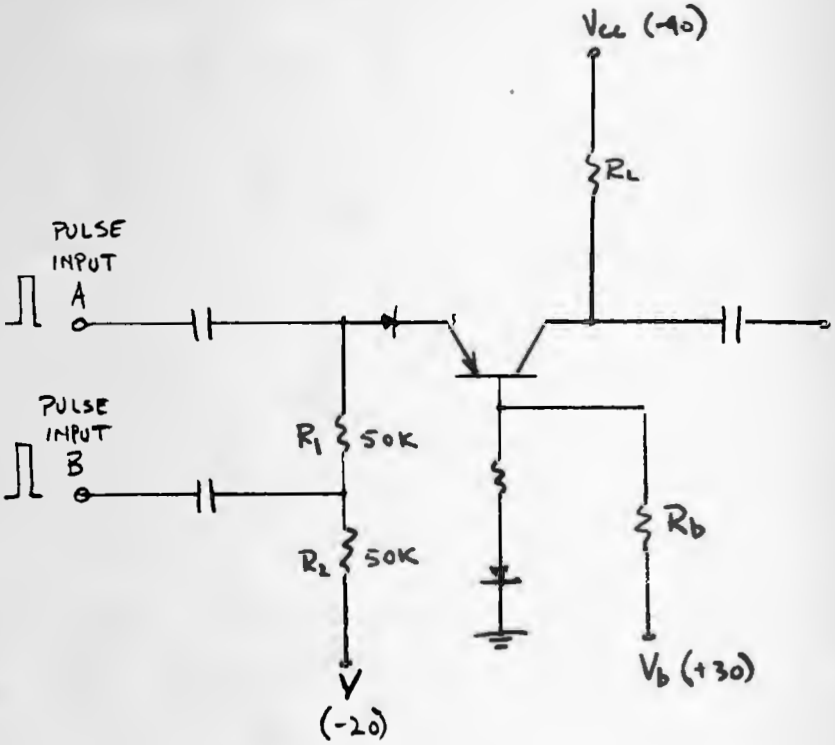


Fig. 5. Circuit for Transistor Gated Amplifier with AND Pulse Gating

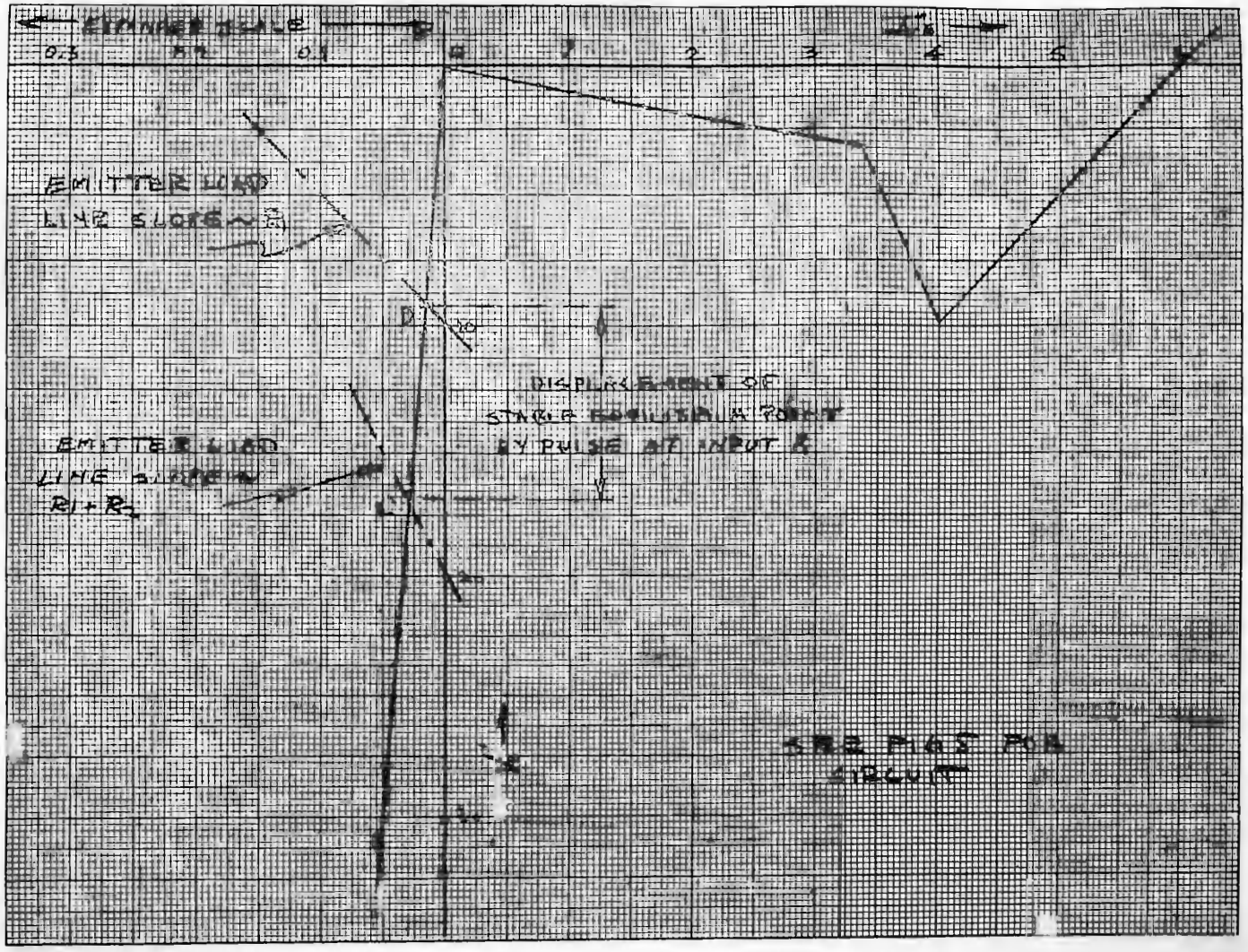


Fig. 6. Characteristics of Transistor Gated Amplifier with AND Pulse Gating

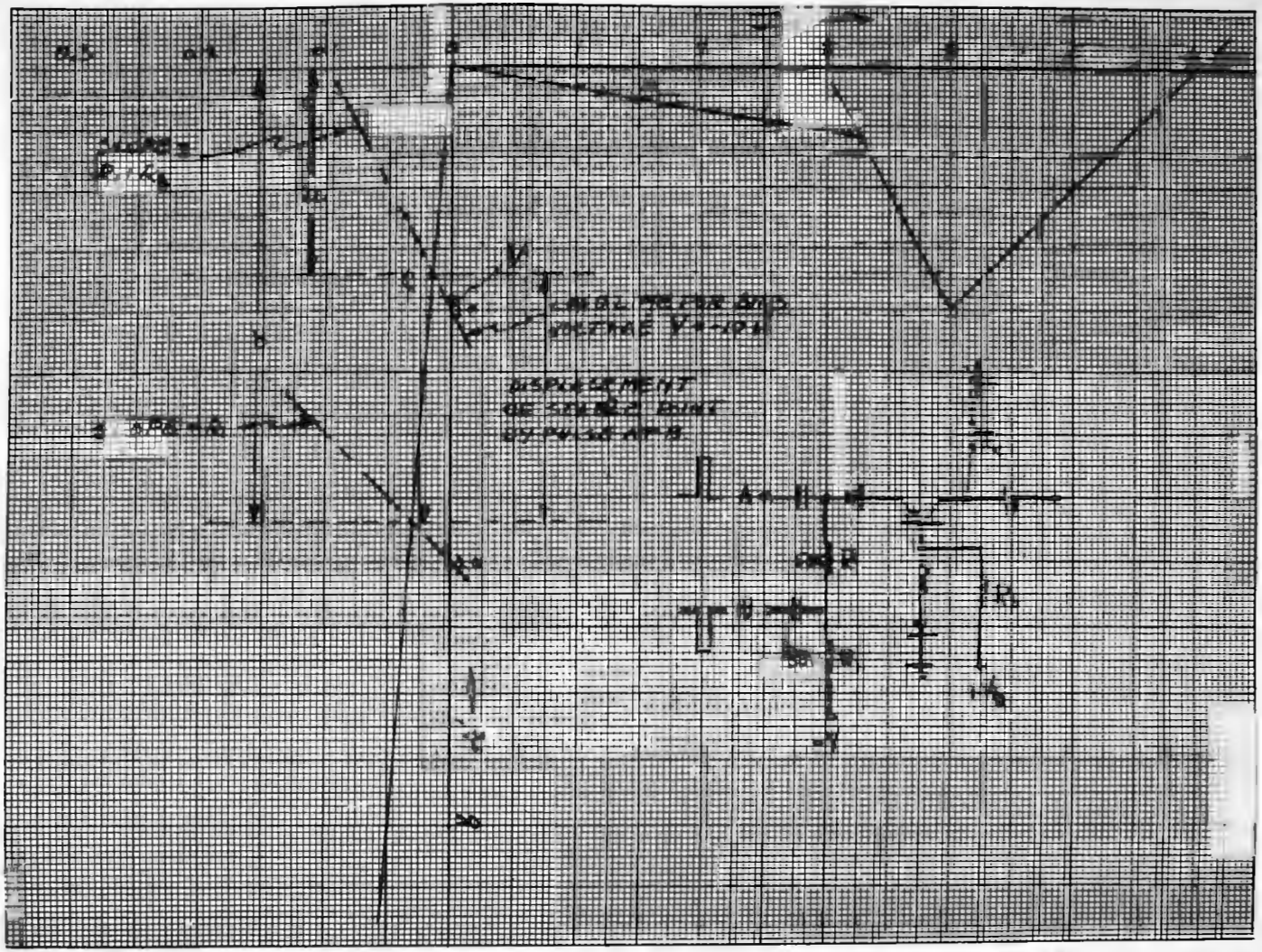


Fig. 7. Circuit and Characteristics of Transistor Gated Amplifier with BUT NOT Pulse Gating

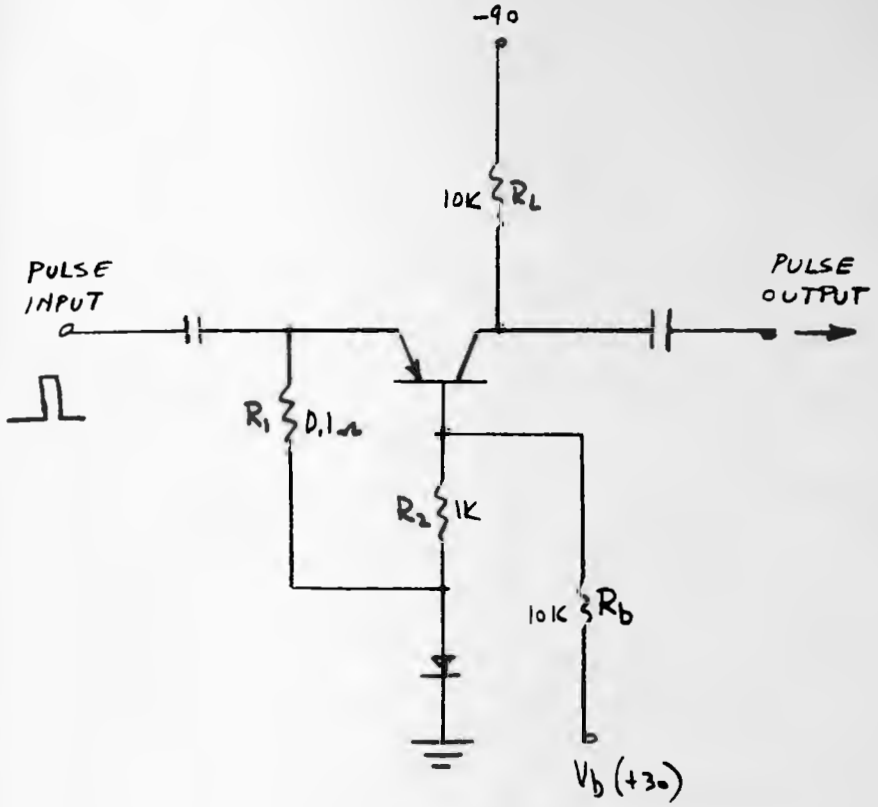


Fig. 8. Circuit for Pulse Regeneration in Transistor Gated Amplifier

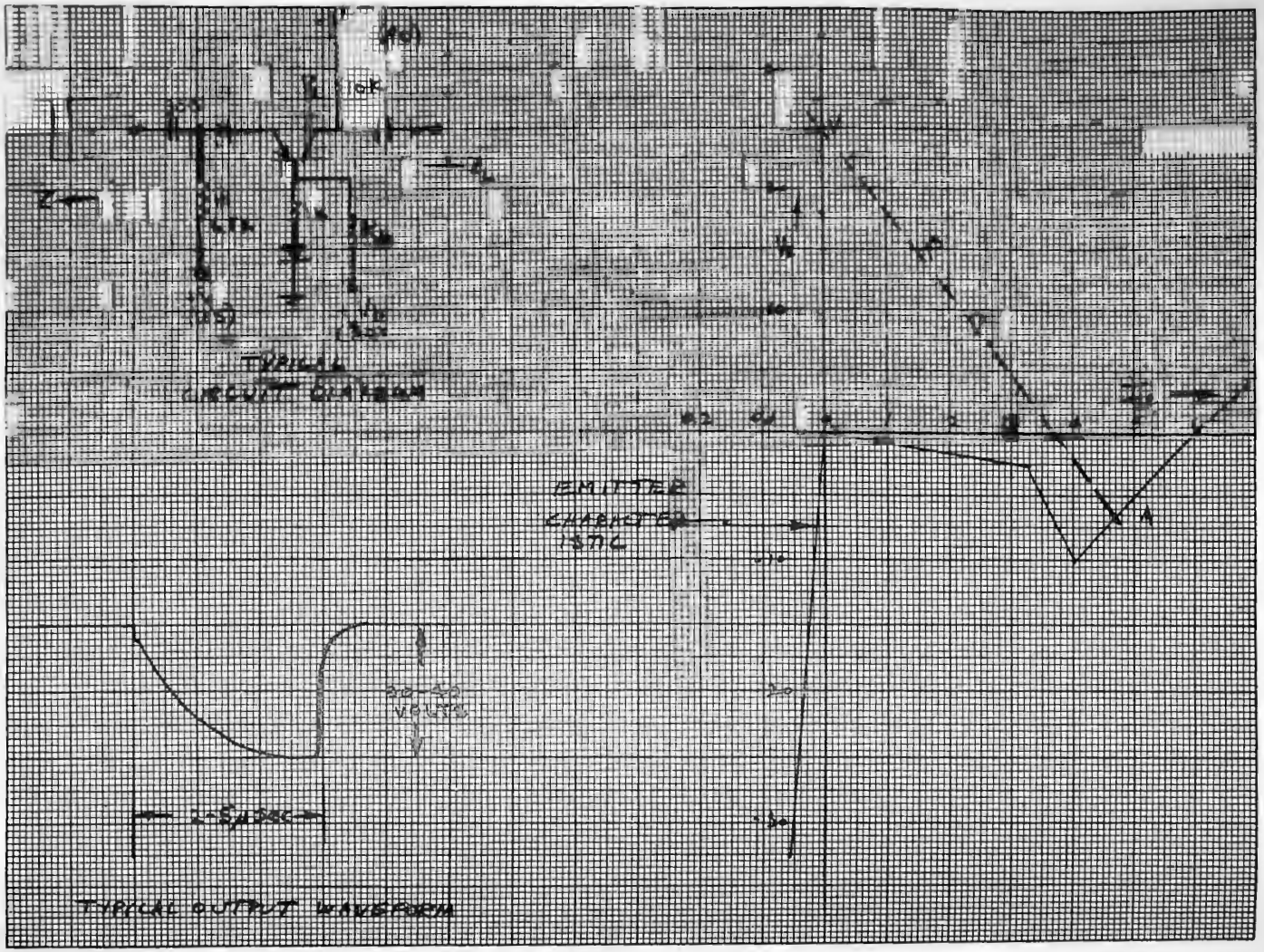
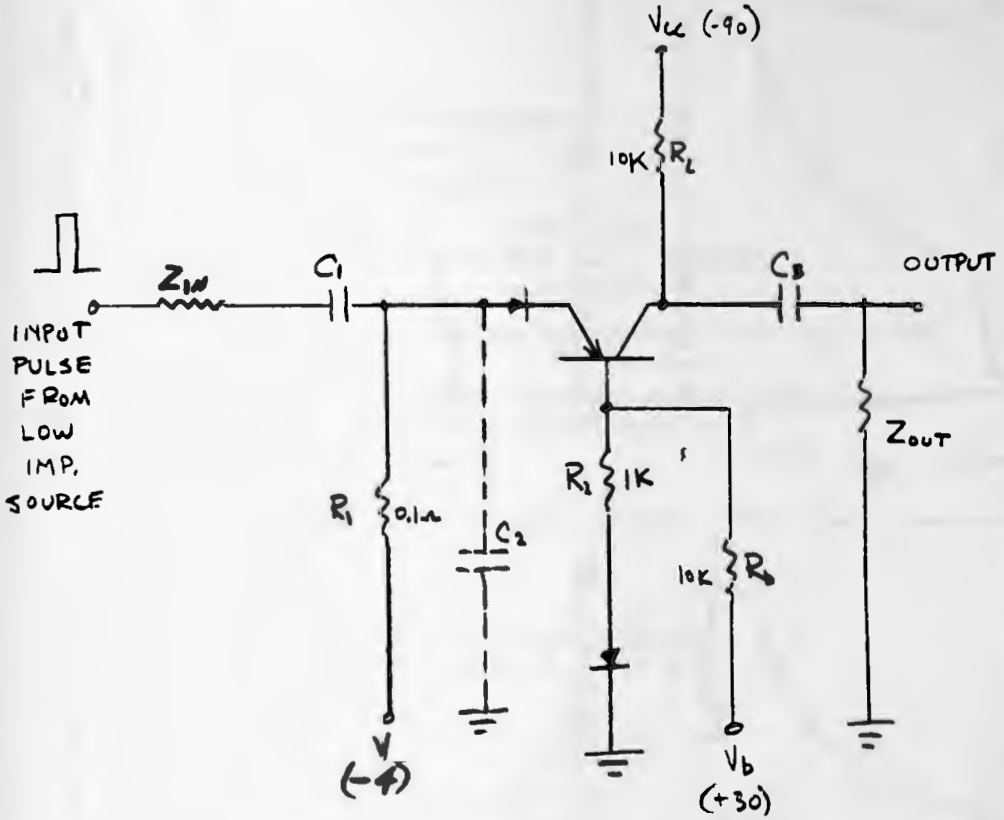


Fig. 9. Circuit and Characteristics of Transistor Gated Amplifier with Added Delay



$Z_{IN} = Z_{OUT} \sim 2400\ \Omega$

Fig. 10. Circuit for Modification of Transistor Gated Amplifier Output Waveform

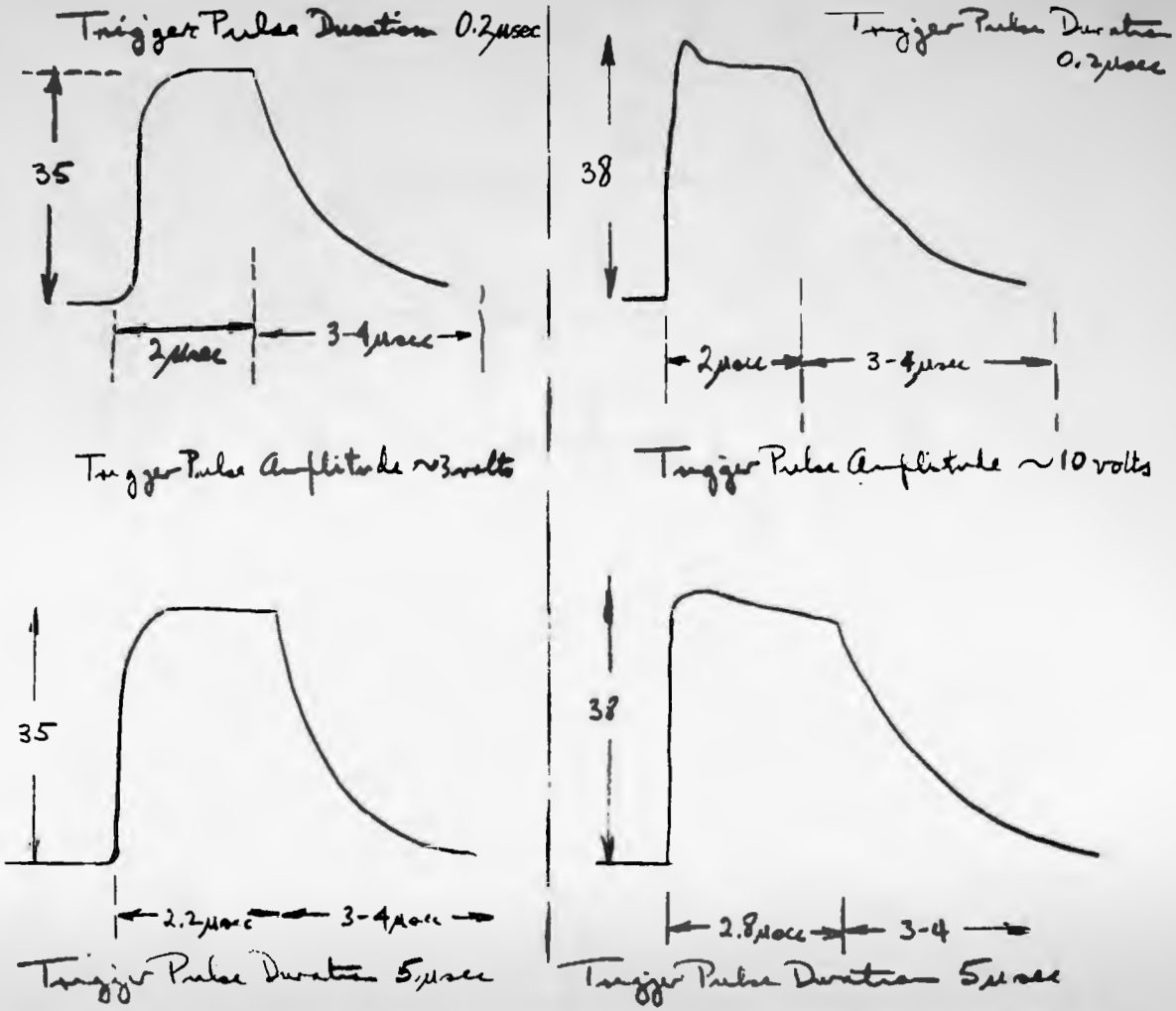
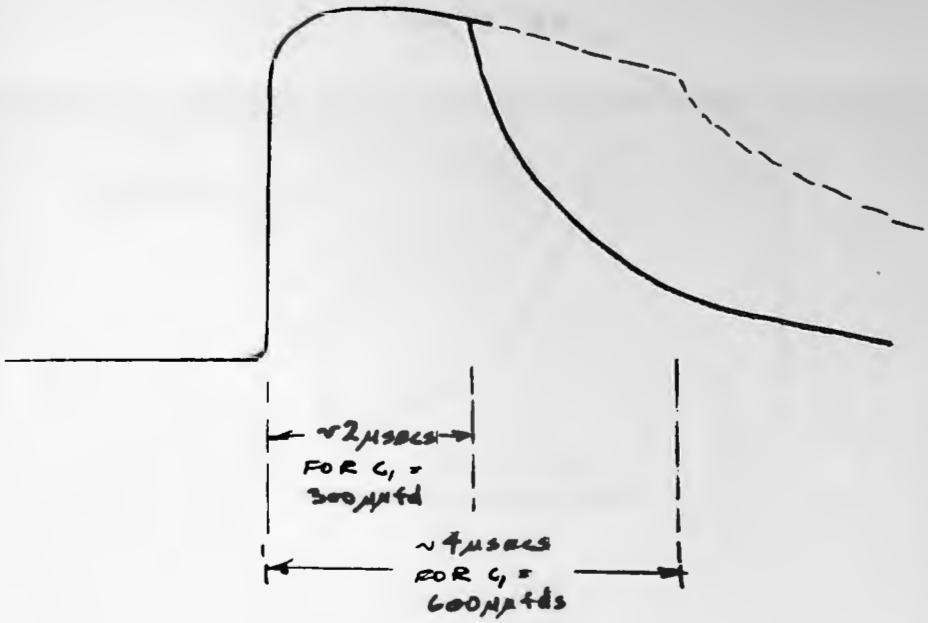


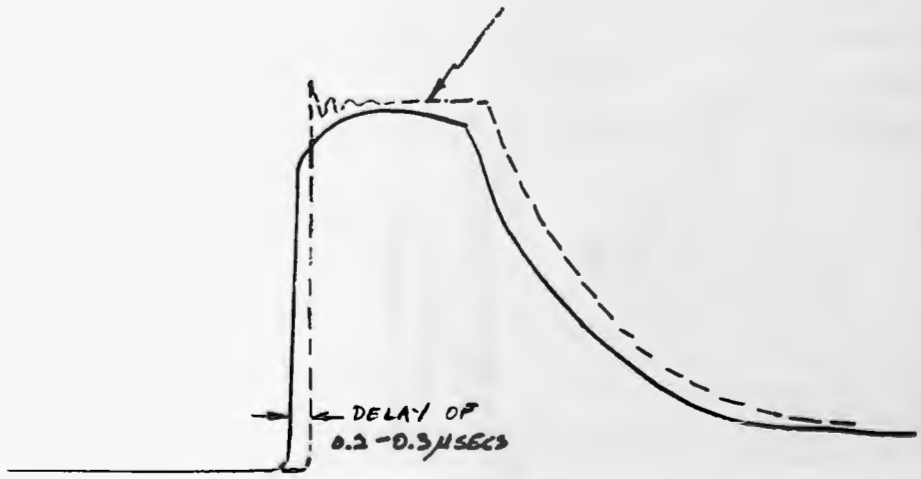
Fig. 11. Effects on Waveform of Amplitude and Duration of Triggering Pulse



Note: Wave Front and Trailing Edge Waveform Substantially the Same

Fig. 12. Output Waveform at Collector of Transistor Gated Amplifier

Effect of adding 100 mmf
to stray capacitance C_2



Note: Duration and Trailing Edge Waveform Unaffected

Fig. 13. Effect on Putput Waveform of Increasing Stray Capacitance C_2

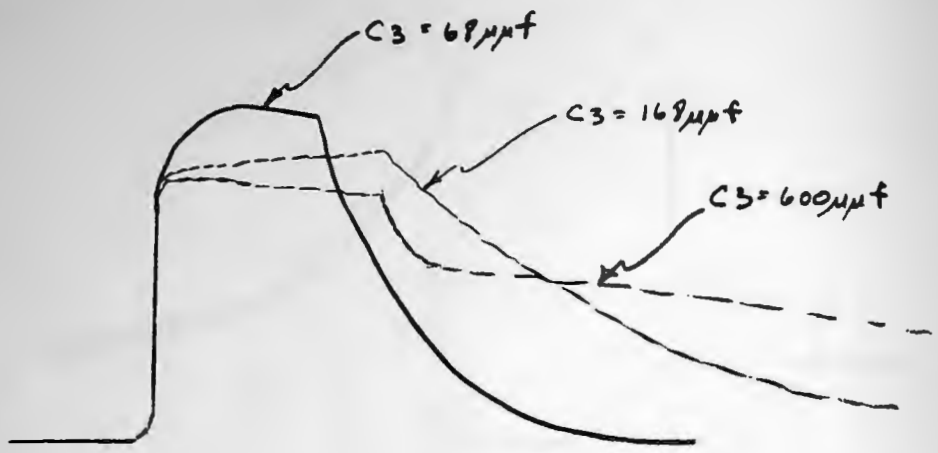


Fig. 14. Effect on Output Waveform of Changing Capacitance C_3

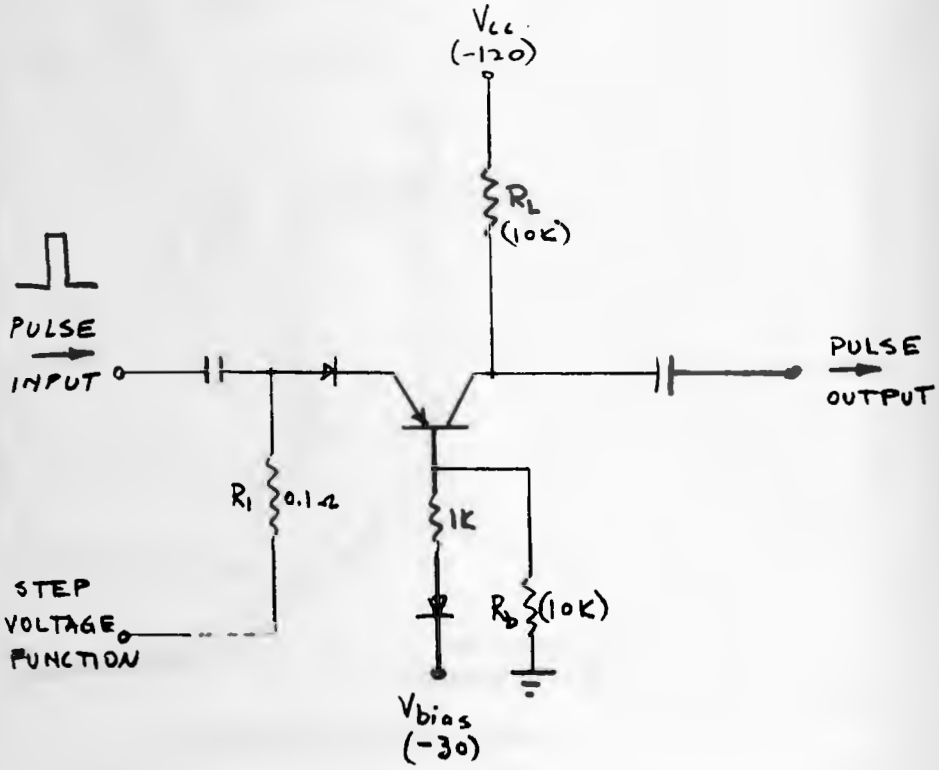


Fig. 15. Circuit for Transistor Gated Amplifier with Displaced Characteristic

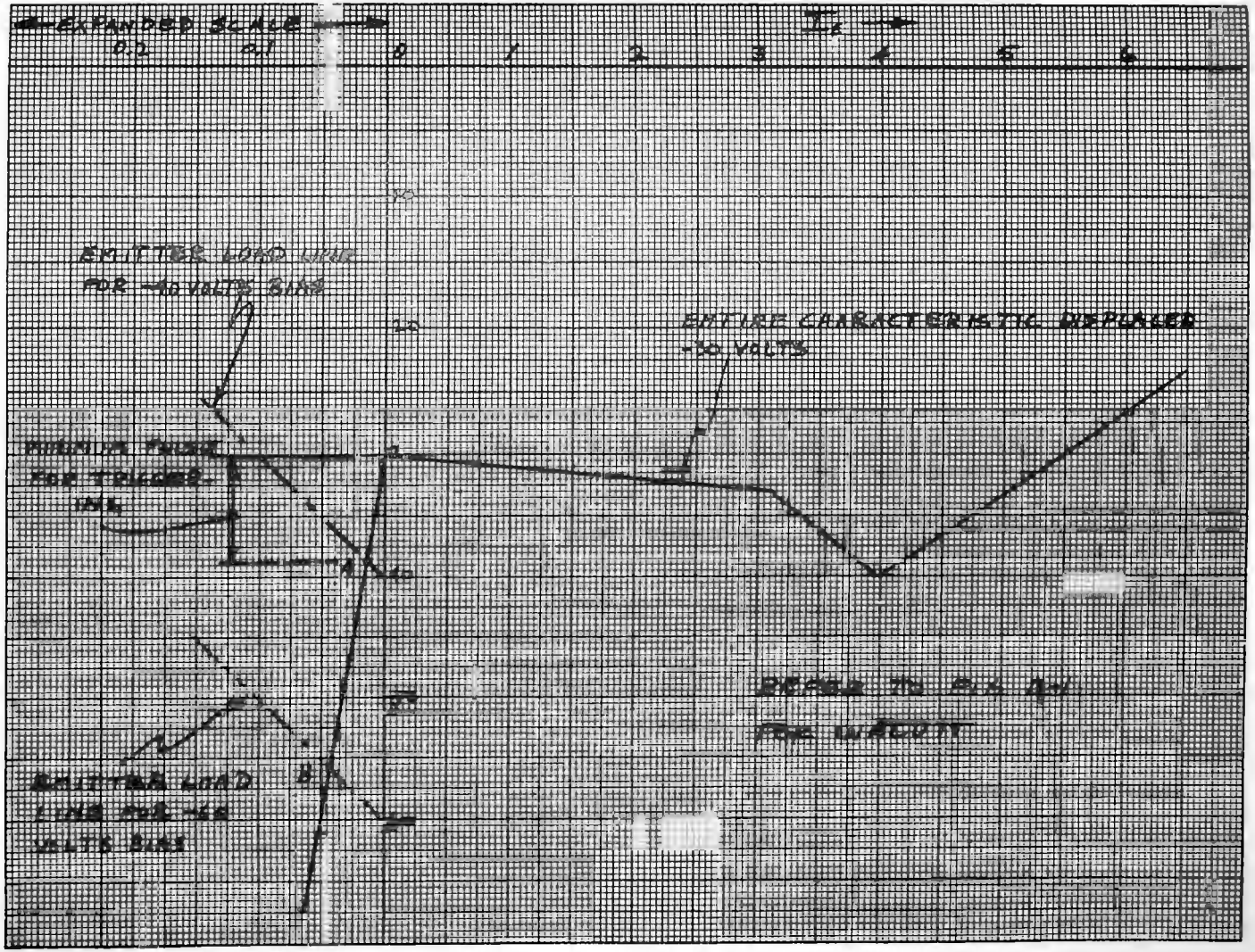


Fig. 16. Displaced Characteristic of Transistor Gated Amplifier

STABILIZED SINGLE TRANSISTOR BINARY COUNTER

R. L. TRENT

Produced under contract W 36-039 sc-44497

STABILIZED SINGLE TRANSISTOR

BINARY COUNTER

The stabilization means previously described * has been applied to a single transistor binary counter stage. The basic circuit has been described by J. T. Bangert, of the Bell Telephone Laboratories, and is illustrated in Figure 26, together with the modified stabilized stage.

The principles of operation of the typical unstabilized stage are as follows:

1. Referring to Figure 27, which shows the negative resistance characteristic for the stage, the OFF condition is defined as stable points A or A', when the emitter current is nearly zero or negative, corresponding to a low collector current; and the ON condition is defined as stable point B, corresponding to positive emitter current and high collector current. In the OFF condition, the base is slightly less negative than the junction point X, so that the varistor is in its conducting low resistance state. In the ON condition the base is more negative than the junction point X and the varistor is in its high resistance non-conducting state.

2. Assuming that the transistor is in the OFF condition, a negative pulse applied at the input terminal will divide between the emitter and the base; a much larger proportion of the pulse will be impressed upon the base due to the lower impedance of the series path to the base. Since a negative increment on the base is equivalent to a positive increment on the emitter, the emitter will be triggered positively into the negative resistance region of the device. In this region the current gain is high and more collector current will flow than emitter current. Since the emitter and collector currents flow in opposite directions in the base resistor, the larger collector current causes the base to become more negative. the same direction as the initiating pulse. This process continues until equilibrium between collector current increments and emitter current increments is obtained and the ON condition stable point B is reached. The next negative pulse appearing at the junction point X sees a high impedance path to the base and a low impedance path to the emitter, so that the pulse is applied principally on the emitter, causing a reduction in emitter current and a consequent triggering to the OFF condition. Thus the device changes state each time a negative pulse appears at the input.

*"Idealized Negative Resistance Characteristics of the Transistor", by R. L. Trent.

3. Figure 27 illustrates the variation in the device characteristic which will result from unit to unit for values of V_{co} between 120 and 30, as discussed in Report 5. The difficulty in triggering the higher V_{co} units for these biasing conditions is immediately apparent, since the emitter current increment required for triggering ranges from about 0.5 mA to 3.0 mA, in the OFF condition.

The principles of operation of the stabilized stage are as follows:

1. In addition to the circuit elements for the unstabilized stage, the base resistor is split up into R_3 and R_4, 1K and 9.1K, respectively, with the junction point Y connected to ground through a diode varistor poled with its positive terminal towards the base. The device characteristic for this circuit configuration is illustrated in Figure 28. An attempt has been made to keep circuit elements comparable in value to those shown for the unstabilized stage for comparison purposes, and it should be understood that the circuit elements shown are not optimized for best circuit operating conditions.

2. Assuming that the transistor is in its OFF condition, a negative pulse applied at the input terminal will divide as previously described between the emitter and the base, with the larger proportion appearing at the base. This initiates the triggering action as before. However, as soon as the collector current exceeds the emitter current by the amount of the biasing current passed through the series diode in the base lead by the +19 volt biasing voltage, that diode assumes its high resistance condition, and the base resistance becomes essentially the sum of R_3 and R_4, neglecting the effect of the shunting diode resistance. At the same time, the additional biasing voltage causes a vertical shift in axis of the load line. The biasing voltage is so chosen that this vertical axis shift occurs at approximately zero emitter current. The change in value for the load line impedance is due to the combined effects of the two diodes changing state. When the transistor is in the OFF condition; or in the region of negative emitter current, the emitter load resistance may be approximated by:

$$R_L = R_1 + \frac{R_2(RV_1 + RV_2 + R_3)}{R_2 + R_3 + RV_1 + RV_2} \triangleq 1600 \text{ ohms} \qquad (1)$$

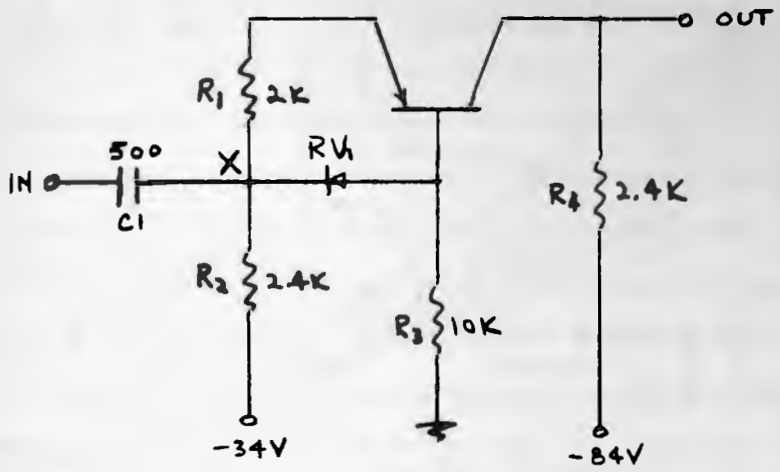
This equation neglects the shunting effects of R_4 and the internal collector resistance r_c in series with R_5, since they are several orders of magnitude greater than the resistances they shunt.

As soon as the diodes assume their high resistance condition (at $I_\varepsilon \triangleq 0$), the emitter load resistance becomes:

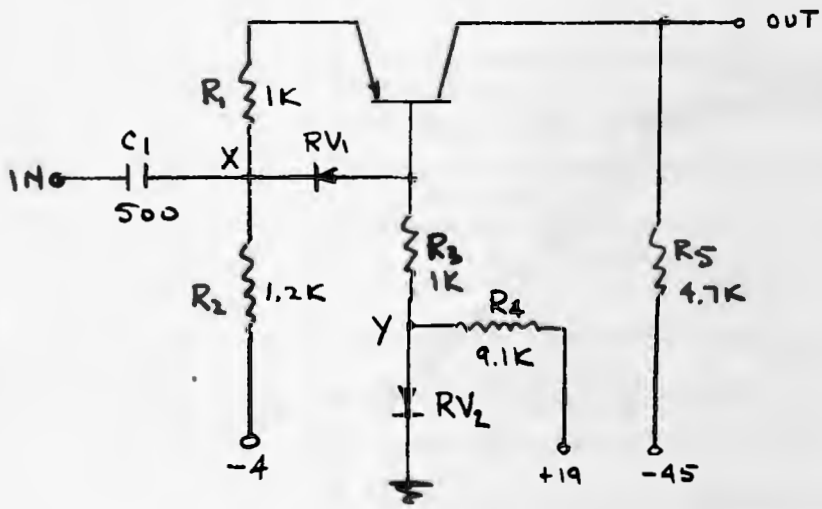
$$R_L = R_1 + R_2 = 2200\,\omega \qquad (2)$$

By the same process as described for the unstabilized stage, equilibrium between collector current and emitter current increments, after triggering from the OFF condition results in attaining the ON condition stable point B. The next negative pulse appearing at junction point X sees a high impedance path to the base and a low impedance path to the emitter, so that the pulse is applied principally on the emitter, causing triggering to the OFF condition. As soon as the negative emitter current region is attained during this triggering period, the diode RV_2 reverts to its conducting state and similarly, as soon as the base becomes less negative than the junction point X, the diode RV_1 becomes conducting, and equilibrium at the OFF condition stable points A-A' will result.

An examination of Figure 28 shows the improvement attained in triggering conditions by the stabilization means, in that the variation in emitter current from point A to A' (the OFF condition stable points) is 0.25 to 1.25 mA compared to 0.5 to 3.0 mA for the unstabilized stage.



SINGLE TRANSISTOR BINARY COUNTER STAGE



STABILIZED SINGLE TRANSISTOR BINARY COUNTER STAGE

Fig. 1 Binary Counter Circuits

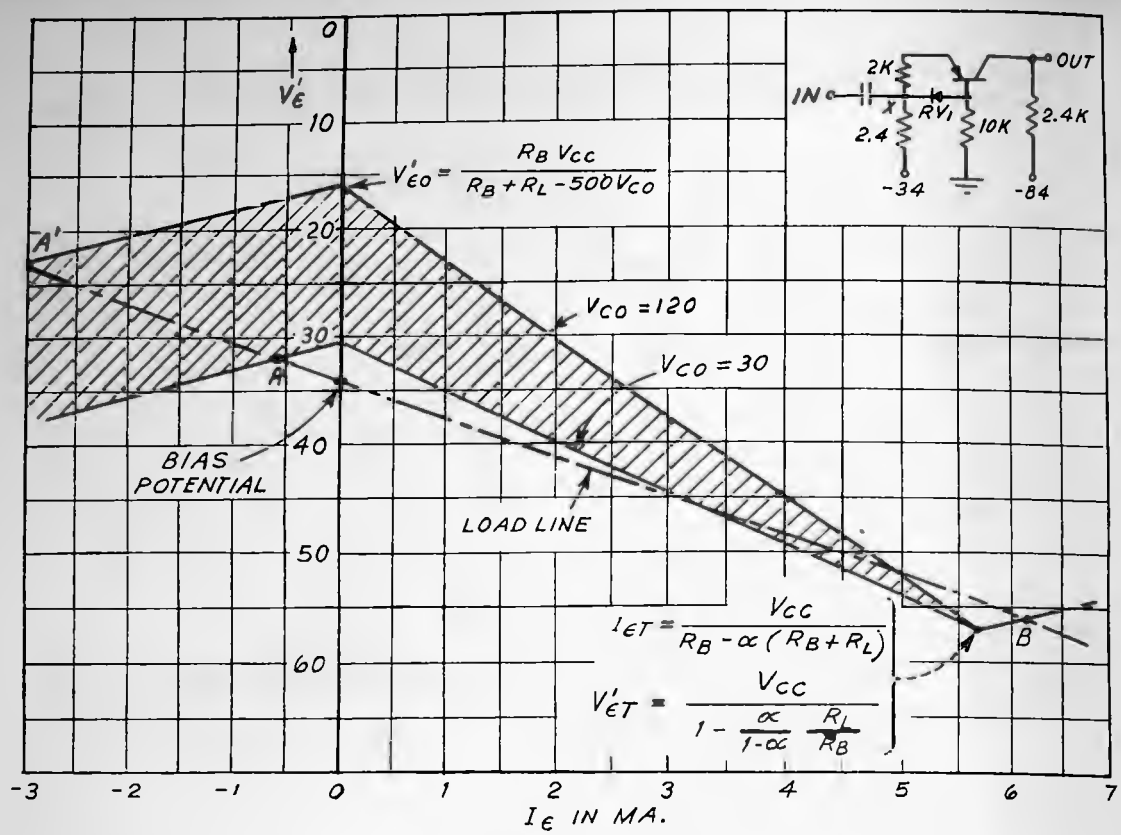


Fig. 2 Binary Counter Characteristics

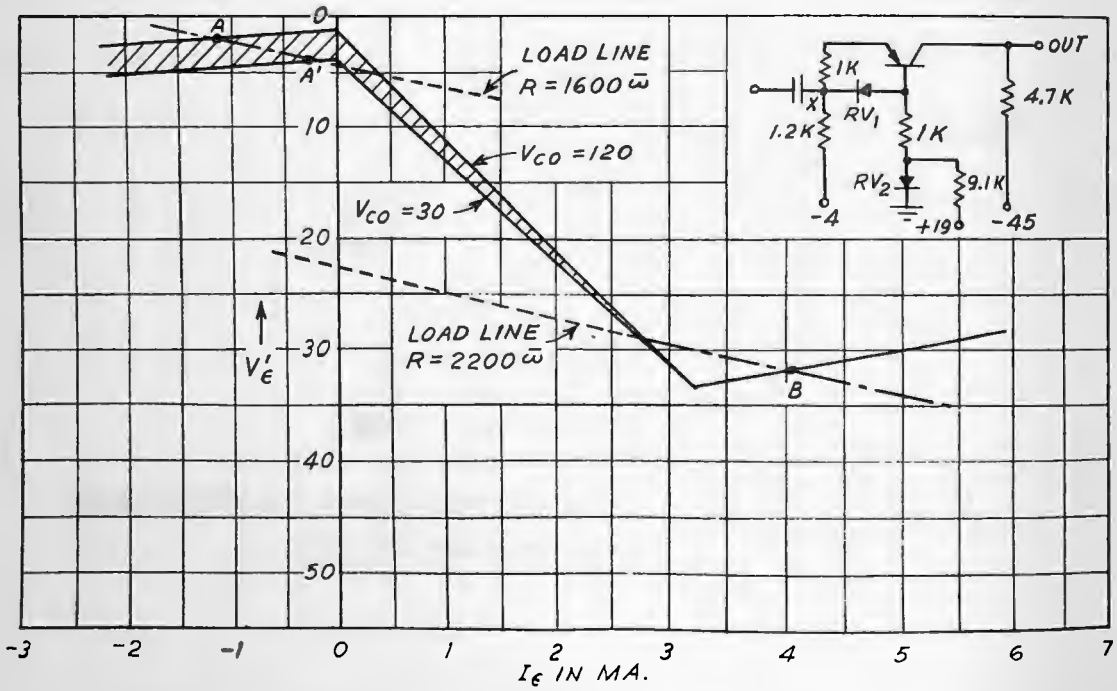


Fig. 3 Stabilized Binary Counter Characteristics

A STABILIZED GENERAL PURPOSE TWO-TRANSISTOR BINARY COUNTER

R. L. TRENT

Produced under contract W 36-039 sc-44497

A STABILIZED GENERAL PURPOSE TWO-TRANSISTOR BINARY COUNTER

1.0 General

The purpose of this report is to describe the theory and operation of a newly developed general purpose binary counter employing two transistors, and to discuss its characteristics and performance.

In this report, the term binary counter is used to describe a circuit having two d-c stable equilibrium points, and which can be triggered from one stable point to the other by the application of pulses of the same polarity to a common input terminal.

1.10 Requirements

The following general requirements were set up as design criteria.

1) Triggering by pulses of either positive or negative polarity.

2) Repetition Rate: 0 to 50 kc.

3) Independence from reasonable power supply variations.

4) Independence from transistor variations due to aging and temperature effects.

5) Average ratio of output pulse amplitude to triggering pulse amplitude of at least 5 to 1.

6) Incorporation of means of setting to either ON or OFF condition by means of suitable pulses applied at independent input terminals.

1.20 Circuit Description

A functional representation of a binary counter incorporating the required features is shown in Figure 1. Figure 2 is a schematic of a developed circuit, incorporating two transistors, meeting the above requirements. The circuit comprises two balanced transistor stages with back couplings between collectors and bases, consisting of resistor-capacitor networks (C_2 and R_4, C_1 and R_5). In addition there is a d-c direct coupling by means of a common emitter load resistor (R_{12}) and an a-c direct coupling between emitters (C_3). These couplings provide regenerative paths for the passage of a-c signals and d-c paths to achieve stability once the transient effects have subsided. Triggering pulses are impressed upon the bases under the control of a diode steering circuit (RV_1, RV_2) which is given direction by the difference of potential existing between the bases when one unit is in the ON condition and the other is in the OFF condition. The necessity for a separate biasing supply for the emitters is eliminated because of the fact that the common emitter resistance (R_{12}) at the two stable equilibrium points is always passing the emitter current of the ON stage to provide the biasing potential for the OFF stage. In addition, the circuit utilizes base stabilization and a feature insuring low triggering requirements.

In order more easily to describe this circuit the various features will be discussed one by one by means of skeleton circuits.

1.21 Stabilization Feature

In the design of transistor trigger circuits, variations among transistors, and variations in the operation of a single transistor at different temperatures, affect adversely the triggering sensitivity and stable operating points. These effects have been traced to the fact that transistors, at present, vary widely from one another in the amount of base current which flows when the emitters are biased negatively in the non-conducting region. This current is denoted as the I_{co} of the transistor. This variable base current, flowing through the large base resistance added in the base circuit to secure the desired negative resistance characteristic, causes variations in the turning point of the characteristic. A stabilization feature* consisting of the addition of a biased

---

*"Idealized Negative Resistance Characteristics of the Transistor", by R. L. Trent.

diode in the base lead to provide a low shunting impedance in the negative emitter current region and a high impedance in the positive emitter current range is illustrated in Figure 3. The low impedance prevents variations in collector current (I_{CO}) from unit to unit from causing a wide range of voltage to be developed across the base resistances: the high impedance condition allows the remainder of the standard negative resistance characteristic to be obtained over a range of positive emitter current. A small resistance has been added in series with the shunting diode to insure obtaining at least a minimum value of negative resistance as soon as positive emitter current obtains on triggering. In addition, this insures the existence of a sharply defined turning point in the characteristic. This sharp transition results in ease of triggering and less loading of the triggering source, and decreases the transition time, making possible higher-frequency response. By proper proportioning of the added resistance, the effects of variations in the I_{CO} on the turning point may be kept within practical allowable limits. The resultant characteristic is shown in Figure 4. This stabilization feature has been applied to both transistors in the binary counter circuit to be described.

1.22 Triggering Sensitivity Features

On of the major stability criteria of a binary counter stage is that once it has been set to one condition, it must maintain that condition indefinitely until the next incident pulse is impressed upon the input lead. Satisfaction of this requirement has often resulted in poor triggering sensitivity. The binary counter under discussion employs circuit means to secure both stability and sensitivity.

In the stripped circuit of Figure 5A the required lockup feature is performed by the common emitter coupling resistance R_1 and the cross coupling resistors shown as R_2. Figure 6A is a simplified characteristic for such a circuit which assumes that both transistors have identical characteristics and that each transistor acts separately as a two-terminal device. This assumption is not necessarily the case, but it will serve to illustrate the point under discussion. The ON unit is assumed to be stable at point B, the intersection of the load line R_1 with the characteristic. Because of the coupling action of R_1, the OFF unit must be stable at point A at the same emitter potential. It can readily be seen that positive triggering pulse must have an amplitude greater than Δ.

The effect on the characteristic and load line of adding component R_3 and RV_1 in the stripped circuit of

Figure 5B is illustrated on Figure 6B. The ON unit has the same operating point B as before. Instead of the OFF unit operating at point A, however, the diode in series with its emitter is biased in the reverse direction, resulting in the new load line, E-E', for the OFF unit. E-E' intersects the normal characteristic at point C with a corresponding decrease in triggering pulse amplitude from greater than Δ to greater than δ. However, the characteristics also has been modified by the introduction of the shunt resistance R_3, causing intersection to occur at D instead, thereby resulting in a further decrease in the required positive triggering amplitude. Triggering now occurs if the pulse is greater than γ. While the sensitivity has been increased materially no concession has been made to the stability of the states of the circuit. This design feature has been applied to both stages of the binary counter circuit as shown.

1.23 Composite Characteristic of Counter

An approximate emitter characteristic for one of the two transistors of the counter is depicted on Figure 7. The base stabilization feature assures that the turning point of the characteristic will be very close to zero voltage, and if the stabilizing current is larger than that required to balance the I_{co}, there will be a region where the negative resistance is essentially determined by the series base resistance (R_8, R_9, Figure 2) and the diode resistance, say 2500 ohms. When the shunting diode switches to its high resistance state, the negative resistance is determined by the total series base resistance (R_8 & R_{10}, R_9 & R_{11}, Figure 2), approximately 12,200 ohms, neglecting the shunting effect of the diode resistance.

If the assumption is made that transistor TR-2 is in the ON condition, the diode in series with the emitter (RV_6) will be in its low resistance condition, and the emitter load line (shown as load line 2 in Figure 7) for TR-2 is then (R_{12} + RV_6), about 2200 ohms, intersecting the origin. The stable ON point for TR-2 is then the intersection of this load line with the characteristic at B. However, since transistor TR-1 is in the OFF condition, its emitter current is slightly negative and its series diode (RV_5) will be in its high resistance state. This determines the emitter load line slope for TR-1, and its biasing potential is set by the emitter current of TR-2 passing through R_{12}. This voltage may be calculated by:

$$I_{e(TR-2)} \left[\frac{R_{12}}{R_{12} + RV_6} \right] \qquad (1)$$

Graphically, the biasing voltage may be determined by taking the emitter voltage represented by the stable point B and multiplying it by the ratio of R_{12} to R_{12} + RV6.
We may then draw the emitter load line for TR-1, from the voltage axis intercept of the determined potential. This load line intersects the characteristic at point A, the OFF stable point for TR-1, very close to the turning point of the characteristic.

1.24 Mechanism of Triggering

The binary counter circuit is triggerable with either positive or negative pulses, provided the proper steering diode polarity is observed. With the circuit as shown in Figure 3-2, the steering diodes are connected for positive pulse triggering. Steering is accomplished by the difference of potential existing between ON and OFF unit base connections. The base of an ON unit is always 5 to 10 volts more negative than the base of an OFF unit. Therefore, a positive pulse appearing at the input terminal will be impressed upon the base of the ON unit. Because of the low impedance from base to emitter in the ON condition, this pulse appears also on the emitter of the ON unit. The coupling capacitor C_3 passes this pulse to the emitter of the OFF unit which is at high impedance. Since the OFF unit is stable at a point very close to the turning point in the characteristic, it will be triggered. The resulting very sharp pulse of about 20 volts amplitude, 0.2 to 0.3 microsecond rise time appearing at the collector of TR-1 due to this triggering action is passed by means of the cross coupling capacitor C_1 back to the base of the ON unit. This results in an abrupt shift in the characteristic of the ON unit in a positive direction. If the characteristic is shifted sufficiently so that the emitter load line 2 (Figure 7) no longer intersects the characteristic, the only stable point remaining is in the OFF stable region. Thus the collector voltage becomes more negative, following the time constant of the cross coupling capacitor C_2 and its associated discharging resistance. It is this time constant which sets an upper limit to the repetition rate of the counter.

The coupling capacitor C_3 performs an important role in addition to passing the pulses from the ON emitter to the OFF emitter. In effect, it maintains the potential differential existing between the emitters (of the order of 5 to 10 volts) during the triggering interval, and thereby assures that the steering diodes will remain polarized correctly until the end of the transition period. Thus there is no possibility of the ambiguity of having the same pulse cause two transitions or changes in state.

Typical waveforms existing at the collectors during a triggering transition are illustrated in Figure 8. These waveforms verify the above analysis of the triggering action. Thus a short time after the incidence of the positive triggering pulse, the OFF unit is triggered toward the ON condition. The duration of the delay depends primarily on the amplitude of the pulse. The sharp positive wavefront which is passed to the base of the ON unit triggers the ON unit towards the OFF condition. The wave shapes illustrate that the triggering is independent of triggering pulse length greater than a certain minimum value, largely determined by the amplitude of the pulse. A portion of the triggering pulse appears on the collector of the ON unit, as shown.

Triggering with negative pulses may be accomplished by reversing the steering diodes thereby insuring that a negative pulse on the input terminal will be impressed upon the base of the OFF unit. A typical waveshape for this type of triggering is illustrated on Figure 9.

2.0 Characteristics of Binary Counter Stage

The binary counter circuit illustrated in Figure 3-2 has the following characteristics:

2.10 Triggering Pulse Requirements

a. Polarity: Positive or Negative

b. Amplitude: >3 to 6 volts (average ≈ 3 volts)

c. Duration: >0.6 microsecond

2.20 Repetition Rate

a. 0 to 50 kc with circuit as shown.

b. 30 kc, approx. 100 kc with C_3 and cross-connecting capacitors C_1 and C_2 reduced to 100 micromicrofarads. Other circuit constants may require some optimizing for maximum performance at higher rates.

2.30 Output Pulse

Amplitude: 16 to 25 volts (average ≈ 20 volts)

Rise Time: 0.2 to 0.4 microsecond

Fall Time: 2 to 4 microseconds

2.40 Power Supply Stability

Collector Supply: -45 ± 15%

Base Bias Supply: +33 ± 15%

2.50 Transistor Requirements

In general, transistors meeting the tentative specification for the A1698 Switching Transistor will operate successfully in the circuit. In addition, degradation due to temperature effects and aging may be tolerated to a large degree. Units having I_{CO} values in excess of 3 ma have operated successfully in the circuit, whereas the specification maximum is 2 ma.

2.60 Power Input

Total Power: 0.5 watt approx.

3.0 Discussion of Working Applications

Approximately 10 of these binary counter stages have been constructed as breadboard models in the laboratory for general usage. In addition, 4 stages incorporating a modification to permit direct operation of a reed type relay have been incorporated in a demonstration model.

It is proposed to use the counters to perform various timing and registry functions , and to serve as elements in a reversible binary counter. For these applications, the counter stages will be packaged using the "Auto-Sembly" process developed by the Signal Corps. The package will occupy a volume of approximately 1 by 1 by 2 inches.

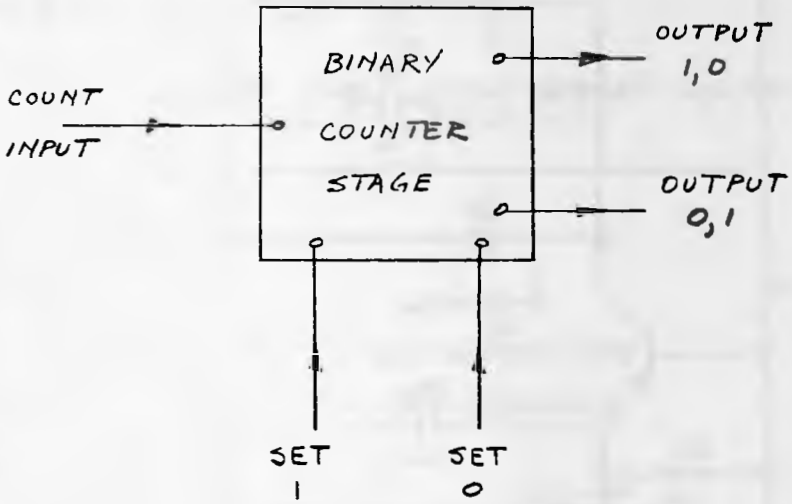


Fig. 1. Functional Representation of Binary Counter Stage

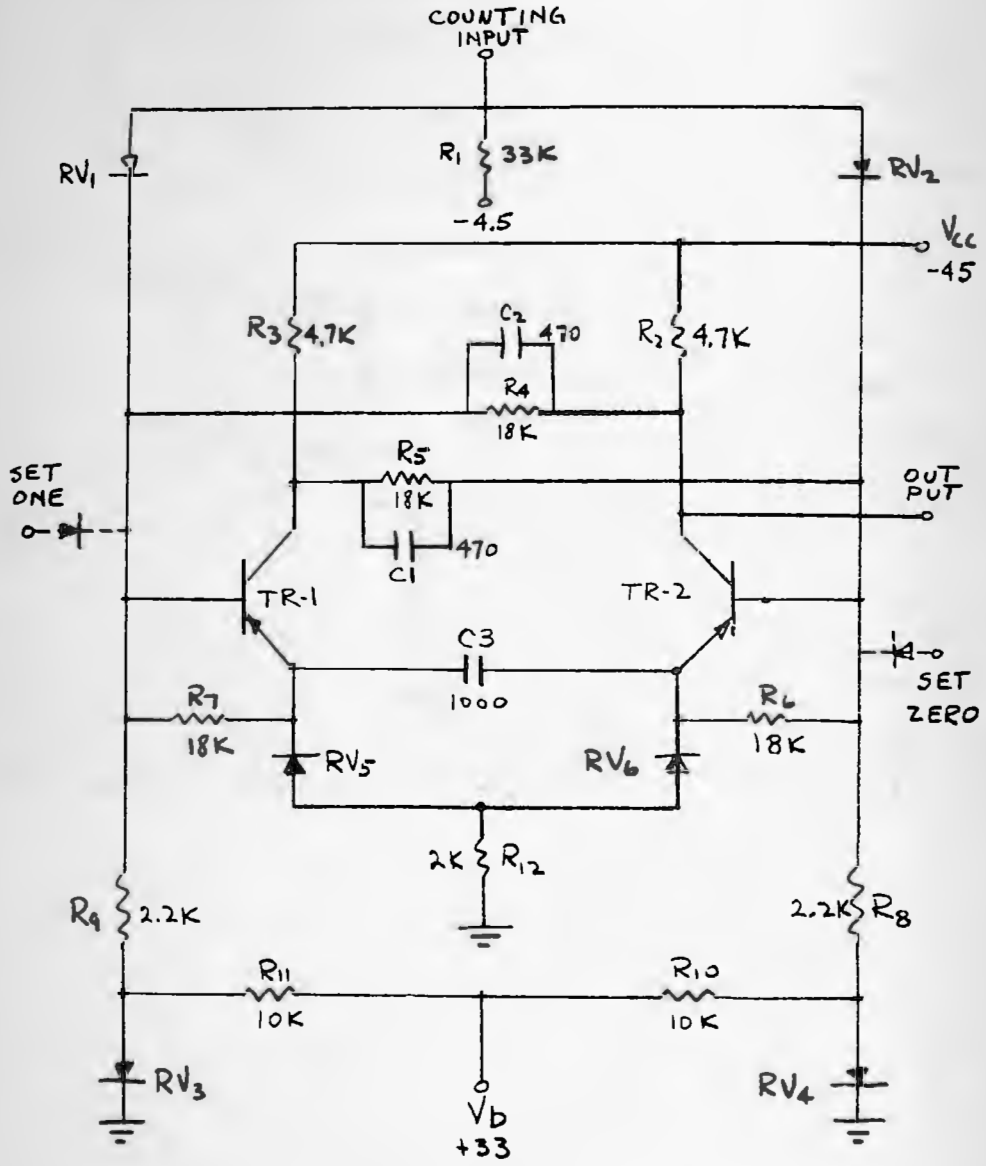


Fig. 2. Circuit of Stabilized Binary Counter

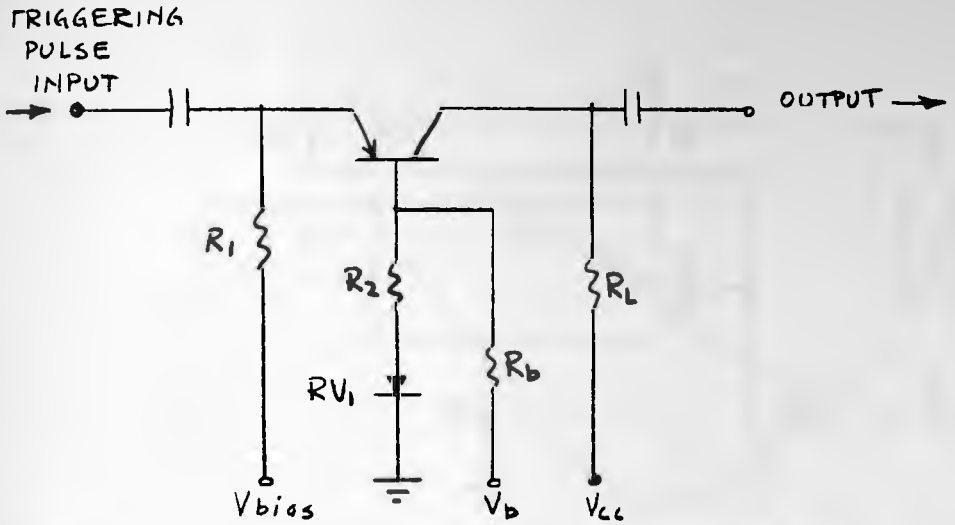


Fig. 3. Circuit of Stabilized Single Transistor for Triggering Pulse Input

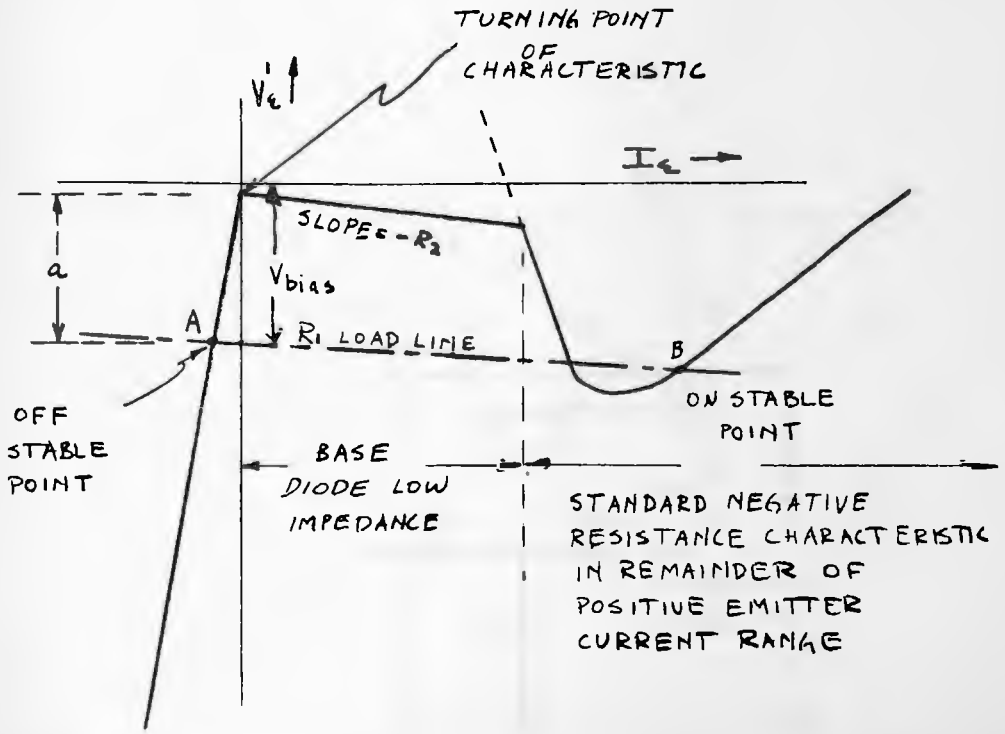


Fig. 4. Characteristic and Load Line for Stabilized Single Transistor Trigger Circuit

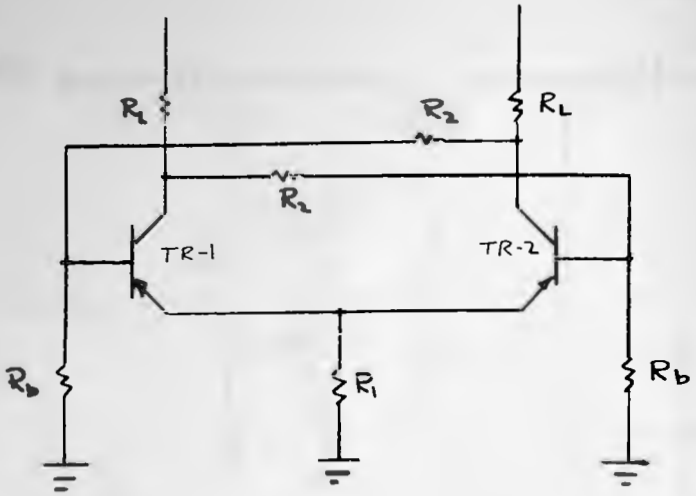


Fig. 5A. Trigger Circuit with Lockup Feature Provided by Common and Cross Coupling Resistors

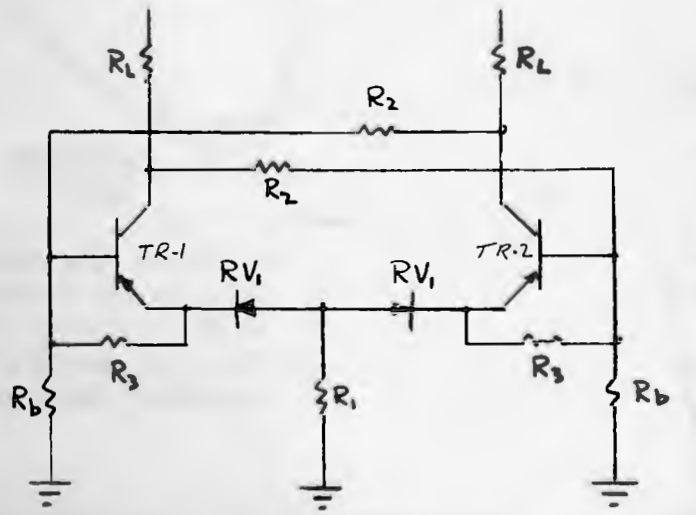


Fig. 5B. Modified Trigger Circuit of Increased Sensitivity

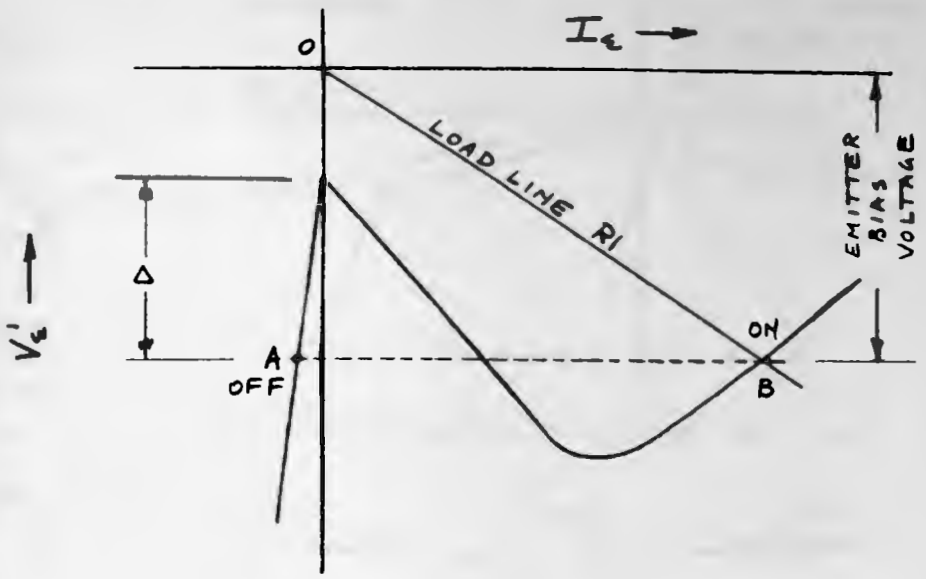


Fig. 6A. Emitter Characteristic of Trigger Circuit with Simple Lockup Feature

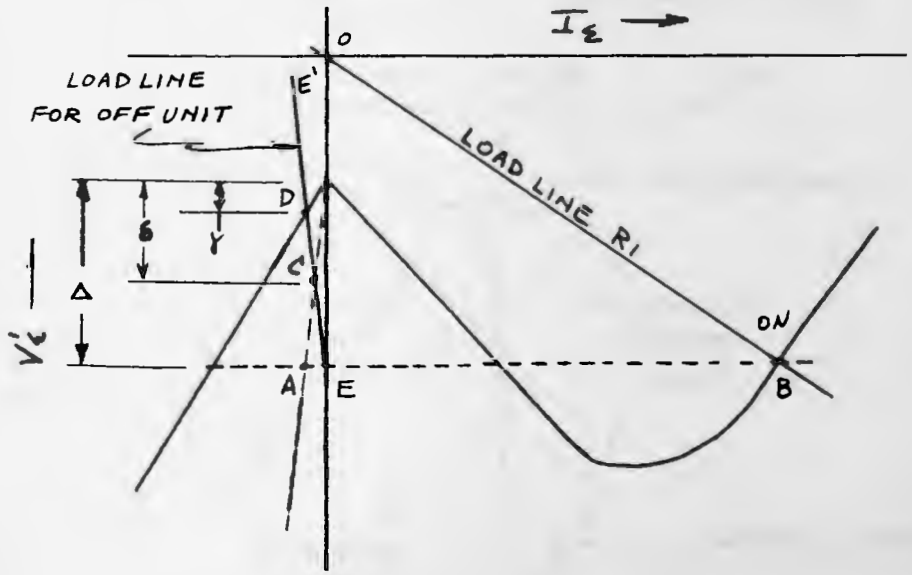


Fig. 6B. Emitter Characteristic of Trigger Circuit with Lockup Feature Modified for Increased Sensitivity

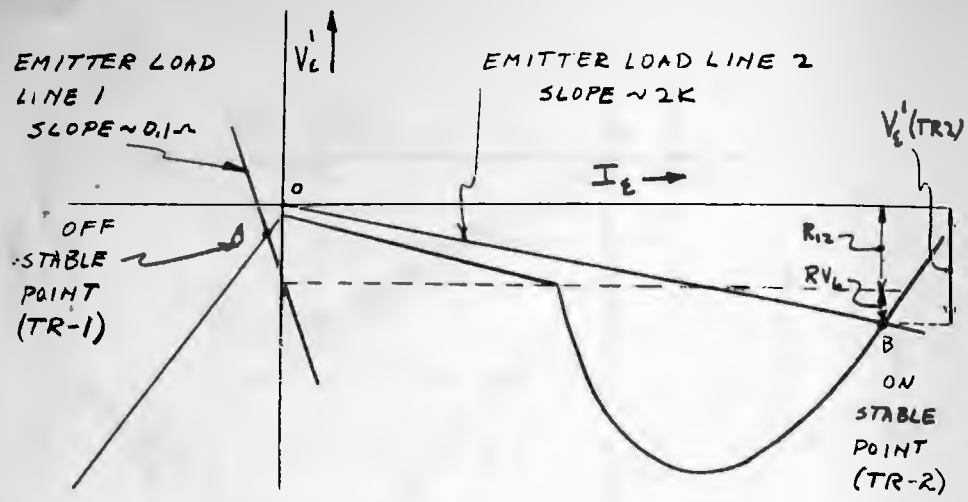


Fig. 7. Emitter Characteristic for Binary Counter

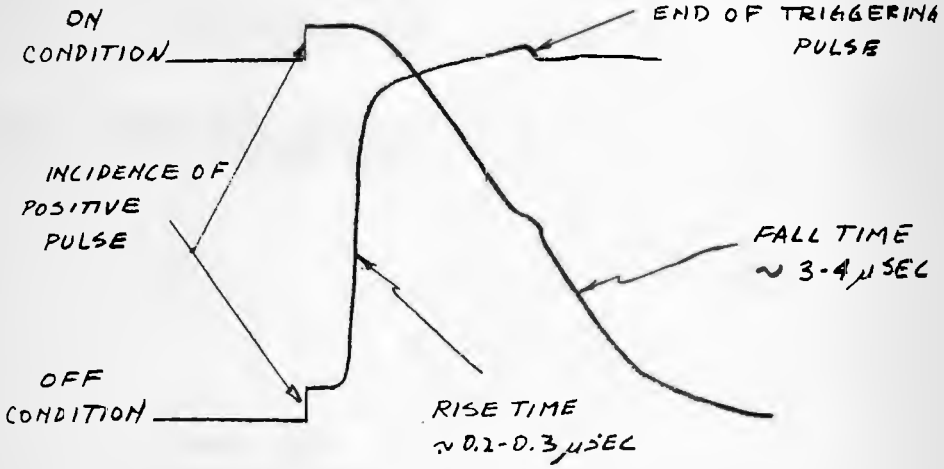


Fig. 8. Typical Waveforms at Collectors during a Positive Triggering Pulse

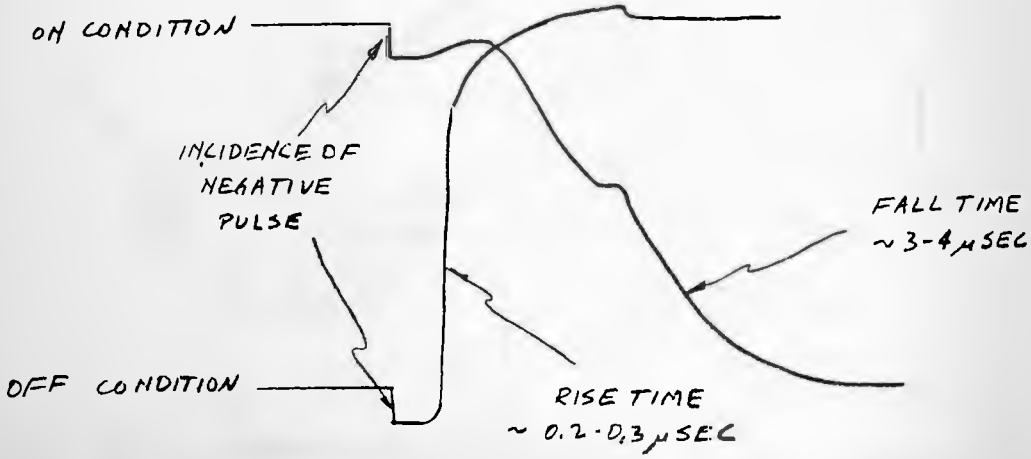


Fig. 9. Typical Waveforms at Collectors during a Negative Triggering Pulse

A TRANSISTOR REVERSIBLE BINARY COUNTER

R. L. TRENT

Produced under contract W 36-039 sc-44497

A TRANSISTOR REVERSIBLE BINARY COUNTER

The purpose of this report is to explain the operation of a Reversible Binary Counter and to describe the application of various transistor circuits to the development of such a counter.

1.0 General

A binary counter may be defined as a counting mechanism which is capable of translating a series of input pulses into a binary code corresponding exactly to the sum of the number of pulses impressed. A Reversible Binary Counter (RBC), in addition, should be capable of performing the function of subtraction as well as addition under the influence of some control or memory function. This means that if the control function indicates that addition to the binary sum already present in the counter is required, the next triggering pulse impressed on the input to the device will cause an additional binary digit unit to be stored. If the control function indicates that subtraction is required, the next impressed triggering pulse will cause a binary digit unit to be subtracted from the stored code.

2.0 Theory of Operation of a RBC

2.10 Functional Conventions

An elementary functional diagram of a RBC is shown in Figure 1. The following conventions are assumed in describing this figure:

1. The leads which register the state of a particular binary counter stage are designated A and B. The states between which a lead will change are defined as 0 and 1, where 1 is more positive than 0.

2. The binary counter stages are assumed to contain two active devices, such that if leads designated A register a state corresponding to 1, the leads designated B will register a state corresponding to 0, and vice versa.

3. Assuming that a counter stage is arranged so that it responds only to a positive pulse impressed upon its count input lead, a change in state from 1 to 0 of the output lead to which a count input connection to the following stage is made, will cause no change in the state

of the following stage. However, a change in state from 0 to 1 will cause a positive pulse to be impressed upon the count input lead of the following stage, causing it to change states.

4. The gating functions represented on the figure by simple switches are assumed to act in unison, making simultaneous connections to all A or all B leads under some control stimulus. In addition, this operation of switching is assumed to occur a sufficient time after a triggering pulse has appeared at the count input to the RBC so that all resultant changes in state of the counter will have already occurred. It is also implicit that switching between all A or all B leads must have occurred before the incidence of the next triggering pulse.

5. By the rules of translation between decimal and binary numbers, a 1 appearing on the B lead of the first stage has a weight of 1, a 1 appearing on the B lead of the second stage has a weight of 2, and a 1 appearing on the B lead of the third stage has a weight of 4, etc., so that a 4 stage binary counter can register all numbers between 0 and 15.

2.20 Discussion of Operation

Considering Figure 1, and the binary code representation illustrated, we may assume that the state present on the registry leads B for all stages is 0, and that under the influence of the control function, the interconnecting gates are all connected to the A leads to perform the addition function. Then a positive triggering pulse impressed upon the counting input lead will cause Stage 1 to assume the 1 state on its registry lead B, and since the A lead changed from 1 to 0, no positive pulse is transmitted to the next counter stage and no further action follows.

Count pulse 2 causes stage 1 to change state again so that a 0 condition appears on the registry lead. However, in this case the A lead changes from 0 to 1 and a positive pulse is passed to stage 2, causing it to assume the condition 1 on its registry lead. A 1 on the B lead of stage 2 has a weight of 2, corresponding to the number of pulses impressed on the count input lead.

Count pulses 3 and 4 similarly cause further charges in state of the counter, resulting in the code being set up as illustrated in Fig. 1.

At this time assume that the control function receives information to the effect that the succeeding pulses should be subtracted from the binary sum already present on the counter stages. The control function therefore causes all the switches to make simultaneous connections to the B leads, in order that subtraction may be performed.

Count pulse 5 causes stage 1 to change state so that a 1 appears on its B lead, and since this is a change from 0 to 1, a positive pulse is passed to the second stage, and a 1 appears on the B lead of the second stage. This results in a positive pulse to the third stage and a 0 appears on its B registry lead. Since the B lead of the third stage changed from 1 to 0, no positive pulse is passed to the fourth stage and no further action follows. If we examine the binary code appearing on the registry leads of the counter, it corresponds to the code set up for the third count pulse. Thus the fifth count pulse has resulted in subtraction. Similarly, count pulses 6, 7 and 8 cause the binary counter stages to operate in such a manner as to revert to the starting condition.

Thus a binary counter, with switches so connected that the count input lead for the following stage may be switched from one output lead to another, may be made reversible so that it performs the arithmetic function of subtraction as well as addition.

3.0 Circuit Element of the RBC

3.10 Gating Function

With the aid of the neuron notation, a functional representation of the RBC is shown on Figure 2. From this it can be seen that one method of performing the required switching utilizes threshold two gates in which one of the enabling inputs is a step voltage from a Memory function, and the other is the positive pulse obtained from the binary counter lead. A transistor gated amplifier which is capable of performing this gating function has been developed and is described in Report 2.

3.20 Memory Function

The Memory function for the control of the gated amplifiers must provide the proper biasing potentials in the form of two step voltages. One of these is applied to all the A gates in parallel; the other to all the B gates in parallel. It must be capable of changing state, that is, reversing the

step voltages, under the control of operate and clear functions, usually designated Set 1 and Set 0. The operate or Set 1 will impress a step voltage on one input of the A gates of a value such that a positive pulse impressed upon the other input connection of an A gate will be transmitted to the next binary counter stage. The corresponding step voltage impressed simultaneously on the B gates will be of a value such that the B gates will be inhibited. In a similar fashion, a clear or Set 0 signal will result in inhibition of the A gates, and correspondingly the excitation of one of the B-gate leads such that a positive pulse impressed upon the other B-gate input connection will be transmitted to the next binary counter stage. A two-transistor stabilized bistable circuit which is capable of meeting these requirements has been developed.*

3.30 Counting Functions

3.31 For Usage as a Simple Binary Code Register

The requirements placed upon the counting stages where the RBC is to provide simple indication of the binary code registered upon the respective weighted stages, for example, by means of current indicating meters in the register leads, are as follows:

a. The stage must be capable of being triggered by the positive pulse output of a transistor gated amplifier. This requirement is not difficult to meet, since the amplifier output waveform may be adjusted as to duration and rise time, within limits, and the amplitude (25 to 35 volts) is sufficient to operate succeeding transistor circuits with large margins.

b. The voltage step function present upon each output lead must be capable, when differentiated, of providing a positive pulse of sufficient amplitude to trigger a transistor gated amplifier, when the latter is enabled and must not be large enough to trigger it when it is inhibited.

The two transistor stabilized binary counter stage* meets these requirements. In particular, the margins of operation in meeting requirement (b) are discussed in Appendix A.

3.32 For Usage as a Direct Arithmetic Register

The requirements placed upon the counting stages where the RBC, in addition to providing simple indication of

*"A Stabilized General Purpose Two Transistor Binary Counter Stage", by R. L. Trent.

the binary code registered upon the respective stages, must also provide a direct arithmetic indication of the binary code, are as follows:

a. Same as for 3.31a.

b. Same as for 3.31b.

c. The stage must also activate some additional gating function permitting a weighted current or voltage to be racked up in a supplementary indicating device.

4.0 Circuit Diagram of Reversible Binary Counters

4.10 Binary Code Register

The circuit elements as described in section 3 were interconnected as shown in Figure 3 to provide an experimental verification of the feasibility of application of the particular transistor circuits described. In addition, decoupling networks were incorporated in power supply leads, and in step voltage gating leads from the Memory function to the transistor gated amplifiers, to prevent common impedance coupling between collectors.

The circuit as shown operated successfully as a RBC at repetition rates up to approximately 30 kc. The input triggering pulse was about 1 microsecond long, of 3 to 6 volts amplitude.

4.20 Direct Arithmetic Register

The circuit elements as described in section 3 were interconnected as shown in Figure 4 to provide an experimental model of a 16-count RBC with arithmetic translation of the binary code. The circuit element building blocks making up this RBC are shown on Figures 5, 6, and 7, and consist of:

"A Stabilized General Purpose Two Transistor Binary Counter Stage", by R. L. Trent.

a. Gating Function (Figure 5) - The gates are transistor gated amplifiers operating under the conditions discussed in Appendix A.

b. Memory Function (Figure 6) - The memory function is a bistable circuit basically the same as the binary counter except that the input is arranged for setting (OPERATE and CLEAR) instead of scale of two counting.

c. Count Pulse Input (Figure 7) - The counting pulse is obtained from a one-shot multivibrator which impresses a positive triggering pulse of about 20 volts, 10 microseconds long on the count input connection of the first binary counter stage, under the manual control of a pushbutton.

d. Arithmetic Register - The register circuit consists of resistors, in series with the contacts of the relays, so chosen in value that the first relay when operated, places an increment of 1 ma on the indicating meter, the second relay an increment of 2 ma, the third and fourth relays increments of 4 ma and 8 ma respectively. Thus increments from 0 to 15 ma, corresponding to a 4 digit binary code, may be registered on the meter.

The experimental model, illustrated in the photograph of Figure 4-8, has been operated manually and in addition has been operated dynamically from an external triggering source up to approximately 320 cycles per second. It has been observed that variations in the relay contact characteristics play a large part in determining the operate time of a relay. In general, consistent operation at manual pulsing rates has been observed and this method of operating the relays directly from the binary counter stages may be practical.

Appendix A

The purpose of this appendix is to describe and illustrate the margins of operation of a transistor gated amplifier when used as a gated interstage between two stages of binary counters. The general characteristic of the gated amplifier is shown on Figure 9 in which the negative emitter current scale is enlarged for illustrative purposes. If we assume that the transistors used in the binary counter stages and in the Memory function will have I_{co} values between 1 ma and 3 ma, this assumption, together with a knowledge of the circuit constants, will enable the calculation of the variations in ON and OFF condition collector voltages. This in turn gives a direct indication of the step voltage output from the Memory, which acts as the sensitizing-inhibiting means for the gated amplifier, and also the voltage output of the positive pulse from the counting stages.

A. OFF Condition Collector Voltage Variations

The calculation of OFF condition collector voltage variations may be made as illustrated on Figures 10 and 11. Experimental verification of these calculated values has been made with excellent agreement. We may therefore assume that the variations lie between -30 and -35 volts.

B. ON Condition Collector Voltage Variations

The calculation of ON condition collector voltage variations is somewhat more complex since it involves both emitter and collector circuit currents. However, this calculation has been made and experimentally verified resulting in variations between -11.2 and -15.6 volts.

From the above results we may therefore calculate the minimum and maximum pulse amplitude obtainable from a collector. These amplitudes are +14.4 and +23.8 volts respectively.

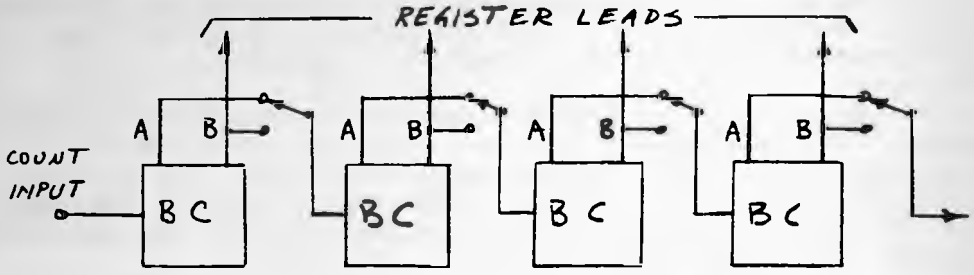
On Figure 9 then, we may indicate the range of variation between the ON and OFF condition biasing voltages for the emitter load lines. These may be translated, by plotting the load lines, into variations in triggering voltage amplitude required.

The conditions which must be fulfilled to prevent ambiguity are:

(1) The lowest triggering voltage obtained from a binary counter stage (calculated as +14.4 volts) must be capable of triggering the threshold two regenerative gate under all conditions of biasing set up by the Memory step voltage when transmission is desired. (-11.5 ± 2.0 volts)

(2) The highest triggering voltage obtained from a binary counter stage (calculated as +23.8 volts) must not trigger the threshold two regenerative gate under any conditions of biasing set up by the step voltage control function when it is desired to inhibit the gate. (-27.5 ± 2.5 volts)

Since both of these conditions are met with margins of about 1 volt under the most adverse conditions to be expected the usage of the transistor amplifier as a gate is a feasible solution. Obviously, condition (2) may be improved by the insertion of a diode limiting network to clip the triggering pulse amplitude at some level, say 20 volts.



COUNT PULSE	A	B	A	B	A	B	A	B
—	(1)	0	(1)	0	(1)	0	(1)	0
1	(0)	1	(1)	0	(1)	0	(1)	0
2	(1)	0	(0)	1	(1)	0	(1)	0
3	(0)	1	(0)	1	(1)	0	(1)	0
4	(1)	0	(1)	0	(0)	1	(1)	0
	SWITCHES CHANGED TO B LEADS							
5	(0)	1	(0)	1	(1)	0	(1)	0
6	(1)	0	(0)	1	(1)	0	(1)	0
7	(0)	1	(1)	0	(1)	0	(1)	0
8	(1)	0	(1)	0	(1)	0	(1)	0

Fig. 1. Functional Diagram of Reversible Binary Counter

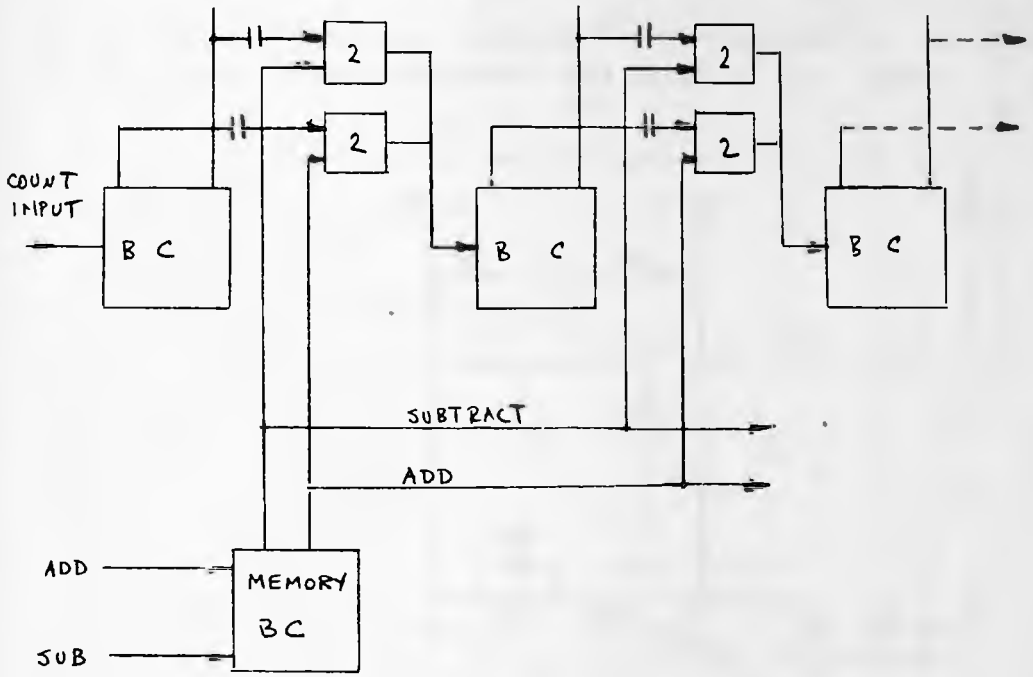


Fig. 2. Functional Representation of
Reversible Binary Counter Using Neuron Notation

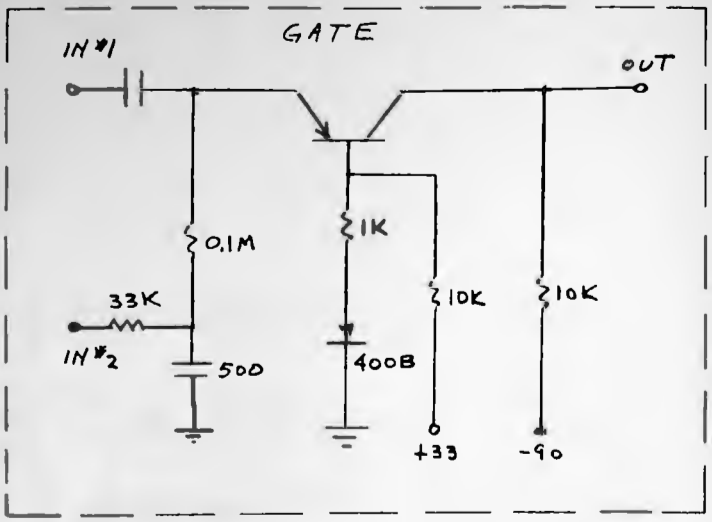


Fig. 5. Circuit for Transistor Gated Amplifier

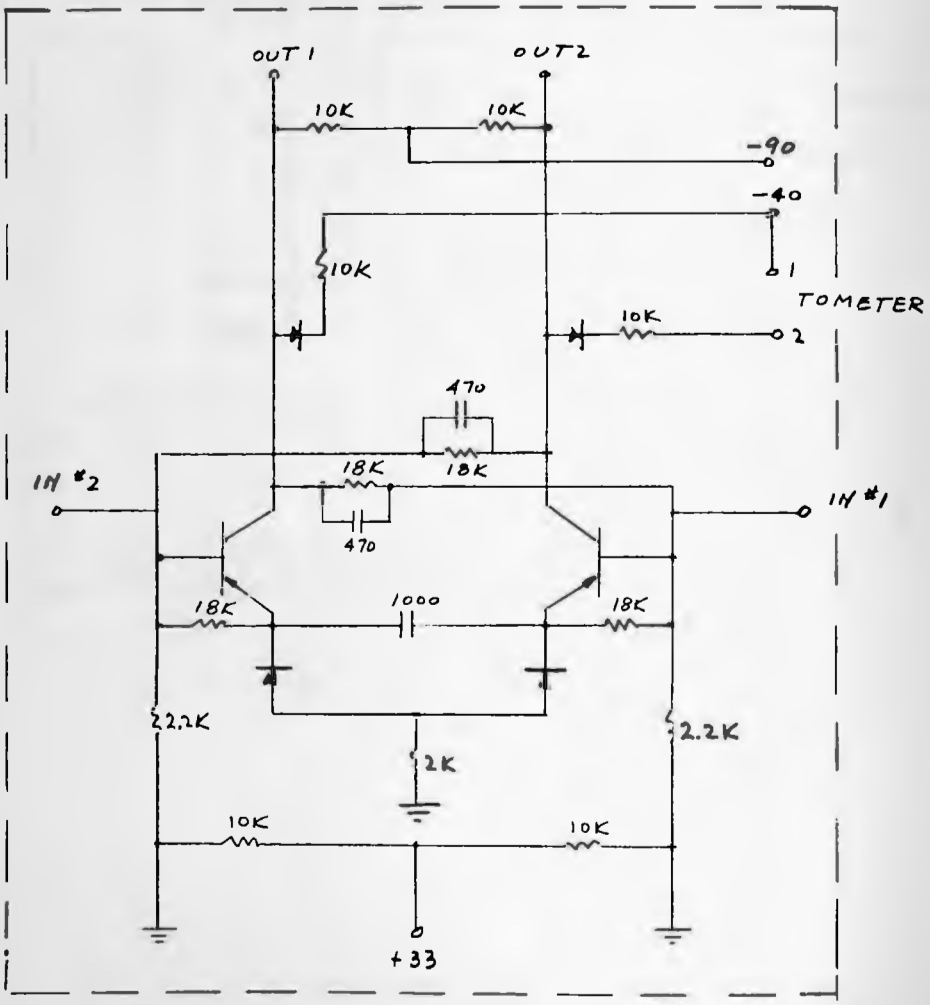


Fig. 6. Bistable Circuit for Memory Function

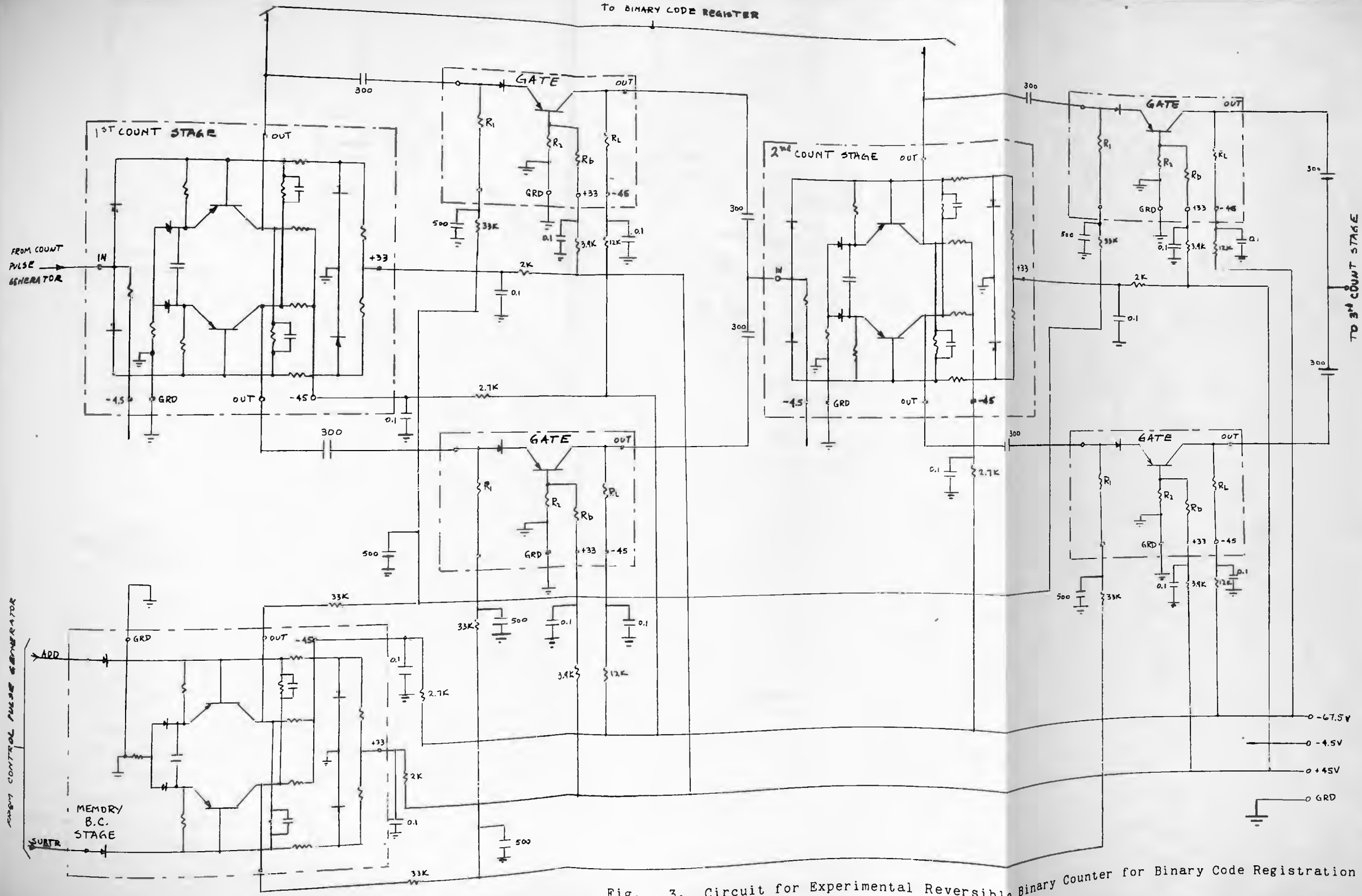


Fig. 3. Circuit for Experimental Reversible Binary Counter for Binary Code Registration

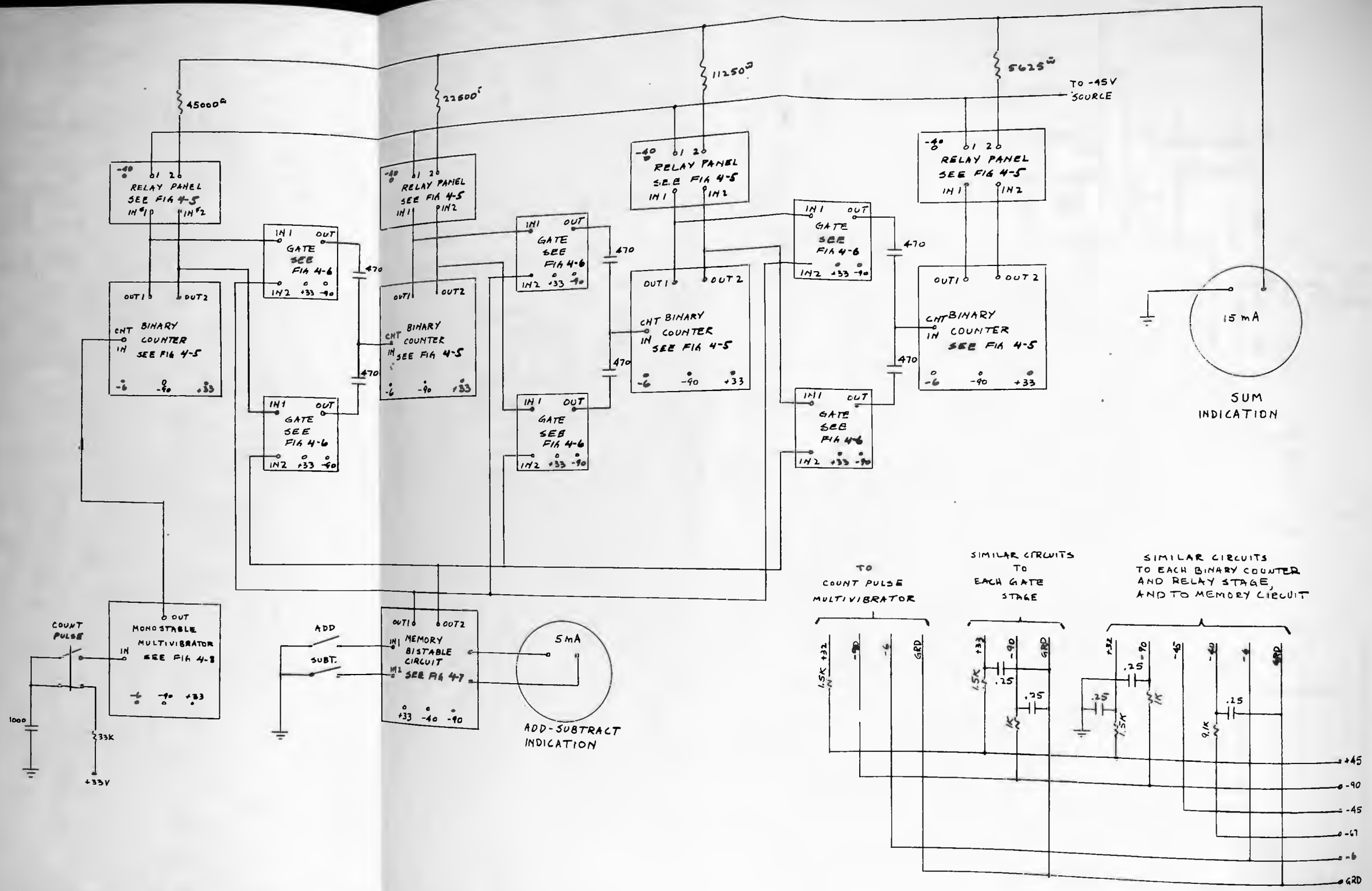


Fig. 4. Circuit for Experimental Reversible Binary Counter Arithmetic Register

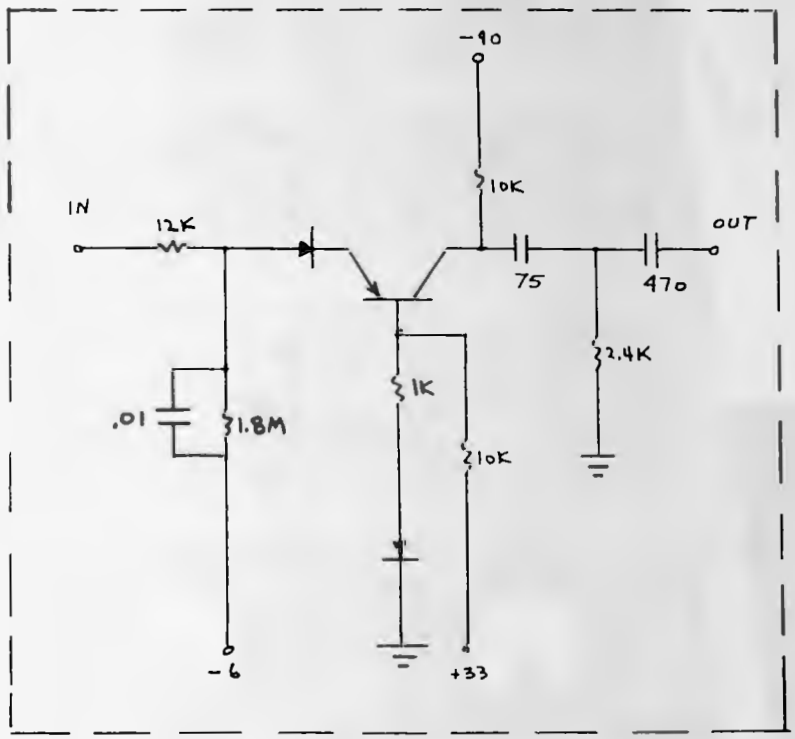


Fig. 7. Circuit for Monostable Multivibrator for Count Pulses

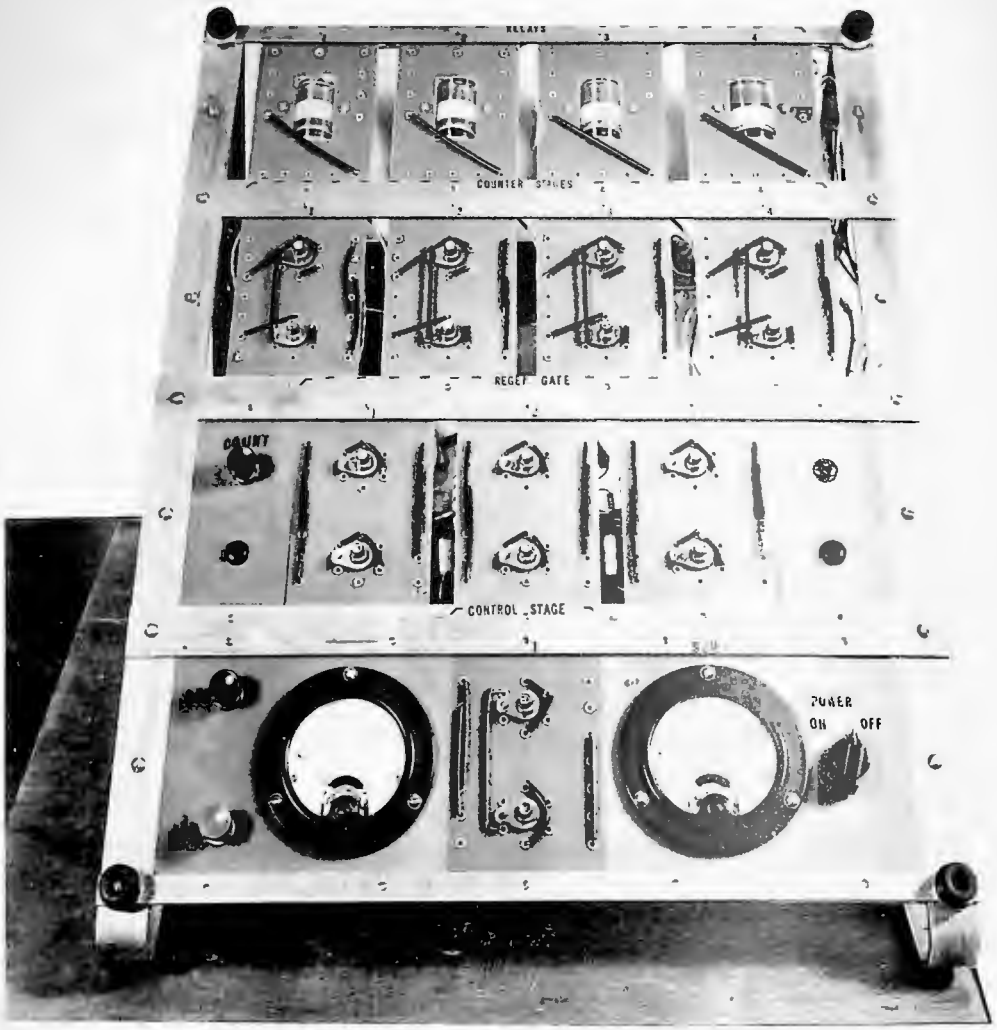


FIG. 8 EXPERIMENTAL MODEL OF BINARY COUNTER

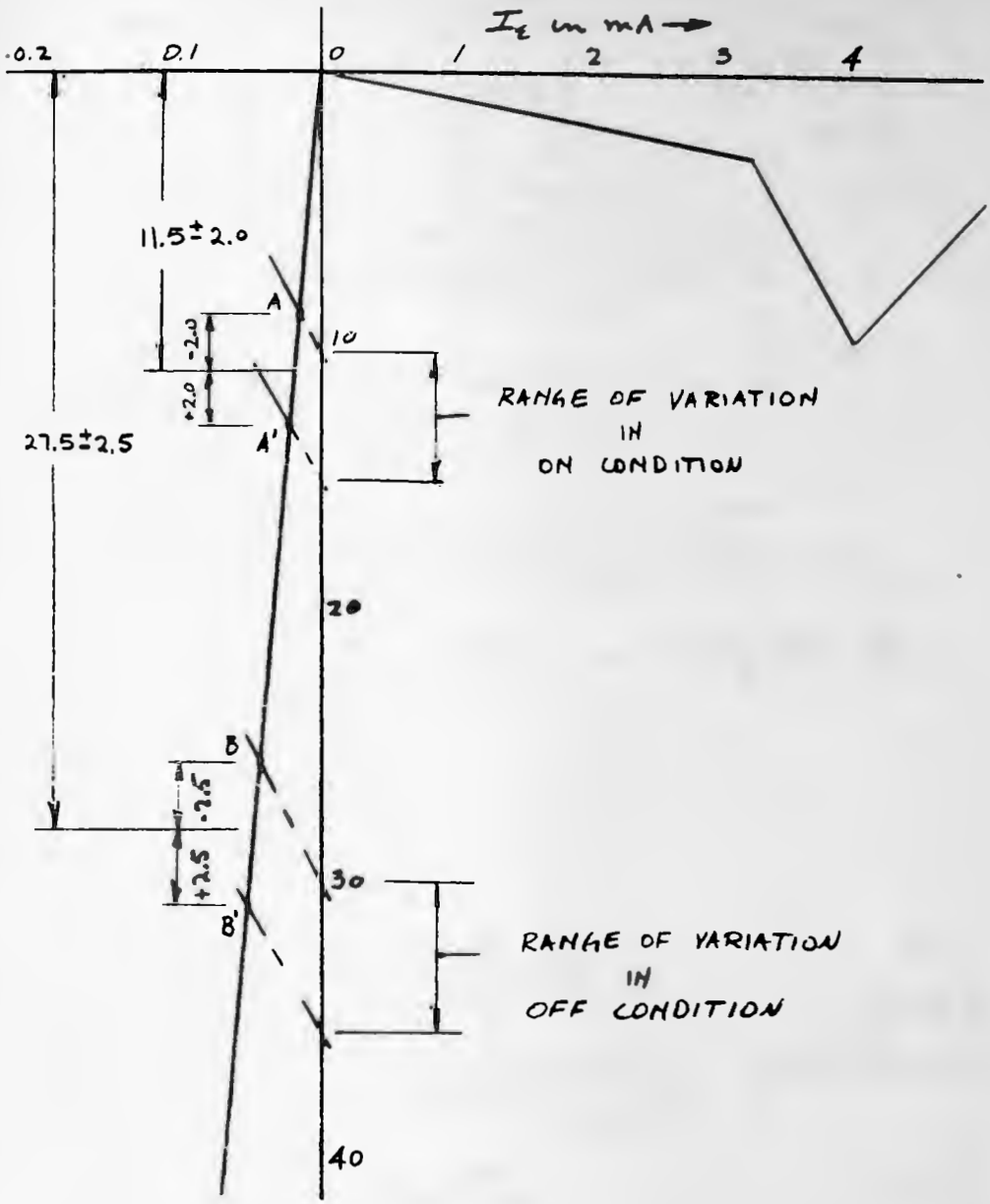


Fig. 9. Operating Margins for
Transistor Gated Amplifier Switching-Type Gate

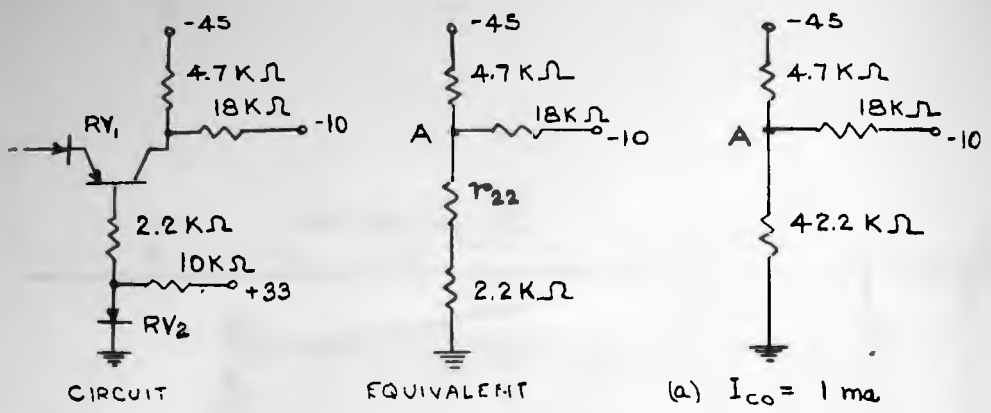


Fig. 10 (Also see Fig. 11)
Calculation of OFF Condition Collector Voltage for I_{co} = 1 ma

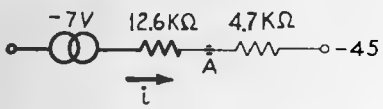
Assumptions

(1) RV_1 and RV_2 are perfect switches: RV_1 open, RV_2 closed.

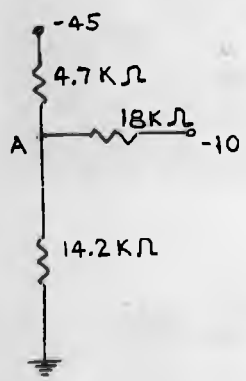
(2) The ON condition base voltage is constant at -10 volts.

(3) Variations from 1 ma to 3 ma in I_{co} result in variations from 40 KΩ to 12 KΩ in r_{22}.

Replacing 18 KΩ and 42.2 KΩ in (a) with equivalent generator and internal resistances results in the circuit shown below, which permits computing the voltage at A.



Then i = 2.2 ma, and the voltage at A is the sum of -27.8 volts and -7.0 volts, a total of -34.8 volts.

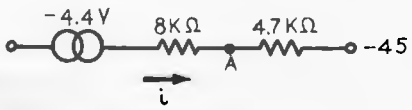


(b) I_{co} = 3 ma

FIG 11

Fig. 11 (Also see Fig. 10)
Calculation of OFF Condition Collector Voltage for I_{co} = 3 ma

Replacing 18 KΩ and 14.2 KΩ in (b) at left with equivalent generator and internal resistances results in the circuit shown below, which permits computing the voltage at A.



Then i = 3.2 ma, and the voltage at A is the sum of -25.6 volts and -4.4 volts, a total of -30 volts.

Thus the OFF condition voltage varies between -30 volts and -34.8 volts.

FOUR STAGE BINARY COUNTER

R. L. TRENT

Produced under contract W 36-039 sc-44497

FOUR STAGE BINARY COUNTER

A demonstration model of a four stage binary counter, employing packaged transistor circuits, was constructed. The model was prepared to illustrate the practicability of a packaged M1745-1 Binary Counter stage driving another stage without intervening amplification.

The counter consists of 4 binary counter stages, each of the last three being driven by the preceding stage, and the first being driven by a transistor pulse generator. Accordingly, the counter has 16 states (counts 0-15). The state of each stage, and consequently of the counter as a whole, is indicated by a set of neon lamps.

The pulse input source consists of a packaged free-running multivibrator (M1731-1 Regenerative Gated Amplifier) providing pulses at approximately 20 cps. Visual indication of the binary counter states is provided by neon lamps. The model is operated from unregulated rectifiers of -90 volts and +33 volts respectively. A voltage divider is used on the -90 volt supply to obtain the -45 volts required for the M-1745-1 Binary Counter stages.

The complete circuit of the model is shown on Figure 1.

Circuit Description

The M1731-1 Regenerative Gated Amplifier is caused to operate free-running by biasing the emitter load resistance positive such that the load line will intersect the characteristic in the negative resistance region. The slow rate of 20 cps is achieved by insertion of a 1.0 megohm resistance in series with the 0.1 megohm emitter load provided in the package, and adding the 0.25 mf capacitor from emitter to ground.

The output pulse from the free-running MV is connected to the first binary counter stage through a simple diode steering circuit designed to pass the pulse alternately to the bases of the two transistors depending upon the prior state of the binary counter.

The OUT 2 leads of all the binary counter stages are connected to neon lamp indicators. The circuit is arranged so

that when a collector is ON (at about -10 volts) the voltage drop across the lamp is sufficient to initiate the glow discharge. When ignited the lamps have a fairly constant sustaining drop across the discharge in the order of 55-60 volts. When a collector changes from ON to OFF (at about -30 volts) almost the entire change is impressed across the lamp, thus decreasing the voltage drop across the lamp below the sustaining value, and extinguishing the glow discharge.

In order to cause the binary counters to act as pulse adders, the output to a following stage is taken from the OUT 1 collectors in each stage. The steering circuits interposed between binary counter stages have the advantage of inherently suppressing the small positive pulse which appears at the collector of an ON unit when it is triggered OFF, thus avoiding ambiguity.

Results and Conclusions

The model has operated successfully in several demonstrations, and has also been in continuous operation for several 48 hour periods without failure. The photographs of Figure 2 and 3 illustrate package and wiring sides of the model.

It has been shown that M1745-1 Binary Counters which meet the test requirements may be successfully cascaded without intervening amplification.

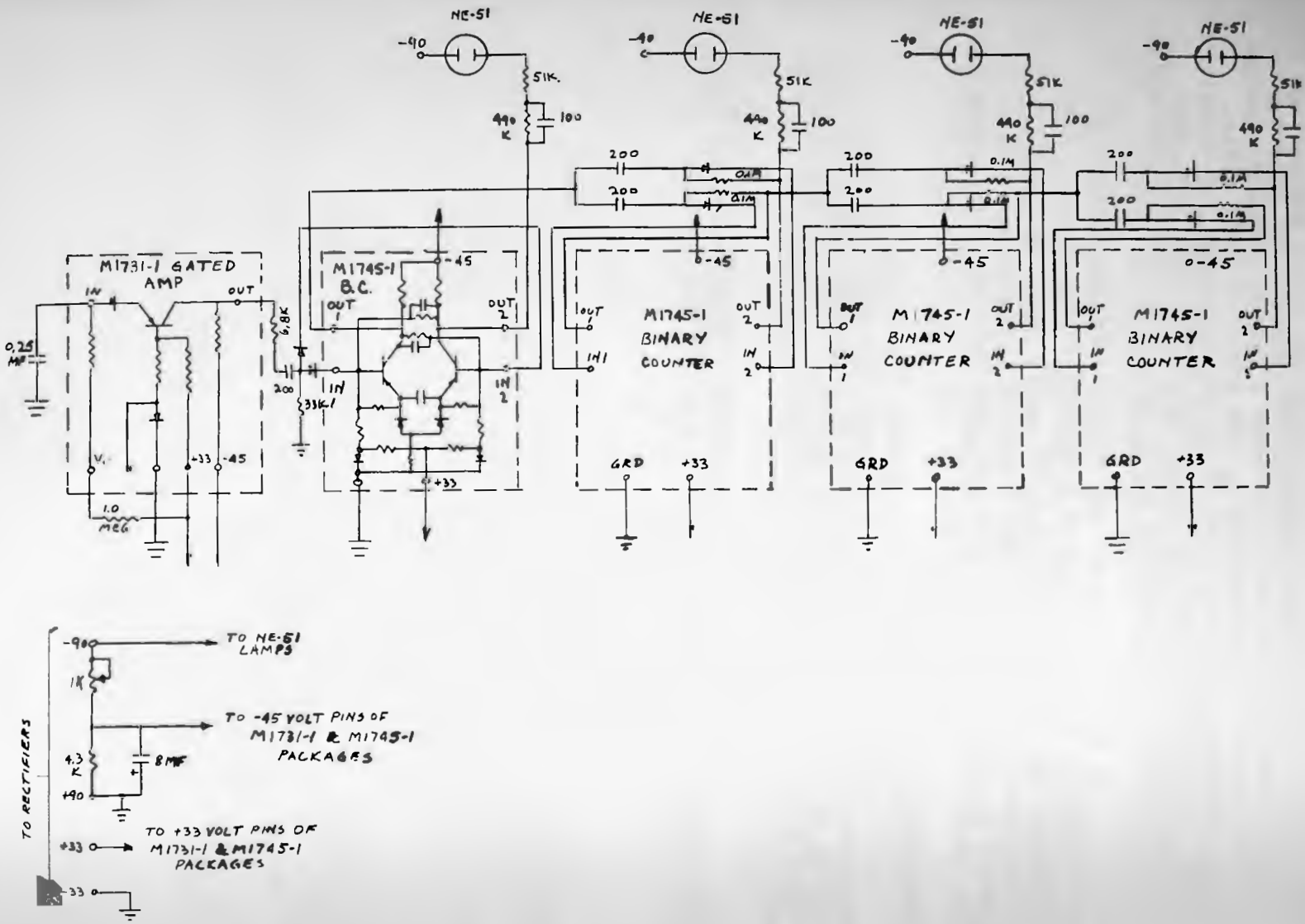


Fig. 1. 4-Stage Binary Counter Using M 1745-1 Binary Counter Packages

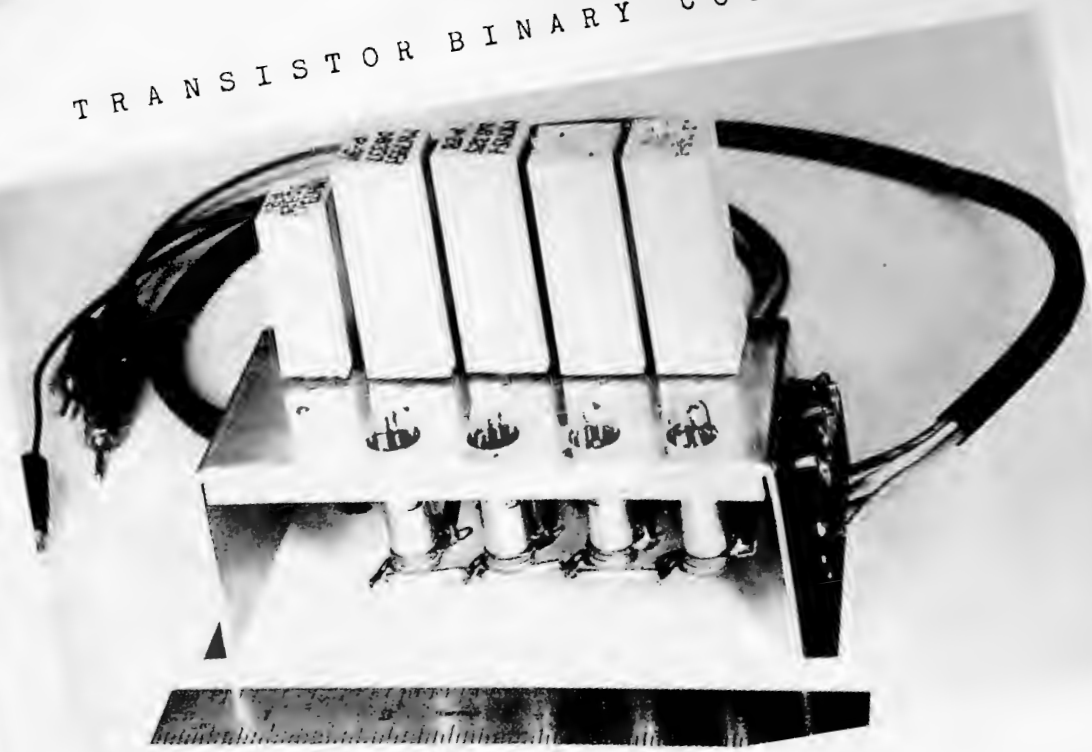


FIGURE 2 - FRONT VIEW

FIGURE 3 - WIRING VIEW

NO.

DIODE GATES

L. W. HUSSEY

Produced under contract W 36-039 SC-44497

DIODE GATES

This report and the one on "Anticoincidence or 'Not And' Gate" constitute a preliminary investigation of devices which should satisfy most of the basic gate requirements of the system. They are prospective "And" or "Inhibited" gate and "Not And" gate packages.

Both the transmission type and the switching type gate (discussed here) meet several of the probable system requirements with ample margin; for example -

1. Discrimination

The ratio of output voltage when the gate is enabled to that where the gate is disabled can be made 50 db or better. Since 30 db is probably adequate, this puts very mild limits on thermal and manufacturing variation of the diodes.

2. Levels

Input and output voltages of the order of 20 volts should be adequate for the system and are practical for the gates.

3. Impedances

The 2K load used in the experimental set up is about average for the system now projected. The gates are apparently operable over a range of terminations from 1K to 10K and presumably over a wider range if p-n junction diodes are used. This covers all but rather extreme requirements.

4. Band Width

No quantitative information has been obtained on band width, but the band width of the standard germanium

diodes seems to have a liberal margin over requirements and from what evidence is available, the p-n junction diodes should also be adequate.

The limitations of the gates are:

Transmission Type

The control pulse "pedestal" effectively limits us to a 2/1 range of useful output and adds a rather critical minimum triggering requirement to devices following the gate. This could be avoided by auxiliary circuits but involves considerable complication.

Switching Type

The switching type gate has no pedestal difficulties but when working into and out of equal impedance, has a loss of about 6 db in voltage. This looks much less difficult from an engineering viewpoint than the pedestal.

A gate which has not been investigated, is the "Transistor Gate". This is essentially a transistor amplifier with a gate of the type described here built into the base. Our present feeling is that a separate buffer amplifier package is preferable to combining the two in a single unit.

The "Not And" gate described could be a collection of packages rather than a single unit. In any case, requirements and specifications can better be discussed in terms of the parts.

Switching Type Gate

In a previous investigation a transmission type diode gate was studied both analytically and theoretically. The conclusion was that this gate is a valuable device but for switching and pulsing techniques it has one serious limitation; the control pulse produces an output of the same order of magnitude as the signal or, in other words, the signal output is superimposed on a "pedestal".

Two-control switching type gate is discussed here which eliminates this defect. (Figure 1) This circuit has characteristics quite different from the transmission gate but seems to perform the essential functions (as far as switching circuits are concerned) very well.

Here we have three diodes $\{D_k\}$ as shown, a high impedance or constant current, bias $\{i_b\}$ and two control inputs.

R_1 and R_2 are the generator impedances of the control sources and V_1, V_2 the voltages which are actually impressed on the diodes.

If either V_1 or V_2 is sufficiently negative the corresponding diode, D_1 or D_2, is conducting and V_b is negative. Thus i_b flows through D_1 or D_2 and the load is isolated from all potentials by the high impedance D_3. If both E_1 and E_2 are suitably positive, D_1 and D_2 are biased high impedance, V_b is positive, and i_b flows in R_L giving an output. The two controls (and additional ones) can be used either for enabling or disabling the gate. For example if E_1 and E_2 are both negative biases, positive signals are necessary on both controls to produce an output. In this case we have an "And" or "Coincidence" gate. If E_1 is negative and E_2 positive, control 2 is an inhibiting control. A positive signal on 1 produces an output, but a simultaneous negative signal lets i_b flow through D_2 and holds the gate disabled.

One can readily set up for idealized diodes (switching between zero and infinite impedance) the characteristics and requirements. (A better approximation and an analysis is appended.)

The output when the gate is enabled is independent of signal amplitude:

$$V_L = R_L \, i_b$$

For disabling the gate, the bias must satisfy:

$$E_1 < -R_1 i_b$$

This insures that V_b is negative and D_3 non-conducting.

The bias in the inhibiting control must satisfy:

$$E_2 > + R_L \, i_b$$

If this bias is smaller, V_L and V_b cannot rise to maximum output without D_2 becoming conducting. A smaller bias produces a limiter action, the output rising to the bias value and stopping there.

The signals have requirements similar to those on the biases. To enable the gate, a signal must satisfy:

$$E_1 > R_L\, i_b,$$

and to disable it the inhibitor voltage must satisfy:

$$E_2 < -R_2 i_b.$$

These requirements show the main defect of the device. In return for eliminating the "pedestal" of the transmission gate we must take a voltage loss of around 6 db when the gate is working between equal impedances. For example, if $R_1 = R_L$ the bias must be,

$$E_1 < -R_L\, i_b$$

and the maximum signal must be,

$$E_1 > + R_L\, i_b$$

Thus the signal swing must be greater than 2 $R_L\, i_b$ in order to obtain an output $R_L\, i_b$.

Figure 2 illustrates the behavior of the gate. The diodes are 400A's and the load impedance 2000 ohms. Two cases are shown, the normal operation ($V_2 = 15$) and operation when the output is limited by a small inhibitor control bias ($V_2 = 5$). The crosses and circles are experimental values and the solid lines computed curves based on the assumption that the diodes switch from 200 ohms to 1.25 megohms. At around 10 volts signal the discrimination is of the order of 50 db in voltage.

The new p-n junction diodes are promising for this application. They can be made with lower forward resistance and higher back resistances. They should give us greater discrimination and a much wider voltage range.

Analysis

As a matter of convenience, the circuit of Figure 1 may be expressed in terms of voltages and conductances

$$(G_1 + G_2 + G_3)\, V_b - G_3 V_L = i_b + G_1\, E_1 + G_2\, E_2$$

$$- G_3\, V_b + (G_L + G_3)\, V_L = 0$$

In these equations G_1 is the conductance of the series combination of R_1 and the series diode; G_2 is the similar combination for R_2; G_3 is the conductance of the diode in series with the load and $G_L = \frac{1}{R_L}$.

The diodes are assumed to switch between two values $g = \frac{1}{R}$ and $G = \frac{1}{r}$ such that

$$g \ll \frac{1}{R_K} \ll G \qquad K = 1, 2\ L$$

The general solution is

$$V_b = \frac{i_b + G_1 E_1 + G_2 E_2}{G_1 + G_2 + \frac{G_3 G_L}{G_3 + G_L}}$$

$$V_L = \frac{G}{G_L} V_b = \frac{R_L}{R} V_b.$$

There are several cases, depending on the magnitudes of V_b. The cases may be considered in connection with Figure 2 . Starting with the disabled case we have

Case 1 $E_2 > 0$, $E_1 < V_b < 0$

Here

$$G_1 \doteq \frac{1}{R_1},$$

$$G_2 \doteq g,$$

$$G_3 = g,$$

$$V_b = \frac{i_b + \frac{1}{R_1} E_1 + g E_2}{\frac{1}{R_1} + g + \frac{g G_L}{g + G_L}} \doteq R_1 i_b + E_1,$$

$$V_L = \frac{R_L}{R}(R_1 i_b + E_1)$$

This stops when $V_b = 0$ or $E_1 - = R_1 i_b$. On Figure 2 this point is $V_1 = -1$, $V_L \doteq 0$.

Case 2 $V_b > E_1 > -R_1 i_b$, $V_b > 0$.

$$G_1 \doteq \frac{1}{R_1}$$

$$G_2 \doteq g$$

$$G_3 = G$$

$$V_b = \frac{i_b + \frac{1}{R_1} E_1 + g E_2}{\frac{1}{R_1} + g + \frac{G G_L}{G+G_L}} \doteq \frac{R_1 i_b + E_1}{1 + \frac{R_1}{R_L+r}}$$

$$V_L = \frac{R_L}{R_L + r} V_b$$

When E2 is sufficiently large this phase continues until $V_b = V_1$ (V_1 = 11 volts on Figure 2). At this point the current through R_1, D_1 must be zero so

$$V_b = E_1 = \frac{R_1 i_b + E_1}{1 + \frac{R_1}{R_L+r}}$$

$$\therefore E_1 = (R_L + r) i_b$$

Case 3 $E_1 > (R_L+r) i_b$, $V_b > 0$, $V_b < V_i$

$$G_1 \doteq G_2 \doteq g$$

$$G_3 = G$$

$$V_b = \frac{i_b + g(E_1 + E_2)}{2g + \frac{1}{R_L + r}} \doteq (R_L + r)\, i_b$$

$$V_L = \frac{R_L}{R_L + r}\, r_b = R_L\, i_b$$

In case E_2 is not greater than $R_L\, i_b$, V_b becomes greater than E_2 before full output is attained, as illustrated in the case $V_Z = 5$ on Figure 2 . This gives two different combinations, first with all diodes conducting and second with the signal path blocked but the other two diodes conducting. These cases are readily analyzed in the same manner as the others.

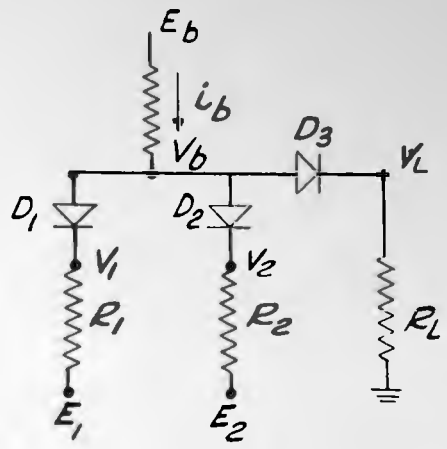


Fig. 1 Diode Gate

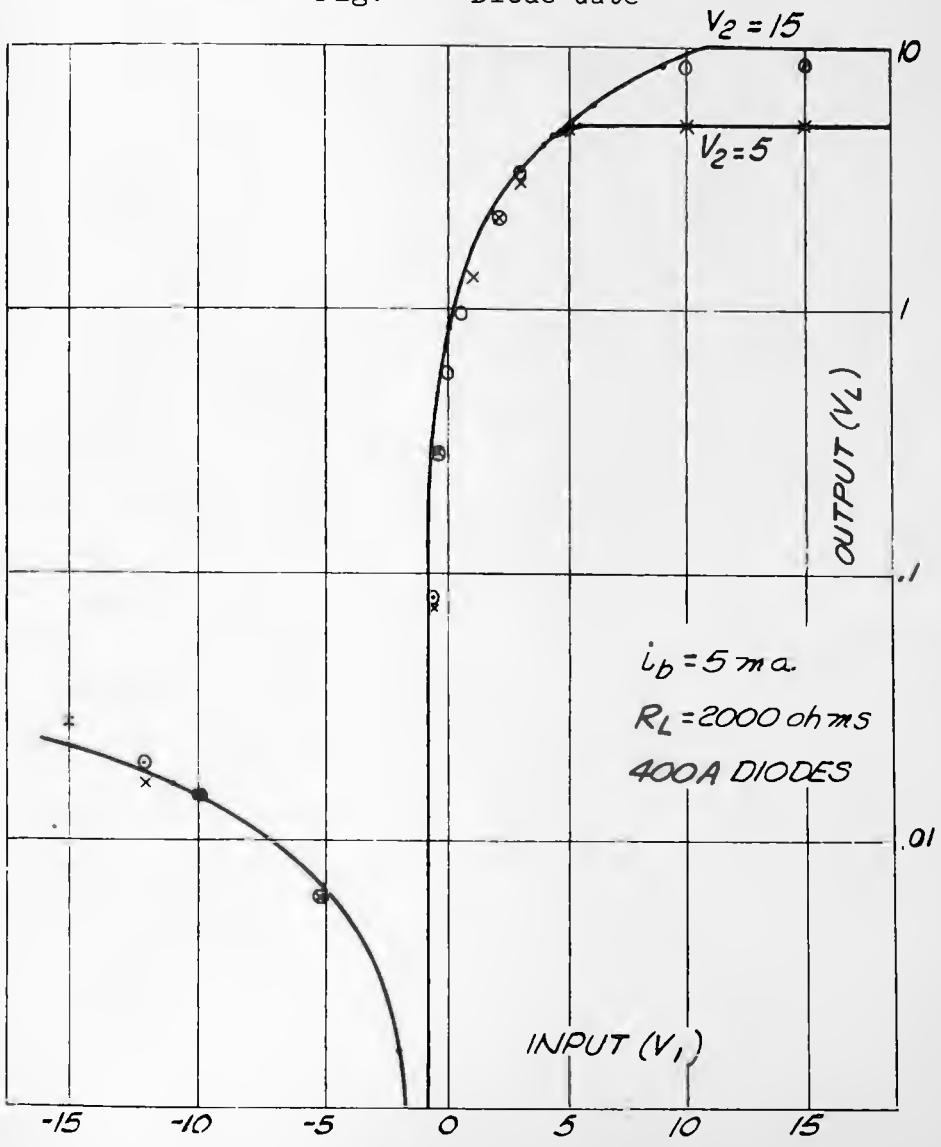


Fig. 2 Diode Gate Characteristic

A N T I C O I N C I D E N C E

O R

" N O T A N D " G A T E

L. W. HUSSEY

Produced under contract W 36-039 sc-44497

ANTICOINCIDENCE OR "NOT AND" GATE

An anticoincidence or "Not And" gate is a gate with two inputs. It produces an output when a single enabling input is impressed on either input but no output where there is no input or when there are two coincident inputs. The usual serial adder system uses this type of gate.

Some study has been made of these gates - in particular, considering the difficulties that the lack of exact coincidence of the two inputs may cause. The basic unit used (Figure 1A) is a simple switching type diode gate with one enabling control and one inhibiting control. This unit has previously been discussed. It serves our purpose very well except for one inconvenience - the inhibitor control pulse must have opposite polarity to the enabling control pulse. There would be some advantage to being able to operate both controls from pulses of the same polarity.

In devices such as computers where the control pulses may be made coincident with a high degree of precision an anticoincidence gate can be made with two such units. The gates are connected with their outputs tied together. Signal 1 enables gate 1 and signal 1 (inverted) inhibits gate 2. Similarly signal 2 enables gate 2 and inhibits gate 1. Thus either signal, by itself, produces an output but the two together block each other. This obviously assumes such accurate coincidence of the control pulses in time and form that no pulse could produce an output before a "coincident" pulse inhibited the gate.

Since this high precision of synchronization is difficult or impossible in some cases, an anticoincidence gate has been developed and tested which defines a coincidence interval and produces no output when the signal pulses are within such an interval. Figure 1B shows the circuit schematically and Figure 1C shows the time sequence of the pulses at various points in the system. In this system signal 1 is impressed on a monostable flip-flop (FF1) which puts out a negative pulse of duration 2τ. This pulse is impressed on the inhibiting control of gate 2 (G 2). Signal 1 is also passed through a delay network (D 1) which delays it a time τ

and impresses it on the enabling control of gate I (G 1). Signal 2 is similarly impressed on flip-flop 2 (FF2) and delay 2 (D 2). A study of the time sequence diagram (Figure 22C) shows that if the time interval between the two signals is less than τ, they will effectively block each other. In Figure 22C the dashed pulses show the outputs which each signal would have produced in the absence of the other.

In the actual embodiment FF1, FF2 D1 and D2 were all monostable flip-flops of the type described by A. E. Anderson*. The delayed pulse was obtained by differentiating the trailing edge of the flip-flop output. The gates G1 and G2 used W.E. 400A diodes and operated into a 2000 ohm load. The system operated and gave evidence that the device is a practical one. It had certain limitations due to the type of flip-flop. This flip-flop will produce only a positive pulse with steep leading and trailing edges and that pulse is far from flat topped. For this reason the inhibiting pulse is necessarily positive so the signal pulse must be negative. The imperfections of the wave form of the inhibiting pulse also produced an undesirable transition interval during which a partial pulse appeared at the output.

For this purpose a monostable flip-flop package should produce a pulse with steep edges and a flat top and it is desirable that pulses of either polarity be obtainable from it.

An alternative circuit, in cases where the operation of the system is under the control of a "shift pulse" is shown (Figure 2). Symmetric flip-flops which simultaneously produce a positive and a negative potential are needed for this. This sort of bistable unit always presents, at two different terminals, one of which can be used for enabling one control of a gate and the other simultaneously inhibiting another gate. The gates have three controls, two enabling and one inhibiting. There must be enabling signals on both the enabling controls and no inhibiting signal for the gate to produce an output. When a signal is present the flip-flop impresses an enabling potential on one gate and a disabling potential on the other.

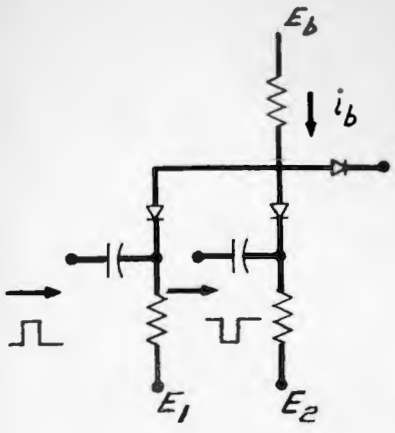
The sequence of operations is as follows. The shift pulse is impressed on one enabling control of each gate. If there is no signal present, neither gate has its second enabling control activated so there is no output; if there are two signals present, both gates are inhibited; when only one signal is present, one gate is uninhibited and has both enabling controls activated so that an output is produced. After a brief and

- -

*"A Stabilized Transistor Delay and Switching Circuit", by A. E. Anderson.

non-critical delay the shift pulse operates to set up the next pair of signals.

The flip-flops may be the output units of shift registers which are operated by the shift pulse or the shift pulse may enable gates to let the next pair of signals operate the flip-flops. In case of synchronous operation the flip-flops would be reset by the next pair of signals without the aid of the shift pulse.



BASIC GATE UNIT

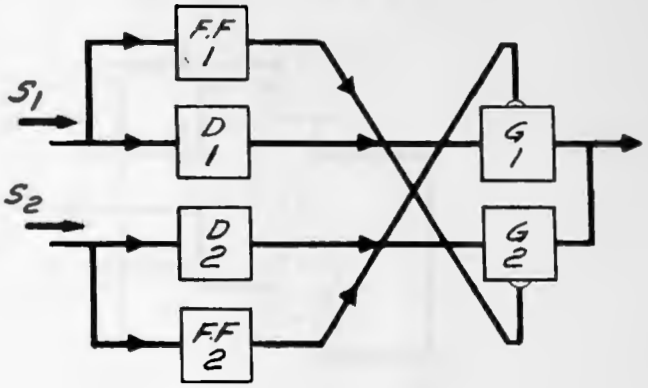
$E_1 < 0$ CONTROL INPUT

$E_2 > 0$ INHIBITOR INPUT

FIG. 1A

BASIC "NOT AND" GATE

FIG. 1B



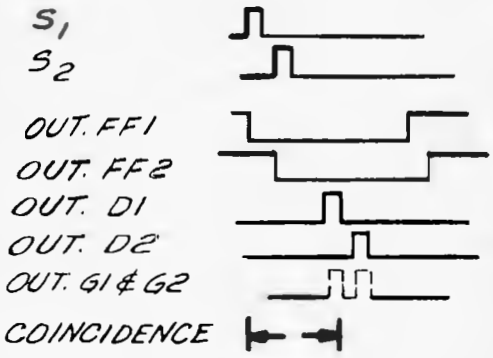


FIG. 1 C

Fig. 1 Basic Gate Unit and "Not And" Gate

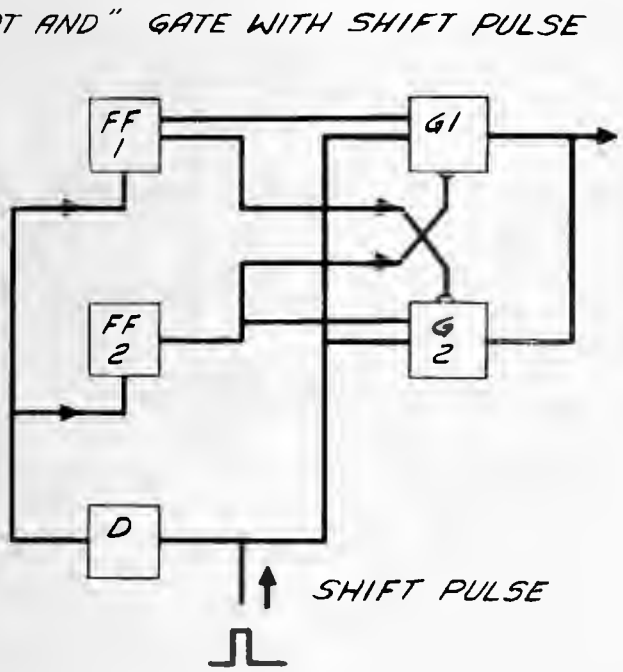


Fig. 2 "Not And" Gate with Shift Pulse

STORAGE AND ADDITION OF BINARY NUMBERS

J. R. HARRIS

Produced under contract W 36-039 sc-44497

STORAGE AND ADDITION OF BINARY NUMBERS

0.0 Problem

This report describes means for storing, shifting, and adding binary numbers. The immediate application is to fill a need for accepting two large binary numbers presented instantaneously in space division, adding these numbers, and passing out the sum, digit by digit in time sequence. The design is based on the use of transistors.

0.1 Evaluation of Work

Work on the necessary circuits and semiconductor devices was undertaken because such work was judged to show promise of giving smaller equipment with longer life and less power drain than would be practicable with a design based on electron tubes. An adder and nine stages of shiftable (shift) register have been built and operated. It appears that such equipment has very good reliability and the desired small size and power drain. Life performance has not yet been fully determined.

The functions of registering (remembering or storing) digits, shifting digits, and adding numbers are widely used. It appears that equipment like that described herein could be used with similar advantages in systems other than the immediate one.

0.2 Outline of Report

Part 1 of this report covers the adder and part 2 covers the shift register. Part 2 is divided into sections covering; the shift register as a whole, the one-bit register, the pulse amplifier and amplifier-delay, and interstage coupling networks. The pulse amplifier, amplifier-delay, and interstage coupling are used to shift all digits one stage to the right in the shift register in response to the program generator. "To the right" is used in the same sense as with numbers written on paper. Each digit moves into a spot previously occupied by a less significant digit.

A block diagram of the registers and adder is given in Figure 1.

1.0 Adder

1.1 Outline of Treatment in Report

The adder work will be treated by covering first the broad choices of possible adders and then narrowing down to the adder as finally built.

1.2 Choice of Serial Addition

The decision to concentrate on serial addition came about as a result of a comparison of the equipment requirements for parallel and serial addition. On the basis of this it appeared that serial addition would require less complex comparison equipment than parallel addition to perform the same general function in this transistor system.

It is interesting to note that serial addition can correspond very closely to adding by hand on paper. One first writes down the numbers to be added, corresponding to parallel insertion of digits into the registers. Then three operations are made for each pair of digits: (1) add the two input digits and the previous carry (and write down, or store, the sum) (2) find the new carry, and remember it, (3) shift attention leftward one digit space. The electronic serial adder follows this sequence for (1) and (2). In the case of (3), instead of shifting attention leftward, the attention of the adder is kept (wired in) at a certain place and the input numbers are shifted one digit space to the right. In the two-step, or Boolean, adder, step (1) is divided into two steps.

1.3 Requirements on the Serial Adder

To make the adder problem concrete and useful for the transmission system in point, the following requirements were placed on the adder: The adder should be capable of taking digits of the two input numbers from a pair of transistor one-bit registers and putting the sum digit into a third transistor one-bit register. Any program pulses necessary for the operation of the adder should be of a kind that can be produced readily by transistor circuits. The adder must accept new input digits presented at the rate of 600 pairs per second, and all operations for adding and carrying should take not over 200 microseconds. These requirements on frequency response are readily met by transistors.

1.4 Possible Adder Arrangements

Several kinds of adder operation are possible, and the broadest choice will be considered first.

1.5 Timing

The most basic alternatives are whether the timing of the sequence of operation is controlled by circuits within the adder itself or whether it can be controlled by an

external program generator. It may be that the sequence timing is controlled by passive circuits within the adder itself. Such an adder is called "free-running" and operates at its own characteristic rate. It puts close tolerances on the times when digits have to be presented at its input and also on the times when sum digits are available. In the free-running adder to have any tolerance at all on the closeness to perfect time coincidence of the two input digits, one needs complex arrangements such as pulse stretching and delay within the adder. Close tolerances on timing make it troublesome to synchronize the adder with the rest of the system. For this reason, and to simplify the adder itself, it may be preferable to cause the adder sequence to operate under control of a master program generator. This more tractable version may be called a "program-controlled" adder. The advantages of program control seem to apply to both two-step (Boolean) and one-step (Von Neumann) adders.

1.6 Storage

In a program-controlled adder, arrangements can be made to store digits permanently until acted upon. An example of such storage is a simple bistable flip-flop, d-c coupled circuit. It is conceivable that a primarily a-c coupled circuit could be used, say an oscillator that would stay quiescent to store a zero and stay oscillating to hold a one. Here still there is considerable dependence on d-c characteristics.

Permanent storage may not always be essential in a program-controlled system. In the system in question, most parts are not left undisturbed longer than 1.67 milliseconds, so that any storage longer than this minimum is essentially equivalent to permanent storage. Under such circumstances the memory can be entirely a-c coupled, internally if the time constants are made long enough, with some of the advantages of a-c amplifiers over d-c amplifiers; namely, simpler power supply requirements and reduction in the effects of drifts in operating point. Such an arrangement is program-controlled under normal conditions, but may clear itself spontaneously if the program is made too slow. All a-c coupling in a two-state memory needs considerable further investigation.

The choice between these types of storage was based on the following considerations: Permanent storage seemed attractive because it was (1) applicable to program control with great flexibility of timing, (2) easy to make completely symmetrical as to the "one" and "zero" states (this is an advantage for the switching-type gates of an adder)* and (3) easy to test by programming slowly by hand. There has not been time yet to study all a-c coupling at all well. On these considerations, it was judged that permanent storage should be tried first.

*"Diode Gates", by L. W. Hussey.

1.61 Electrical Representation of Digits at the Adder Input

If, to facilitate synchronization and aid simplicity, the adder is to be the program-controlled type with permanent storage operation, the simple and natural representation of a digit is by the level of a voltage or current. Thus the other possible representations, such as modulation on a carrier, were not considered. To determine just what the digit levels should be, it is necessary to draw on information on the nature of the transistor-register digit source. In the register described in the Part 2 of this report, the collector terminal (of the output stage) seems to be the best digit source. Looking into this terminal, one sees in one state about -30 volts in series with 4000 ohms and in the other state about -10 volts in series with 1500 ohms. Both conditions are available, one on one collector and one on the other, for either state of the one-bit register. This then is the way digits from both input registers and the carry register are to be presented to the adder. The 20 volts' difference in the two voltages is the nominal stimulus voltage of the adder gates.

There is a further consideration of the relation of the adder to the digit source. The adder will cause some current to flow in the digit source, and this current must not impair the ability of the register to change state and remember properly. It was determined experimentally that 2 ma in each collector terminal lead of the register, each current in the most unfavorable sense, was not an unduly large load in a register under laboratory conditions. Current was chosen to make the test because the switching-type gates considered for the adder have a constant bias current that is switched between input and output.

1.62 Choice of the Basic Systems for Performing Add and Carry

Two basic systems known to us for serial addition are the Boolean and the Von Neumann. The Boolean system adds only two digits at a time, so that two steps must be taken to add a pair of input digits (giving partial sum) and then to add the partial sum to the previous carry. The Von Neumann system adds two input digits and carry digit in a single operation, giving equal weight to all three digits. Functional block diagrams of the two systems (with means for obtaining carry omitted) are shown in Figure 2. Each has its standard carry arrangement. In addition there is a new carry arrangement described that is applicable to either system. This will be called "Hussey carry". All these carry arrangements are given as functional block diagrams in Figure 3. There are four possible combinations of add and carry.

The important functions within any of the adders are (1) gate operation, (2) storage of carry, and perhaps (3) amplification. The Switching Type Gate* seems ideal for adder operation. This gate has the very good feature that the input level required to prevent transmission need not lie between a pair of limits of level, but rather needs only to lie on the proper side of a single limit. Also the input level for full transmission needs only to lie on the proper side of a single limit. Gate outputs may be paralleled without added parts.

The carry storage and amplification (level restoration) can readily use the register and pulse amplifier (or amplifier-delay) described in the Part 2 of the present report.

With this information to go on, the four adder systems of interest were drawn in enough detail to compare their merits. Functional block diagrams of the four systems are given in Figures 4 through 7 inclusive. Each register stage shown can store one bit (one binary digit) of information permanently. Each register has two symmetrical outputs whose conditions reverse when the stored information is changed. In the drawings, the following convention will be observed: The signal to which a gate responds is the signal that stands on the upper right output of the register when the register is storing a one. The same signal stands on the other output when the register is storing a zero. This convention is an aid to using the functional diagrams as simplified diagrams of the complete, detailed, system.

Because of the symmetry of the digit sources, we have the choice of representing a one at the adder output by a pulse (coincident with the "add" pulse) and a zero by no output pulse at the time of the "add" pulse; or vice versa. We have chosen the former as being intuitively more natural. This choice fits in with the use of a common "set-to-zero" for shift in the shift register. Shift systems are discussed in Part 2 of this report. Most efficient use of the adder output is also discussed in the same section.

One may note that the Von Neumann adder, regardless of the type of carry, uses a 3 input, threshold "1 only" function. This function is realized differently for standard carry and for Hussey carry, as shown in Figures 8A and 8B. The reason for using a somewhat more complex arrangement in the standard Von Neumann adder is that the "3 input, threshold 2" function needed for carry is thereby obtained without extra parts.

- -

*"Diode Gates", by L. W. Hussey.

The relative merits of the four systems are compared in a quantitative way in Figure 9. This is a table giving a numerical comparison of eleven features deemed important for the data system application. In judging which adder to settle on, considerations given greatest weight were; quantity of equipment needed, whether or not tandem gates would be needed (one switching type gate operating directly into another one), number of program pulses required, and magnitude of current from digit source.

Switching type gates can be operated in tandem by means of suitable coupling circuits. Such as arrangement was used in the adder built before all the alternatives of the present report were considered. This was the one-step adder with standard carry. The adder worked quite successfully, but the use of tandem gates caused some transmission loss and put closer requirements on gate bias resistors and voltages. It also was felt that tandem gates for control of transmission by the program generator were less desirable than direct control of the adder gates, because the latter may allow lower loading of the digit source. The last point is treated later in this report. Fewer program pulses for the adder decrease the complexity of the program generator. Relatively low current from the digit source (measured in terms of the bias current to a single gate, as is shown on the comparison chart) means that the adder can have higher bias current, therefore more freedom from diode leakage effects and more output, with a given loading of the digit source. Emphasis is placed on the bias current of a single gate because the maximum output current of an array of gates (such as the array needed to produce the sum) is very nearly the bias current of a single gate of the array. Thus low source loading for a given gate-array output.

On the basis of these facts and comparisons, it was judged that the one-step, or Von Neumann, adder with Hussey carry was the best choice.

1.7 Circuit Design of the Adder

1.71 Gate Circuits

In the adder chosen, all the gate circuits (switching type gates) must be of the same polarity. The only decision to make is which polarity. In the positive (or positive responding) gate, the gate puts out a positive output (transmits) when all inputs are positive and puts out a negative output for any other condition of inputs, mixed or all negative. In the negative gate, the gate puts out a negative output (transmits) when all inputs are negative and puts out a positive output for any other condition of inputs. "Positive" and "negative" are relative, not necessarily measured with respect to ground. To decide which gate to use, we draw on the fact that the input

sources of a switching type gate supply current only when the gate is not transmitting, and the gate current is divided among those sources which are of a polarity to prevent transmission. From the figures previously quoted on the internal output impedance of the register stage, it appears that the -10 volt, or more positive, condition has much better voltage regulation than the -30 volt, or more negative, condition. If negative responding gates are used, the gates will draw no current when the source is in its more negative (poor regulation) region. This seems desirable. It means, however, that a negative program pulse must be used, and the present transformerless monostable transistor circuits used as pulse amplifiers do not produce negative pulses as well as they produce positive pulses. If negative gates are used, the gate output resulting from an enabling pulse on the gates (assuming the d-c gate input voltages are such as to permit transmission) is an immediate negative step. This fits in well with the use of the pulse amplifier described in Part 4.0 of this report following the adder to raise its output level. This amplifier requires a negative step to trigger. Thus it seemed reasonable to design the one-step adder with Hussey carry on the basis of negative gates.

1.72 Gate Bias Level

The value of bias current for a switching type gate in the adder must be chosen. As shown in Figure 8, the largest adder input current in a register output for the one-step adder with Hussey carry is 1-1/3 gate bias currents. Based on the fact that the register can stand at least 1 ma and perhaps 2 ma, a gate bias current of about 0.9 ma was chosen. The bias current for each of the six switching type gates is obtained from a -90 volt supply with an 82,000 ohm resistor for each gate. This supplies the desired current, and the supply voltage chosen is large compared with the changes in voltage on the gate (-10 to -30 volt range), giving almost constant current.

1.73 Gate Loads

When the nature of the gate stimuli and the gate bias current are known, it is easy to determine an optimum load for a gate. This load is also optimum if fed by an array of gates. The process for finding the best load is shown in Figure 10. A complete V-I characteristic of the gate output is measured with gate transmitting and again with it not transmitting. Then one chooses a load line slope and a voltage to which the load line is returned in such a way that the change in voltage, current, or power in the load is maximized as required by the load sensitivity. One must also consider that the gate characteristic in the non-transmitting condition may

change enough under adverse conditions for the intersection to slide around the knee of the curve and allow output current to flow. In the present case, the load has been returned to -14 volts, so that the nominal -10 volt gate input (not transmitting) can vary several volts before any output current appears. With the load line as chosen, the load voltage steps from -14 to -28 volts and the load current steps 700 microamperes when the gate is switched on. These values fit well with the input impedance and trigger requirements of the pulse amplifier discussed in Part 2 of this report. The effective gate load impedance in this application is the load impedance seen by a step function. The capacitor coupling the gate output to the pulse amplifier can be considered zero impedance for this purpose. The load line shown represents the resistance of gate-output bias resistor and amplifier input resistance in parallel. Being capacitance-coupled, the amplifier input does not determine the quiescent position of the load line, that is, crossing zero current at -14 volts.

The program pulse applied to the gates need be only as long as the pulse required by the gate load. The pulse amplifier works well with a 1 microsecond input pulse.

1.74 Requirements on Gate Diodes

The requirements on the diodes used in the gates are determined mainly by the gate bias current and the nature of digit source.

The forward resistance of the diodes must be low enough that the bias current flowing through a diode will develop a voltage much smaller than the voltage of a stimulus at the input. With 1 ma gate bias, the forward resistance could be as much as 100 ohms (giving 1 volt drop with the full bias current) and still be tolerably small. This requirement appears easy to meet.

Requirements on maximum reverse current of diodes are as follows:

An input diode has reverse voltage when the corresponding input voltage is set to transmit. If all inputs are set to transmit, the reverse current augments the gate output. If one or more inputs are set to prevent transmission, the reverse current flows to the input thus set. Thus reverse current in the input diodes acts to increase the apparent gate-bias current. In a 4-input gate with only one input set to prevent transmission, the one input takes the bias current plus reverse current of the other three inputs. With a 1 ma bias current, a maximum total reverse current of 0.3 ma, or 0.1 ma per diode, seems reasonable.

Low reverse current in the output diode is essential when gate outputs are connected in parallel. In the adder gates, only one of the four gates with paralleled outputs ever transmits at one time. The reverse current of the other three gate output diodes is current robbed from the load. The nominal maximum available load current is 0.7 ma. A total diversion of perhaps 30 per cent of the output current, or 0.2 ma, can be tolerated. This corresponds to 0.065 ma maximum per output diode. It is interesting to note that increased reverse current in the input diodes gives greater gate output, tending to compensate the output signal for increased reverse current in output diodes of other gates connected to the same load.

The requirements on reverse current appear to be too strict to be met by present point contact diodes at extreme temperatures. P-N junction diodes show good promise of meeting the requirements.

1.75 Most Efficient Use of Power from Digit Sources

If it should become desirable to operate the adder gates at a much higher current level (bias current and output current), it seems likely that the circuit could be modified to permit this and at the same time make the current loading on the digit sources even smaller. This arrangement depends on the fact that, in a switching-type gate, the input whose voltage is farthest in a sense to prevent transmission takes substantially all the gate bias current. The collector of the digit source register will have some voltage beyond which it will never go more positive, say -8 volts. Every switching-type gate in the adder has one program-pulse input. If the program source has low d-c resistance (as a pulse transformer winding) and is returned to say -6 volts, then the program source will carry the bias current of all gates in the adder as long as no program pulse is present. Then the digit inputs to the adder will be required to furnish current only during the program pulse.

This makes the average resistance of the digit inputs to the adder very high, substantially the back resistance of the diodes connected to a given digit input. This allows us to benefit from the integrating effect of a resistance-capacitance L-section filter of low-pass type between the digit source and the adder (shown in Figure 11). In operation, the capacitor has some 1500 microseconds to assume the potential of the digit source, then it must furnish current for only 1 microsecond or so. The 1500 microseconds represent the time between shifting digits into position and the time of the "add" and "carry" program pulses.

It has been mentioned that the current from a digit source with the proposed circuit is substantially the reverse current of diodes connected to the source. This reverse current must be low enough that it will develop a voltage small compared with the stimulus-voltage when it flows through the series arm of the R-C filter. The value of R might be 10,000 or 15,000 ohms, and the voltage should not exceed say 2 volts. Thus the total reverse current should not exceed 0.2 ma or 0.065 ma per diode where three diodes are connected to a digit source. If an inductance were used to replace the resistor in the filter, much higher leakage could be tolerated.

Note that if a transformer is used in the program-pulse path, the transformer can invert the program pulse, and the need for a negative program pulse at the adder is no longer a disadvantage.

1.76 Cross-Inhibition of Pulse Amplifiers

There is one remaining feature, the need for which is not self-evident. The Hussey carry has a register that is set to one when one carry gate transmits and set to zero when the other carry gate transmits. In the model adder, a pulse amplifier was put in between each gate output and the register input. It was found that when a pulse from one amplifier changed the state of the register, the voltage change on the idle input of the register, applied to the output of the idle amplifier, would sometimes trigger the idle amplifier. The trouble could no doubt have been cured by means of diodes together with a clamping voltage at the register input, similar to the circuits in the shift register. However, to try out as many arrangements as possible, a different system was used. Making use of the fact that the two amplifiers should never fire simultaneously in normal operation, each amplifier output was connected to the opposite input through a network of appropriate loss. Input is a negative step and output is a positive step. Thus when one amplifier fires, its positive output disables the other for an instant. The networks are simple and non-critical. They are shown on Figure 12, the schematic of the model adder.

1.8 Summary of Adder Work

Many adder systems and sub-systems have been investigated and evaluated, and a model adder has been designed and built. The adder is simple and reliable and fills the requirements for the data transmission system with the possible exception of failure due to diode leakage at high temperature. This shortcoming can be corrected by use of p-n junction diodes and/or by a circuit trick presented. The requirements on the internal amplifiers and register are no more severe than requirements in shift register use. The internal amplifiers and register and the adder as a whole can be expected to work under the limit conditions given in Part 2.

2.0 Shift Register

2.1 Requirements on the Shift Register

2.11 General

In the data transmission system under consideration, a shift register is required to present serially to the adder 16 digits that it has previously taken in simultaneously in space division. Concurrently with feeding digits out serially, it may also have to receive digits serially from an adder output into the space vacated each time a shift occurs; and, after taking in 16 digits serially, feed these out simultaneously in space division. Digit shifting should occur in response to a program signal and should not occur otherwise; specifically, digits fed in space division should stay put and not shift as a result of the read-in.

Sixteen digits that together form a number are called a "word". The adder requires that the serial output begin with the least significant digit of the word. By convention, the output is taken from the right side in a functional diagram, so that when the register is storing a word, the least significant digit appears at the right. This convention is consistent with writing numbers on paper, where the least significant digit is at the right. A block diagram of the register-adder is shown in Figure 1.

Conversion between space division and time division is accomplished in a shift register by storing a word in a string of one-bit registers (as many as there are digits in the word) and, in response to a shift signal, shifting all stored digits, each digit moving one place to the right. In accordance with the reasoning in Part 1, it was decided that the one-bit registers should give permanent storage.

The register should consume a minimum of power and have maximum reliability under adverse conditions, particularly high temperature.

2.12 Digit Output

The output of the register is simply the condition of the two collector terminals of a one-bit register. It seemed reasonable that the output should be taken as whatever the best practical register known at the time could produce, rather than setting up limits based on the load fed by the register. The output capabilities of the register are described in Part 1.6. This represents the output characteristics both for serial feed to the adder and for simultaneous output. In simultaneous parallel output, the word stored in the shift

register must be impressed on a set of 16 one-bit registers in response to a program signal. Detailed circuits to accomplish parallel read-out in a transistor system have not been studied.

2.13 Time Sequence Digit Input - Design for Maximum Efficiency

An unsymmetrical source of input digits (in this case the output of an adder) can be used most efficiently in the following way: The source has a certain difference in voltage and current between its two states. Arrange clamping circuits so that the source produces no output in one state and in the other state produces an output equal to the change between the two states. Let an output represent a one and no output represent a zero. The source cannot then read a zero into the register, but the power available for reading a one into the register is 6 db greater than it would be if one attempted to accomplish both "set to one" and "set to zero" from the source. This trick is successful if one has a program such that the register can always be set to zero by the program generator in advance of the time at which the source is scheduled to put out a digit (a zero or a one). A pulse from a program generator, at least in the present application, is relatively cheap. It is relatively powerful because it is not the result of logical operations which generally introduce loss. A digit pulse from an adder, on the other hand, has come from a complex array of gates and two-state circuits. Here an increase in signal level is more dearly brought, and it pays to use the adder output signal to the fullest efficiency permitted by the program.

As noted in Part 1, the electrical representation of digits at the adder output consists of no output for a zero and a 0.5 ma negative pulse into any load up to 27,000 ohms for a one. The pulse lasts as long as the "add" program signal is fed to the adder, say 1 to 5 microseconds. The serial input of the shift register must operate from these pulses.

2.14 Simultaneous Digit Input in Space Division

In the present application, a register must accept 16 digits simultaneously and be stable in its new state within about 100 microseconds. It appears that a far tighter limit on time can be met.

There are two basic ways of storing digits in parallel simultaneously) depending on whether the register has any ones standing in it at the time or not. Registers are used in the present system under both conditions. The difference in ways of reading-in arises because a register stage in general has two separate input leads for storing a one and storing a zero.

(As opposed to, say, a single lead with different polarities to store one and zero.) If the system design assures that there will be no ones, only zeros, in the register at the time of parallel read-in, only 16 leads (to the "set to 1" inputs of the register stages) need be used. However, if there are any ones in the register, these will sometimes need to be changed to zero. In this case 16 additional leads to the "set to zero" inputs of the register stages can be used. Or, all stages can be set to zero with a common lead, and use made of only 16 leads to set the appropriate stages to one. It appears that the cost of equipment to feed signals to the 16 extra zero-setting leads is low in this case. The digit source is a set of key switches, 16 for each different stored word that must be read in. Read-in is accomplished by applying a pulse to a group of 16 switches. If only 16 leads are used, single-pole-single-throw switches are needed. The pulse either passes through the switch or not. To add 16 set-to-zero leads, one needs only to make the switches double-throw. The pulse then is fed to "set to one" or "set to zero" according to key position. On the other hand, some complexity would be added by a common "set-to-zero", even if such a system were already used for shifting. (Methods of shifting are treated in a later section.) An extra program pulse on the set-to-zero lead, out of rhythm with shift pulses, would be needed. Thus the 16 extra leads seem attractive for the case where "set to zero" is needed.

The electrical representation of digits at the parallel inputs of the register depends mainly on the one-bit register circuit. With the register circuits described herein, a register input looks like about 2000 ohms and requires a positive pulse of about a microsecond (or greater) and 10 volts (or more) appearing at the one-bit register input.

2.2 Methods of Shifting Digits in the Register

The problem is to change the state of one-bit registers so that the stored information moves one place to the right along the linear array of one-bit registers. There are two basic ways of accomplishing this. The first is to sample the state of each register and cause the next register to go to this state. The second is to set all registers to "zero" and then set to "one" those registers where a "one" was stored in the preceding register. The two systems are shown in Figures 13A, 13B1, and 13B2.

The complete shift cannot be instantaneous in either case. As an example, consider the first system. Here the state that we must sample is the state of a register just before it starts to change under the influence of the preceding stage. Thus some delay or memory is required.

The use of the first system for transistors has been investigated by Hussey and Trent. They concluded that gain was needed in the interstage and proposed the second system as most appropriate for transistors. Their work was covered in a report which included a shift register block diagram. This block diagram (similar to Figure 13B2) was followed in designing the present model shift register. However, since the model was built, Trent has introduced a "regenerative gate"** that gives gating and regenerative gain in a simple structure. This makes it desirable to take another look at the ways of shifting digits in a transistor system.

The regenerative gate can be used both in the first system, Figure 13A and in a form of the second system, Figure 13B1. To repeat, Figure 13B2 is the form of the second system used in a present model. Here a two-transistor monostable circuit ("amplifier-delay") was used. The gate shown in Figure 13B2 is needed to disable digit shift due to registers changing from one to zero during parallel read-in. The gate was omitted from the first form of the model. The function of prevention of digit shift during parallel read-in is very simple when regenerative gates are used. Here the absence of a pulse to the gates prevents digit shifts.

The regenerative gate functions as the two-input gate plus amplifier in Figures 13A and 13B1. The gate has one steady input and one pulse input. It is important to note that the pulse inputs are connected together to a common program source. (In Trent's reversible binary counter with regenerative gates, it is the steady inputs that are connected together to a common source.)

Even though the transistor systems may need interstage gain (regenerative gate or amplifier-delay) where the electron-tube version does not, the transistor system has a strong advantage in power consumption. A stage of the electron-tube version draws about 7.5 watts total, whereas the transistor stage complete with amplifier-delay draws only 1 watt or less. With a single regenerative gate in the interstage, the drain per stage would be about 0.75 watt. In comparing power drain and quantity of equipment it is of interest to note that the buffer amplifiers frequently used on the output of an electron-tube two-stage circuit have not been found necessary with transistors.

- -

**"A Transistor Packaged Circuit for Multifunctional Switching Applications", by R. L. Trent.

In making a practical choice among the three systems comparisons must be made first between those illustrated in Figures 13A and 13B1, then between the latter and that in Figure 13B2. Where regenerative gates are used in the interstage, the choice between the first and second systems is as follows: The arrangement shown in Figure 13A has the simplest programming. Only one pulse per shift is needed, and this pulse must operate regenerative gates only, two gates per stage. The shift pulse must operate 64 regenerative gates when there are two 16-stage shift registers. The arrangement in Figure 13B1 uses two program pulses (on two wires) per shift. One pulse must operate 32 register stages and the other must operate 32 regenerative gates. Were only a few stages used, simplicity of program would outweigh extra equipment in the interstage, thus favoring the first system. With many stages, say 16, the choice depends on whether many registers can be set to zero simultaneously from a simple program generator. In the model system, 9 one-bit registers are fed from a single transistor, and these are not the most sensitive registers that can be built. Thus, the arrangement shown in Figure 13B1 seems more desirable at present.

The choice between the systems shown in Figures 13B1 and 13B2 depends upon the complexity and reliability of the interstage. There are two rather different types of regenerative gates, with different characteristics, available. These have not yet been tested in a shipped register. The interstage circuits for realizing the system shown in Figure 13B2 have been studied fairly closely. With the exception of the gate, these circuits have been tested in operating shift registers. The choice between circuits in Figures 13B1 and 13B2 is discussed in Part 2.4 of this report.

2.3 The One-Bit Register

2.31 General

Requirements on the one-bit register are outlined in Parts 1.3 and 2.1 of this report. In general, the input is to be as sensitive and output as powerful as we can attain, consistent with stability of registry in all adverse conditions, with a simple circuit.

The maximum transition time allowable in the present application (say 1 or 2 microseconds) is relatively long in terms of the time constant of a transistor. Thus it seemed reasonable to concentrate effort on maximum reliability of storage rather than fast transition.

The register output should have push-pull symmetry for feeding the adder. Separate "set-to-one" and "set-to-zero" inputs are required.

At the start of this investigation, the most promising bistable transistor circuit was a unsymmetrical two-transistor binary counter developed by R. L. Trent. This circuit is the analog of the electron tube "cathode-coupled bistable multivibrator". The transistor circuit has stabilization added to take care of variations in base current due to temperature change and differences among transistors. Trent's two-transistor circuit seemed more reliable than the best available one-transistor circuit, indicating that there might be an inherent advantage in using two transistors. This indication was strengthened by later work.

The problem of finding a suitable circuit configuration was attacked as follows:

(1) By attempting to draw all interconnections (using passive elements) of two transistors which would give positive loop gain and such that a disturbance in the system would cause the two transistors to go towards opposite conditions (one towards cutoff, or OFF, one towards saturation, or ON). The last condition was put on so as to get symmetry in the push-pull sense. Also, if both transistors were to proceed towards the same condition, there seems to be no advantage over a single transistor except increased loop gain. The extra gain might give faster operation, but probably no better permanent storage.

(2) By starting with successful electron-tube circuits, drawing the transistor analog, and adding stabilization appropriate to transistors. The idea of using the electron-tube analog looked attractive in this case because of the relation of Trent's unsymmetrical binary counter to the known electron-tube circuit.

(3) By comparing the resulting circuits with successful transistor circuits to see if any known features could profitably be added.

2.32 Circuit Configurations

There are three basic circuit configurations for a single transistor; grounded base, grounded emitter, and grounded

collector. The grounded collector arrangement can furnish two possibilities because it has gain in both directions. The ten combinations of two transistors with one element grounded are shown in Figure 14. For permanent d-c storage, no transformers are allowed. All but six combinations can be thrown out on the basis of degenerative feedback. Two more can be thrown out because the transistors go the same direction. The case of two grounded-collector transistors (8) and (10) of Figure 14 does not seem promising because transmission around the loop in one direction tends to cause the units both to go towards OFF or towards ON. (If a unilateral device were used between transistors this configuration might be very attractive.) Thus there remain three configurations of interest.

At this point it is well to consider what other interconnections can be made. If the grounded electrode of the transistor can be raised off ground without changing the phase relation of the other electrodes, there is another terminal available. In (4) of Figure 14, we can add a path between the base of the left unit and the collector of the right unit. This makes (4) symmetrical and gives it the qualities of (5). Or we can arrive at the same configuration by putting common impedance in the emitter-to-ground paths of (5). This combination of (4) and (5) is a rather general configuration for our purpose, and circuit work has been concentrated on it.

The configuration (6) has not been studied thoroughly. Perhaps this can also be made symmetrical. This arrangement has the property that each half inverts the signal phase. The phase inversion in the grounded-collector half comes about because the resistance looking into the emitter is a negative resistance, lower than the resistance of the generator feeding the emitter. The voltage appearing on the emitter is inverted in phase from the internal voltage of the generator. This property depends on the transistor's being in its active region. With the transistor in the cutoff or saturation region (as it might be in a bistable circuit) the phase of the transmission might reverse. This could be good or bad. Here further investigation is needed.

Starting out from electron-tube circuits, one finds that the basic (symmetrical) bistable multivibrator is the analog of (5) of Figure 14, and the cathode-coupled bistable multivibrator is the analog of (4) of Figure 14. Thus the combination of (4) and (5) seems worth our attention. One should note that there is still the possibility of adding paths on an individual half of the transistor circuit to give positive feedback at d-c. This is not possible with electron tubes.

The symmetrical two-transistor binary counter developed by Trent .

has the general configuration of (4) and (5) combined. The binary counter has circuit refinements that give the memory required for operation with single-lead input and also give especially easy triggering. Its development was concurrent with register development, with mutual benefit.

2.33 Circuit Design

In choosing the circuit elements for the configuration, it is convenient to treat the a-c and d-c transmission separately. In the present case, the circuit must store information indefinitely, so that the d-c characteristic is of primary importance. It seemed reasonable to set up the d-c transmission and then make only necessary modifications to it for a-c transmission.

Perhaps the most familiar approach to transistor circuits is the plot of the 2-terminal V-I curve seen by breaking the circuit at an emitter and looking in between emitter and ground. In the present case, the load line for the plot is the internal resistance and voltage of the emitter coupling network together with the emitter of the other transistor. The treatment was very successful in this case. The other interconnecting paths, that is, the collector-to-base paths, are fully taken into account by the way they modify the V-I characteristic.

It seemed reasonable to choose first the collector-to-base circuits, find the V-I curves, and then design the emitter network in the light of what was found. Two V-I curves should come out of this - one "other half off" and one "other half on". A pair of these measured curves is shown together with a circuit of half a register in Figure 15. Circuit constants were chosen with the aid of a system worked out by B. G. Farley for computing emitter V-I curves from known transistor characteristics. Base stabilization is used (diode and positive voltage) to restrict the range of variation of the intersection on the voltage axis.

The 3.6 ma bias current in the diode is chosen equal to the sum of current from A (other side ON) plus the maximum expected leakage current of the transistor. Some series resistance is inserted base-to-diode. This resistance results in negative resistance looking into the emitter in the low-emitter-current region (base diode conducting) and may aid triggering. The point A represents the collector of the other

side, and the two voltages at A represent the voltages expected with other side ON and OFF. The resistor from A to base is chosen as low as possible without loading the other collector unduly. The collector-to-minus-supply resistor R_c is chosen low enough to allow good negative resistance looking into the emitter and high enough to (1) give good collector-voltage change for output purposes and (2) limit ON current for economy of power and transistor protection. To give good negative resistance, R_c must be somewhat lower than the $(\alpha - 1)\, r_c$ of the transistor in the active region.

In the normal operation of the register, the reliability of registering depends on having the ON side stay on and keep the other side OFF. Note how this shows up on the V-I curves. In order for the ON side to stay on most safely, its load line should intersect only in the ON, or saturation, region. This can be accomplished by having a load line of high enough resistance intersecting the axis at a positive enough voltage. The line shown, 2,200 ohms returned to ground, accomplishes this. The other side is assumed OFF, with high emitter resistance, so the load line is determined by the coupling resistor only. It is essential that the voltage at which the load line intersects the curve (in the ON region) shall be negative enough to keep the other side safely OFF under adverse conditions. Here the effect of the collector-to-base coupling shows up clearly. The emitter characteristic of the ON unit is the more negative one, giving a relatively negative intersection. Yet the OFF unit will stay safely off as long as the intersection is more negative than the turnover point of the emitter curve of the OFF unit (point near zero current). The curve of the OFF unit is the more positive curve, representing added safety. The criterion for choice of load line is, then, that it shall intersect at the most negative voltage consistent with always intersecting in the ON region and only in the ON region. It appears from this analysis that the use of two transistors in a register may give inherently better reliability under adverse conditions than a single transistor.

A complete register circuit is shown in Figure 6-16. The circuit was found to operate as required without the addition of any capacitors or inductances, so none were added. It did not seem necessary to add any of the various diode circuits (other than base stabilization) described by Trent.

Power dissipation in the transistor should be limited to about 100 milliwatts (except during transition, which is so short that higher power can be tolerated). Power with the resistance values and voltages shown is as follows: The side that is ON carries about 7-1/2 ma. The ON resistance can be expected to be less than 1000 ohms. If it were 1000

ohms, the power would be only 50 milliwatts. The side that is OFF sees a circuit of some 6000 ohms and 45 volts between collector and base. Maximum power is dissipated by a low V_{co} transistor. For a limit V_{co} transistor (limit for performance in the present circuits) with about 10,000 ohms collector resistance, the power would be about 80 milliwatts.

2.34 Triggering

With this register design there is the possibility that if both sides were in the ON condition simultaneously, they would stay there and give improper operation. This follows from our trying to attain absolute maximum reliability of storing a digit in this circuit. Fortunately, we can prevent both sides being on simultaneously by proper choice of triggering. If the trigger always turns the ON side OFF, the other side must go on, and we have the desired result.

In the circuit given, there is no provision for "steering" triggering pulses. Rather, the pulses must come in on two leads. This is consistent with register use. Steering could be added with some extra elements.

Triggering of the one-bit register from a slowly changing voltage was investigated by measuring V-I curves looking into a base or a collector (circuit not opened). Measurements were made with the register supplying no power, and measurements on the base were made with collectors supplying output currents of 1 ma in all combinations of polarities. This output current did not interfere appreciably with operation. A typical set of curves is shown in Figure 17. From these curves one can see that a positive trigger applied to the ON collector will work, but 7 or 8 milliamperes is required. Remember that there are no "speedup" or "memory" capacitors or elements of similar function to couple collector to base tighter for high frequencies. A negative trigger on the base of the OFF side may or may not work. It can be seen that it is possible to get both sides ON with this kind of trigger. A positive trigger on the base of the ON side seems foolproof, both sides switching state.

Trigger requirements including taking account of the internal resistance of the trigger generator, can be studied by means of Figure 17. The trigger generator internal resistance and internal voltage are represented by a load line (really a generator line) and intersection of load line on the voltage axis. As the internal voltage changes (the rise of trigger voltage), the intersection of load line and V-I curve moves. When this intersection reaches the limit of the stable portion of the V-I curve, the register changes state. Using this approach, one can find the internal trigger voltage

necessary with a given trigger generator resistance or vice versa. It is assumed that the (resistive) trigger source is coupled with a capacitance large enough to have negligible reactance at the high frequencies present in the trigger. Then the "load line" will be resistive, and as shown. The effect of the capacitor is to cause the load line to drift back after triggering until it intersects the V-I curve at the zero-current axis.

The V-I curve may be near enough a straight line in the region of interest to use a straight line approximation. The curve for positive triggering on the ON base is 2000 or 3000 ohms, and the register changes state when the trigger source is putting 2 ma or more into the register.

The mechanism of triggering can be grasped by referring to Figure 14. A positive trigger on the ON base raises the whole emitter curve of the ON side towards positive voltage until the sides switch state.

Using pulses longer than about 2 microseconds, the trigger requirements seem to be about the same as with a slowly varying voltage. With shorter pulses, a little more voltage is needed.

The polarity of pulse needed fits very well with the present pulse amplifiers and amplifier delay circuits. These circuits can give a positive output pulse that is more powerful than the negative output pulse.

2.35 Comparison of Electron-Tube and Transistor Registers

It seems fruitful to compare electron-tube registers and transistor registers at this point. A symmetrical electron-tube register can be treated by the "emitter characteristic" method just used. Using analogous elements, we must open the cathode lead of one stage and look in. The curves are as shown in Figure 18. Note that the curves are very similar to the transistor curves rotated 180 degrees. There are two striking differences. The electron-tube curve has no negative slope. Also the electron-tube V-I curves can be moved vertically with negligible expenditure of power to move them. The vertical position follows the grid voltage. In the transistor case, vertical position follows the base voltage, and a relatively large and somewhat variable current must flow to change the base voltage. The difference in current required to move the characteristic can be thought of as an advantage for the electron tube. However, the negative resistance shown by the single transistor (and by it alone) allows us to choose a load line such that relatively

small vertical motion of the V-I curve is needed. (In some single-transistor registers, there is no vertical motion at all.)

2.36 The Register as a Generator

The equivalent output circuit of the register is well represented in this application by Figure 17. The ON collector has a certain internal voltage (about -11 volts) and its internal resistance is almost linear at about 1250 ohms in the example. The OFF collector shows -33 volts and 3600 ohms. The ON collector can be expected to be about -10 ±2 volts and the OFF collector about -32 ±4 volts with extreme variations in transistors.

2.37 Requirements on Transistors for the Register

The two primary requirements on transistors are that the collector-to-base d-c resistance shall be high enough in the OFF condition (with emitter current zero) and low enough in the ON condition.

The circuit is stabilized against collector leakage in the OFF condition up to 3 ma. This corresponds to 10,000 ohms d-c collector resistance, or V_{CO} = about 20 volts. Attempting to stabilize against higher leakage current reduces the voltage span of the negative resistance range of the emitter characteristic so that the voltage used to keep the OFF side off is not negative enough. V_{CO} of 20 volts is a very low value for switching transistors, which can be made with V_{CO} above 40 volts and on up to 100 volts or more. However, V_{CO} is the transistor characteristic that is most dependent on temperature, going downwards at high temperature. Transistors have been made with V_{CO} that stays high, say 40 volts or more, even up to 80° C. However, these are scarce at present, and it seemed best to make the circuit workable with as low a V_{CO} as present circuit art would permit.

The resistance in the ON condition must be held low enough (under control of the limited emitter current available) to keep the other side off. The available emitter current is determined largely by the emitter-to-ground resistor. This means that the transistor needs to be one for which characteristics on the V_C, I_C plane hug the zero-voltage line at low collector current and break away from this line at some value of collector current. A value that seems attainable in a switching transistor and is satisfactory for the circuit is that V_C at 3 ma emitter and 5.5 ma collector shall be below 4 volts. The V_C at 3 ma has been called V_{C3}. The V_{C3} test insures that the "average" α (current amplification) exceeds some minimum.

It is possible for a transistor to have a defect, such as a collector-to-emitter leakage path, that is not caught by these two tests. This leakage and also the base resistance controlled by a requirement that the r_{12} (common resistance for backward transmission) shall be less than 1000 ohms.

These three tests do not cover the active region of the transistor. The best single figure of merit in this region known at present for the register is $(r_m - r_c)$ or $(\alpha - 1) r_c$. This figure is, among other things, the slope of the constant-base-current curves on the V_c, I_c (r_{22}) plane. It has not been necessary to invoke such a test for the present application of the register. The d-c tests tend to insure that $(\alpha - 1) r_c$ is not unreasonably low.

The model adder-shift register uses 20 transistors in registers. It has been possible in every trial so far to get reliable system operation with transistors that pass only the three tests, $V_{co} > 20$, $V_{c3} < 4$, $r_{12} < 1000$. Thirty or forty transistors have been tried. In a typical case, about half the transistors in use were within 5 volts of the V_{co} limit and perhaps one fourth of them within one volt of the V_c limit at 3 ma emitter current. The system was operated several times as fast as the present requirements and gave many millions of successive operations without an error. The -45 volt and +36 volt supplies to the register could be varied individually ±15% without errors.

It should be understood that transistors must pass the three specified tests at all extremes of temperature, etc. under which the system must work.

2.38 Summary of Register Circuit Performance

Input and output are completely symmetrical in a push-pull sense. When the circuit is storing a zero, a certain pattern of voltages and impedances appears. When it is storing a one, the same pattern appears but it is reversed in space.

Input and output are described here by giving an approximate equivalent circuit for each state. The circuit is an internal resistance in series with a voltage. See Figure 16.

D-C Power Requirements:

8 to 10 ma at -45 volts

7-1/2 ma at +36 volts

Input:

Terminal 1, State A or Terminal 2, State B } 2000 ohms, -10 volts

Terminal 1, State B or Terminal 2, State A } 2000 ohms, -3 volts

To change state, move the input that is at -10 volts to -3 volts

Output:

Terminal 1, State A or Terminal 2, State B } 1500 ohms, -10 volts

Terminal 1, State B or Terminal 2, State A } 4000 ohms, -28 to -36 volts

Maximum output terminal current without ill effect: 1 ma in each output simultaneously. With regard for characteristics shown on Figure 17, larger currents can be used.

Speed of Operation:

This has not been well determined, but it is known that the register responds in about a microsecond with a step at the output. The maximum repetition frequency for change of state has not been well determined and needs further study.

2.4 Pulse Amplifier and Amplifier-Delay

2.401 General

A pulse amplifier is needed in the adder; its system function is treated in the Part 1 of this report. Its broad function is to raise the level of gate output and operate a register.

The amplifier-delay is used in one of the two methods of shifting digits from one register to another which are treated in Part 2.2 of this report. The other method uses a regenerative pulse amplifier that can be enabled or disabled by a control voltage; this is called a regenerative gate.* To make a full treatment of the means for transferring digits in the shift register, one would need to compare the best available

- -

*"A Transistor Packaged Circuit for Multifunctional Switching Applications", by R. L. Trent.

amplifier-delay with circuits of the regenerative gate type. Circuits to give the amplifier-delay function have been studied, designed, and evaluated to the point where they have been used successfully in the model shift register, and a certain amount of design theory has been worked out.

The regenerative gate circuits for this particular application have not been investigated much. The base stabilized regenerative gate has been developed and studied by Trent, but performance has not yet been determined with many pulse inputs connected in parallel, the way the gate would be used in a shift register. Very little attention has yet been given to the new regenerative gate, or gated-pulse regenerator, consisting of a switching-type gate followed by the type of pulse amplifier treated herein.

In view of this, the only circuit for digit transfer covered in the present report is the amplifier-delay.

It is interesting to note that the pulse amplifier and amplifier-delay as designed for the adder and shift register were usable (and used) to build a program generator to supply add, carry, and shift pulses to the model adder-shift register for high speed operation.

The pulse amplifier is treated first. It turns out that the desired amplifier-delay action can be obtained by adding simple circuits following the amplifier.

The input and output requirements are the same for the amplifier and amplifier-delay. The main requirements are that input sensitivity shall lie between limits and output power shall lie above a limit. (This statement of requirements applies especially to regenerative circuits.) These requirements are treated in following sections. In addition, we want the input to look as much like a passive resistance as possible and we want the units to refuse to respond to stimuli applied to the output. Failure to meet these additional conditions perfectly can be taken care of by auxiliary networks in most cases.

In accordance with system operation at 3000 digit-pair additions per second, the amplifier and amplifier-delay must receive signals at least this fast. The 3000 per second speed is five times as fast as the data system requires and was chosen to give considerable margin without introducing any troubles.

The circuits for the amplifier and amplifier-delay functions should use a minimum of components and draw minimum power consistent with other requirements. They should have maximum reliability, particularly at high temperature.

An attempt was made in first-order design work to use the same transistor limits laid down in Part 2.1 of this report, that is, $V_{co} > 20$, $V_{c3} < 4$, $r_{12} < 1000$. This has been successful but for more refined work a pair of tests to replace the r_{12} measurement is indicated. These are (1) α versus V_ε at fixed I_c and (2) α versus I_ε at fixed V_c. These tests measure base resistance and collector-to-emitter leakage, respectively, and are useful in very close control of trigger threshold.

2.402 Input Requirements

The most severe requirements on sensitivity fall on the amplifiers fed from gates - the shift-register input amplifier and the amplifiers within an adder. Here the amplifier must respond to a step of voltage (negative in the model adder) from a source whose maximum power output is a 500 microampere step applied to a 27,000 ohm load. Gate output V-I curves and load line are shown in Figure 10. The 20,000 ohm gate output load shown there is composed of 82,000 ohms bias resistor and 27,000 ohm amplifier input in shunt. It can be seen that a lower value of load resistance will draw at least 500 microamperes. With the present adder, it seemed reasonable that the pulse amplifier should have an input impedance (nearly resistive) of 27,000 ohms or less and should respond, even under adverse conditions, to a step of 300 microamperes or more with step rise time of 1 microsecond or less. Variations in sensitivity due to changes in temperature, transistor characteristics (especially V_{co}), etc., should be held to an absolute minimum.

The preceding requirements will also allow the amplifier input to function properly in the register interstage. Here the register furnishes at least a 14 volt step (either positive or negative as desired) with at least 1 ma current capacity. In this case the collector that steps negativeward when the one-bit register is set to zero will be used as a source for the amplifier-delay.

2.403 Output Requirements

It seemed reasonable to try for as much pulse output power as present transistors in a simple circuit would allow. One thing uncertain at this stage is the relation of peak power, pulse length, etc., to life of transistors. It is necessary to make some guesses in circuit design and then see if life is satisfactory. For applications other than program generator, a positive step with 0.1 microsecond or less rise time and 10 volt or more amplitude applied to a 2000 ohm load is satisfactory. For program generator use, it is desirable to apply such a step to a 100 or 200 ohm load (many register inputs in parallel). A network will generally be added at the output to convert the step to a pulse. The amplifier should put out such

a step delayed no more than about 0.1 microsecond from the input. The amplifier-delay, when triggered, should put out no positive step until the end of the delay period, and then put out the required positive step.

2.404 Delay Period for Amplifier-Delay

The system will tolerate variations in delay from a few microseconds (just longer than the stabilization time of a register) up to periods of a substantial fraction of the interval between operations, which is 333 microseconds at 3000 operations per second or 1666 microseconds at the normal operating speed of the data system. In the type of circuits considered later in this report, the delay period is the ON period of a normally OFF monostable circuit used as an amplifier. This ON period is some fraction (determined by the circuit) of the recovery time of the circuit. It seemed likely that designing the circuit to fit a recovery time of about 333 microseconds would give a delay in the acceptable range.

2.405 Choice of Basic Circuits

The term "Pulse Amplifier" or "Amplifier" has been used frequently without further description. Such a function is not necessarily regenerative, but trials of linear and triggered, regenerative circuits indicated that the latter showed more promise for this two-state system. The information presented at the input is of such simple form ("yes" or "no") that there is no need to transmit it without change in waveform, but rather just to regenerate it. Also, the information arrives at spaced, predictable times. This is of essential importance because it means that between input signals the regenerative amplifier has time to accumulate and store energy for a relatively large output pulse when it is triggered. The gain can be made far higher than that of a linear amplifier of equivalent stability. One might think that such a circuit would have to be made insensitive to prevent triggering due to changes in temperature, etc. This has been overcome in a circuit invented by Anderson and treated in another report.* Here the circuit can be arranged so that only a sudden change exceeding the threshold will trigger the circuit. The threshold can be varied by design over a wide range. Slow changes in the transistor, circuit elements, or input voltage have little effect, even the threshold level staying constant. The circuit can be arranged so that the transition time of an information signal is treated as "sudden". The circuit with values chosen for the present application is shown in Figure 19.

- - - -

2.406 Attaining Maximum Reliability in a Regenerative Amplifier

In the Part1.0 of this report, a-c versus d-c operation was discussed briefly. Permanent d-c coupled storage was chosen for the digit sources for the adder gates. However, it was mentioned that a-c operation had advantages of freedom from drift of circuit characteristics. The amplifier circuit is an illustration of this point. One can note in the circuit (Figure 19) that the input is d-c coupled in a circuit sense. This is misleading because the circuit functions like an a-c coupled circuit and has the advantages of an a-c coupled circuit. Functionally, the input is sensitive to fast changes and is not sensitive to slow changes or drifts. One can conclude that in two-state work it may be more appropriate to talk about the advantages of "d-c insensitive" or "d-c sensitive" than of "a-c coupled" over "d-c coupled". The circuit performance comes about because to be effective the input signal must appear between emitter and base. If the input to the base changes slowly, the emitter follows, giving little or no emitter-to-base voltage. If the input steps, or changes rapidly, the emitter is anchored at its previous voltage and an emitter-to-base voltage appears. "Fast" or "slow" is measured against the time constant of the capacitance and the emitter-to-base resistance. In the circuit shown, this is $.01 \times 10^{-6} \times 10^{4} = 10^{-4}$ second or 100 microseconds.

This circuit is normally OFF. It is held off by current from the -90 volt supply flowing through the parallel combination of emitter back resistance and 10,000 ohm emitter-to-base resistance. This is the current that the trigger must overcome. Note that the current supply is very nearly "constant current". This fact is at the heart of the circuit stabilization of required trigger amplitude. The quiescent base voltage could vary, say, from -10 to -20 volts (representing a variation of 1 ma in collector leakage) with only about 15% change in the hold-off current. This order of stabilization is ample for the present application. If it were necessary to improve the circuit stabilization, one would use a higher voltage and resistance for the hold-off current supply.

2.407 Circuit Operation and Choice of Circuit Values

Considerations for the value of collector-to-minus-supply resistor R_C are about the same as in the one-bit register. This resistance must be low to give the necessary negative resistance looking into the emitter, otherwise the circuit would not trigger. For this, R_C must be somewhat less than the $(\alpha - 1)\, r_c$ of the transistor. R_C must be high to give the required amplitude of step of collector (output) voltage

when triggered. In the present application, the desired output current does not flow through R_c so that R_c does not directly limit output current.

The base-to-ground resistance R_b must be high, (1) so as not to rob too much trigger current from the emitter-to-base path and (2) so as to give desirable negative resistance looking into the emitter. It must be low so as not to take too large a part of the collector supply voltage (when OFF) with a transistor of low V_{co} (low d-c collector-to-base resistance).

The emitter-to-base resistance must be large enough (1) not to rob current from the emitter in the ON condition, and (2) not to require an excessively large constant current to produce the hold-off voltage. It must be small to let the circuit recover soon - that is, be ready soon to be triggered again.

The ON time is controlled by R_b and R_c together with C and the transistor (assuming no output load). When the circuit goes ON the capacitor sees an approximate equivalent circuit consisting of a resistance and internal voltage determined by R_b and R_c and the transistor. A current flows in the capacitor through the emitter as the capacitor tends toward the internal voltage of the equivalent circuit. The current slowly drops until it no longer will keep the transistor ON. As the transistor starts to go OFF, the base becomes more positive, accelerating the going off.

Immediately after going OFF, the emitter is quite negative, having been charged during the ON time by the negative supply. The circuit is thus insensitive at this time. The emitter drifts positive as the capacitor discharges. Most of the discharge current flows through the emitter-to-base resistor. A low value of this resistor gives quicker recovery. It is evident that the recovery time is dependent on the amplitude of trigger. With a strong trigger, the circuit will trip before it has recovered full sensitivity.

The trigger current into the base terminal divides among three paths, R_b, collector path, and the emitter-to-base resistance. A low V_{co} transistor gives greatest loss of trigger current. In the circuit shown in Figure 6-19, with a transistor having V_{co} = 20 volts, only 37% of the input trigger current is useful. To trigger on 300 microamperes, the hold-off current should be about 110 microamperes. (This is a rough approximation that is true if R_1 is very low.)

The resistance values shown in Figure 19 were chosen to fill the terminal requirements previously given for

the circuit. For circuit design, it was assumed that the transistor would barely meet the requirements given in Part 2.3 of this report. The capacitance shown gives an ON time of some 10 to 40 microseconds and a recovery time of about 300 microseconds maximum. The ON time is not especially important for amplifier use, but it can be used to determine a delay.

The circuit has some sensitivity to a negative step applied at the collector terminal. It is most sensitive (worst condition) with low V_{CO}. With V_{CO} = 20, about 12% of a current step fed into the collector terminal flows through the emitter-to-base path. In the present circuit, about 900 microamperes into the collector terminal will trip the circuit. Looking into the collector terminal one sees about 3500 ohms. Thus a negative voltage step of about 3 volts at the collector terminal will trigger in the case of V_{CO} = 20 volts. It is interesting to note that the collector responds with a positive voltage step, opposite to the voltage applied. However, the trigger current and output current are in the same sense. This latter point makes it rather difficult (but not impossible) to prevent backwards transmission to the collector by means of diodes.

A positive step on the input does not trigger the circuit, but causes it to start to stabilize with reference to the new, more positive, input voltage. After it stabilizes, the circuit is again fully sensitive to a negative step, this time referred to the input voltage at which it stabilized.

A series of negative steps, stairstep fashion, each going farther negative, will trigger the circuit again and again, provided the steps are far enough apart to allow the circuit to recover and provided the process is not carried too far.

The maximum value of collector supply voltage can be determined by maximum permissible dissipation in the transistor. It was considered that it would be safe if the collector dissipation for periods exceeding 10^{-5} second did not exceed 100 milliwatts with nominal voltage and resistances, and least favorable transistors. Somewhat higher peak power can be handled as the period becomes shorter than 10^{-5} second. With circuit values shown, the dissipation with no output load comes out as follows: The maximum dissipation in the ON condition occurs with high V_{CO} and high V_{C3}. The former causes the emitter to run near ground potential when OFF; giving higher currents when the unit goes ON. High V_{C3} means high ON resistance. The highest voltage to appear between -45 volt supply and emitter will be about 40 volts. Current is limited by R_C and the ON resistance and comes out at about 7 ma. The maximum ON resistance allowed in the V_{C3} test is

730 ohms. We can expect that the resistance is considerably lower when the unit has just gone ON and may be somewhat higher before it has gone OFF. If the ON resistance is 1000 ohms, the dissipation will be 50 milliwatts. In the OFF condition, dissipation will be maximum when the d-c collector resistance matches the 15,000 ohms of $R_b + R_c$. Here the dissipation is only 35 milliwatts. The collector supply voltage could be raised to 60 or 65 volts without exceeding the power limit. Adding a load increases the ON current and thereby increases the pulse power dissipation.

2.408 Considerations for Maximum Power Output

When a step occurs in the output of a two-state circuit, the output can sometimes be represented by a very simple equivalent circuit before the step, and a different simple circuit after the step. This is possible and very fruitful in the present case. When the amplifier is OFF we see looking into the collector terminal a resistance of some 3500 ohms and an internal voltage of, say, -36 volts. Immediately after it has triggered, we see a few hundred ohms in series with the capacitor and an open circuit voltage of say -18 volts. (These values are with a limit V_{co} transistor.) Now if we are going to use this transition to apply a step (following a quiescent condition) to an a-c coupled load, we can make use of a way of thinking that seems widely true, as follows: There is only one item of significance in the equivalent circuit before the step, namely, its open circuit voltage. The important equivalent circuit is made up of the internal impedance (assumed resistance) after the step together with a voltage step of the difference of internal voltage before and after. A low internal impedance after the step corresponds to high available power. The present amplifier circuit has excellent performance in this respect. If the initial ON resistance is say 400 ohms, with an 18 volt step, the available power is $E^2/4R = 0.2$ watt. With higher supply voltage, high V_{co} and low V_{c3}, a power of more than a watt is readily attainable. Maximum pulse power would be delivered to a 400 ohm load with the assumed ON transistor resistance. The power starts to fall as the emitter capacitor voltage changes. The output voltage falls much faster than when there is no load. A higher resistance load dissipates the output energy more slowly. The available output energy is proportional to the capacitance in the emitter lead.

When this circuit puts out maximum pulse power, the duty cycle is rather poor. The duration of the output pulse is roughly the ON resistance multiplied by the capacitance. The duration of the recovery period is controlled by the values of emitter and base resistances which are much larger than the ON resistance. In the present case the duration of the maximum power output pulse would be about $.01 \times 10^{-6} \times 400 = 4$

microseconds as against a recovery time of 300 microseconds. It appears that the present design can be somewhat improved if maximum power output together with good duty cycle are needed.

The maximum possible pulse current with the assumed constants is determined by the 18 volt step and 400 ohm internal resistance - that is, 45 milliamperes (a-c short-circuit load). This flows through the transistor from emitter to collector, and corresponds to 800 milliwatts dissipation. If the capacitance were large enough to maintain this for 5 or 10 microseconds or more, one could expect the transistor to be impaired. It is not yet certain how short such pulses must be to allow long transistor life.

It has been mentioned that the internal output resistance immediately after a transition is of most interest for maximum pulse power output. This consideration bears on our choice of whether an amplifier consisting of a triggerable monostable circuit should be normally ON or normally OFF. We would like to have the transition that occurs upon triggering to have the largest step of open circuit output voltage and lowest internal resistance following the transition. We would also like to have transistor dissipation and power supply drain low as much of the time as possible. The present circuit fits these requirements well. It seems true of all n-type transistor-triggered circuits that, if the circuit is to draw acceptably low steady current in the quiescent period before triggering, then high output power from the collector terminal is obtainable only from the positive step.

2.409 Summary of Amplifier Circuit Performance

Input and output are described here by giving an approximate equivalent circuit for each state. The circuit is resistance, capacitance (if any) and voltage in series. Variations shown are due mainly to V_{CO}, which is assumed to lie between 20 and 200 volts.

D-C Power Requirements:

110 microamperes at -90 volts

1 to 2-1/2 ma at -45 volts

Input:

Before triggering, 3500 to 5000 ohms, -18 to -5 volts

During ON time, 200 to 1000 ohms, .01 microfarad with variable voltage.

Input can be built out with a resistor to avoid large steps in input resistance.

Output:

Before triggering, 3500 to 4500 ohms, -36 to -42 volts

During ON time, starts at 200 to 400 ohms, .01 microfarad, -18 to -5 volts respectively. Drifts negative for 10 to 40 microseconds, then snaps negative. Output load reduces the 10 to 40 microsecond ON time.

Delay and rise time of positive output step: roughly 0.1 microsecond.

Signal required for triggering:

Input: Negative current step of 300 microamperes or more with rise time of 10 microseconds or less.

Output: Negative current step of as little as 900 microamperes.

2.410 Additional Circuits for Delay

Perhaps the simplest delay circuit would use the transition at the end of the ON time of a normally off triggerable circuit (such as the amplifier discussed in this report) as the output. This fails in the present case because the ON-to-OFF transition of the amplifier is of wrong sense (at the collector) and is relatively weak in terms of voltage step and internal impedance after the step. Also, it is found that this step has a longer rise time than the OFF-to-ON step in the circuits considered.

It seemed reasonable to take advantage of the stable sensitivity of the present amplifier and to use the amplifier, with its delayed negative step, to control a circuit which would give the desired delayed output. It seemed that a suitable circuit would be a two-state monostable arrangement, stable when the amplifier is stable, and driven to its unstable condition when the amplifier is triggered. At the end of the delay time it should fall back to its stable state and simultaneously give the desired output. A positive voltage step is needed at the end of the delay time. If output is taken from the collector, this means that the added circuit must go ON at the end of the delay time, that is, it must be the normally ON type. This may seem wasteful of d-c supply current, since the circuit would be OFF only a small fraction of the time. However, the

alternatives seem less attractive. Even if we could use the negative voltage step obtained from the collector of a normally off unit, we would have the following trouble: Suppose we needed a 20 ma output pulse lasting two microseconds (such as needed for programming). A transistor can handle such current and duration. But if a transistor is to apply such a current to a load upon switching OFF, it must act by removing a shunt from the load. The transistor must have been passing the heavy current during the period before switching. In the present case this period may last as long as 40 microseconds, which is too long to pass 20 ma through the contemplated transistor. It appears that the step at the collector resulting from the OFF to ON transition has an important advantage for high-power output in that the transistor need pass high current only as long as the load needs high current. This step is a positive voltage step with an n-type transistor.

The primary requirement for the normally ON circuit is that it shall be extremely stable in the ON condition under adverse conditions. The simplest circuit that meets this is one with emitter grounded and base open. This circuit can be pulsed off by means of a positive pulse on the base. This fits perfectly with use of the amplifier to drive it. If the base is capacitor-coupled, the circuit will fall ON of itself after the base drifts negative. If a negative step (such as the delayed negative step from the amplifier) comes before this time, it will go ON in response. During the OFF period the collector drifts negative along with the base, so that a large positive voltage step is available. The complete circuit, which is used in tandem with the amplifier, is shown in Figure 19.

There is one serious fault in the delay circuit. This fault has not prevented its being used and there are circuit means for improving performance. The fault comes about because the collector-to-base resistance remains low for a very short interval of time after the emitter current is cut off. The circuit starts out in its stable ON condition, with about 8 ma flowing emitter-to-base and base-to-collector. The amplifier is triggered and the base is driven positive. The base-to-collector resistance remains low for some fraction of a microsecond. During this time the positive step on the base shows up as a positive pulse on the collector. This pulse is highly variable in its duration and amplitude. It may be as long as a microsecond and as high as 10 volts. The presence of this pulse violates one of the requirements. The pulse can be reduced at some cost in reduced output by clamping the base so that it never goes but a volt or two positive. In the shift register, a stronger than minimum set-to-zero pulse will override the spurious pulse. Also, in the model shift register, the coupling circuits reject a spurious pulse of less than

7 volt amplitude, while passing the desired pulse with little attenuation.

2.411 Delay Circuit Performance

D-C Power Requirements:

7 to 8 ma at -45 volts

Triggering:

Controlled by amplifier output.

(In the complete amplifier-delay, trigger sensitivity is the same as given under amplifier. The complete unit is not sensitive to triggering on its output.)

Output:

Before positive step - about 3000 ohms, OCV -30 to -42 volts.

After positive step - 200 to 400 ohms, OCV -2 to -3 volts.

Transition time - about 0.1 microsecond or less.

2.5 Shift Register Coupling Networks

It is necessary to add a network between each delay output and the set-to-one input of the one-bit register to accomplish several functions. The main function is to transmit a positive step forward while blocking direct current. A series capacitor is used. The negative step at the beginning of the delay period must leak off the capacitor before the positive step comes along; this is in order that the input voltage of the register shall go positive with respect to its quiescent value. Looking out from the register, one should see a high impedance in order that a signal on the set-to-zero input can set the register easily. It is desirable to prevent the negative step at the beginning of the delay from appearing on the set-to-one input. Finally, it is desirable to prevent any spurious positive pulse at the beginning of the delay period from appearing at the set-to-one input when the register is storing a one. (In the case of a stored one it would hinder the set-to-zero. If the register were already storing a zero, no harm would come from its being set to one earlier than normal by the spurious pulse.) All these functions are taken care of by the simple circuit shown on the delay output on Figure 20. The shunt diode dissipates the negative step.

The series diode prevents forward transmission of the negative step and with the -10 volt bias shows a high resistance looking backward from the register. The resistor connected to -10 volts allows transmission of a positive step with no loss when the register is storing a zero, and clips away the lower 7 volts of a step when the register is storing a one.

Coupling between bit-register output and amplifier input must transmit a negative step forward. It should give some a-c isolation (so that change of amplifier input impedance upon triggering will not disturb the register) and also d-c blocking. A series capacitor and resistor as shown in Figure 20 accomplish this.

The coupling in the common set-to-zero circuit from the program generator must transmit a positive pulse to all registers that are in the one condition. It must be such that each bit register sees a high resistance when it is being set to one by the delay circuit. Also, the negative step that appears on the set-to-zero input of a bit register when the stage is set to one should not be transmitted to any other register. The negative step from the program generator (the step or decay following the desired positive step) should be dissipated before the next program pulse occurs. These functions are accomplished by the simple diode and resistance circuit shown in Figure 20.

3.0 Operation of the Model Adder-Shift Register

A model has been built in which an adder is fed by two shift registers, all of the type described herein. The system block diagram is given in Figure 1. One shift register has 5-bit capacity and the other 4-bit capacity. The system can be controlled by a program generator, using transistors for active elements which has been built to operate the system at high speed. Rates of 3000 complete additions (of a pair of digits) per second were easy to attain. This rate is about five times as fast as that required; it was chosen to give a reasonable margin. The program is such that a single pulse fed into the program generator initiates the following events: (1) An immediate negative pulse on the "add" input to the adder; (2) after roughly 50 microseconds, a negative pulse on the "carry" input of the adder; (3) after a further 50 microseconds, a positive pulse on the "shift" lead common to the two shift registers. After a wait of 350 microseconds, or any longer time after the initiation pulse, the system is ready for another initiation pulse.

One way to demonstrate operation is to store two binary numbers in the two shift registers, supply a shift register to store the output of the adder, and supply initiation

pulses slowly by hand. The sum should move into the output register, and remain there if pulses are stopped. Or, one could connect the output of each source register back to its input (say four stages of each source register used) so that the input numbers would not be lost, but would pass through a register over and over. The output to input connection is in addition to the connection from shift-register output to adder input. Then one could supply initiation pulses at high speed and monitor the adder output with an oscilloscope. The sum should be repeated over and over.

A method was worked out which would not only check high speed operation, but would also give a permanent indication in case an error occurred in any operation whatever. This error-detecting setup is a very satisfying way of determining whether the model can run for long periods without a single failure of any part on any operation. Such performance is not essential in the data transmission system, which can tolerate occasional errors, but it is a valuable test. To operate error-detecting, we must choose input numbers a and b such that a + b = 2b (or 4b or some integral multiple of 2b). Suppose we choose a + b = 4b. In binary notation, 4 is 100. The sum (4b) is binary 100 x 1 of the input numbers (b). This means that the sum is the same as one input number except that it is displaced two places towards more significance. Therefore the sum, moved two places towards less significance, is identical with one of the input numbers. For example:

a =	10101
b =	00111
a+b	11100

Now if we recirculate "a" in its register, and supply the sum serially at the proper place in the other register, we can supply as many initiation pulses as there are digits in a, and we will end up with the original numbers a and b standing in the input registers. In general, this will be true only if addition, carry, shifting, and storage have been done with no error. If we supply initiation pulses steadily at high speed, the number b will pass over and over through the register. This can be shown as a pattern on an oscilloscope. However, if there is a single error at any time, the adder will not supply the number 100b, and the pattern on the scope will change permanently.

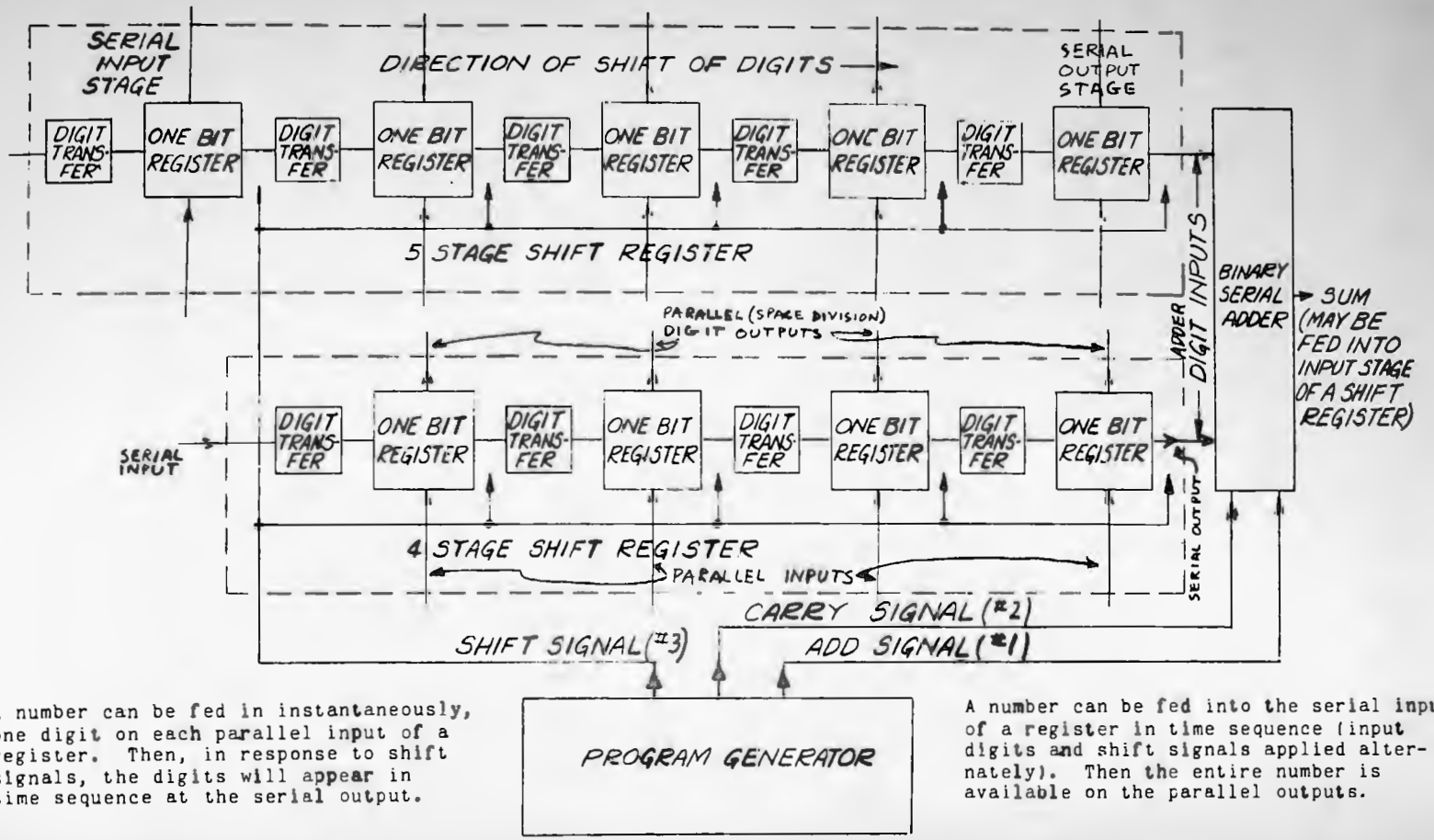
With about half the transistors in the model just on the limits given in Part 2.3, the model has performed many millions of successive operations without error. The limits represent values that transistors can meet over a very wide

temperature range. With these transistors, any supply voltage can be varied about ±15% without causing an error in the high-speed error-detecting operation.

4.0 Conclusions

Transistor circuits have been designed and built which perform very well the functions of binary addition, storage, and space-division to time-division conversion needed in a data transmission system. The power drain is about eight to one better than in present electron-tube design, and it appears that the volume should also be smaller by about the same factor. Even as they stand, the transistor circuits show promise of rather wide use. Considerable insight into systematic design of two-state transistor circuits has been gained.

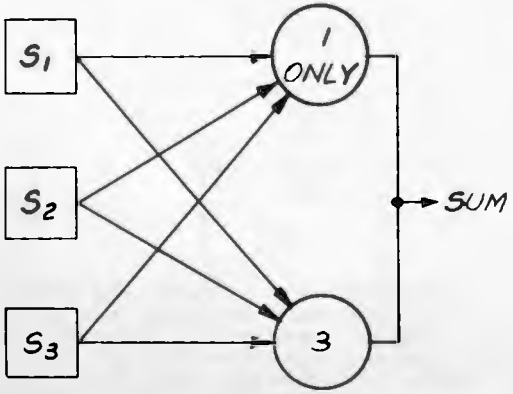
There are several things in the immediate field that need to be investigated as follows: (1) Can a regenerative gate be a better shift register interstage than the amplifier-delay? (2) What is the upper limit on speed of operation of the one-bit register, and why? (3) Can temporary storage be designed to be more attractive than permanent storage? (4) What is the expected life of transistors operating in the present register-adder circuits at various ambient temperatures?



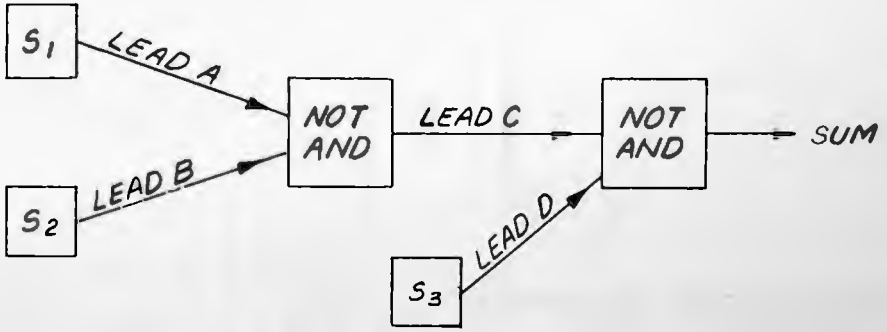
A number can be fed in instantaneously, one digit on each parallel input of a register. Then, in response to shift signals, the digits will appear in time sequence at the serial output.

A number can be fed into the serial input of a register in time sequence (input digits and shift signals applied alternately). Then the entire number is available on the parallel outputs.

Fig. 1. Functional Diagram of Model Binary Adder - Shift Register Systems



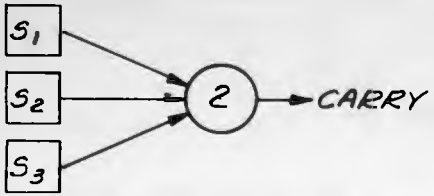
I One-step or Von Neumann type.
This requires a Threshold "1 only"
and a Threshold 3 Gate.



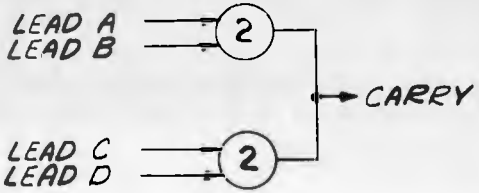
II Two-step or Boolean type.
This requires two "Not-And" Gates.

Two methods of simple addition of three digits (such as two input digits and a third digit representing carry from a previous operation).

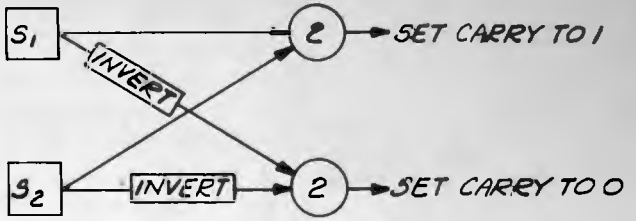
Fig. 2. Block Diagrams of Von Neumann and Boolean Methods of Addition



I Carry especially applicable to the one-step adder. This requires a single three-input Threshold 2 Gate.
Note: The three-input Threshold "1 only" function in the one-step adder can be realized by using tandem gates in such a way that the three-input Threshold 2 Gate needed for carry is also produced.



II Carry applicable only to the two-step adder.
This requires two two-input Threshold 2 Gates.



III Carry applicable to either one-step or two-step adder. This requires two two-input Threshold 2 Gates ("Hussey Carry").
Note: The invert function can be dispensed with by using symmetrical (push-pull) output from S_1 and S_2.

Three methods of obtaining carry. S_1 and S_2 are input digits, and S_3 is previous carry digit.

Fig. 3. Functional Diagram of Three Methods of Obtaining Carry

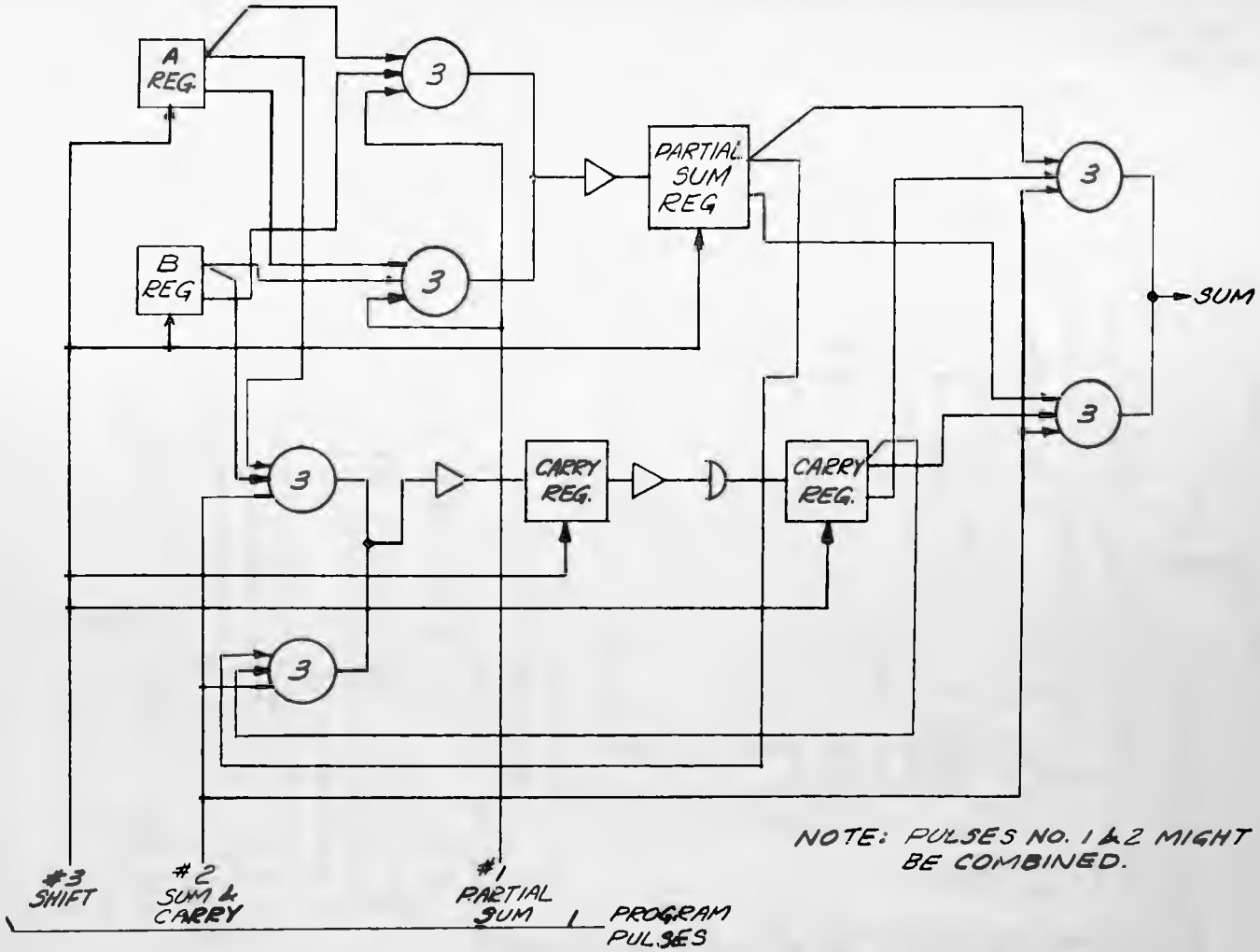


Fig. 4. Functional Diagram of Two-Step (Boolean) Adder

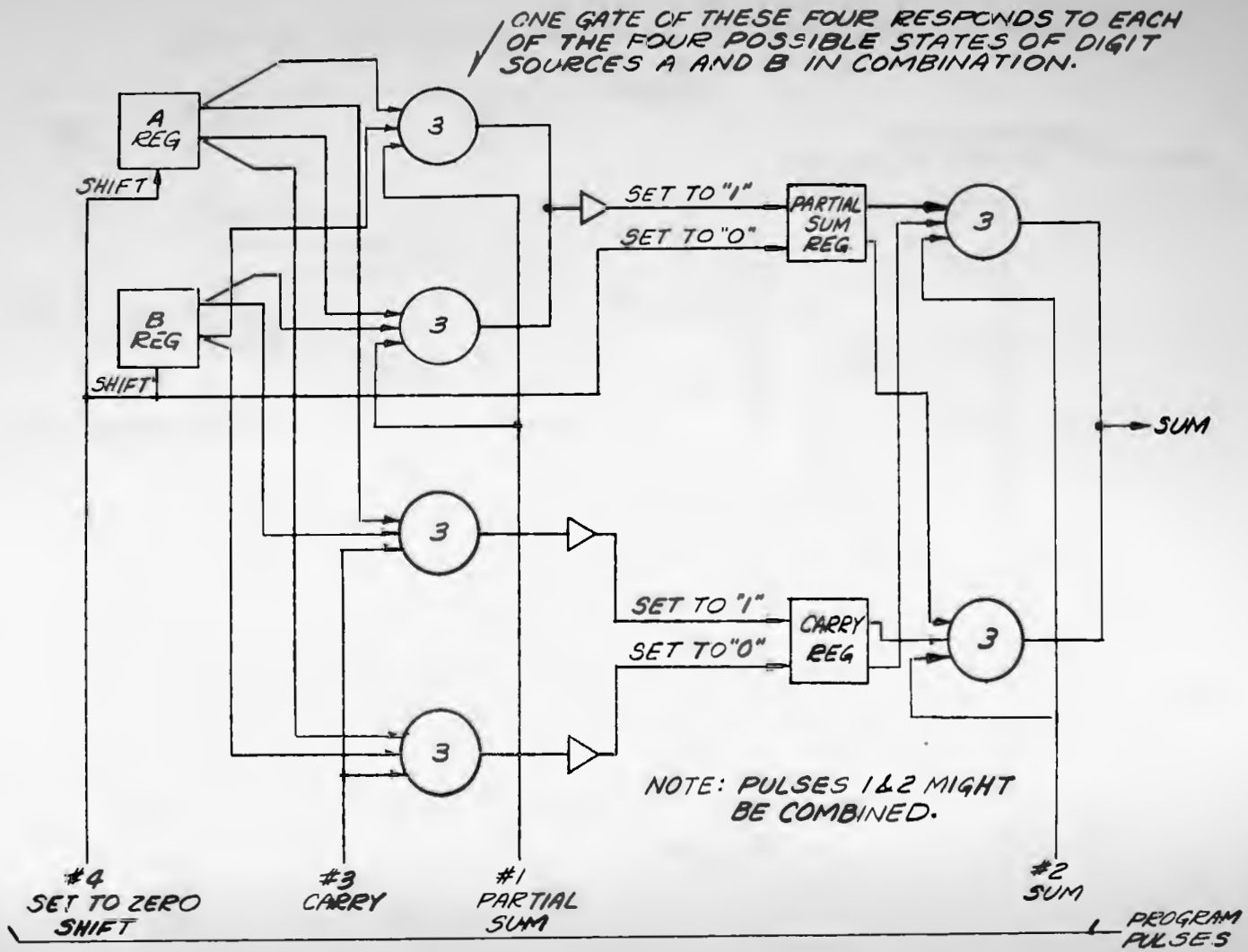
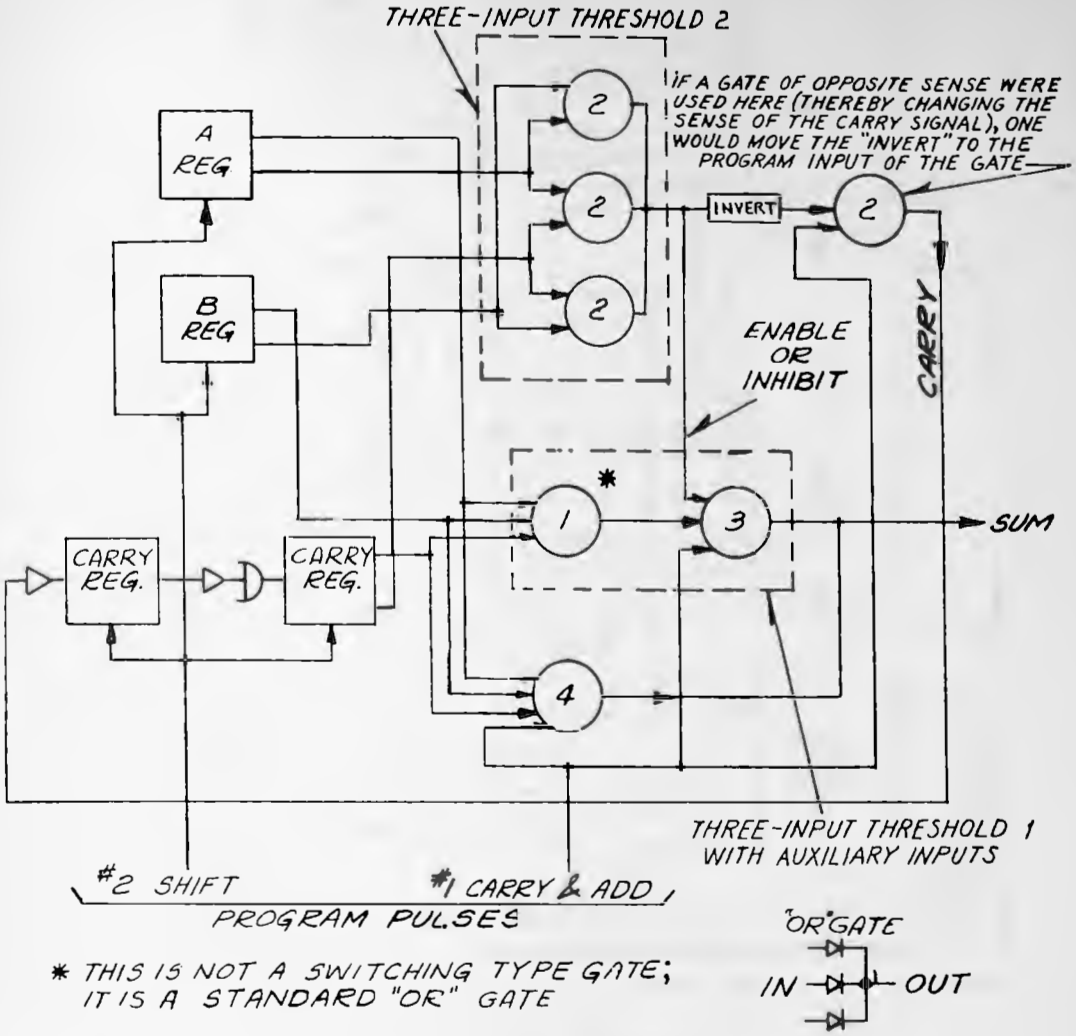


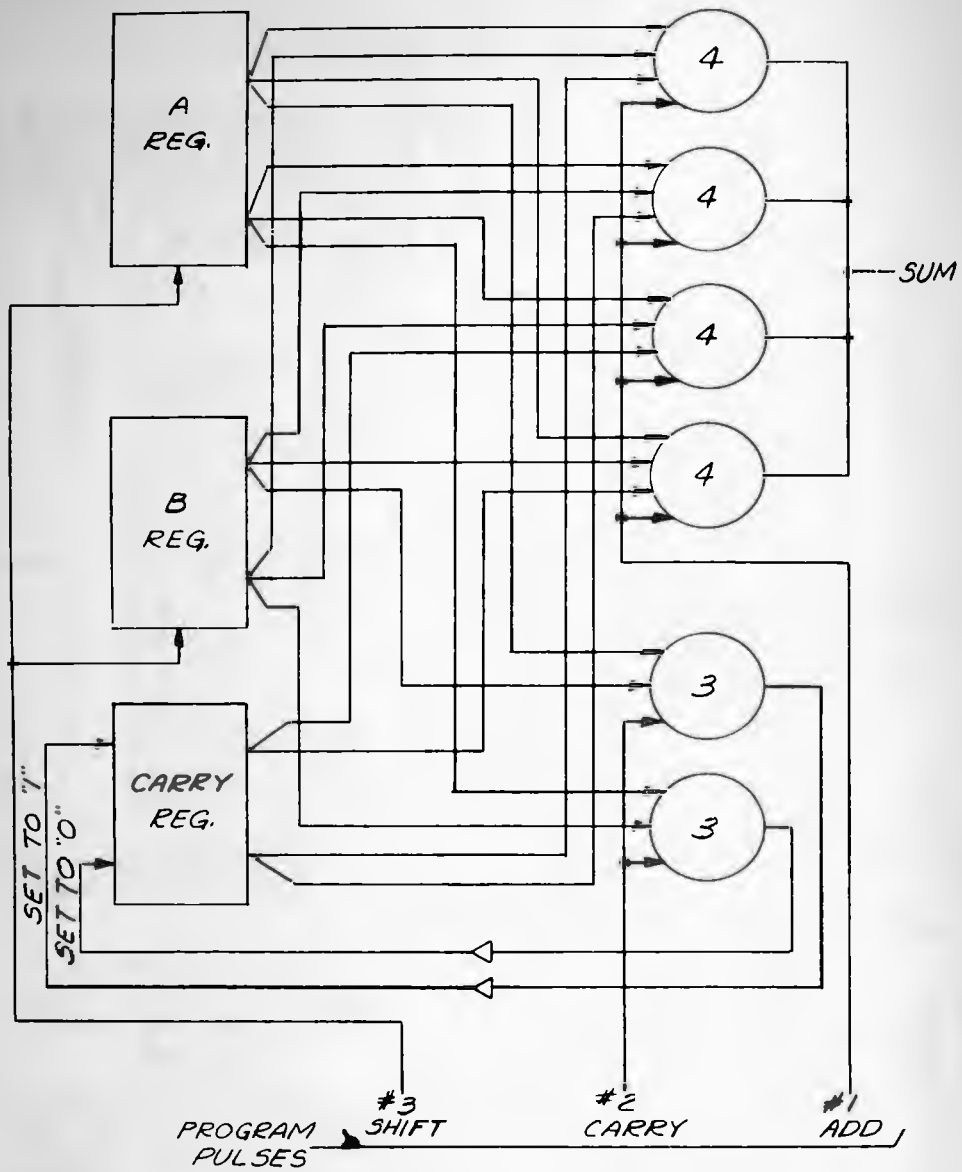
Fig. 5. Functional Diagram of Two-Step Adder With Hussey Carry



NOTES:

1. The (three-input Threshold 2) and (three-input Threshold 1) together perform the "three-input Threshold 1 only" function, and also produce the carry.
2. The three gates of the (three-input Threshold 2) function can be chosen either all positive responding or all negative responding with no other changes in the system. The basic reason is that the threshold is numerically one half the number of significant input states. The four significant states of the three registers feeding the gates are: no "ones"; a single "one"; two "ones"; and three "ones".

Fig. 6. Functional Diagram of One-Step (Von Neumann) Adder



NOTE: The gates shown are realized by "switching-type gate" circuits. In all gates there is one extra input and one higher threshold (beyond the basic needs of add and carry) to take care of programming.

Fig. 7. Functional Diagram of One-Step Adder With Hussey Carry

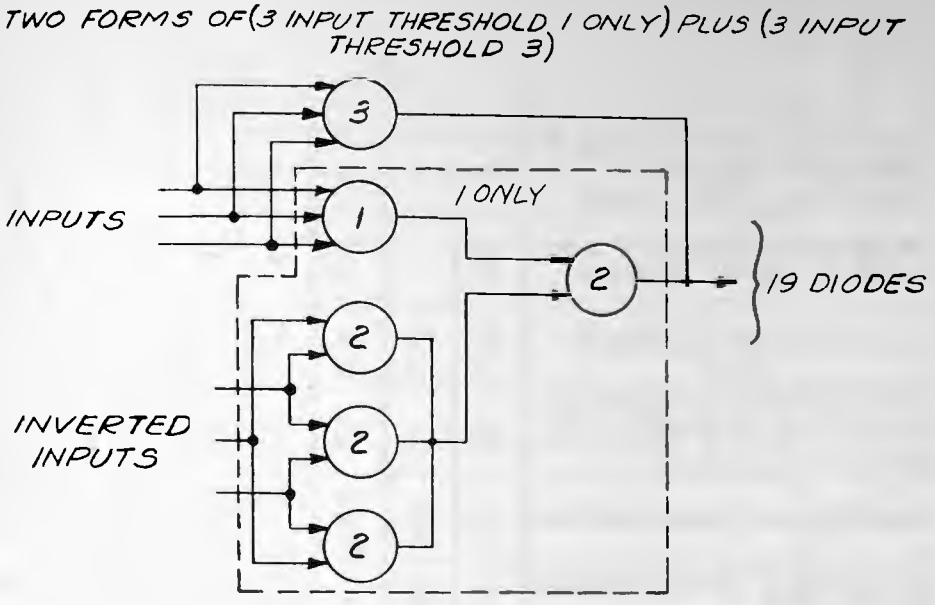


Fig. 8A. Functional Diagram of Three-Input, Threshold 1 Only, Gating Function

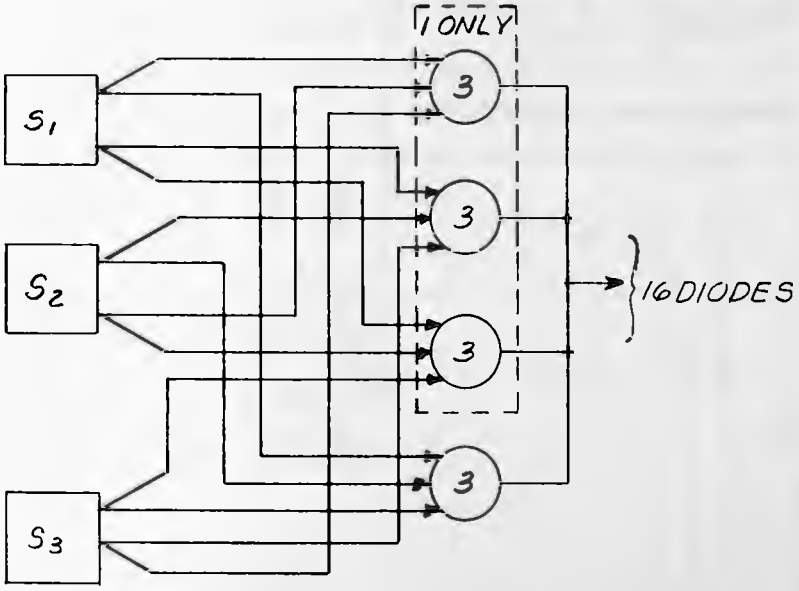


Fig. 8B. Functional Diagram of Three-Input, Threshold 3, Gating Function

	NUMBER OF REGISTERS	NUMBER OF PULSE AMPLIFIERS	NUMBER OF DELAYS (INCL. AMP)	NUMBER OF PROGRAM PULSES	MAX NUMBER OF GATES ON A SINGLE PROGRAM PULSE	NUMBER OF BIAS CURRENT SUPPLIES (INCLUDING TANDEM GATE LOADS)	NUMBER OF TANDEM GATES	MAX NUMBER OF GATE BIAS CURRENTS IN A REGISTER OUTPUT	NUMBER OF DIODES	*RELATIVE QUANTITY REG., AMP. & DELAY PARTS	REGISTERS ON A SINGLE PROGRAM PULSE (INCL. A&B DIGIT REG.)
2 STEP WITH USUAL CARRY	3	2	1	2 OR 3	6 OR 4	6	NO	$1\frac{1}{2}$	24	$9\frac{1}{2}$	5
2 STEP WITH HUSSEY CARRY	2	3	0	3 OR 4	4 OR 2	6	NO	1	24	7	3
1 STEP WITH USUAL CARRY	2	1	1	2	3	8	YES	2	24	$6\frac{1}{2}$	4
1 STEP WITH HUSSEY CARRY	1	2	0	3	4	6	NO	$1\frac{1}{3}$	28	4	2

* REGISTER WEIGHTING = 2
AMPLIFIER WEIGHTING = 1
AMPLIFIER-DELAY WEIGHTING = $1\frac{1}{2}$

Fig. 9. Tabular Comparison of Adders

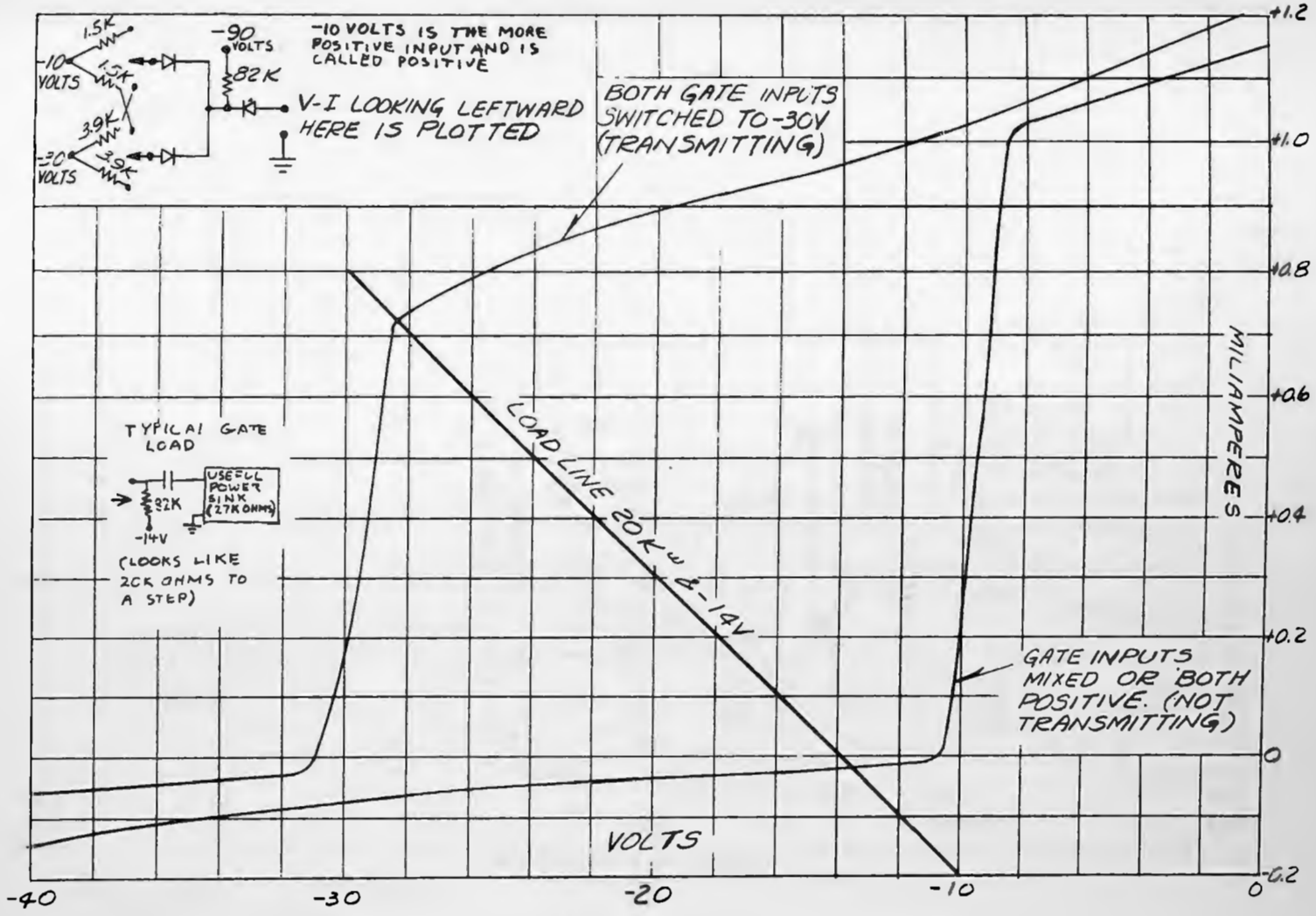


Fig. 10. Gate Output V-I Characteristic

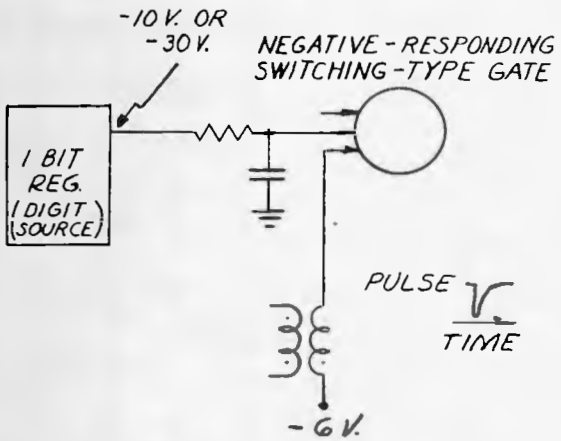


Fig. 11. Circuit of R-C Filter for Reduction of Loading of Digit Source

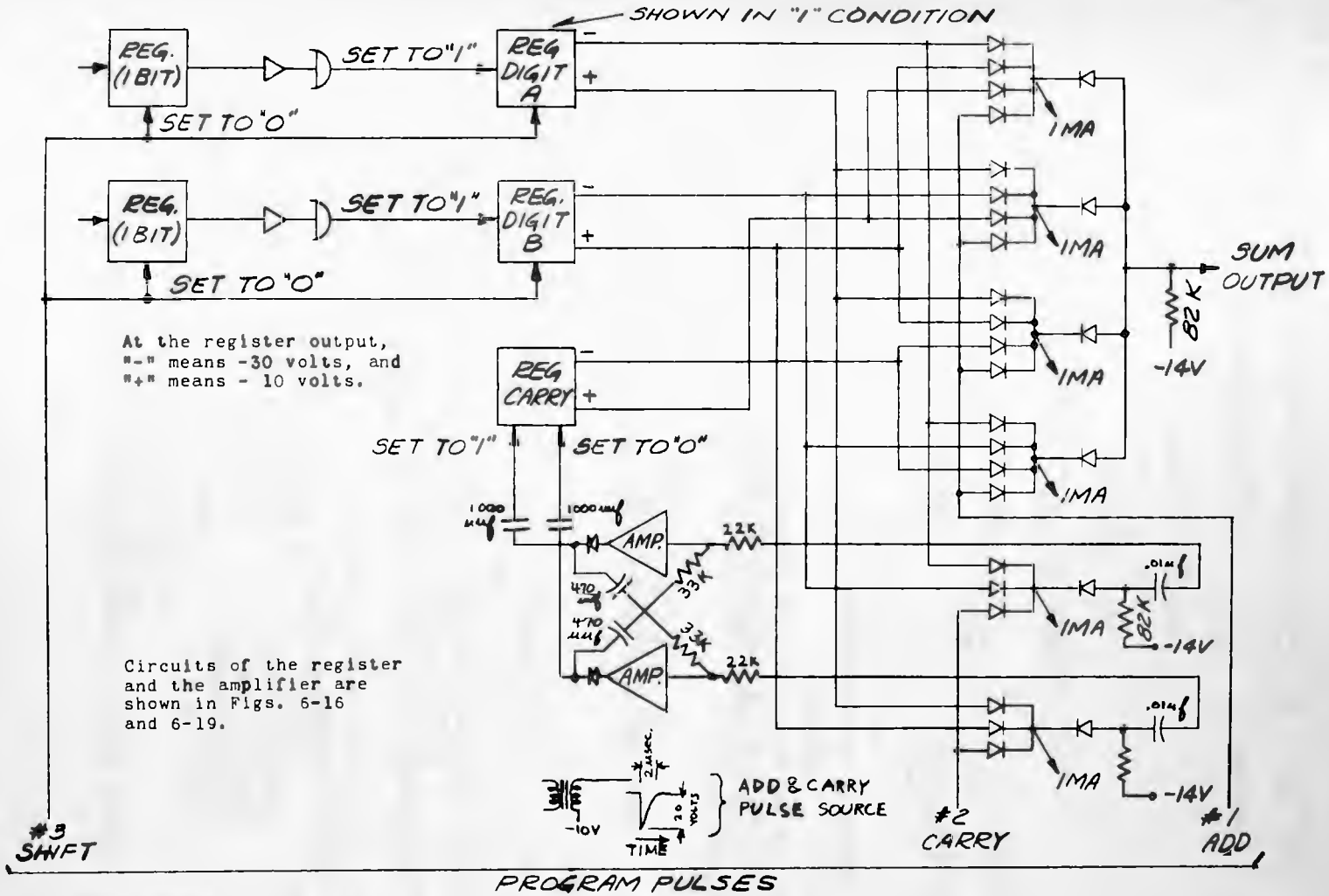


Fig. 12. Block Diagram of Model Adder

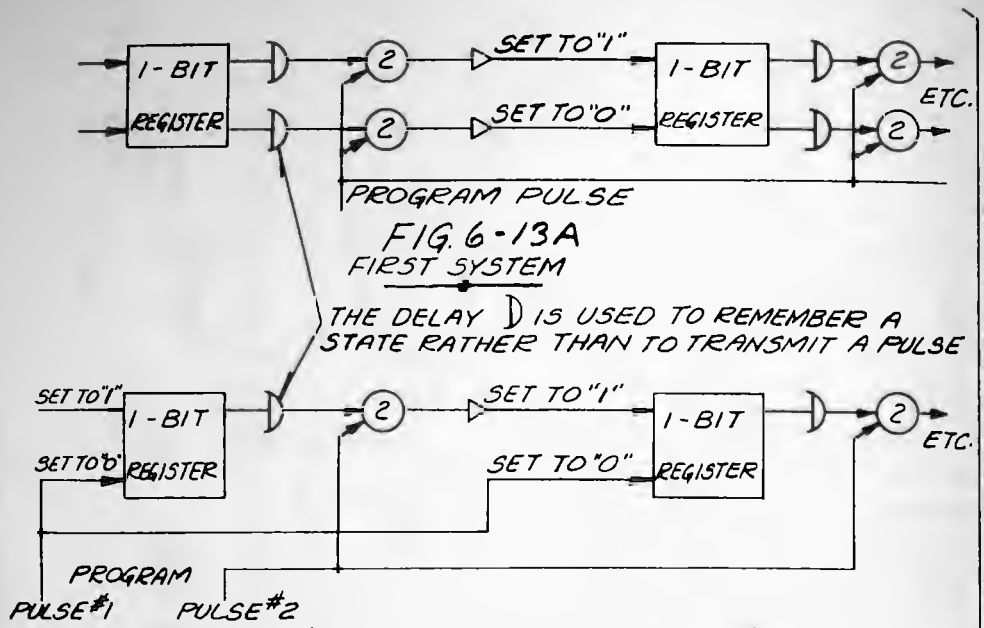


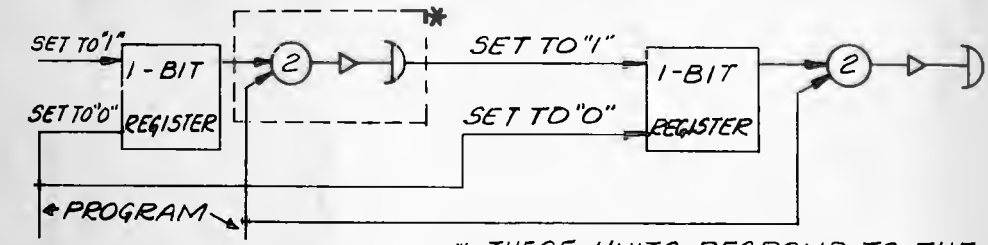
FIG. 6-13A
FIRST SYSTEM

THE DELAY D IS USED TO REMEMBER A STATE RATHER THAN TO TRANSMIT A PULSE

DELAY OF PULSE #2 WITH RESPECT TO PULSE #1 MUST BE LONG ENOUGH TO LET REGISTERS GO TO ZERO AND MUST BE SHORTER THAN THE LENGTH OF TIME THAT D CAN REMEMBER.

GATES IN THESE FIGURES WILL RESPOND TO A PROGRAM PULSE WHEN THE SIGNAL ON THE NON-PROGRAM INPUT OF THE GATE IS THE SAME AS THE SIGNAL ON THE UPPER RIGHT OUTPUT OF A REGISTER WHEN THE REGISTER IS STORING A "1".

FIG. 6-13B1
SECOND SYSTEM
WITH DELAY CONTROLLED BY PROGRAM



* THESE UNITS RESPOND TO THE STEP OR PULSE PRODUCED BY THE CHANGE OF STATE OF THE REGISTER FROM "1" TO "0". D MUST BE SHORTER THAN REPITITION PERIOD OF PULSES.

FIG. 6-13 B2
SECOND SYSTEM
WITH DELAY CONTROLLED BY INTERSTAGE

Fig. 13. Functional Diagrams of Digit Shifting Systems

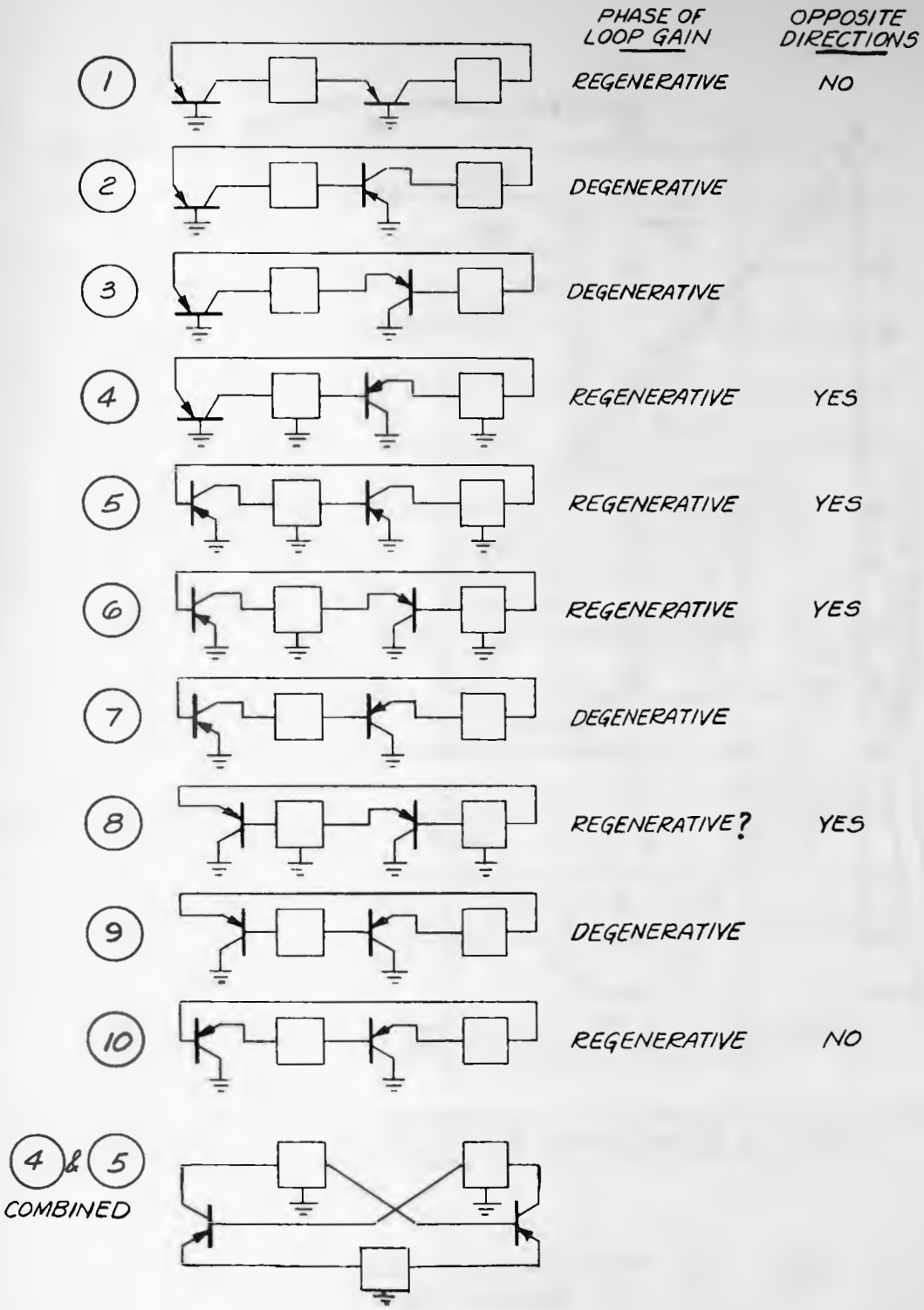
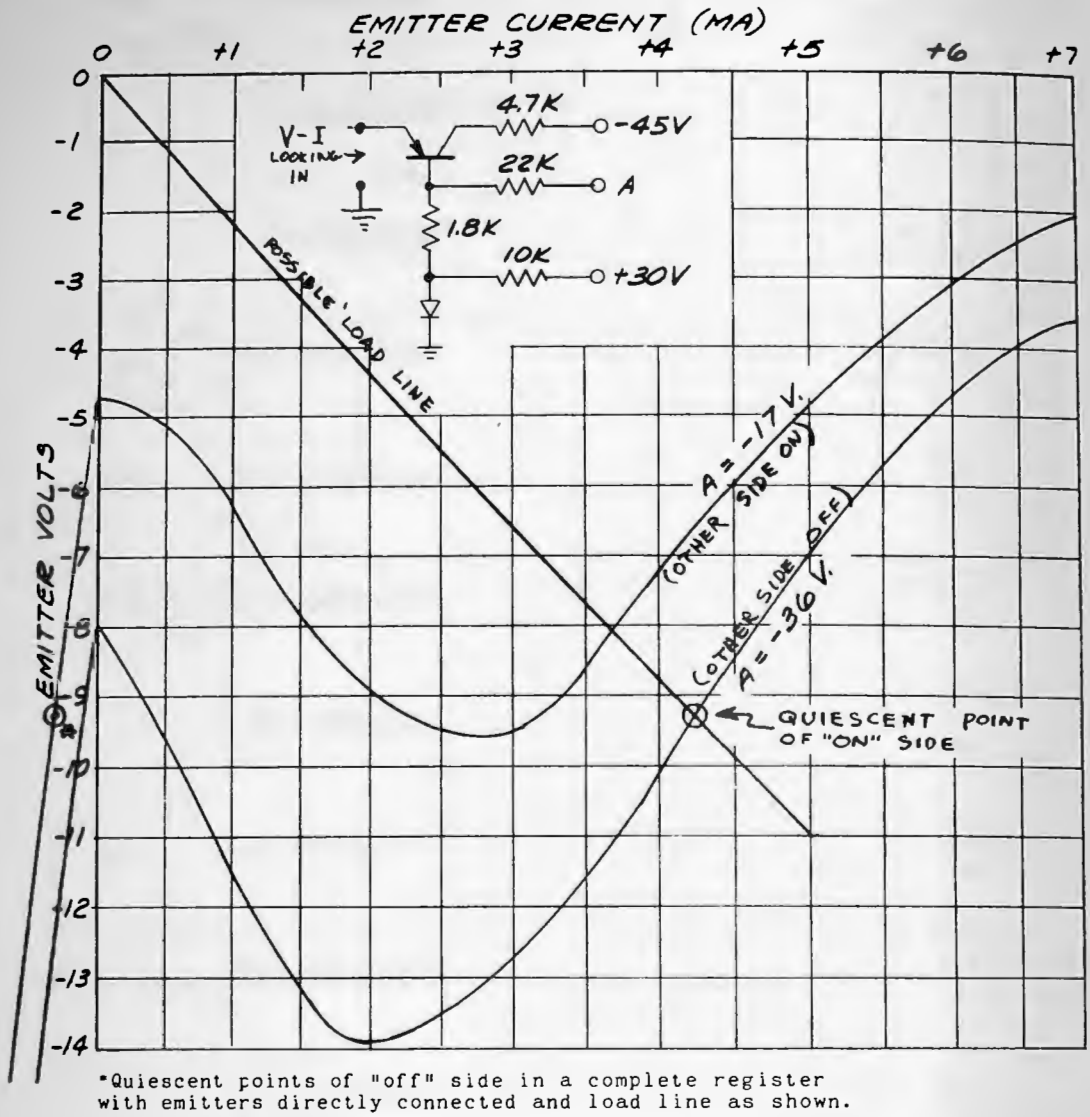


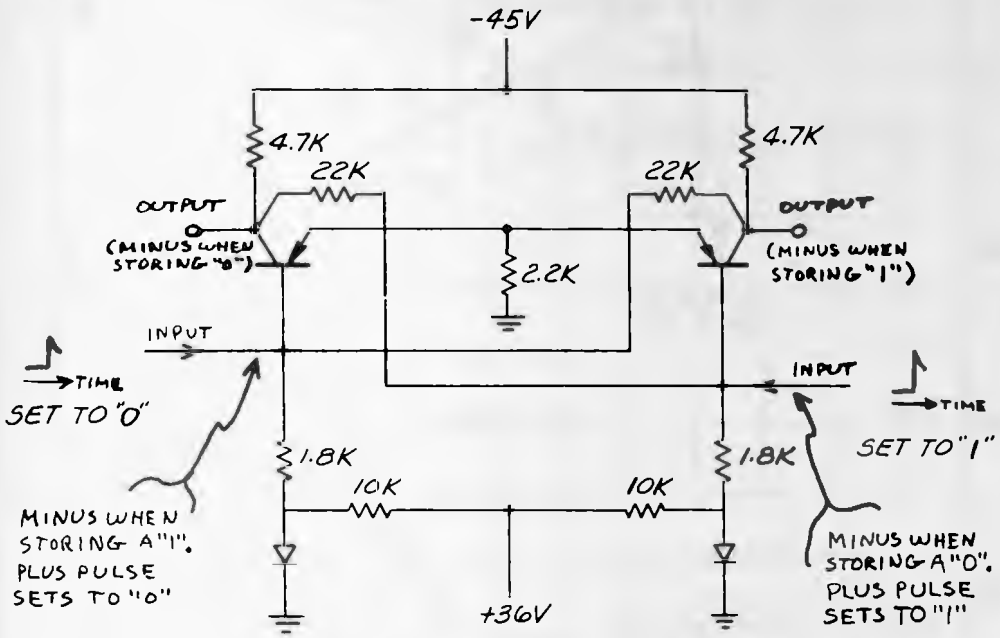
Fig. 14. Basic Circuit Configurations of Two Transistors Each With One Element Grounded



*Quiescent points of "off" side in a complete register with emitters directly connected and load line as shown.

NOTE: The circuit represents half of a symmetrical two-state register. (This is an early set of measurements. "A" should be nearer -10 volts and -30 volts in the respective curves, and the positive supply should be +36 volts.

Fig. 15. Characteristic Emitter V-1 Curves for Register Circuit



NOTE: The open circuit voltage of an input is either about -3 volts or about -10 volts. The -10 volt condition is called "minus."

Fig. 16. Circuit of One-Bit Register

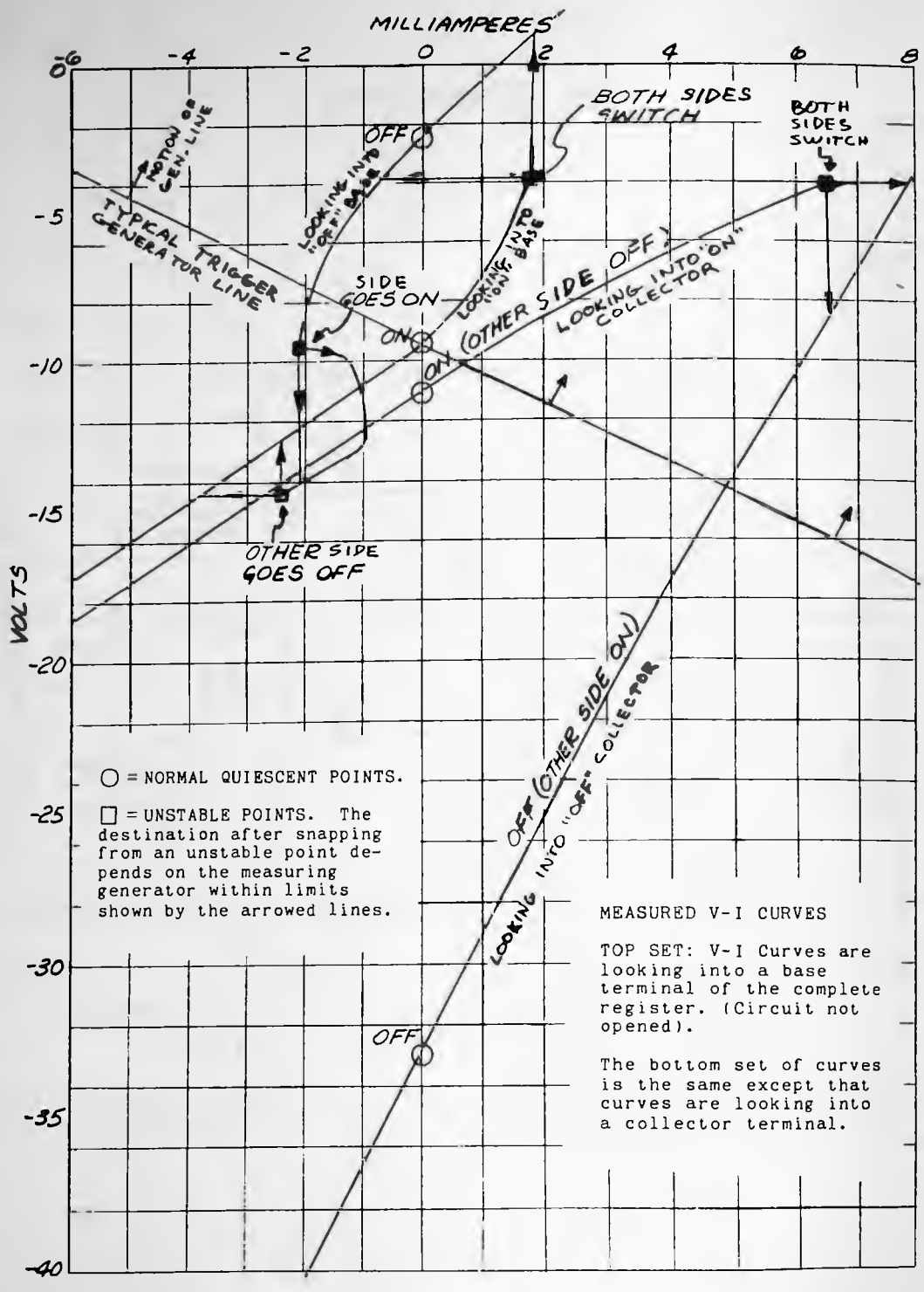
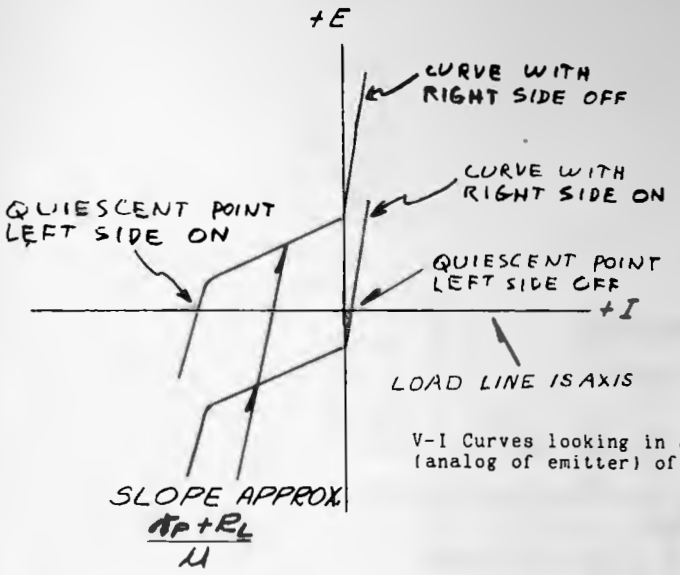


Fig. 17. Characteristic Curves for Triggering One-Bit Register



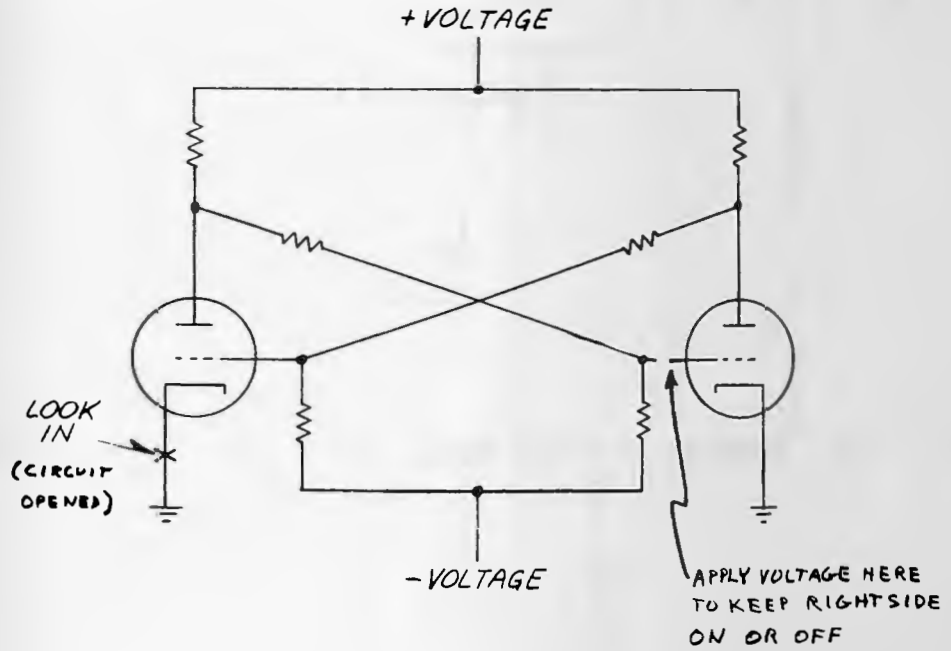


Fig. 18. Circuit and Characteristics of Symmetrical Electron-Tube One-Bit Register

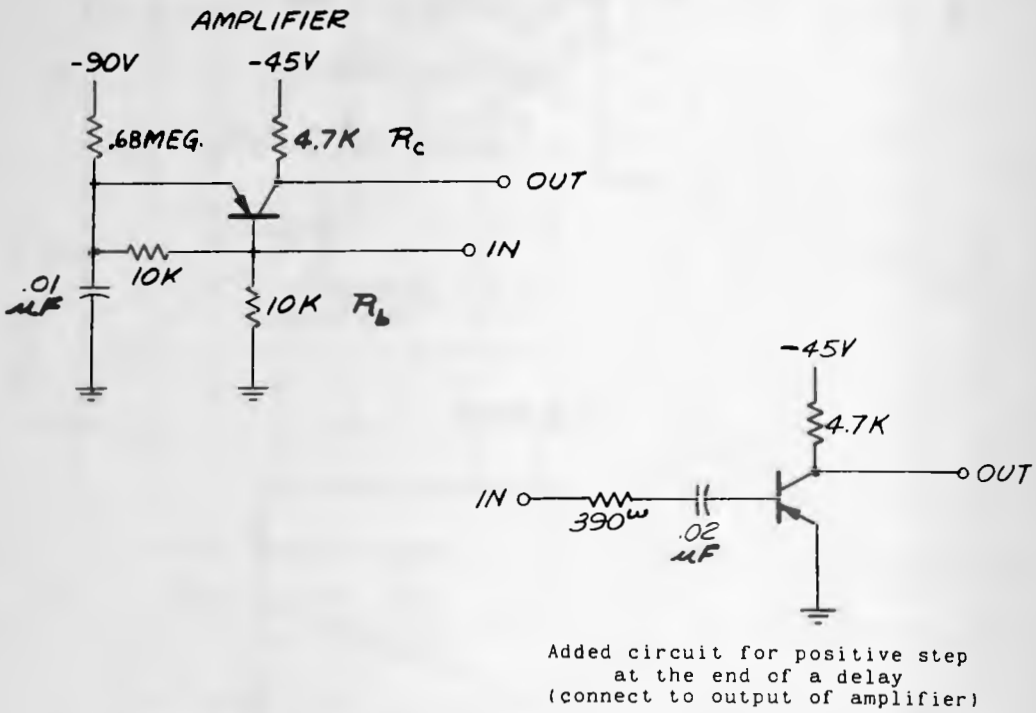
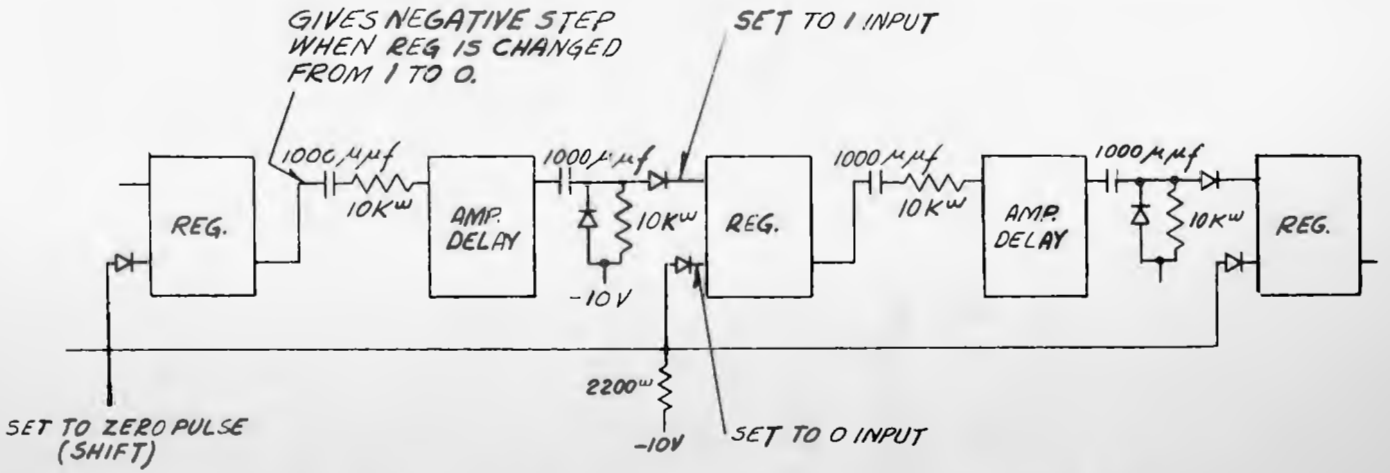


Fig. 19. Circuits of Pulse Amplifier and Amplifier Delay



Register and amplifier-delay circuits are shown in detail in Figs. 6-16 and 6-19.

Fig. 20. Coupling Circuits for Shift Register

A PACKAGED ANGULAR POSITION ENCODER

H. G. FOLLINGSTAD

Produced under contract W 36-039 sc-44497

A PACKAGED ANGULAR POSITION ENCODER

1.0 Introduction and Purpose

In a previous study, a zero order system analysis of an optical angular position encoder was made in which three alternate forms of such an encoder were discussed and compared. As a result of the analysis two encoder types (B and C) were selected for zero order circuit, device, and package phase investigation. Although changes did not alter the general system functions, Encoders B and C were relabeled D and E, respectively, for convenience in identification. First order system feasibility studies of Encoders D and E (and their predecessors B and C) were conducted on the basis of analog-to-digital encoding of the least significant digit of a 15 digit code wheel system.

During the course of those investigations it was determined that final feasibility proof would include demonstration of the angular position encoder using miniaturized functional transistor packages wherever possible. In particular, the feasibility demonstration was to include the analog-to-digital encoding of the 6 least significant digits of a 15 digit "Gray" code wheel, storing of the 6 parallel code digits in a shift register, conversion of the "Gray" code digits to binary form, and finally the shifting of the resulting binary code into a dummy line in time sequence.

The purpose of this report is to summarize and evaluate the results of first order studies of the angular

position encoder (type E) in all phases with respect to its use as the encoding function in the data transmission system specified above.

2.0 System Phase

2.1 Encoder E

The basic function of Encoder E is to translate the angular position (analog information) of an input shaft into a parallel digital code in the form of electrical pulses by photoelectric reading means. The separate block functions to accomplish this encoding process are shown in Figure 3-1. A light source provides an electronically pulsed light of high intensity with a duration short enough to decay before one quantum(3) of the Gray code wheel has passed a radial defining aperture at the fastest desired angular rate of motion.(5) Such a light pulse can act as a reading gate to sample the instantaneous angular position of the code wheel or shaft. Behind the one-half quantum radial defining apertures are placed p-n junction photocells to obtain an electrical reproduction of the light pulse as seen through the code elements and the defining aperture.

A typical code resolution envelope of peak p-n photocell output current versus relative position of the code wheel is shown in Figure 3-2. The envelope deviates from an ideal rectangular (on-off) form because of diffraction effects and the finite width of the defining aperture and light source. In addition, light leakage in the least significant digits is sufficient to create p-n photocell pulse currents in opaque code positions. To minimize the uncertainty in resolution of a quantum step, a transistor regenerative pulse amplifier is

- -

(3) One Quantum is $\cong$ .001 inches on a 15 digit code wheel at a radius of 5 inches.

(5) A 90°/sec. angular rate is equivalent to approx. 8000 quanta/sec. in a 15 digit system. One quantum passes the defining aperture in 125μ sec.

employed which will give an output pulse only when its trigger threshold is exceeded. Ideally, this threshold should be equal to the peak p-n current pulse amplitude occurring at equal on-off code distances as shown on Figure 3-2. An increase in trigger threshold (light intensity and p-n sensitivity constant) is seen to decrease the "1" or transmission distance and increase the "0" distance, whereas a decrease in trigger threshold reverses those changes. Similarly, a decrease in peak light intensity and/or p-n photocell sensitivity (trigger threshold constant) causes a decrease in the "1" or transmission distance.

Therefore, a basic system problem is to properly divide the operating margins between peak light intensity, p-n photocell current sensitivity, and trigger threshold of the regenerative pulse amplifier to insure that under specified temperature changes and aging changes, the combination of these variables shall not cause loss of resolution of any single quantum step. Studies indicate that variability is least in the p-n sensitivity, greater in the triggering threshold, and greatest in the peak light intensity, or that the present limitation is the life of the light source. A discussion of these variables is given in Sec. 3.

Encoder E was preferred over Encoder D(7) as the type to be used for a final feasibility demonstration for the following reasons:

1. The problem of readout gating is simpler.

2. The peak light intensity is much higher than in D, relaxing the requirements on photocell and regenerative pulse amplifier sensitivity.

3. The light source does not heat the p-n photocells appreciably, while heating effects in D were a major problem.

It should be noted, however, that Encoder E may be restricted to slower speed systems than is D because of the present life limitation on arc discharge flash tubes using high repetition frequencies.

- - - -

A final system requirement for Encoder E is that the output pulse for each Gray code digit have an amplitude and energy sufficient to operate reliably suitable output circuits for eventual transmission as time sequential binary code. First order studies resulted in design of a particular Gray to Binary Converter-Shift Register configuration which set the exact encoder output requirements. A discussion of this configuration is given in a report (8) and a list of Encoder E system characteristics based on use in the system is compiled in Sec. 1A, Appendix I.

2.2 Gray-to-Binary Converter - Shift Register

Gray to binary conversion(8) is basically a "NOT AND" comparison process performed in the following manner:

The most significant digit (x) in Gray code is the same as the most significant digit (x) in the binary code. The resulting (x) binary digit is compared with the next most significant (x-1) Gray digit on a "NOT AND" basis. That is, if the (x) binary and the (x-1) Gray digits are alike (o,o or 1,1), a "0" will result for the (X-1) binary digit, but, if unlike (o,1 or 1,0), a "1" will result for that binary digit. This process is continued down to and including the least significant digit, resulting in the binary equivalent of the original Gray code. Various circuital methods are available for accomplishing the conversion process. A unique arrangement was developed by R. E. Yaeger which combined the conversion process with the shift register function.

2.3 Programming

The flow chart of Figure 3-3 shows how the encoding function (Encoder E) supplies parallel Gray code information to the converter - shift register. The operation of the system with programming may be described sketchily but simply as follows:

A "READ" pulse applied to the light source causes instantaneous sampling of the Gray code digits on the simulated code wheel, which digits appear at the encoder output in the form of parallel Gray code pulses and are stored simultaneously in the bit registers. After stabilization of the register

- -

(8) "A Gray to Binary Translator and Shift Register", by R. E. Yaeger

circuits, a CONVERT pulse applied to the regenerative gates through appropriate delays (Regen. Amp.) permits the "NOT AND" comparison to take place, with the result that the stored Gray digits are converted into stored binary digits. After this process is complete, a series of SHIFT pulses applied to the bit registers shifts the binary digits out of the bit registers in time sequence. The program pulse train is shown in Figure 3-4 with a sequence of READ, CONVERT, and SHIFT, respectively. The time interval between the READ and CONVERT pulses is expanded to show the timing delays in Encoder E and the shift register which must elapse before the CONVERT pulse may be applied. The time required for the photocell current pulse to reach its peak value is about 5-15 μ sec. after the READ pulse. Storage delay in the input of the regenerative pulse amplifier adds 15-25μ sec. to the former delay with a resulting total delay of 20-40μ sec. between the rise of the READ pulse and the rise of an encoder output pulse. After the encoder output pulses are applied to the bit registers, stabilization time is required for the bit register and pulse steering circuits, after which the CONVERT pulse may be applied.

Specific programming requirements are given with the description of converter-shift register operation.

3.0 First Order Circuit and Device Phase Results

3.1 Encoder E

3.11 Light Source

The light source for Encoder E (see Figure 3-10) consists of a high intensity arc discharge lamp (SA309) which supplies a short pulse of light when triggered. The light pulse has a rise time of about 5μ sec. and a duration of 40μ sec.(9) Breakdown current for the gap is supplied by a 600 volt, 20 ma. power source. The lamp remains deionized, however, until an external high voltage trigger coil encircling the breakdown path supplies the rate of change of field necessary to break down the gap. A pulse modulator consisting of an 0A5 thyratron and a pulse transformer, when triggered by a sampling or read-out pulse, supplies the HV breakdown pulse. Deionization is accomplished with an RC circuit

---

(9)See footnote 3 for a typical pulse duration requirement.

consisting of R6 and C3. (R6C3 = .02 sec. or one-fifth of the repetition frequency, allowing almost complete charge between flashes.)

An equipment design precaution must be observed in that the modulator and arc discharge lamp must be placed in one housing to prevent external radiation from the high current and high voltage carrying components and leads. Such a precaution reduces external noise voltage by about 40 db and makes it possible to eliminate shielding of the p-n photocell, regenerative amplifier, and output circuits.

The life of the arc discharge lamp is probably the limiting factor in encoder reliability of operation. Life studies of the SA309 are in process using a sampling rate of 10 pps, and a p-n photocell (calibrated under standard optical conditions) to measure peak light intensity at prescribed time intervals. At this writing one sample has had a half-amplitude life of about 700 hours. A reference list of light source characteristics and specifications is compiled in Sec. 2A, Appendix I.

3.12 Code Wheel Simulator and Optical System

The code wheel simulator reproduces in all essential respects the specifications set for a 15 digit angular position encoder by the Signal Corps.(10) A glass code plate containing 64 quantum steps of .001 inch each and employing the 6 least significant digits of the 15 digit reflected binary code was obtained by photographic reduction of a greatly enlarged master drawing. The code plate (See Figure 3-5) and a .0005 inch engraved defining aperture are mounted on a micromanipulator as shown in Figure 3-6. Lateral motion of the code plate past the defining slit is measurable to the nearest .0001 inch (1/10 quantum) by means of an enlarged micrometer scale. Provision is made for mounting the photocell and defining slit assembly in a holder permitting alignment of the defining slit parallel to the radial code elements. Micrometer adjustment of the distance between the surface planes of the code plate and defining slit makes possible the simulation of wobble effects in an actual code wheel. This distance is specified to be .010 ±.005 inches. Distance between the emulsion side of the code plate and the protruding end of the SA309 bulb (Figure 3-10) has been set at 5/8 ±1/16 inch on

- -

(10) "Angular Position Encoder", Signal Corps Tentative Specification, No. SCL-1210, January 10, 1949.

the basis of tests for optimum code resolution. A detailed reference list of simulator-optical system characteristics and specifications is compiled in Sec. 3A, Appendix I.

3.13 P-N Junction Photocell

The rectangular geometry of the sensitive area of the p-n photocell and its small size make the cell ideally suited for alignment in this encoder, where the requirement is that active areas be .1 inch apart along a single straight .0005 inch defining slit. A typical 6 digit package less defining slit is shown in Figure 3-7. The potting of 6 p-n photocells in a bio-plastic package (coded M1749-1) has been successful. In a representative sample, no perceptible change in ac current sensitivity occurred after potting. The electrical stability with age of cells in the package appears excellent, and although not tested under extreme variations of temperature, it is expected from earlier tests on individual p-n cells that ac current sensitivity will not be affected substantially.

The p-n photocell package with its connecting circuits is indicated in Figure 3-10. Bias is supplied to each photocell through bias resistors R8 to R13. The bias resistance value was chosen to be small in comparison to the high reverse impedance of the p-n cell but large enough not to load appreciably the amplifier input impedance of about 15 kilohms.

The photocell output currents are coupled to the regenerative amplifiers through condensers C5 to C10 to insure that slow dark current increases with temperature will not trigger the amplifiers falsely and produce spurious encoder output pulses. If this restriction were not present, photocell bias could have been supplied directly by the negative d.c. voltage at the regenerative pulse amplifier input. An advantage accruing from the use of a pulsating light source is that the ac sensitivity characteristics of the p-n photocell, which are fairly constant with temperature, can be utilized.

A reference list for p-n photocell, package, and connecting circuit specifications is compiled in sections 4A and 6A, Appendix I.

3.14 Regenerative Pulse Amplifier

In previous first order studies the particular regenerative amplifier employed was of the type shown in Figure 3-8. The general operation of this circuit has been described in a discussion of base stabilization of negative impedance characteristics of transistors.(12)

Negative emitter bias is obtained by a current through R_1 which opposes and is greater than I_{co}, the collector current existing with zero emitter current. The negative bias voltage corresponding to the trigger threshold is approximately $I_{bias} R_1 - I_{co} R_1$ or

$$R_1\left(\frac{E_b}{R_2+R_1} - I_{co}\right).$$

It is evident that variation in trigger threshold is very dependent upon changes in I_{co} when I_{bias} is comparable to it in magnitude. This condition exists at low trigger thresholds and is a disadvantage since I_{co} in switching type transistors is somewhat variable from unit to unit and with time and temperature. In previous encoder studies using this circuit, low trigger bias voltages (E_{R_B} or 4 volts) were required; in those cases the variation in trigger threshold between units was 2/1 or more. Such a variation was considered too severe to permit its use as a level reading device in the encoder (Figure 3-2).

The monostable multivibrator circuit of Figure 3-9, however, has a trigger threshold which is largely independent of I_{co} changes(13). The bias current flowing through R_1 is essentially constant, determined largely by the values of E_b and R_1, since R_1 is much higher than the transistor input impedance to ground. Trigger threshold is stabilized by shunting between emitter and base a resistance R_2 whose value is low in comparison with the back resistance of the emitter. Variation in trigger threshold is caused by a complex combination of currents flowing through collector to emitter leakage resistance and base resistance but is usually small. In use with trigger thresholds suitable for encoder operation, a

---

(12) "Idealized Negative Resistance Characteristics of the Transistor", by R. L. Trent.

(13) "A Stabilized Transistor Delay and Switching Circuit", by A. E. Anderson.

maximum threshold variation of ±25% is typical (see Figure 3-9) and is tolerable as a level reading variation.

The triggering characteristics of this circuit may be described using the emitter characteristics as shown. With negative emitter bias the input impedance is high (~15,000 ohms) and remains high until the triggering point is reached. At the transition a zero impedance voltage source (a Condenser C) is needed to supply current during the short but finite zero impedance region. In order that C have no effect on the critical triggering value, its value should be such that it would be almost completely charged to the desired triggering value during the time of the p-n current pulse. For the type of p-n pulse in Encoder E that value of C is about 300 μμf, but such a value results in an output pulse of rather short duration (1μ sec.). To increase the output pulse energy somewhat at a slight cost in input sensitivity, C was raised to 500 μμf. The resulting average dynamic and quiescent characteristics are shown in Figure 3-9. The output pulse maintains a minimum amplitude V_2 for an average duration of 4μ sec. which is sufficient to operate the bit registers of the converter-shift register. An additional regenerative amplifier of the same type but with a much larger C may be used in tandem in applications requiring a longer output pulse. The regenerative amplifier of Figure 3-9 has been coded M1733-2 and packaged using the Auto-Sembly technique. A reference list of its characteristics is compiled in Sec. 5A, Appendix I.

3.2 Converter-Shift Register

The converter-shift register was designed to use in as many places as possible functional packages already developed or in process of development. In particular, packages used for the main functions shown in Figure 3-3 are: M1736-1 Bit Register, M1731-1 Regenerative Gate, M1733-1 Regenerative Amplifier. Development work consisted largely of the design of appropriate connecting circuits. One passive connecting circuit function shown as "Pulse Steering" was repeated so often in the system that it was also packaged and coded M1751-3. A simplified schematic of the converter-shift register is shown in Figure 3-12 for reference only. Detailed characteristics of this circuit are given in another report by R. E. Yaeger.*

4.0 Summary of Package Phase

In the introduction it was stated that feasibility of the 15 digit encoder would be demonstrated by the

*"A Gray to Binary Translator and Shift Register", by R. E. Yaeger.

analog-to-digital encoding of the 6 least significant digits of a 15 digit "Gray" code wheel, storing of the 6 parallel code digits in a shift register, conversion of the Gray code digits to binary form, and finally, the shifting of the resulting binary code into a dummy line in time sequence. During the course of investigation of the encoder-converter-register principles, first order feasibility of this 15 digit system seemed assured on the basis of extensive tests on breadboarded circuits working separately and together in such a system. However, final feasibility proof was to consist of demonstration of packaged functional packages working separately and together. Functional packages comprising this system have already been described in detail, the specifications and operating characteristics of the packages have been set, and the packages have been coded. The packaged Encoder E model is shown in Figure 3-11. From left to right are the light source, code plate assembly, defining slit, a 6 digit p-n photocell package coded M1749-1, an assembly of p-n bias resistors, and 6 regenerative pulse amplifiers coded M1733-2, respectively. The packaged converter-shift register model is shown in Figure 3-13. The package-to-circuit relation may be seen by referring the M numbers on the packages back to the simplified schematics of Figure 3-10 and 3-12.

Figure 3-14 shows the encoder model and the converter-shift register model assembled as a working angular position transmitter. The resolution of 64 quantum steps of .001" each has been demonstrated by an oscilloscopic presentation of the time sequential binary code representing each individual quantum step. The CONVERT pulse was removed at times to allow the Gray code to be shifted out in time sequence. This allowed a comparison of the Gray and binary equivalence for any of the 64 quantum steps.

On the basis of existing specifications for and operating characteristics of the packaged model it is assumed that first order feasibility of 15 digit angular position encoding using functional packaged circuits has been proven.

Appendix I

Reference List for Encoder E Characteristics

1A. System (General)

Input READ Pulse

Amplitude ... +85 v. min.

Duration .... 10μ sec. min.

Rise time ... 1μ sec. max.

Output Pulses (Gray or reflected binary)

Amplitude ... 20 v. min.

Duration .... 2μ sec. min.

Rise time.... 0.2μ sec. max.

Code Resolution Tolerances (Figure 3-2)

"1" Total transmission distance .... nominal ± .0005 inch

"0" Total no transmission distance .... nominal ± .0005 inch

Power Drain (typical values)

Supply Voltage (Volts)	Current (ma)	Power (Watts)
-45	5.0	.225
-90	.06	.005
-135	1.8	.24
+600	15.0	9.0
	Total Power	9.47

2A. Light Source

Block diagram ..... Figure 3-1

Circuit schematic ..... Figure 3-10

Photograph ..... Figure 3-6, 3-11, 3-14

Circuit components list ..... Appendix II

Power Supply Spec.

No load ..... 650 v. max.

Full load ..... 600 v. max., 20 ma. min.

SA309, Electrical Ratings

Anode Voltage

425 v. min., 600 v. max.

Trigger Voltage

4KV min., 12 KV max.

Max. dissipation ..... 3.5 watts*

Repetition frequency .... 100 pps max.

Input READ Pulse

Amplitude ..... + 85 v. min.

Duration ..... 10 μ sec. min.

Rise time ..... 1μ sec. max.

Dynamic Characteristics

Repetition frequency ..... 10 pps

Output light flash

Rise time (10% to 90% peak) 5μ sec.

- -

*Tube dissipation in watts equals $\frac{fC(V)^2}{2}$ where

C is discharge capacity in μμf
V is voltage in kilovolts
f is number of flashes per second

Duration ..... 40 μ sec.

SA309 dissipation ..... 3.5 watts

Total power drain ..... 9 watts

3A. Code Wheel Simulator and Optical System

Block diagram ..... Figure 3-1, 3-10

Photograph ..... Figure 3-6, 3-11, 3-14

Parts list ..... Appendix II

Code wheel Specifications

(a) "Angular Position Encoder," Signal Corps Tentative Specification, No SCL-1210, January 10, 1949.

4A. P-N Junction Photocell Assembly (M1749-1)

Block diagram ..... Figure 3-1, 3-10

Photograph ..... Figure 3-7

Power Supply Spec. ..... -90 v ±10%

5A. Regenerative Pulse Amplifier (M1733-2)

Block diagram ..... Figure 3-1, 3-10

Circuit schematic ..... Figure 3-9, 3-10

Photograph ..... Figure 3-11, 3-14

Circuit components list ... Appendix II

Power supply spec.

-45 v. ± 10%

-135 v. ± 5%

Operating characteristics ... Figure 3-9

6A. Connecting Circuits for Encoder E

Parts list ..... Appendix II

APPENDIX II

Partial Stock List for Encoder E (Figure 3-10)

2.0 Light Source

Quan.	Part No.	Description
1	R_1	240K ±5%, 1/2 w
1	R_2	470K ±5%, 1/2 w
1	R_3	10 meg. ±5%, 1/2 w
1	R_4	120K ±5%, 1/2 w
1	R_5	20 meg ±5%, 1/2 w
1	R_6	10K ±5%, 10 w
1	R_7	33K ±5%, 2 w
1	C_1	.01μ f, 150 v
1	C_2	.5 μf, 1000 v
1	C_3	2 μf, 1000 v
1	T_1	PF-3, UTC Pulse Transformer 15 KV peak
1	V_1	0A5
1	V_2	SA309

3.0 Code Wheel Simulator and Optical System

Quan.	Part No.	Description
1	B3S-750820	Code Wheel Simulator Ass'y.
1		Code Plate, Figure 5
1		Defining Aperture, .0005" (Lyman Nichols, Nutley, New Jersey

4.0 P-N Photocell package

Quan.	Description
1	M1749-1, 6 Digit P-N package

5.0 Regenerative Pulse Amplifier

Quan.	Description
6	M1733-2 Regenerative Pulse Amplifier

6.0 Connecting Circuits for Encoder E

Quan.	Part No.	Description
12	C5-C16	.01 µf, 150 v.
6	R8-R13	330K ±10%, 1/2 w

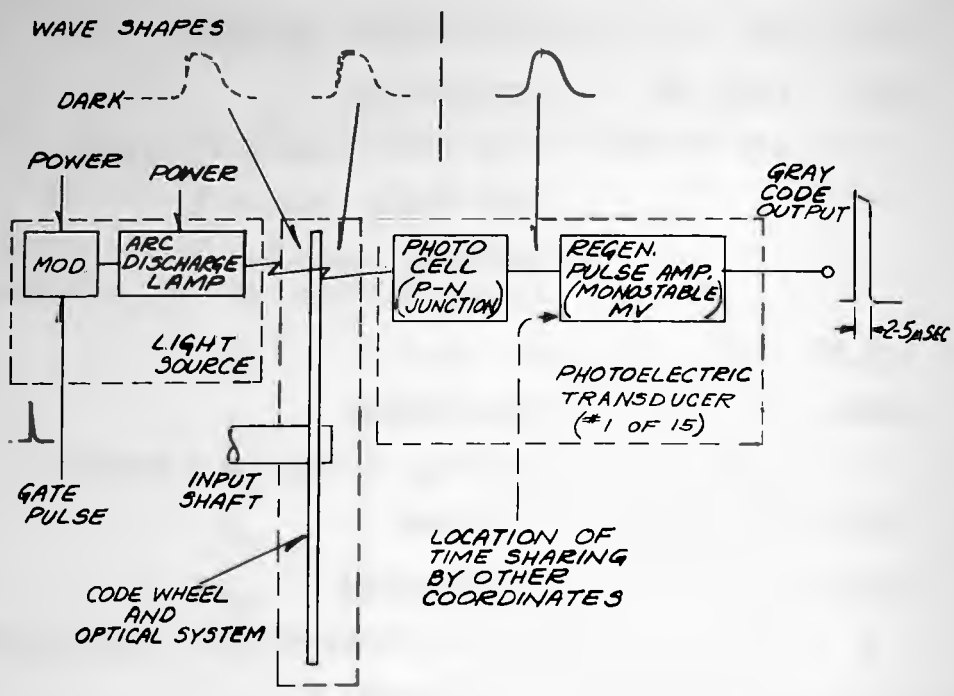


Fig. 3-1. Encoder E. Block Diagram

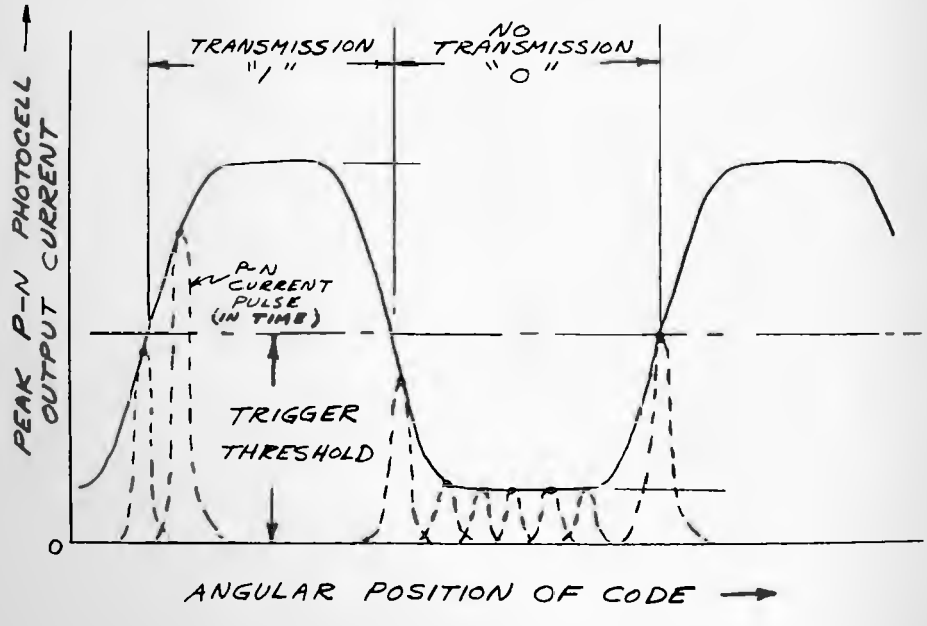


Fig. 3-2. Code Resolution Envelope

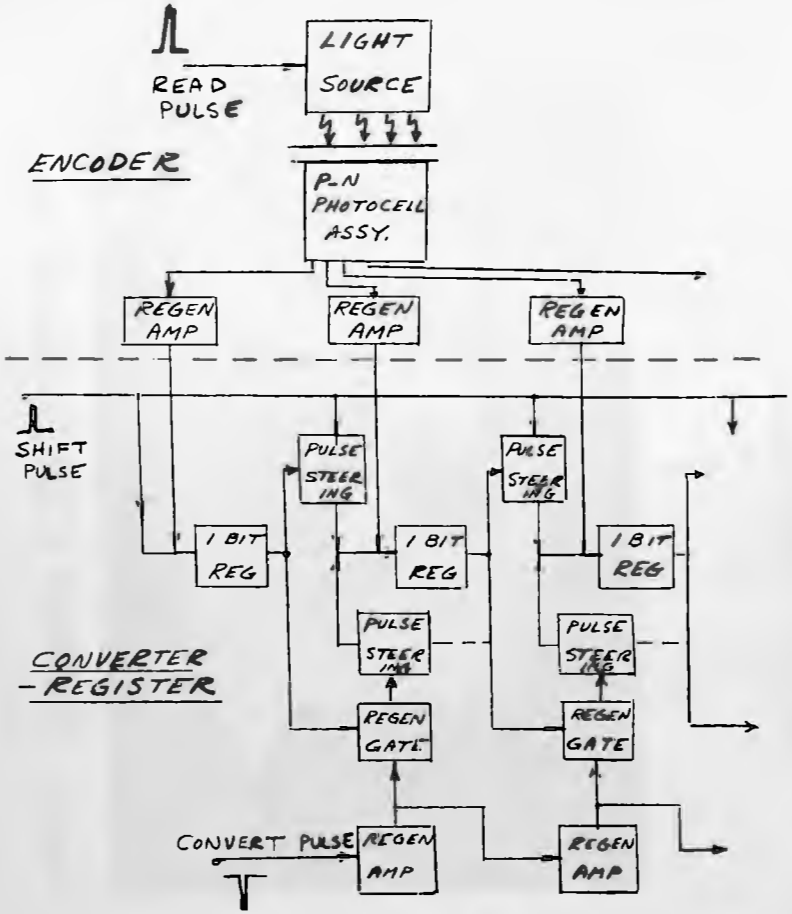


Fig. 3-3. Encoder-Converter Register Flow Chart

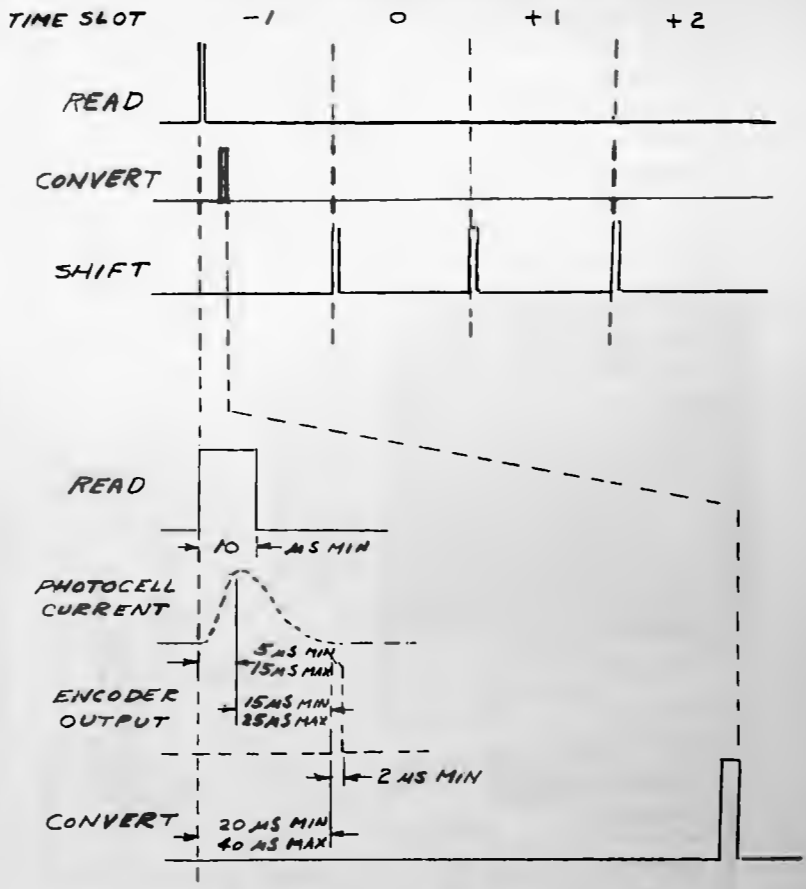


Fig. 3-4. Program Pulse Sequence

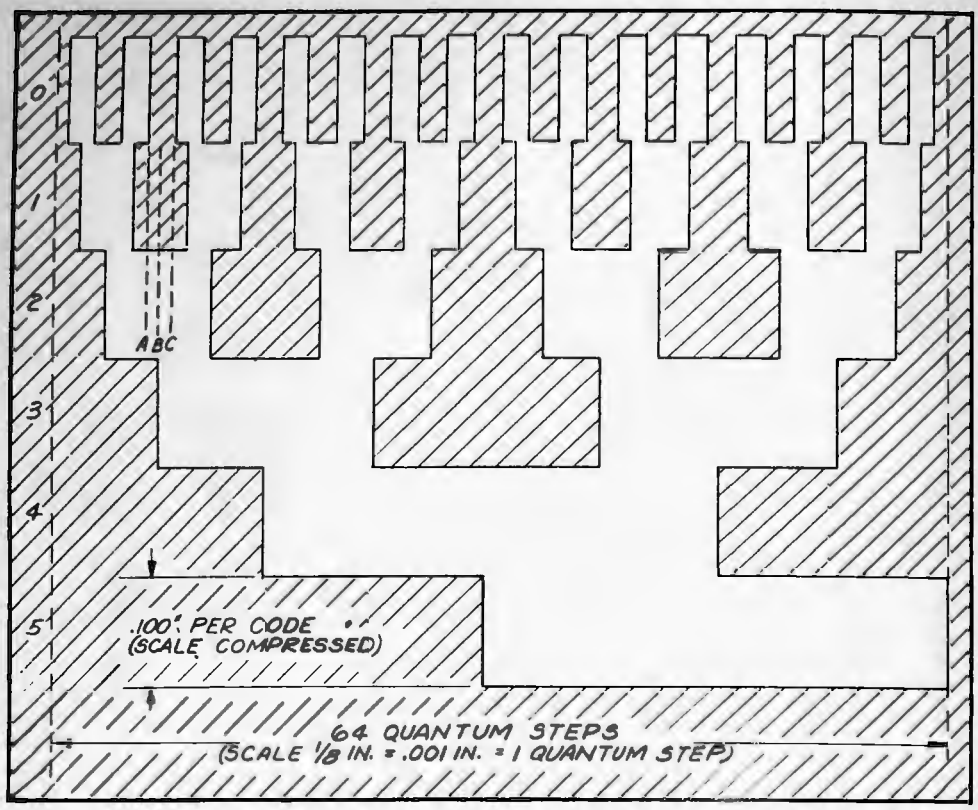


Fig. 3-5. 6 Digit Gray Code Plate

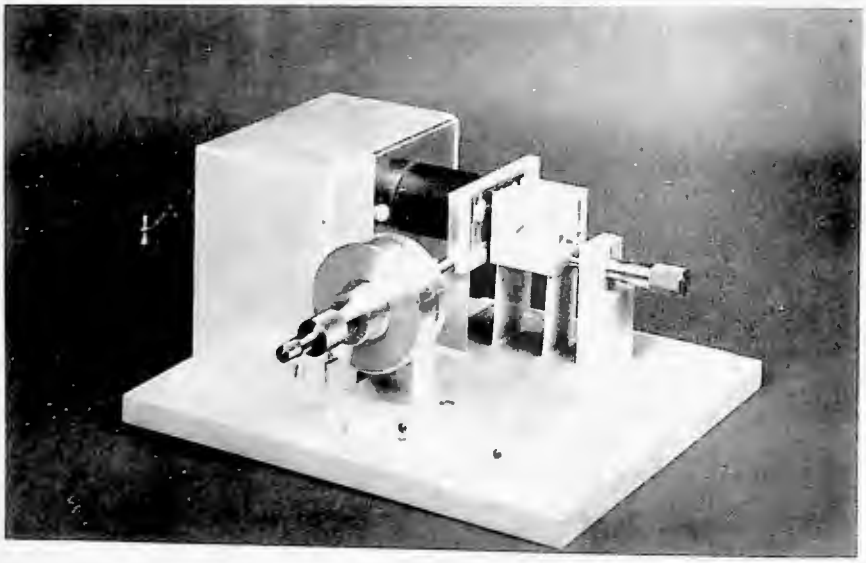


Fig. 3-6. Light Source and Optical System for Encoder E

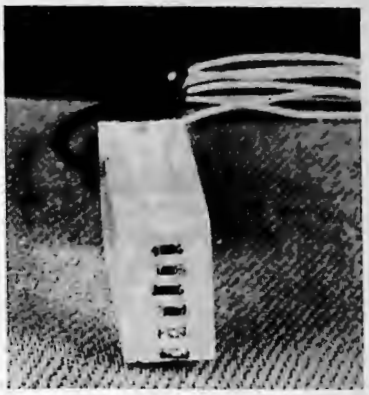
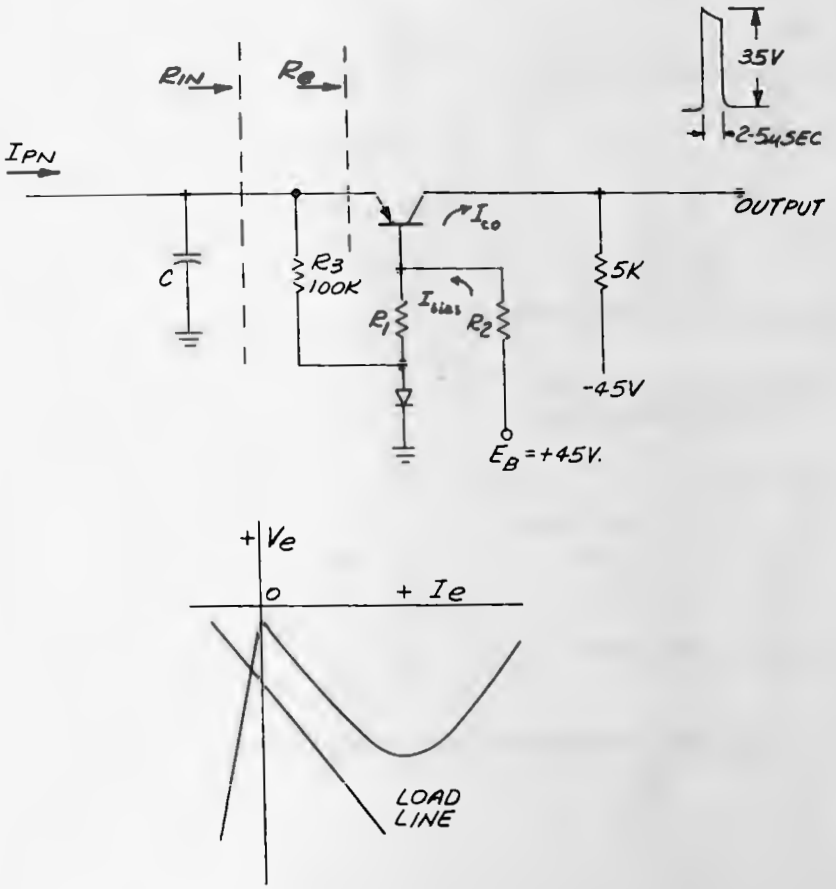
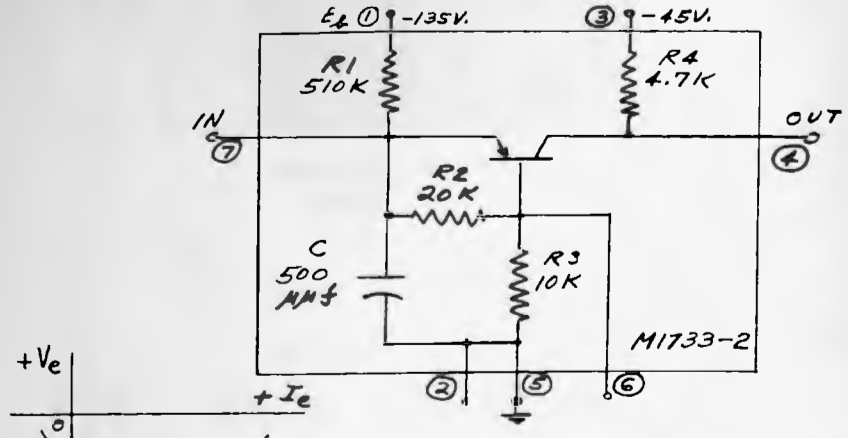


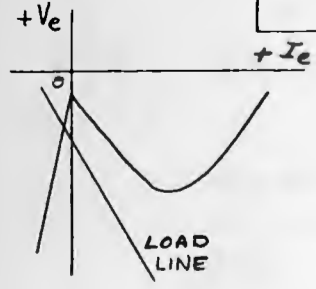
Fig. 3-7. 6 Digit p-n
Photocell Package



EMITTER CHARACTERISTIC

Fig. 3-8. Monostable Multivibrator





EMITTER CHARACTERISTIC

AVERAGE QUIESCENT CHARACTERISTICS

I_{-45}	0.9 MA.
I_{-135}	0.24 MA.
V_7	−15.2 V.
V_6	−12.0 V.
V_4	−43.3 V.

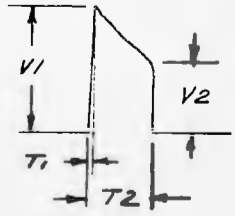
AVERAGE DYNAMIC CHARACTERISTICS

INPUT THRESHOLD

CONSTANT CURRENT SOURCE @ 10 P. P. S

I_T = 175-275 μa

OUTPUT



V_1	42 V.
V_2	20 V.
T_1	2.1 μsec.
T_2	4 μsec.

Fig. 3-9. Regenerative Pulse Amplifier M 1733-2

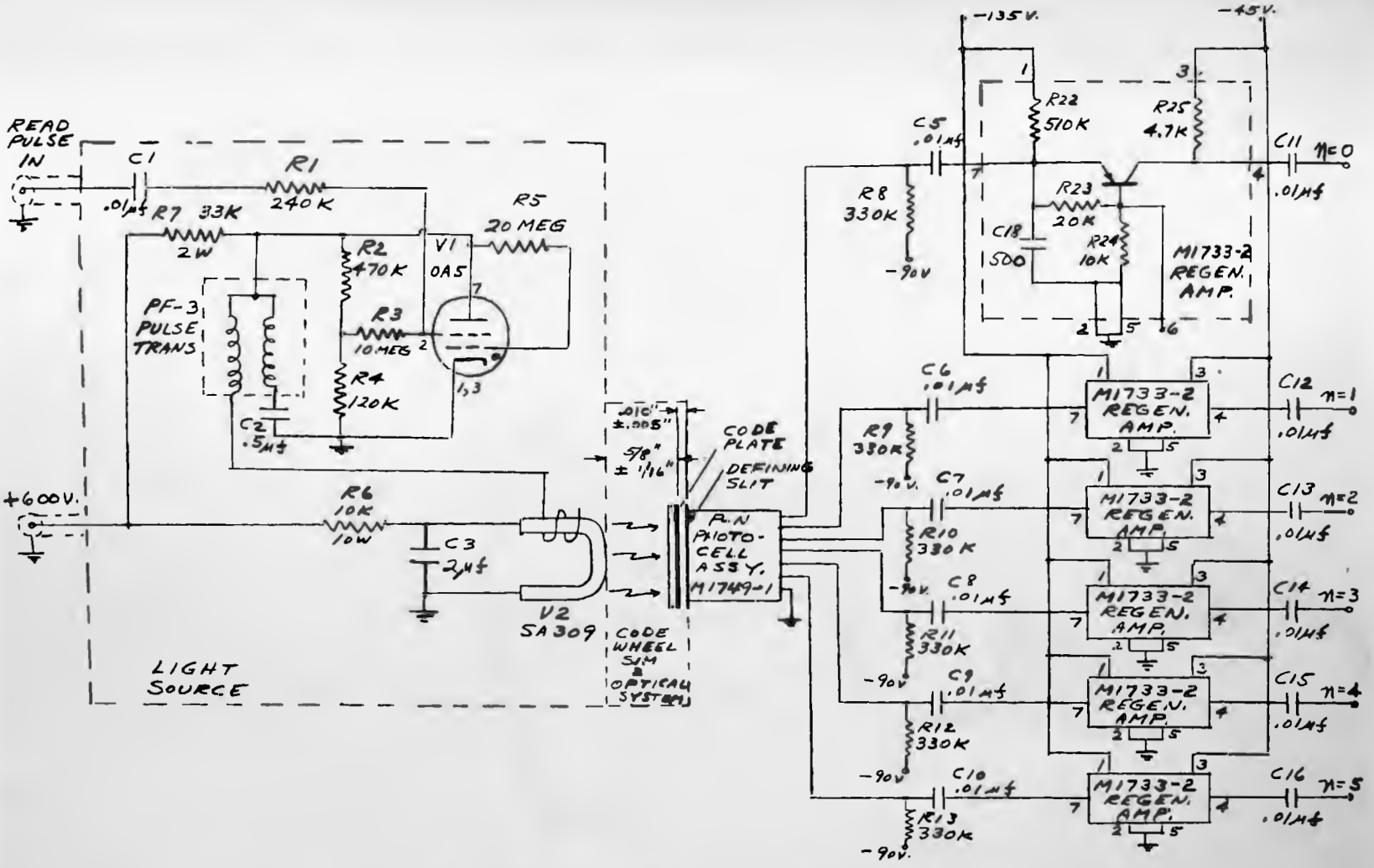


Fig. 3-10. Encoder E. Simplified Schematic

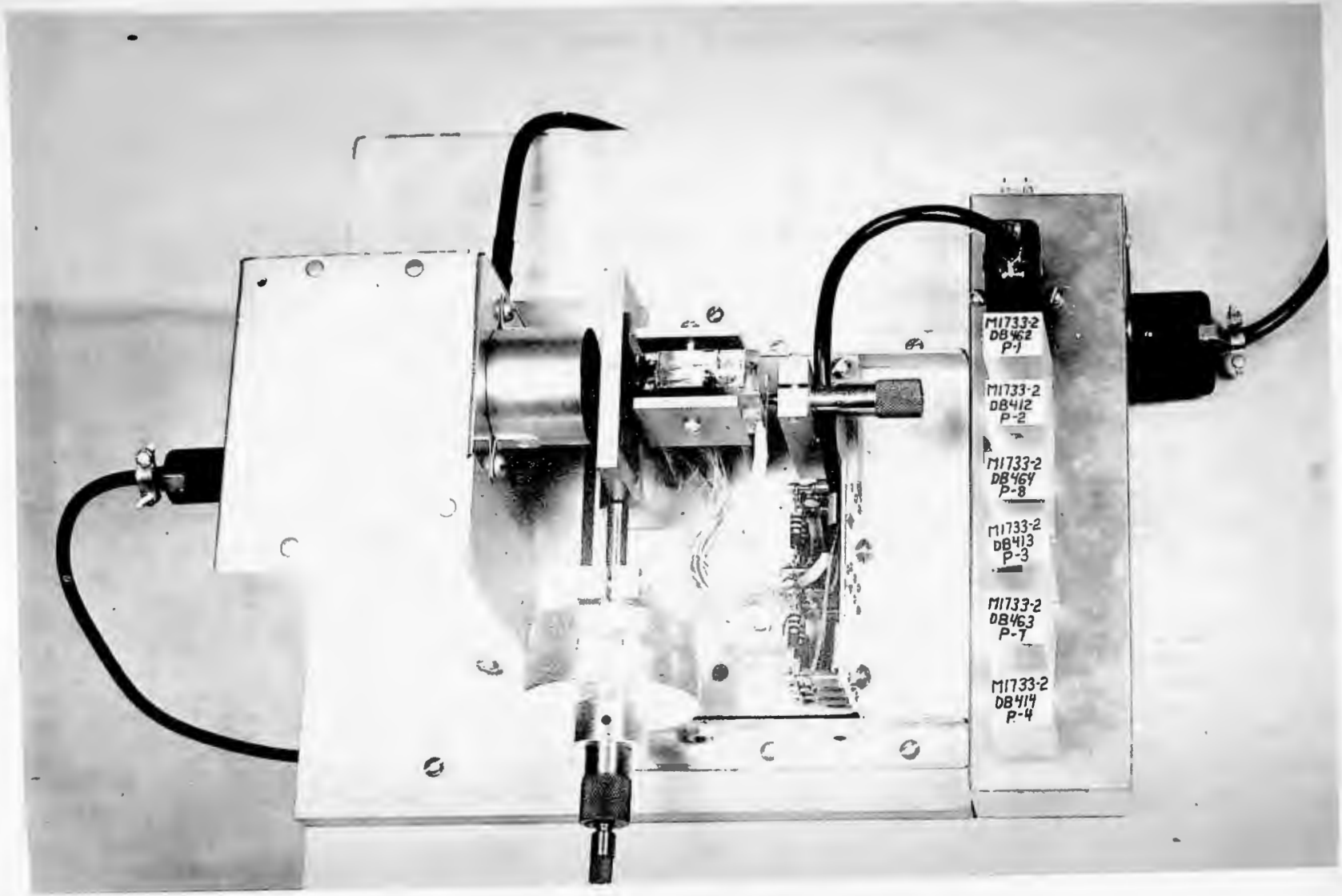


FIG. 3-11 PACKAGED ENCODER E MODEL

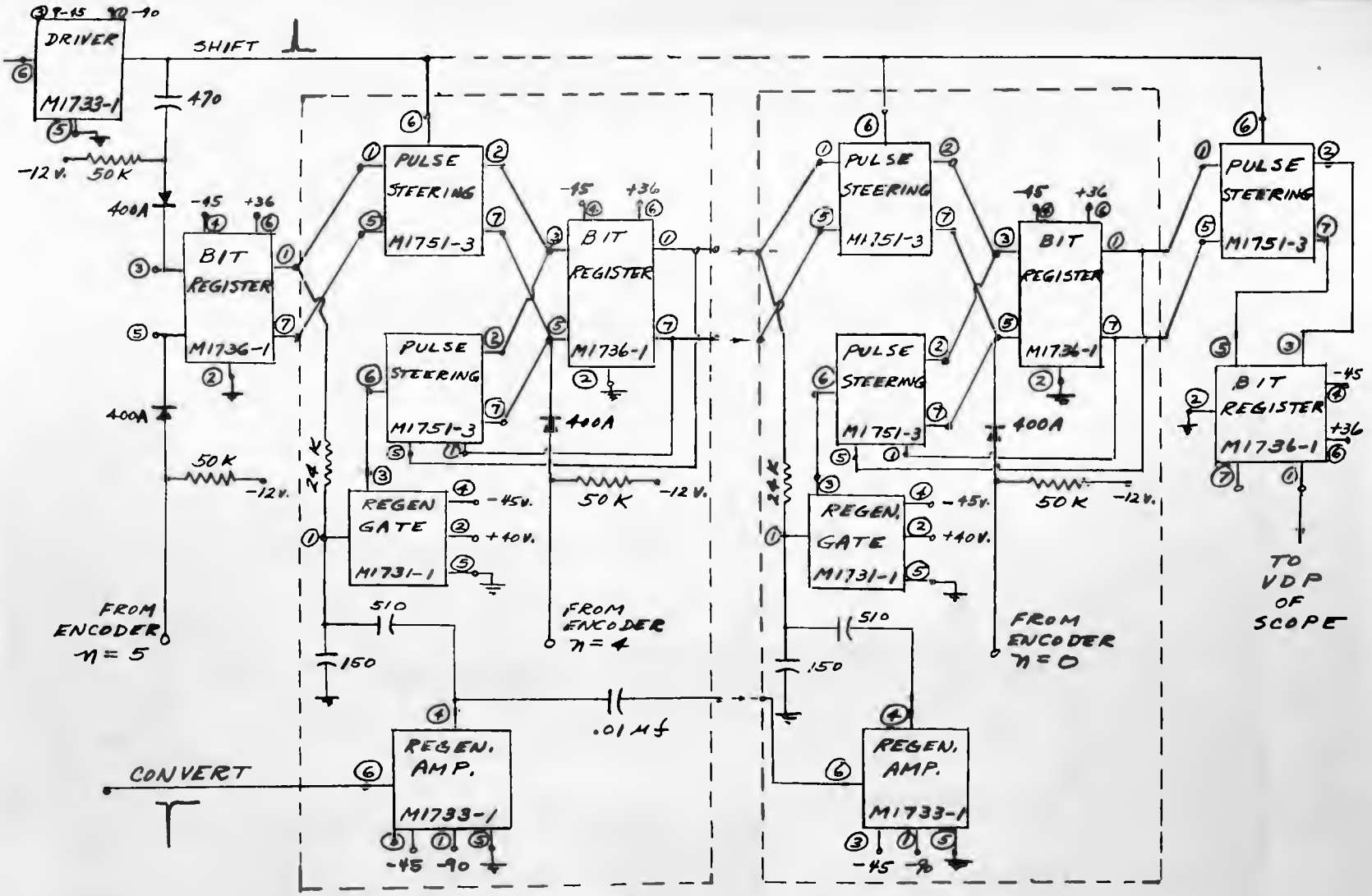


Fig. 3-12. Converter-Shift Register. Simplified Schematic

FIG. 3-13. PACKAGED CONVERTER-SHIFT REGISTER MODEL

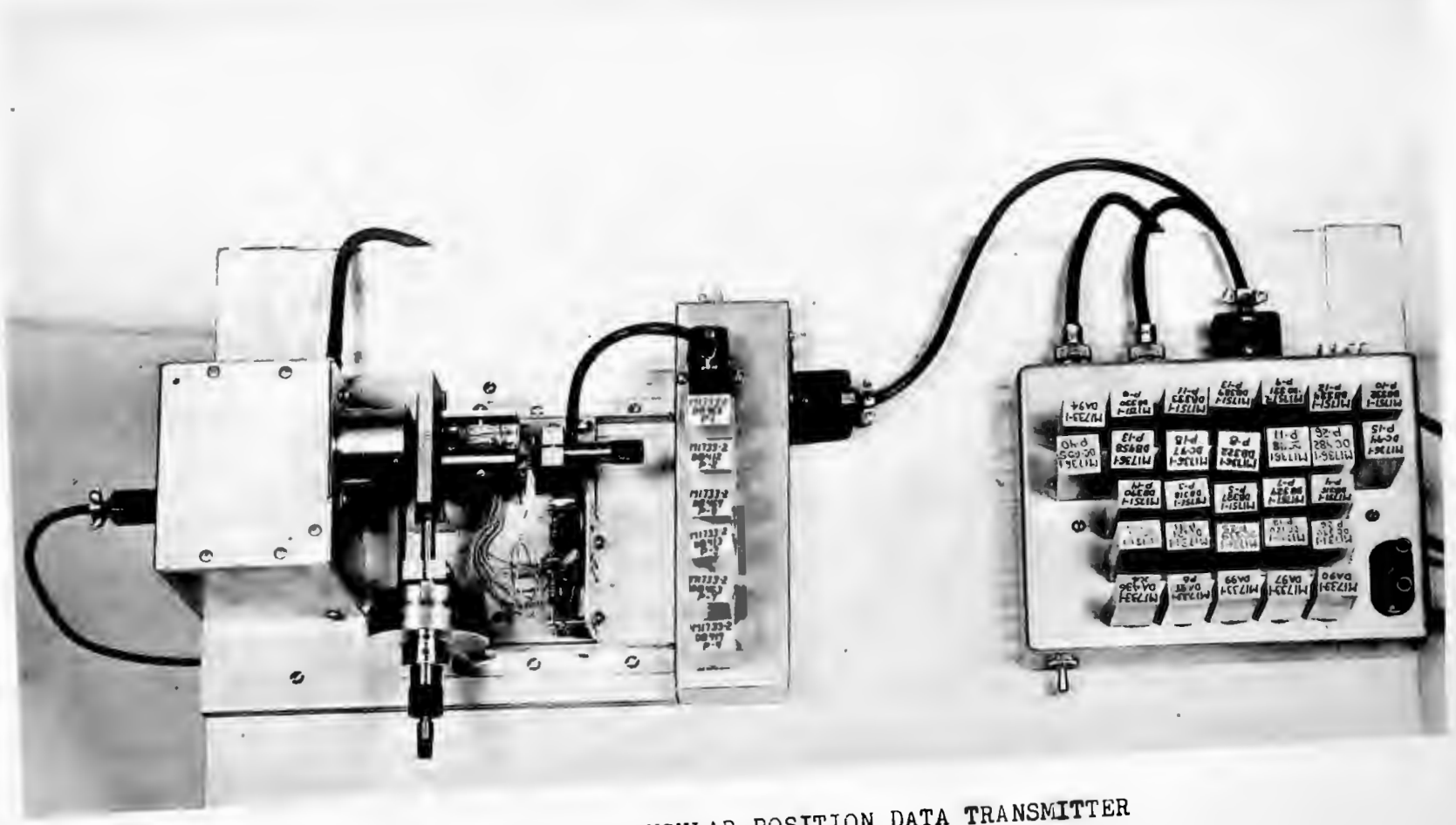


FIG. 3-14. PACKAGED ANGULAR POSITION DATA TRANSMITTER

A GRAY TO BINARY TRANSLATOR AND SHIFT REGISTER

R. E. YAEGER

Produced under contract W 36-039 SC-44497

A GRAY TO BINARY TRANSLATOR AND SHIFT REGISTER

The purpose of this report is to describe the theory and operation of a combined Gray (reflected binary) to binary code translator and shift register. The Gray or "Reflected Binary" code has the unique advantage of presenting only a single digital change at the transistion of one quantum step to the next. Translation of the Gray code output to binary code is required as computing and decoding in Gray code are not possible by the usual binary circuit configurations. Gray code computing circuits can be devised, but each of such circuits would require some system of operation on the Gray code similar to the translation process.

Translation Process

Although it is possible to describe the arithmetical gymnastics which show the required translator process, it is simpler to describe here the basic process by the comparison of the representative groups of Gray and binary numbers shown in Figure 1. Only 4 digits are shown in the Figure; the general rules apply, however, to any number of digits. In examining the two different codes it is readily seen that the most significant digits of the codes are slways the same in equivalent numbers. In the example shown in Figure 1 the first translation step shows the most significant digit (#1) becoming binary without change. The next or 2nd digit of the binary number depends, however, on the comparison of the 2nd Gray digit and the 1st binary digit. If in this comparison the digits are alike (0,0, or 1, 1) the resulting binary digit is a 0; if they are different (0,1 or 1,0) the resulting 2nd binary digit is 1. In turn, the 3rd binary digit depends on the same comparison of this resulting 2nd binary digit with the 3rd Gray digit. When these comparisons have been completed down to the least significant digit, the result is the equivalent binary number.

A basic functional configuration for translation of parallel Gray code is shown in Figure 2. Here the NOT AND gates provide the series of comparisons described above. The NOT AND gate is a device such that if either one, and only one, of the 2 inputs is stimulated with 1, the output will be a 1. If both or neither are stimulated, the output is a 0. The

connections to these gates are such that the nth binary digit is the result of the NOT AND comparison of the nth Gray digit and the (n+1) th digit resulting from a similar previous comparison. It is apparent that the binary input to the NOT AND gate is delayed with respect to the Gray input from the encoder by the time required for the previous comparisons of the more significant digits. The NOT AND gate must therefore be capable of comparing two input pulses that are not necessarily coincident.

Figures 3 and 4 show how these two methods of translation may be used in an Optical Data Transmitter System. In the serial method of translation of Figure 4 the two directional shift register is required to reverse the transmitter sequence to least significant digit first, this being necessary for the serial addition of aprallax correction. The requirement of this additional register appears to outweigh the advantages of using the simpler serial translator. Methods for translations of the parallel code were therefore primarily considered.

NOT AND Gate

In a previous report* there is discussed the problems of NOT AND gates for input signals which are not exactly coincident. In his report several types of transistor and diode NOT AND gates were shown which are applicable to a parallel code translator. Investigations of the shift register circuits, however, showed the possibility of combining the functions of translator and shift register, reducing the total number of components required for these two functions. This resulted in a program controlled NOT AND gate configuration shown in Figure 5. The essential parts of this gate are the "one bit" register of the shift register** and a supplementary threshold 2 AND gate. As indicated in this figure, steering control has been added to register #2. The function of the steering control is to control input pulses in such a

- -

*Anticoincidence or 'Not And' Gate", by L. W. Hussey.

**"Storage and Addition of Binary Numbers", by J. R. Harris.

way that they always change the state of the register. That is, if the register is in a state of 1, a pulse on the input will change it to a 0; if the state is a 0, a pulse will change it to a 1. This action is identical to the count action of a binary counter stage.

To examine the operation of this gate, assume that both registers have been cleared by a previous operation and that a new pair of digits is read into the register states by setting those stages to one where the digit is a "1" and leaving the other stages in a zero condition. A "compare" pulse is then applied to one of the inputs of the AND gate. This gate is connected to register #1 in such a way as to transmit the compare pulse only if register #1 is in a 1 condition. If the gate does transmit a pulse, it will then change the stage of register #2. The final state of register #2 is the NOT AND comparison of the original numbers. This can readily be checked by trying the four possible combinations: 0,0 and 1,1 which result in a 0; and 1,0 and 0,1 which result in a 1. A desirable feature of this gate is that the input signals to be compared do not have to be coincident. The only timing requirement is that the registers be set before the "compare" pulse is applied.

Figure 6 shows the configuration for a parallel Gray to binary translator incorporating this NOT AND Gate. Each register stage is also a part of a complete shift register, the connecting circuits of which are not shown for reasons of simplicity. To examine the translation operation assume that the shift register has been cleared initially so that all "One bit" registers are in a state of 0. Under the control of the system program a Gray code number is "read" in the encoder and the digits, in short duration pulse form, are applied to the inputs of register. When the condition of the register has stabilized, storing this parallel Gray code number, a "translate" program pulse is applied at AND gate #1. The NOT AND comparison described above is then made for the most significant digit in register #1 and the next digit in register #2. The resulting condition of register #2 is the binary equivalent of the original Gray digit. After a time delay of sufficient length to insure translation and stabilization in register #2 the translate pulse is applied to gate #2, comparing the state of registers #2 and #3. After this comparison, the digit in register #3 is in binary form. Again, after an appropriate

delay the translate pulse is applied to gate #3 with the result that all registers are then storing the binary equivalent of the original Gray number.

Circuits

Figure 7 shows in schematic form the details of the translator described above. For simplicity, only 3 storage registers are shown and the interstages for digit shifting are omitted. The circuit for the storage register has been previously described *. The method for steering control for this register is shown in Figure 4.1-9. In this figure assume transistor #1 of the register is in the "on" condition, in which case its base and collector are at approximately -10 and -12 volts respectively. The difference between these voltages, 2 volts, is the effective back voltage on diode #1 of the steering circuit. At the same time the base and collector of the "off" transistor are -4 and -35 volts respectively, biasing diode #2 approximately 31 volts in the back direction. It is apparent that under these conditions a positive input pulse on the control circuit, having an amplitude between 20 and 35 volts, would be readily steered through diode #1 to the base of transistor #1, turning this transistor off and transistor #2 on. The storage effect of C_1 and C_2 tends to keep the input pulse correctly steered until the register has completely changed state. After several time constants of R_1C_1 (or R_2C_2) diode #2 becomes the readily conducting diode to an input pulse, enabling the next input pulse to turn transistor #2 "off" and #1 "on". The six passive components for the steering control have been packaged into a single functional package, M1751-3, the specifications of which are given in Fig. 13. To provide sufficient triggering level for the register stage a transistor regenerative gate is used for the AND gating function.*

The delays for the translate pulses are obtained from the cascaded stages of regenerative pulse amplifiers in the following manner: at the time the negative program translate pulse is applied, the first regnerative amplifier (amp. #1) is triggered, providing a positive output pulse of about 60 μsec. duration. The positive leading edge of this pulse is applied to gate #1, while 60 μsec. later the negative trailing edge triggers amp. #2. The positive leading edge of the output pulse of amp. #2 is applied to gate #2; while an additional 60 μsec. later amp. #3 is triggered by the trailing edge. This action continues along the entire chain of amplifiers, providing in each step a delay of approximately 60μ sec.

- -

*"Storage and Addition of Binary Numbers", by J. R. Harris.

As seen in the diagram of Figure 7, the output of the regenerative amplifier is differentiated and attenuated from approximately 25 volts to 20 volts by the condenser divider network, C_1 and C_2, providing the proper critical triggering level for the regenerative gate. When the regenerative gate is triggered, the emitter current maintaining the transition of the gate in the "ON" condition is supplied by the charge stored in C_1 and C_2. C_1 and C_2 in parallel, therefore, determine the charge available in the output pulse of the gate. In determining the value for the parallel combination of C_1 and C_2 two things must be considered. If the register controlling the regenerative gate has just changed state due to the translation of the digit it was storing, the change must be applied to the gate before translation of the next digit by charging C_1 and C_2 in parallel through R_1. The higher the value of C_1, plus C_2, the greater the time required per digit for translation. Yet C_1 and C_2 must be sufficiently large for the gate to supply an adequate signal charge for setting its associated register stage. As high speed translation was not required in this system it was possible to make C_1 plus C_2 large enough to meet this requirement. Where higher speed translation may be required a non-storage type gate might be used in conjunction with regenerated gain.

The dotted enclosures on Figure 7 indicate the functional packages used in the translator. References to their specification are given in Table 1.

Digit Shifting Register

In a report previously referred to* there is described various configurations for digit shifting registers. One of these types is shown functionally in figure 9. In this diagram, it is seen that the program shift input is applied simultaneously to all the interstage threshold 2 AND gates and the zero set input of register #1. The first shift pulse, therefore, sets register #1 to zero and enables one of the two AND gates in all the interstages. That AND gate which is enabled is determined by the original state of the register just ahead of the interstage. The delay provides the memory of this state even though the controlling register may already be in a changing condition. Therefore, if register #1 was in a state of 1 just before the first shift pulse, the gate on the set 1 lead of register #2 is enabled, setting the register to 1 if it was a zero or leaving it in a state of 1 if it was a 1. Likewise, the stage of register #2 is transferred to register #3 and the state of register #3 to a following stage at the time of the first shift pulse. In effect, the original number stored in the register is shifted one digit to the right. Successive shift pulses perform the same operation, each pulse shifting the stored number one digit to the right until all

- -

*"Storage and Addition of Binary Numbers", by J. R. Harris.

digits of the number have been shifted out of the register. As the shifting process for this type of register is entirely controlled by program pulses, the interstages cause no interference with the translation process. In the actual shifting circuit used for the combined translator-register circuit, the first two stages of which are shown on Figure 10, the delay and gating functions are accomplished with a simple R, C, and diode network, identical with the steering control network for the translator. The connections to the registers differ, however, in that the steering of the input pulse is controlled, not by the same register condition, but by the adjacent register. Successive shift pulses, therefore, transfer the states of one register to the adjacent register, providing the digit shifting process described above.

The loss through the interstage is small enough that the registers may be set directly from the shift driver output without the additional interstage gain indicated in the functional diagram, Figure 9. Seven register stages have been operated from one M1733-1 driver satisfactorily. For a full 15 digit system, it is expected that the register load would be divided between 3 such driver stages.

Packaged System

The translator-shift register was assembled using the functional packages developed under the packaging program. The complete interconnecting schematic for this system as used for demonstrating the operation of the Optical Encoder is shown in Figure 11, together with component values and tolerances not covered by the package specifications. In this demonstration model for the six digit optical encoder a 7th bit-storage register (Register A) was used, corresponding to the modulator gate for an actual data transmitter system. The oscilloscope indication of the output of this register represented the envelope coded signal output of a data transmitter system.

The control program sequence for the demonstration model is shown in Figure 12.

Here it is seen that the reading of the code in the optical encoder and the storing and translation of the code in the register are performed during the 1667μ sec. time slot before the first digit is to be transmitted out of the system.

Overall system specifications for the translator-register and the individual package specification references are given in Table I of this report.

Conclusion

For an optical encoder with requirements based on a typical data transmission system the presented configuration of circuits for code translation and digit shifting appears to be the simplest from the standpoint of the number of active components required. The model constructed for use with the 6 digit encoder has operated satisfactorily for over two months and has been demonstrated many times during that period. Although rigorous life tests have not been made on the overall system, such tests are being made on the individual components package. From the statistical results of these tests, the performance of the overall system can be resolved.

The feasibility of encoding by optical means was proved with the demonstration of the 6 digit optical encoder. Its demonstration with this translator-shift register proves the practicality of optical encoding in a data transmission system.

TABLE I

Requirements for 6-digit Translator-Shift Register System

Power Supply	Voltage	Tolerance	Current	Power
	+40	± 5 v	80 ma	3.2 watts
	-45	± 5 v	80 ma	3.6
	-10	± 1 v	1 ma	.01
	-150	± 25 v	2 ma	.3
			Total	7.2 watts

Program Input	Pulse Amplitude	Pulse Length	Input Impedance
Translate	-3 v min.	1 u sec. min.	5000 ohms
Shift	-3 v min.	1 u sec. min.	1000 ohms
Encoder			
Set "1"	+10 v min.	5 u sec. min.	1000 ohms

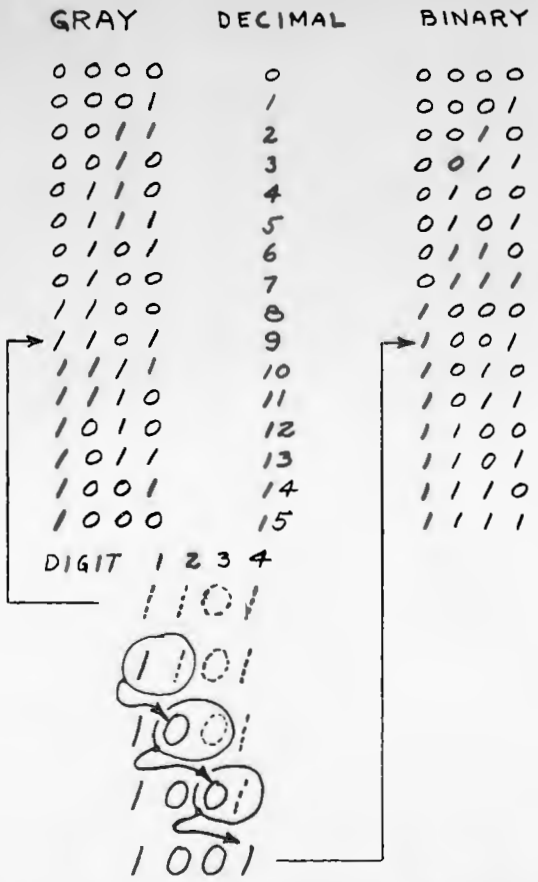


Fig. 1 Comparison of Gray and Binary Codes

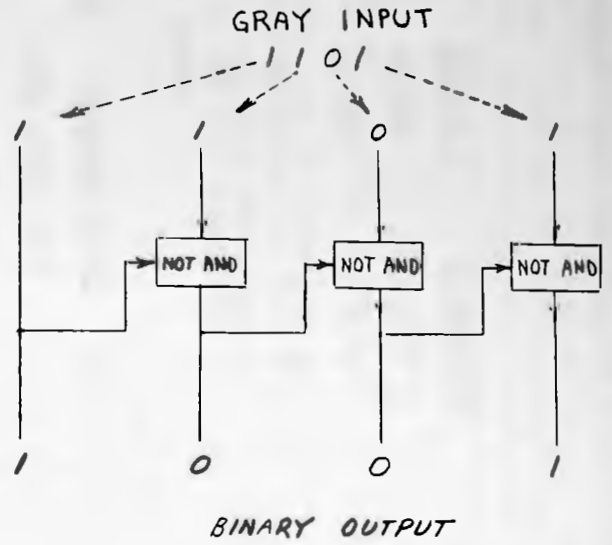


Fig. 2 Parallel Translation

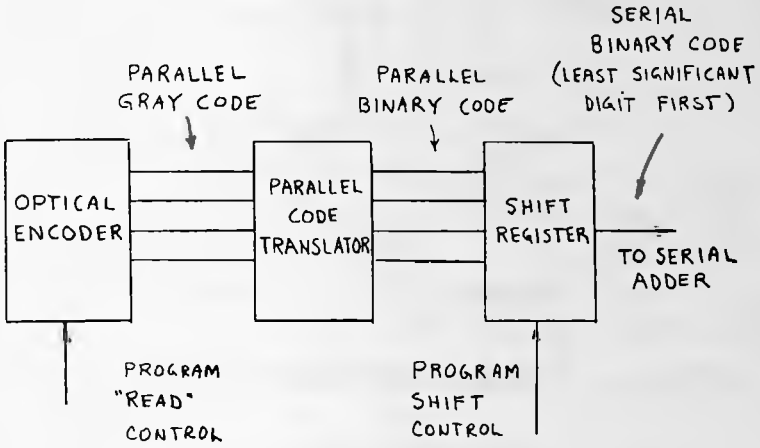


Fig. 3 Optical Data System Using Parallel Translation

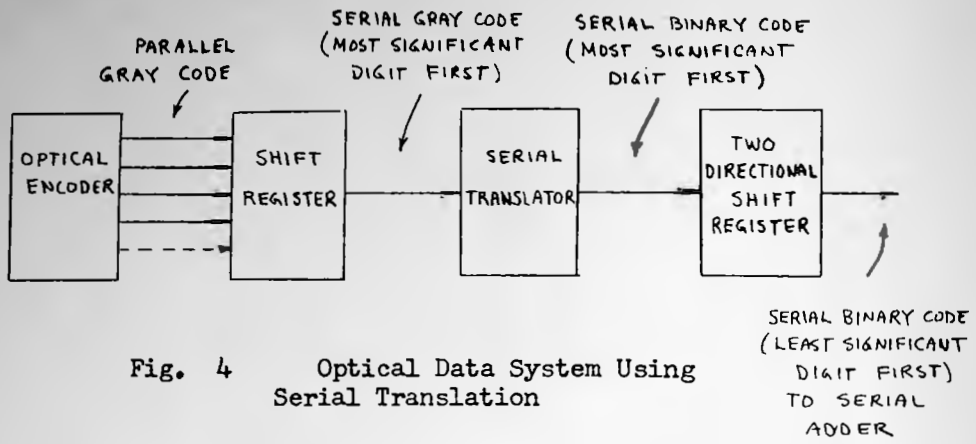


Fig. 4 Optical Data System Using Serial Translation

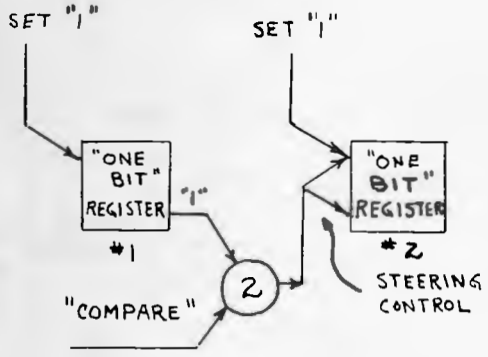


Fig. 5 Program Controlled "NOT AND" Gate

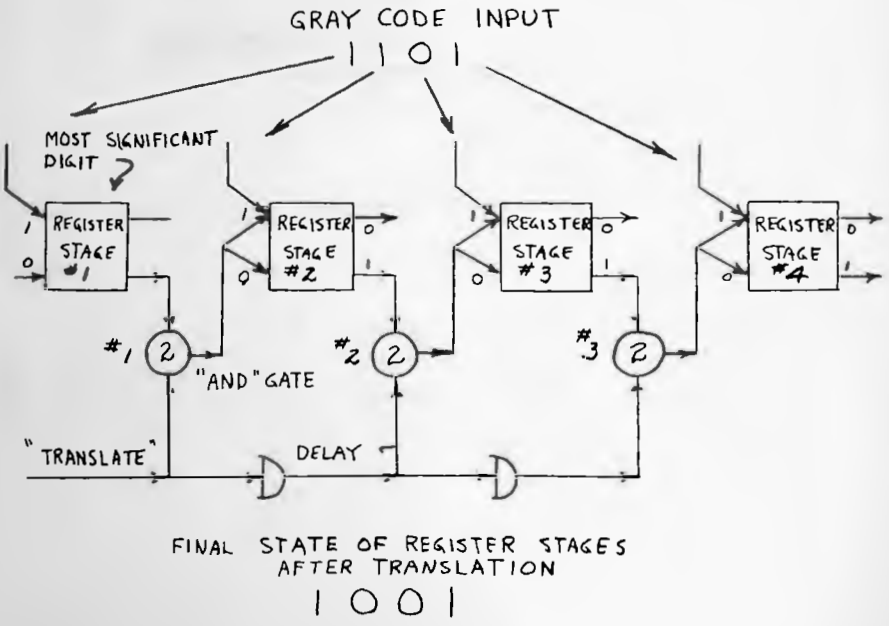


Fig. 6 Functional Schematic of Gray to Binary Translator

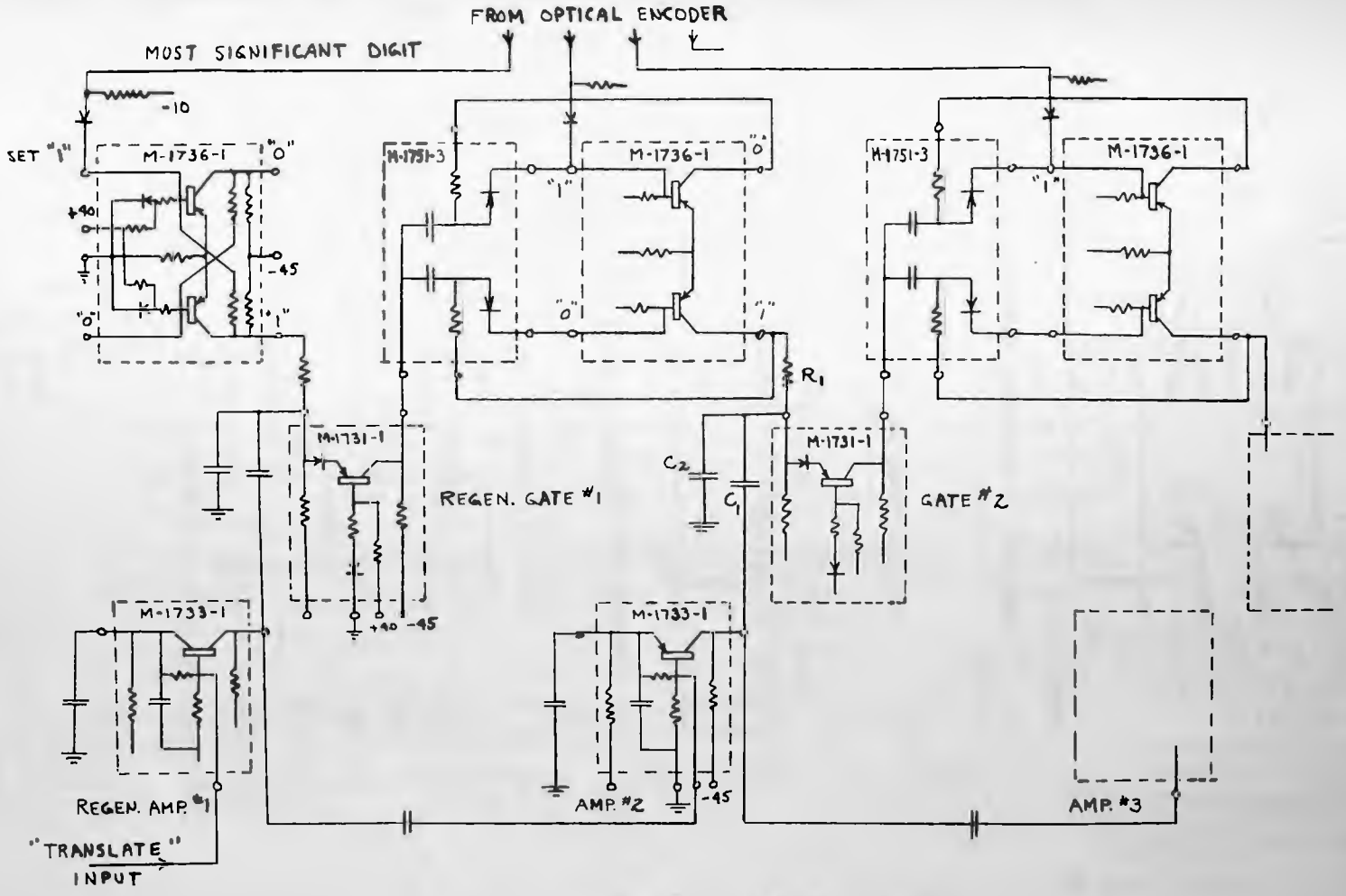


Fig. 7 Partial Translator Schematic

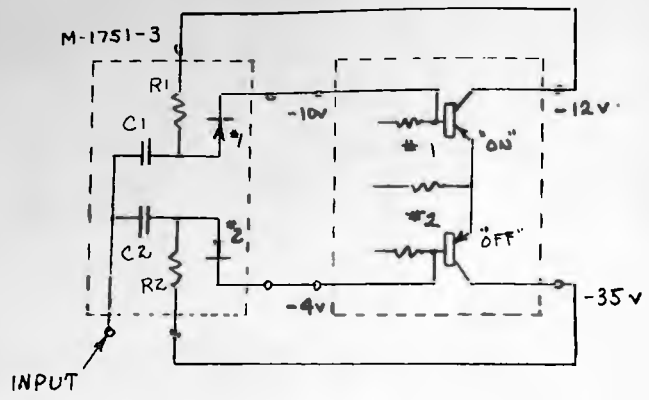


Fig. 8 Steering Control

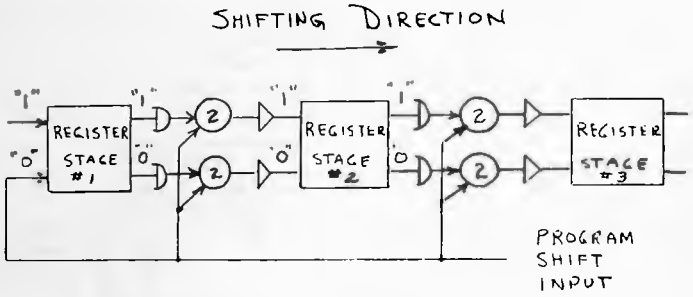


Fig. 9 Functional Diagram of Digit Shifting Register

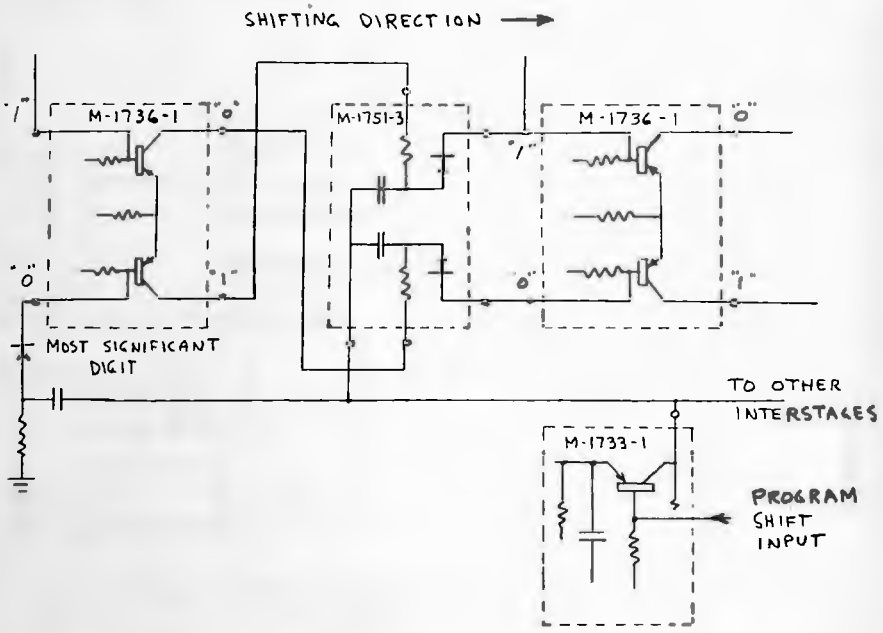


Fig. 10 Digit Shifting Register Schematic

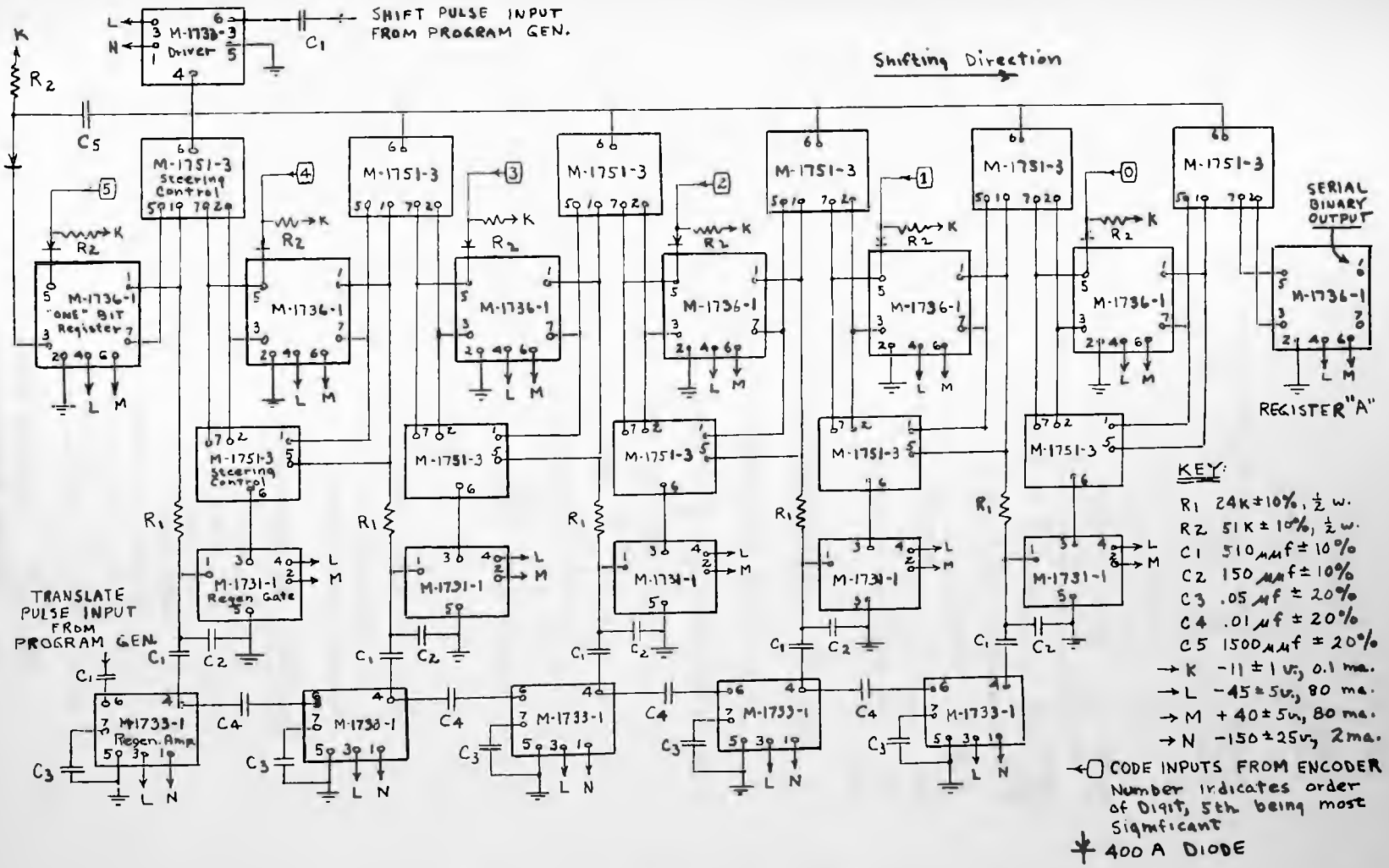


Fig. 11 Translator-Shift Register Schematic

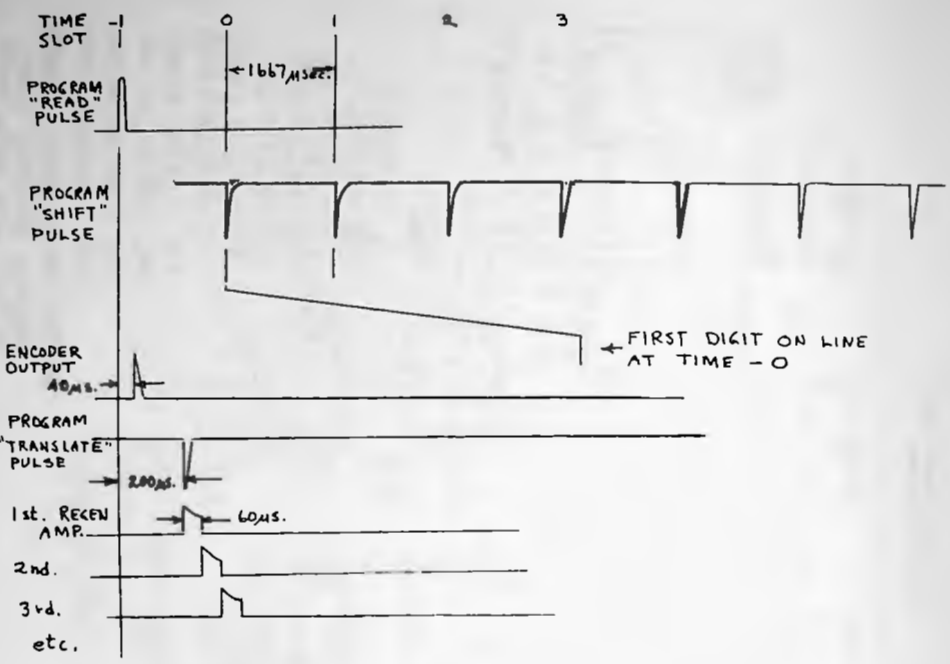
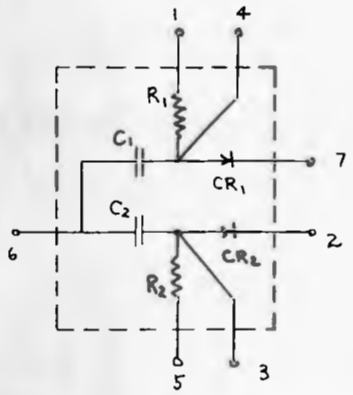


Fig. 12 Translator-Shift Register Programming



Numbers designate Plug pins on packaged assembly

R_1, R_2 51 K ± 10% Allen Bradley – ½ w.
C_1, C_2 1500 μμf ± 10% Erie GP2B
CR_1, CR_2 400A Diode

Post-Potting D.C. Resistance Tests

5 to 3		51 K ± 20%
1 to 4		51 K ± 20%
6 to 3		Open Circuit
6 to 4		Open Circuit
3 to 2	(Forward)	1.0 volt drop max. at 2.0 ma.
2 to 3	(Back)	500 K min. at 20 volts
4 to 7	(Forward)	1.0 volt drop max. at 2.0 ma.
7 to 4	(Back)	500 K min. at 20 volts

Fig. 13 M 1751-3 Diode Control Package Specifications

NOTES ON THE DESIGN OF HIGH SPEED DIGITAL COMPUTERS USING TRANSISTORS

J. H. FELKER

INTRODUCTION
CATALOG OF BLOCK DIAGRAMS
A TRANSISTOR BLOCKING OSCILLATOR
COUNTING DOWN WITH BLOCKING OSCILLATORS
BASIC TRANSISTOR AMPLIFIER
A WORD GENERATOR
DELAY LINE STORAGE
BIT STORAGE WITHOUT FLIP-FLOPS
A BINARY ADDER

Produced under contract Nord-10422

INTRODUCTION

This section reports work that has been done towards a high speed digital computer based on transistors and semi-conductor diodes. Specific components that have been built are discussed. The material is largely a series of independent memoranda written at different times and for separate purposes. This accounts for some of the inconsistencies and the lack of integration that the reader may detect.

1.1 WHAT A COMPUTER DOES

'A statement to the effect that a digital computer is a giant brain is a very helpful statement to make when one is trying to get a group of people interested in digital computers. However, once the people are interested, this statement has lost most of its usefulness. If someone really wants to know how a digital computer works, what kind of processes are used by the computer, and what kind of components are in the computer it doesn't do the person much good to say that a computer is a giant brain. Perhaps a better statement to make is that a digital computer is a desk calculator with an operator built in. The operator is a person of very limited resources, very great speed, capable of following very long instructions exactly as those instructions are given, and incapable of doing anything useful not planned when the original instructions were made up.

The arithmetic unit of the computer will be able to do the same kinds of things that a desk calculator can do except that it will be able to do them much faster. It will be able to add, subtract, multiply, divide or perform any other of those numerical operations that can be described in terms of the elementary ones.

As an aid to visualizing the kind of things that a digital computer has to do, it is convenient to imagine oneself faced with a particular problem in numerical computation and see how it could be done on a desk calculator.

Suppose that it is desired to evaluate

$$\int_1^2 \text{Log } (x + \sqrt{\sin x})\ dx \qquad (1)$$

The first step would be to decide how the integration is to be performed. Suppose that it were known that Simpson's rule for integration would give sufficient accuracy for tabulation intervals of 0.05. One would then obtain a large sheet of paper and lay out the computation as on Figure 1. In column A, the values x are listed. In column B, the sine of the entries in A is tabulated from some table. In column C, the square roots of the entries in B are tabulated after being computed with the calculator. In column D, the sum of A and C is tabulated. In column E, the logarithm of the entries in D is tabulated. This column then contains the value of the integral at intervals of 0.05. Column F contains weighting factors which are used to weight the values before they are summed (these coefficients come from Simpson's rule). The product of column E by F is tabulated in G. The entries in G are then summed and the result, when divided by three and multiplied by the tabulating interval is the desired integral.

It is clear that one of the essential features of computation based on a desk calculator is the large sheet of paper to keep track of computation. The piece of paper stores the program for the computation so that the user can merely "follow instructions" as he operates the calculator and consults tables. The sheet of paper also serves as a storage or memory medium for intermediate results.

An automatic computer, which is what we usually mean by a digital computer, must have not only the arithmetic unit corresponding to the desk calculator, but it must have a memory (corresponding to the large sheet of paper) and a control unit (corresponding to the operator).

In distinguishing between an analogue computer and a digital computer it has been said that a digital computer is a computer which has a language. One of the main uses of the language is for instructions. A common instruction scheme is based on the four address code shown in Figure 2. At any

time the control unit in the machine is concerned with four addresses plus an operation instruction. The operation, abbreviated OP, is the arithmetic operation to be performed next. Addresses one and two are the locations in the memory at which the numbers to be operated on are found. Address three is the address at which the result obtained is to be stored. Address four is the address at which the code group for the next step will be found. What takes place in the computer is: The control unit pulls out of memory a four address code. This tells what is to be done in the next step. The machine takes the two relevant numbers out of storage, it performs the indicated operation, it puts the result back into storage in its appropriate place, then it observes the fourth address, looks in that address, and pulls out the details of the next step.

In computer language, numbers are represented as polynominals just as in other languages. Figure 3 illustrates polynominal notation while Figure 4 illustrates that the base need not always be 10. Because electronic apparatus generally has only two states, open or closed, conducting or not conducting, binary numbering systems are used in most electronic computers. Thus only two symbols are admitted, zero and one.

All words or numbers must be stated in terms of their symbols. The control unit of a computer must be able to take sequences (in time or in space) of pulses (ones) and absence of pulses (zeros) and recognize that a certain connection is to be set up. The arithmetic unit must be able to take the same kind of information and do arithmetic. In the next section of this report block diagrams of components for switching, storage, and computing will be presented, while in later sections the specific details of typical transistors designs will be given.

1.2 COMPONENTS

Existing digital computers were, in most cases, built without major effort to control their size. However, concentrated effort on size reduction will be required in order to design a digital computer that will be small enough to compete with the analogue computers now available .

Techniques for miniaturization have been widely developed, but a fundamental difficulty with reduction in the size of components is the problem of what to do with the heat that the components generate. To get really important reduction in size we must reduce the heat generated as well as the size of the components.

The basic reason that digital computers have to dissipate large amounts of power is that they use components that were originally developed for use where power must be developed to drive output devices. A digital computer may go through a million alterations of state before any output device is operated. With this very loose coupling between the internal operations and the output, it is extremely wasteful to perform every internal operation at a power level sufficient to drive an output device. All internal operations in a computer can be reduced to driving a device from one distinct state to another distinct state. If a small digital computer is to be built, it is essential, therefore, to look for components, in which the power involved in going from one state to the other is a minimum. In voltage-operated devices this power is determined by the voltage swing required to change states and the current required to charge up the parasitic capacitances in the available operate time. In current-operated devices the power is determined by the current swing required and the voltage necessary to drive that current through the parasitic inductances in the time allowed for a change of state. As we reduce power dissipation we will be able to reduce size, which will lower the parasitic capacitance and inductance which, in turn, will permit further reduction in power. Not only must the power required to change state be reduced, but so must the power lost in cathode heaters and similar auxiliaries to active devices.

1.21 CRYSTAL DIODES

A step forward was made when it was realized that active elements such as relays and vacuum tubes are not required to perform the operations essential in digital computers.[1] Any non-linear device has two states, one in which the impedance is high and one in which the impedance is low. Crystal diodes, for example, have two clearly defined states: one in which the impedance is about 200 ohms and one in which the impedance is greater than 50,000 ohms, and crystal diodes have the advantage that they can be driven from one state to the other by a one-volt signal. From the point of view of power dissipation, crystal diodes are far superior to vacuum tubes. The power dissipated in the 50,000 ohm state is negligible. In the 200 ohm state the power dissipated is only 5 milliwatts, while the equivalent vacuum tube diode dissipates one watt continuously in its heater. This lower power dissipation of the crystal, plus its mechanical simplicity makes it an ideal component for miniaturization.

- -

1. The University of Pennsylvania EDVAC and the Bureau of Standards SEAC are computers reflecting this principle.

Computers have been built in which all the logic (switching and arithmetic operations) is performed in crystal diode networks. Being passive elements, crystal diodes have a high degree of reliability, but because they are passive, the signal is attenuated as it progresses through a network of crystals. When the power level has fallen to where it can no longer operate other circuits, an amplifier is inserted to restore the signal to its original strength. A recent machine[2] used about twenty crystals for every tube. It has been found practical to drive as many as thirty-two crystals at a mega-cycle rate with one tube.

1.22 TRANSISTORS

A major forward step was made when it became possible to replace the tube amplifiers with small devices that dissipate less power. The transistor is such a device. It can be driven from one state to the other by a very low voltage. In addition, it works well with crystal diode circuits, since the transistor itself is a crystal device. Also, it is well suited to use in regenerative amplifiers. In an amplifier for use in a digital computer there is no premium on linearity; in fact, regeneration is desirable because it decreases rise time and increases the effective power gain. An all "semi-conductor" computer would represent an advance in the art and is to be attainable now. A transistor amplifier, described in a following part of this report, has been developed which will amplify half-microsecond pulses with rise times of less than 0.05 microsecond and which requires less than 50 milliwatts of power.

---

2. The Bureau of Standards SEAC

TYPICAL COMPUTATION SHEET

A	B	C	D	E	F	G
x	SIN A	$\sqrt{B}$	A & C	LOG D	COEF.	E x F
1.0					1	
1.05					4	
1.10					2	
1.15					4	
1.20					2	
1.25					4	
1.30					2	
1.35					4	
1.40					2	

FIG. 1-1

FIGURE I-2 FOUR-ADDRESS CODE

FOUR-ADDRESS CODE

(OP, A1, A2, A3, A4)

OP = OPERATION TO BE PERFORMED

A1, A2 = ADDRESSES OF NUMBERS TO BE OPERATED ON

A3 = ADDRESS AT WHICH RESULT IS TO BE STORED

A4 = ADDRESS AT WHICH CODE GROUP FOR NEXT STEP WILL BE FOUND

FIGURE I-3 NUMBER AS A POLYNOMIAL

$$a_n x^n + a_{n-1} x^{n-1} --+ a_1 x^1 + a_0 x^0 + a_{-1} x^{-1} + a_{-2} x^{-2} ----$$

WRITE IN POSITIONAL NOTATION AS:

$$a_n a_{n-1} --- a_1 a_0 . a_{-1} a_{-2} ---$$

EXAMPLE: RADIX 10

$\pi = 3.1415$

$a_2 = a_1 = 0 \quad a_0 = 3 \quad a_{-1} = 1 \quad a_{-2} = 4 ---$

FIGURE I-4 EXAMPLE OF RADIX SEVEN

7^3	7^2	7^1	7^0	7^{-1}	7^{-2}
343	49	7	0		

DECIMAL 1000 IN SEPTENARY CODE

2 6 2 6

COUNTING

2	6	2	6	=	1000
2	6	3	0	=	1001
2	6	3	1	=	1002
2	6	3	2	=	1003
2	6	6	6	=	1028
3	0	0	0	=	1029

CATALOG OF BLOCK DIAGRAMS

2.1 INTRODUCTION

Starting from schematics for certain basic logic elements this memorandum presents a fairly complete catalog of diagrams of the units required in a digital computer. The designs have the common feature of performing all logic without the use of active elements. Active elements are employed only as gain producing devices (repeaters) to make up for attenuated signals. Thus the designs can be implemented with vacuum tubes, transistors, magnetic amplifiers, thermistors, or any other devices that have gain at the frequencies requiring amplification.

Serial operation has been used in all the designs. In serial operation all the digits of a word or number are transmitted in series along a single wire with the least significant digit first. The interval between the rise of successive digit pulses is referred to as one digit time. The interval between successive numbers or words is commonly referred to as the word time. Digit times of one microsecond have proved to be very practical and words as long as fifty binary digits have been employed (equivalent to 15 decimal digits).

Transistors make excellent amplifiers for pulses occurring at a megacycle rate and the reader can equate one digit time to one microsecond whenever he wants an estimate of the computing times required by the various machines described.

The reasons for choosing serial operation instead of parallel are quite simple. In a parallel system, each digit of a number appears in a different circuit. In a parallel adder a separate adder stage is used for each digit, whereas in a serial adder only one adder stage is used, and this one stage handles all the digits. Not only are the arithmetic operations performed with fewer components in the serial machine, but switching is also simplified. To switch an n-digit parallel number requires n switches to switch the separate wires, whereas in the serial machine only one switch is required.

All of the digital computers that have thus far been put into operation have exhibited the common defects of undesirably high failure rate and very large size. These particular defects are easier to avoid in a serial design than in parallel machines, because a serial machine uses many less components.

2.2 BASIC BUILDING BLOCKS

The proposed computer elements are based on an assembly of a relatively small number of different kinds of basic units. The basic units have been limited to an or-circuit, an and-circuit, an inhibitor-circuit, an amplifier, and a storage cell based on a delay line. The designs suggested are similar to those in the EDVAC designed at the University of Pennsylvania and the SEAC designed at the Bureau of Standards. Some of the basic components have been described in the literature.* They are described again in the following paragraphs to provide a specific basis for the estimates of numbers of parts that are a goal of this study.

2.21 Or-Circuit

An n-terminal or-circuit (see Figure 2-1) develops an output when any one of its input terminals is energized. Crystal diodes are put in series with each input to prevent a pulse at one input from feeding back to any of the other inputs. An n-terminal circuit, therefore, requires n crystal diodes.

2.22 And-Circuit

An and-circuit having n input terminals develops an output only when all n of the input terminals are energized. In figure 2-2 each of the inputs is returned to a negative voltage and the output is clamped slightly below ground by X_{n+1}. Only when all of the inputs rise above

- -

*D. R. Brown and H. Rochester, "Rectifier Networks for Multiposition Switching", Proceedings of the I.R.E., Feb. 1949; and Tung Chang Chen, "Diode Coincidence and Mixing Circuits in Digital Computation", Proceedings of the I.R.E., May 1950.

ground will X_{n+1} be cut off, permitting the output to rise. Thus the output consists of the overlapping part of the inputs. An n-terminal circuit is seen to require n + 1 crystals.

Two terminal and-circuits are used extensively in serial computers for retiming signals. One terminal of the circuit is fed by the signal to be retimed while the other is fed by digit pulses from a master clock. There will be an output from the circuit only for the overlap of the master digit pulse and input signal. The circuit whereby the output is made a replica of the pulse from the clock is discussed in paragraph 2.24.

2.23 Inhibitor Circuit

An inhibitor terminal can be added to any and-circuit or or-circuit. Such a circuit operates as though there were no inhibitor terminal when the inhibiting pulse is not transmitted. When the inhibiting pulse is present, however, the circuit prevents any output from being developed. The inhibiting circuits used in the designs proposed herein are of the simple variety shown in Figure 2-3 where positive inputs synchronized in time are required. Note that the signal to be inhibited is passed through an eighth-digit delay line, while the inhibiting pulse is passed both through and around a quarter-digit delay line. This insures that the inhibitor pulse will, in effect, arrive earlier than the signal pulse and last longer. In the absence of input pulses, crystal X_4 will clamp the output at ground because input B is returned to a negative potential. Note that X_1 and X_2 are returned through the transformer to a positive potential. If input B goes positive (without an inhibiting pulse appearing on input A) X_4 will be cut off and the output voltage will rise until it is clamped at the positive potential to which X_1 and X_2 are returned. If there is an inhibiting pulse (positive) it is inverted by the transformer and will carry the cathodes of X_1 and X_2 negative, which will keep X_4 conducting, no matter what happens at B. Thus, if pulses A and B were written as a two-digit binary number AB, the circuit translates 01 into a 1 at the output. It translates 00, 10, and 11 into zero at the output.

2.24 Active Elements

Active elements will be used not as flip-flops or switches but as repeating amplifiers to make up for attenuation in crystal circuits and delay lines. The standard use will include a retiming feature as well as amplification. This is shown in Figure 2-4.

In the design of the machine, wherever a pulse is likely to suffer intolerable attenuation, deformation, or a variable delay, a circuit like that of 2-4 is inserted.

The assembly shown has two inputs, A and B. Input A is the pulse to be retimed and amplified. Input B comes from the master clock. This component supplies reference pulses (known as digit pulses) every digit time. These pulses are available in various phases, that is with various but accurately controlled delays of a fraction of a digit time. The pulse fed to B is selected to rise sometime between the expected rises and falls of the pulses on A.

If there is no input on A, there will be no output from the amplifier because of the and-circuit. If there is an input on A, when the digit pulse arrives the amplifier output will rise with a rise time determined by the digit pulse (assuming that the amplifier pass-band does not limit it). Part of the amplifier output is fed back through an or-circuit to the and-circuit. This insures that the output pulse will not fall until the reference digit pulse does, even though pulse A may have ended after B rose.

This reshaping with crystal circuits and an amplifier is the way in which every pulse is maintained with the desired time synchronization. Pulse A may vary somewhat in the delay it has suffered but the output pulse will still leave the amplifier at a time determined only by the reference pulse from the master clock. Thus, the pulses in the computer are made to have fixed durations and to occur at designated times.

In the tabulations of components in the following paragraphs an active element is intended to represent as one unit the tube, or transistor, with transformer. The term will include neither the and-circuit nor the or-circuit shown in Figure 2-4.

2.25 Storage Cell

The basic storage cell proposed is not a static device like a flip-flop but is an electric delay line plus an amplifier. When vacuum tubes are used, this type of storage saves one active element in a one-digit storage cell and is believed to be a more reliable use of active elements. In larger storage units more elements will be saved.

A block diagram of a cell is shown in Figure 2-5. The unit has three inputs: digit pulses, the signal to be stored, and an erase pulse. The digit pulses are received from the master clock every digit time and are used to retime (as discussed in paragraph 2-4) the output of the delay line before it is amplified and recirculated. The erase signal is received whenever new data are to be stored. It serves to erase the data in storage, blocking the delay line output from its input until the new data have been inserted.

The delay line may be long enough to store one word or just one digit of data. It is believed that up to fifteen-digit delay lines with lumped impedances can be built to hold the delay constant to within a small fraction of one digit time. Depending on the length of a word, it may be necessary to break one-word lines into sections and insert an amplifier between sections to retime and regenerate the pulses stored. To insure conservative estimates, the estimates made in following sections are on the basis of regeneration after every eight digits of delay.

2.26 Summary of Components Required in Basic Units

The circuits which have been discussed use the components listed below. This table will be used frequently in estimating the parts required in the larger assemblies:

TABLE 1

Unit	Crystals	Digits of Delay	Active Elements
N-Term'l Or-Circuit	N	0	0
N-Term'l And-Circuit	N+1	0	0
Inhibitor-Circuit	4	3/8	0
N-digit Storage Cell	10*	N	1*

2.3 SWITCHES

The switches are planned to combine a switching and a storage function. When a switch is given instructions to go to the k'th position, it goes there and it remembers that it is to remain there (self-locking operation). All the elements of the switches have been discussed in paragraph 2. How these elements are combined to make switches is described in the following:

2.31 Double-Throw Switches

A single-pole double-throw switch as shown in Figure 4-6 consists of a storage cell and a switch unit. When a "one" is stored in the storage unit, as the result of a pulse on the switching instruction lead, the left-hand and-circuit of the switch unit will pass signal "a" while the inhibitor blocks signal "b". When a zero is stored, the and-circuit will block signal "a" and the inhibitor-circuit in the switch unit will pass signal "b".

Whenever the switch is to be reset, an erase signal is fed to the storage unit, which then drops its old instruction and goes to position "b" unless the new instruction sets it to "a".

- - - -

The switch unit itself uses thirteen crystals 3/8 digit of delay, and one active element, as can be ascertained from Figure 2.6 and table 1. A complete single-pole double-throw switch has 23 crystals, 1-3/4 digits of delay, and two active elements. A two-pole double-throw switch would have two switch units and one storage unit. A three- or four-pole switch would have three or four switch units, one storage unit, and perhaps an extra amplifier to prevent the switch units from loading down the storage unit excessively.

2.32 Multiposition Switches

Switches with more than two positions can be assembled in a slightly different manner from the double-throw switches. Suppose, for discussion, that an eight-position switch is needed. A three-digit code must be sent to the switch to specify the position it is to select. A convenient way to operate such a switch is to translate the three-digit code into a six-wire code so that each digit is represented by signals of opposite phase on a pair of wires. In this system, a one is represented by a positive pulse on the positive bus and a negative pulse on the negative bus. A zero is represented by a negative pulse on the positive bus and a positive pulse on the negative bus. A three-terminal and-circuit is provided for each of the eight lines that may be selected. The and-circuits are operated by positive pulses, and all are connected to either one or another of each pair of wires representing the three digits of the position code. The and-circuit on the fifth (101) wire, for example, is connected to the positive bus of the pair representing the coefficient of 2^2, to the negative bus of 2^1 and the positive bus of 2^0. The three leads of the and-circuit will be energized by positive buses only when 101 is sent to the switch as instructions.

A single-pole eight-position switch is shown in Figure 2-7. The eight three-terminal and-circuits (one for each horizontal lead), plus the nine-terminal or-circuit make an easily studied crystal matrix which feeds another and-circuit for retiming of pulses and an amplifier on the output. The and-circuits are fed by switches S_0, S_1, and S_2, which are driven by the switching instructions. These three switches are like the one shown in Figure 4.6. S_0 is

operated by the coefficient of 2^0, S_1 by the coefficient of 2^1, and S_2 by the coefficient of 2^2. As the figure is drawn, only the and-circuit on line 7 (111) is energized. A pulse will be received at the output whenever a pulse is put on line 7. Pulses can be put on any of the other lines without getting to the output, because every other line is held negative by at least one lead from S_0, S_1, or S_2. If an erase signal were sent to the three storage cells of S_0, S_1, and S_2, the number 111 would be erased leaving 000, and the switches would be free to move to new positions and select another one of the eight inputs.

Based on this type of mechanization, an r-position switch requires n single-pole double-throw switches, where n is the smallest integer equal to or greater than $\log_2 r$. The switch requires r (n + 1) crystals plus four crystals for retiming the output, and an output amplifier.

2.33 Summary of Components Required for Switches

The table below gives estimates of the crystals, digits of delay, and active elements required in switches of different degrees of complexity.

TABLE 2

Type of Switch	Crystals	Digits of Delay	Active Elements
1P.2T	23	1-3/4	2
2P.2T	36	1-3/4	3
4P.2T	63	6-1/2	6
1P.4T	61	2-3/4	4
1P.8T	105	6	7
1P.16T	160	7	9
1P.32T	311	8-3/4	11

where, 2P.2T = two-pole, two-throw switch.

2.4 ARITHMETIC UNITS

The following discussions show how the basic units discussed earlier are combined to perform the necessary arithmetic operations. The discussion of arithmetic units is preceded by some information on the handling of negative numbers.

2.41 Handling of Negative Numbers

The last digit place of every number is reserved to indicate the sign of the number. Positive numbers have a zero in the last place. A negative number is obtained by taking the two's complement of the positive number. This results in every negative number having a one in its last place.

A point of interest is that even though either or both the multiplicand and multiplier are negative the correct sign will be obtained for a product provided the negative numbers are increased from their normal length W to 2W-1 by filling in ones before multiplying. This is necessary because the product of two W-digit numbers, where the last digit specifies sign, is a number 2W-1 digits long. Unless the multiplicand and multiplier are increased to length 2W-1 (when negative) the 2W-1 place may be incorrect. This lengthening is not required for positive numbers, because the digits from W to 2W-1 would be zeroes if they were filled in and would contribute nothing to the product.

A consequence of the above is that the product XY can be obtained to 2W-1 places with W(W-1) elementary additions provided X and Y are positive while W(2W-1) additions will be required if both are negative. Where fast multiplication is desired it may be advisable to convert negative numbers to positive ones, multiply, and then adjust the sign of the product.

This system is equivalent to the ten's complement method used in decimal calculators. In a decimal calculator operating with three significant figures, a fourth place might be provided for the sign. The number -187 might be represented by its ten's complement 9813. Then, for example, if -187 were required to be added to 500 the operation would be to add 9813 to 500 which gives 10313 and thus is recognized as 0313 since the machine was assumed to have only four digit places.

A negative number (two's complement) can be obtained in the binary computer by first forming the one's complement (changing all zeroes to ones and vice versa by means of an inhibitor circuit), and then adding one. Figure 2-8 shows several examples of binary arithmetic performed with negative numbers.

2.42 Adder

The adder can be considered as a translator with three inputs: addend, augend, and carry. It is a simple translator in that its output is a function only of the number of ones among its three inputs, as can be seen from the table below:

TABLE 3 - BINARY ADDITION

INPUTS			OUTPUTS	
Addend	Augend	Carry	Sum	New Carry
0	0	0	0	0
0	0	1		
0	1	0	1	0
1	0	0		
0	1	1		
1	1	0	0	1
1	0	1		
1	1	1	1	1

The combination 0 0 0 is automatically taken care of in the adder shown in figure 2-9. The three dashed circuits at the left of the block diagram recognize the other three situations among the three inputs. The situations are: at least one "one", at least two "ones", and three "ones" among the inputs. If there is only one "one", it will go through the bottom or-circuit, the following inhibitor circuit, and then another or-circuit. After being reclocked and amplified it will provide a "one" as the sum. In this case none of the and-circuits on the A, B, and carry leads will have operated. If there are at least two "ones" on the A. B. and carry leads, at least one of the three two-terminal and-circuits in the dashed box will operate, with two results. The output of the three-terminal or-circuit at the bottom left of the diagram will be inhibited so that it makes no contribution to the sum. In addition, a carry signal will be developed which is delayed one digit, reclocked, and amplified to serve as the carry for the next

augend and addend. If there are three "ones", the three-terminal and-circuit at the top and left-hand side of the diagram will operate and develop a sum of one. The three two-terminal and-circuits on the left of the diagram will also have operated and provided the carry. Thus, the adder table is mechanized with two active elements (amplifiers).[1]

The inhibitor circuit in series with the carry lead should be noted. This circuit is fed by a word pulse as well as the carry digits. The word pulse is received in synchronism with the first digit of every number. The word pulse will inhibit the carry pulse if one is present and will prevent a carry developed in one problem from being used in the next. This feature is required in the addition of negative numbers.

The carry lead is brought outside the adder to facilitate subtraction. Suppose x is to be subtracted from y. The number y might be fed to the addend terminal and the number x fed through an inhibitor circuit to the augend terminal. The inhibitor would also be fed by digit pulses from the master clock, and the augend would therefore be the one's complement of x. A one would be inserted into the carry terminal in synchronism with the first digits of y and of the one's complement of x. The sum out of the adder would then be x-y (see par. 4.1). The adder produces a sum within a fraction of a digit time after it receives an input. Thus there is only a small delay in obtaining the sum of two numbers.

Table 1 can be consulted to show that the adder of Figure 2-9 requires 38 crystals, 1-3/8 digits of delay, and two active elements.

2.43 Accumulator

A block diagram of an accumulator is shown in Figure 2-10. The output of the adder is fed back to its input through a w-digit delay line. The output of the delay line is continuously reclocked in the and-circuit in

---

[1] The operation of so many crystal circuits in series without amplifiers may be questioned but it is believed to be feasible at a microsecond rate. Experience may show, however, that another amplifier is required.

accord with digit pulses from the master clock. The timed pulses are then amplified in a one-stage amplifier to make up for attenuation in the delay line. Whenever a new accumulation is to be started, an erase signal is fed to the inhibiting circuit. This signal is w digits long and blocks the output of the delay line from the adder and insures that the new accumulation will start from zero. An accumulator to accumulate sums 48 digits long will require 76 crystals, 48 digits of delay, and 8 active elements.

2.44 Multiplier

The multiplier that is discussed in the following paragraphs is designed to multiply two positive members, each having W digits, in W(W+1) digit times.[1] Operating at a megacycle rate, the product of two 48 binary digit numbers would be obtained in 2352 microseconds. It is believed most efficient to convert all negative members to positive ones before multiplying and to adjust the sign of the product according to the rules of algebraic multiplication (see paragraph 4.1). The components to be added to the multiplier to make the sign correction have not been included in this study. The multiplicand (x) is multiplied by the first (least significant) digit of the multiplier (y) in the first (W+1) period and by the W'th digit of y during the w'th (W+1) period. The w partial products are added together, with each partial product moved over one place with respect to the preceding one as it is accumulated.

The multiplication table in binary arithmetic is very simple. If it is desired to multiply the binary number x by a single binary digit, m, x is connected to the input of a switch that is held open if m = 0, and closed if m = 1. The output of the switch will be mx. Thus, in multiplying x by the successive digits of the number y, it is only required to recirculate x through a switch that is opened and closed at the appropriate times according to whether the successive digits of y are zeroes or ones.

Part of the multiplier is shown in Figure 2-11. The multiplicand (x) is stored in delay line 1, and the multiplier (y) is stored in delay line 2. After the first word period, S1-A and S1-B are thrown to position "b" and

---

[1] This multiplier would obtain the correct sign for multiplication by negative members W/2 digits long.

the multiplier and multiplicand recirculate in their delay lines until the multiplication is finished. DL1 (delay line 1) storing the multiplicand has one more digit of delay than line DL2 in which the multiplier is stored. Therefore, after the first circulation of x through DL1, the least significant digit of x leaves DL1 just as the second least significant digit of y leaves DL2. After the W-1'th circulation of x, its least significant digit leaves DL1 just as the most significant digit of y leaves DL2.

The output of DL2 is fed continuously to an and-circuit. Every W+1 digits, an examining pulse is fed to the and-circuit, and if the digit of y coming out of DL2 is a one at that time, the and-circuit develops a one in its output. If, when the "examining" pulse comes along, the coefficient of y leaving DL2 is a zero, the and-circuit will have no output. Because DL2 has a delay of w digits, the inhibitor circuit in effect filters out the successive digits of y, choosing a new digit every w+1 digits times. The examining pulse is also sent to S2, where it is the erase signal of the storage cell associated with S2 (see paragraph 3.1). The output of the and-circuit goes to S2 as switching instruction and is stored there. Thus, S2 stores either a one or a zero, according to whether the digit of y examined is a one or a zero. The switch is arranged so that when a one is stored, S2 is closed, and when a zero is stored, it is open.

A new partial product is obtained at the output of S2 every w+1 digit times by the process described above. To accumulate them, an adder can be provided which accepts as one input the output of S2, and as the other input its own sum output delayed by w digits. The one digit place difference in the arrival of the two inputs of the adder automatically provides the "shift" feature necessary in adding the partial products. The only difficulty is that the product of two w-digit numbers may be a number 2w digits long which cannot be stored in the one-word delay line between the output of the adder and its input. Therefore, it is necessary to provide two delay lines and to use one of them to accumulate half the answer and the other for the second half.

As the complete block diagram (Figure 2-12) shows, at any time either DL4 or DL5 is receiving its input from the adder and feeding its output back into the adder. The

other delay line, at the same time, is feeding its own output into its input. Switching of the delay line is performed by the four ganged sections of S4. The switch is operated in such a way that the adder is connected to DL4 all of the first word period and to DL5 for one digit time longer of each succeeding word period, with the result that during the W+1'th word time the adder is connected to DL5 for the complete word period.

The switching instructions for S4 are developed in the following manner. The storage unit associated with S4 receives the word pulse from the clock every w digit times and uses it as an erase signal to empty its one-digit delay line. Every w+1 digit times the "examining" pulse is fed to the storage unit of S4 and stored there. S4 is arranged so that with a one in its storage unit, it selects position "a", and with a zero in its storage unit, it selects "b". Thus, S4 is at "a" from pulse 1 to pulse w, when the word signal at w erases the one in the storage unit. The switch goes to "b" and stops there for one digit time, then the w+1 pulse is written into the storage unit of S4 and the switch goes once more to "a". The second word pulse at 2w again erases the one in the storage unit and sends S4 to "b", where it stays for two digit periods until the second w+1 pulse at 2(w+1) writes in another one and pulls S4 to "a". Thus, S4 is held at "b" for one digit time longer each word period.

Figure 2-13 gives a digit-by-digit account of the operation of a four digit multiplier in multiplying 1001 by 1101. In studying the figure it is recommended that the reader keep in mind the position of S4 at each step so that he can tell where DL4 and DL5 are connected. It should also be noted that the outputs of DL4 and DL5 are always their inputs of four digits earlier.

A multiplier based on Figure 2-12 can be mechanized for 15 digit words with 204 crystals, 5w[1] digits of delay, and 20 active elements.[2]

- -

1. The digits of delay required in inhibitors and switch memory units have not been counted. These uncounted delays might total eight or so digits.

2. A multiplier for longer words would have longer delay lines and consequently would require additional retiming circuits. It is estimated that a 48 digit multiplier could be built with 324 crystals, 5w digits of delay, and 40 active elements.

2.45 Divider

The divider operates with repeated subtractions and shift operations much as one would do in pencil and paper division. Figure 2-14 shows two division examples. In the first one, three is divided by eight. First, it is seen that 1000 will not go into 11, so the first zero is written in the quotient. 100.0 will not go into 11, so the second zero is written. 10.00 will go into 11.0, so a one is written in the quotient and 10.00 is subtracted from 11.00. The remainder is 1.00. The binary point is again shifted in the divisor, giving 1.000 which goes exactly once, and another one is written in the quotient giving an answer of 0.011 or 0 x 1/2 + 1 x 1/4 + 1 x 1/8 equal to 3/8.

It should be noted that at each step the divisor is either subtracted from the remainder or nothing is subtracted and that a shift to the right is made each time before the divisor is subtracted. In effect, the successive subtrahends are obtained by repeated divisions of the divisor by two. For machine operation, the awkward part of the above process is that it cannot be ascertained whether a subtraction should be made without making it and seeing whether or not the remainder is positive or negative. If the remainder is negative, the subtrahend should be added back. This necessity for sometimes adding back the subtrahend may be avoided. Let x by the subtrahend and R the minuend. Suppose x is subtracted from R and the remainder found to be negative. During the next step the machine should subtract x/2 from R, but the machine has lost R by subtracting x from it. R can be regained and x/2 subtracted at the same time by adding the factor x/2 to the negative remainder R - x which gives the desired result R - x/2. Thus, on the successive steps the machine either subtracts half the previous subtrahend or adds half the previous subtrahend, depending on whether the previous remainder was positive or negative. The process is further simplified by avoiding subtraction through adding the two's complement of the subtrahend rather than subtracting the subtrahend.

An example of the simplified division process is shown in Figure 2-14. The machine is assumed to have five digit places. The first step is to fill out the divisor and dividend to 2w-1 places, and then to form the complement of the divisor (according to the methods discussed in par. 4.1).

The divisor complement is first added to the dividend. The 2w-1'th digit of the sum is noted. In the example, the digit is a one, indicating a negative result. The machine then writes a zero in the quotient, drops out the 2w-1'th digit of the previous sum, and adds half the divisor (divisor shifted one place to the right). The 2w-1'th digit is again examined and is found to be a one. Another zero is written in the quotient, the 2w-1'th digit is dropped out, and the divisor is again shifted and added. This time the 2w-1'th digit is zero, indicating a positive sum. A one is written in the quotient, and the complement of the divisor is shifted and added. This process is continued until the five-digit answer 0.0110 is obtained. Note that the first operation was made with the binary points of the divisor and dividend lined up. In general, it would be necessary to start with the least significant digit of the divisor written under the most significant digit of the dividend. 2w(2w-1) digit times are required for division in a general purpose machine.

The divider block diagram is shown in Figure 2.16. The divisor is stored in delay line 1 and recirculates there becoming available every 2w-1 digit times. The one's complement of the divisor is taken by the inhibitor circuit fed by DL1 (delay line 1). Every 2w-1 digit times switch S2 selects either the one's complement of the divisor or the divisor itself to go into the adder. The switch selects the one's complement when it is desired to subtract. However, to obtain subtraction it is necessary to add the two's complement. This is done by bringing out the carry terminal of the adder, and whenever S2 is thrown to select the one's complement a one is sent into the carry terminal and, in effect, makes the augend the two's complement.

The storage cell of Switch S2 receives an erase signal every 2w-1 digit times. At the same time, the switch receives a one or a zero from the inhibitor circuit on the adder output. If this inhibitor sends a one to the switch, the 2w-1'th digit of the sum just completed was positive and switch S2 goes to position "b" to select the one's complement of the divisor for the next addition. The output of the inhibitor circuit is also sent to the carry terminal of the adder, so that when S2 is at "b", a one will be injected to convert the one's complement to a two's complement. Each output of the inhibitor is also a digit of the quotient, for if the 2w-1'th digit of the sum is a zero, a one should be written in the quotient and the inhibitor supplies that one.

It is written in delay line 3, which has a circulation time of 2w. Since a new quotient digit is received every 2w-1'th digit time, delay line 3 has as its output the digits of the quotient in reverse order to that in which they are obtained, which puts the least significant digits first as they should be.

The above discussion shows how the augend is obtained for the adder. The addend is obtained from S1B. This switch selects the dividend during the first 2w-1 period, and thereafter selects the previous sum from delay line 2. Note that an inhibitor circuit between the adder and DL2 blocks out the 2w-1'th digit each time to keep the remainder from growing longer than 2w-1. Since delay line 2 has 2w delay and delay line 1 has 2w-1 delay, the augend arrives 1 digit early each 2w-1 period with respect to the addend. This feature supplies the successive divisions by 2 that are required.

Delay line 4 exists merely to supply the 2w-1 pulse that is sent to S2 and the inhibitor circuit on the adder output. Switch S1 is driven by the master clock. Its sections are switched to "a" at the end of the computation period to empty the lines of the dividend, divisor, and 2w-1 pulse preparatory to solving a new problem. After 2w-1 digit times, all sections of S1 are thrown to "b" and recirculate their data.

An estimate based on Tables 1 and 2 of the parts required by the divider of Figure 2-16 shows that to handle 48 digit numbers, 371 crystals, 384 digits of delay, and 51 active elements would be required. This assumes that the delay lines have their data regenerated every w/2 digits. If the data are regenerated only every w digits, only 36 active elements are required. The divider is seen to be of about the same complexity of the multiplier.

2.46 Square Rooter

This machine obtains the square root of a number to w places in $2w^2$ digit times. It mechanizes the ordinary arithmetic method of taking square roots. Before describing this unit, it is well to consider the algebraic basis of the arithmetic method normally used.

2.461 Development of Operating Principle

In Figure 2-17 an arithmetic problem is solved in the right-hand column, and the justification for each step is shown in the left. Note that a new digit place is obtained in each step, and that the only processes involved, apart from adding zeroes, are multiplication by a number less than ten and subtraction. The method can be extended to binary operation as shown in Figure 2-18. Binary operation is somewhat simplified because the only processes needed besides adding zeroes are multiplication by one or zero and subtraction.

For machine operation, there is a drawback to the method illustrated in Figure 2-18. Sometimes, as in the fourth step, the subtrahend must be multiplied by zero, not one, before it is subtracted. The only way the machine can find out it should have multiplied by zero is for it to assume it should multiply by one, form the subtrahend on that assumption, and subtract it. If the machine obtains a negative remainder it knows that it should not have subtracted anything at all. The simplicities of binary arithmetic make it possible to devise a method of operation in which the necessity of adding back the subtrahend is avoided. [1]

Suppose, with reference to the algebra of Figure 2-17, that the machine is taking the square root of x and has just established the partial answer

$$y_{n-1} = h_1 + h_2 - - - - h_{n-1}$$

and is about to determine the value of h_n, where h_n will be either a one or a zero multiplied by 2^p where p is an integer.

At this stage the remainder in the machine is

$$R_{n-1} = x - y^2_{n-1}$$

---

[1] This procedure is believed to be a new contribution.

and the next operation will be to form the trial subtrahend

$$(2\ y_{n-1} + 2^p)\ 2^p$$

and subtract it from R_{n-1}, which will leave as the new remainder

$$R_n = x - (y_{n-1} + 2^p)^2 = x - y^2_n$$

If R_n is a positive number the machine will enter a one as the coefficient of 2^p in the answer and proceed to obtain the coefficient of 2^{p-1} in a similar manner.

If, on the other hand, R_n turns out to be negative, the machine should write 0 as the coefficient of 2^p in the result (thus $y_n = y_{n-1}$) and subtract

$$(2\ y_n + 2^{p-1})\ 2^{p-1} = (2\ y_{n-1} + 2^{p-1})\ 2^{p-1}$$

from R_{n-1}, which would give the correct value

$$R_{n+1} = x - (y_n + 2^{p-1})^2 = x - (y_{n-1} + 2^{p-1})^2$$

Unfortunately, the machine does not have R_{n-1} available because it has subtracted and obtained R_n. However, to get the desired R_{n+1}, the machine need only add to the erroneous R_n.

$$[\ 2\ y_n + 3\ (2^{p-1})\ ]\ 2^{p-1}$$

instead of subtracting

$$(2\ y_n + 2^{p-1})\ 2^{p-1}$$

which would have been done if R_n had been positive. There is only a slight difference in the two operators above. When a positive remainder is obtained, the machine adds 01 times 2^{p-1} to 2y and then adds 2^{p-1} zeroes at the end of the sum to obtain the new operator. In the event that a negative remainder is obtained, the machine adds 11 (3) times 2^{p-1} to 2y and then adds 2^{p-1} zeroes at the end of the sum. On Figure 2-19 two examples are given of the simplified method.

The square root machine, Figure 2-21, obtains a new digit of a w-digit answer every 2w digit times. In each 2w period the machine either subtracts or adds, depending on whether the digit determined in the previous period is a one or a zero. The minuend or the addend for a given 2w period is obtained from the remainder or the sum of the previous period. The subtrahend or the augend is formed from the answer digits already obtained. The manner of forming the operator is dependent upon the value of the digit determined in the preceding step as is the decision to whether the operator is added or subtracted.

Before going into the discussion of the block diagram it is well to consider in detail how the square root of a specific number would be taken. For this purpose a relatively short number is more convenient than a larger one. In Figure 2-20 the second example of Figure 2.19 (for which W = 5) has been rearranged in the sequence in which the digits would be operated upon by the machine. The top row lists the time (measured in digit times from the start of the problem) at which the digits listed arrive. The second row lists the digits of the operators that are the digits that make up the operators used as subtrahends or augends. The fourth row shows the results of subtracting or adding the operators above, and the next row shows the answer digits as they are obtained. The final row shows when the answer digits are required to reappear in order to be used in the third row.

As Figure 2-20 shows, in the first period the first operator is the number (of 2w digits = 10) whose square root is desired, while the second operator is a one inserted at digit time 2w - 1 (9). The last digit of the remainder is examined at time 2w (10). Since the digit examined is a zero, C_1 is written as one, the most significant digit of the answer. If the 2w'th digit were one, zero should be written in the answer as the coefficient of 2w; since in the example C_1 is written as one, the remainder of the first step is used as the minuend for the next period. The subtrahend is formed by the digit C_1 delayed 2w-1 (9) digit times and by the one subtracted in period one delayed by 2w-2 (8) so that it arrives at digit time 17. The digit C_1 was a one, so in the second period subtraction is performed and the 2w'th digit of the remainder examined. This digit is seen to be a zero, so the remainder obtained on the second 2w period is used as the minuend for the third period and a one is written in the answer as C_2.

The subtrahend for the third step is formed from the digit C_1, delayed 2w-1, the digit C_2 delayed by 2w-3, and a one delayed 2w-2 (8) from the time (17) it was used in step 2 to arrive at 25.

The remainder in the third step has a one in the 2w'th place, indicating it is negative. The remainder is, therefore, used as the addend for the fourth 2w period and a zero is written as C_3. The augend for the fourth period is formed by delaying the number C_1 C_2 by 2w-1 digit times from the time it was used previously, by delaying digit C_3 2w-5 digits (so it arrives at time 35), and at digit times 33 and 34 writing ones. The one at time 33 is the one written at time 25 delayed by 2w-2. The one written at time 34 is written because the remainder of the third period was negative.

From the figure it is seen that the sum in the fourth step is still negative. The sum is used as the addend for the fifth step and another zero is written in the answer as C_4. The augend for the fifth step is formed by the word C_1 C_2 C_3 delayed by 2w-2 from the time it was used in the

fourth step, the digit C_4 delayed by 2w-7, so that it arrives at time 43, and the word 11 delayed by 2w-2 from the time it was used in the fourth period. When this augend is added, the 2w'th digit of the sum is a zero, and a one is written in the answer as C_5.

2.462 Rules of Operation for Square Root Machine

The procedure can be generalized into the following rules for using an adder to obtain the square root of x, where the square root is written as $C_1\ C_2\ C_3\ C_4$ etc., and each coefficient is a one or a zero.

Rule 1 - Examine every 2w'th digit of the output of the adder, and if the examined digit is zero, write one as the answer digit. If examined digit is one, write zero as answer digit.

Rule 2 - In the first 2w period, use x as the addend. In all other periods, use the output of the adder delayed by 2w digit times.

Rule 3 - In the first 2w period form the augend for the adder by taking the two's complement of a number that consists of a single one written at digit time 2w-1. In all other periods determine the augend according to rule 4.

Rule 4 - If the answer digit C_{n-1} determined in the n-1'th 2w period is a one, form the augend for the n'th period by rule 5. If C_{n-1} is a zero, form the augend by rule 6.

Rule 5 - Form the augend for the n'th period by taking the two's complement of q where q consists of the three parts

(a) The word $C_1\ C_2$ - - - C_{n-2} delayed by 2w-1 from its use in the previous step.

(b) The digit C_{n-1} delayed by 2w+3-2n from the time it was obtained in the previous step.

(c) The digit one written two digits earlier than C_{n-1}.

Rule 6 - Form the augend for the n'th period from the three parts listed below:

(a) The same as in Rule 5.

(b) The same as in Rule 5.

(c) The same as in Rule 5, but with an extra one inserted at digit time 2w+2-2n.

2.463 Block Diagram

One way of studying the block diagram (figure 2-20) is to take each of the rules stated in 4.62 and see how they are implemented.

Rule 1 - The machine receives a pulse every 2w digits from the master clock. This pulse goes to the inhibitor circuit that is connected to the output of the adder. The inhibitor, therefore, examines every 2w'th digit of the adder output and develops a one if the adder digit is zero and develops a zero if the adder output is a one. Thus the output of the inhibitor can be taken as the successive digits of the square root of x. Each output of the inhibitor is stored temporarily in a one-digit storage cell until it can be written into DL4, which stores the complete answer.

The examining pulse that occurs every 2w digits also goes to the one-digit storage cell to act as the erase signal and remove C_{n-1} when C_n is stored.

Rule 2 - In the first 2w period, S1 is at position "a", and the addend for the adder is x transmitted through this position. Thereafter, S1 is at "b" and the addend is the adder output delayed by 2w digits in DL1.

Rule 3 - Coincidental with the transmission of the least significant digit of x into the machine, a start signal is received from the master clock. The start signal sets S2 to position "a". The start signal also progresses through DL5 and DL2 to arrive at an inhibitor circuit feeding S2 at time 2w-1. The inhibitor circuit has as output, therefore, the one's complement of 0100 --- 0. To convert this to a two's complement, the start signal is fed directly to the carry terminal of the adder.

Rule 4 - The operating instructions for switch S2 are obtained from the inhibitor in the adder output as well as from the start signal. S2 is of the type discussed in 3.1. Every time a new digit of the answer is determined, it is fed to the input terminal of the storage cell associated with the switch and sets it to "a" if the digit is a one or to "b" if the digit is a zero. Coincidental with the answer digit, the 2w pulse is fed to the erase terminal of the switch storage cell to remove the previously stored instruction, so that the switch can move to the new position. Setting switch S2 to "a" is equivalent to saying, "Follow Rule 5" and setting switch S2 to "b" is equivalent to saying "Follow Rule 6".

Rule 5 - When digit C_1 is determined it goes into the one-digit delay line DL7 and recirculates there until C_2 is determined. At time 2w+1 the start signal leaves DL6. It is inhibited, however, by the 2w pulse delayed one digit in DL3. The recirculated start signals from DL2 are all delayed two digits by DL6 and reappear at the output of DL6 once in each 2w period at the times listed below:

Period	Time	Digit entered in DL4
2	2w+1	none
3	4w-1	C_1
4	6w-3	C_2
5	8w-5	C_3
-	---	--

The output of DL6, in each case above except the first (when it is inhibited), goes to an and-circuit in the one-digit storage cell and lets a digit of the answer into DL4. The input of DL4 is also fed through an inhibitor to the "a" input of S2 and back to the adder to mechanize part "b" of Rule 5. Since C_1 first enters DL4 at 4w-1 it will re-enter it at 6w-2 just behind the first entry of C_2, which is as it should be to fulfill part "a" of Rule 5.

Part "c" of Rule 5 is fulfilled by the recirculated start pulse from DL2. It may be noted that DL6 provides the two-digit delay required between C_{n-1} and the "one" digit of part "c" of Rule 5.

The inhibitor at terminal "a" of S1 will provide the one's complement of q (see Rule 5). The two's complement is obtained by feeding the output of the inhibitor on the adder output back into the carry terminal of the adder. The one-digit delay line (DL9) is required to get the carry at 2w+1 rather than 2w.

Rule 6 - Parts "a" and "b" of this rule are fulfilled by the connection of the input of DL4 into terminal "b" of S2. Part "c" is fulfilled by the connection of the output of DL2 into terminal "b" of S2 both directly and through the one-digit delay line DL8.

In all discussions above it has been assumed that x is a 2w digit number. This may not be the case in some computers. To convert x into a 2w number it may be delayed w digits, which adds w zeroes to it. The delay may be obtained through a tap in DL1. Another expedient would be to take x as a w-digit number and send the first 2w pulse at time w, the second at 3w, etc.

The block diagram has been analyzed for its component requirements. It is believed that a square rooter to give 48 digit answers can be mechanized with 338 crystals, 312 digits of delay and 45 active elements. It may be noted that a square root machine is no more complex than a divider.

2.47 Summary of Components Required for Arithmetic Units

Two sets of estimates have been prepared. Table 3 lists the components required for units handling 15 binary digit numbers (4-1/2 decimal place). Table 4 lists the components for 48 (14 decimal place) digit numbers. It should be noted that an adder would be the same in either case.

TABLE 3

COMPONENTS REQUIRED FOR 15 DIGIT ARITHMETIC UNITS

UNIT	CRYSTAL DIODES	DIGITS OF DELAY	ACTIVE ELEMENTS
Adder	38	2	2
Accumulator	52	15	4
Multiplier	204	85	20
Divider	179	128	19
Square Rooter	182	98	19

TABLE 4

COMPONENTS REQUIRED FOR 48 DIGIT ARITHMETIC UNITS

UNIT	CRYSTAL DIODES	DIGITS OF DELAY	ACTIVE ELEMENTS
Accumulator	76	48	8
Multiplier	324	85	40
Divider	371	384	51
Square Rooter	338	312	45

2.5 MISCELLANEOUS UNITS

In addition to the units discussed above, certain others are useful in a computer. Some are discussed below.

2.51 Translator

This unit translates one set of numbers or words into another set. Storage may be regarded as translation in which addresses are translated into the words stored at the addresses.

Translators are useful in developing switching orders. At the completion of some process, it may be necessary to develop a new set of orders. When the orders must be developed in a certain sequence, a binary counter can be used to develop the input for the translator. Whenever the next set of orders are required, a pulse is fed to the counter. The binary count number is used as the switching instruction for a switch. Every time the counter receives an input pulse the switch will be advanced one position and will put a pulse on the selected lead. The various output leads of the switch will be connected to a group of or-circuits. One or-circuit will be required for each digit of the translator output. Figure 2-22 shows a scheme for generating 8 sets of 3 digit words. To generate n digit words n or-circuits would be required. No more than 8 leads would be required for any or-circuit. The translator shown develops the output word in parallel. If the or-circuits are fed to taps in a delay line, the output word will be received serially from the output of the delay line. The or-circuits have the same function as the coil maze in the Laboratories' Mod 6 computer.

2.52 Binary Counter

Figure 2-23 shows how each stage of a binary counter can be made from a one-digit memory cell (see paragraph 2.5) plus an inhibitor and an and-circuit. In Fig. 2-23, if the count is at 01 and a one is sent into the counter, the one in the first stage will be admitted to the second stage. The one that was in the first stage inhibits the one input and the first stage therefore goes to 0 and the count is 10. If another one is sent into the counter, it will enter the first stage and the count will advance to 11. If another one is fed in, the first two stages will set to zero and the third will receive a one making the count 100. It is seen that each stage is a dynamic flip-flop. This type of counter can be used only to count synchronized input pulses but that is all that a counter is expected to do in a serial computer.

In a serial machine binary counters can often be replaced by simpler devices. Suppose that it is desired to time some operation and know when 64 digit times have elapsed.

A six-stage counter could be used or a digit pulse could be circulated in a seven and a nine digit delay line. If an and-circuit were fed by the two delay lines, 63 digit times would elapse before the and-circuit were operated. If the output were fed through a one digit delay line the desired 64 digit delay would be obtained.

2.53 Shift Registers

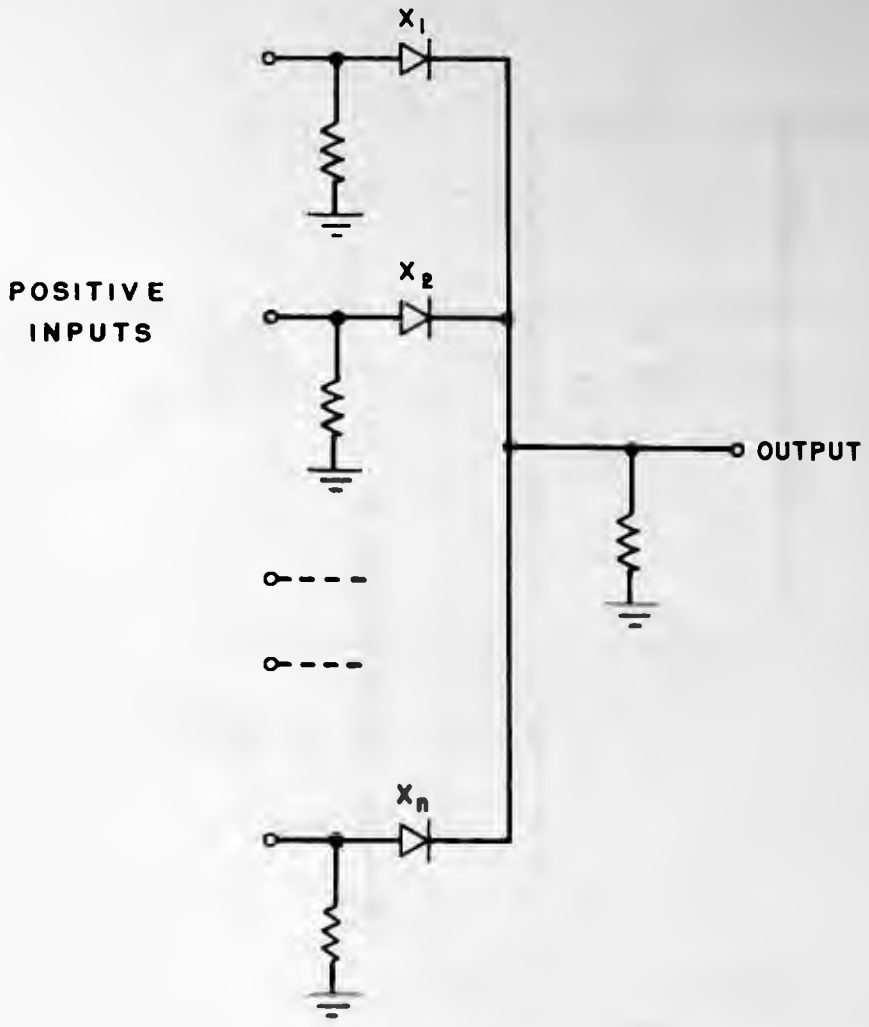
Shift registers are sometimes needed to take data at a low rate, from a magnetic drum for example, and re-transmit it at a rapid rate. A shift register that is loaded serially may be read out in parallel on w leads or in series on one lead. Each stage of such registers can be built around a one-digit memory cell. Fig. 2-24 shows two stages of each kind of register. The numbers in each cell are erased and advanced to the next cell whenever an advance pulse is received.

Advance pulses will be received every time a new digit is to be loaded. When a number is in the register, the advance pulses can be stopped and the number can be stored until it is required. In the case of the parallel-read-out registers, extra and-circuits are used at each stage. These are fed by the read-out signal.

Besides their use in changing the rate of transmission of data, the shift registers described above can store words for integral multiples of a digit time. Where the storage period is an integral multiple of a word time it is more efficient to use delay lines as discussed in paragraph 2.5. It may be mentioned that in a synchronous computer tapped delay lines can replace shift registers for the conversion of serial to parallel data and vice versa.

2.6 CONCLUSIONS

A catalog of computer units has been presented. The units make use of active elements only as repeating amplifiers. It is shown that multipliers, dividers, etc. can be built with less than 50 such amplifiers.



SYMBOL

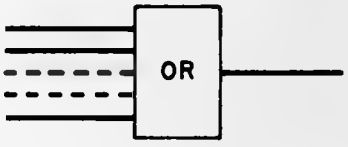


FIGURE 4-1
N-TERMINAL
OR-CIRCUIT

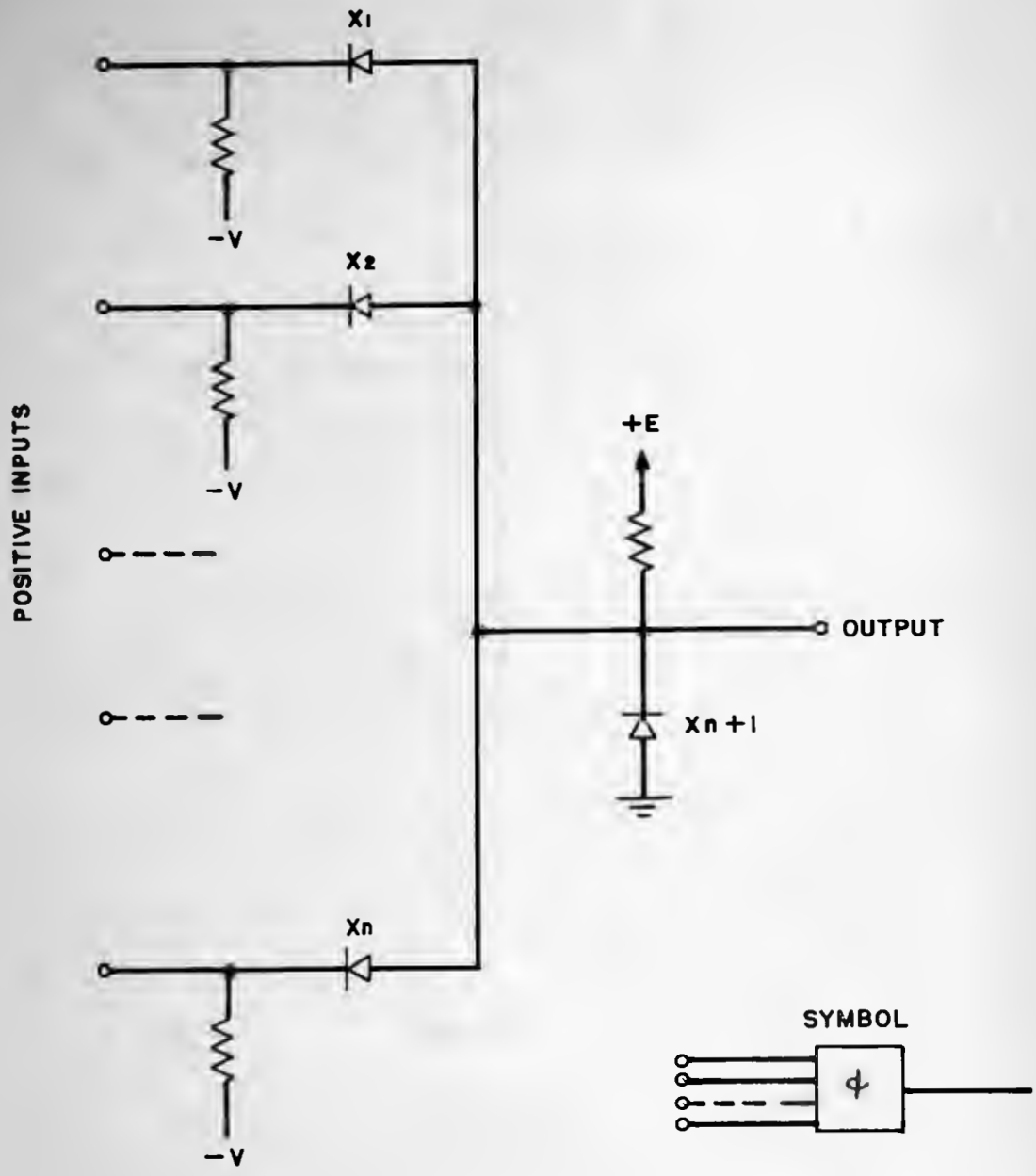
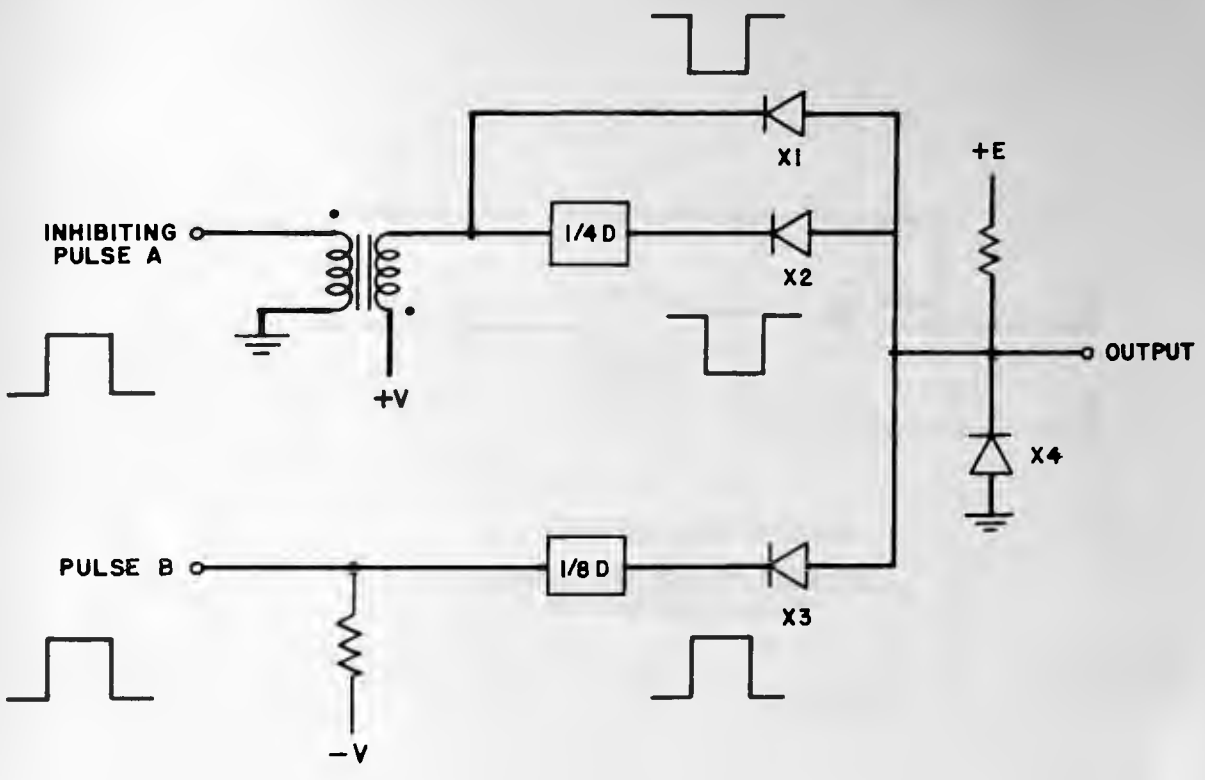


FIGURE 2-2
N-TERMINAL
AND-CIRCUIT



YD = DELAY OF Y- DIGITS

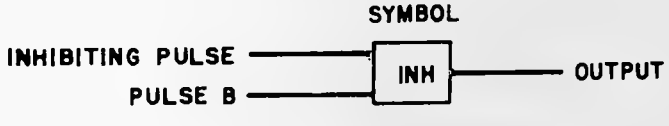
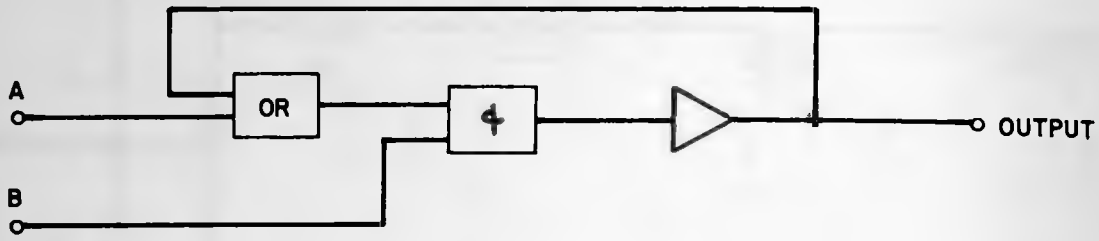


FIGURE 2-3
INHIBITOR
CIRCUIT



A = PULSE TO BE RETIMED

B = DIGIT PULSES – REFERENCE PULSES FROM MASTER CLOCK

FIGURE 2-4
USE OF ACTIVE ELEMENTS

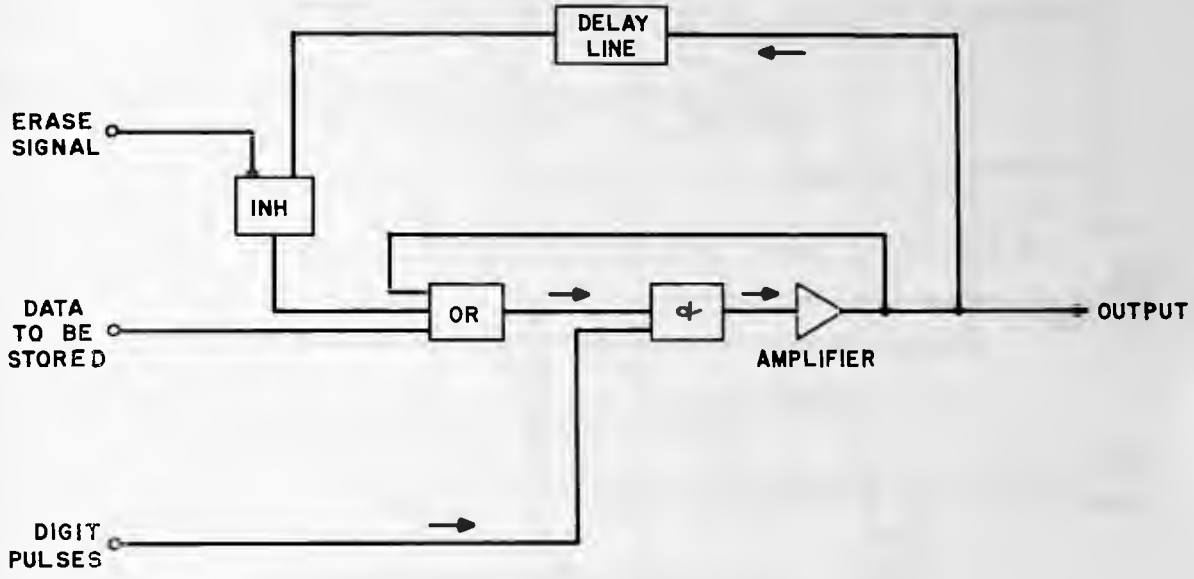
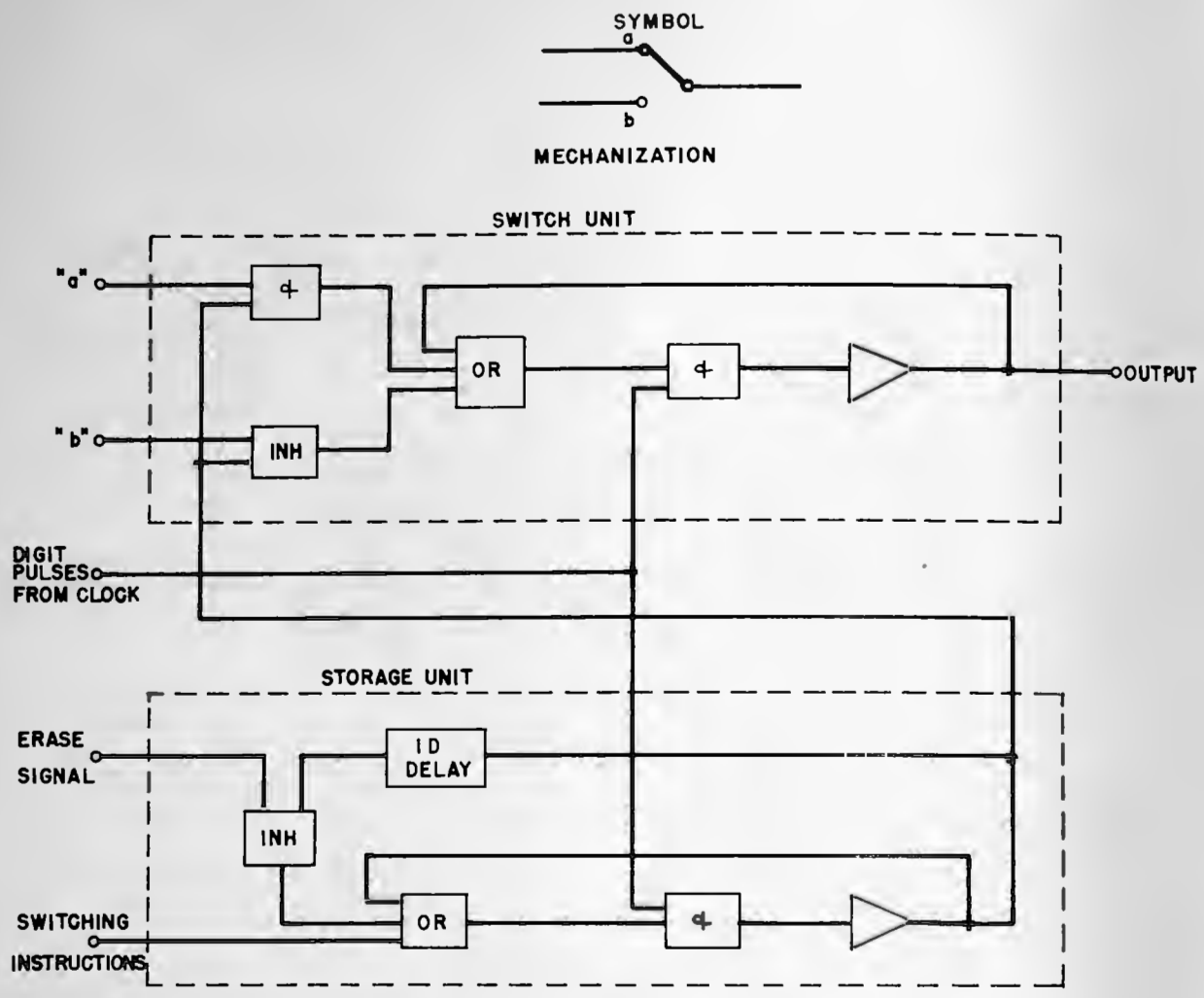


FIGURE 2-5
STORAGE CELL



0 SETS SWITCH TO "b"

1 SETS SWITCH TO "a"

FIGURE 2-6
SINGLE POLE
DOUBLE THROW
SWITCH

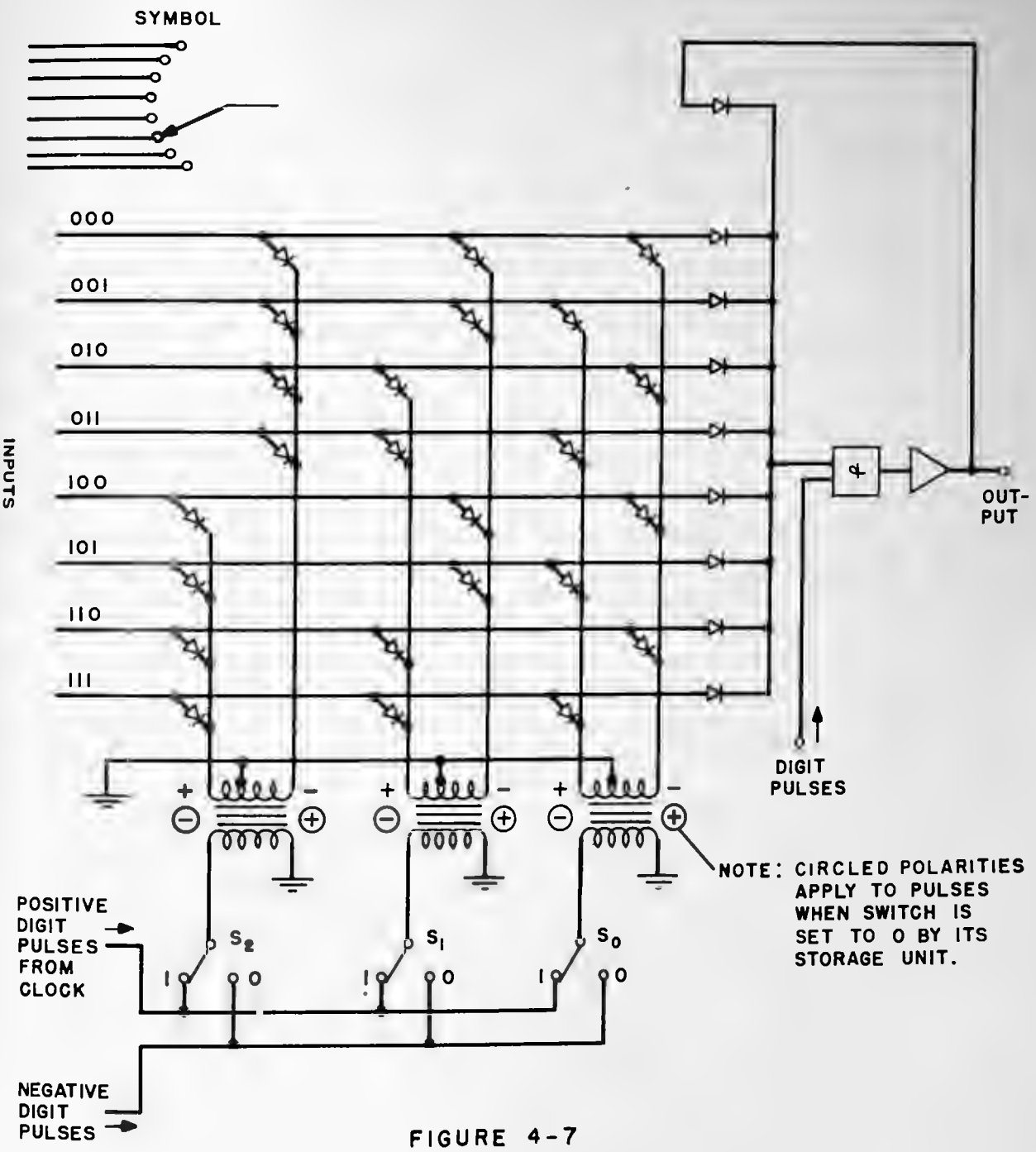


FIGURE 4-7
SINGLE-POLE
EIGHT-POSITION
SWITCH

NEGATIVE NUMBERS IN A 5 SIGNIFICANT DIGIT SYSTEM WITH ONE IN SIXTH PLACE SIGNIFYING THAT THE NUMBER IS NEGATIVE

(A) FORMATION OF NEGATIVE NUMBERS

```
+8 =              001000
ONES
COMPLIMENT =      110111
ADD ONE                 1
                  ------
∴ -8 =            111000

CHECK -8 +8=0
+8 =              001000
-8                111000
                  ------
SUM =           1 000000 = ZERO
TO THE MACHINE
```

(B) ADDITION

```
-8+5=-3
-8 =              111000
+5 =              000101
                  ------
SUM =             111101 = -3
CHECK +3 =        000011
                  ------
SUM =           1 000000 = ZERO

-8 +15 = 7
-8                111000
+15               001111
                  ------
SUM =           1 000111 = 7
```

(C) MULTIPLICATION

```
-8X7=-56
-8 =              1000
 7 =               111

                  1111000
                  0000111
                  -------
                  1111000
                 1111000
                11111000
               000000000
              0000000000
             00000000000
            000000000000
           0001001000
```

SHOWS ANS. IS NEG.

```
CHECK
-56+56=0
-56               1001000
+56                111000
                  -------
                 10000000
```

(D) ANOTHER MULTIPLICATION

```
-8 X -7 = +56
-8 =              1111000
-7 =              1111001
                  -------
                  1111000
                 0000000
                00000000
               111111000
              1111111000
             11110111000
            111100111000
           0111001111000
```

INDICATES POSITIVE NUMBER

FIGURE 2-8
HANDLING
NEGATIVE NUMBERS

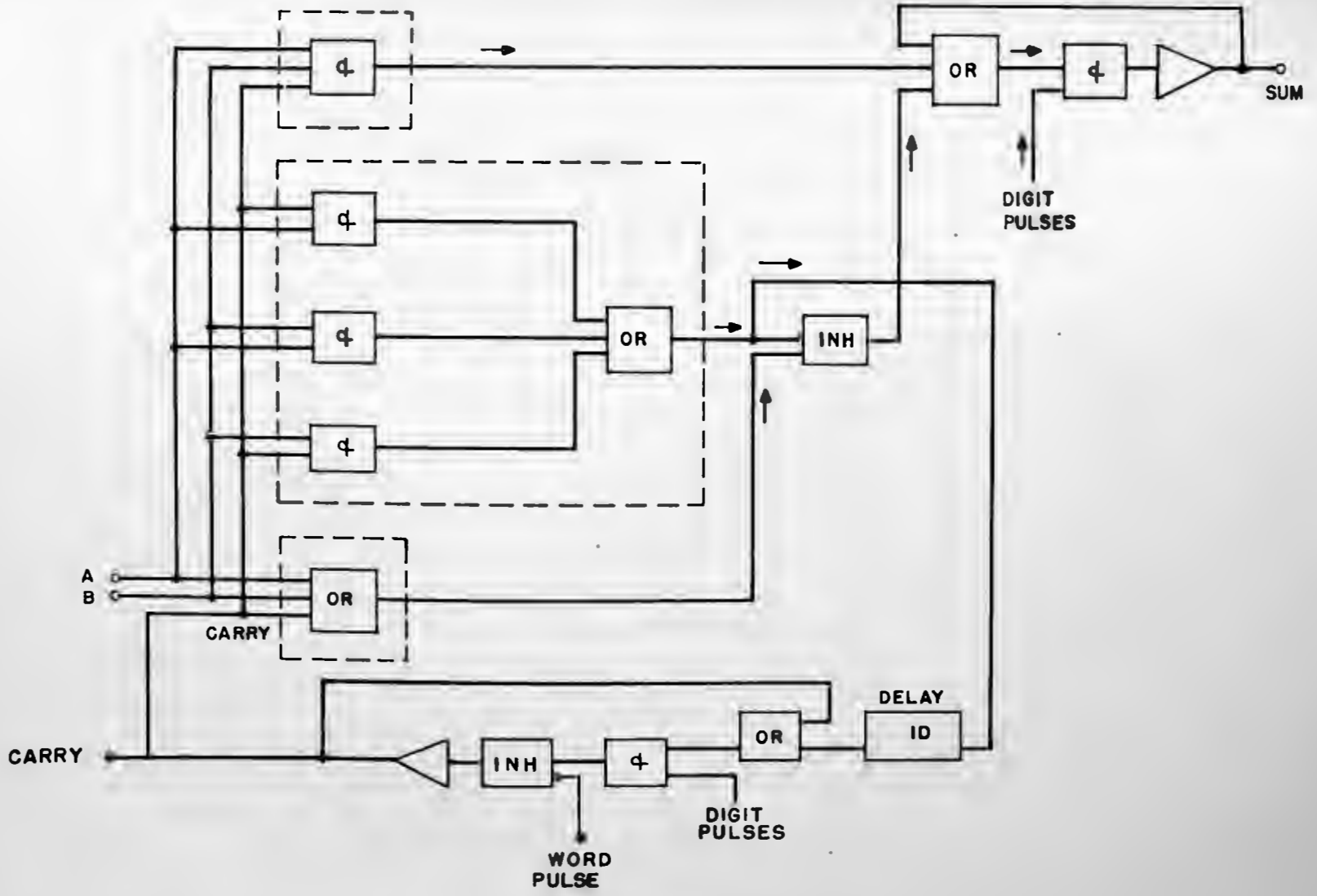


FIGURE 4-9
ADDER BLOCK
DIAGRAM

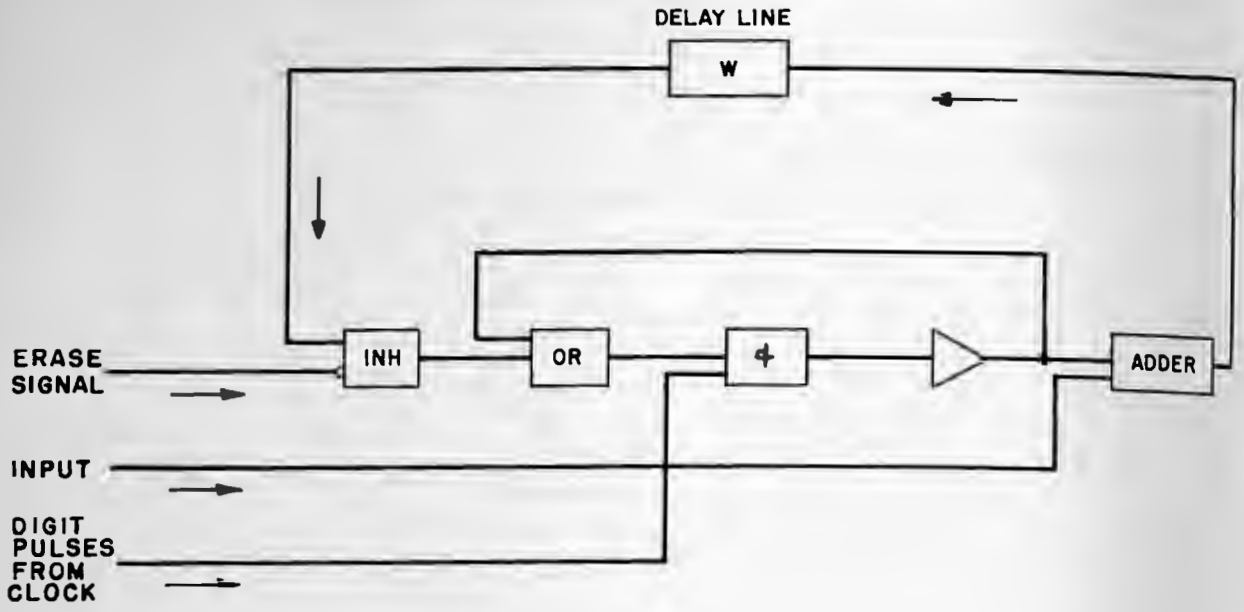


FIGURE 4-10

BLOCK DIAGRAM

OF

ACCUMULATOR

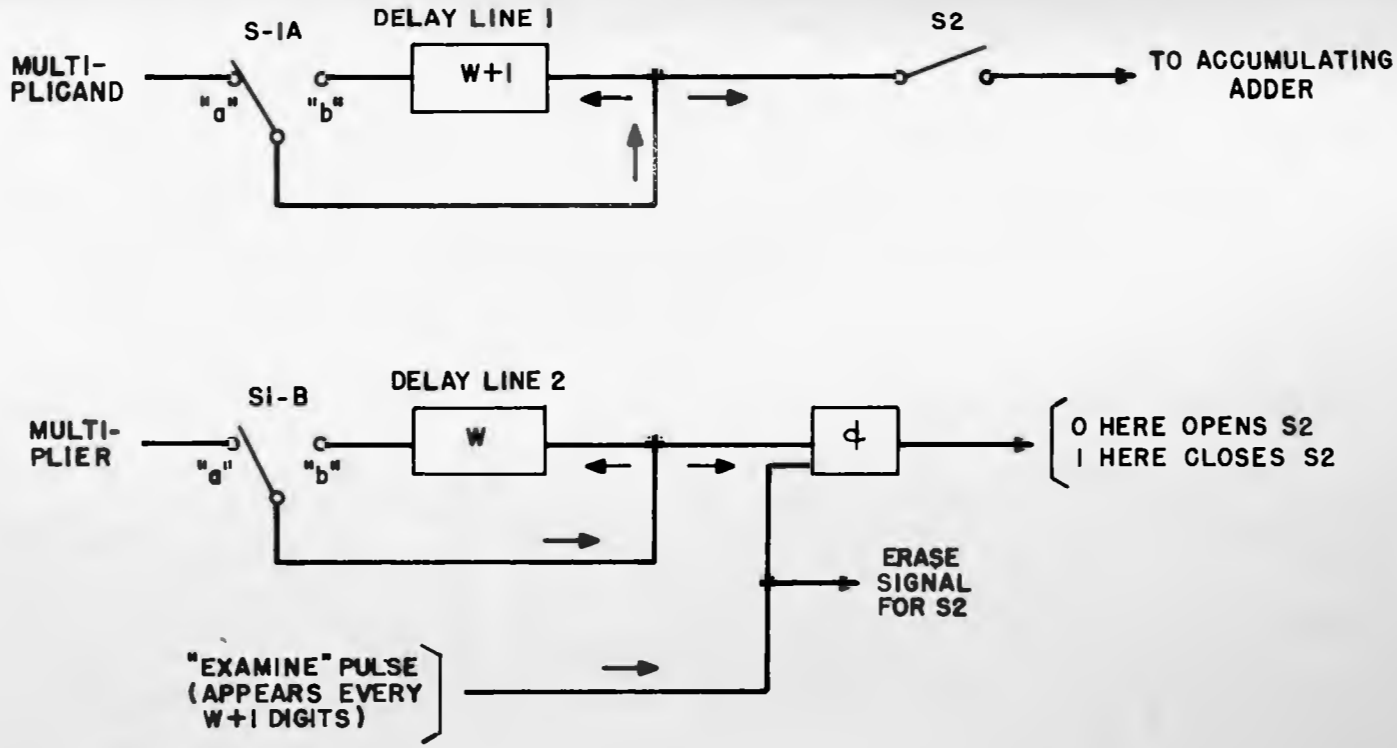


FIGURE 2-11

PARTIAL BLOCK

DIAGRAM OF MULTIPLIER

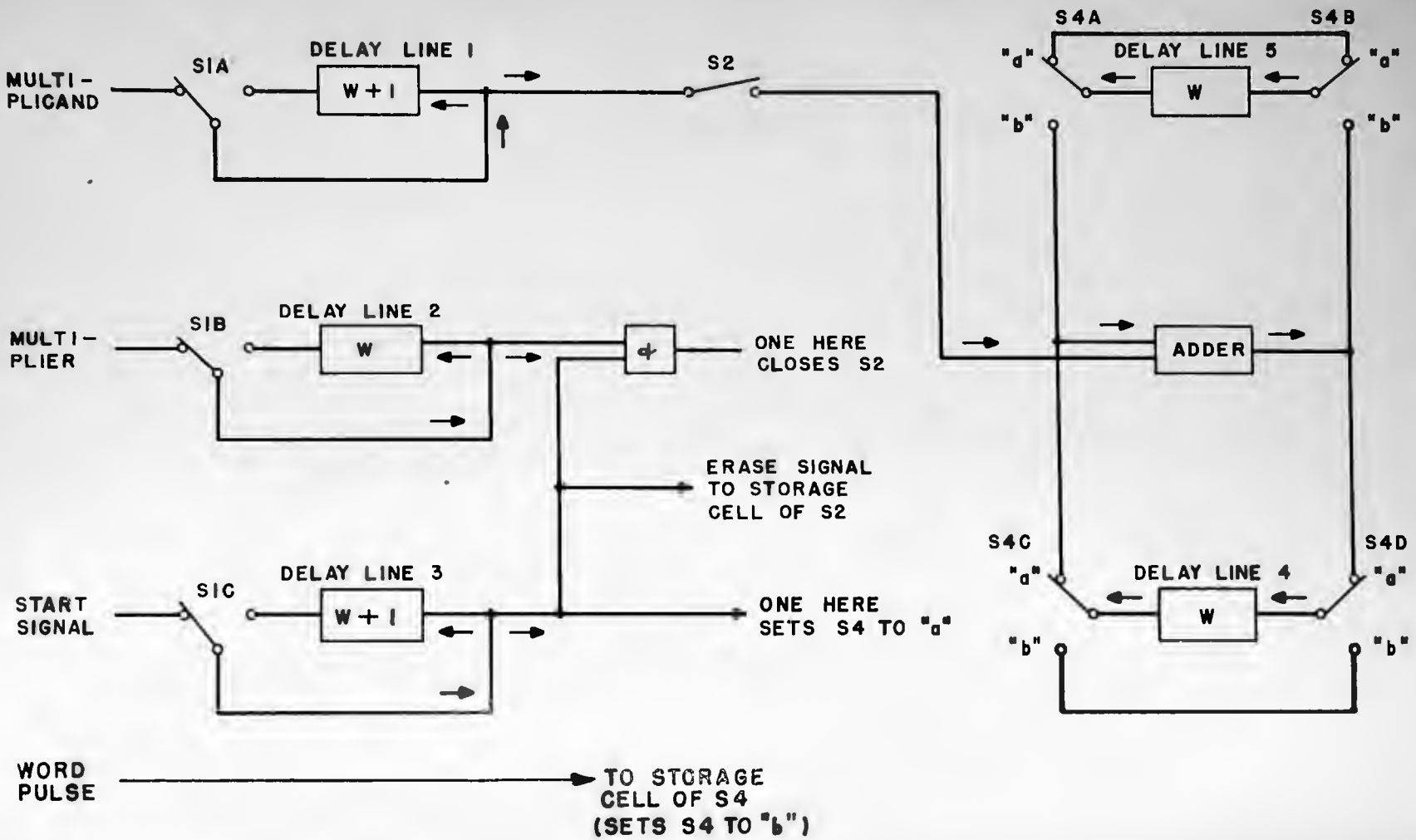


FIGURE 2-12
BLOCK DIAGRAM
OF
MULTIPLIER

MULTIPLICATION OF 1001 BY 1101 IN W=4 SYSTEM

20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	← TIME	
			1				1				1				1				1	WORD PULSES	
	1	0	0	1		1	0	0	1		1	0	0	1		1	0	0	1	MULTIPLICAND (DL1)	
1	1	0	1	1	1	0	1	1	1	0	1	1	1	0	1	1	1	0	1	MULTIPLIER (DL2)	
				1					1					1					1	W+1 PULSES	
← CLOSED →				← CLOSED →				← OPEN →							← CLOSED →					POSITION OFS2	
	1	0	0	1		1	0	0	1	0	0	0	0	0	0	1	0	0	1	FROM S_2	ADDER INPUT
				1				1	0			1	0	0		0	0	0	0	FROM DL4	
	0	1	0	0	0	0	0			0	0				0					FROM DL5	
				0				1	1			1	0	0		1	0	0	1	TO DL4	ADDER OUTPUT
0	1	1	1		0	1	0			0	0				0					TO DL5	
				0				1	1			1	0	0		1	0	0	1	FROM ADDER	INTO DL4
0	1	0	1		1	0	1			0	1				1					FROM DL4	
0	1	1	1		0	1	0			0	0				0					FROM ADDER	INTO DL5
				0				0	0			0	0	0		0	0	0	0	FROM DL5	
b	b	b	b	a	b	b	b	a	a	b	b	a	a	a	b	a	a	a	a	POSITION S_4	

ANSWER = $\underbrace{0111}_{DL5}\ \underbrace{0101}_{DL4}$

FIGURE 2 - 13
EXAMPLE OF
MULTIPLICATION

BINARY DIVISION

DIVIDE 3 (11) BY 8 (1000)

```
        0.011
1000 | 11.000
       10 00
       -----
        1 000
        1 000
        -----
```

DIVIDE 15 (1111) BY 24 (11000)

```
        0000.101
11000 | 1111.000
        1100 0
        ------
          11 000
          11 000
          ------
```

FIGURE 2-14
BINARY DIVISION

DIVIDE 3 BY 8

WORD LENGTH = 5 PLACES. 5TH PLACE INDICATES SIGN.

3 =	00011
8 =	01000
ONES COMPL. =	10111
ADD ONE	1
−8 =	11000

CHANGE NUMBERS TO 2W-1 OR 9 DIGIT LENGTH.

3 = 000000011

8 = 000001000

-8 = 111111000

DIVISION RULE: IF (2W-1)th (9th) DIGIT OF REMAINDER IS ZERO ADD - DIVISOR; ADD + DIVISOR IF DIGIT IS ONE.

		0.01100
000001000	000000011.00000	
ADD −8	111111000	
NOTE 9TH DIGIT → 1 AND ADD +8 WITH SHIFT	111111011	000001000
NOTE 9TH DIGIT → 1 AND ADD +8 WITH SHIFT	111111110	000001000
NOTE 9TH DIGIT → 0 AND ADD −8 WITH SHIFT	000001100	111111000
NOTE 9TH DIGIT → 0 AND ADD −8 WITH SHIFT	000000000	111111000
NOTE 9TH DIGIT → 1 AND ADD +8 WITH SHIFT	111111000	000001000
NOTE 9TH DIGIT → 1	111111000	

THE QUOTIENT IS THE ONES COMPLEMENT OF THE SUCCESSIVE 9TH DIGITS OF THE REMAINDERS

FIGURE 2-15

SIMPLIFIED

BINARY DIVISION

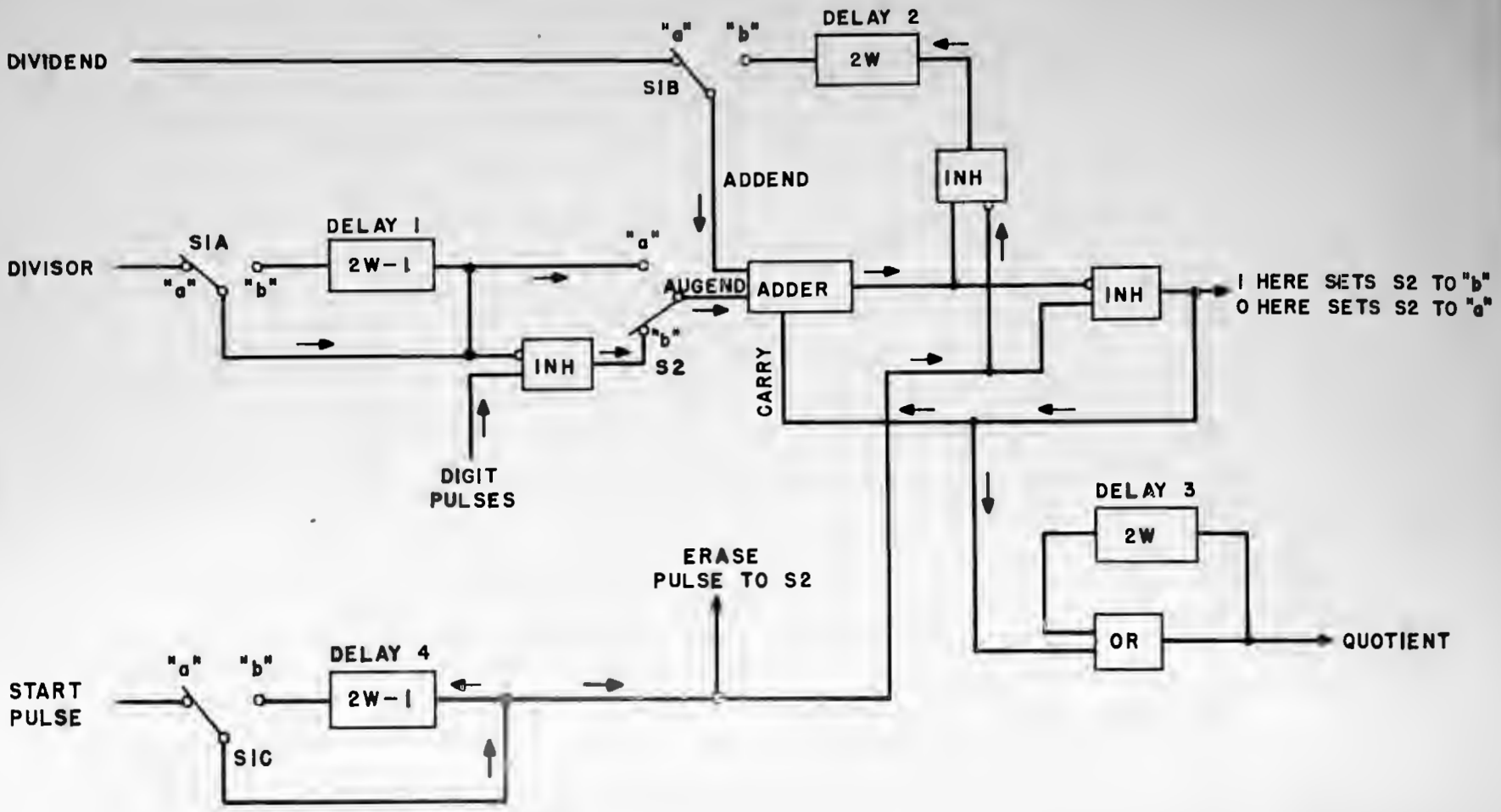
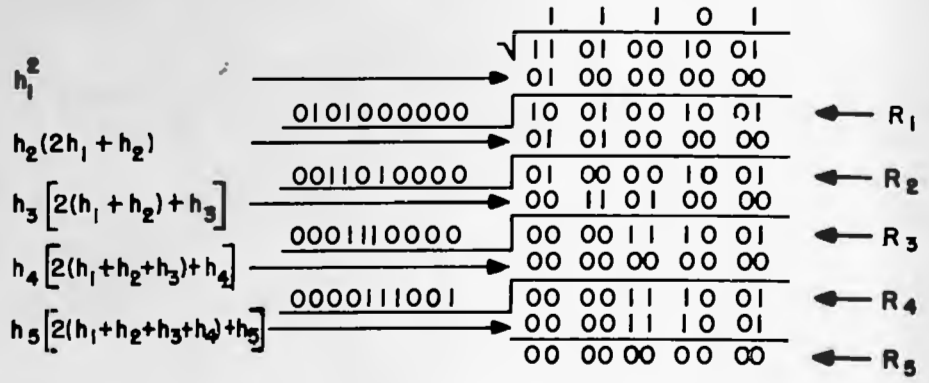


FIGURE 2-16
DIVIDER
BLOCK DIAGRAM

ALGEBRAIC PROCEDURE	ARITHMETIC PROCEDURE
FIND $\sqrt{x}$	FIND $\sqrt{123,904}$
STEP 1 MAKE FIRST ESTIMATE h_1 AND OBTAIN FIRST REMAINDER. $R_1 = x - h_1^2$	STEP 1 $h_1 = 300$ $123904 - 90000 = 33904 = R_1$
STEP 2 MAKE ESTIMATE OF INCREMENT h_2 SUCH THAT $h_2(2h_1 + h_2) \leq R_1$ AND COMPUTE SECOND REMAINDER. $R_2 = R_1 - h_2(2h_1 + h_2)$ [NOTE: $R_2 = x - (h_1 + h_2)^2$]	STEP 2 FIND h_2 SUCH THAT $h_2(600 + h_2) \leq 33904$ $50 \times 600 = 30,000$ $\therefore h_2 = 50$ $50(600+50) = 32,500$ $33904 - 32500 = 1404 = R_2$
STEP 3 FIND h_3 SUCH THAT $h_3\left[2(h_1 + h_2) + h_3\right] \leq R_2$ AND COMPUTE THIRD REMAINDER $R_3 = R_2 - h_3\left[2(h_1 + h_2) + h_3\right]$ [NOTE: $R_3 = x - (h_1 + h_2 + h_3)^2$]	STEP 3 FIND h_3 SUCH THAT $h_3(2 \times 350 + h_3) \leq 1404$ $2 \times 700 = 1400$ TRY $h_3 = 2$ $2(700 + 2) = 1404$ $1404 - 1404 = 0000 = R_3$ $\therefore (h_1 + h_2 + h_3)^2 = 123,904$ $\sqrt{123904} = 352$
STEP N $R_n = R_{n-1} - h_n\left[2\sum_{j=1}^{j=n-1} h_j + h_n\right]$ $\sqrt{x} \cong h_1 + h_2 + h_3 \cdots h_n$	

FIGURE 2-17
ALGEBRAIC
JUSTIFICATION OF
SQUARE ROOT
METHOD

FIND SQUARE ROOT 1101001001



h_1 = 10000 = 2^p

h_2 = 1000 = 2^{p-1}

h_3 = 100 = 2^{p-2}

h_4 = 00 =

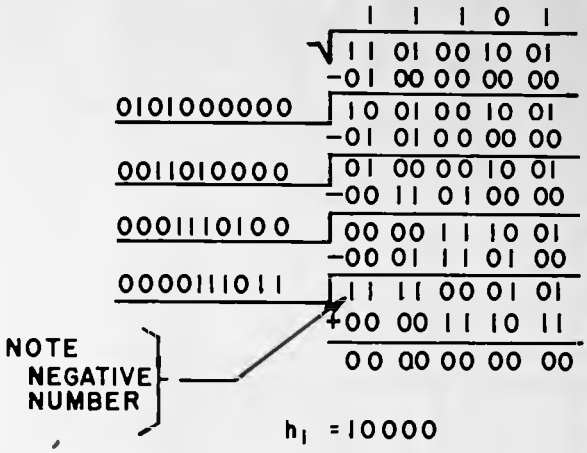
h_5 = 01 = $2^{p-4} = 2^0$

$\sqrt{x}$ = 11101

FIGURE 2-18
BINARY SQUARE
ROOTING

EXAMPLE ONE

FIND SQUARE ROOT 1101001001



h_1 = 10000

h_2 = 1000

h_3 = 100

h_4 = 00

h_5 = 01

11101 = RESULT

EXAMPLE TWO

FIND SQUARE ROOT 1001110001

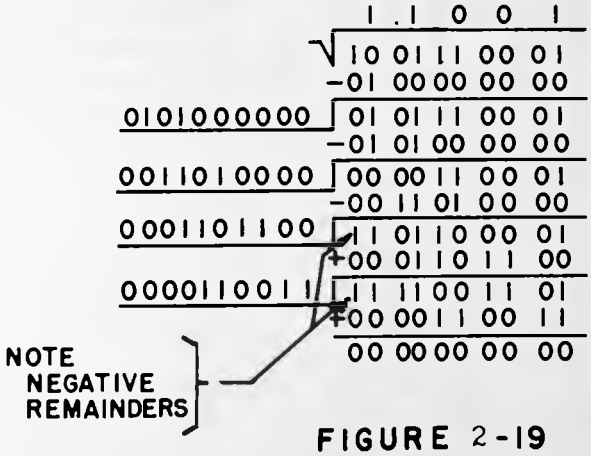


FIGURE 2-19
SIMPLIFIED BINARY
SQUARE ROOTING

FIND SQUARE ROOT OF 100111001

30 29 28 27 26 25 24 23 22 21	20 19 18 17 16 15 14 13 12 11	10 9 8 7 6 5 4 3 2 1	DIGIT TIME
0 0 0 0 1 1 0 0 0 1	0 1 0 1 1 1 0 0 0 1	1 0 0 1 1 1 0 0 0 1	MINUEND OR ADDEND
0 0 1 1 0 1 0 0 0 0	0 1 0 1 0 0 0 0 0 0	0 1 0 0 0 0 0 0 0 0	SUBTRAHEND OR AUGEND
1 1 0 1 1 0 0 0 0 1	0 0 0 0 1 1 0 0 0 1	0 1 0 1 1 1 0 0 0 1	REMAINDER OR SUM
0 C_3	1 C_2	1 C_1	ANSWER DIGIT
C_1 C_2	C_1		REAPPEARANCE OF ANSWER DIGITS

50 49 48 47 46 45 44 43 42 41	40 39 38 37 36 35 34 33 32 31	DIGIT TIME
1 1 1 1 0 0 1 1 0 1	1 1 0 1 1 0 0 0 0 1	MINUEND OR ADDEND
0 0 0 0 1 1 0 0 1 1	0 0 0 1 1 0 1 1 0 0	SUBTRAHEND OR AUGEND
0 0 0 0 0 0 0 0 0 0	1 1 1 1 0 0 1 1 0 1	REMAINDER OR SUM
1 C_5	0 C_4	ANSWER DIGIT
C_1 C_2 C_3 C_4	C_1 C_2 C_3	REAPPEARANCE OF ANSWER DIGITS

FIGURE 2-20
DETAILS OF
SQUARE ROOT PROCESS

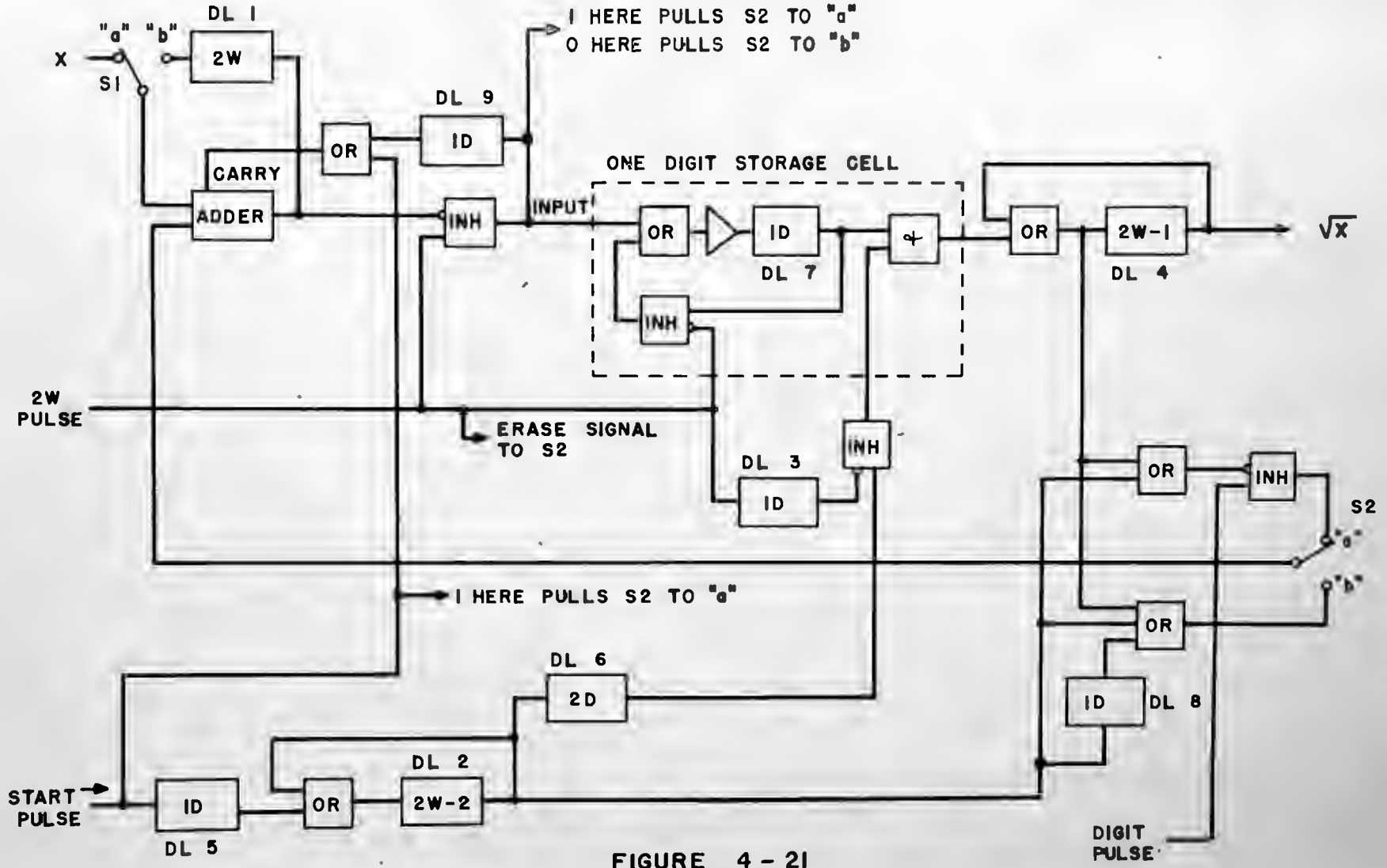
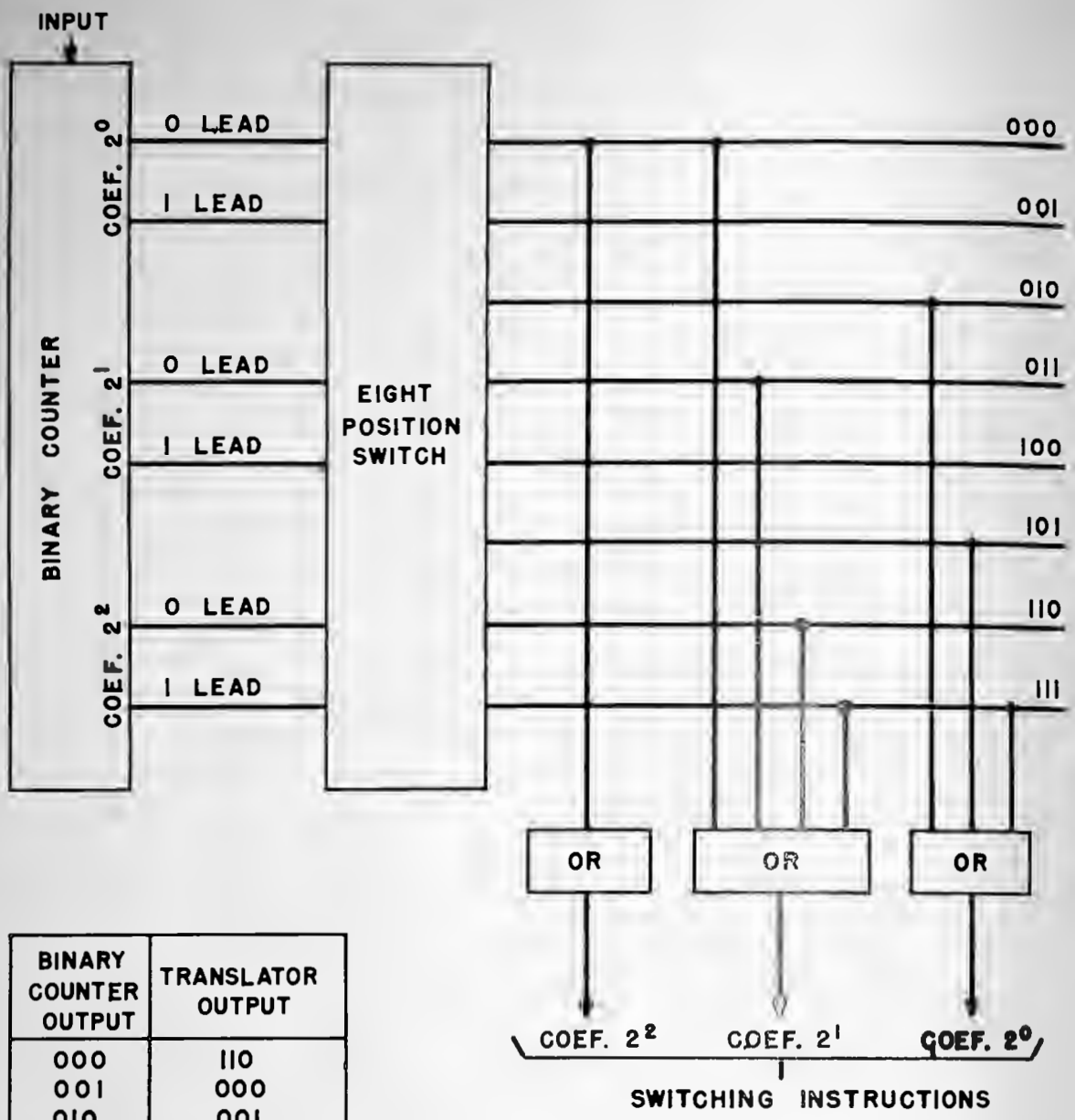


FIGURE 4 - 21
SQUARE ROOTER
BLOCK DIAGRAM



BINARY COUNTER OUTPUT	TRANSLATOR OUTPUT
000	110
001	000
010	001
011	010
100	000
101	001
111	011

FIGURE 2-22
TRANSLATOR

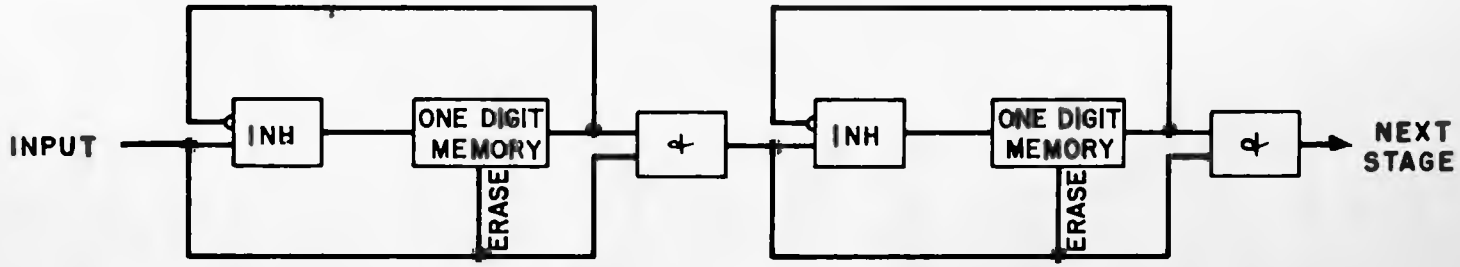


FIGURE 2-23
BINARY
COUNTER

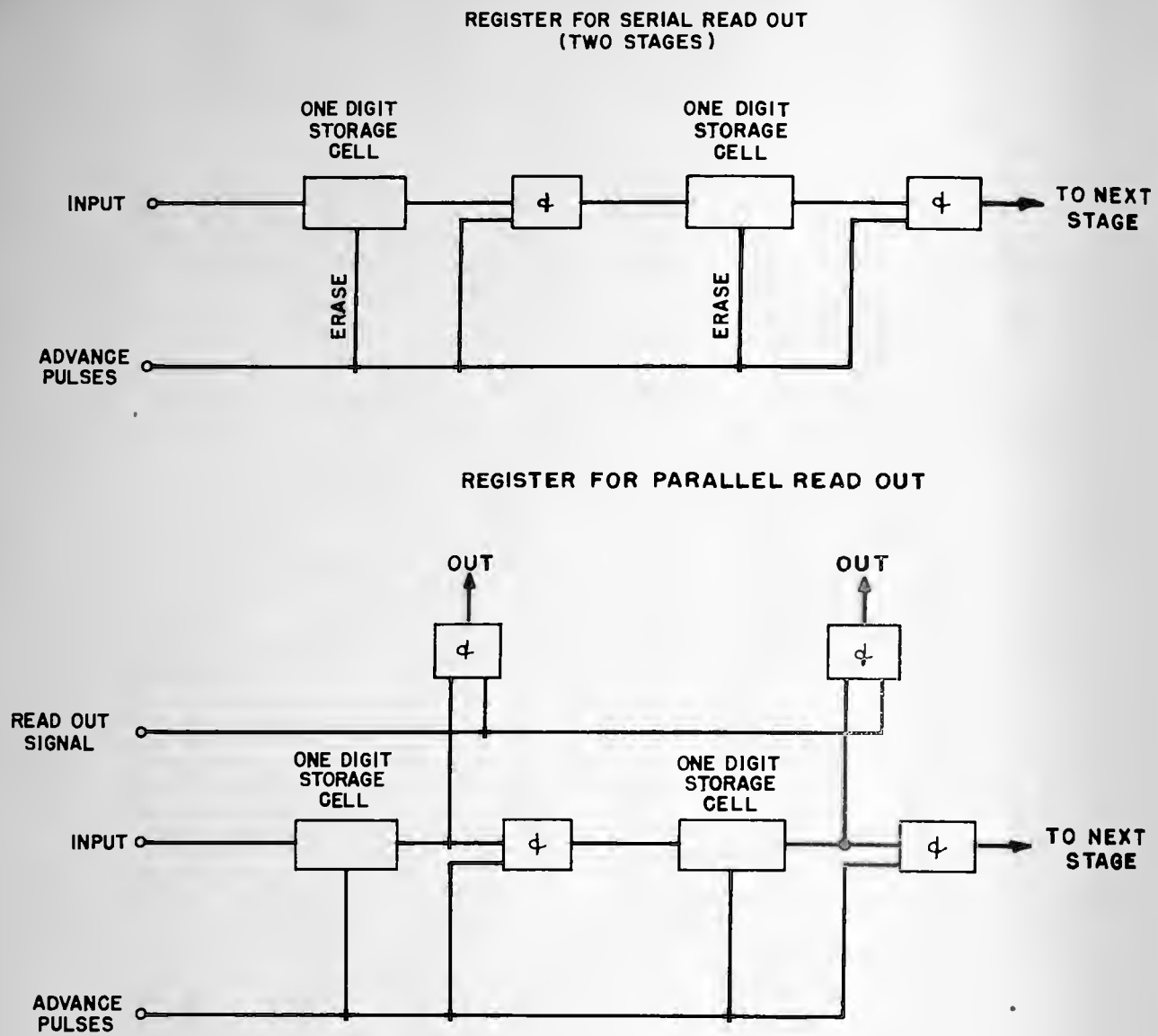


FIGURE 2-24
SHIFT REGISTER

A TRANSISTOR BLOCKING OSCILLATOR

3.1 INTRODUCTION

Because the transistor has a much lower input than output impedance, it appeared that a transistor pulse-generator could be designed in which the output is coupled back to the input through an impedance changing transformer. Experiments along these lines show that it is possible to build a transistor blocking oscillator that will work with almost any transistor. Rectangular pulses of less than a microsecond duration are easily generated, and rise times as short as 0.01 microsecond have been observed in 20-volt pulses. Pulse repetition frequencies from the low audio range up to two megacycles have been obtained. Analysis of the circuit reveals that the negative resistance required for oscillation can be obtained even though transistor current gain is less than one.

3.2 BASIC CIRCUITS

The basic circuit is shown in Figure 1A where the transformer couples the inverted collector pulse back to the base at an impedance level comparable to the base impedance. A variation of the circuit is shown in 1B, where the collector voltage is fed back in phase to the emitter. Both circuits oscillate, but most of the work has been done on circuit 1A.

When the oscillator is to be free-running, R (Fig. 1) is returned to a positive voltage in such a way that a constant current of about one milliampere is supplied to the emitter. If the circuit is to operate only when triggered, R is generally returned to a negative voltage.

3.3 CIRCUIT OPERATION

The circuit operates much like the vacuum tube equivalent. During the "ON" period, the collector voltage is nearly at ground, the base is held negative by the pulse transformer, and the emitter condenser (Fig. 2) is charged negative by the emitter current to a voltage nearly equal to the base voltage. During the "OFF" period, the collector voltage is negative, the base is at ground potential, and the charge left on C holds the emitter at a negative potential.

For purposes of a more detailed discussion, assume that the emitter condenser has been charged negatively and is discharging through R towards a positive voltage. Until the condenser reaches ground potential, the transistor is cut off. When the condenser reaches ground potential, emitter current begins to flow, releasing holes to the collector. The collector current pulls the collector potential up, and transmission through the inverting transformer causes the base to fall in potential which increases the emitter current. This regenerative transition from no emitter current to high current may occur in from one-hundredth of a microsecond to several tenths of a microsecond, depending on the particular transistor. Transistors of the type currently supplied for experimental use have rise times of less than 0.1 microsecond.

During the transition the emitter current (see Fig. 2) is charging C negatively, but as C charges the emitter current descreases because the emitter becomes less positive with respect to the base. At the same time that the emitter current is falling, the collector current required in the transformer to maintain the negative pulse at the base increases because of the low frequency cut-off (or time constant) of the transformer. When the emitter current has fallen to where it no longer releases the holes necessary to support the demanded collector current, the base voltage will rise towards ground which will cause a further decrease in emitter current, and the transistor will regeneratively cut itself off. C is left at a negative potential and begins again to discharge through R towards ground.

The above discussion shows that the recurrence rate is determined by C and the current supplied through R. The pulse duration is determined primarily by the transformer and C, and only in part by the transistor.

In appendix I, an analysis of the circuit is presented on the basis of linear circuit theory.

3.4 EXPERIMENTAL RESULTS

Figure 3 shows an experimental circuit built to develop a trigger pulse in advance of a main pulse, so that the main pulse could be observed on an oscilloscope. Transistor 1 was connected as a free-running blocking oscillator and its output differentiated and delayed 0.15 microseconds to trigger transistor 2.

Using transistor AN2891, the free-running circuit had a pulse duration of 1.5 microseconds, a rise time of 0.15 microseconds, a collector-pulse-amplitude of 30 volts, and developed a 10-volt signal across 560 ohms in the tertiary of the transformer. The 400A crystal was put across the collector winding of the transformer to clip the negative tail that occurs at the end of the pulse when the transformer is not fully loaded down.

The second circuit was tried with 41 transistors and worked with 40, while it was unstable with one. Five of these transistors were of miscellaneous previous history, while thirty-six were loaned from the spares of a matrix switch. Of the latter group of 36, 9 had rise times of 0.03 microseconds or less. The average duration of the pulse was 0.8 microseconds, and the standard deviation was less than 0.1 microseconds.

Photographs were made of the extremely steep wave forms that can be generated. Figure 4 shows the circuit used, and Figure 4A shows the collector output pulse. Figure 4B shows the collector output when C is returned to the collector, rather than ground. This circuit variation has been found to flatten the collector pulse but lengthen the fall time.

In these photographs, one small division equals 0.01 microseconds. Thus the transistor is seen to be capable of generating pulse energy over an extremely wide video band. It may be that in spite of its restricted bandwidth as a linear amplifier, the low inductance of its leads and their very small shunt capacitances may make it a better generator of extremely rapid pulses than the vacuum tube.

3.5 DESIGN OF CIRCUITS

An advantage of the circuit is that its performance is predictable in advance. The selection of the components to obtain the desired results are discussed below:

a. Transformer - For pulses of from one-half to two microseconds duration, two types of transformer cores have been tested. One is a rectangular core of 4-79 Molybdenum Permalloy having a cross-sectional area of 1/16 square inch and overall dimensions of 1 x 1-1/8 inches. The

second core tested is a 3/4 inch ferrite ring with a cross-sectional area of 9/256 square inches. The writer believes that the 0.01 microsecond rise time observed with some transistors might have been even less if better transformers were available.

Some of the factors influencing the selection of the turns ratio are discussed in the appendix. A turns ratios of approximately five to one has been found to give good results with all transistors. Lower ratios do not work with some units.

b. Pulse Duration - The pulse duration, as discussed in paragraph 3, is determined by the transformer low-frequency response, the value of the emitter condenser and in part by the transistor. A 430 mmf condenser and a transformer with three turns in the base winding and 16 in the collector (using the 4-79 Molybdenum Permalloy core previously mentioned) gives a 1/3 microsecond pulse. Using the same core with nine turns in the base and 50 turns in the collector winding, and a 2200 mmf condenser, a duration of 1-1/2 microseconds was obtained.

It should be noted that C can be made so large that the emitter current required to charge it to the base voltage may burn out the transistor.

c. Recurrence Rate - For predictable results, the emitter resistor R should be returned to a large positive voltage (E_e), so that a constant current is supplied to the emitter. If the emitter swings during the pulse in ΔE_e, the recurrence rate will be

$$f_r = \frac{1}{\frac{\Delta E_e}{E_e} RC + \tau}$$

where τ is the pulse duration. This equation has been found to give results as good as could be checked on an oscilloscope. The emitter swing is equal to the base swing and is given approximately by

$$\frac{\Delta E_c}{N}$$

where ΔE_c is the collector swing, and N is the transformer turns ratio. If the collector winding is returned to a battery of E_c volts, and N is three or more, the swing in collector voltage will almost equal the battery voltage and will be very stable. An approximation to the recurrence rate is, therefore, given by

$$f_r \approx \frac{1}{\frac{E_c}{E_e}\frac{RC}{N} + \tau}$$

Where stability of frequency is desired, it is important to supply a constant E_c, rather than obtain it from a series resistor. Six switching transistors were tried in a circuit designed, according to the above equation, to have a period of 14.5 microseconds. The actual periods were measured to be 14, 12, 12.5, 12.5, 12.5 and 14 microseconds.

d. Triggered Oscillator - If the emitter is returned to a negative voltage, a pulse will be obtained only when the circuit is triggered (usually through an auxiliary winding of the transformer). When rapid triggering rates were desired, the emitter resistor R was made 1200 ohms and returned to -1.5 volts. When R and the emitter condenser are both made small, the maximum triggering rate as well as the pulse duration will be influenced by the low frequency cut-off of the transformer.

3.6 CONCLUSIONS

A new transistor circuit has been described whose characteristics are stable and almost independent of the transistor characteristics. Because of its stable and circuit-controlled recurrence frequency, it should be useful as a frequency dividing device. Since it can be operated at a megacycle rate, it should be useful wherever fast powerful pulses are required. The writer is chiefly interested in the device as a regenerative amplifier for digital computer applications.

APPENDIX

ANALYSIS OF TRANSISTOR BLOCKING OSCILLATOR

Analysis of the impedance (R_i) shown at the emitter of the blocking oscillator shows a negative resistance of the series type and the magnitude of this negative resistance may be taken as some evidence of the performance of the circuit as an oscillator. Referring to Figure 5, we may write the following equations (assuming that the transformer is ideal and inverting with a turns ratio n).

In Mesh I

$$i_e(R_e + R_b) - i_c R_b = E - e_b \tag{1}$$

and in Mesh II

$$- i_e(R_b + R_m) + i_c (R_b + R_c) = e_b (n + 1) \tag{2}$$

Since the transformer is assumed to be ideal, we may write

$$i_c = \frac{i_e - i_c}{n} \text{ or } i_c = \frac{i_e}{n+1} \tag{3}$$

Substituting (3) in (1) and (2) gives

$$i_e \left(R_e + \frac{n}{n+1} R_b\right) + e_b = E \tag{4}$$

and

$$- i_e \left(R_m - \frac{R_c}{n+1} + \frac{R_b n}{n+1}\right) - e_b (n + 1) = 0 \tag{5}$$

which can be solved for

$$\frac{E}{i_e} = R_i = - \frac{R_m}{n+1} + \frac{R_c}{(n+1)^2} + R_b \left(\frac{n}{n+1}\right)^2 + R_e \tag{6}$$

If one tries to maximize R_i by varying n, it is found that

$$n_{opt} = \frac{2(R_c - R_b)}{(R_m - 2R_b)} - 1 \qquad (7)$$

or

$$n_{opt} \approx \frac{2R_c}{R_m} - 1$$

This shows that if R_m/R_c is large, the transformer should not invert in order to maximize R_i.

If $n = \frac{2R_c}{R_m} - 1$ is substituted in (6), we obtain

$$R_i = -\frac{R_m}{4}\,\frac{R_m}{R_c} + R_e + R_b\left(1 - \frac{R_m}{2R_c}\right)^2 \qquad (8)$$

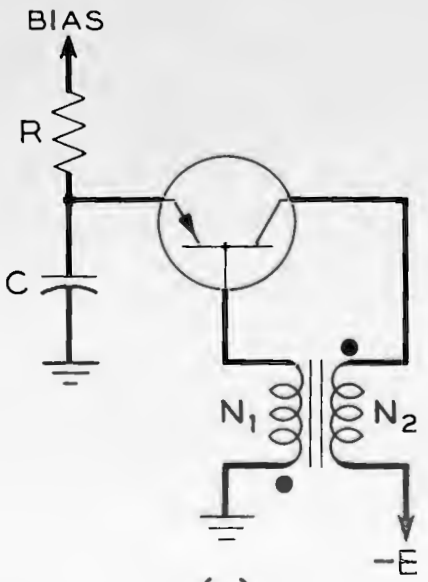
which shows that a negative resistance can be obtained when R_m is less than R_c.

When choosing the transformer, one has to consider the values of R_m and R_c over the entire range of collector current. A negative resistance is desired over as large a portion of the collector current curve as possible.* At the boundaries of the collector current range, R_m will be less than R_c, and this suggests that one ought to make the transformer inverting and with a turns ratio of three or so to obtain regeneration even when $R_m = 1/2\ R_c$. Experience to date has shown that any lower turns ratio is not desirable, but no upper limit has been investigated chiefly because the performance of the circuit is not very sensitive with regard to n, except in the respects mentioned in paragraph 5.

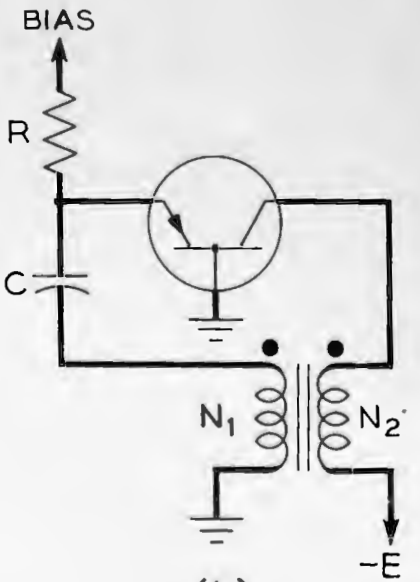
- -

* In order to get the maximum possible collector current swing and minimum voltage rise time.

TRANSISTOR BLOCKING OSCILLATOR CIRCUITS



(a)



(b)

TYPICAL WAVE FORMS IN BLOCKING OSCILLATOR

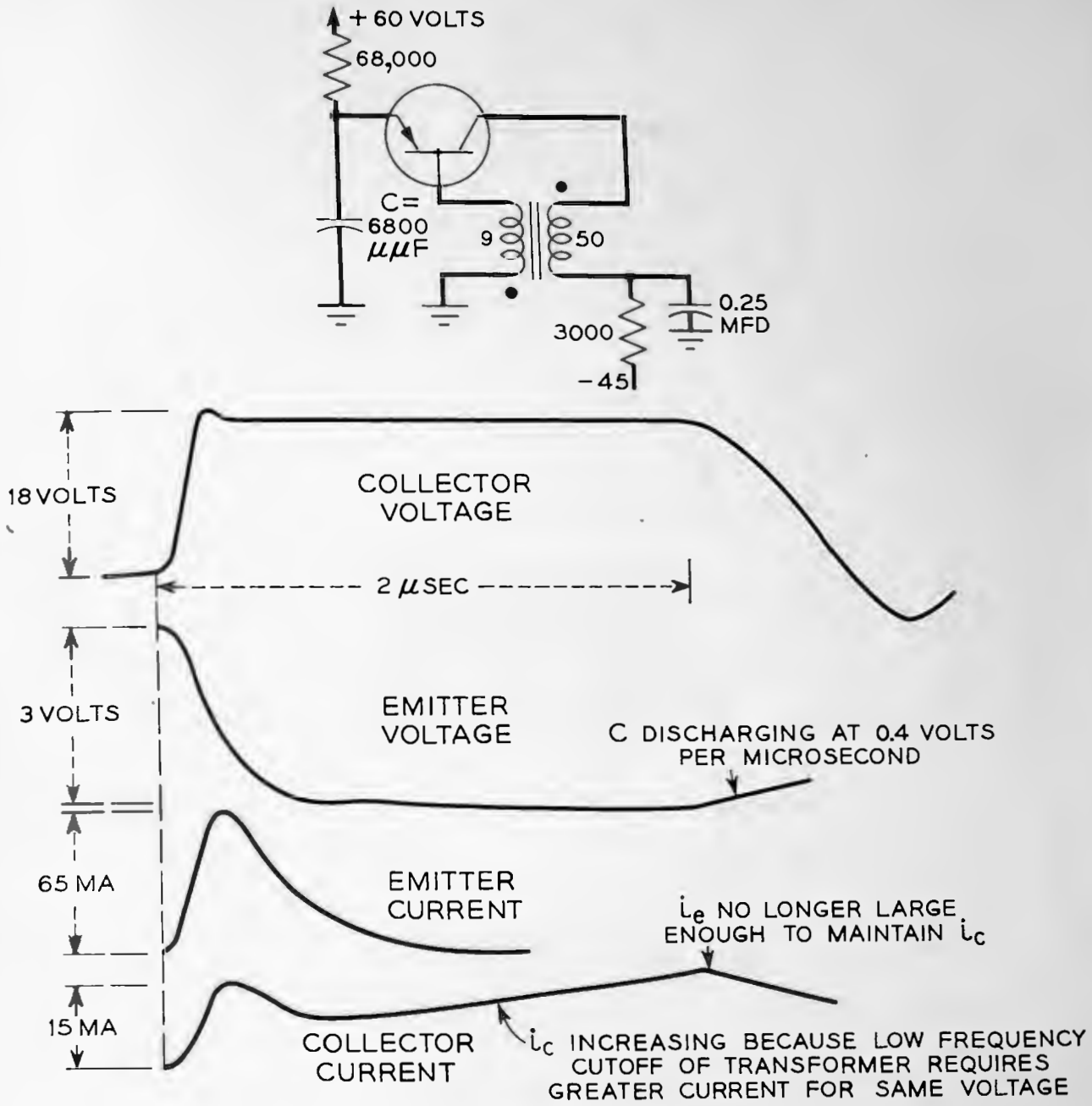


FIG. 2

TEST CIRCUIT FOR OBSERVING MAIN OSCILLATOR PULSE

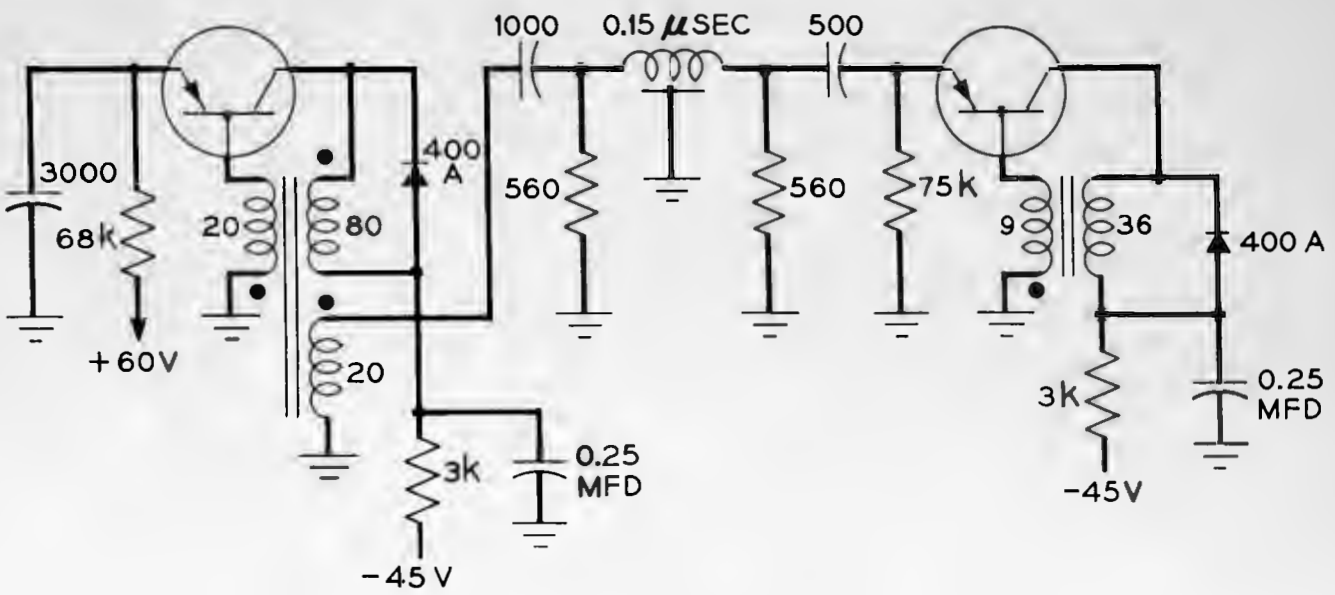
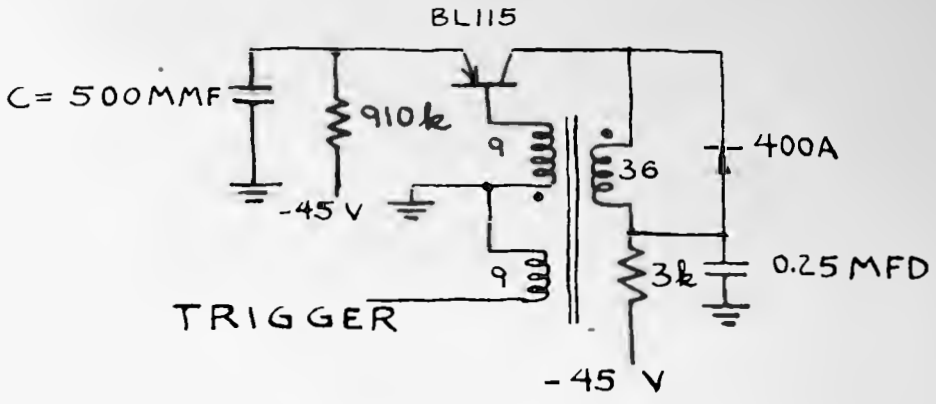
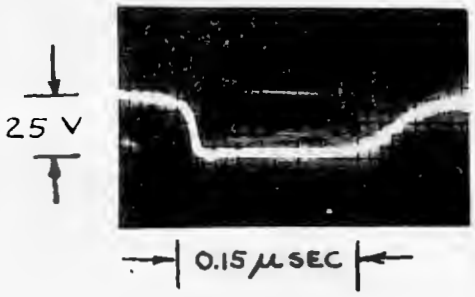


FIG. 3

STEEP WAVE FRONTS IN TRANSISTOR BLOCKING OSCILLATOR



COLLECTOR WAVEFORMS
(INVERTED)



A

B

C RETURNED TO COLLECTOR

FIGURE 4

EQUIVALENT CIRCUIT OF TRANSISTOR
BLOCKING OSCILLATOR

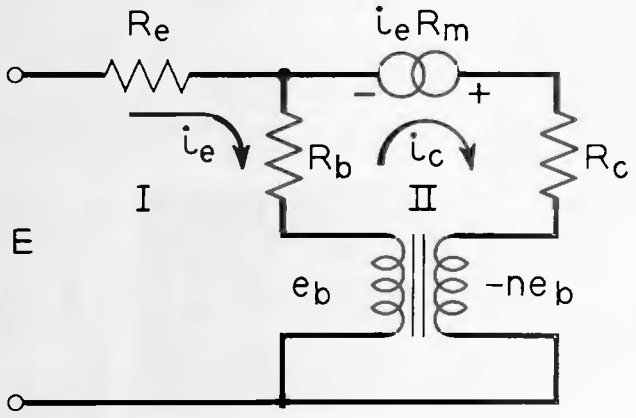


FIG. 5

COUNTING DOWN WITH BLOCKING OSCILLATORS

4.1 INTRODUCTION

In serial computers a word pulse is required to mark the beginning of each new word or number. For this project 16 digit words were desired at a digit rate of one megacycle. The word-pulse generator therefore is a frequency divider having a one megacycle input and a 62.5 kc output. A chain of four blocking oscillators has been used for the frequency division.

4.2 SCHEMATIC

The schematic of the word-pulse generator is shown in figure 4-1 and in figure 4-2 the emitter and collector wave forms are shown. The first stage was set to free run at about 350 kc. The saw-tooth at the emitter caused by the discharge of C1 through R3 has superimposed upon it the one megacycle clock signal. This signal synchronizes the firing of the first stage at 500 kc as can be seen from figure 4-2. A portion of the output of the first stage is coupled to the second stage and synchronizes it at 250 kc. The third stage is similarly synchronized at 125 kc and the fourth at 62.5 kc.

The output of the fourth stage is used to synchronize test oscilloscopes. It also drives an output amplifier of the type described in paragraph 5. This amplifier standardizes the duration of the word pulse at 1/2 microsecond.

Front and back views of the unit are shown on photograph 192402. One of the coaxial cables shown carries the clock signal. A second cable carries a synchronizing signal to test oscilloscopes and a third cable carries the word pulse to the computer.

4.3 SUMMARY

The word pulse generator described herein has been operated since January of 1951 without transistor replacements. It has a total battery drain of only 80 milliwatts.

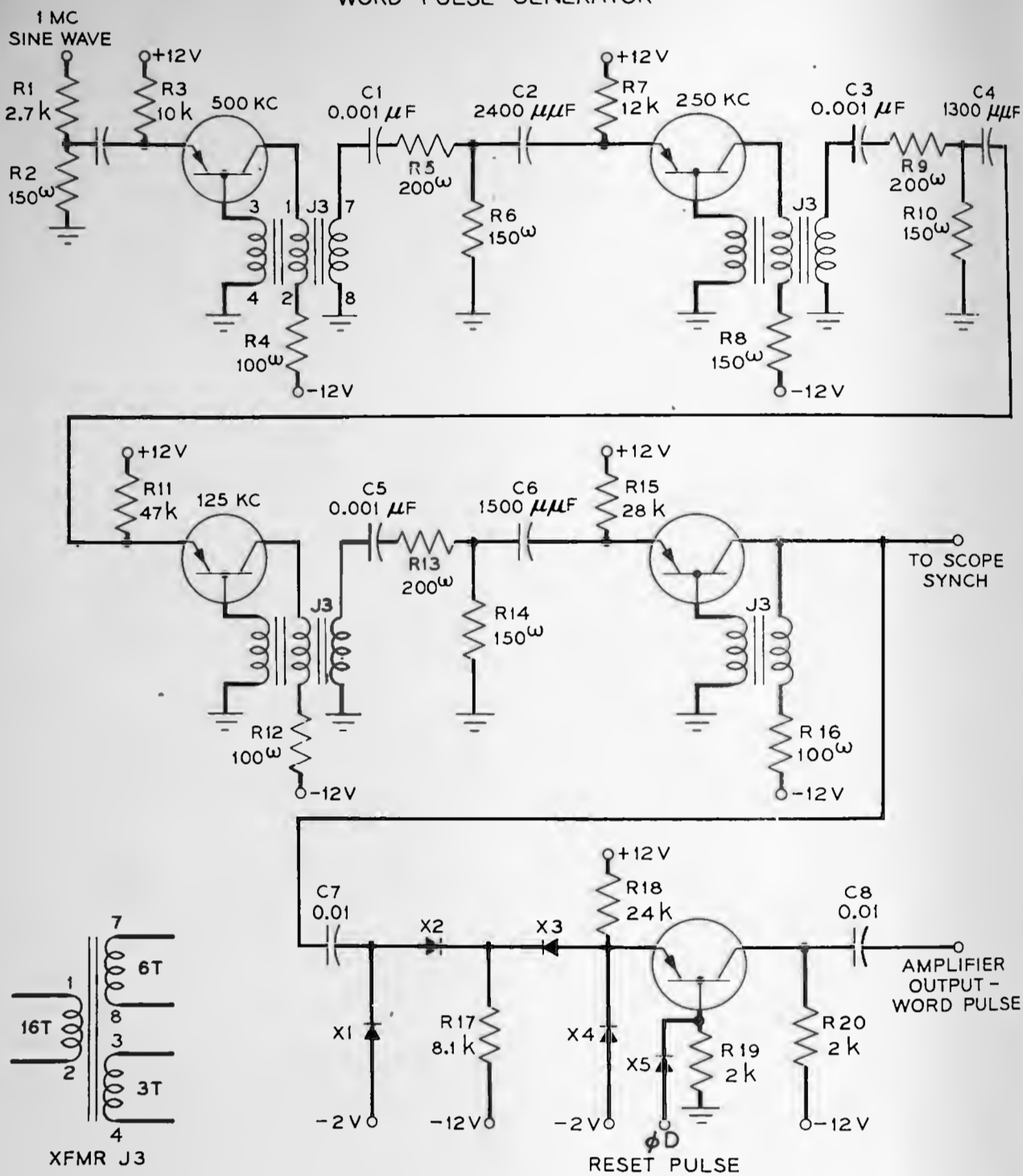
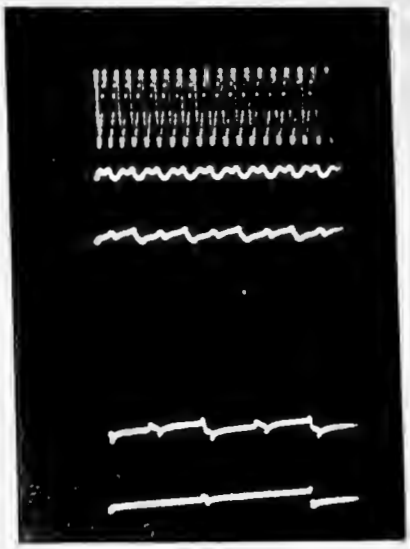
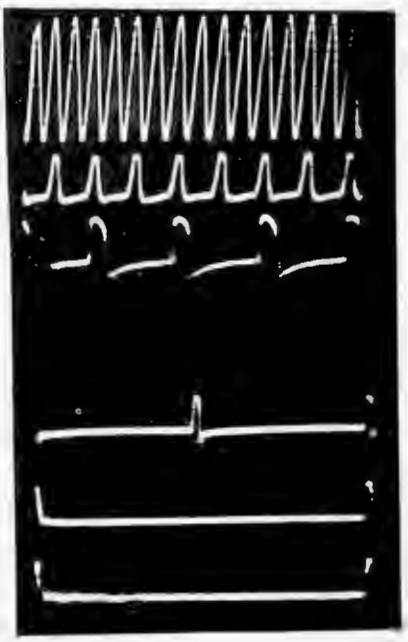

WORD - PULSE GENERATOR
1 MC
SINE WAVE
+12V
R1
2.7k
R2
150ω
R3
10k
500 KC
C1
0.001 μF
R5
200ω
C2
2400 μμF
R6
150ω
+12V
R7
12k
250 KC
C3
0.001 μF
C4
1300 μμF
R9
200ω
R10
150ω
J3
R4
100ω
-12V
R8
150ω
-12V
+12V
R11
47k
125 KC
C5
0.001 μF
R13
200ω
C6
1500 μμF
R14
150ω
+12V
R15
28 k
TO SCOPE
SYNCH
J3
R12
100ω
-12V
R 16
100ω
-12V
+12V
C7
0.01
X2
X3
R18
24 k
C8
0.01
AMPLIFIER
OUTPUT -
WORD PULSE
X1
R17
8.1 k
X4
X5
R19
2 k
R20
2 k
-2V
-12V
-2V
-12V
φD
RESET PULSE
7
6T
1
16T
8
3
3T
2
4
XFMR J3

FIG. 4-1

WAVEFORMS IN WORD-PULSE GENERATOR

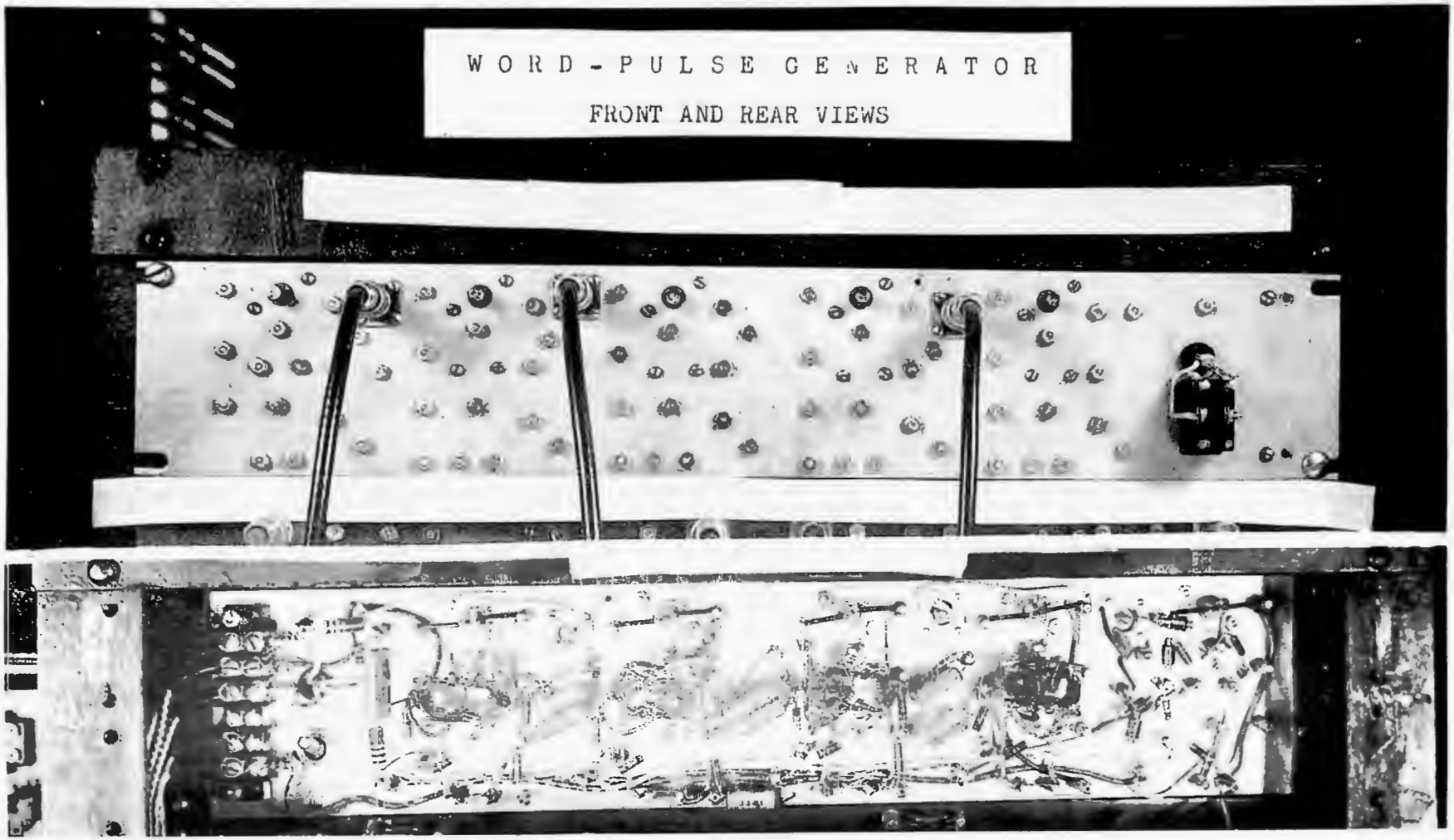


EMITTER WAVEFORMS



COLLECTOR WAVEFORMS

FIGURE 4-2


WORD - PULSE GENERATOR
FRONT AND REAR VIEWS

BASIC TRANSISTOR AMPLIFIER

5.1 INTRODUCTION

A good vacuum tube may not have serious transit time limitations upon its use at frequencies below 1000 megacycles, but the parasitic impedances of the same vacuum tube will limit its use in wide-band circuits to band widths of about 20 megacycles. High-speed transistors are limited by transit time considerations to frequencies below about 50 megacycles, but the parasitic limits are at frequencies above the transit-time limits, and for that reason the slow "hole" of transistor electronics may be a better carrier for the production or amplification of high-speed pulses than is the much faster electron of vacuum tube electronics. The M1734 type transistor in the circuit to be described is a fast pulse amplifier competitive with vacuum tubes in high-speed switching systems. In the applications described one may say that a vacuum tube would be a poor transistor substitute.

5.2 INPUT CHARACTERISTIC OF THE TRANSISTOR

Figure 1 illustrates the negative input impedance of the transistor for a current gain (alpha) greater than 1 and a high base impedance. Figure 2 shows a typical characteristic for an M1734 transistor.

The relationships shown on Figure 1 for the valley voltage and current are approximate and based on a number of assumptions about the relative magnitudes of the transistor parameters. The interested reader is advised to start with the "broken-line" assumptions of Mr. B. G. Farley and derive his own relationships. The broken line assumptions are:

1. R_e, the emitter impedance, is high and constant (comparable to the back impedance of a diode) when negative emitter current flows and low and constant (comparable to the forward impedance of a diode) when positive emitter current flows.

2. R_b, the base impedance, is constant.

3. R_c, the collector impedance, is high and constant for negative or positive emitter currents unless the unit is saturated, in which case the collector impedance drops to a constant value of the same order as the emitter impedance.

4. R_m, the mutual impedance, is zero for negative emitter current, constant and high for positive emitter current unless the unit is saturated, in which case R_m is zero.

The most significant feature of the characteristic is the peak voltage; the voltage, that is, at which the input resistance becomes negative. The peak voltage is generally negative and the peak point is reached after the current gain becomes greater than 1, which is generally assumed to occur when the emitter current ceases to be negative and becomes positive. There is some evidence that the peak point may not be reached until the emitter current is quite positive. In any event, the peak point is a first-order effect of base impedance and the collector current that flows when the emitter current is zero. In the M1734 transistor efforts are being made to hold the peak point between 0 and-1/2 volt for a transistor with no external base impedance and a collector voltage of -8 volts*.

Another significant point on the emitter characteristic is the valley point, the point at which the transistor saturates. As the emitter current is increased, the collector current increases at a more rapid rate (if $\alpha > 1$) and the resulting base current will carry the internal emitter-base-collector node negative. At the same time the collector current will carry the external collector terminal in a positive direction towards ground (if $R_1 > 0$). When the internal node has fallen to a voltage near that of the collector terminal, the valley point has been reached and the transistor is said to be saturated. The transistor has in effect become a passive network because the holes emitted at the emitter have reduced the effective collector impedance to a minimum value and R_m has therefore been made zero. Further increments in the emitter current will divide between the base and collector circuits in accordance with their impedance levels, and, as seen in Figures 1 and 2, will require that the emitter move positive.

The peak point may be compared to the cut-off point of a vacuum tube, while the valley point is somewhat like the saturation point that is reached in a vacuum tube circuit when the plate voltage has fallen to the cathode voltage. Those working with conditionally stable units must grow accustomed to the cut-off being above the saturation point.

- -

*Various arrangements have been described to make the peak point independent of the external base impedance. These circuit techniques cannot be used to nullify the effect of the internal base impedance.

5.3 ELEMENTARY CIRCUIT

An elementary circuit used to test the pulse performance of transistors is shown in Figure 3. The 1-megacycle sine wave applied to diode X1 carries the emitter positive towards the peak point. When the peak point is exceeded, X1 cuts-off and a current determined by R1 will flow in the emitter circuit. In the circuit shown, the current will be approximately 2 milliamperes and will cause the emitter to go negative along a path that depends upon the emitter capacitance to ground and the bandwidth of the transistor. In any event, the path proceeds either directly or circuitously towards the intersection of the load line of R1 with the emitter v-i characteristic displayed in Figure 1. If the intersection of the two load lines lies between the peak and valley points (a negative resistance region) the unit will oscillate provided the stray capacitance from emitter to ground is large or the bandwidth of the transistor is very wide. If the intersection is at a current greater than the valley current (a positive resistance region) the circuit will rest at a constant emitter voltage below the peak voltage as long as the input sine wave is positive. When the input voltage falls below the emitter voltage, diode X1 conducts and pulls the emitter negative. When the emitter falls below the valley voltage, the emitter current changes from a value greater than the valley current to a value less than the peak current and the transistor remains a high impedance until the input rises once more to a value greater than the peak voltage.

The ratio of i_c to i_e, for ie = 2 ma, as a function of time, has been defined as the pulse alpha; satisfactory M1734 units are expected to have a pulse alpha greater than 1.75 after 0.1 microseconds. Some units develop the required alpha in one or two hundredths of a microsecond and the excellent rise and fall time in the wave form shown in Figure 2 is typical of the M1734 transistor.

5.4 REGENERATIVE AMPLIFIER CIRCUIT

In the circuit of Figure 3, the over-all current gain* is limited to be of the same order as the current gain of the transistor. The current gain can be increased by the addition of diode X2 as in Figure 4. Diode X2 is returned

- -

* Defined as the ratio of the collector of current that flows when the circuit has reached its stable high-current state to the input current required to trigger the circuit.

to a voltage slightly more negative than the peak voltage. Resistors R1 and R2 are chosen so that the emitter will rest (when there is no emitter current) at a voltage below the bias on X2. This is point A in Figure 4. When the input goes above the peak point, diodes X1 and X2 will be cut off and the current set by R1 will start to flow into the emitter. Because of the negative input impedance the positive emitter current causes the emitter voltage to fall towards C. If the negative resistance is high enough and the unit has sufficient high frequency gain the emitter current will snap out to point B. The circuit will then stay locked up at B indefinitely unless the emitter is pulled below the valley point C or the base is pushed positive. Note that because of the negative impedance at the emitter, the load line can be switched from the high impedance offered by R2 to the low impedance offered by X2 when conducting. This gives the circuit a large current gain at the emitter. The emitter current is multiplied by the alpha of the transistor to give a practical current gain at the collector of the order of 20. The minimum triggering current will be influenced by the slope of the curve between point A and C. It may be noted in this connection that if the load line of R_1 does not intersect the load line of X2 before it intersects the load line of the transistor, in some cases (presumably those involving transistors with restricted bandwidths) the transistor will get "hung up" on the negative impedance portion of the characteristic and the collector signal will be very small.

The above circuit is actually a flip-flop rather than an amplifier. In order to make it an amplifier it is customary to replace diode X2 with a capacitor. When the emitter current has snapped out to line DC, the capacitor charges down line DC. When the emitter voltage reaches the valley point the current snaps to the low-current state, the capacitor discharges to A, and the unit locks at the low-current state. Thus, the circuit develops an output pulse in return for each input trigger. Two disadvantages of using a condenser to provide automatic shut off in high-speed applications are discussed below. One disadvantage is that the duration of the output pulse depends upon the slope of the transistor characteristic between points C and D. Another disadvantage is that in triggering the circuit, the condenser must be charged from voltage A to the peak

point. This means that if significant delays are to be avoided in high-speed applications, a larger triggering current is required than when diode X2 is used. The circuit shown in Figure 4, can be triggered by a quarter of a milliampere without delays significant to a computer operating at a megacycle rate. (A disadvantage of the circuit of Figure 4 is that it does require a reset pulse.)

Turning off the circuit shown in Figure 4 by a voltage applied at the emitter is extremely difficult since the emitter has to be pulled negative to the valley voltage and the voltage drop has to be developed across the low impedance of the conducting diode X2. However, the circuit can be shut off by a positive reset pulse applied at the base of the transistor through diode X3. It will be seen that the transistor can go to its high current state in this circuit only when the reset pulse is not present and will always be driven back to the low-current state by the next reset pulse applied to the base. Therefore, the onset and the duration of the input pulse is under the control of the reset pulse and is to a degree independent of the transistor, the shape of the input pulse, and the circuit parameters.

This feature makes it a very simple matter to obtain the rigid synchronism required in a serial computer or switching system. Sine-wave clock signals are applied to the bases of the transistor amplifiers through diodes. The phase of the signal fed to the base is chosen so that the base will fall to ground just after the latest time at which the input pulse, if present at all, could have risen. Then the output, if any, will be in synchronism with the clock and not the input. The circuit is therefore a pulse standardizer as well as an amplifier.

These matters may be made clearer by reference to Figure 5. Note that a collector pulse is not produced by a base pulse alone. Only the signal at the emitter can produce an output, and it can produce a signal only during intervals set by the base or clock signal. Diode X4 has been added to the base circuit in order to present a high base impedance when the transistor base is driven positive and to present a low impedance when the transistor is triggered. R4 is added to provide a path for discharging stray capacities. The type of load shown in Figure 5 provides d-c restoration of the signal so that the pulse going into the next diode network

will always start at -2 volts. The load has another purpose in that the conducting diode X5 provides a very low load impedance for the transistor when it first starts to flip to the high current state. This speeds up the regenerative operation. Once the collector voltage has risen appreciably, diode X5 is cut off.

It may also be mentioned that the circuit can be used for indefinite storage by holding back the base pulse until the stored data is to be removed from the circuit. Difficulties encountered with the circuit are described in a later paragraph.

5.5 SOME APPLICATIONS

In Section II "Catalog of Block Diagrams" an extensive group of block diagrams were presented in which all logic operations were to be performed in passive diode circuits and active elements were used only as gain-producing devices to make up for attenuation in the diode circuits. The amplifier circuit described in Paragraph 4 is a circuit recommended for such an application.

Figure 6 shows three common logic circuits. The bottom circuit has an output whenever at least one of its three inputs are energized. The middle circuit will develop an output whenever at least two of its three inputs are energized. The top circuit has an output only if all three of its inputs are energized. Note that the same regenerative amplifier circuit is used in each case.

The top photograph of Figure 7 shows three trains of input pulses; the bottom photograph shows the corresponding outputs of the circuits of Figure 6. The reader can check the operation of the circuits by casting his gaze vertically upwards and downwards across the two photographs to see that the indicated discriminations have been made at each pulse period. As is the case with all the waveforms shown in this memorandum, the width of each pulse represents one-half a microsecond.

Another common logic operation is that of inhibition. A three-terminal and-circuit to which an inhibitor terminal has been added is shown in Figure 8. This circuit has an output only when all three and-inputs are present and the

inhibitor pulse is not. The inhibiting pulse is applied to the base and in effect fills in the blanks in the clock pulses (which, otherwise, would leave the transistor ready to be operated). A complete binary adder has been built using the circuits described in this paragraph and has been operated for over a month at a megacycle rate.

5.6 DIFFICULTIES

The one major difficulty that has been encountered in this project is that some M1734 transistors do not work in the circuit, and it is not always apparent what the basic differences are between a unit that works and one that doesn't. In a group of 42 M1734 transistors recently received, 37 worked in the circuit of Figure 5 and 5 didn't. Of the 37 which worked, some are considered marginal.

In testing units for operation in this type of circuit three voltages are measured. In the circuit of Figure 5, the bias on X2 is made variable. The bias voltage is set at -4 and then gradually brought up towards ground. At first the output will not be a half-microsecond square pulse but will be a packet of sawteeth each lasting about 0.1 microsecond, with the packet lasting a half microsecond. This indicates that the bias on X2 is more negative than the valley voltage. As the bias is carried towards ground, the packet will become a square wave, indicating that the transistor is sticking in a high current state until reset by the base pulse. This means that the load line of X2 is intersecting line CD (see Figure 4) at a voltage more positive than the valley voltage. The voltage at which the square wave is first obtained is called E_1, (a measure of the valley point). As the bias on X2 is carried closer to ground, it may occur that an output pulse is developed at the end of every base pulse regardless of whether or not an emitter trigger has been received. This voltage is called E2 and is taken as a measure of the peak point, since the fact that the circuit "fires" at the end of every base pulse indicates that the load line of X2 does not intersect the emitter characteristic in the negative current region. Since, in the absence of an emitter input, the emitter rests at a voltage more negative than the bias on X2, E2 is more positive than the peak voltage. The bias on X2 is then lowered and the voltage E3 is noted at which the circuit responds only when there is an emitter trigger. E3 is generally quite near E2 if it is much below E2 hole storage troubles are suspected.

The effects of hole storage are apparent in the positive pips on the trailing edges of almost all of the wave forms. The effect is clear in Figure 5. Note that the collector voltage goes up to (and in some cases above) the emitter voltage before the base pulse succeeds in resetting the transistor. This is not due to a lack of amplitude in the base pulse but results from the fact that the collector remains a low impedance circuit until all of the stored holes have been cleared out.

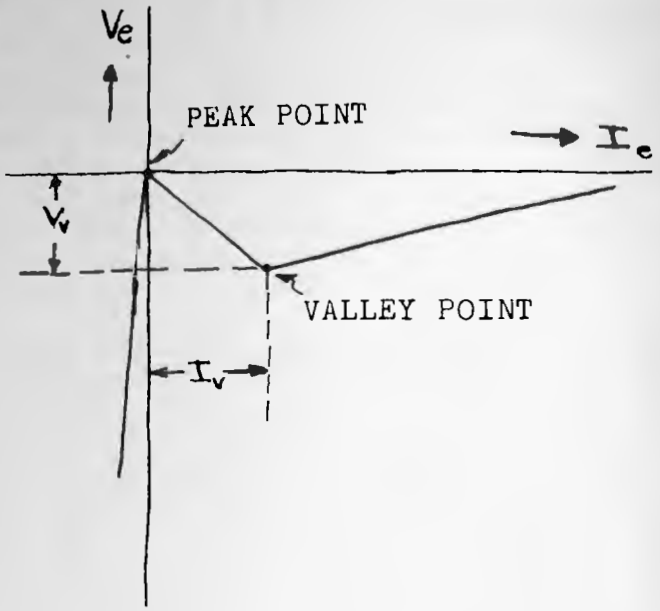
These hole storage effects are undesirable because they make the duration of the output pulse a variable, they load down the resetting signal, and if very bad they persist so long that the transistor may not be turned off by the time the base pulse has ended. Thus the circuit becomes locked up permanently in the high-current state. This is particularly serious when transformer coupling is used. The low d-c impedance of the transformer permits large average currents to flow and thereby increases the hole storage. Thus the E3 defined previously may be much lower than E2. In some cases E2 may be lower than E1. As a result, at the present time transformer coupling cannot be used with more than a few transistors. (Transformer coupling is desirable because it permits the major part of the signal power to be used in the load rather than dissipated in the collector resistor (R4 of Figure 5).)

For successful use it is expected that in the circuit of Figure 5, that E1, defined as above, will be below -2 volts, that E2 will be above -1/4 volt and E3 be above -0.4 volt. This gives good margins in the circuit when X2 is biased at -1 volt. It is found that units work satisfactorily and interchangeably, if E1, E2 and E3 are as stated above. A source of dissatisfaction is that it has not been possible to correlate the values of E1, E2 and E3 measured in the circuit of Figure 5 with the values that d-c measurements indicate. The writer feels that what is needed is more knowledge about transistor parameters in the megacycle region and more knowledge about how such parameters will affect the negative resistance. It should be emphasized that not only must more be known about transistors but more must be known about negative resistance devices when operated at frequencies where transit time considerations and storage phenomena must be considered.

5.7 CONCLUSIONS

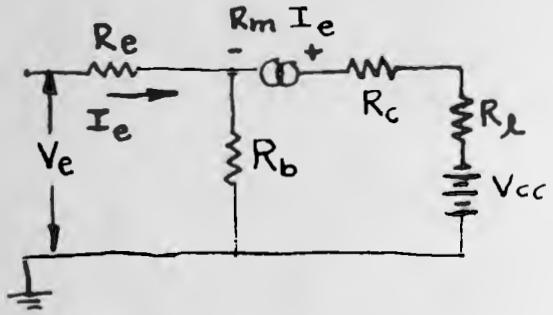
A transistor amplifier circuit can be made to work well in high-speed switching and computing circuits. It is believed that such a transistor circuit can be used advantageously to replace many vacuum tube circuits.

IDEALIZED INPUT CHARACTERISTIC



$$V_v \cong \frac{V_{cc}}{1 + \left(\frac{\alpha}{\alpha - 1}\right)\frac{R_l}{R_b}}$$

$$I_v \cong \frac{-V_{cc}}{(\alpha - 1)R_b + \alpha R_l}$$



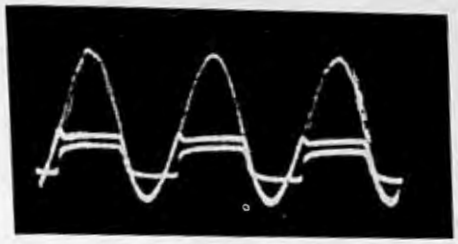
$$\alpha = \frac{R_m + R_b}{R_c + R_b} = \left[\frac{i_c}{i_e}\right]_{R_l = 0}$$

FIGURE 1

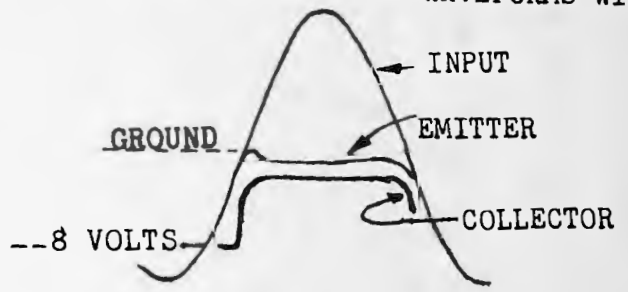
EMITTER CHARACTERISTIC

FIGURE 2

TEST CIRCUIT FOR PULSE∝



WAVEFORMS WITH DA498



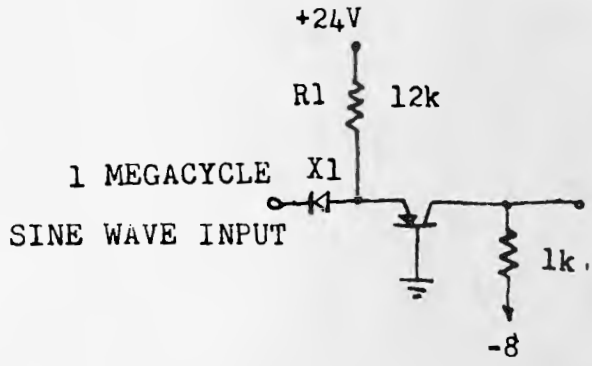


FIGURE 3

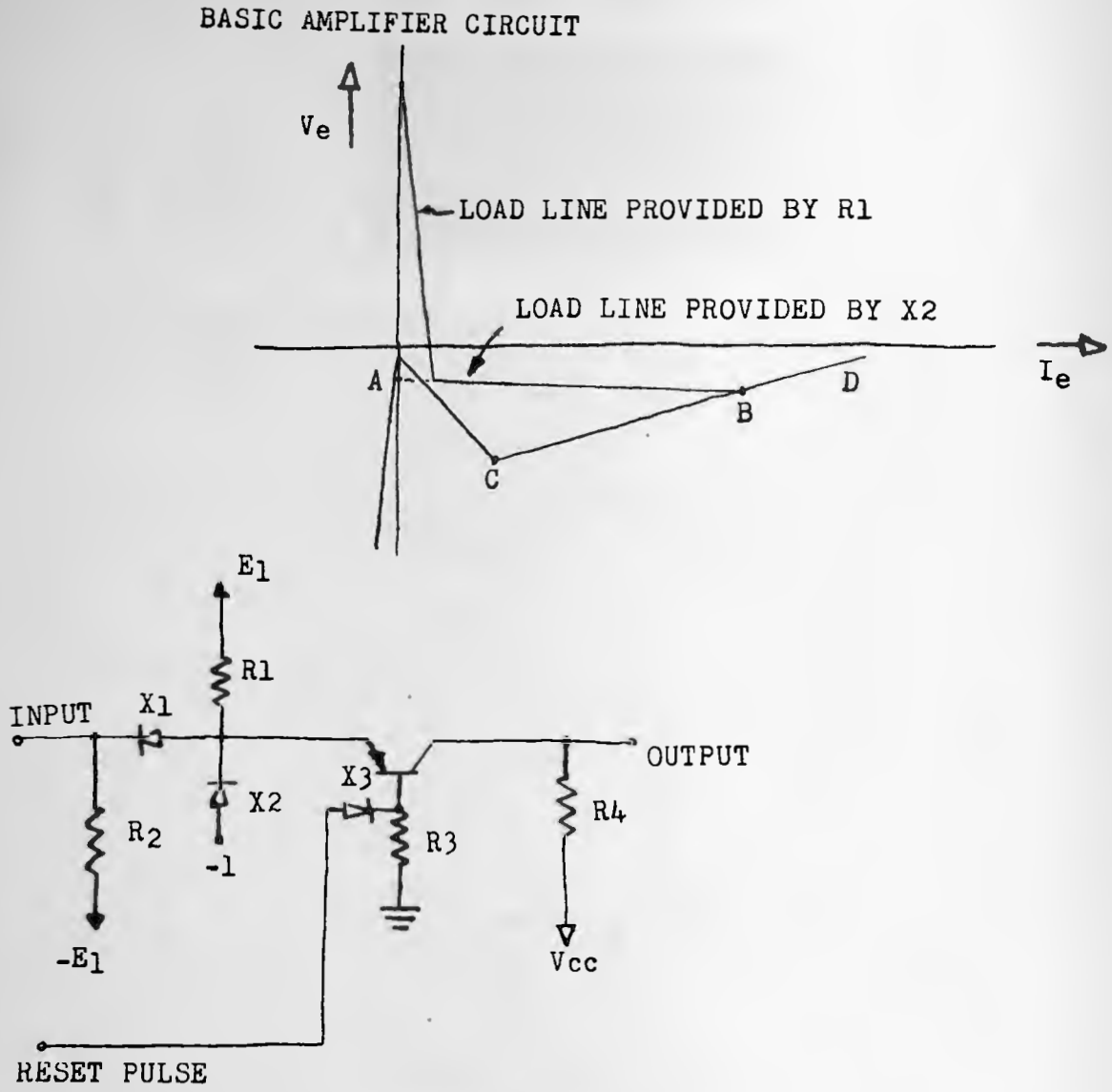
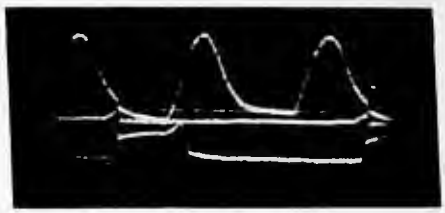
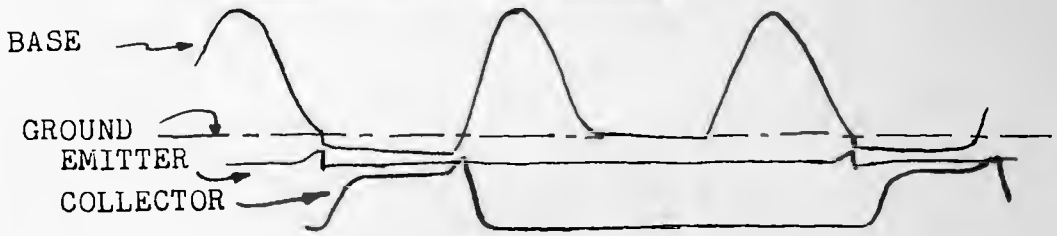

BASIC AMPLIFIER CIRCUIT
Ve
LOAD LINE PROVIDED BY R1
LOAD LINE PROVIDED BY X2
A
B
C
D
Ie
E1
R1
INPUT
X1
OUTPUT
X3
X2
R2
R3
R4
-1
-E1
Vcc
RESET PULSE

FIGURE 4

AMPLIFIER WITH WAVEFORMS





WAVEFORMS WITH DA498

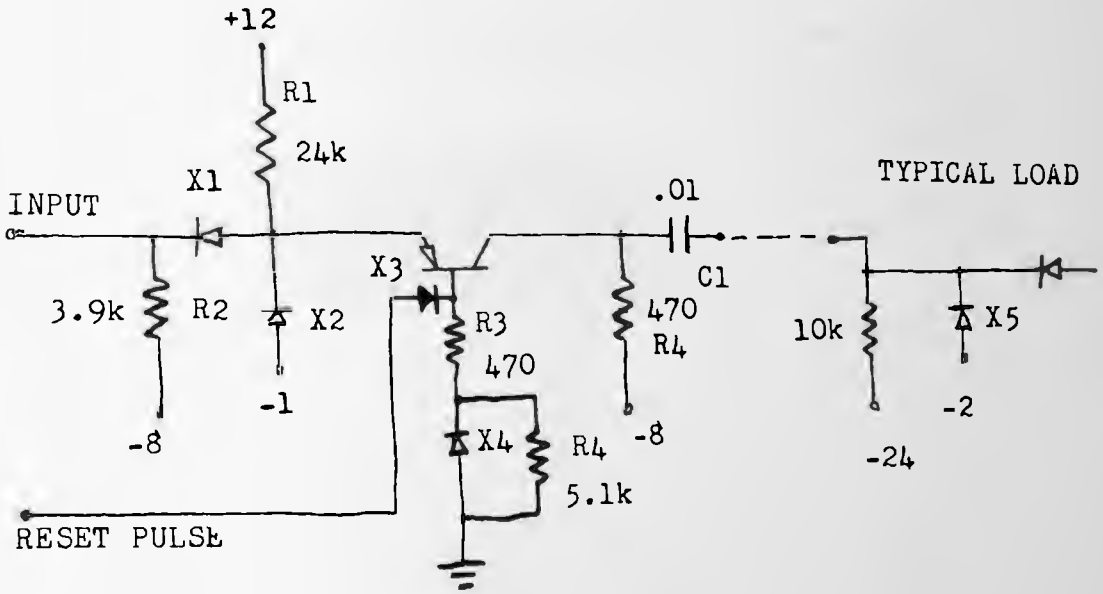


FIGURE 5

SOME ELEMENTARY 3-TERMINAL LOGIC CIRCUITS

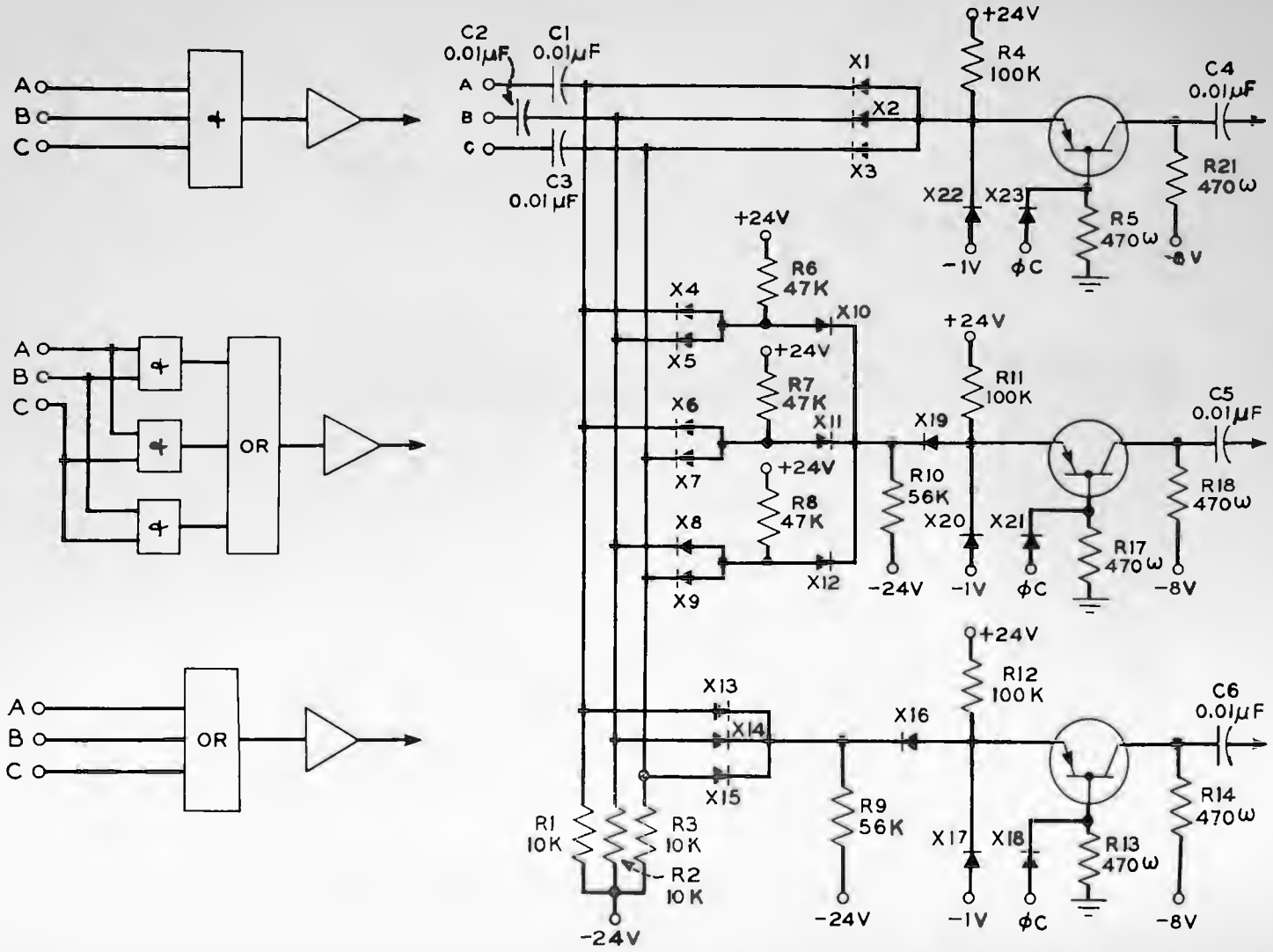


FIG. 6

INPUTS AND OUTPUTS OF 3-TERMINAL LOGIC CIRCUITS

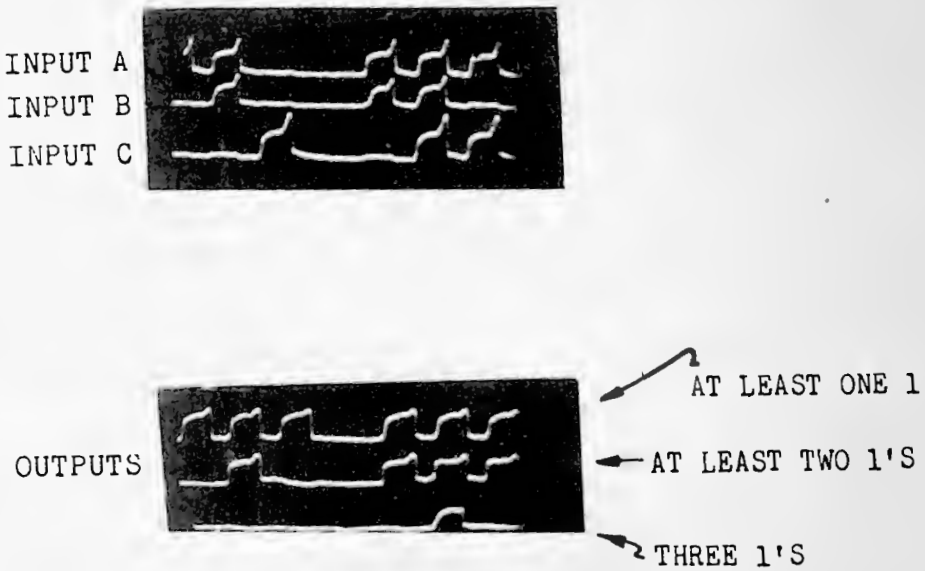
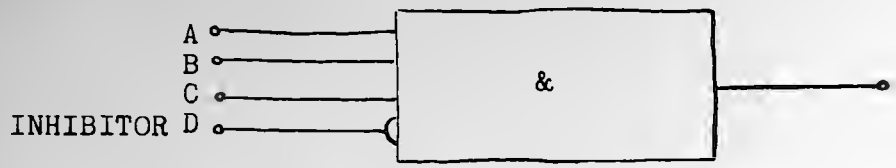
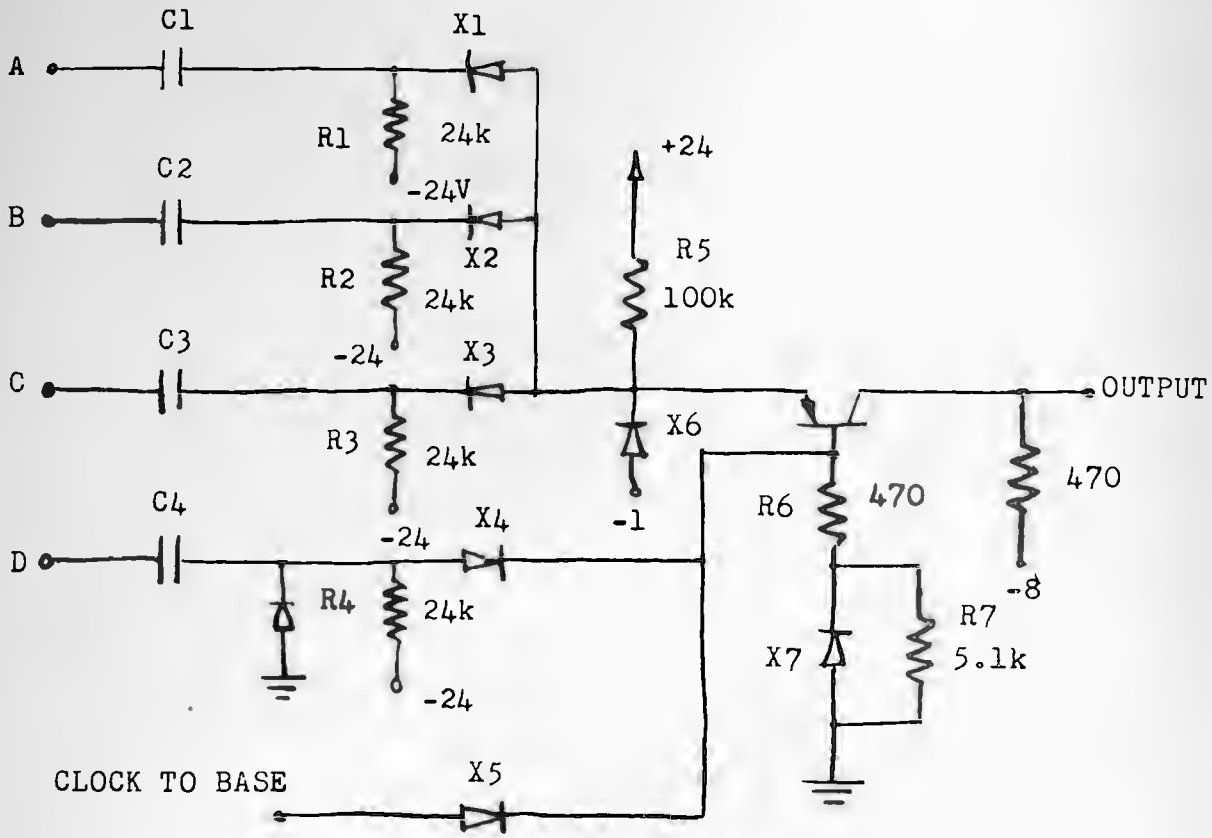


FIGURE 7

3-TERMINAL AND-CIRCUIT WITH INHIBITION





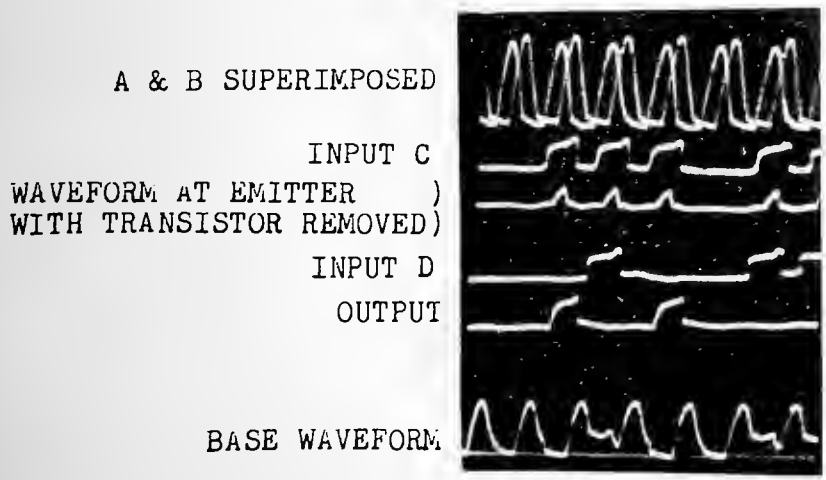


FIGURE 8

A WORD GENERATOR

6.1 INTRODUCTION

When digital computers or switching systems are developed, a fundamental device needed is a generator of numbers or words* to be used in testing the operation of the system. The device described herein generates 16 digit serial numbers in which each digit is represented by a 1/2 microsecond pulse. The generator can be set by means of manually operated switches to generate simultaneously any 4 numbers between 0 and 2^{16}-1. Other uses of the generator are discussed below. A front view of the word generator is shown on photograph 191565 while a back view is shown in photograph 191567.

6.2 BLOCK DIAGRAM

The generator generates the desired numbers every 16 microseconds. It is kept in rigid synchronism with the output of a 1 megacycle clock oscillator. The word pulse that constitutes the input to the unit is obtained from the megacycle clock by means of a transistor blocking-oscillator frequency divider which puts out a pulse every 16 microseconds. In the word generator this word pulse is regenerated 16 times and delayed by one microsecond 15 times (see Figure 1). The pulse appears at the output of amplifier one at zero time and represents 2^0. It appears at the output of amplifier 16 at time 15 and represents 2^{15}. In addition to feeding a delay line each amplifier feeds one of the vertical busses shown in Figure 1 and photograph 191567. A group of 4 horizontal busses are shown in Figure 1 and each horizontal bus may be connected to any of the desired vertical busses by a manually operated switch with a diode in series to give isolation between the vertical busses (photograph 191567). At the end of each horizontal bus there is an output amplifier. The user operates the bank of 64 switches to generate the specific code groups he wants. The generated words will be available at the output amplifiers at a level suitable for driving other equipment.

- -

*This device is called a word-generator in keeping with the nomenclature common in the digital-computer field. It might as well be called a number-generator or a code-group generator.

6.3 CIRCUIT DIAGRAM

Each of the twenty amplifiers shown in Figure 1 is exactly like all the others. The type was described by the author in paragraph 5. The circuit of two of the amplifiers, showing the delay line between them is given in Figure 2 which represents two successive stages in the chain of amplifiers through which the word pulse progresses.

The delay line used was obtained from the James Millen Company and has a nominal impedance of 1300 ohms and a delay of 1 microsecond per 20 inches. The delay line is made only 16 inches long so that the emitter of the nth stage will start to rise 3/4 of a microsecond after the n-1'th stage "fires". One quarter microsecond later the base pulse from the clock will have fallen to ground and the n'th circuit "fires" exactly one microsecond after the n-1'th. Thus the exact delay of the cable can vary considerably without changing the times at which the circuit operates.*

Figure 3 shows the emitter and base waveforms of one of the stages. (The positive peaks of the base waveform were lost in reproduction.) The figure also shows a typical number at the output of one of the word generator amplifiers.

Due to the compact nature of the amplifier construction, little stray capacitance is present. It has been found advisable to add 15 mmf from the emitter to ground in order to make the circuit less critical to transistors. This capacitance helps transistors with restricted band widths to switch from the low current to the high current state. It has also been found beneficial to operate the collector into about 50 mmf.

6.4 OTHER USES OF UNIT

The reader can see that the word generator has all the components required to convert serial codes to parallel codes. If instead of a word pulse, a serial number or code were fed into the first regenerative amplifier of Figure 2, after 16 microseconds the 16 vertical busses would carry the parallel representation of the code. Similarly if the 16 wires of a parallel system were connected to the vertical busses and a parallel number transmitted, the same number in

- -

*The clock pulse fed to the emitter is not necessary since the base pulse locks each transistor with the clock. The diodes X1, X2, and resistor R1 would be omitted in any future design.

serial form would be available at the 16th amplifier. For this application, it would be desirable to place a diode between diode X6 and resistor R5 to prevent the word backing up in the amplifier chain.

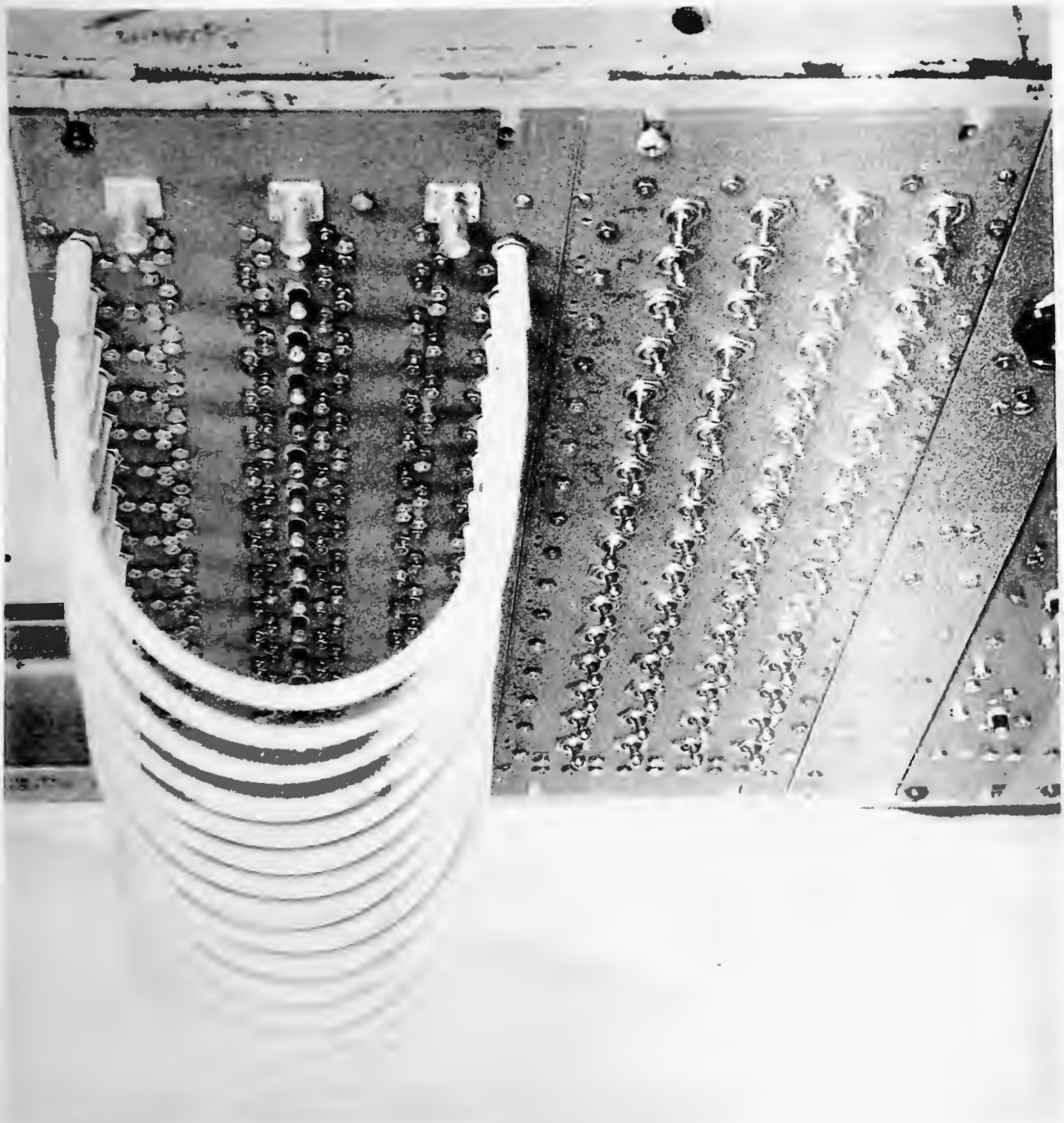
The unit might also be used to read punched cards. Consider the 16 switches connected to the top horizontal bus of Figure 2 as being feeler springs reading a punched card. If a half microsecond pulse were put into the word generator it would generate the number stored on the punched card. If the entire group of 64 switches of Figure 2 were arranged to read punched card data, the unit could read four words simultaneously. It is apparent that the switches might also be part of a commutator system on a shaft and the unit could then convert commutator positions to binary numbers. The low impedance of the transistor output circuits make such applications appear entirely practical.

Another use is for the permanent storage of numbers. If the group of 64 switches were replaced with appropriate permanent connections, the generator would always generate the same words. The generator might be extended to develop more than four words without difficulty. The numbers stored might, for example, be the numerical coefficients of some polynominal frequently needed in a computation.

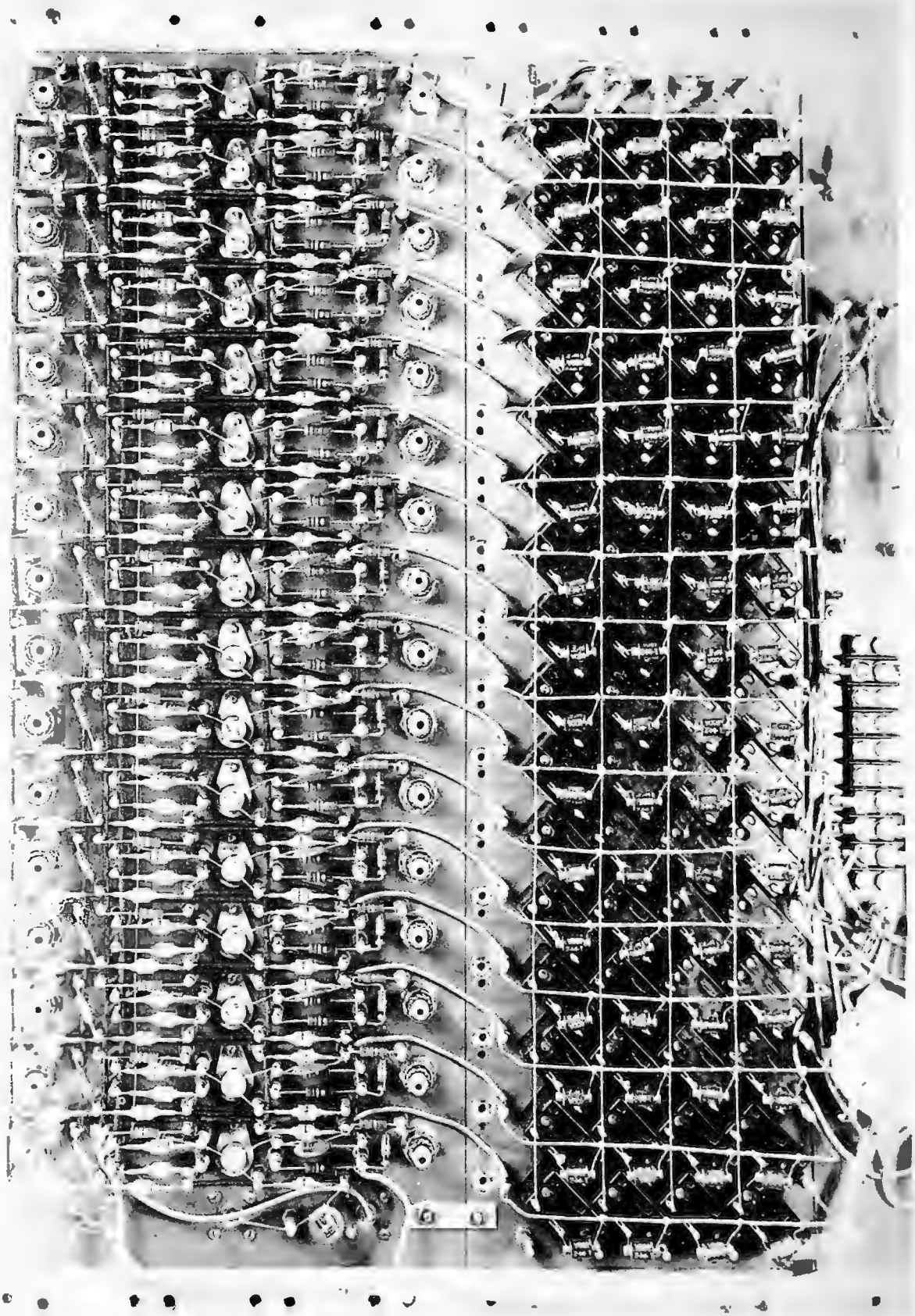
6.5 CONCLUSIONS

A transistor operated word-generator has been described in which 20 transistors, approximately 180 diodes, and 16 delay lines are operated on a total power of less than one watt and on battery voltages of +6, -1 and -8 volts. The unit works well and has been operated 8 hours a day since May 10, 1951. Only one unit has been replaced and it is believed that this unit was marginal when put in the generator. The word generator is being used to test transistor adders and memory units.

WORD GENERATOR



FRONT VIEW



REAR VIEW

WORD GENERATOR
BLOCK DIAGRAM

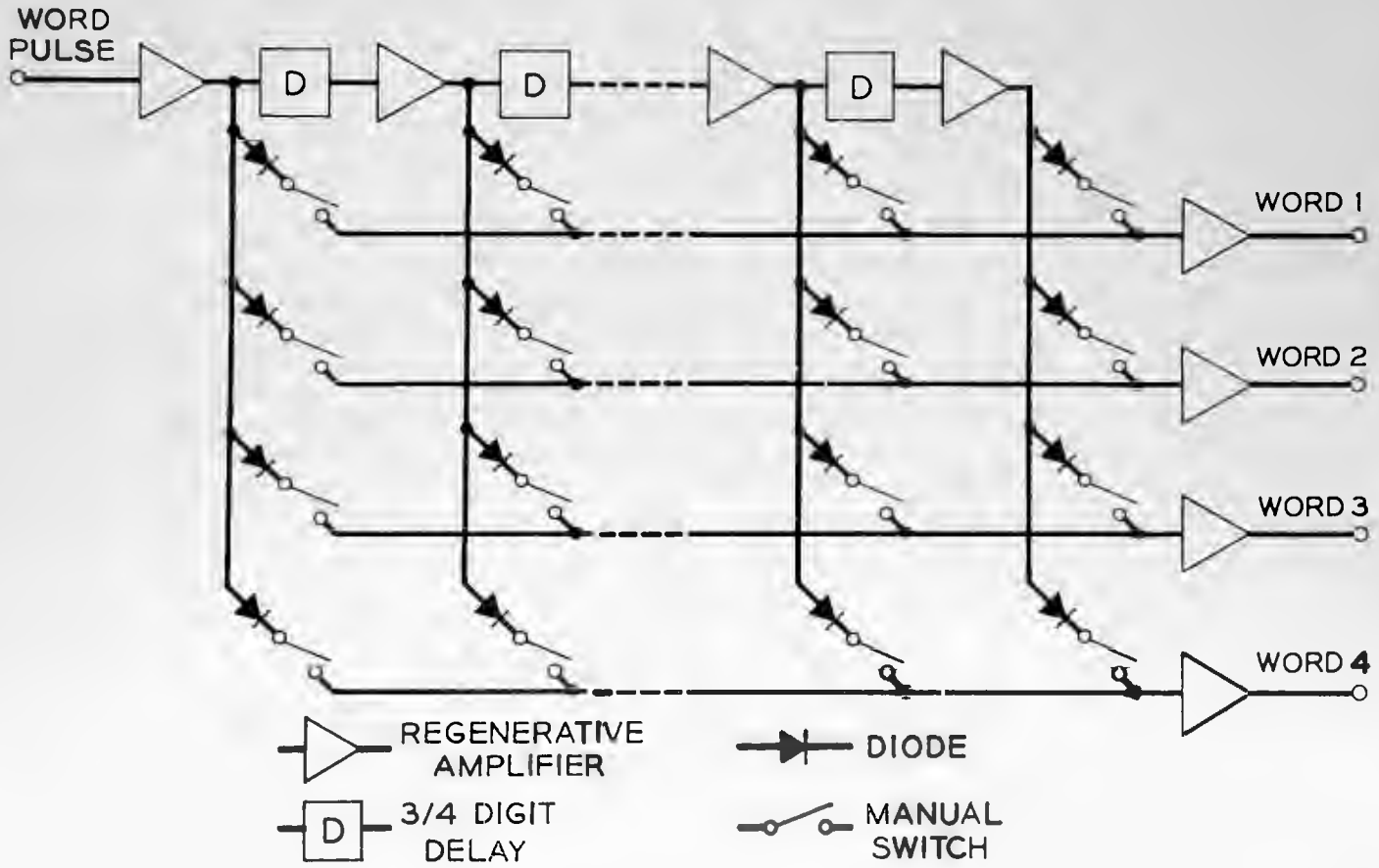


FIG. 1

REGENERATIVE AMPLIFIER WITH DELAY LINE

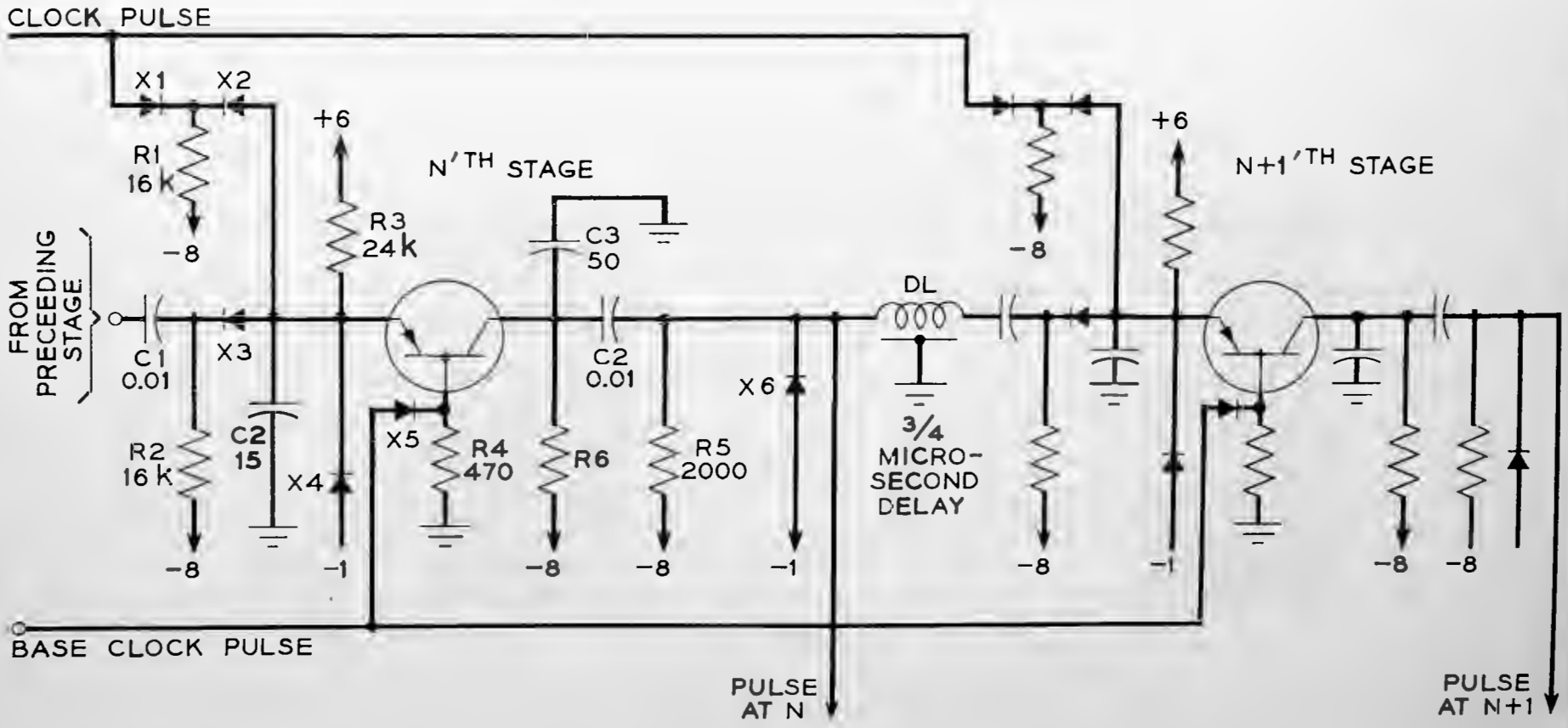
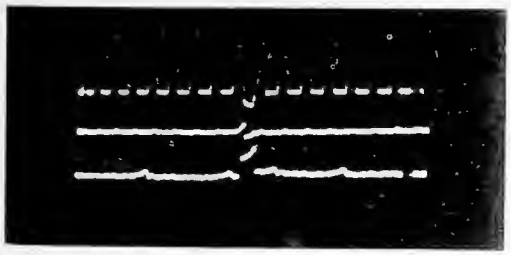


FIG. 2

WAVEFORMS IN WORD GENERATOR

A
B
C

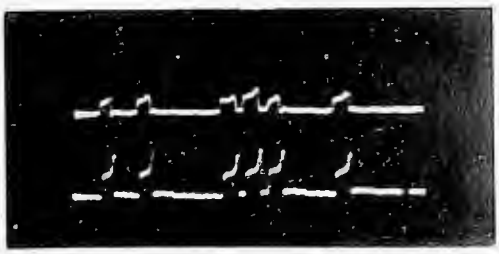


A. BASE WAVEFORM AT 10 TH STAGE

B. EMITTER WAVEFORM AT 10 TH STAGE

C. COLLECTOR " " " "

D
E



00101000 111 001 DECIMAL 10,004

D. TYPICAL NUMBER OUT OF DIODE MATRIX

E. NUMBER AFTER REGENERATION

FIGURE 3

DESIGN OF HIGH SPEED DIGITAL COMPUTERS USING TRANSISTORS

DELAY LINE STORAGE

7.1 INTRODUCTION

In section II, the writer described a method of building a digital computer based on a transistor amplifier, diode logic circuits, and electric delay line storage. In section V, the basic transistor amplifier was described together with the basic diode circuits. In this memorandum the work is continued with the description of a delay line storage cell (Figures 1 and 2) which will store 16 digit numbers at the price of only three active elements. Photographs 191566 and 191568 show front and back views of two storage units now in use.

7.2 DELAY LINE

The delay cable used was made by the James Millen Company and has a delay of about one microsecond per twenty inches. It is not the voltage attenuation of the line that requires the use of three amplifiers in the device pictured but is the time distortion suffered by a half-microsecond pulse after 5 microseconds delay. With a lumped-parameter delay line of greater bandwidth it is believed that a 16 digit word could be stored with only one regenerating amplifier.

7.3 CIRCUIT DIAGRAM

The complete diagram of the unit is shown in Figure 1. The transistor amplifiers have been described in detail in paragraph 5. It may be noted that the amplifier is turned on primarily by the signal from the preceding delay line but it cannot be turned on until the clock pulse at the base has fallen to ground and it will be turned off when the clock pulse rises one-half microsecond later. Each delay line is made 1/4 microsecond shorter than the total delay desired so that the output pulse will have been at the emitter one-quarter of a microsecond before the clock pulse at the base has fallen to ground. This arrangement permits the delay lines to vary in delay and rise time without upsetting the synchronism imposed by the clock.

The switch S1 is used to empty the delay line by inhibiting the first transistor. In a full-fledged computer application S1 would be removed and an inhibiting circuit would be placed at the output of DL3. Whenever a new word is to be stored, the inhibitor circuit would be pulsed and prevent the old word from circulating any longer.

The output of delay line 3, and the emitter, base and collector waveforms of the transistor it feeds are shown on Figure 2. The word being stored was developed by the word generator described in paragraph 6.

7.4 CONCLUSIONS

A 16 digit delay line storage device has been described which offers a more efficient storage facility than flip-flop or shift-register storage. It has another advantage in that it can be emptied of stored data by turning off one transistor rather than 16. This is very important where hole storage effects make it difficult to turn off more than one transistor with the output of another transistor.

The device has been operated 8 hours a day for over a month (since its construction) without transistor replacements. The unit makes a convenient device for the testing of transistors for random errors since if a word is introduced into the unit just once and then left to circulate, one transistor error will multilate the word permanently.* It has been possible to operate the unit for over 48 hours without failure.

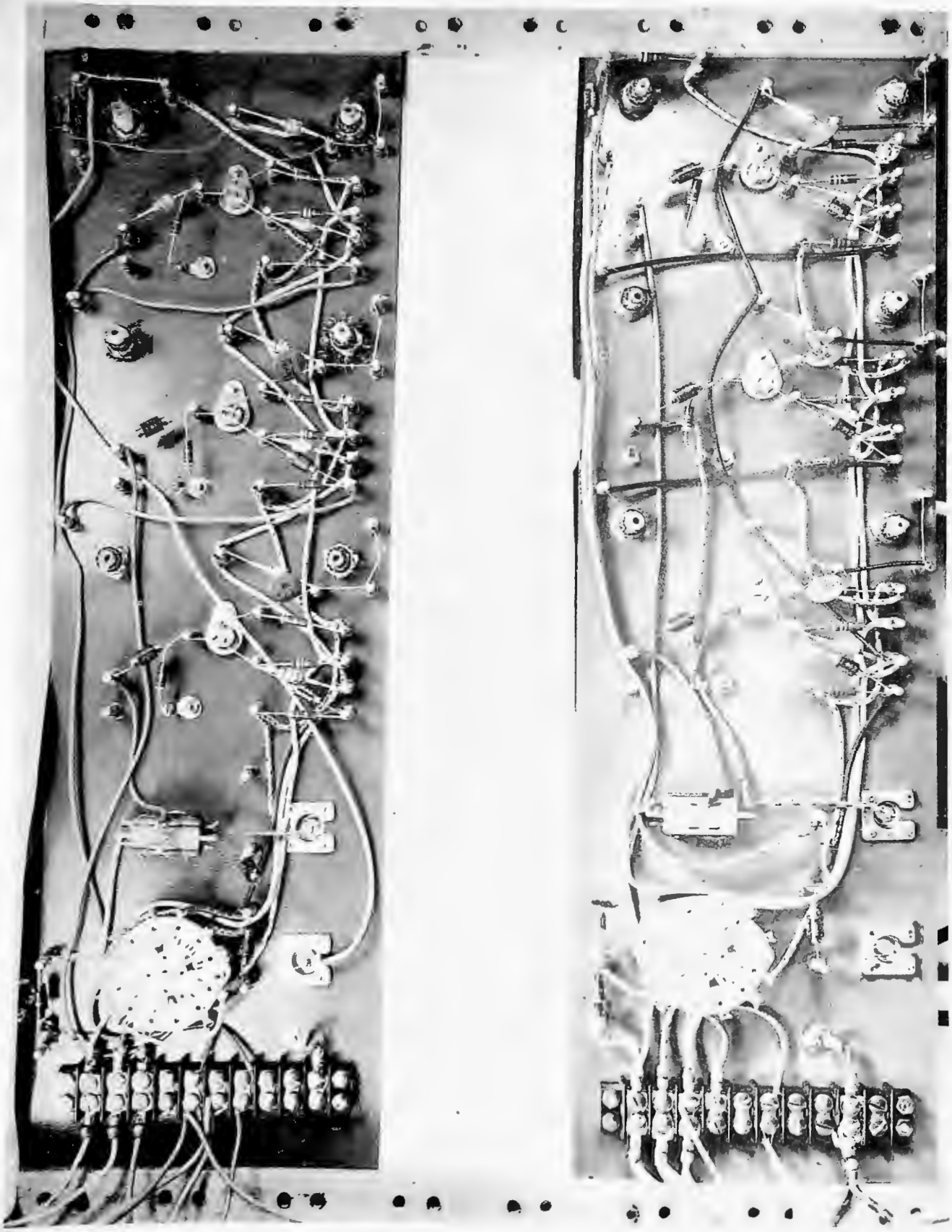
This is the equivalent of 500 billion successful transistor discriminations. When a megacycle clock is available which will operate in spite of a-c power shutdowns, it is hoped that long term tests can be run.

- -

*Transistors produced for digital applications might profitably be "aged" in such a circuit before they are delivered to a prospective user.

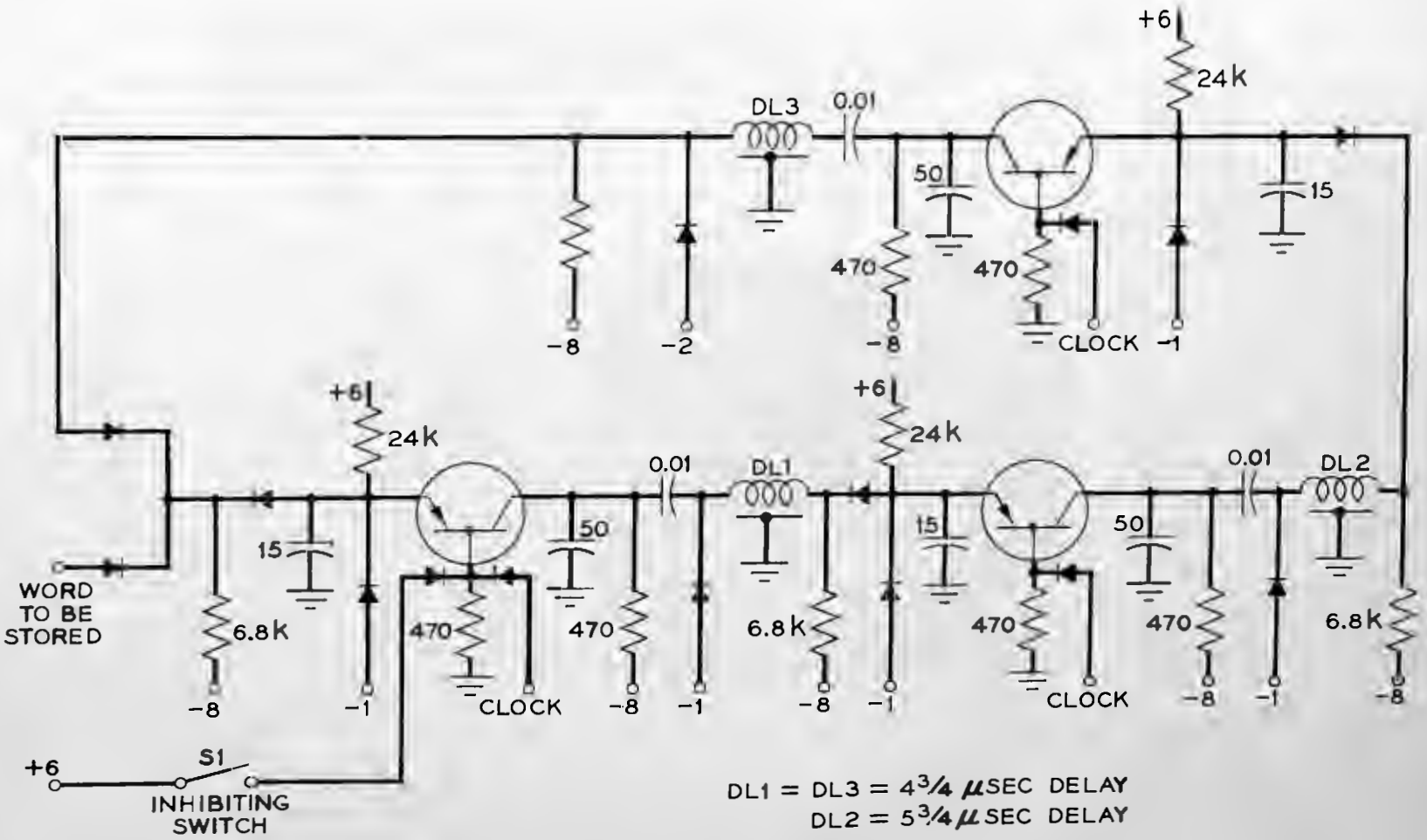
D E L A Y L I N E S T O R A G E U N I T

DELAY LINE STORAGE UNIT



REAR VIEW

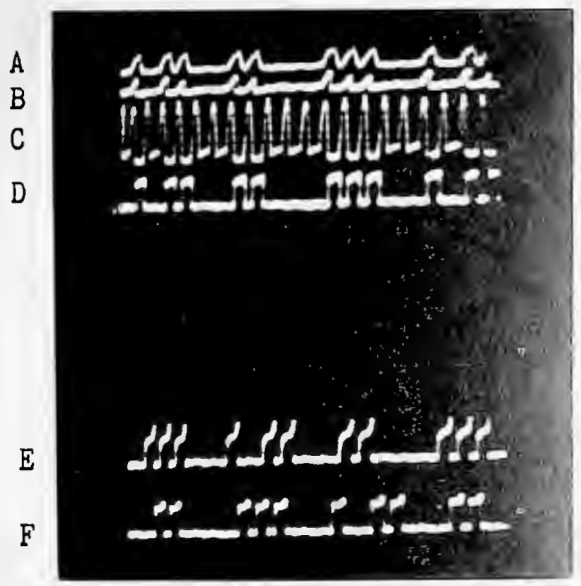
DELAY LINE STORAGE



DL1 = DL3 = 4 3/4 μSEC DELAY
DL2 = 5 3/4 μSEC DELAY

FIG. 1

WAVEFORMS IN STORAGE UNIT



A - OUTPUT OF DL3
B - EMITTER OF FIRST STAGE
C - BASE OF FIRST STAGE
D - COLLECTOR OF FIRST STAGE
E - COLLECTOR OF SECOND STAGE
F - COLLECTOR OF THIRD STAGE

FIGURE 2

BIT STORAGE WITHOUT FLIP-FLOPS

8.1 INTRODUCTION

Many of the storage requirements of a serial computer can be met by delay line storage in which a word to be stored is re-circulated in a one word delay line and kept in synchronism with the rest of the computer by means of regenerative amplifiers. This type of storage is quite efficient since for example, 16 digits can be stored with the use of only three transistors as regenerative amplifiers. A mechanization of this type of storage was described by the author in paragraph 7 above. Even in serial computers, however, this form of storage cannot be used for all storage functions. There are some places in a computer where the stored information has to be continuously available and cannot be available only once every word period. For example, there are places in a computer where switches are required. These switches normally must be self locking; that is, when a switch has been closed it is supposed to remain closed until it receives an instruction to open. When it is opened, it must remain open until it receives an instruction to close. There must be some memory element associated with the switch which stores the close instruction or stores the open instruction and keeps the switch in the desired state. It has been customary to use flip-flops for this type of storage.

This memorandum describes a different storage circuit for such use. It is called a bit storage cell because it can store one binary bit of information, either a 1 or a 0. In Figure 1, the schematic representation of the circuit is shown as a box with a M standing for memory, with an output terminal, and 2 input terminals: one input terminal is labeled "one" and the other labeled "zero". This device has the property that if a pulse is put on the one terminal, a 1 will be regenerated continuously at the output until a pulse is put on the zero terminal, in which case the output will be 0 and remain 0 indefinitely unless a new pulse is put on the one terminal. Thus it is seen that the output of this device is a train of pulses occurring at the clock frequency when a 1 is stored and when a 0 is stored there is no output at all.

A typical use for this type of storage is also shown in Figure 1 where the mechanization of a single pole switch is given. The memory circuit feeds one terminal of an and-circuit which has its other terminal connected to the signal to be switched. If a pulse is fed to the one terminal of the memory circuit, the switch will be closed and whatever data is put on the input to the switch will appear on the output. If on the other hand, a pulse is put on the zero terminal of the memory circuit, the memory circuit will have no output thereafter and the switch will be open because the and-circuit cannot respond to the input on the in terminal.

8.2 POSSIBLE MECHANIZATIONS

A mechanization for this type of storage is shown in Figure 2A. If a pulse is put on the one terminal of the device, it will go into the amplifier through the or-circuit and the inhibitor-circuit and will appear as an amplified pulse at the output of the amplifier. This pulse will travel back through the one digit delay line into the amplifier and continue to recirculate indefinitely appearing once every digit period at the output of the amplifier. The inhibitor-circuit is put in series with the circulating path so that if a pulse is put on the inhibition lead it will prevent the output of the delay line from getting back into the amplifier. This scheme is not considered very efficient because it requires a 1 digit delay line and if this is made out of the delay line materials now available, it will be a very bulky component. Since it is only necessary that 1 digit be delayed by 1 digit time in order to implement this device, a delay line isn't really required. Suppose that the output pulse of the amplifier, which is normally a positive pulse, is inverted to be made a negative going pulse and this negative going pulse is put across a parallel tuned circuit. The output pulse will cause the parallel circuit to ring negative and then after a delay ring positive. This positive ring can be fed back into the or-circuit, of Figure 2, to retrigger the amplifier. To implement this scheme it is necessary to have a transformer, to convert the positive going amplifier output to a negative pulse, and a resonant circuit. A little thought reveals that the resonant circuit can be combined with the transformer since the transformer that inverts the pulse may itself be made to ring. This scheme is shown in Figure 2B where the transformer is fed through a series resistance.

The series resistor in effect decouples the transformer from the amplifier and prevents the low output impedance of the amplifier from damping the transformer. This makes it a poor transformer but a good ringing device. It also prevents the transformer from loading down the amplifier appreciably.

8.3 CIRCUIT DIAGRAM

In Figure 3 a specific circuit is shown. The amplifier used was described in paragraph 5 above. When a 1 is put into the one terminal of the memory circuit it fires the transistor which causes the collector to go positive. The clock pulse at the base turns the transistor off 1/2 microsecond later. The pulse at the collector passes current into transformer T2 through R5 and causes the transformer inductance together with the shunt capacity presented by the circuit to ring. After the pulse at the collector has disappeared the transformer secondary rings positive and through diode X6 retriggers the transistor provided that an inhibitor pulse has not been fed to transformer T1. The clock pulse at the base of the transistor insures that the transistor will not fire until the preceding digit period is finished. An inhibitor pulse fed to T1 will be inverted and through diode X2 will hold the emitter negative thereby preventing a stored pulse from recirculating.

The two transformers T1 and T2 are miniature transformers designed by Mr. A. H. Bobeck. These transformers are designed to pass a 1/2 microsecond pulse into a 470 ohm load. When operated from a high impedance generator and into a light load, these transformers ring at about one megacycle in response to a step pulse.

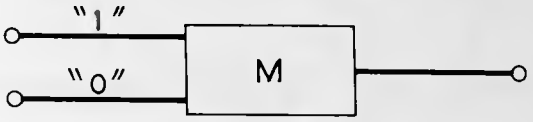
Figure 4 shows typical wave forms in a memory circuit. The input pulses applied to the one terminal and the zero terminal of the memory device are taken from the word generator described in paragraph 6. A 1 is stored in the memory device at time one. Four microseconds later, a pulse is put on the zero terminal and empties the storage cell. At time 6 another pulse appears on terminal one but does not get through the amplifier because at the same time a pulse is received on the zero terminal. The pulse at time 7 goes through the amplifier, but it does not recirculate because at time 8 a pulse is put on the zero lead. The output of the device is 0 until time 11 when another pulse appears on the one terminal. This pulse is stored for 4 digit periods until another pulse is put on the zero terminal and empties the cell.

Since the regenerative amplifier used in this device is itself a bi-stable device and is only made to look like an amplifier by being turned off every digit time by a clock pulse, the reader may ask why isn't the bi-stable transistor circuit used for memory. The reasons are as follows: The bi-stable transistor circuit can be turned on quite easily since the emitter represents a high impedance when the transistor is in the low current state. However, once the transistor is locked up in the high current state, both the emitter and base terminals present very low impedances and it takes considerable current to shut off the device. In the computing schemes the writer has proposed it is never necessary to turn off one transistor by another. Transistors are used only to turn on one another and the shut-off pulses are developed from a clock. Since the clock signal represents steady state power, it is felt that it can be generated efficiently. If the bi-stable transistor circuit were used for a memory circuit it would be left locked up in its high current state until it was desired to store a 0 and a transistor some place in the computer would have to develop enough current to shut off the transistor at that time. In the scheme shown in Figure 3, it is no more difficult to store a zero than a one, because the storage of a zero merely requires that a positive pulse be applied to T1. T1 converts the pulse to a negative going pulse, which causes diode X4 to conduct. It is only necessary for diode X4 to carry slightly more than a quarter of a milliampere to prevent the transistor circuit from firing. The merit of the circuit can be seen, therefore, to lie in the fact that it can be made to store a 1 or a 0 equally well and the sensitivity to a change in either direction is the same.

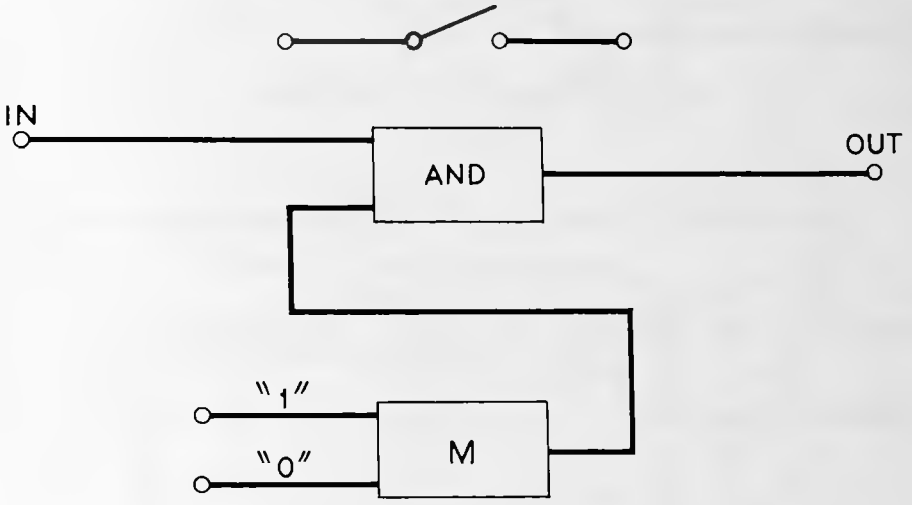
8.4 USE OF CIRCUIT FOR DELAY

The device described herein makes a useful delay element where only one digit of delay is required. In Figure 5 a chain of amplifiers are shown which are connected to one another through ringing transformers. If an input pulse is put into amplifier one, it will appear at amplifier two at time 2, amplifier three at time 3, etc. The circuit might be an important substitute for delay lines where long delays are required and delay lines themselves might be too bulky.

BIT STORAGE CELL
SYMBOLIC REPRESENTATION AND TYPICAL USE



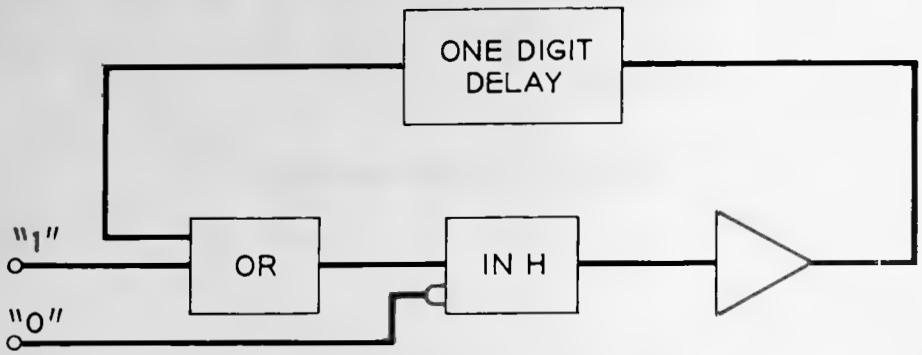
(a) SYMBOLIC REPRESENTATION



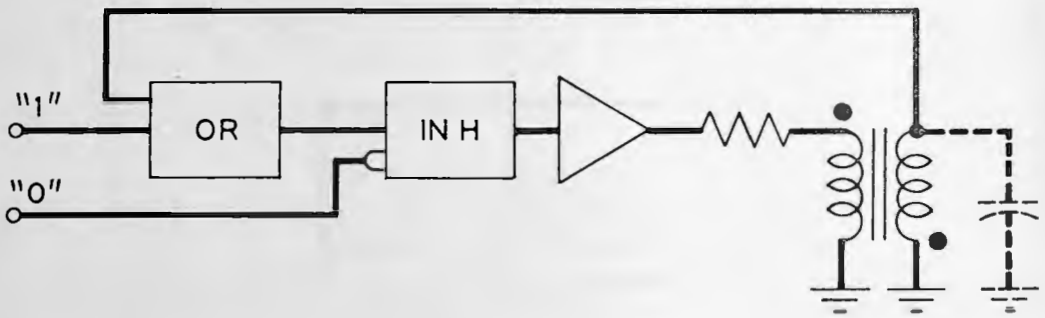
(b) SINGLE-POLE SWITCH

FIG. 1

BIT STORAGE CELL
POSSIBLE MECHANIZATIONS



(a) DELAY LINE STORAGE



(b) RESONANT CIRCUIT FOR DELAY

FIG. 2

CIRCUIT OF MEMORY CELL

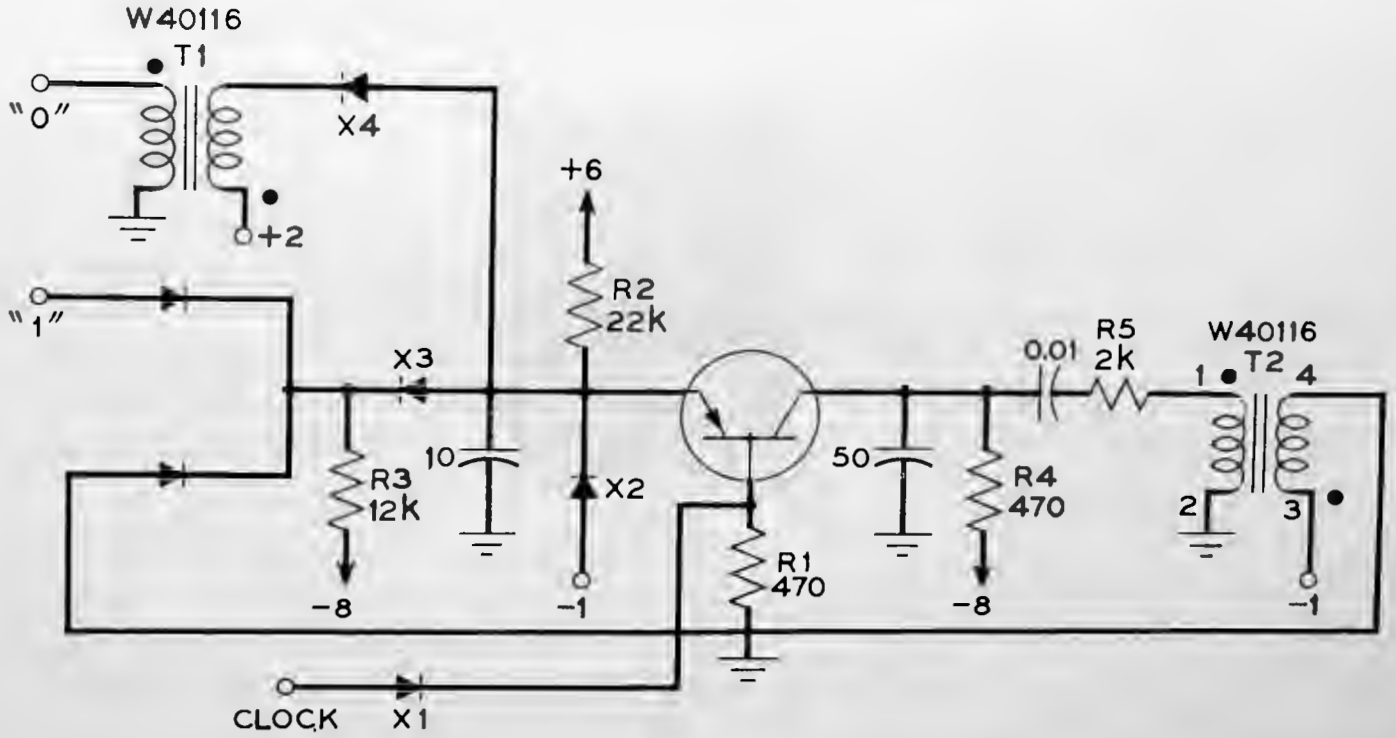


FIG. 3

WAVEFORMS IN STORAGE CELL

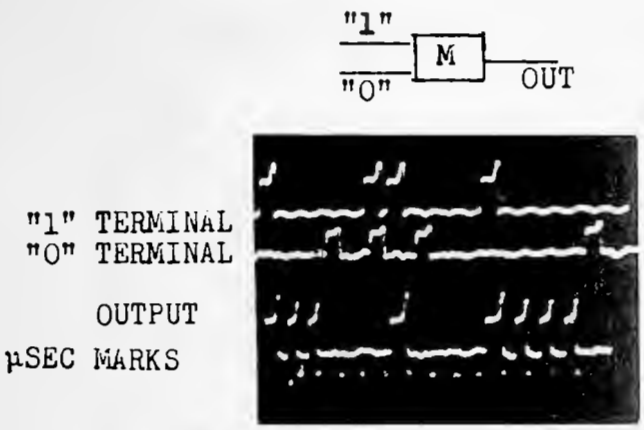


FIGURE 4

BIT STORAGE
DELAY SCHEME

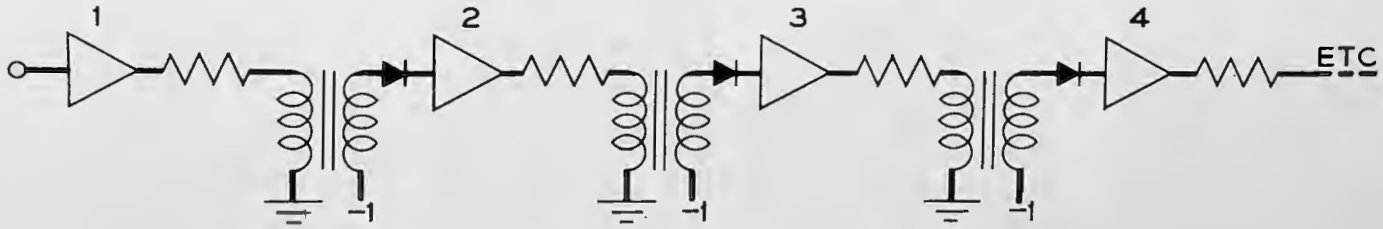


FIG. 5

A BINARY ADDER

9.1 INTRODUCTION

The basic amplifier of paragraph 5 has been combined with the appropriate diode logic circuits to produce a serial adder. The adder can accept inputs at a megacycle rate and uses six interchangeable plug-in amplifiers. The total battery drain is 0.264 watts.

9.2 BASIC AMPLIFIER PACKAGE

From the experience obtained with the word generator and delay line storage units described in preceding paragraphs it was apparent that a standard amplifier package would speed up the construction of new devices such as the adder described herein. The package that was evolved represents only a crude approximation to the package that would be used in production because time did not permit recourse to such techniques as auto-assembly and printed circuitry.

The circuit of the amplifier is shown in figure 9-1. The normal input and output are at terminals 2 and 12. Terminals 3 and 11 are provided so that direct connections can be made to the emitter and collector when desired.

Front and back views of the assembled amplifier are shown in photographs 192144 and 192143. The small cylindrical object shown in 192144 is the miniature transformer that is used for inhibition as described in the detailed discussion of the adder.

9.3 ADDER CIRCUIT

The adder can be regarded as a translator with three inputs, addend, augend and carry. It is a simple translator in that the output is a function only of the number of ones among its inputs. The logic circuits for performing addition were described in paragraph 2.41 above. Figure 9-2 shows the schematic of a complete adder in which each of the amplifiers in 9-2 has the same circuit as the one in figure 9-1.

If there is at least one "one" among the inputs either diode X9, X10, or X11 will pass current into amplifier B3 which will produce a half-microsecond output. If there is only one "one" among the inputs, 4 out of the 6 diodes X3, X4, X5, X6, X7, and X8 will not cut-off and amplifier B2 will not fire. Similarly, X1 or X2 or the internal diode X3 of amplifier B1 will continue to conduct and B1 will not fire.

If there are two ones among the inputs either X3 and X4 or X5 and X6 or X7 and X8 will be cut off which will put 3/4 of a milliampere through X12, X13, or X14 to fire amplifier B2.

If there are three "ones" among the inputs, X1, X2, and X3 (internal to B1) will be cut off and amplifier B1 fires.

When amplifier B3 is triggered a half microsecond pulse passes current through X22 which cuts off the diode X3 internal to B5. Amplifier B5 will fire when the diode X18 is cut off by the clock provided amplifier B2 has not fired. If amplifier B2 has fired the transformer T1 will invert the positive output pulse and cause X19 to conduct which will prevent B5 from firing. The purpose of feeding the clock pulse to X18 is to insure that B5 will not fire on the input of B3 until after B2 has had an opportunity to inhibit B5. It should be recognized that B5 cannot fire until X18, X19, and the X3 internal to B5 have all been cut-off. The diode X19 will permanently be cut-off, unless B2 fires, because it is returned to +2 volts.

The output of B5 is restored to a DC level of -2 volts by diodes X20 and R8. When B5 fires it will trigger B4 through X16 to produce a one for the sum. The firing of B1 will trigger B4 through X15 whether or not B5 has operated. The carry is obtained by feeding the output of B2 through a half microsecond delay line to trigger B6.

It was explained in paragraph 5 on the basic amplifier circuit that the phase of the base pulse supplied by the clock determines when an amplifier will be able to fire. The clock used has a 4 outputs, Φa, Φb, Φc, and Φd, each of which is delayed by 1/4 microsecond from the preceding one. The inputs to the adder are obtained from amplifiers whose bases were operated on Φc. The amplifiers B1, B2 and B3 are

operated on Φd to permit the Φc outputs to pass through the diode circuits. This results in the outputs of B1, B2 and B3 being delayed exactly 1/4 microsecond from the addend, augend and carry.

Amplifier B5 is operated on Φa to permit the outputs of B2 and B3 to rise and amplifier B4 is therefore operated on Φb. This results in the sum signal being delayed 3/4 microseconds from the addend, augend and carry. Delay line one is required to delay the output of amplifier B1 by 1/4 microsecond so it will be in phase with the output of B5.

The carry pulse has to be in phase with the addend, and augend which are Φc. Since 1/4 of a microsecond is lost in both B2 and B6 delay line 2 need be only 1/2 microsecond long.

The addend, augend, and carry inputs are obtained through capacitors (C3 of figure 9-1) and their DC level is set at a voltage below -1 1/2 volts by the internal diode X2 of amplifier B1. This diode X2 is returned to -1 1/2 volts whereas the diode X2 of each of the other amplifiers is returned to the normal value of -1. The reason for setting the DC level of the inputs at below -1 1/2 volts is to prevent the voltage drop across diodes X3, X4, X5, X6, and X7 and X8 from raising terminal 2 of B2 above -1 volt in the absence of signals. It should be noted that R4 is made 2.7 k rather than 2.2 k because it has 12 k, R3 of amplifier B1, in parallel with it whereas R5 and R6 are not shunted.

The input signal required to operate the adder can be computed by studying the addend requirement alone since the augend requirements are the same. In the absence of a signal the junction of R5 and X1 rides at a voltage below -1 1/2 volts determined by the drop across diode X1. Assuming that the drop across X1 is about 1/2 volt the current through R5 can be calculated as

$$\frac{8 - 2}{2.2}$$

or 2.73 milliamperes. The drop across X1 (a W.E. 400A varistor) for a 3 milliampere current will average 1/2 volt and this checks that the addend terminal will ride at -2

volts in the absence of a signal. To insure that the emitters of B1, B2 and B3 can be carried above their turning point (assumed to be no higher than ground) the addend terminal has to be carried to about +0.25 volts. The current required to do this is

$$\frac{8 + 0.25}{2.2}$$

or 3.75 milliamperes. The voltage change required is 2.25 volts. The input requirement on the adder is therefore a 2.25 volt pulse and a current of 3.75 milliamperes.

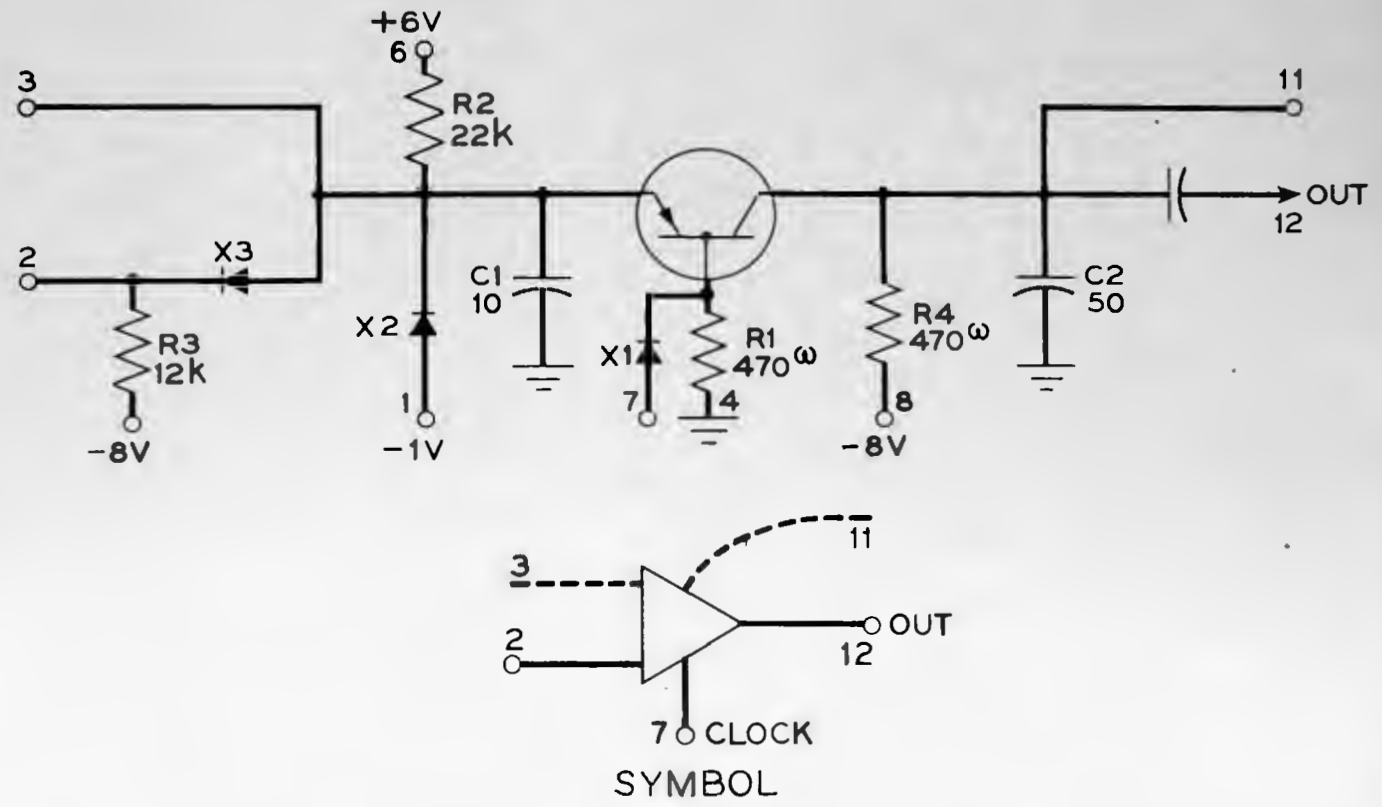
If the diodes available has been ideal (zero forward impedance) the input voltage would have been only 1 volt and the current only 2/3 of a milliampere. In this case if the peak point of transistors could be held to a spread of 0.25 volts the input requirement to the adder could be only 0.25 volt. It is hoped that the above discussion indicates how the forward impedance of diodes and the dispersion in transistor peak points affect signal power required.

Figure 9-3 shows the wave forms at various points in the adder for the addition of two typical numbers. Photograph 192399 shows a front and back view of the adder. Photograph 192400 shows a front and back view of the diode package designed for the logic circuits at the front end of the adder.

9.4 SUMMARY

A successful all transistor high speed adder has been described. This is the third such adder that the writer has built. The first two have been operating since May 1951 without any transistor replacements.

PACKAGED AMPLIFIER



SYMBOL

FIG. 9-1

BINARY ADDER

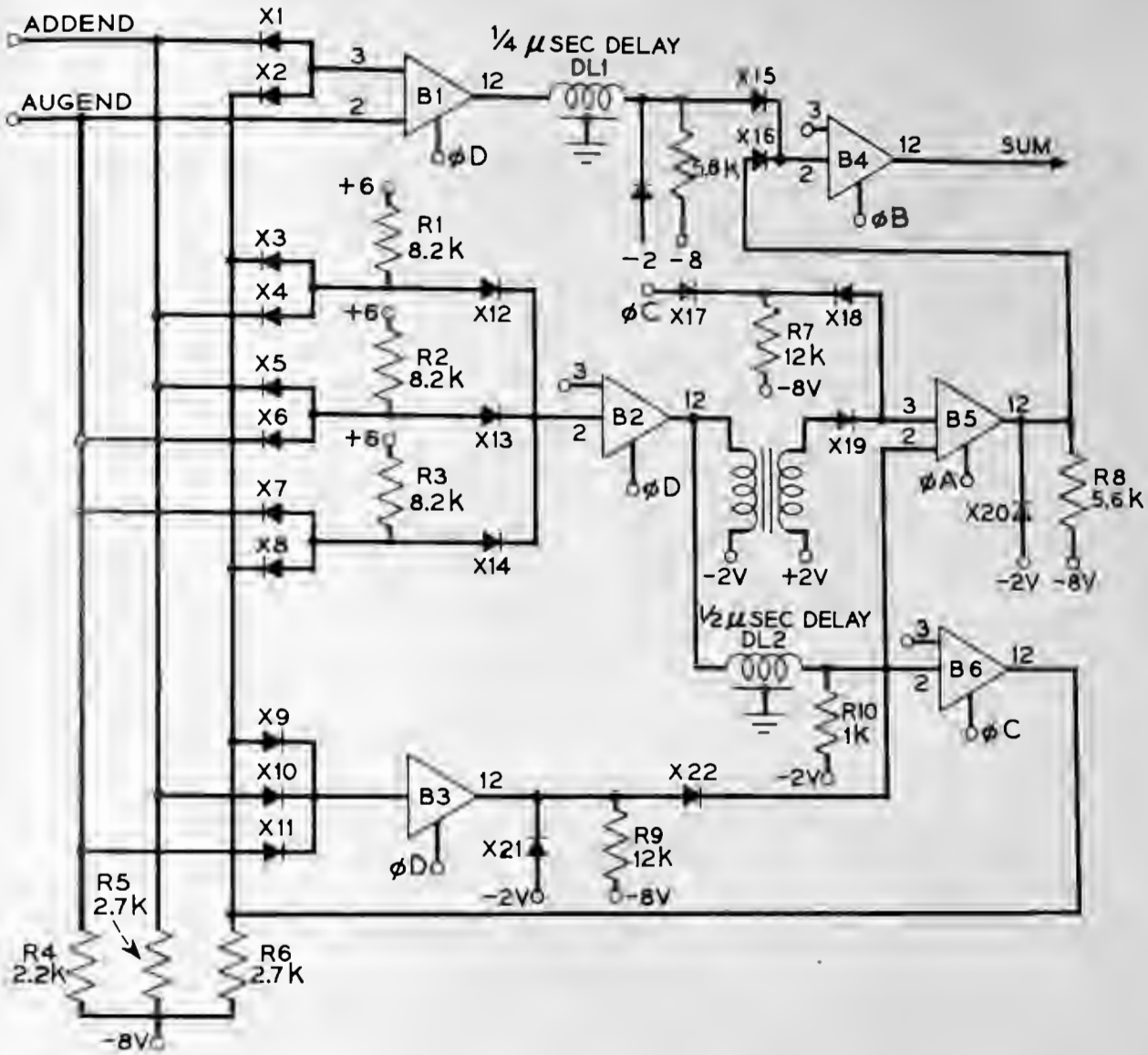
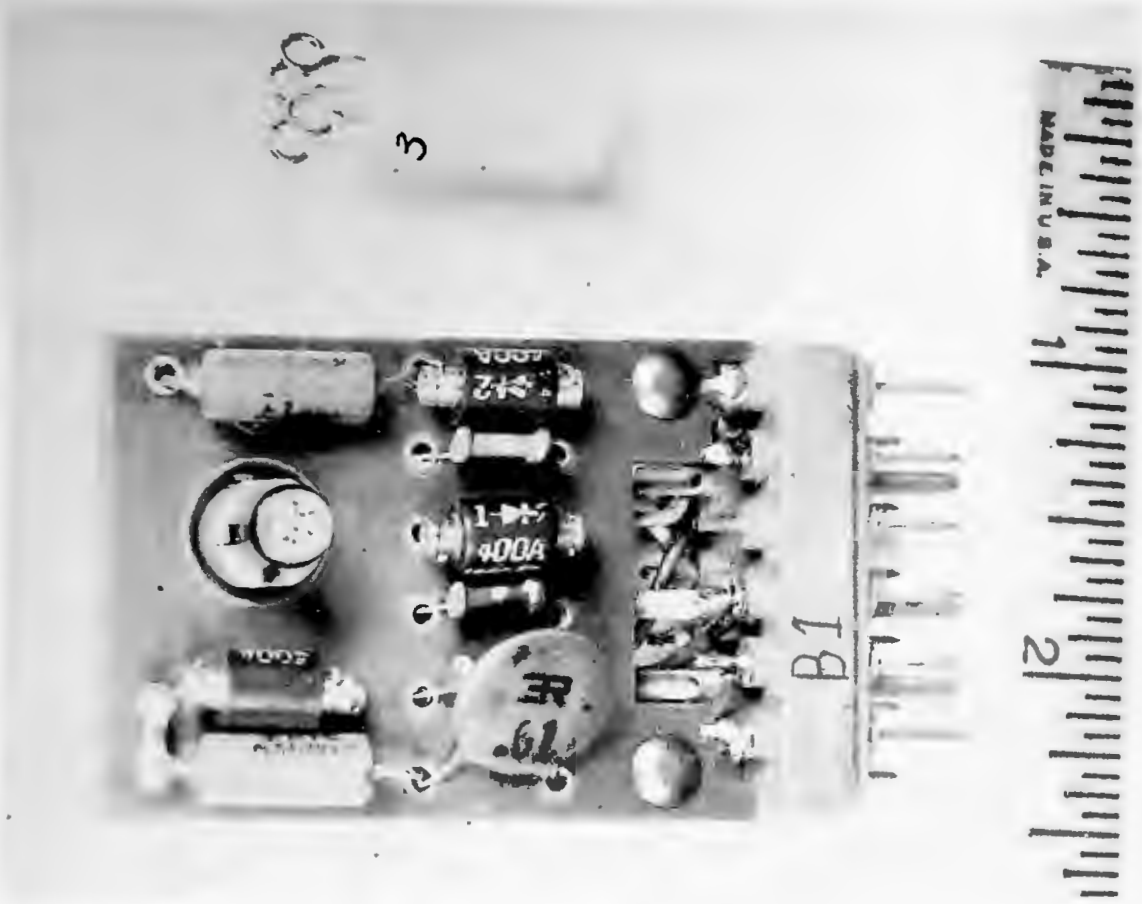


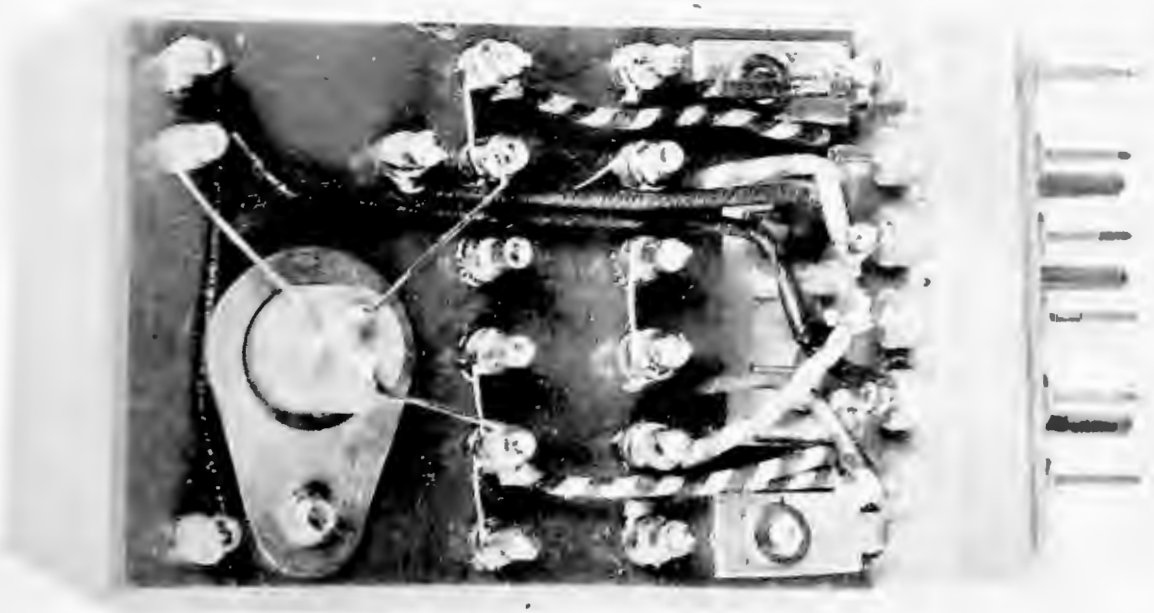
FIG. 9-2

B A S I C A M P L I F I E R F O R B I N A R Y A D D E R

FRONT VIEW

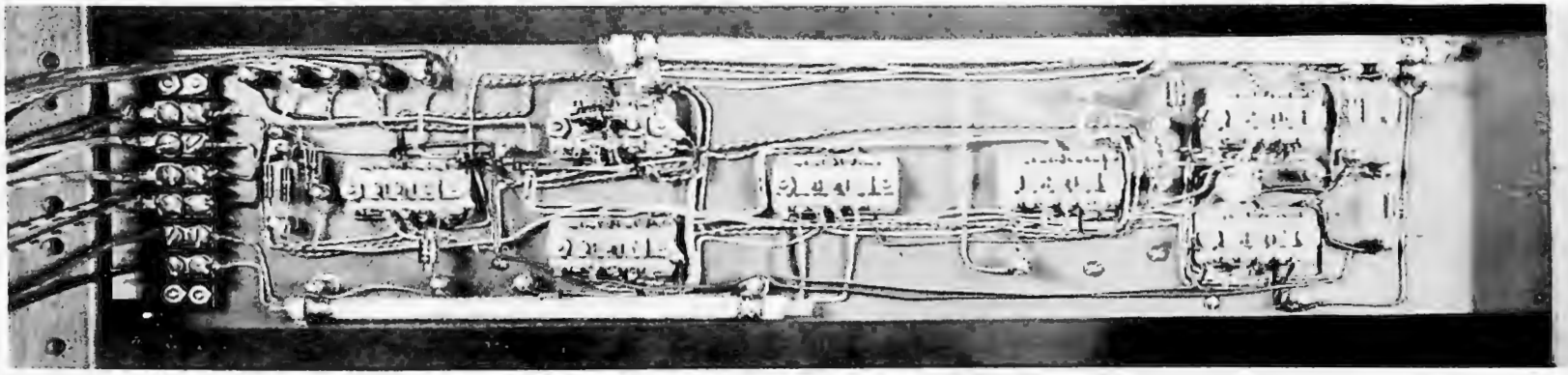


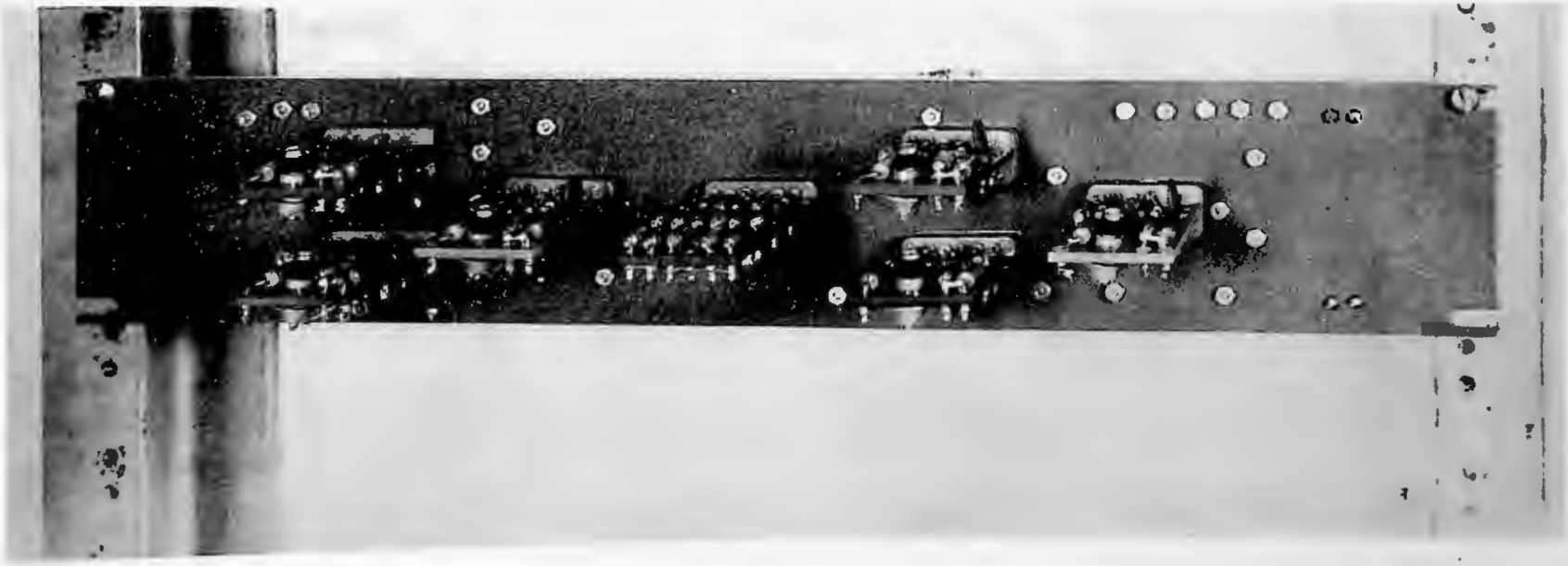
BACK VIEW



B I N A R Y A D D E R

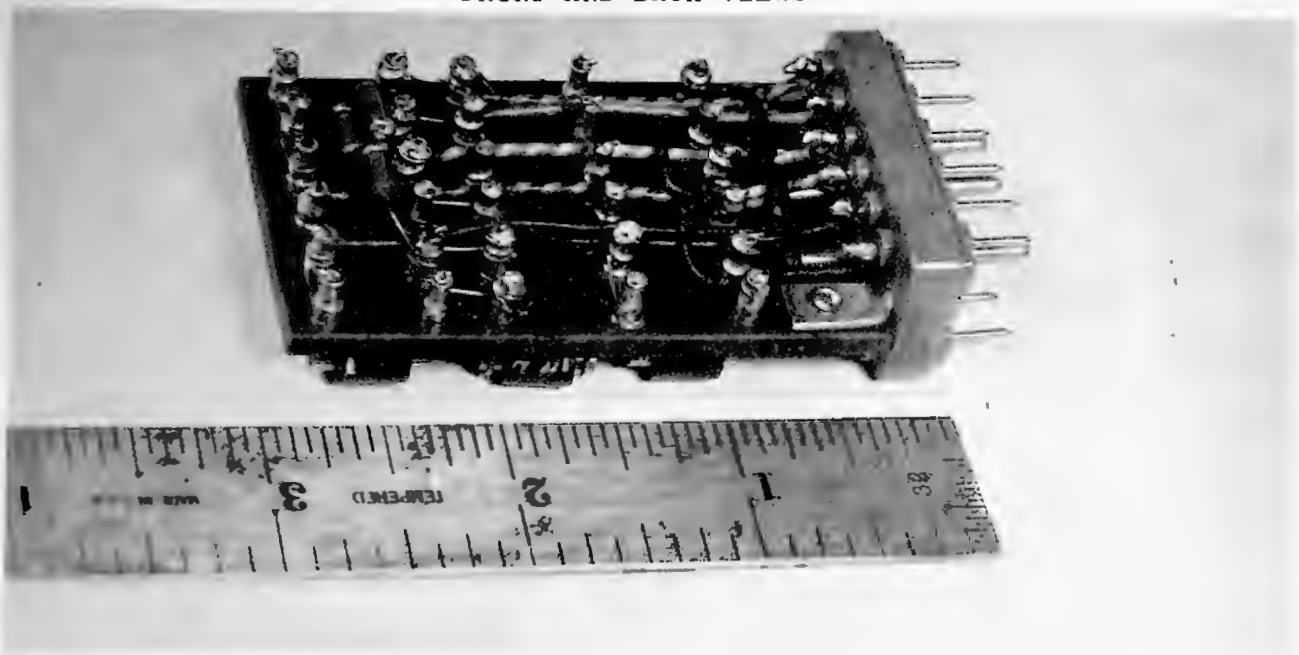
FRONT AND REAR VIEWS

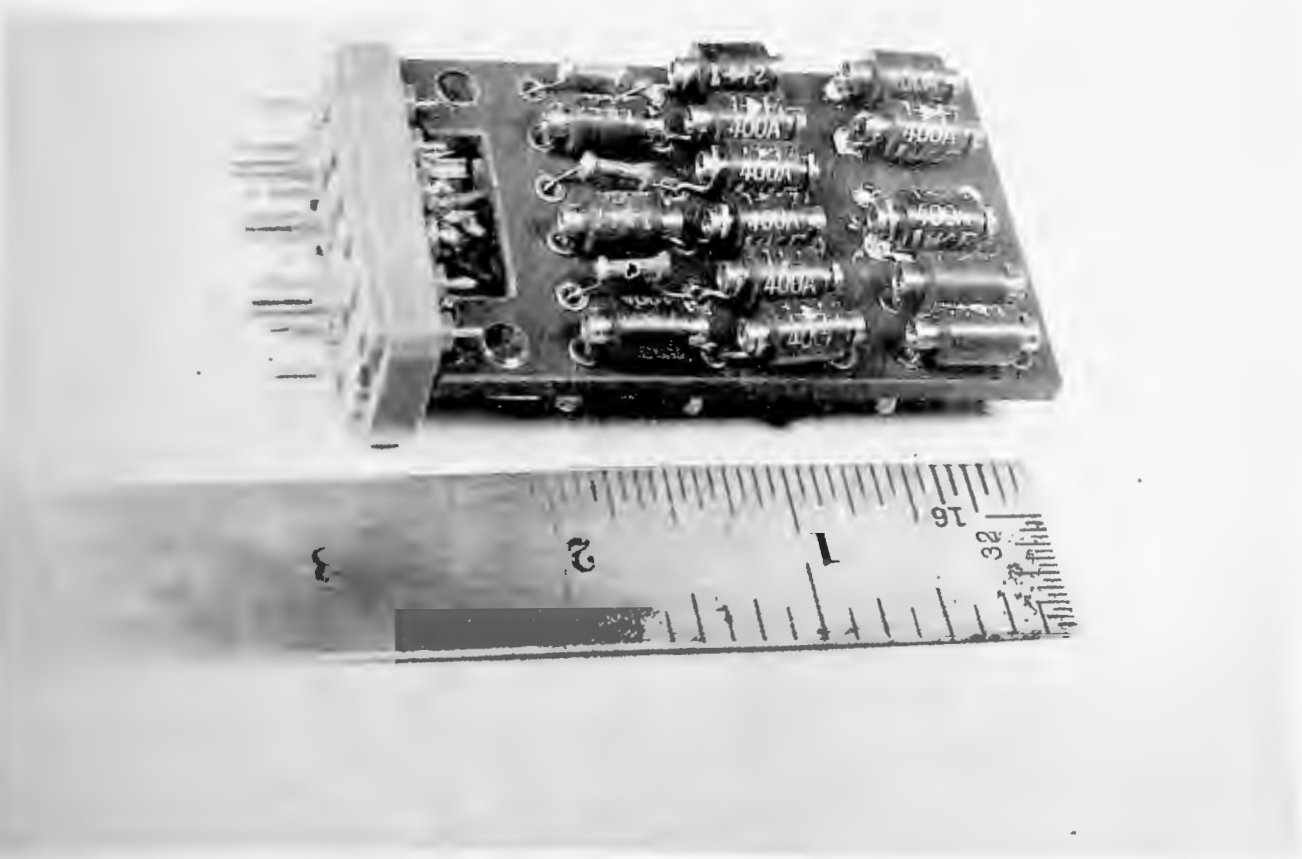




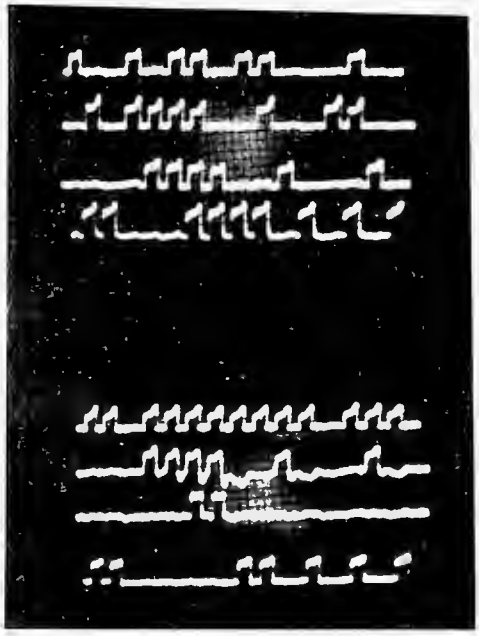
D I O D E P A C K A G E F O R B I N A R Y A D D E R

FRONT AND BACK VIEWS





ADDER WAVEFORMS



= 10010110110001

= 01011110010011

= 1100011110101 01
$2^0 2^1 2^2 2^3 2^4 2^5$ ETC

FIG.9-3

DATA ON EXPERIMENTAL TRANSISTOR TYPES

BTL M1679 POINT-CONTACT TRANSISTOR	J. J. KLEIMACK
W.E.CO. A1698 POINT-CONTACT TRANSISTOR	
BTL M1725 - BTL M1729 POINT-CONTACT TRANSISTOR	R. J. KIRCHER R. M. RYDER
BTL M1727 - BTL M1728 P-N JUNCTION DIODES	JOHN N. SHIVE
BTL M1734 POINT-CONTACT TRANSISTOR	B. G. FARLEY
BTL M1740 P-N JUNCTION PHOTOCELL	JOHN N. SHIVE
BTL M1752 NPN JUNCTION TRANSISTOR	K. D. SMITH
BTL M1754 - BTL M1755 P-N JUNCTION DIODES	JOHN N. SHIVE

produced outside the work performed under this contract

DATA ON EXPERIMENTAL TRANSISTOR TYPES

It is now possible on an exploratory development level, to make transistors to several sets of prescribed characteristics with good tolerances and satisfactory yields.

For applications involving one or more of the following features, transistors may be especially suitable.

1) Economy of Space - Minaturization

2) Economy of Power

3) Ability to Stand Shock and Vibration

4) Long Life

Current transistors have certain limitations as to frequency, power and temperature range. However, with better understanding and improved techniques, it is expected that these will be, at least partially, overcome. Some indication of what may be expected from transistors may be had from inspection of the tables below:

REPRODUCIBILITY

Point Contact Transistors

α	± 20%
r_c	± 30%
r_e	± 20%
r_b	± 25%
Cut-off frequency	± 30%
Noise figure	± 3db
Large signal characteristics	± 25%

PERFORMANCE

LOW-FREQUENCY GAIN	<45 db/stage	1752
VIDEO GAIN	20 db/stage	1729
I-F GAIN	20 db/stage	1734
GAIN CONTROL		1752
POWER - CLASS A	20-50 mw	1729,1752
CLASS B	400 mw	1729
CLASS C	1-3 watts peak	1689,1698
FREQUENCY RESPONSE	1-3 mc	1752
	7 mc	1729
	20 mc	1734
NOISE FIGURE, 1 kc	15 db	1752
	45 db	1768

LIFE	>70,000 hours	PT. CONTACT UNITS
TEMPERATURE RANGE	<80°C	1729
VIBRATION	>100 g	PT. CONTACT UNITS
SHOCK	~20,000 g	PT. CONTACT UNITS
UNIFORMITY	±30%	PT. CONTACT UNITS
SIZE	$\frac{1}{1000}$ cu. in.	1689
	$\frac{1}{100}$ cu. in.	1752
	$\frac{1}{50}$ cu. in.	CASED UNITS
MINIMUM POWER	1 microwatt	1752
	few milliwatts	1768
NEGATIVE RESISTANCE		1689, 1698, 1734
NEGATIVE FEEDBACK		1752

A more detailed discussion is given below.

Gain: At audio and carrier frequencies, gain up to about 45 db per stage is offered by the 1725 junction-type unit in grounded emitter connection. At video frequencies, gain of about 20 db per stage is offered by the 1729 point contact unit in grounded base connection. Even though the grounded emitter connection can give somewhat more gain, it is the

moment not recommended for point-contact transistors because of stability and variability difficulties. At high frequencies (up to about 20 mc/s) the 1734 point contact unit offers some possibility of gain of about 20 db per stage; but stability problems make this use more exploratory than the 1729 or 1752.

Gain Control: Transconductance proportional to operating current is offered by the 1752 units.

Power Output:

a) Undistorted Class A power output from either 1752 or 1729 is about 20 to 50 mw. The 1752 has higher efficiency, but only slightly more gain.

b) Class B output of about 400 milliwatts has been obtained from a pair of 1729's in push-pull connection, using duality-type circuitry. The collector power efficiency is good, but the overall efficiency may be debilitated if a resistor is used in the power supply circuit. Data on 1752's is not available, but they would probably give about the same output with somewhat better gain and considerably better overall efficiency.

c) Class C output with rectangular wave-forms in on-off trigger-type service permits efficiency well above 90%. Consequently 1698 units can give powers in the range 1 to 3 watts peak.

Relative to what can be obtained from tubes, power output is lower but efficiency is often considerably higher.

Frequency Response: The 1752 can be operated up to about 3 megacycles as a grounded base video amplifier: up to only about 200 kilocycles in the higher-gain grounded-emitter connection. The point contact 1729 units average 7 megacycles cutoff frequency; the point-contact 1734 units average about 20 megacycles, but with more feedback, so that the last are usefue primarily in trigger circuits. These frequency figures are quite variable at present, but by selection a range of ±30% can be met by 1729 units.

Noise: A noise figure of about 15 db at 1000 cycles for the 1752, with 1/f frequency response, is indicated by very preliminary data. Noise of point contact units is about 30 db higher.

Vibration and Shock: Vibration to at least 100 g does not affect units and shock to the order of 20,000 g does not damage them. In applications where vibration is severe, even point contact transistors may give noise performance superior to tubes.

Size: At present only the 1689 and 1752 units are available in beads. (Less than .01 cubic inch volume). However, in many applications the overall size is limited by passive components, even for cased transistors (1/50 cubic inch), which are about an order of magnitude smaller than comparable tubes.

Power Economy: Operation with only microwatts of input is possible with 1752 junction-type units. The 1768 point-contact units give 20% efficient oscillations at 35 milliwatts input, and will operate with less.

Low Voltage Operation: For 1752's a few tenths of a volt bias and a few microamperes of current will produce operation; for 1768's, a volt or two, with about a hundred microamperes, is needed.

Life: Extrapolated estimates give about 70,000 hours or more for the life of point-contact units under moderate operating conditions. Data are not available for junction units, but comparable diodes indicate even better performance is likely.

Uniformity: Units are held to limits which are in most cases within 20 to 30 per cent of nominal values for the type. These figures are comparable with the ordinary run of unselected vacuum tubes, and are sufficient for a useful degree of interchangeability. For example, the range of gain of 1729 grounded-base amplifiers is about ± 1.5 db from nominal.

Miscellaneous Circuit Properties:

a) Negative feedback for extreme uniformity and low distortion is applicable, especially to 1752 amplifiers.

b) A useful high-to-low impedance transformation, similar to the cathode-follower tube, is offered by the analogous connection of the 1752.

c) Negative resistance effects obtainable from point contact units, especially the 1689 and 1698, are very useful in trigger-type switching circuitry.

d) Stability of the operating point, particularly in transistors having alpha greater than one, requires attention to terminating impedances at all frequencies, including those outside the useful band, both high and low. In particular, self-biasing requires care because the bias polarity corresponds to positive rather than negative d-c feedback. To reduce variations in operating point, it is well to have the d-c resistance in the emitter lead much larger than that in the base lead.

Finally, there follows information on a number of transistor types which are either in pilot production or in an advance development stage.

DEVELOPMENT CHARACTERISTICS OF TRANSISTOR TYPE B.T.L. M1689

The M1689 bead transistor is a miniaturized transistor having electrical characteristics suitable for switching circuits. It consists of a molded bead, approximately one-eighth inch in diameter, with a heavy base lead and two fine wires as the emitter and collector connections. It is primarily for use in miniaturized molded circuits in which small size and direct connection into circuits by soldering is required. Application may also be made without potting.

An exploratory development specification for the transistor is attached.

EXPLORATORY DEVELOPMENT SPECIFICATIONS

Description: Transistor - Switching Applications

Ratings:	V_c	I_c	V_e	I_e	P_c	T_a	Note 1
Absolute	V	mA	V	mA	mW	°C	
Maximum	-50	-40	-40	+40	80		
Test Conditions:	-	-	-	-	-	25	

Dimensions: Dia. 1/8" Approx. Mounting Position: Any

Connections: Base - .025" Dia. Wire

Collector - Bent Lead - .005" Dia.

Emitter - Straight Lead - .005" Dia.

Test	Conditions	Min.	Max.
Handling Precautions	Note 2		
Aging Period	48 hrs., Note 6		
Vibration & Shock	Note 3		
OFF Collector Current	V_c = -35V dc; I_e = 0 Note 7	I_c:0	-2 mA dc
ON Collector Voltage (1)	I_c = -2 mA dc: I_e = 1 mA dc; Note 4	V_c:0	-3 V dc
ON Collector Voltage (2)	I_e = 3 mA dc; I_c = -5.5 mA dc	V_c:0	-4 V dc
Emitter Resistance	V_e = -10V dc	I_e:0	-0.2 mA dc
Alpha (1)	V_c = -30V dc I_e = +1.0 mA dc Note 5	α:1.5	
Alpha (2)	V_c = -30V dc; I_e = +0.5 mA dc Note 5	α:2.0	-
Alpha (3)	V_c = -30V dc; I_e = -0.1 mA dc Note 5	α: -	0.3
Open Circuit Input Resistance	V_c = -10V dc; I_e = +1 mA dc	r_{11}: -	800 Ohms
Open Circuit Feedback Resistance	V_c = -10V dc; I_e = +1 mA dc	r_{12}: -	500 Ohms
Open Circuit Forward Resistance	V_c = -10V dc; I_e = +1 mA dc	r_{12}:15,000	- Ohms
Open Circuit Output Resistance	V_c = -10V dc; I_e = +1 mA dc	r_{22}:10,000	- Ohms
Turn-Off Time	Note 8	0	1 microsecond

Note 1 In ward direction of current is taken as positive for both emitter and collector. Voltages are measured with respect to the base. Subscripts "c" and "e" refer to collector and emitter respectively. Voltages are taken on an absolute basis so that a voltage of $V_c = -40$ is greater than a voltage of $V_c = -30$ volts. DC ratings are on the basis of any duration longer than the order of 5 microseconds. Transients in excess of the ratings, but with durations much less than 5 microseconds, may not injure the unit; it is to be understood, however, that such service is experimental. Currents of 50-100 mA have been employed.

Note 2 The transistor should not be subjected to service in which there may be excessive transients as in plugging in and out with power on, but this may be done if under any case the ratings are not exceeded. Base contact should be made first. If solder connections are made, heat sink protection on the transistor side of the joint should be provided as with flat nose pliers.

Note 3 The same degree of ruggedness experienced with germanium diodes may be expected. Transistors have been shocked and accelerated to 20,000 G without failure.

Note 4 Limiting resistors of 100,000 ohms in the emitter circuit and 50,000 ohms in the collector circuit should be employed.

Note 5 Alpha tests (1), (2) and (3) are small signal measurements.

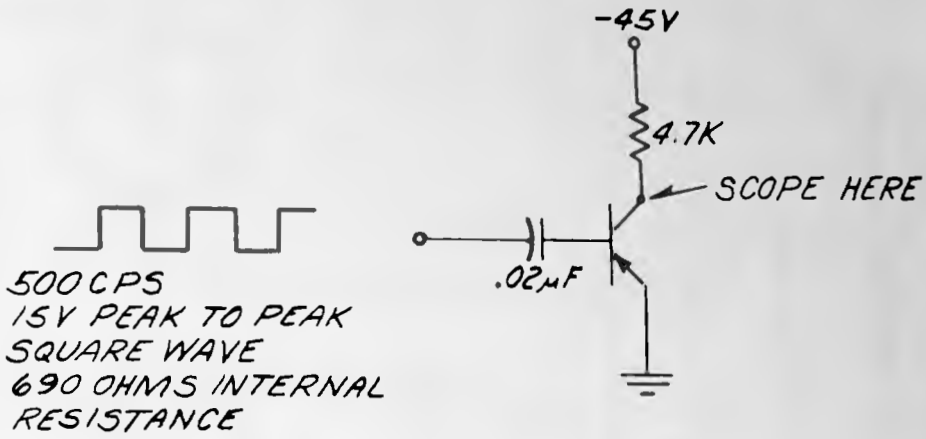
Note 6 All tests should be made after aging at 75mW, collector dissipation for 48 hours and after four temperature cycles of one-half hour duration between room temperature and 75°C.

Note 7 For the I_{co} test, -35 V from a constant voltage source should be applied for at least 5 minutes at the end of which period I_{co} is not to exceed -2 mA. I_c shall be not greater than -2 mA at $V_c = -20$ volts at 55°C.

Note 8 This test and requirement is measured in the circuit shown in Fig. 13-1 and is defined as the time required for the collector voltage to return to the value it had in the ON condition also illustrated. The transistor is in a "fully saturated" ON state when the turn-off step is applied. The emitter and collector currents are equal in the ON state. (This is a tentative method for measuring turn-off time.)

J. J. KLEIMACK

TEST FOR TURN-OFF TIME



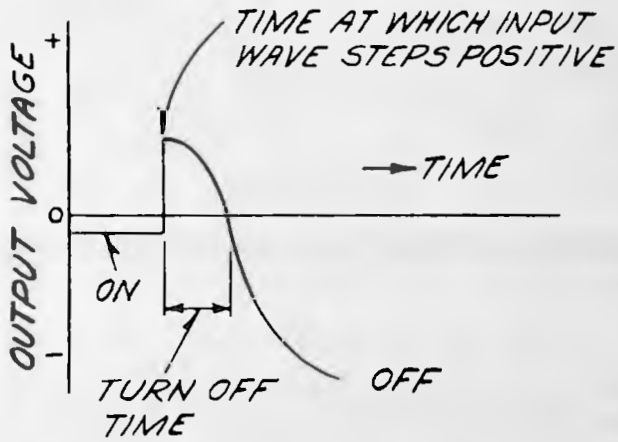
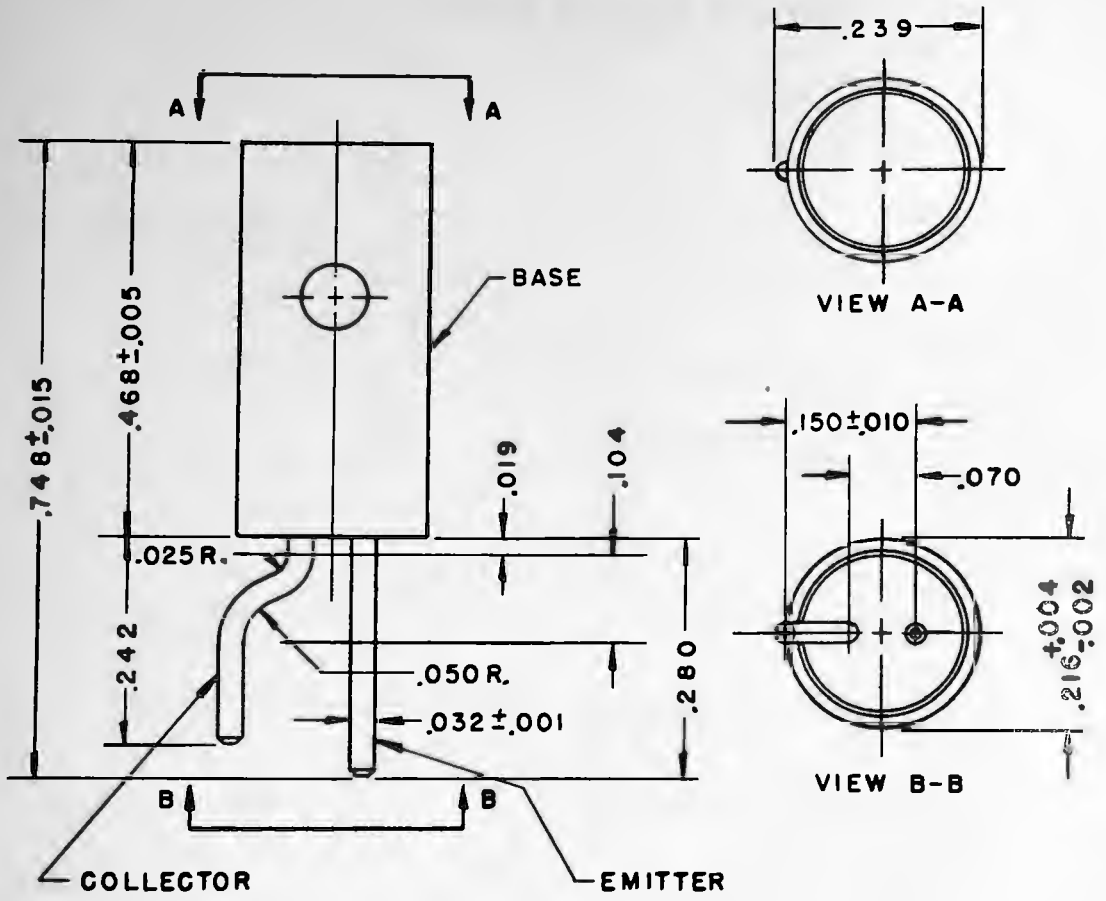


FIG. 13-1

ADVANCE TRANSISTOR DATA SHEET
WESTERN ELECTRIC TRANSISTOR
DEVELOPMENT MODEL A1698



DESCRIPTION

The A1698 is a point contact transistor triode in cartridge form. It is designed for use in switching circuits where the large-signal parameters of the active device are of primary interest. The external connections are arranged to fit the Cinch #EXP8672 or equivalent socket.

(Operating conditions and characteristics on Pages 2 and 3)

Issue 1, 9-7-51

A1698

GENERAL CHARACTERISTICS

MECHANICAL DATA

Mounting Position	any
Dimensions and Pin Connections - See Outline, Page 1	

MAXIMUM RATINGS

Collector Voltage (D-C)	-100 volts
Collector Current (D-C)	-15 milliamperes
Collector Dissipation	120 milliwatts
Emitter Voltage (D-C)	-100 volts
Emitter Current (D-C)	15 milliamperes
Ambient Temperature	130 F

TYPICAL OPERATING CONDITIONS AND CHARACTERISTICS

Stable End Points	
High Collector Impedance Condition	
Emitter Current (D-C)	0 milliampere
Collector Voltage (D-C)	-40 volts
Collector Current (D-C)	-1.5 milliamperes
Low Collector Impedance Condition	
Emitter Current (D-C)	1 milliampere
Collector Current (D-C)	-2 milliamperes
Collector Voltage (D-C)	-2 volts
Small Signal Characteristics	
Current Amplification	
When Collector Voltage=-30;Emitter Current=1ma	2
When Collector Voltage=-30;Emitter Current=0.05ma	3
When Collector Voltage=-40;Emitter Current=-0.1ma	0.2
Cut-off Frequency, Nominal [1]	5 megacycles
Base Resistance, Nominal[2] (r_{12})	
At V_c=-30;I_e=1.0 ma	200 ohms

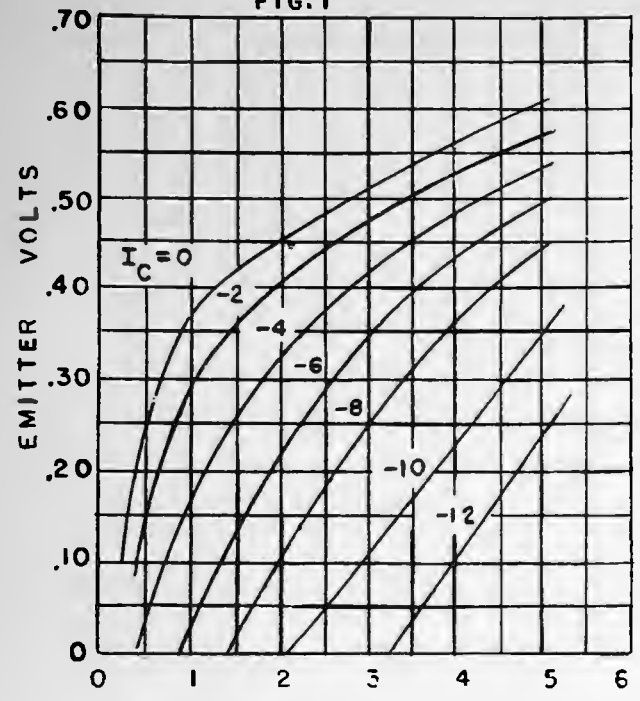
Note 1 Cut-off is defined as the frequency at which the current amplification is 3 db below its low frequency value.

Note 2 Base resistance is defined as the ratio of the voltage developed across the input terminals, to the collector current causing it. In other words, it is a measure of the positive feed back inherent in this type of device.

Operating Precaution

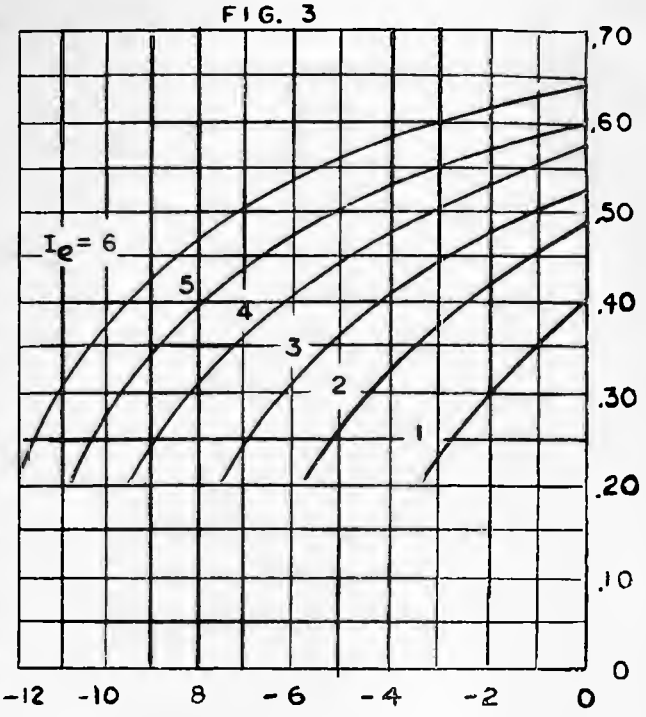
Where possible the transistor should not be connected to its circuit with bias voltages applied. Where this is not possible, the base connection should always be made first. Violation of this precaution may result in permanent damage to the transistor.

FIG. 1



EMITTER CURRENT (MA)

FIG. 3



COLLECTOR CURRENT (MA)

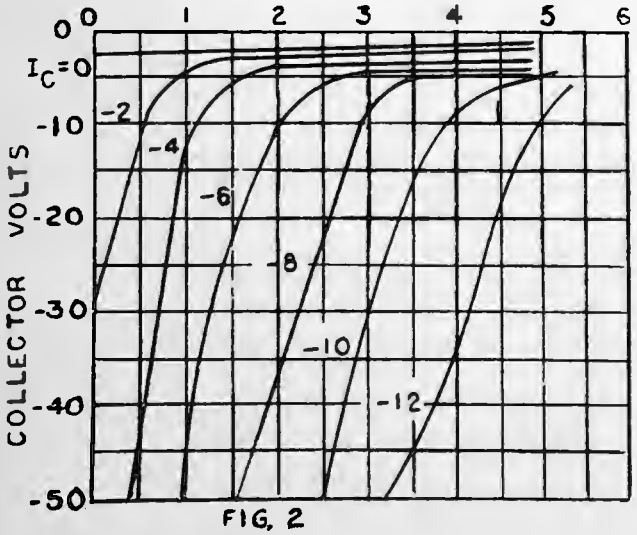


FIG. 2

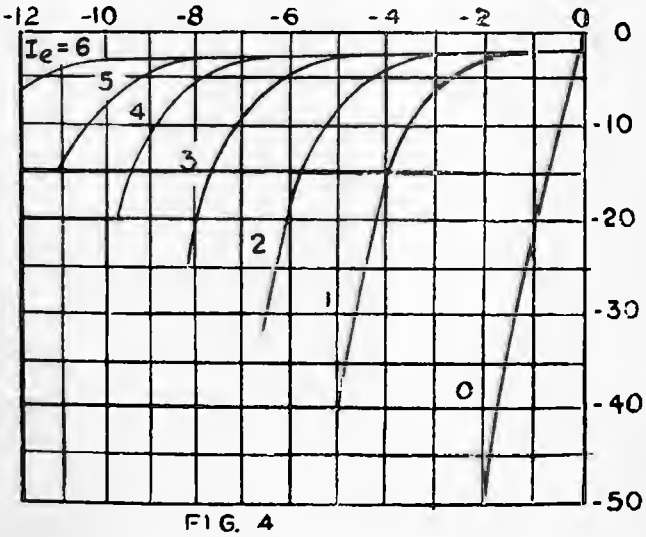


FIG. 4

EXPLORATORY DEVELOPMENT CHARACTERISTICS OF BTL M1725. BTL M1729 TRANSISTORS

Recent improvements in uniformity of transistors have permitted exploratory production of units with variations within about ±25% of nominal characteristics. While this degree of uniformity is somewhat inferior to that of standardized vacuum tubes, a useful degree of interchangeability is now realizable. Temporary specifications on two types of transistors, the BTL M1729 and the BTL M1725 are attached herewith.

For general purpose audio and carrier frequency application, the BTL M1729 unit has 20db gain, 50mw class A output power, 3 mc. high frequency cutoff, and 54db noise figure (at 1000 cycles).

For economical low voltage operation, the BTL M1725 unit has 18db gain, 20% output power efficiency, 48db noise figure (1000 cycles), with only 4ma at 5 volts required of the power supply.

At ambient temperatures approaching 60°C the properties of these transistors may be adversely affected, the units are therefore given reduced dissipation ratings above 50°C ambient, and are not recommended for operation above 80°C ambient temperature.

The predicted life of these units, when operated at 125 milliwatts dissipation, (before the gain falls 3db) is of the order of 70,000 hours.

The same degree of ruggedness experienced with germanium diodes may be expected. Accelerations up to 10,000 G in any direction can be tolerated by these transistors.

EXPLORATORY DEVELOPMENT SPECIFICATIONS

BTL M-1725 TRANSISTOR

Reference - JAN-1A

Description - Transistor

Maximum Ratings (Note 1)

V_c	I_c	V_e	I_e	P_c	P_e	Ambient
V	ma	v	ma	mW	mW	°C
-50	-20	+5	+15	200	50	50 (Note 4)

Test Conditions

-5.0	+1.5	20

Dimensions: Per Outline 2 Mounting Position: Any

Connections: Per Outline 2

Test	Conditions	Min.	Max.
Handling Precautions	Note 2		
Holding Period (48 Hours)	Note 5		
Vibration and Shock	Note 3		
Input Resistance R_{11}	Note 6	160	230 ohms
Feedback Resistance, R_{12}	Note 6	90	160 ohms
Forward Resistance, R_{21}	Note 6	12000	24000 ohms
Output Resistance R_{22}	Note 6	6000	12000 ohms
Current Amplification Factor, α	Note 6	1.8	2.6
Noise figure (1000 cycles)	Note 7	----	53 db
Frequency cutoff (alpha 3db down)		3.0	---- mc

Note 1 Subscripts c and e refer to collector and emitter respectively. Voltages are measured with respect to the base. DC ratings are on the basis of any duration longer than about 5 microseconds. Transients of shorter duration somewhat in excess of the ratings may not injure the unit; it is to be understood, however, that such service is experimental.

Note 2 The transistor should not be subjected to excessive transients such as may occur on plugging in or out with power on. This may be done if in any case the ratings are not exceeded; base contact should be made first. If solder connections are made, heat sink protection on the transistor side of the joint should be provided, as with flat-nose pliers. It is recommended that solder connections not be made to the base.

Note 3 The same degree of ruggedness experienced with germanium diodes may be expected. Transistors have been shocked and accelerated to 20,000 G with no difficulties, except that an acceleration of more than 10,000 G, applied in such a direction as to lift the points, may cause failure. Suitable packaging is used to protect the units against normal atmospheric conditions.

Note 4 For ambient temperatures from 50°C to 80°C, the maximum collector dissipation should be 100 milliwatts.

Note 5 All tests should be made after holding period. Ref. F3, JAN-1A.

Note 6 Operating point for measurement of circuit parameters is I_e = 1.5 ma, V_c = -5v.

Note 7 Noise figure varies inversely with frequency.

EXPLORATORY DEVELOPMENT DATA SHEET

BTL M-1725 TRANSISTOR

Triode Amplifier for Minimum Power Drain

MAXIMUM RATINGS not to be exceeded in continuous operation

Emitter Current, I_e	15 ma
Emitter Voltage, V_e	-50 v
Emitter Dissipation	50 mW
Collector Current, I_c	-20 ma
Collector Voltage, V_c	-50 v
Collector Dissipation	200 mW (up to 50°C ambient)
	100 mW (up to 80°C ambient)

AVERAGE OPEN CIRCUIT RESISTANCE PARAMETERS

Grounded base connection, with	I_ε	= 1.5 ma, V_c = 5.0 v
Input Resistance	R_{11}	= 195 ohms
Feedback Transfer Resistance	R_{12}	= 115 ohms
Forward Transfer Resistance	R_{21}	= 16,000 ohms
Output Resistance	R_{22}	= 8,000 ohms
Current Amplification Factor	a	= 2.1

TYPICAL LOW-DRAIN OPERATION CONDITIONS

Class A power amplifier, grounded base conditions

Emitter Current, I_e	1.5 ma
Emitter Dissipation	less than 1 mW
Collector Voltage, V_c	-5 v
Collector Current, I_c	-4 ma
Collector Dissipation	20 mW
Input Termination	500 ohms
Output Termination	1000 ohms
Available Power Gain	18 db
Power Output	4.5 mW
Cutoff Frequency (alpha 3db down)	5 mc
Noise Figure (1000 cycles)	48 db

Dimensions:

Outline drawing Fig. 14-1

EXPLORATORY DEVELOPMENT SPECIFICATIONS

BTL M-1729 TRANSISTOR

Reference - JAN-1A

Description - Transistor

Ratings:	V_c	I_c	V_e	I_e	Load	Dissipation P_e	Dissipation P_c	Ambient
Units	Volts	ma.	Volts	ma.	Ohms	Milliwatts		Temperature degrees, C
Maximum	-50	-20	+5	+15		50	200	50 (Note 4)
Test Conditions (small signal)	-30		+1.0		20,000			Approx. 30

Dimensions: Per Outline 2 Mounting Position: Any

Connections: Per Outline 2

Ref.	Test	Conditions	Min.	
	Handling	Note 3		
	Precautions	Note 2		
F3	Holding Period	48 hrs., Note 5		
	Vibration and Shock	Note 3		
	Collector Voltage	-30V, Emitter Current 1.0 ma.		
	Input Resistance,	R_{11}	150 ohms	225 ohms
	Feedback Transfer Resistance,	R_{12}	30	100
	Forward Transfer Resistance,	R_{21}	20,000	40,000
	Output Resistance,	R_{22}	10,000	20,000
	Current Amplification Factor,	α	2.0	3.0
	Cutoff Frequency, f_c		4.0	--mc
	Noise Figure (1000 cycles)	(Note 7)	--	56.0 db

Note 1 Subscripts c and e refer to collector and emitter respectively. Voltages are measured with respect to the base. DC ratings are on the basis of any duration longer than about 5 microseconds. Transients of shorter duration somewhat in excess of the ratings may not injure the unit; it is to be understood, however, that such service is experimental.

Note 2 The transistor should not be subjected to excessive transients such as may occur on plugging in or out with power on. Such plugging in or out may be done if in any case the ratings are not exceeded; base contact should be made first. If solder connections are made, heat sink protection on the transistor side of the joint should be provided, as with flat-nose pliers. It is recommended that solder connections not be made to the base.

Note 3 The same degree of ruggedness experienced with germanium diodes may be expected. Transistors have been shocked and accelerated to 20000 G with no difficulties, except that an acceleration of more than 10000 G, applied in such a direction as to lift the points, may cause failure. Suitable packaging is used to protect the units against normal atmospheric conditions.

Note 4 For ambient temperatures from 50°C up to 80°C, the maximum collector dissipation should be 100 milliwatts.

Note 5 All tests should be made after holding period. Ref. F3, JAN-1A.

Note 6 Operating point for measurement of circuit parameters is I_e = 1.0 ma, V_c = -30v.

Note 7 Noise Figure varies inversely with frequency and is measured at I_e = 0.5 ma, V_c = -8v.

EXPLORATORY DEVELOPMENT DATA SHEET

BTL M-1729 TRANSISTOR

Triode Amplifier for Audio and Carrier Frequencies

MAXIMUM RATINGS not to be exceeded in continuous operation

Emitter Current I_e	15 milliamperes
Emitter Voltage, V_e	-50 volts to +5 volts
Emitter Dissipation	50 milliwatts
Collector Current, I_c	-20 milliampers
Collector Voltage, V_c	-50 volts
Collector Dissipation	200 milliwats (up to 50° ambient)
Collector Dissipation	100 milliwatts (for ambient up to 80°C)

AVERAGE OPEN CIRCUIT RESISTANCE PARAMETERS, GROUNDED BASE CONNECTION AT THE OPERATING POINT I_c = 1.0 ma, V_c = -30 volts

Input Resistance	R_{11} = 190 ohms
Feedback Transfer Resistance	R_{12} = 75 ohms
Forward Transfer Resistance	R_{21} = 32,000 ohms
Output Resistance	R_{22} = 15,000 ohms
Current Amplification Factor	α = 2.5

TYPICAL OPERATING CONDITIONS: Small Signal Class A Amplifier

Emitter Current, I_e	1.0 milliamperes
Emitter Dissipation	Less than 1 milliwatt
Collector Voltage, V_c	-30 volts
Collector Current, I_c	5.0 milliamperes
Collector Dissipation	150 milliwatts
Input Termination	300 ohms
Available Power Gain	20 db
Load Resistance	15,000 ohms
Cutoff Frequency (alpha 3 db down)	5 Mc
Noise Figure (1000 cycles)	54 db

Dimensions:

Outline drawing Fig. 14-1

NOTE Noise Figure is inversely proportional to frequency. In operation at I_e = 0.5 ma, V_c = 8 volts, the noise figure improves to 48 db.

TYPICAL OPERATING CONDITIONS: Large Signal Class A Amplifier

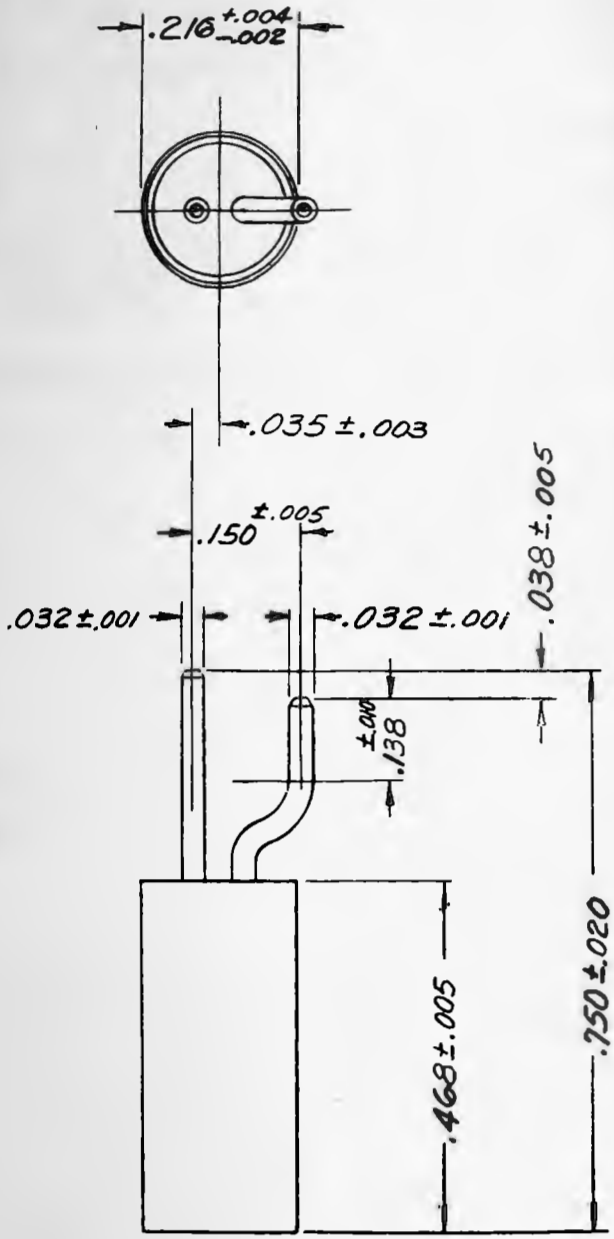
Emitter Current, I_e	2.0 milliamperes
Emitter Dissipation	Less than 2 milliwatts
Collector Voltage, V_c	-30 volts
Collector Current, I_c	7 milliamperes
Collector Dissipation	210 milliwatts
Available Power Gain	18 db
Load Resistance	5000 ohms
Power Output	50 milliwatts

R. J. KIRCHER

R. M. RYDER

OUTLINE DRAWING OF TRANSISTOR

APPLICABLE TO TYPE Nos	
A 1698	
A 1723	
M 1725	
M 1729	
M 1734	



STATIC CHARACTERISTICS OF M1725 TRANSISTOR

FIG. 1

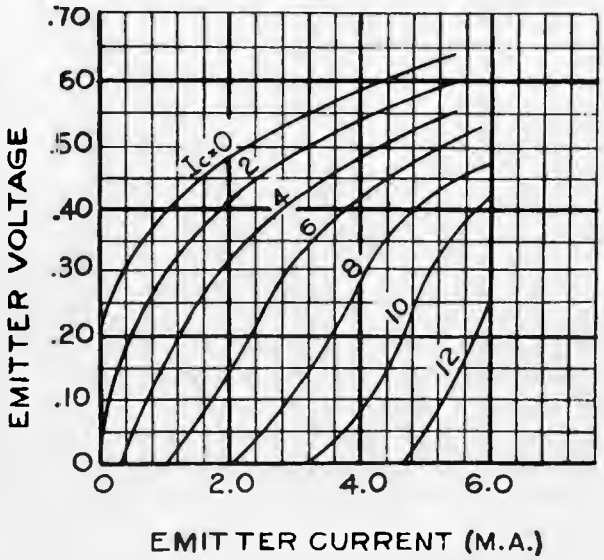


FIG. 2

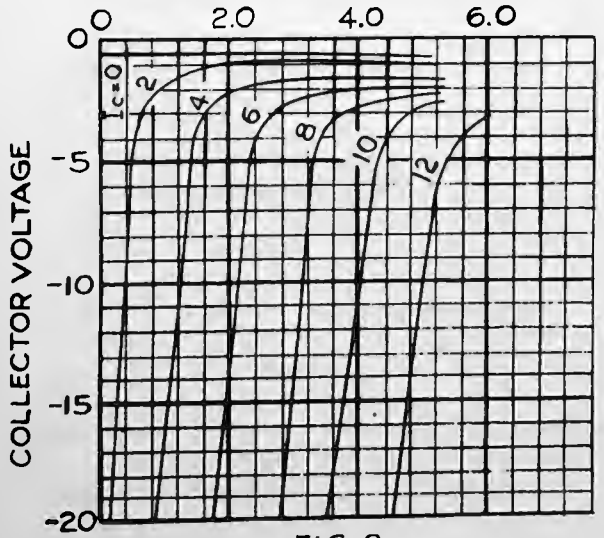


FIG. 3

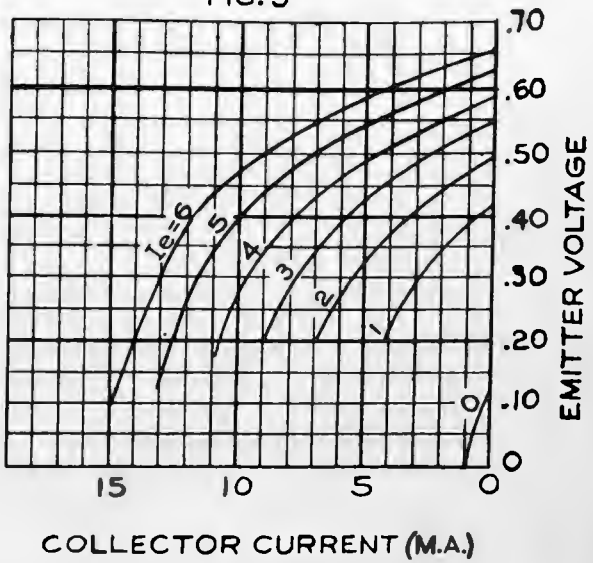
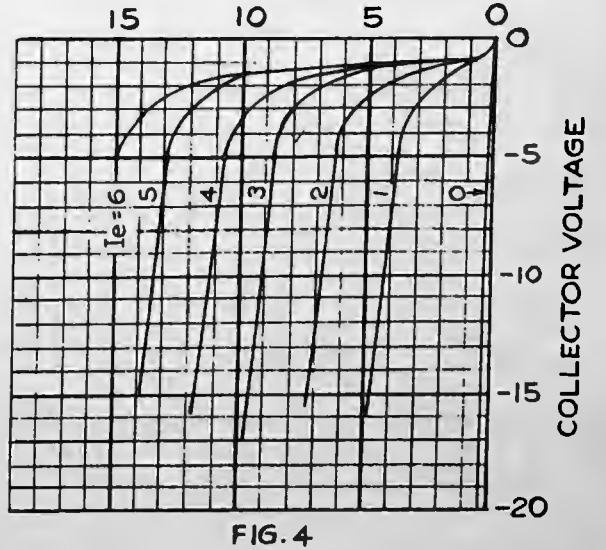


FIG. 4



B0755501

STATIC CHARACTERISTICS OF M1729 TRANSISTOR

FIG. 1

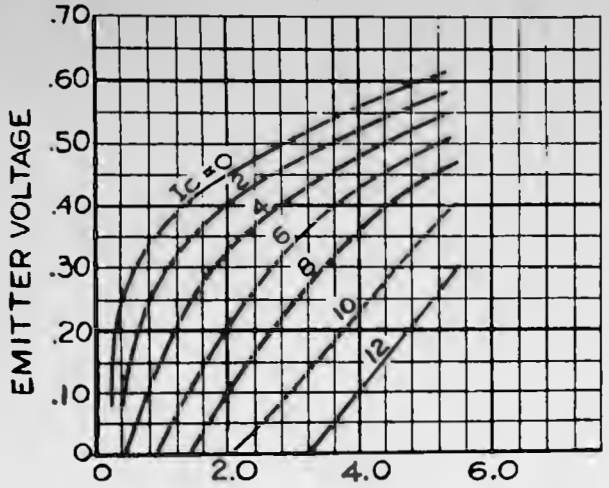


FIG. 2

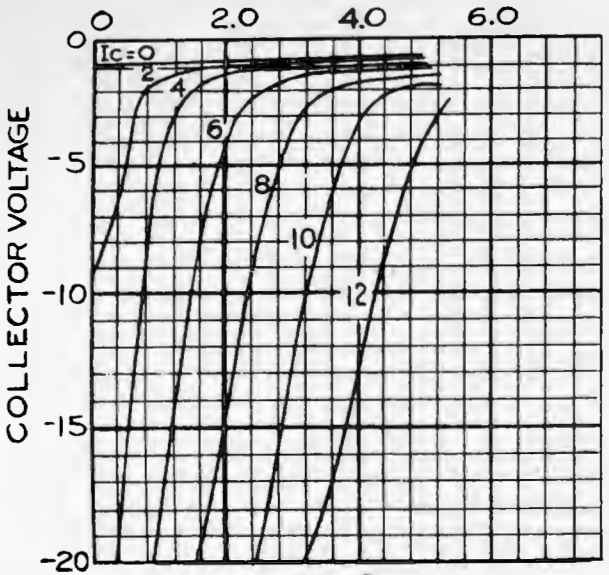


FIG. 3

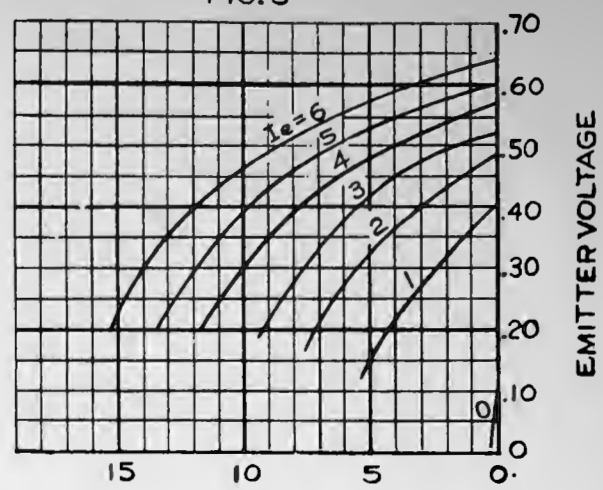
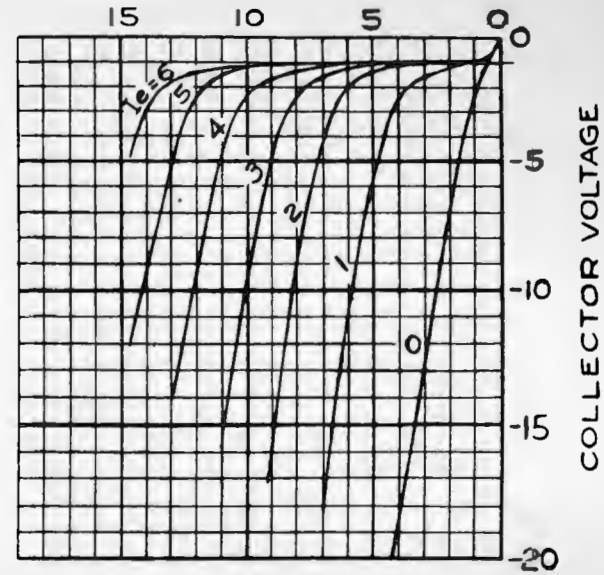


FIG. 4



BO 755502

EXPLORATORY DEVELOPMENT CHARACTERISTICS

BTL M1727 AND BTL M1728 P-N JUNCTION DIODES

The BTL M1727 and BTL M1728 P-N junction diodes, are designed for low forward impedance and high reverse impedance, respectively. Their particular advantages are low forward impedances, high reverse impedances, high back voltage capabilities, and small size. Their major limitation is comparatively low (100 KC) frequency cutoff. Summaries of the terminal characteristics of these two devices are given in the following sheets. -

BTL M1727 P-N JUNCTION DIODE

Reference: JAN-1A

Description: Germanium P-N junction diode, low forward impedance

Ratings:	epx	il	Io	Pp	TA
	v	mA	mA	mW	oC
Absolute Maximum	-	-	-	200 (note 1)	80
Test Conditions	-	-	-	-	25

Dimensions: as per outline Figure 15-1

Mounting position: any

Connections: red spot on shell marks positive terminal for easy flow

BTL M1727 P-N JUNCTION DIODE - CONT.

Ref.	Test	Conditions		Min.	Max.
-	Handling precautions	Note 2			
F-6b(1)	Vibration	-			
F-6c(1)	Forward current	Eb = 1Vdc	Ib:	10	- mAdc
F-6c(1)	Reverse current	Eb = 40Vdc Eb = 100Vdc	Ib: Ib:	- -	10 µAdc 20 µAdc
-	Reverse current	Note 3			
F-6c(1)	Reverse voltage	Ib - 100µAdc	Vx:	150	- Vdc
-	Frequency cutoff	Note 4			
-	Noise	Note 5			

BTL M1728 P-N JUNCTION DIODE

Reference: JAN-1A

Description: Germanium P-N junction diode, high back impedance

Ratings:

	epx	il	Io	Pp	TA
Absolute	v	mA	mA	mW	°C
Maximum	-	-	-	200 (note 1)	80
Test Conditions	-	-	-	-	25

Dimensions: as per outline Figure 15-1

Mounting position: any

Connections: redspot on shell marks positive terminal for easy flow.

BTL M1728 P-N JUNCTION DIODE - CONT.

Ref.	Test	Conditions	Min.	Max.
-	Handling Precautions	Note 2		
F-6b(1)	Vibration	-		
F-6c(1)	Forward current	Eb = 1Vdc Ib:	3	-mAdc
F-6c(1)	Reverse current	Eb = 100Vdc Ib:	-	10 µAdc
F-6c(1)	Reverse current	Eb = 200Vdc Ib: Note 3	-	15 µAdc
F-6c(1)	Reverse voltage	Ib = 100µAdc Vx	300 -	Vdc
-	Frequency cutoff	Note 4		
-	Noise	Note 5		

BTL M1727 and BTL M1728

Note 1 For ambient temperatures above 50°C the maximum dissipation rating becomes 100 mW.

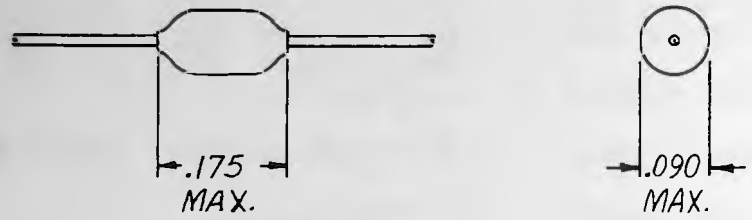
Note 2 When soldering to or bending the lead wires they should be gripped with pliers between the operation and the connections to the element.

Note 3 At an ambient temperature of 60°C the maximum reverse leakage current ratings are 10 times higher than those given for 25°C.

Note 4 Rectified dc output for continuous wave applications may fall off considerably in the range below 100 kc and prohibitively in the range from 100 kc to 1 MC. Step-function action in which the element is carried from reverse to forward is more rapid in response.

Note 5 The short circuit noise current in the band from 200 cps to 200,000 cps should not exceed 10^{-7} amp rms at a dc bias of 90 v reverse.

JOHN N. SHIVE



LEAD WIRES .010 NICKEL, LENGTH 1½ IN. MIN.

POLARITY INDICATED BY RED SPOT ON SHELL
AT POSITIVE END FOR EASY FLOW DIRECTION.

EXTERNAL DETAIL OF M-1727, M-1728, M-1754, M-1755
P-N JUNCTION DIODES.

FIG. 15-1

EXPLORATORY DEVELOPMENT CHARACTERISTICS FOR

B. T. L. M1734 TRANSISTOR

Numerous important system applications need a device combining good switching capabilities, and low voltage and power requirements with high speed operation. A transistor unit, designated the BTL M1734, is under development to exploit the possibilities of the point-contact type in these directions. Briefly, as an indication of performance, the BTL M1734 makes possible computer-type switching operations at rates of at least a megacycle per second with voltage supplies of 8-24 volts and average power drains per unit of the order of 25-50 milliwatts.

Complete specifications for the BTL M1734 have not yet been set, but development ones are attached.

A curve of current gain (α) versus frequency is taken with V_c = -7 v., and I_e = 0.5 mA. The frequency at which α has fallen to 0.707 of its low-frequency value (3 db down) is designated the frequency cutoff, f_c.

A maximum time of response at a low voltage (-7 v) is assured by requiring that $f_c \geq$ 10 mc/sec.

Output or collector large signal switching characteristics are fixed by cutoff and saturation requirements. These are 1) Cutoff,

$$I_e = 0,\ V_c = -7;\ I_c \leq -0.7 \text{ ma.}$$

(This value of I_c called I_{co})

2) Saturation:

$$I_e = 3 \text{ ma.},\ I_c = -4 \text{ ma.};\ V_c \leq -1.2 \text{ volts}$$

(This value of V_c called $V_{c3,4}$)

These specifications assure that a considerable proportion of a low collector supply voltage can be obtained as output voltage with low load resistances.

The cutoff input resistance is fixed by requiring that the cutoff resistance of the emitter to be greater than 100 K. at V_e = -10 V. Also, the large signal characteristic is held by fixing the minimum large signal current gain. This is done by introducing a 2 mA. step into the emitter (from cutoff) with -8 v. collector supply and 1000 ohm load. It is required that $\alpha \geq 1.8$ under these conditions, and as a further check on response time, this minimum value of α must be reached within 0.1 μsec. or less.

SPECIFICATIONS B. T. L. M1734

Reference - JAN-1A

Description: Transistor-High Speed Switching Applications

Ratings	V_c	I_c	V_e	I_e	P_c	T_a	Note 1
Absolute	V	mA	V	mA	mW	°C	
Maximum	-100	-40	-40	+40	120		
Test Conditions	-	-	-	-	-	25	

Dimensions: Per Outline Mounting Position: Any

Connections: Per Outline Fig. 14-1

Ref.	Test	Conditions	Min.	Max.
-	Handling Precautions	Note 2		
F-3	Aging Period	24 hrs, Note 4		
	Vibration & Shock	Note 3		
	OFF Collector Current	V_c = -7 Vdc; I_c: I_e = 0	0	-0.7 mAdc
	ON Collector Voltage	I_e = 3 mA V_c: I_c = -4 mA	0	-1.2 Vdc
	Emitter Resistance Reverse	V_e = -10 Vdc; I_e: I_c = 0	0	-0.1 mAdc
	Frequency Cutoff	I_e = 0.5 mAdc; f_c: V_c = -7Vdc Note 5	10	- mc sec.
	Alpha after 0.1 μsec	I_e = 2 mA α:1.8 V_c = -7 Vdc Note 6		

Note 1 Inward direction of current is taken as positive for both emitter and collector. Voltages are measured with respect to the base. Subscripts "c" and "e" refer to collector and emitter respectively. Voltages are taken on an absolute basis so that a voltage of $V_c = -40$ is greater than a voltage of $V_c = -30$ volts. DC rating are on the basis of any duration longer than the order of 5 μs. Transients in excess of the rating, but with durations much less than 5 μs, may not injure the unit; it is to be understood, however, that such service is experimental.

Note 2 The transistor should not be subjected to service in which there may be excessive transients as in plugging in and out with power on, but this may be done if under any case the rating are not exceeded. Base contact should be made first. If solder connections are made, heat sink protection on the transistor side of the joint should be provided as with flat nose pliers. Solder connections to the base are not recommended.

Note 3 The same degree of ruggedness experienced with germanium diodes may be expected. Transistors have been shocked and accelerated to 10,000 G with no difficulties, except, when applied in such a direction as to lift the points, there may be failure.

Note 4 All tests should be made after aging at 100 mW collector dissipation for 24 hours.

Note 5 Frequency at which oc, the small signal current gain, is down 3 db from its low frequency value is designated, f_c.

Note 6 For this test, a 2 mA step function is applied to the emitter from cutoff. The collector load is 1000 ohms.

B. G. FARLEY

EXPLORATORY DEVELOPMENT CHARACTERISTICS FOR B.T.L.

M1740 P-N JUNCTION PHOTOCELL ELEMENT

During the development of an encoding system the P-N junction photocell emerged as the preferred unit for use in the photodetection circuit of the angular position encoder. Reasons for this preference are: high reproducibility from unit ot unit in manufacture, low dark current, low noise, and small size.

Development data sheets and specifications are attached.

DATA SHEET ON B.T.L. M1740 PHOTOCELL

Maximum ratings Not to be exceeded in continuous operatior

Bias Voltage, Ebb		200 Vdc
Bias Current, Ib		10 mAdc
Dissipation, Pp	T < 50°C	100 mWdc
	T > 50°C	50 mWdc
Ambient Temperature, T		80°C

Nominal AC Equivalent Circuit Parameters

At operating point Ebb = 90 Vdc;

Flux = .00625 lm*

Junction Resistance, rj	10 megohms
Body Resistance, r_b	200 ohms
Transfer Constant, K	.035 milliamps / millilumen

- -

*Black body source at color temperature 2400°K, focussed into spot 0.5 mm in diameter on junction.

Typical Operating Condition: Small signal photoelectric transducer, 20°C

Supply Voltage	180 Vdc
Load Resistance	.5 megohm
Light Flux, Steady Component	.006 lm
Bias Current, Ib	190 μAdc
Bias Voltage, Ebb	90 Vdc
Dissipation, Pp	17 mWdc

Typical Operating Conditions: Light-operated on-off switch, 20°C

Supply voltage	135Vdc	135 Vdc
Load Resistance	1 megohm	0.5 megohm
Dark Current	<20μA	<20μA
Light Flux	>.005 Im	>.009 Im
Light Current	135μA	270μA

Short Circuit Noise Current*, in 1 cycle per sec. band at 1000 cycles per sec.

Ebb = 90 Vdc; Ib = 7μAdc (dark) $I_n = 2 \times 10^{-5}$ μAac rms

Ebb = 90 Vdc; Ib = 450μAdc (light) $I_n = 1 \times 10^{-4}$ μAac rms

Frequency Cuttoff, at which output is -3 db from low frequency value $f_{co} > 25$ kc

Spectral Range

Approximately 75% quantum efficiency for incident light in the visible and infra red to 1.6 micron wave length. See Figure 17-2.

- -

*The noise power per unit band approximates a $\frac{1}{f}$ relationship with frequency at which the noise is measured.

SPECIFICATIONS ON B.T.L. M1740 PHOTOCELL ELEMENT

Reference: JAN-1A

Description: photocell element - germanium p-n junction (visible and infra-red sensitive)

Ratings:	Ebb	Light Source	Light Flux	I_b	P_p	T
Absolute	Vdc	-	lm	mAdc	mWdc	°C
Maximum	200	-	-	10	100 Note 6	80
Test Conditions:	90	Note 1	.00625	-	-	20

Spectral Response: Note 3 Mounting: any position

Dimensions: per outline Figure 17-1.

Connections: per outline Figure 17-1.

- - - -

Note 1 The testing light source is a tungsten filament operated at a color temperature of 2400° Abs.

Note 2 This flux is focused into a small spot 1/2 mm in diameter, conetered on the p-n junction tract on the .040 x .125 face of the cell element.

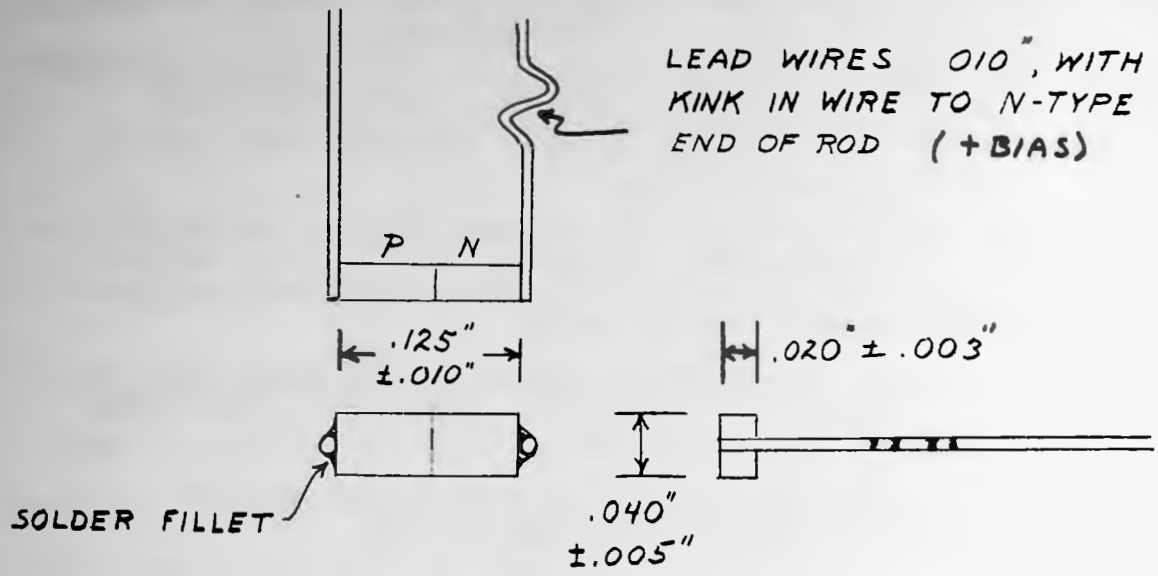
Note 3 The spectral response corresponds to a quantum efficiency of approximately 75% in the region from visible blue to 1.6 microns in the infra red. See Figure 17-2.

Note 4 The waxed surfaces of the element should not be handled. When the wires are to be bent, they should be gripped between the bend and the soldered electrodes, so as to place no stress on the latter.

Note 5 The short-circuit noise current in the band from 200 cps to 200,000 cps should not exceed 10^{-7} ampere rms at a dc bias of 90 volts.

Note 6 At ambient temperatures above 50°C, the maximum dissipation rating is reduced to 50mW.

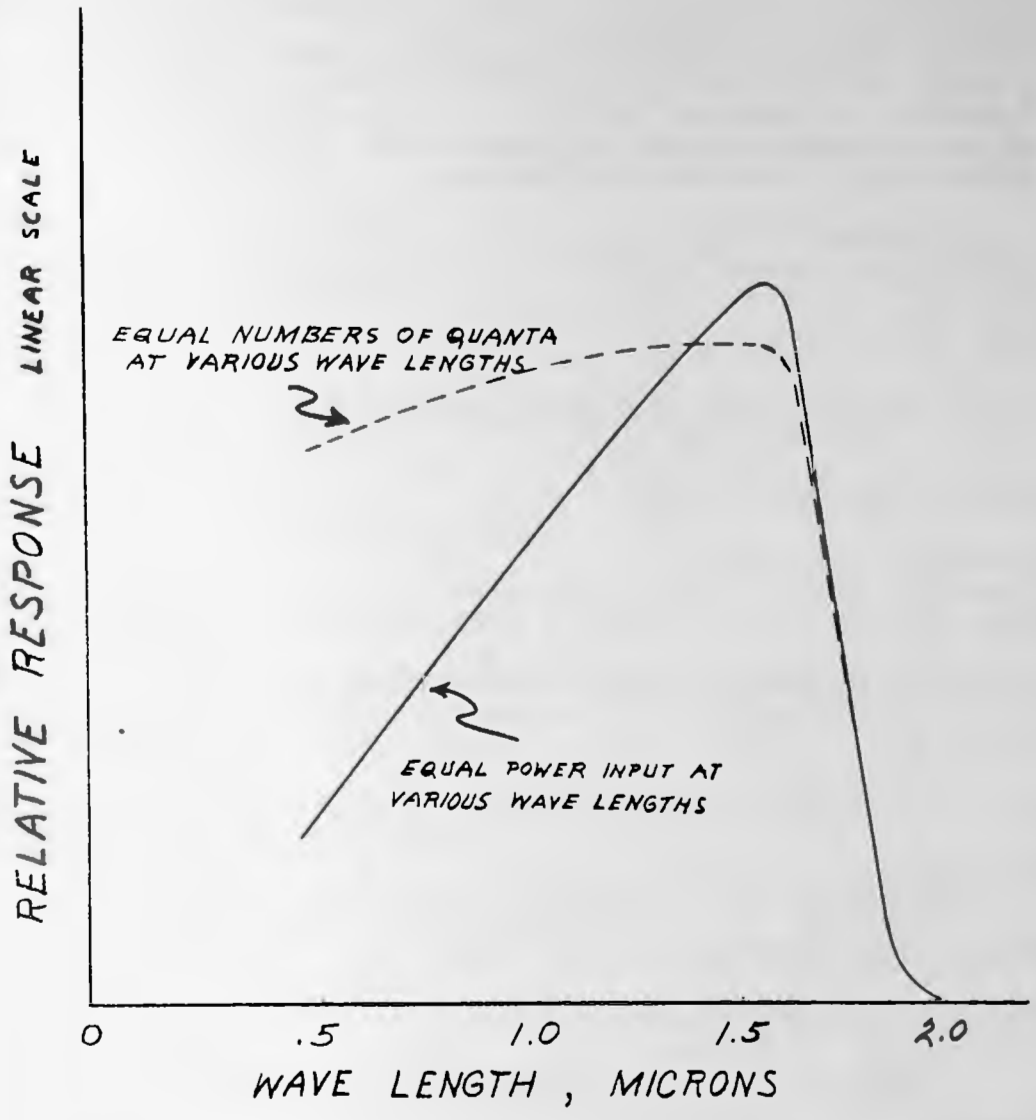
JOHN N. SHIVE



SURFACES OF ROD COVERED WITH
PROTECTIVE COAT OF WAX (NOT SHOWN)

DETAIL OF P-N JUNCTION ELEMENT

FIG. 17-1



M-1740 SPECTRAL RESPONSE

FIG. 17-2

EXPLORATORY DEVELOPMENT CHARACTERISTICS FOR THE BTL M1752 TRANSISTOR

The BTL M1752 transistor is an NPN junction transistor which can give high power gain at high efficiency, even at very low operating voltages and currents. In addition, the noise figure of the device is very much lower than that of point contact transistors of conventional design.

A limited number of development models of the BTL M1752 have been made for preliminary study in several possible circuit applications. Figure 18-1 shows outline dimensions and lead arrangement of these models, and a development specification is attached.

SPECIFICATIONS BTL M1752 TRANSISTOR

Reference - JAN-1A

Description - Transistor

Ratings:	V_c	I_c	V_e	I_e	Load	Dissipation P_e	Dissipation P_c	Ambient Temp. degrees C
Units	Volts	mA	Volts	mA	ohms	milliwatts		
Maximum	+50	5		5			50	50
Test Conditions								
Small Signal	+4.5			1.0				Approx. 25

Dimensions: per Outline Dwg. Fig. 18-1

Connection: per Outline Dwg. Fig. 18-1

Test	Conditions	Min.	Max.
Handling	Note 2		
Holding Period	48 hrs., Note 5		
Vibration and Shock	Note 3		
Collector Voltage		0	+50 volts
Input Resistance		-	1000 ohms

Test	Conditions	Min.	Max.
Output Resistance		100,000	
Current Amplification Factor	2	.95	1.0

Note 1 Subscripts c and e refer to collector and emitter respectively. Voltages are measured with respect to the base. D-c ratings are on the basis of any duration longer than about 5 microseconds. Transients of shorter duration somewhat in excess of the ratings may not injure the transistor.

Note 2 The transistor should not be subjected to excessive transients such as may occur on plugging in or out with power on. This may be done if in any case the ratings are not exceeded; base contact should be made first. When solder connections are made, heat sink protection on the transistor side of the joint should be provided, as with flat-nose pliers.

Note 3 The same degree of ruggedness experienced with germanium diodes may be expected.

Note 4 For ambient temperatures from 50° C up to 80° C the maximum collector dissipation should not exceed 25 milliwatts.

Note 5 All tests should be made after holding period. Reference F3, JAN-1A.

Note 6 Operating point for measurement of circuit parameters is I_e = 1 mA, V_c = + 4.5.

K. D. SMITH

OUTLINE DRAWING OF M1752 TRANSISTOR

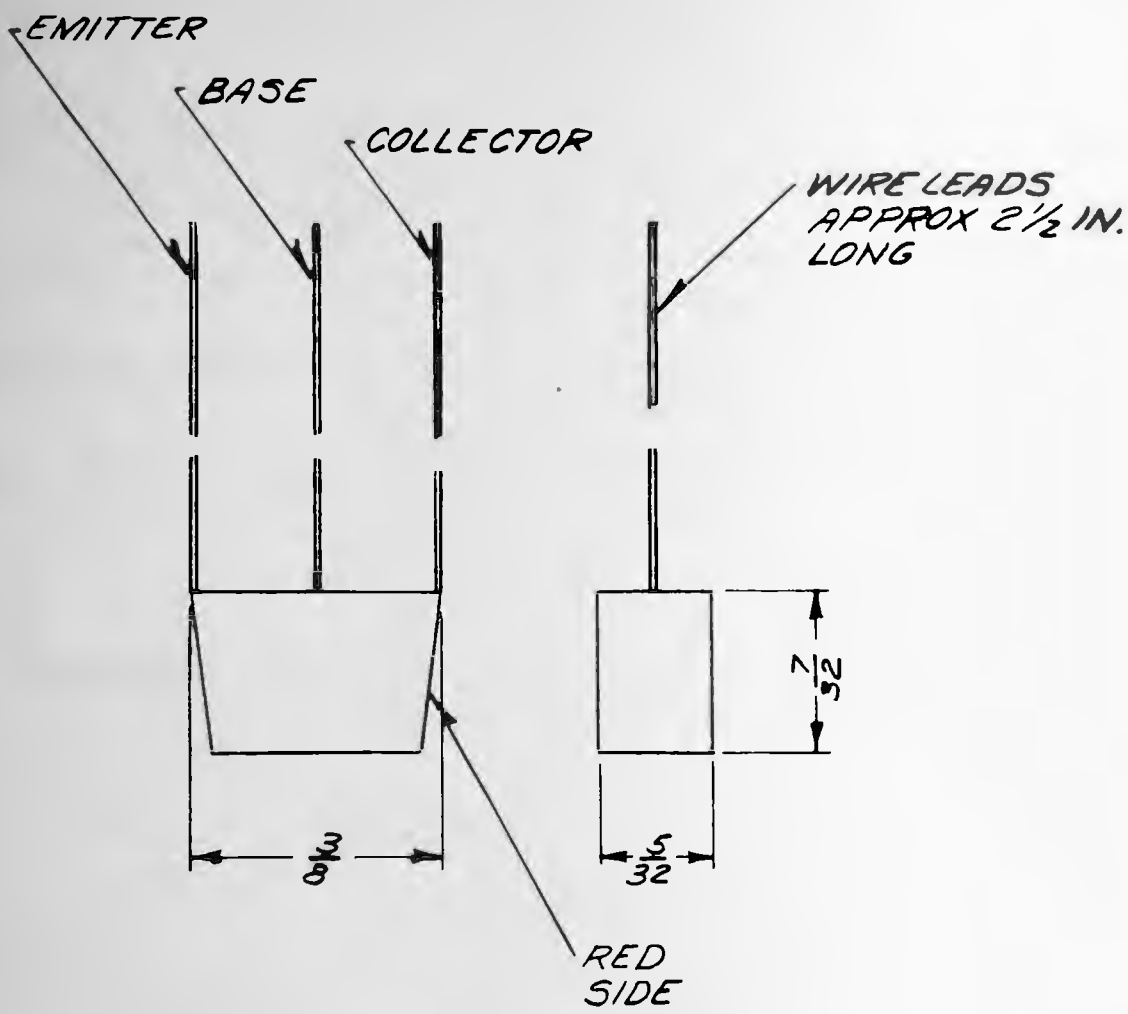


FIG. 18-1

EXPLORATORY DEVELOPMENT CHARACTERISTICS FOR BTL M1754 and M1755 P-N JUNCTION DIODES

The BTL M1754 is a new p-n junction diode which is intended for applications where unusually high forward-to-back ratios are required in the operating range below 50 volts. It does not have a high back voltage sustaining ability in the range above 50 volts. The BTL M1755 is a p-n junction diode for which the reverse characteristic exhibits a sharp "break-away" from the saturation current and proceeds to higher current along a Zener line of rapidly decreasing impedance. We thus have a current-saturation region of several megohms a-c impedance sandwiched between the forward characteristic region and the Zener region, both of which may have a-c impedances of the order of tens of ohms.

The equation of the Zener region portion of the reverse characteristic is:

$$I = V \times 10^{(11 - \frac{10^7}{E})} \text{ amps/sw.cm., where } E = \frac{V}{d},$$

the field in the barrier, with d being the barrier thickness. The reciprocal of the slope of the Zener characteristic, which gives the a-c impedance in the Zener region, is:

$$Z_z = \frac{V^2}{i(V + 2.3 \times 10^7 d)} \text{ ohms per sq. cm.,}$$

showing that this impedance is a function of the saturation voltage (which remains practically constant over this region), the current, and the barrier thickness. The properties of the barrier in this region are independent of temperature. However, the Zener region characteristic of a practical diode is temperature dependent insofar as the series resistance of the germanium body of the element is so dependent. If the diode is to be of much practical use the onset of the Zener region should occur below 100 volts in order to give a usefully long current range below the 200 mW maximum allowable dissipation point. To date Zener voltages have been found in the range from 12V up to over 100V.

For neither of these devices have enough samples been made and studied to justify the presentation here of full specifications. Accordingly there are given below only a list of those limits which have been tentatively established for measurement purposes.

M1754 TEST LIMITS

Forward current at 1 Vdc	25 mA min.
Reverse current at 25 Vdc	$.5 \times 10^{-6}$ max.
Reverse current at 40 Vdc	1.0×10^{-6} max.
Reverse voltage at 100×10^{-6} dc	75 V min.
Power dissipation	200 mW max.

M1755 TEST LIMITS

Forward current at 1 Vdc	5 mA min.
A-c resistance in reverse current saturation region	1 megohm min.
A-c resistance at 2 mA reverse (Zener region)	1000 ohms max.
Zener limiting voltage range	to 100 V max.
Power dissipation	200 mW max.

JOHN N. SHIVE

Santa Clara University
3 5098 10804669 2