THE BELL SYSTEM

Technical Journal

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

VOLUME XXXIX JANUARY 1960 NUMBER 1

THE BELL SYSTEM TECHNICAL JOURNAL

THE BELL SYSTEM TECHNICAL JOURNAL is published six times a year by the American Telephone and Telegraph Company, 195 Broadway, New York 7, N. Y. F. R. Kappel, President; S. Whitney Landon, Secretary; L. Chester May, Treasurer. Subscriptions are accepted at $5.00 per year. Single copies $1.25 each. Foreign postage is $1.08 per year or 18 cents per copy. Printed in U.S.A.

THE BELL SYSTEM
TECHNICAL JOURNAL

VOLUME XXXIX JANUARY 1960 NUMBER 1

The Ferreed — A New Switching Device

By A. FEINER, C. A. LOVELL, T. N. LOWRY
and P. G. RIDINGER

(Manuscript received September 14, 1959)

An experimental switching device is described that has the following properties: (a) sealed metallic contacts, (b) control times in the microsecond range, (c) coincident selection, (d) memory without holding power and (e) small size. The device, named the ferreed, may be used as a crosspoint in telephone switching networks of the space-separation type. The development of the ferreed is traced from a conceptual model, through realization of a practical model, to possible applications in switching networks. Two methods of coincident control are discussed, and three devices related to the conceptual ferreed are described briefly.

I. INTRODUCTION

The spectacular success of the electronic technology during the last decade, particularly in the field of computers and semiconductors, has provided a challenge to the communication industry. The problem of realizing the promise of electronics in telephone switching systems has been taken up by many communications laboratories.[1,2,3] Many different approaches were and are being taken. But no component of the telephone office has received a more varied treatment than has the usually most voluminous and costly part that permits the telephone customers to be interconnected — the switching network.

The multiplicity of solutions considered for the network problem can be grouped into three main categories:

i. space-separation networks combining electronic controls with conventional electromechanical switches;

ii. space-separation networks making use of electronic devices such as gas tubes or semiconductor elements as crosspoints; and

iii. time-division techniques that attempt to utilize fully the switching speeds of the electronic devices.

It is perhaps due to the complex nature of the network problem that none of these solutions has shown very clear and conclusive evidence of economic superiority. Solutions involving conventional electromechanical networks with electronic controls suffer from the time incompatibility between the two. While removing this difficulty, the purely electronic solutions bring with them handicaps of their own: the need for different and perhaps more costly telephone sets leading to difficult and unprecedented cutover procedures, new protection problems and limitations in transmission properties.

It was in this climate that a new class of switching devices was conceived. The ferreeds, as the new devices were named, are characterized by providing metallic contacts while being controllable at electronic speeds. Furthermore, the ferreeds can be left operated without holding power being required, have sealed contacts and can be selected by coincident current methods. These properties make the devices attractive for use as network crosspoints in electronic switching systems.

II. THE CONCEPTUAL DEVICE

2.1 *Ferrite + Magnetic Reeds = Ferreed*

The ferreed, as its name implies, comprises a magnetically hard ferrite member in combination with a magnetic reed switch. The switch consists of two soft magnetic reeds, which are responsive to the remanent magnetic field of the ferrite and which also serve as electrical contacts.

A conceptual model of the ferreed may be developed from the basic device shown in Fig. 1(a). In this simple structure a pair of overlapping reeds is fastened between the ends of a semicircular ferrite member. An exciting winding is uniformly distributed over the length of the ferrite. A short pulse of current applied to the exciting winding magnetizes the ferrite, which in turn induces unlike magnetic poles at the overlapping ends of the reeds. The reeds are mutually attracted, and, if the remanent flux is sufficient to overcome the spring force, the reeds will close, thereby establishing an electrical connection.

The connection is released by means of an exciting current pulse smaller in magnitude and of opposite polarity, which reduces the remanent flux of the ferrite member below that value necessary to hold the reeds closed. Since the force of attraction is proportional to the square of the flux between the reeds, the release current must not be allowed to exceed a certain maximum value; otherwise, sufficient reverse remanent flux will be produced to reclose the reeds. The rather precise control of release current required is therefore a fundamental limitation of this basic structure.

In Fig. 1(b) the ferrite hysteresis loop resulting from the asymmetrical excitation necessary for cyclic closure and release of the reeds is compared with the major loop produced by symmetrical cyclic excitation. Following a closure pulse, I_c, the operating point resides at C, and there is sufficient remanent flux to close the reeds. The reeds are held closed

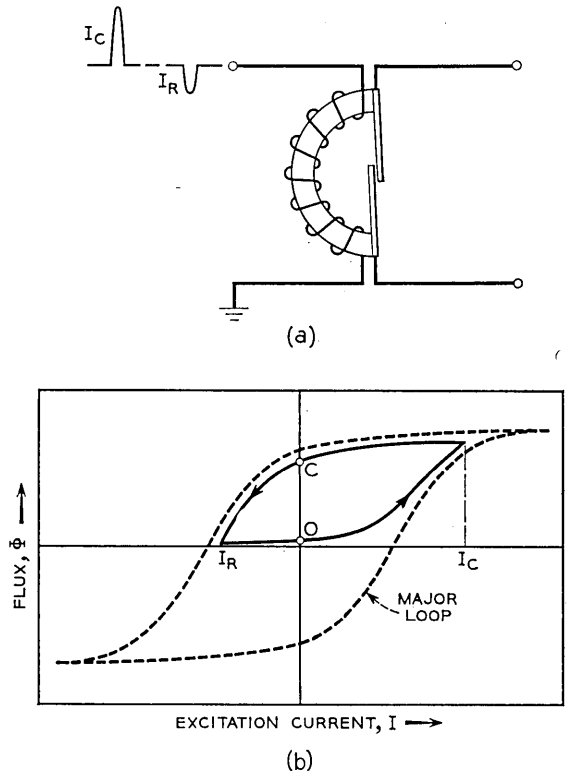

Fig. 1 – (a) The single-branch ferreed; (b) its magnetization characteristic.

until a release pulse, I_R, reduces the remanent flux to that at point o.

At first glance, the simple device of Fig. 1(a) appears similar to well-known types of reed relays. Perhaps most closely related is the magnetically latched reed relay, which is actuated by means of the additive fields produced by an exciting current and a permanent magnet. On closer examination, however, several important differences emerge. In the conventional magnetically latched relay, the remanent flux of the permanent magnet is only sufficient to hold the reeds closed and is essentially unaffected by the exciting current. Consequently, the closure excitation must be maintained throughout most of the reed closure interval, which may be in the order of one millisecond.

In contrast, the ferrite in the structure of Fig. 1(a) is controlled by the exciting current, and its remanent flux alone is sufficient to close the reeds. Furthermore, the ferrite may be switched from one magnetic state to another in a few microseconds, and as a consequence, this device may be readily controlled by currents applied for a like period of time.

Another important characteristic of the ferreed, contributed by the square-loop ferrite, is the well-defined threshold of magnetomotive force that must be exceeded before sufficient remanent flux is produced to close the reeds. This threshold can be made sufficiently large to mask variations in reed sensitivity, thereby permitting use of the device in coordinate arrays employing conventional methods of coincident current selection.

The ferreed affords an unusual opportunity to determine indirectly the state of its output contacts. Because the device incorporates internal magnetic memory, it can be interrogated without its electrical output circuit being disturbed. Methods for the nondestructive readout of the memory may be adapted from the magnetic core technology. However, the delay between ferrite switching and reed operation suggests a different approach. This method requires application of a short current pulse to the ferrite and observation of induced voltage in another winding to determine the memory state. The memory state can then be restored before the contact state is altered.

2.2 *Two-Branch Ferrite Structure*

The attractive features of the ferreed shown in Fig. 1(a) are considerably offset by the difficulty encountered in demagnetizing the ferrite to effect release. This basic limitation may be overcome by the addition of a second ferrite branch, as shown in Fig. 2. In this structure the reeds are closed by exciting the two windings so as to produce parallel mag-

netization in the two ferrite branches, and they are released by an excitation that causes series magnetization in the two branches.

The two-branch or parallel ferreed is adaptable to various methods of excitation, one of which is illustrated in Fig. 2. In this case, parallel magnetization is produced by equal coincident currents of the same polarity, whereas series magnetization is obtained by equal currents of opposite polarity. With equal parallel magnetization, the two-branch structure is equivalent to the closure condition in the single-branch structure of Fig. 1(a); with equal series magnetization, however, the two-branch structure appears demagnetized and the terminal magnetomotive force applied to the reeds is reduced to zero. It is significant that the condition of zero terminal magnetomotive force is obtained for arbitrarily large excitations above a certain minimum value and, as a consequence, a maximum limit is no longer imposed on the release current.

The basic concepts described above may be extended to structures having more than two ferrite branches. As an example, a structure of four parallel branches might be realized in which each branch is independently magnetizable in one direction or the other. In this structure, the reeds will be closed when three or four branches are magnetized in the same direction and released when pairs of branches are magnetized in opposite directions.

2.3 *Possible Magnetic States*

Prior to discussing methods of exciting a parallel ferreed, it may be appropriate to define the magnetic states that are to be produced by such excitation. Since its remanent flux is a function of the applied magnetomotive force, each ferrite member may assume an indeterminate number of magnetic states. However, if exciting currents sufficient for

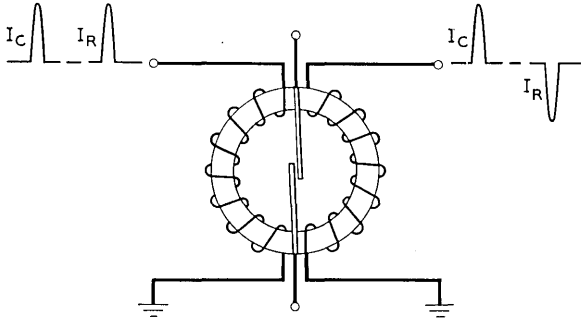

Fig. 2 – The two-branch or parallel ferreed.

saturation of the ferrite members are assumed, the number of possible states reduces to four. Fig. 3 illustrates the four saturated magnetic states: two modes of parallel magnetization that result in reed closure, and two modes of series magnetization that cause the reeds to release.

2.4 *Coincident Methods of Ferrite Excitation*

Establishing a path through a switching network is accomplished by the activation of relatively few among a large number of crosspoint switches. Individual selection of crosspoints is not economically attractive. Instead, it is desirable that the crosspoint device respond only to a coincidence of two or more input conditions. To this end, two coincident control schemes have been devised for the ferreeds: the additive and the differential excitation methods.

2.4.1 *Additive Excitation*

The fact that a square hysteresis loop offers a means for coincident current selection has been widely exploited in magnetic core memory arrays.[4] If a magnetic member of material with a square hysteresis loop is surrounded by two identical windings, and if the excitation in each is limited to a value slightly below the coercive force, then the magnetic state of the material will be altered only if both windings are excited simultaneously and for a sufficient length of time.

One of the ways in which this principle can be applied to the ferreed is shown in Fig. 4(a). The left-hand ferrite member is surrounded by

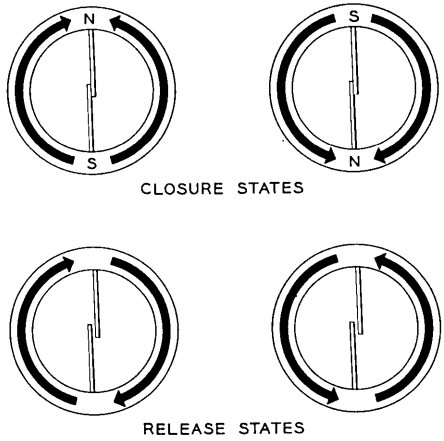

Fig. 3 – Saturated magnetic states of the parallel ferreed.

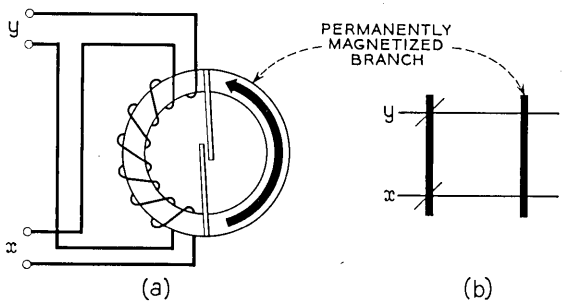

Fig. 4 – (a) Additive excitation applied to one branch of a parallel ferreed; (b) symbolic representation.

two identical overlapping coils, while the right-hand member is assumed to remain in a permanently magnetized state as indicated. Fig. 4(b) shows a symbolic representation of the parallel ferreed adapted from magnetic core work.[5]

Fig. 5 shows a typical relationship between the total remanent magnetomotive force developed across the ferreed structure and the excitation. It can be seen that, even with a two-to-one variation in reed switch sensitivity, coincident current levels can be established for successful closure operation.

Conceptually, the release of the ferreed can be accomplished on a coincident current basis by inverting the pulse polarity on the two wind-

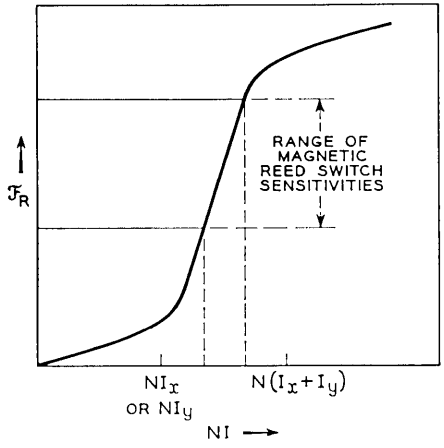

Fig. 5 – Remanent magnetomotive force applied to the reeds as a function of closure excitation.

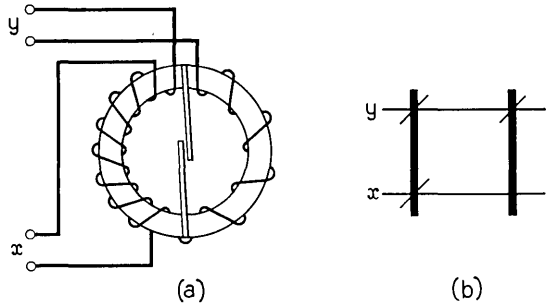

Fig. 6 – (a) Additive excitation with additional winding for maintaining magnetization of permanent branch; (b) symbolic representation.

ings. It has been found, however, that release thresholds are not well defined, due to the bias applied to the device by its own magnetic poles in the closure state. This bias increases the tendency toward magnetic "walk-down"; i.e., cumulative decrease in remanent flux upon repetitive application of partial release currents. Fortunately, the selection problem for release in switching networks is simpler than that for closure. In a two-dimensional array of crosspoints, the release function can be associated with one of the coordinate drives rather than with a specific crosspoint.

A preferred form of the additive case is shown in Fig. 6. The added winding placed around the permanently magnetized leg prevents the demagnetization that might result from finite magnetic coupling between it and the switched leg. The release function in this case is noncoincident and requires a current pulse in winding x of polarity opposite to that used for closure.

2.4.2 *Differential Excitation*

Consider a toroid of remanent magnetic material wound as shown in Figs. 7(a) and 7(b). With application of a sufficiently large current pulse into winding y, clockwise remanent flux is produced in the toroid; the direction of this flux can be reversed by subsequently pulsing the x coil. The magnetic states so produced correspond to the two possible release states of the ferreed. Either closure state can be brought about by simultaneously driving both windings with current pulses of like polarity and approximately equal amplitude.

The basic selection concept explained above, although plausible, cannot be easily implemented without certain modifications. Actually, if the

toroid consists of a square-loop ferrite with high coercive force, a large portion of the flux generated in one of the windings will return through the air, bypassing the other half of the toroid and failing to influence its magnetic state. Also, physical separation of the two ferrite members, which facilitates assembly of the device, tends to further reduce the magnetic coupling.

The practical implementation of the selection principle implied above will be referred to as *differential excitation*. The insufficient magnetic coupling between the two ferrite members is compensated for by an additional winding on each of the ferrite members that is connected as shown symbolically in Fig. 7(c). The auxiliary windings contain only a fraction (typically one-third to one-half) of the turns in the main windings, and the drive current is chosen so that the magnetomotive force produced by the auxiliary windings is equal to or greater than that required to saturate the surrounded ferrite member. In addition, when both pairs of coils are excited to produce a closure state, the differential ampere-turns must also be sufficient to saturate both ferrite members.

Unlike the additive case, differential excitation places no upper limit on the exciting currents. Apart from sufficient amplitude, the two drive currents need only exhibit reasonable amplitude tracking and time coincidence. The tracking requirement can be relaxed, at the cost of increased driving power, by resorting to larger turns ratios between the main and the auxiliary windings.

It should be observed that only one current polarity is required. Actually, as will be discussed more fully in Section V, use of this method removes the necessity for a separate release action. When closure is produced in a ferreed element at an intersection of two coordinates, all other ferreeds located along these coordinates receive only single drives and are left in a released state.

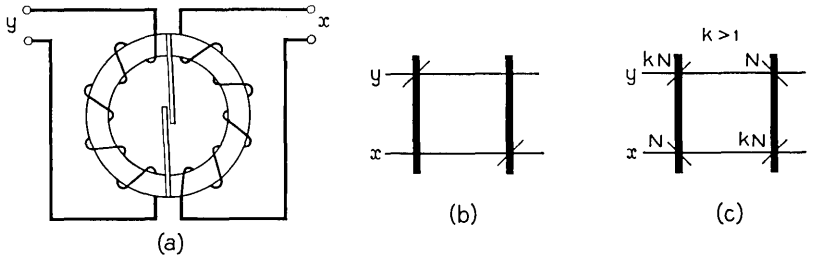

Fig. 7 – (a) Hypothetical model for differential excitation; (b) symbolic representation; (c) winding pattern required for practical realization.

III. A PRACTICAL FERREED

3.1 *The Glass-Sealed Magnetic Reed Switch*

Mention was made before of the structure of the ferreed's contacts and armature members — the magnetic reeds. Successful reduction to practice of the ferreed concept was facilitated by the availability of the glass-sealed magnetic reed switch.

A magnetic reed switch, currently manufactured by the Western Electric Company, and a miniature version of the switch, now under development, are shown in Fig. 8. The miniature switch was adopted for inclusion in the ferreed structure because of its smaller size and greater sensitivity — the excitation required for reed closure is in the order of 30 ampere-turns.

The glass seal allows use of relatively small contact force for a contact life in excess of one million operations, which is adequate for a typical network application. The contact life in the switching network can be preserved by not requiring the reed switch to close or open circuits having battery connected to them. This precaution is also observed in most crossbar networks.

Fig. 9 shows the relationship between applied magnetomotive force and the flux through the gap of the reed switch as obtained by a recording fluxmeter measurement. This plot gives a qualitative picture of the behavior of a typical magnetic reed switch.

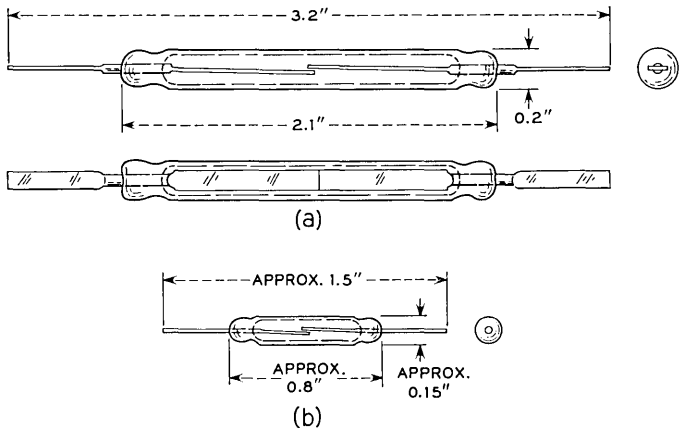

Fig. 8 – (a) Magnetic reed switch in current manufacture (Western Electric Type 224A); (b) miniature magnetic reed switch.

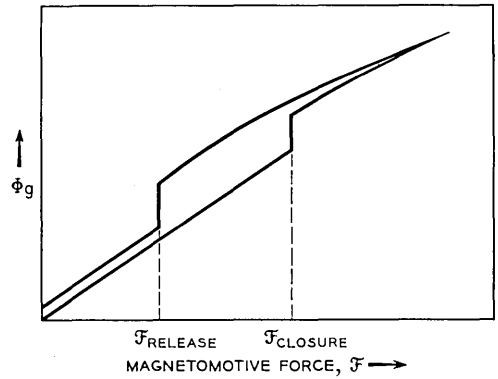

Fig. 9 – Plot of flux in the gap of the magnetic reed switch, showing discontinuities at closure and release.

3.2 *The Ferreed Structure*

A practical model of the ferreed, in which the exciting windings are omitted to present a clearer view of the structure, is shown in Fig. 10. Two magnetic reed switches are associated with the structure to provide two-wire switching as desired for the intended application. The opera-

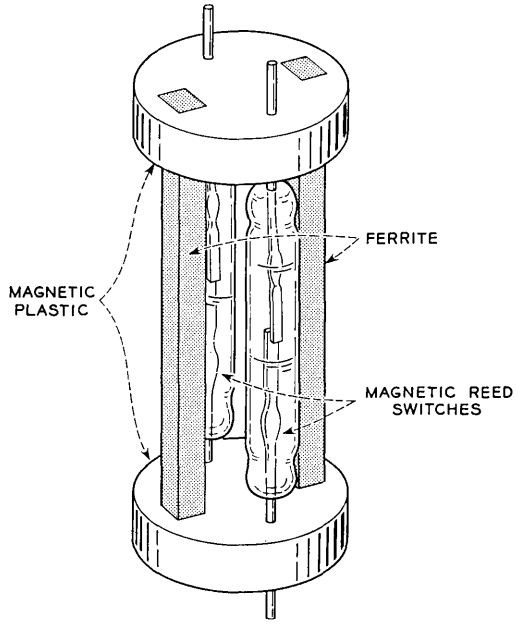

Fig. 10 – Parallel ferreed structure, with windings removed.

tion of this model is based on the concept described above for the two-branch ferrite structure; thus, two ferrite members can be seen in the drawing.

The function of coupling the ferrite bars to the magnetic reed switches is accomplished by the two end-pieces shown in the drawing. Several magnetically soft materials suitable for this purpose have been considered. The material adopted for the model shown consists of a plastic in which a sufficient amount of ferrite powder is suspended to produce adequate permeability (about 20) for this application. This material provides electrical isolation of the reed switches and good mechanical support for the assembly.

The ferrite members of Fig. 10 may be wound prior to assembly for either additive or differential excitation. Fig. 11 is a photograph of a typical experimental ferreed wound for differential excitation.

3.3 *The Ferrite*

The sensitivity and geometry of the reed switches define the length, cross section and magnetic properties of the ferrite bars. In order to operate the most insensitive reed switches, the magnetomotive force existing between the ends of the reeds in the active state of the device should

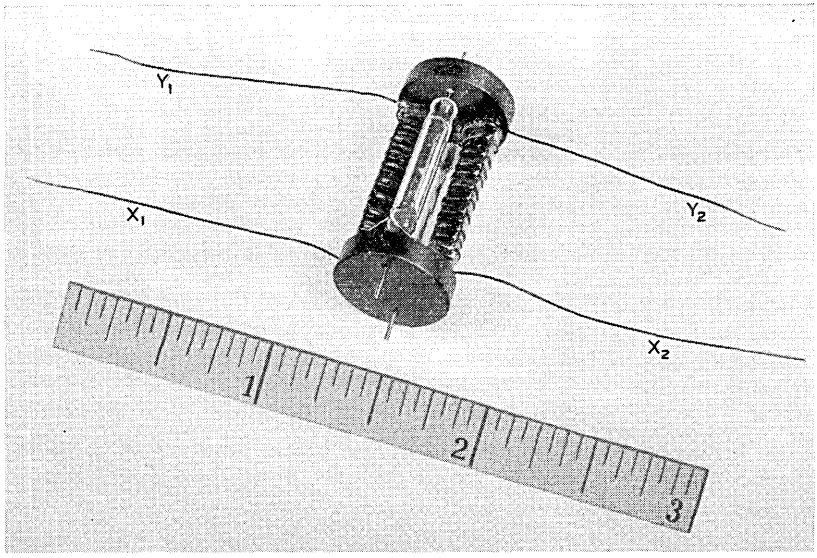

Fig. 11 – Experimental model of parallel ferreed, wound for differential excitation.

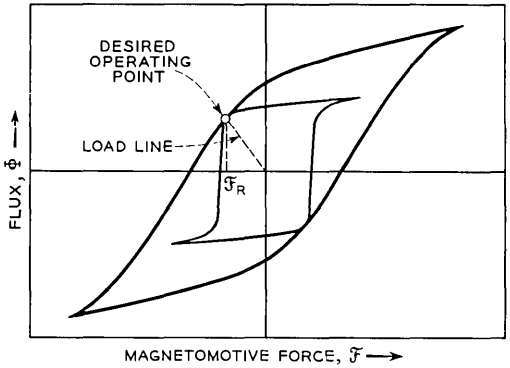

Fig. 12 – Comparison of two hysteresis loops sharing the same operating point.

be in the order of 60 oersted-centimeters. For lengths comparable with that of the reed switch, this implies a coercive force in the two magnetizable members of approximately 30 oersteds.

The desirability of squareness in the the hysteresis loop is illustrated by Fig. 12. If a given operating point is to be established, the material with the squarer of the two loops will permit lower excitation power — an important consideration in large arrays. Also, if the additive scheme of excitation is employed, the squareness of the hysteresis loop has a direct bearing on the obtainable margins.

While it is convenient to make the ferrite body compatible in length with the magnetic reed switches, the cross-sectional area is determined by the maximum flux density available at the point of operation and the over-all magnetic efficiency of the structure. The latter is defined as the ratio of the flux in the gaps of the magnetic reed switches to the total flux flowing through the center cross section of the ferrite. The efficiency of most ferreed models amounts, at best, to about 25 per cent.

Among the ferrites found to be applicable to the design were a cobalt ferrite and a cobalt-zinc ferrite. The latter of these has higher resistivity, a property sought for early ferreed models employing metallic end pieces. A hysteresis loop of a representative cobalt ferrite is shown in Fig. 13.

Apart from ferrites, there exist other materials, such as carbon steel, with suitable coercive forces. Ferrites were found preferable for attaining the ultimate control speeds, since eddy-current delays remain apparent in metallic structures, even when they are laminated. For lower-speed applications, however, metals offer some advantages, e.g., better temperature stability of magnetic properties.

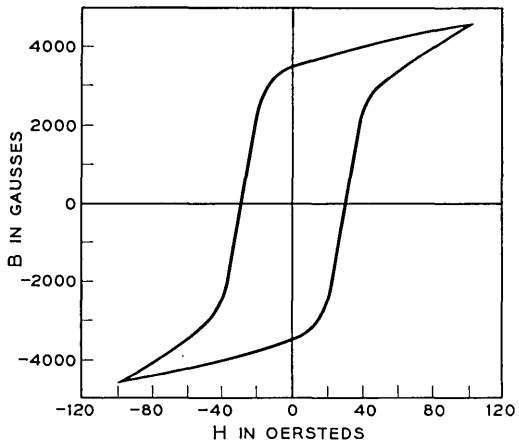

Fig. 13 – Hysteresis loop of a cobalt ferrite suitable for ferreed structures.

3.4 *Operating Characteristics*

The description of a practical ferreed would not be complete without some quantitative account of its behavior. While a detailed report of ferreed performance is not within the scope of this paper, the more important operating characteristics will be summarized.

The closure sensitivity of the ferreed is largely determined by the ferrite magnetization characteristic, the efficiency of the magnetic structure, the sensitivity of the reed switch and, to a lesser extent, by the shape and duration of the excitation pulse. The variation in ferreed closure sensitivity with pulse width for a two-to-one variation in reed sensitivity is given in Table I. Typical values of release sensitivity and

TABLE I — TYPICAL PERFORMANCE CHARACTERISTICS FOR PARALLEL FERREED WITH HALF-SINE PULSE EXCITATION

Sensitivity		
Pulse width, microseconds	Closure, ampere-turns	Release, ampere-turns
10	100 → 145	70 → 95
100	80 → 125	60 → 80

Reed Response Time	
Initial Closure	220 microseconds
Final Closure	450
Release	20

reed closure and release times are included. The time between initial and final closures represents a period of contact chatter that is characteristic of the magnetic reed switch.

Although no precise measurements of ferrite switching times have been made, the response of the ferrite to half-sine pulses as short as 5 microseconds is quite satisfactory. Pulses of this short duration, however, are not considered suitable for driving a series chain of ferreeds (e.g., in a switching array) because of the large back voltages encountered by the pulse source. As a consequence, longer pulses are likely to be used in the control of a multistage ferreed switching network.

The half-sine pulse shown in Fig. 14(a) represents the exciting current that must be applied simultaneously to all windings of a differentially wound parallel ferreed to effect closure. Also shown, in Figs. 14(b) and 14(c), are the voltage developed across the main winding of the ferrite

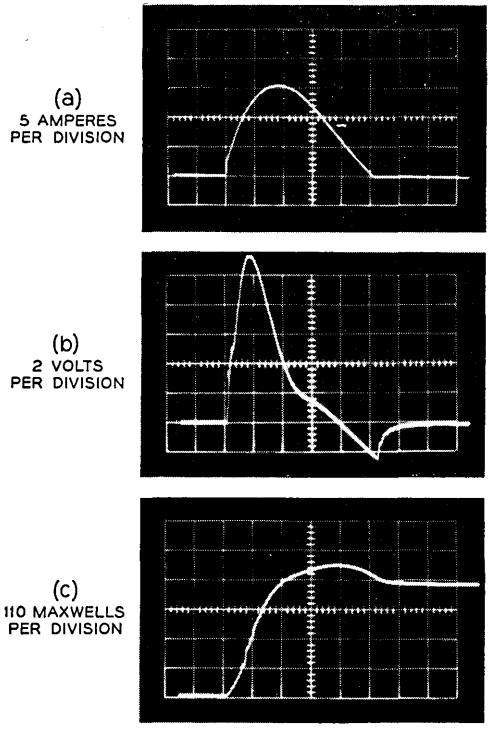

Fig. 14 – Waveforms observed in the switching of a ferrite branch: (a) applied current; (b) voltage induced in the winding; (c) change of flux in the ferrite.

branch being switched and the change of flux in that branch. The average pulse power dissipated in this ferreed is approximately 80 watts, corresponding to an energy requirement of about two milliwatt-seconds.

Variations in current are of particular concern when the ferreed is selectively operated by coincident current methods. The extent to which these variations must be limited for reliable operation represents the operating margins of the device. Fig. 15 presents a graphical comparison of these margins for the additive and differential methods of excitation. In Fig. 15 and in the following development of operating margins, magnetomotive forces applied to the ferrite members have been normalized by assuming the number of turns in each winding to be invariant. This permits reference to the terminal magnetomotive force required for reed closure in terms of a closure current, I_c.

The operating margins for additive excitation are determined from the following considerations: The algebraic sum of the two coincident currents must equal or exceed the closure current; each current must be less

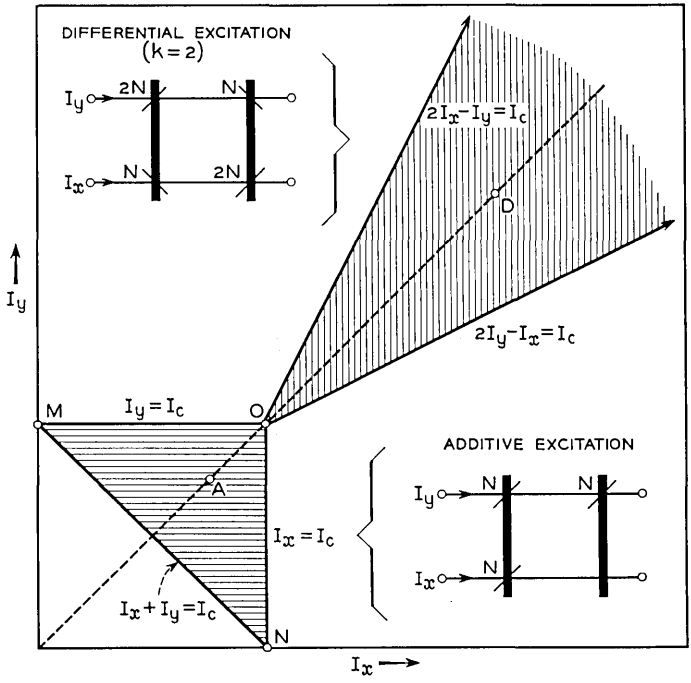

Fig. 15 – Graphical comparison of operating margins: additive excitation valid in horizontally shaded area; differential excitation valid in vertically shaded area.

than the closure current. The limiting conditions are therefore

$$I_x + I_y \geqq I_c , \tag{1}$$

$$I_x < I_c , \tag{2}$$

$$I_y < I_c . \tag{3}$$

These conditions, plotted as equalities in Fig. 15, define a valid region of operation within the triangle MNO. A logical choice for the nominal value of the exciting currents would be $I_x = I_y = \frac{3}{4}I_c$ (point A) which allows a maximum variation of $\pm\frac{1}{4}I_c$, corresponding to $\pm 33\frac{1}{3}$ per cent of the nominal value of I_x or I_y. This is the maximum margin obtainable with additive excitation for a fixed value of closure current I_c. It is evident that variations in ferreed sensitivity will reduce the valid region of operation and decrease operating margins.

A further limitation may be imposed on the margins obtainable with additive excitation by reed vibrations resulting from release. These vibrations occur at a frequency of about 2 kc and decay exponentially with a time constant of approximately 10 milliseconds. Since the closure sensitivity of the reed switch increases with decreasing gap between the reeds, a pulse applied to either winding alone during the post-release period of reed vibration may cause false reclosure of the switch. This situation may arise in coordinate switching arrays using additive excitation unless sufficient time is allowed between the release of one ferreed and the subsequent closure excitation of another sharing the same x or y coordinate.

In the differential mode of operation the condition for closure depends on the preceding release state. If the previous release was obtained by a positive pulse applied to the x winding, the direction of magnetization is down in the left branch and up in the right branch. A subsequent reed closure is effected by reversing the magnetization in the left branch, so the excitation requirement for coincident current closure is

$$kI_y - I_x \geqq I_c . \tag{4}$$

On the other hand, if the previous release was obtained by a positive pulse applied to the y winding, the direction of magnetization is down in the right branch and up in the left. Reed closure is now effected by reversing the magnetization in the right branch, and, in this case, the required excitation is

$$kI_x - I_y \geqq I_c . \tag{5}$$

These two conditions, also plotted in Fig. 15 (for $k = 2$), define a valid region of operation for differential excitation between the two semi-

infinite lines coverging at point o. For the differential case the nominal value of exciting current $I_x = I_y$ may be chosen anywhere along the 45° line that bisects the valid operating region. Starting from a typical operating point, D, it can be seen that the locus for equal variations in I_x and I_y of the same sense is along the 45° line, while the locus for equal variations of the opposite sense is perpendicular to this line. Although the operating margins are most restricted by the latter type of variation, with proper design of the pulse source the former (tracking) type of variation is more likely to occur.

A value of 2 for the turns ratio, k, was arbitrarily chosen for this example. Increasing this ratio increases the angle between the semi-infinite lines, thereby improving the operating margins. However, the accompanying increase in driving power for a given pulse width establishes a practical upper limit on the value of k.

From the graph of I_y versus I_x in Fig. 15, a qualitative comparison of the two methods of excitation can readily be made. The advantage of the open-ended operating region obtained by differential excitation is clearly evident from a comparison with the closed operating region characteristic of additive excitation. With differential excitation, excellent operating margins are obtainable by increasing the exciting current and/or the turns ratio; with additive excitation, good operating margins are realizable only by close control of exciting current and device sensitivities.

IV. RELATED DEVICES

4.1 *The Series Ferreed*

The preceding sections have dealt with a two-branch or parallel ferreed, in which parallel excitation of both branches produced closure states and series excitation produced release states by reducing the flux through the reed switches to zero.

A dual device can be constructed, based on the observation that in a single rod of remanent magnetic material a magnetic state can be induced that corresponds to two shorter magnets connected in series opposition (Fig. 16). The device based on this principle has been named the series ferreed.

A practical form of the series ferreed appears in Fig. 17. The centrally located soft magnetic shunt greatly improves the release sensitivity by reducing the effective air reluctance shunting the individual magnet sections. Both additive and differential methods are applicable to the excitation of a series ferreed.

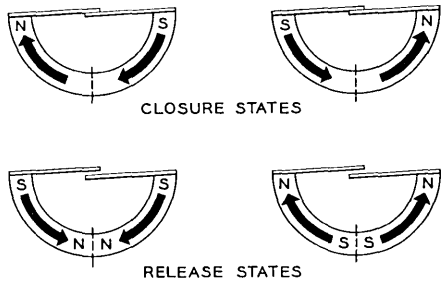

Fig. 16 – Conceptual model of the series ferreed.

4.2 *The Permanent Magnet Reed Switch*

If, in one of the magnetic reed structures shown in Fig. 8, the soft magnetic material used for the reeds is replaced by material of suitably high coercivity and retentivity, a device is obtained that potentially has all the characteristics of the ferreed. It is well known that, for improved magnetic efficiency in permanent magnet structures, the hard magnetic material should be in close proximity to the working air gap. Regarded

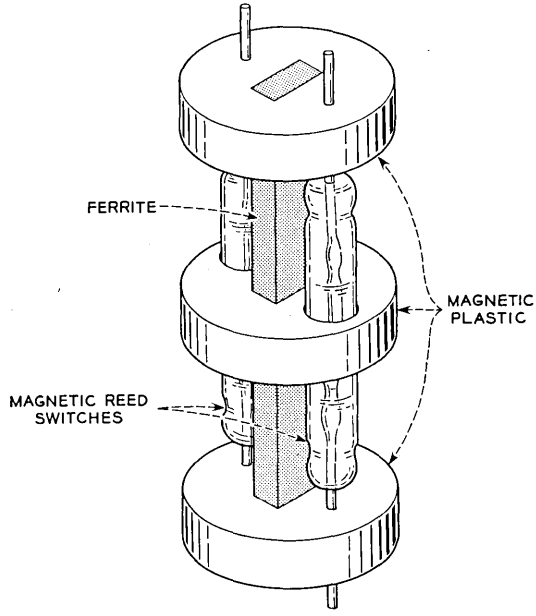

Fig. 17 – Series ferreed structure; windings (not shown) surround upper and lower halves of ferrite bar.

in this light, the permanent magnet reed switch can be viewed as a variation of the series ferreed.

The magnetic states of the device are shown in Fig. 18, together with a differential winding configuration that can produce them. As with the series ferreed, the release sensitivity of the device is improved by a central shunt. A magnetic return path (not shown) helps to increase operate sensitivity.

In experimental units built to test the principle, carbon steel quenched and annealed to produce a remanent flux density of approximately 12,000 gausses and a coercive force of 20 oersteds was successfully used as the reed material.

Although the device can be made to respond to current pulses in the microsecond range, longer pulses, persisting through most of the reed closure interval, result in better current sensitivity and relaxed design requirements.

4.3 *The Polar Ferreed*

The magnetic reed switch of Fig. 8(b) may be replaced in certain ferreed structures by another sealed magnetic switch of the type used in mercury relays. This switch employs a compliant magnetic reed as the transfer contact between two stationary magnetic contacts. By maintaining opposite magnetic poles at the two stationary contacts and vary-

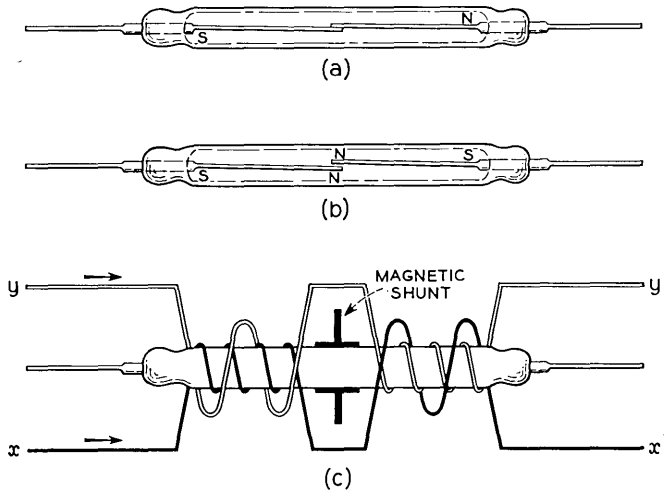

Fig. 18 – The permanent magnet reed switch: (a) closure state; (b) x release state; (c) differential winding pattern.

ing the relation between them and a pole produced in the movable contact, a transfer switching function may be realized. As before, the magnetomotive forces required for operating the switch may be produced by external or internal remanent members.

Ferreeds of this type can be operated by current pulses in the 5-microsecond range, with sensitivities comparable to those of the parallel ferreed. The advantages gained in the polar ferreed are transfer switching action and improved contact performance due to mercury-wetted surfaces.

V. FERREED SWITCHING ARRAYS

5.1 *Space-Separation Networks*

A switching network employing space separation will generally consist of several stages of switching arrays connected in a distributive pattern such as that shown in Fig. 19. A number of alternate paths may be used in the interconnection of two specific terminals. This path redun-

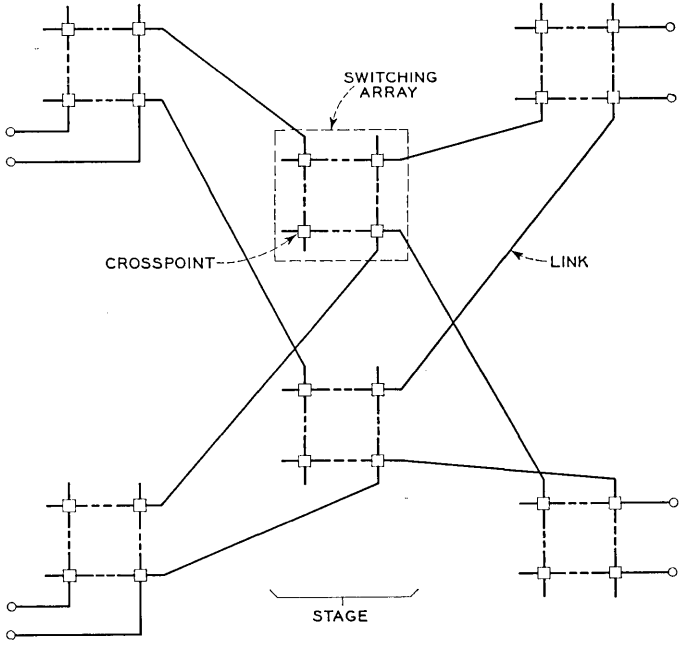

Fig. 19 – Interconnection of switching arrays in a multistage network.

dancy and the number of paths that can be simultaneously maintained are determined by the statistical character of the traffic loads on the lines or trunks to be served.

Four basic control operations may be defined for a switching network of this type. The first of these is the choice of an unoccupied path between two terminals to be connected. Next is the actual closure of the chosen path. When the connection is no longer required by one of the terminals, a third operation re-establishes the identity of the opposite terminal and intervening links. The final operation prepares these terminals and links for subsequent use in other connections.

Switching devices that combine transmission and control circuits, e.g., gas tubes and semiconductor crosspoints, also combine the first two and last two operations. End-marked networks of this type perform path selection in the closure process and path tracing during release.[6] On the other hand, switching devices with separate circuits for transmission and control, such as crossbar switches and ferreeds, employ facilities distinct from the transmission path for the first operation, and usually for the third operation as well.

In a representative crossbar switching network,[7] a third conductor parallels the two conductors of the transmission circuit in each network path. This superimposed circuit, called the "sleeve", permits determination of unoccupied network paths between two terminals and facilitates subsequent release operations. Path selection and tracing may be accomplished in ferreed networks by the addition of a third sealed reed switch to each device, by interrogation of the inherent ferrite memory, or by the use of memory elements external to the switching devices. Of these techniques, the last two are illustrated in the following examples.

5.2 *Array Using Additive Excitation*

The coordinate control paths for a general rectangular switching array of mn crosspoints are shown in Fig. 20. Transmission path multiples (not shown) will be assumed to follow a like pattern, so that simultaneous excitation of a horizontal control path and a vertical control path will effect a transmission connection between the corresponding terminals: one of m to one of n.

Additive excitation may be applied to the ferrite members of a rectangular ferreed array as a means of accomplishing coordinate control. Fig. 21 presents such an array in symbolic form, incorporating the ferreed shown in Fig. 6 for each crosspoint. The right-hand ferrite member of each ferreed is assumed to be switched upward initially and to maintain that polarization indefinitely. When the left-hand ferrite member is

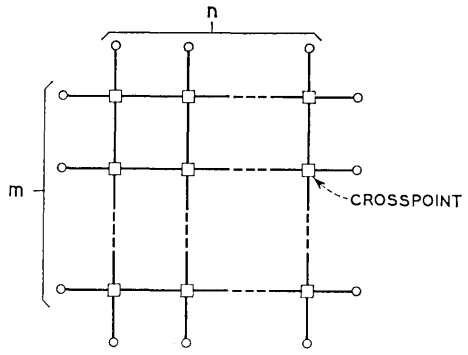

Fig. 20 – Coordinate control paths in a rectangular switching array.

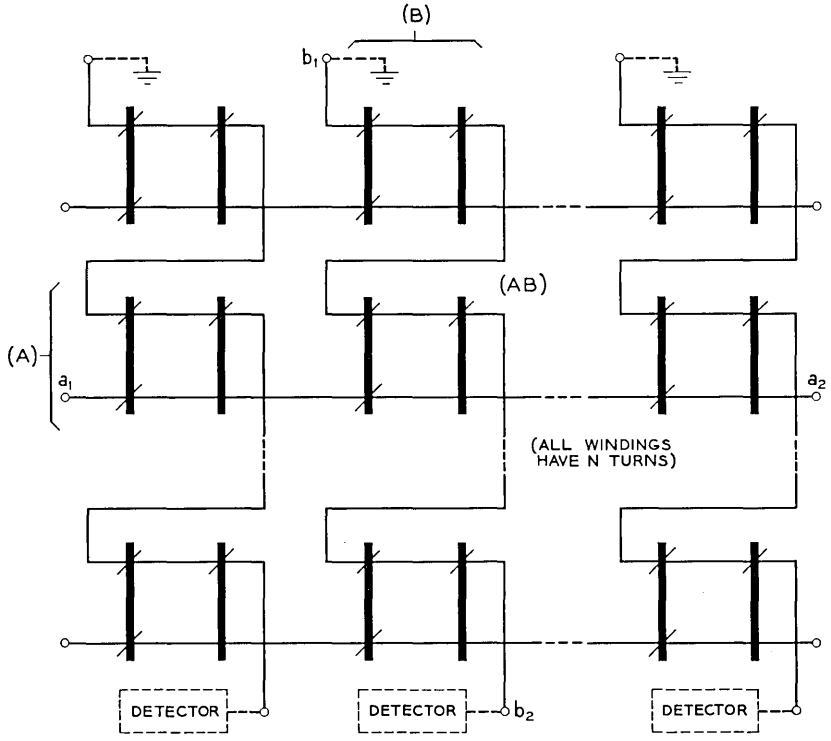

Fig. 21 – Additive excitation applied to a rectangular ferreed array.

switched upward by coincident driving currents, a winding on the right-hand member opposes any tendency for that member to be switched downward due to the finite magnetic coupling between members. The left-hand member of each ferreed is assumed to be switched downward initially, resulting in the release of all contacts.

For illustration, a transmission connection will be established between the terminals marked (A) and (B) in Fig. 21. Initial choice of the path that includes connection (A)–(B) is made by reference to an external memory containing occupancy records for terminals (A), (B) and other affected terminals within the network. Closure of the ferreed (AB) is effected by the simultaneous application of current pulses having amplitudes between one-half and the full ferrite switching excitation along control paths a_1a_2 and b_1b_2.

During application of these pulses, the left-hand ferrite member of ferreed (AB) is switched upward by the addition of the two magnetomotive forces applied. No other ferrite members are reversed, because the others along control paths a_1a_2 and b_1b_2 receive only one-half the required excitation. Control operations within the array are complete when control paths a_1a_2 and b_1b_2 have been pulsed; subsequent closure of the reed contacts (AB) will occur after a delay due to reed inertia.

When it is recognized that terminal (A) no longer requires its present connection, terminal (B) must be identified and both terminals prepared for subsequent use. A current pulse of sufficient amplitude to produce full switching excitation is applied along control path a_2a_1. This pulse switches the left-hand ferrite member of ferreed (AB) downward, but has no effect on the corresponding members of other ferreeds along control path a_2a_1, since they are already switched downward. Before the pulse was applied along control path a_2a_1, sensing elements were temporarily connected to terminals b_2, etc., and terminals b_1, etc., were grounded. The switching of the left-hand ferrite member in ferreed (AB) induces a voltage pulse in control path b_1b_2 that permits identification of the previously connected terminal (B). Control operations are then complete; subsequent opening of the reeds leaves terminals (A) and (B) receptive to new instructions.

In applying this excitation technique to switching networks composed of many such arrays, advantage may be taken of the fact that only one-half excitation is applied along control path b_1b_2. Since this partial excitation will not operate ferreeds that receive no other simultaneous excitation, the vertical control paths shown may be extended into several arrays as a means of conserving access circuits. However, the use of full excitation for release limits control path a_1a_2 to a single array. Fig. 22

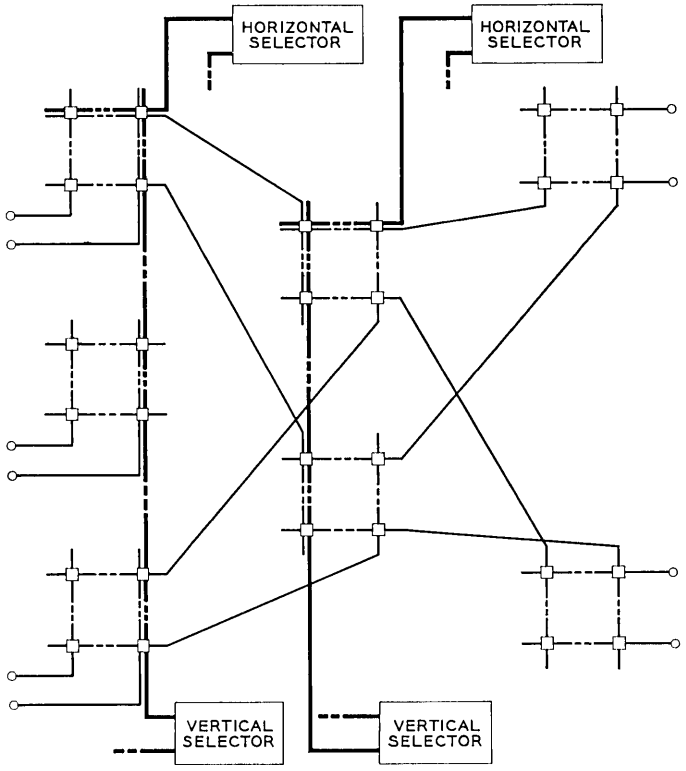

Fig. 22 – Additive excitation in a multistage network with vertical multiple selection (heavy lines represent control paths, light lines represent transmission paths).

illustrates the use of vertical selectors in applying simultaneous excitation along control paths analogous to b_1b_2 in several arrays. In the same way, vertical sensing elements (not shown) may be shared by a number of arrays.

5.3 *Array Using Differential Excitation*

Fig. 23 illustrates the application of differential excitation to the control of a rectangular ferreed array. Horizontal and vertical control paths are provided as before, but with different winding patterns on the individual ferreeds. A typical transmission connection will be established between the terminals marked (C) and (D).

All ferreeds in Fig. 23 are assumed to be released, but this assumption

does not specify the magnetic polarization of the various ferrite members. Because differential excitation produces two release states, the direction of magnetization (i.e., clockwise or counterclockwise) of any released ferreed will depend on the history of its associated control paths. For this example, it is further assumed that all ferreeds in the array initially exhibit magnetic saturation of their ferrite members in a clockwise direction.

The choice of a network path including connection (C)–(D) is again made by reference to an external memory of available terminals (or links) within the network. Closure of the ferreed (CD) is accomplished by the simultaneous application of current pulses along control paths c_1c_2 and d_1d_2. However, in this case, pulse amplitudes are sufficient to produce saturation of ferrite members through windings of either N or $(k - 1)N$ turns.

Due to the opposition of unequal magnetomotive forces, both ferrite members of ferreed (CD) are switched upward. All other ferreeds along

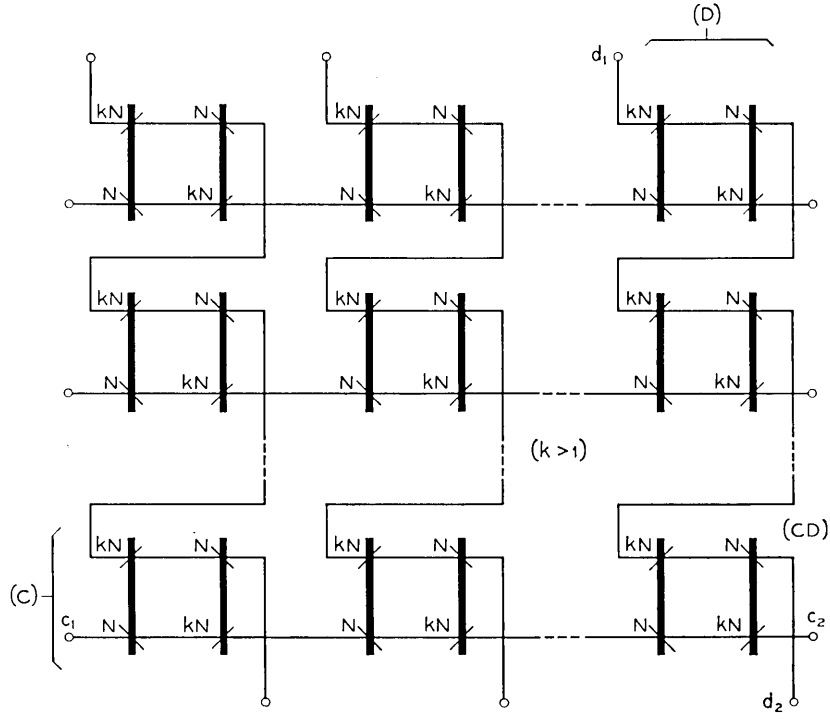

Fig. 23 – Differential excitation applied to a rectangular ferreed array.

control path c_1c_2 experience a reversal in release state, from clockwise to counterclockwise magnetic saturation. The other ferreeds along control path d_1d_2, however, undergo no reversal, because their initial states agree with those imposed by the current pulse along that path. After some delay due to reed inertia, the contacts of ferreed (CD) will close.

When it is recognized that terminal (C) no longer requires its present connection, terminal (D) is identified by reference to another external memory. With the posting of this memory and of the path selection memory, terminals (C) and (D) are prepared for subsequent use. No deliberate release operation is required for ferreed (CD); subsequent path choices that include terminal (C) or (D) will automatically return (CD) to a release state.

The extension of this excitation technique to multistage networks requires a different approach from that employed with additive excitation. In this case, no control path is used for partial excitation; rather, currents in each control path unconditionally release all associated ferreeds

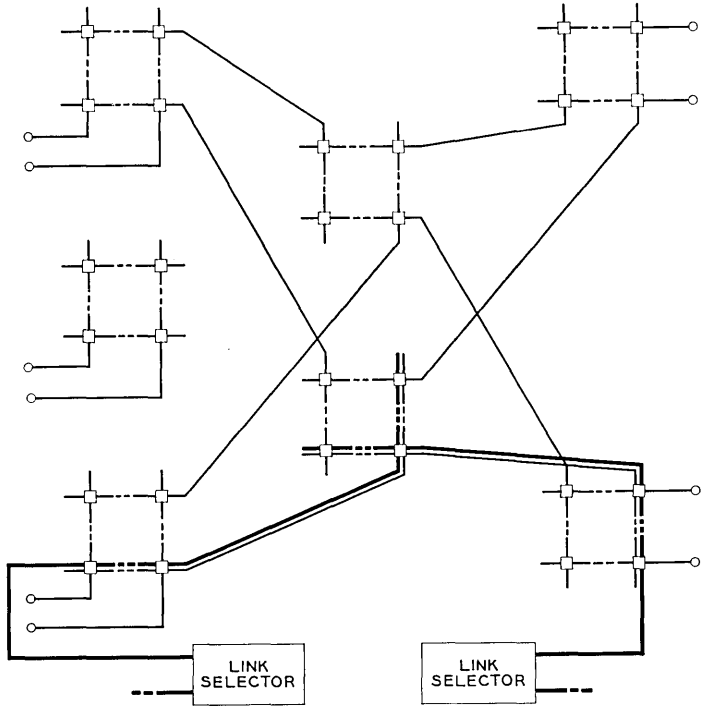

Fig. 24 – Differential excitation in a multistage network with link selection (heavy lines represent control paths, light lines represent transmission paths).

except those located at intersections with other active control paths. For this reason, control paths must conform to transmission paths as a means of assuring that conflicting connections are undesired connections. One significant economy in access circuits can be realized, however: the selection of control paths for two related array terminals may be combined, as in the link selector of Fig. 24.

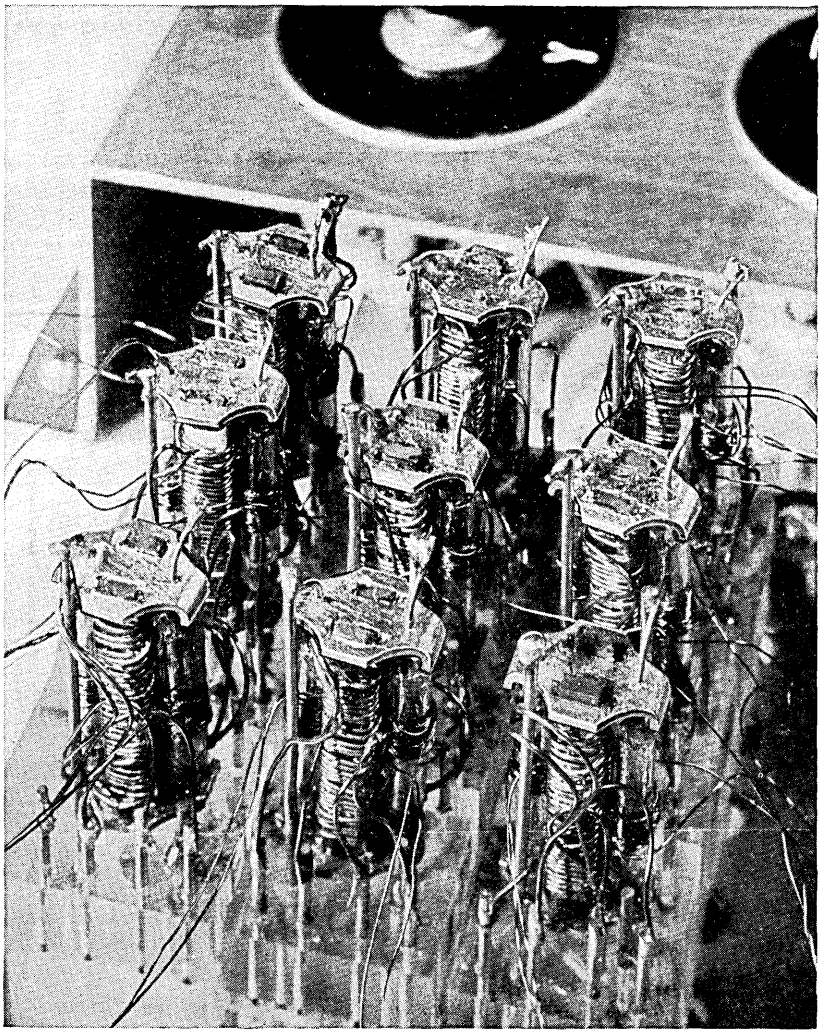

Fig. 25 – An early experimental array of nine parallel ferreeds. (The coils on the glass enclosures are used for flux measurements, and play no part in control.)

Because differential excitation eliminates old connections while establishing new ones, a multistage network of this type can be expected to contain numerous fragmentary connections at any given time. These do not interfere with network operations and do not justify separate release instructions; however, lines and trunks are disconnected from the network to provide isolation of battery and supervision.

VI. CONCLUSION

The union of magnetic reed switches with elements having controllable magnetic remanence has yielded a new class of switching devices. One conceptual model from this class appears to have the properties of microsecond control time, coincident selection, memory without holding power, enclosed contacts and compact structure. Design and construction of experimental forms of the ferreed have borne out predictions of the early concepts and have suggested further variations on the theme.

The ferreed was created to fill a need for a network crosspoint providing high-speed coincident selection and metallic contacts. Because of the specific nature of this objective, primary emphasis has been placed on optimization of the ferreed's characteristics as a crosspoint switch. It appears promising as a network element and may also prove suitable for a number of other applications.

In digital systems, the properties of rapid control, metallic output and absence of holding power suggest the use of ferreeds as memory for display, buffer and input-output functions. Ferreeds can also be used to realize switching functions as combinations of flux patterns in multiple-branch ferrite structures, or as combinations of magnetomotive force in multiple windings such as those used for differential excitation.

VII. ACKNOWLEDGMENTS

The exploratory study of ferreeds and related devices received wide support at Bell Telephone Laboratories. The authors would like to express their appreciation for the efforts of R. A. Chegwidden, who developed the necessary magnetic materials; I. Dorros and R. L. Peek, Jr., who performed much of the analytical and experimental work; and H. J. Wirth, Jr., who constructed numerous device models. Important ideas were contributed by J. T. L. Brown, R. E. Hersey and R. W. Ketchledge.

REFERENCES

1. Steinbuch, K., The International Status of Electronic Techniques in Telephone Exchanges, Nach. Zeit., **10**, 1957, p. 335.

2. Joel, A. E., Jr., Electronics in Telephone Switching Systems, B.S.T.J., **35,** September 1956, p. 991.
3. Joel, A. E., Jr., An Experimental Switching System Using New Electronic Techniques, B.S.T.J., **37,** September 1958, p. 1091.
4. Rajchman, J. A., Magnetics for Computers—A Survey of the State of the Art, R.C.A. Rev., **20,** March 1959, p. 92.
5. Karnaugh, M., Pulse-Switching Circuits Using Magnetic Cores, Proc. I.R.E., **43,** May 1955, p. 572.
6. Feldman, T. and Rieke, J. W., Application of Breakdown Devices to Large Multistage Switching Networks, B.S.T.J., **37,** November 1958, p. 1421.
7. Scudder, F. J. and Reynolds, J. N., Crossbar Dial Telephone Switching System, B.S.T.J., **18,** January 1939, p. 76.

A Remote Line Concentrator for a Time-Separation Switching Experiment

By D. B. JAMES and J. D. JOHANNESEN

(Manuscript received October 30, 1959)

Remote line concentration, time-separation switching and PCM transmission are combined in a communication system experiment called ESSEX (Experimental Solid State Exchange). Organization and design details of the remote line concentrator used in the research model are presented and discussed.

I. INTRODUCTION

An earlier paper[1] has described a research experiment on integrated communications using time-separation techniques. This experiment is called ESSEX (Experimental Solid State Exchange). The purpose of this paper is to discuss the environment and implementation of the remote line concentrators used in the experiment.

ESSEX is an experiment designed to explore the possibilities of using digital systems in exchange area plant. Subscribers are connected to remote units that multiplex and convert analog signals to digital signals. Digital signals are transmitted and switched between remote units and are converted to analog form only at the units to which they are directed. The assemblages that provide the necessary switching and transmission functions for subscribers' lines are called *concentrators*; those that provide the switching and transmission functions for trunks are called *trunkors*. Concentrators and trunkors form the basic building blocks of the system.

A concentrator consists of two units called the *remote line concentrator* and the *concentrator controller*. The trunkor also has two parts, the *trunkor unit* and the *trunkor controller*. Since trunk groups ordinarily have high usage, the trunkor has no concentration, but otherwise it is identical to a concentrator.

The remote line concentrator can serve a maximum of 255 subscribers. In a working system, however, traffic considerations would probably

limit the number of subscribers to about 115, of whom 23 can converse simultaneously. Their voice-frequency signals are selectively switched by time-separation techniques and converted to digital form for transmission. Since digital signals are both sent and received by remote line concentrators, four-wire transmission is employed between them. Exchange cable pairs can be used for this purpose, with regenerative repeaters spaced every 6000 feet or less depending on the kind of cable.

The interconnection of the digital send and receive links from the remote units is made at a unit called the *central stage switch*, which uses four-wire switching on a time- and space-division basis. Time and space division are used to provide the necessary number of paths required when many concentrators are connected to the switch. The send and receive links from each unit are equipped with a time-division switch for each space-division link or *junctor* in the central stage switch. This provides a full-access space-division interconnection between junctors.

In setting up or maintaining a communication channel in ESSEX, the selective switch at the remote unit and the switch at the central stage switch are operated periodically, as is required in time-division systems. Memory is, therefore, required to deliver proper information at the right time to the switches. Several options on the location of the memory were considered, with the one option chosen being to locate all the memory close to the central stage switch. Another digital control pair, called the c lead, was therefore required to deliver address information to the remote unit. The remote unit thus becomes a completely slave-operated unit containing a minimum of equipment.

The memory required by the concentrator is contained in a unit called the *concentrator controller*. In addition to the line number and junctor gate number memory already referred to, this unit contains a third memory to record the state of the call being handled by the concentrator. These three memories are implemented with magnetostrictive delay lines, each arranged in a closed loop. The information in the delay line circulates in serial form and is available for transmission over the c lead to the remote line concentrator.

In addition to the memories, the concentrator controller contains circuits for checking and processing calls, as well as circuits that enable it to deliver and accept information from a common control. A complete description of all the functions performed by the concentrator controller is given in the accompanying paper.[2]

A typical arrangement of concentrators and trunkors is shown in Fig. 1. Here, the remote line concentrators are located some distance from a switching center, which contains the central stage switching, the con-

centrator controllers, the trunkor and possibly the common control. Since the bounds of the ESSEX experiment did not include an experimental investigation of common control, Fig. 1 shows, in its place, a manual console. This manual console, called the *common control simulator*, allows an operator to perform the logic and memory functions normally performed by a common control.

II. TRANSMISSION AND SWITCHING IN ESSEX

A clearer picture of some of the functional requirements of the remote line concentrator can be obtained by following a call through the system shown in Fig. 1. It will, however, be profitable to digress for a moment to give some of the important system numbers and define terms that will be used here and in the detailed discussion that follows.

Voice-frequency signals appearing on a subscriber's line are sampled at an 8000-cycle rate. This allows signal components up to 4000 cps to be reproduced by the low-pass filters in the remote line concentrators. When the signals are sampled at this rate, the period between samples

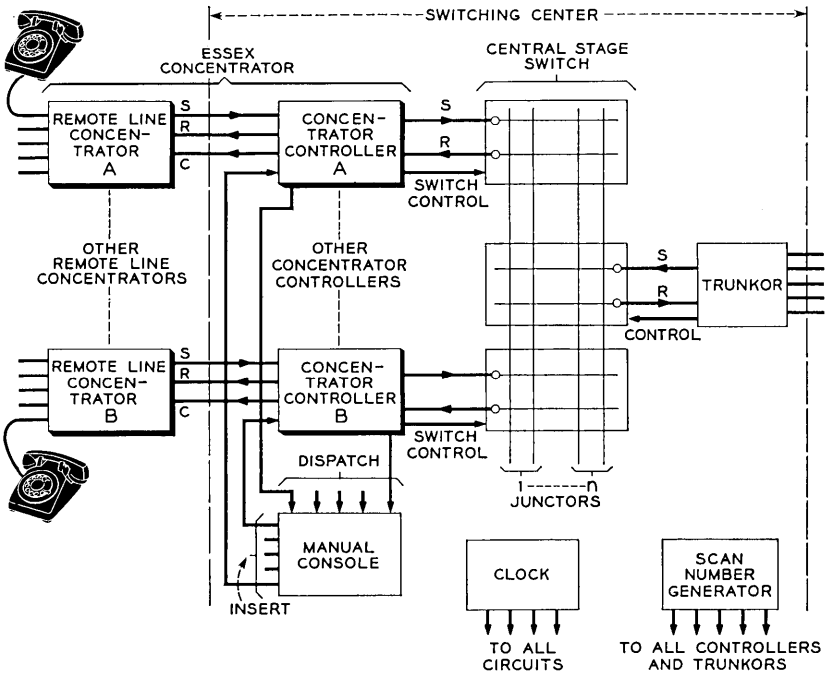

Fig. 1 — The environment of the remote line concentrator.

of the same message is 125 microseconds and is called a *frame*. Frames are subdivided into equal parts, called *time slots*, each of which can be used to set up a path to each concentrator. ESSEX uses 24 time slots, 23 being used for talking and one for supervisory functions. The time slots are numbered 0 to 23 and are each 5.2 microseconds long. This 5.2-microsecond interval is divided into eight pulse or bit positions, each 0.65 microsecond wide, numbered 0 through 7. On the R and S leads, bits 0 through 6 are used for pulse code modulation information (PCM) and bit 7 is used for other functions. The basic bit rate is thus 1.536 mc.

Returning now to Fig. 1, we can see that a voice-frequency signal appearing on a subscriber's line at remote line concentrator A is switched by an address arriving at remote line concentrator A via the C lead. The pulse-amplitude modulated (PAM) sample that results from the switching is then encoded and transmitted over a balanced cable pair, designated S, to concentrator controller A. From here it passes, still in digital form, to the central stage switch, where it is switched onto a cable pair, designated R, to concentrator B. Signals coming from the S lead of concentrator B are switched to the R lead of concentrator A. The junctor gates of concentrators A and B operate at the same time, which means that the same time slot is used in both concentrators in setting up a channel.

Since the junctor gates operate once per frame for each conversation, PCM signals from a concentrator must arrive at the central stage switch in phase with the PCM signals being sent to the concentrator. Remote line concentrators will, in general, be located at different distances from the switching centers, and their loop transmission delays will vary accordingly. If the loop transmission delay is one frame or an integral multiple of frames, this phase requirement will be met. This condition also can be achieved and maintained continuously by the insertion in the S lead of an adjustable delay pad, which is part of the concentrator controller. A more detailed description of the transmission delays is given in the companion papers.[1,2]

From a switching viewpoint, ESSEX uses a four-stage switching plan. Two of these stages are in the central stage switch; the other two are in the concentrators handling the call.

III. THE REMOTE LINE CONCENTRATOR

A block diagram of the remote line concentrator is shown in Fig. 2. The blocks represent multifunctional circuits that handle either analog or digital signals. The elements in the analog class are the subscriber line circuits, the two-to-four-wire converter, the compressor and the ex-

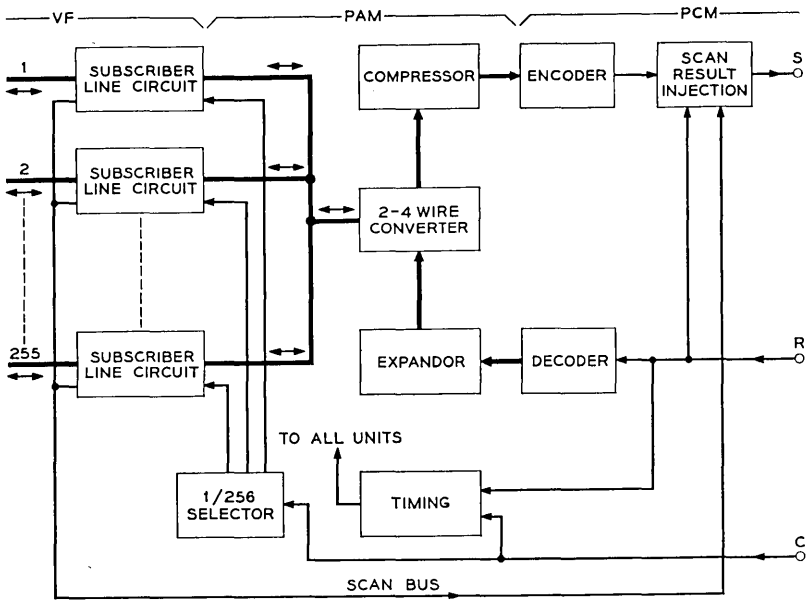

Fig. 2 — Block diagram of the remote line concentrator.

pandor. The remainder, which includes the encoder, the decoder, the selector, the line scan injection and the timing circuits, are digital. The purpose of each of these multifunctional circuits can be made clear by describing the operation of the remote line concentrator itself.

When an eight-bit address is delivered to the 1-out-of-256 selector via the C lead, the selector delivers a pulse to the subscriber's time-division gate. This gate, which is a part of the subscriber's equipment, samples the voice-frequency signal and delivers the resulting PAM signal to the two-to-four-wire converter, which is a time-shared hybrid. The PAM signal is sent through an instantaneous compressor to the encoder, and the resulting PCM code is delivered to the S lead. Signals arriving on the R lead are decoded, expanded, amplified in a common amplifier and sent through the time-shared hybrid to the subscriber's equipment.

Since the remote line concentrator is a slave unit, timing information must be recovered from the signals coming from the concentrator controller. Frequency information is extracted from signals arriving on the C lead, and phase information is obtained from a unique signal that appears on the R lead of each concentrator in the last time slot of each frame (time slot 23).

Line scanning for request-for-service and hang-up is a relatively simple procedure in ESSEX. A *scan number generator*, common to all

concentrator controllers, is provided. This device generates all the binary numbers from 0 to 255 and then repeats, with each number being generated and held for four frame intervals. A particular line gate number thus appears for four frames at the concentrator controller 7.8 times per second. If the line gate number that appears is already recorded in the line gate number memory of the concentrator controller, then the line is scanned in its assigned time slot. In this case, the *scan command*, which occurs at bit 7 time, is sent out on the R lead. If the line gate number is not in the memory, then it is scanned by sending out the line gate number over the C lead in time slot 23 during the first of the four frames. Four frames are used to allow time for the signals to be sent out to the remote line concentrator and return to the controller and for the controller to act on the result.

The lines at the remote line concentrator are scanned when the scan command appears on the R lead. The particular line number that is to be scanned is the one present at the input of the selector at this time. A scan gate is located in each subscriber's line circuit and, when a pulse is delivered from the 1-out-of-256 selector to the subscriber line circuit and the subscriber is "off-hook", the scan gate delivers a pulse to the scan bus. If the scan command is also present on the R lead, the scan flip-flop in the *scan result injection* circuit is set. The scan result is then delivered to the S lead in the bit 7 position of time slot 22.

Table I lists the digital signals that are sent from or received by the remote line concentrator. The implementation of the multifunctional circuits described above will be discussed in the following sections.

IV. TIMING CIRCUIT

The timing circuit performs the following functions:

i. recovers the basic 1.536-mc timing signal from the C lead addresses;

ii. counts this signal down by eight to generate bit pulses at 192 kc, thus defining the time slots;

iii. frames the eight's counter by detecting a unique signal consisting of eight consecutive ones sent out on the R lead in the last time slot;

iv. combines and amplifies the above signals and distributes them to various parts of the remote line concentrator.

The diagram of the timing circuit is shown in Fig. 3. A slave clock extracts the basic 1.536-mc timing signal by passing the C lead bits through a quartz crystal filter. To insure adequate timing signals during low-traffic periods, pulses on the C lead represent "zeros" instead of "ones". The output signals are amplified and two phase pulses, ϕ_0 and

TABLE I—DIGITAL SIGNALS WHICH ENTER OR LEAVE THE REMOTE LINE CONCENTRATOR

From or to	Name of signal	Sent via (see Fig. 2)	Time when sent or received by controller	Purpose
A. To concentrator	1. Line gate number (LGN)	C lead	Bits 0 through 7 of time slots 0 through 22 in every frame	Orders operation of line gate, thereby sampling
	2. Scan gate number (SGN)	C lead	Bits 0 through 7 of time slot 23 in frame 0	Scans idle lines to determine whether they are "off-hook"*
	3. PCM speech or tone	R lead	Bits 0 through 6 of time slots 0 through 22 in every frame	Delivers speech or tone signal to decoder
	4. Scan command	R lead	Bit 7 of time slots 0 through 22 in frame 0	Adds scan order to signal A1*
	5. Framing command	R lead	Bits 0 through 7 of time slot 23 in every frame	Orders reset of counters in slave clock
B. From concentrator	1. PCM speech	S lead	Bits 0 through 6 of time slots 0 through 22 in every frame	Transmits speech signal from encoder
	2. Scan result	S lead	Bit 7 of time slot 22 in frame 1 or 2	Indicates if line scanned (A2 or A4) was "off-hook"

* Either A2 or A4 is used (not both)

ϕ_2, are generated by blocking oscillators. The ϕ_2 pulse advances an eight-stage re-entrant shift register which generates the bit pulses. The framing signal from the eight-ones detector insures that this counter has a single one circulating in the correct phase.

The framing signal is generated by the eight-ones detector, another eight-stage shift register. Whenever a one appears on the R lead, a one is advanced into the register. The arrival of a zero resets the register to zero. A consecutive string of eight ones will advance a one into the eighth stage, and its output is "ANDed" with the last input one, delayed by one-half microsecond, to give the "frame" signal. This signal resets the first seven stages to zero. The last stage is reset when the next one arrives on the R lead; this reset operation is designed to avoid a race condition.

The frame signal sets the inhibit flip-flop, which in turn inhibits the send and receive gates in the framing time slot. This is done to prevent sampling in time slot 23, which would introduce annoying ticks in the telephone being scanned.

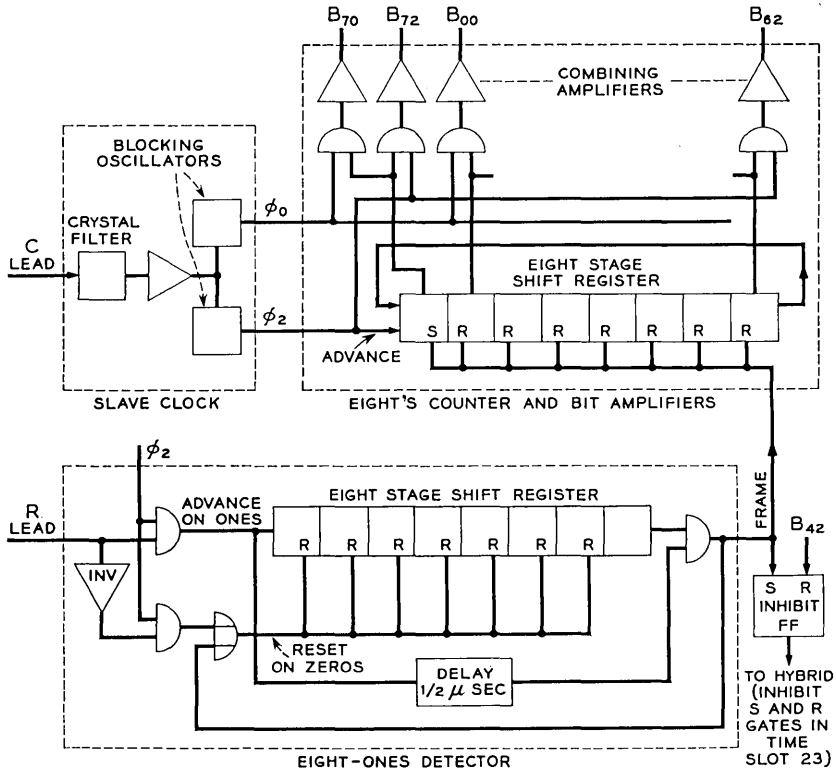

Fig. 3 — Timing circuit.

The eight outputs from the eight's counter are combined with the phase pulses, ϕ_0 and ϕ_2, and amplified to form all the other timing signals required by the remote line concentrator.

V. SELECTOR

Eight-bit addresses carried by the c lead are delivered to the selector (Fig. 4). The serial eight-bit words are advanced through a shift register and, when they have stepped completely into the register, are read out in parallel through AND circuits. Both the bits and their primes are stretched in time to 2 microseconds and amplified by 16 stretching amplifiers. The first four bits and their primes become the inputs to a 1-out-of-16 diode matrix consisting of 16 four-input AND circuits selecting one out of 16 output leads. The second four bits and their primes select one out of 16 output leads of a second diode matrix. The two active leads

from the diode matrices enable two flip-flops, one each in two banks of 16. The setting of the flip-flops is timed by an enabling clock pulse labeled "set". This pulse occurs one microsecond later than the "read" pulse, which interrogated the shift register. The outputs of the flip-flops are transformer-coupled into transistor amplifiers. The collectors of one set of 16 amplifiers in the common emitter configuration provide the verticals of a 16-by-16 coincident-voltage matrix. The emitters of the other set, in the common collector configuration, provide the horizontals of this matrix.

The primaries of the subscriber line gate pulse transformers are connected between the horizontals and verticals of this 256-point matrix.

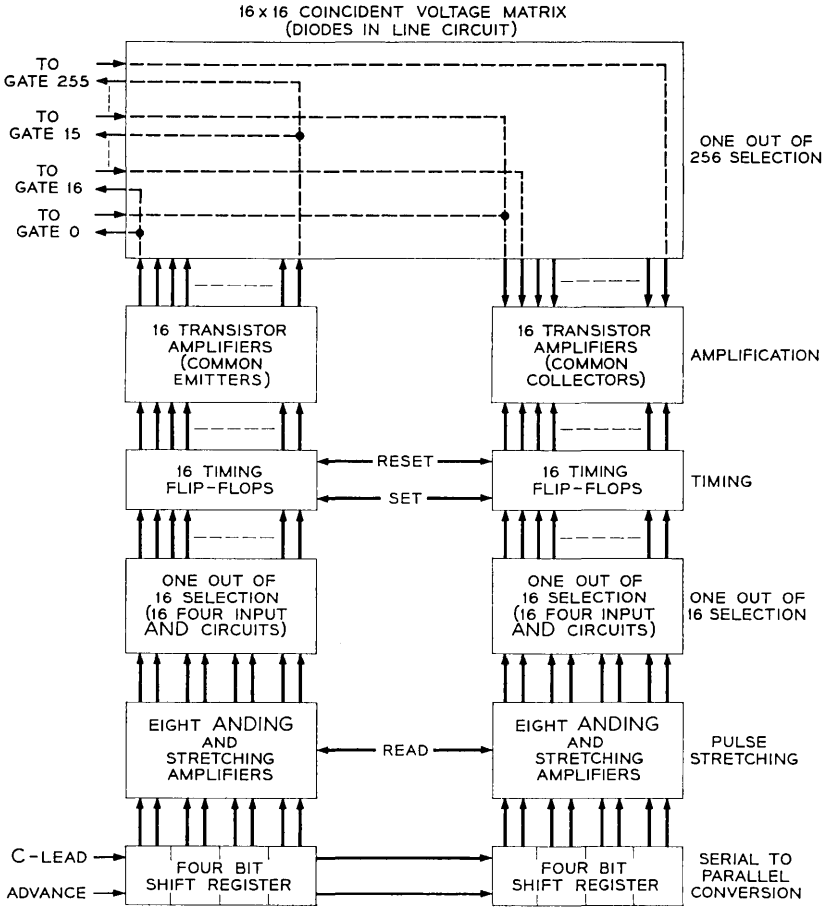

Fig. 4 — 1-out-of-256 selector.

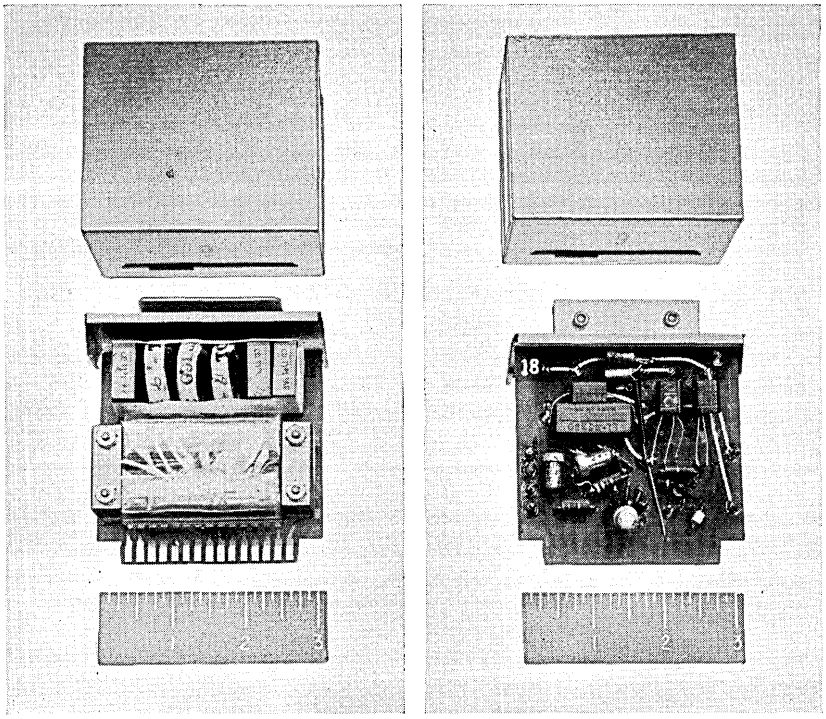

Fig. 5 — Subscriber line circuit package.

A diode to prevent "sneak" paths is located in series with the primary of the pulse transformer in the line circuit module.

The common collectors are connected to −6 volts and the common emitters to ground. When a vertical and a horizontal are selected, the 6 volts applied across two ON transistors, one diode and the primary of the gate transformer allows 80 milliamperes of current to flow in the primary of the transformer. Three microseconds later, the timing flip-flops are reset, interrupting the current flow.

Small signals resulting from parasitic capacities appear on the inputs to unselected gates that share either verticals or horizontals with the selected gate. These unwanted signals are caused by the discharging of wiring capacities through the primary windings of the unselected gates. They can be reduced below the threshold of the gate by adding a resistor (10,000 ohms) from each horizontal and each vertical to a point at −3 volts. Thus, all wiring capacities are charged to the same voltage and all unwanted signals become equal and less than those needed to cause a gate to conduct.

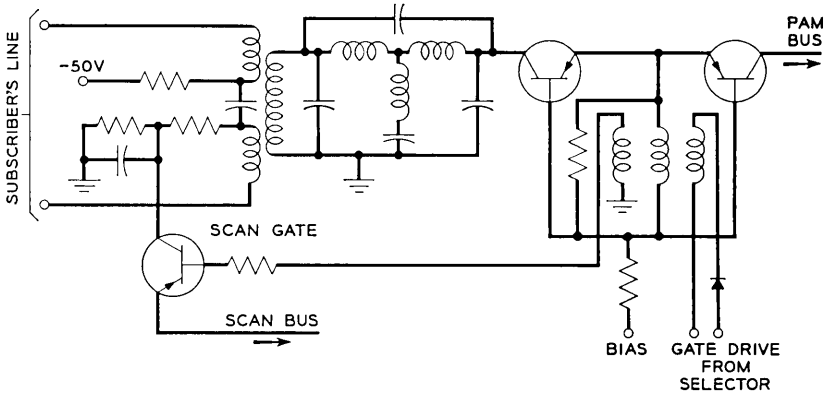

Fig. 6 — Subscriber line circuit diagram.

VI. SUBSCRIBER LINE CIRCUIT

The two-wire subscriber line circuit, one of which is provided for each line, is arranged so that it can be installed as required. It is the smallest module in the system, but one of the most important, since it has such a large effect on the per-line cost of the system. A photograph of the subscriber line circuit package is shown in Fig. 5.

The subscriber line circuit contains a time-division gate, a low-pass filter, a scan gate and a repeat coil. The repeat coil is used to match the 900-ohm telephone to the 2000-ohm filter, and to isolate the time-division gate from the common battery required by the telephone. A circuit diagram of the subscriber line circuit is shown in Fig. 6.

The time-division gate[3] consists of two nearly symmetrical alloy germanium transistors. These are connected emitter to emitter and base to base, with a pulse transformer winding connected between base and emitter. The application of a 80-milliampere current pulse will enable the gate to pass a peak signal current of about 400 milliamperes from collector to collector. In this state, the gate looks like a 2-ohm resistor. When the current pulse is removed, the gate returns to its high impedance state. In the experimental system, which uses 2N417-type transistors, about 2 microseconds is allowed for this purpose, providing the desired adjacent channel crosstalk level of 75 db. In the absence of a pulse, the gate is a high impedance. To increase the isolation obtained in the off condition, a bias voltage is applied to the gate through a resistor, insuring that the collectors are back-biased during the maximum positive swing of the voltage on the filter or on the PAM bus. The resistor also helps to reduce crosstalk produced by the junction capacity of the transistors.

The time-division gate limits the maximum rms sine-wave power-handling capacity of the system. The desired crosstalk level fixes the time that must be allowed for turn-off of the gate, and thus determines the "on" time of the gate, since the time slot width is fixed. The drive available from the selector sets the maximum current the gate can pass. The breakdown voltage of the transistors in the gate, minus the bias voltage, determines the maximum signal voltage swing allowed. Since the maximum voltage swing and peak current are limited by the gate transistors, the power-handling capacity of the gate and the operating impedance level of the filter are specified.

Time-separation systems require a low-pass filter to isolate the subscriber's line from the frequencies generated by the operation of the sampling gate.[4,5] In addition to having good out-of-band rejection, the filter should also have fairly constant impedance characteristics within the pass band. However, since a filter per line is required, it is desirable, from an economic standpoint, to use as few components as possible. These somewhat conflicting filter requirements generally lead to a compromise design. The low-pass filter in the experimental unit is a two-section insertion loss design having a characteristic impedance of 2000 ohms. The impedance level was determined by the factors discussed in the preceding paragraph.

The line scanning procedure was described in Section III. The subscriber line circuit is provided with a scan gate that delivers a pulse when the line circuit is pulsed by the selector and the telephone is "off-hook" and drawing current. The resulting pulse is delivered to the scan bus. If the line is being scanned, the scan command that appears at bit 7 time on the R lead and the pulse on the scan bus will set the scan flip-flop storing the scan result. This result is delivered at the proper time, through the scan result injection circuit (Section X), to the S lead.

The design of the line circuit is influenced to some extent by the type of telephone used. The laboratory models of the subscriber line circuits were designed to work with a transistorized telephone requiring about two-thirds of a watt. This power is obtained from a 50-volt common battery through current-limiting resistors located in the line circuit. The telephone is equipped with a tone ringer, which gives substantially the same acoustic power output for tone voltages that vary from 0.5 to 2 volts.[6] These ringing levels fall well within the power-handling capacity of the subscriber gate. A slightly modified line circuit will allow conventional telephones with electromechanical ringers to be used.

VII. TIME-SHARED HYBRID

Signals from subscribers' telephones are carried to the remote line concentrator by balanced two-wire cables. Signals are carried between the remote line concentrator and its concentrator controller on a four-wire system. A hybrid thus is needed to convert between the two-wire and four-wire systems.

Two alternative arrangements were considered: (a) a voice-frequency hybrid and (b) a time-shared hybrid. A voice-frequency hybrid arrangement requires for each subscriber's line a sending gate and filter, a receiving gate and filter and a voice-frequency hybrid. A time-shared hybrid arrangement requires one filter, one gate and a repeat coil for each subscriber's line plus a share of a common hybrid. The second arrangement was used. The selected line gates connect all active subscribers' circuits via a common two-wire PAM bus to the time-shared hybrid (Fig. 7).

The hybrid consists essentially of a send gate and a receive gate, which are identical to the subscriber gates. Since send signals must be stretched

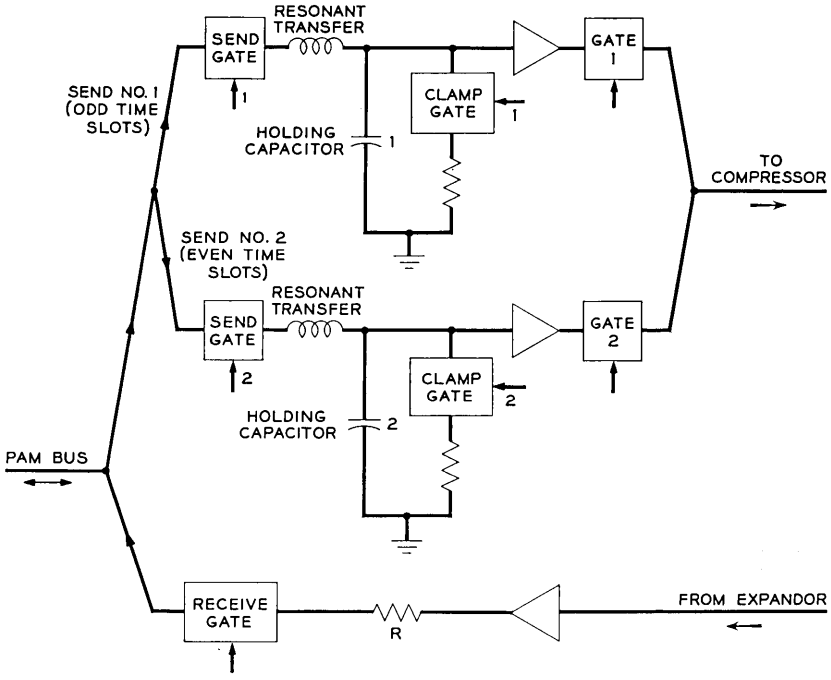

Fig. 7 — Time-shared hybrid.

for nearly one time slot for the coding operation, two send gates are operated in alternate time slots under the control of a binary counter. The hybrid works in the following fashion: the send gate is operated at the same time as the line gate and remains in operation for one microsecond; it then opens up and, after 0.3 microsecond guard space, the receive gate operates, opening 1.6 microseconds later with the line gate (Fig. 8). A received PCM signal is decoded and amplified, goes through the receive gate to the PAM bus, then goes through the line gate and

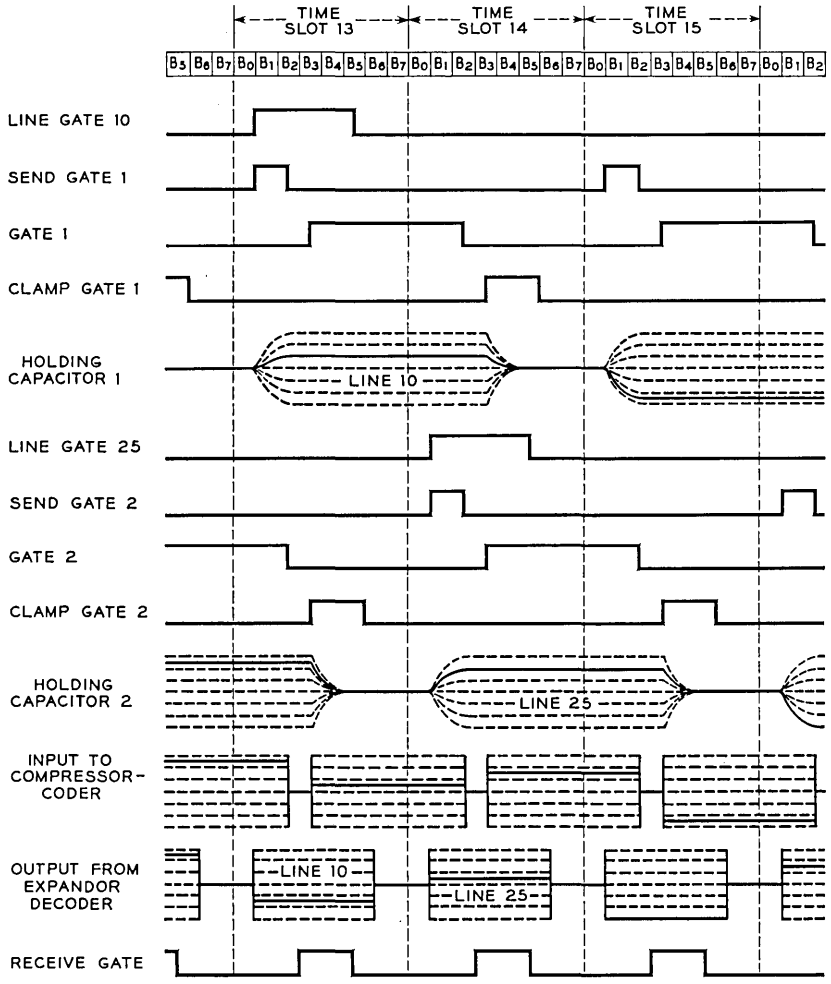

Fig. 8 — Timing of the time-shared hybrid.

charges the shunt filter capacitor. The line gate then opens. During the subsequent 122 microseconds this signal is dissipated in the subscriber's telephone so that very little remains to be sampled when the line gate and the send gate again operate together.

If the subscriber's filter is not matched correctly by the subscriber's telephone, the signal will be reflected and sampled by the send circuits. Thus, the time-division hybrid has properties similar to conventional hybrids except that it is lossless.

The actual circuit (Fig. 7) shows that resonant transfer of charge from the filter capacitor to a pulse-stretching or holding capacitor is used. The stored signal is then amplified and switched through another gate to the compressor. The amplifier has a high input impedance and is used to prevent discharge of the holding capacitor. When the coding action is completed, the stored energy is clamped out with another gate. While PAM signals are being coded in an odd time slot, the second send circuit is sampling in an even time slot.

In the receiving portion of the hybrid, the output of decoder and expandor is amplified by a common amplifier and passes through the receive gate, PAM bus and line gate, and charges the shunt filter capacitor of the line circuit to the value detected by the send sampling in the other concentrator. Resonant transfer is not used in the receive circuit, since any residual signal in the line package should be dissipated, thus terminating it. If the residual signal is dissipated, then the four-wire loop can have unity gain with misterminated line circuits and still not sing. This has been verified both mathematically and experimentally.[7]

Fig. 9 shows the hybrid timing circuits that control the operation of the seven gates in the hybrid. Each gate is driven by a flip-flop controlled transistor amplifier and each amplifier produces a 100-millampere pulse. Another flip-flop, connected as a binary counter, causes the two send circuits to operate in alternate time slots. An inhibiting signal from the inhibit flip-flop located in the timing circuits (Section IV, Fig. 3) prevents the send and receive gates from operating in time slot 23 when scanning takes place.

VIII. COMPRESSOR AND EXPANDOR

The send PAM signals are switched into a shunt compressor, which has a nonlinear resistance characteristic obtained by the use of carefully matched, temperature-controlled diodes. The network is designed so that the loss is increased with increasing signal amplitude. Large signals are attenuated by 26 db relative to small signals. This characteristic, to-

gether with linear encoding, has the effect of spreading small signals over a greater number of levels, thus reducing the quantizing noise.[8] In the receive circuits an expandor uses a similar, but series, network to obtain the inverse characteristic.

The actual circuits are shown in Fig. 10. The compressing network is followed by a high-input-impedance feedback amplifier, which provides sufficient signal for the encoder. The expandor network has both a preamplifier and a postamplifier. The preamplifier, which has a constant-voltage output characteristic, drives the series network, and the resulting current is amplified by a low-input-impedance postamplifier. The postamplifier is followed by the common medium-power amplifier in the hybrid, which charges the line circuit shunt filter capacitor.

IX. ENCODING AND DECODING

The output of the compressor is fed into a 128-level feedback-type encoder[9] (Fig. 11), which, in essence, consists of a summing amplifier

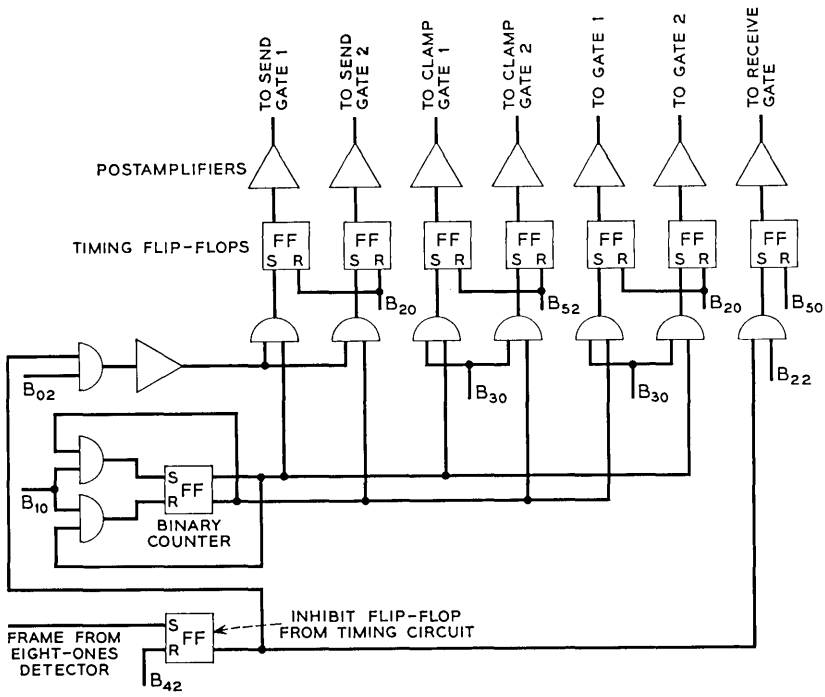

Fig. 9 — Hybrid timing circuits.

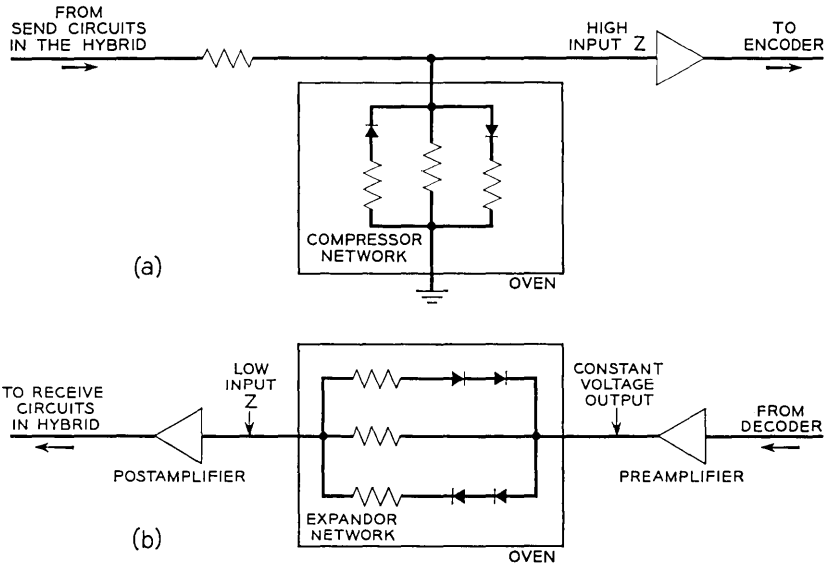

Fig. 10 — Compressor and expandor.

and a programmed decoder. Seven binary weighted resistors, of values R, $2R$, to $64R$, can be connected to either ground or a negative reference voltage according to the state of seven memory elements. The memory elements consist of blocking oscillators so connected that, when triggered, they will give out a series of pulses 0.3 microsecond wide every 0.65m icro-

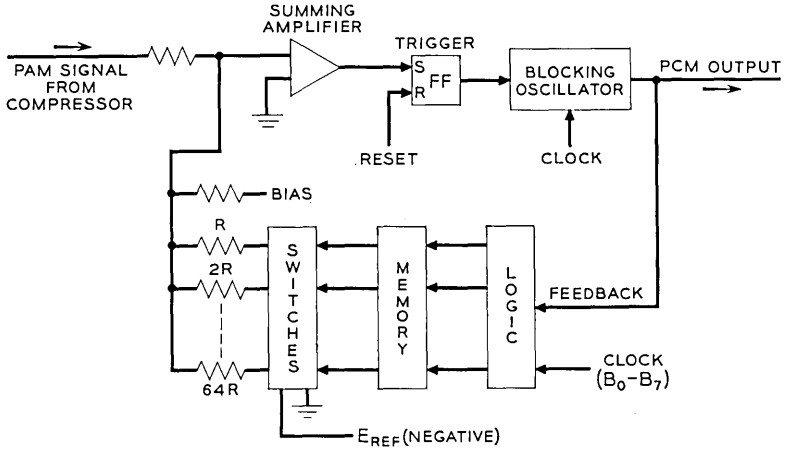

Fig. 11 — Encoder.

seconds. The blocking oscillators can be stopped if an inhibiting signal is applied immediately after the first pulse.

The PAM signal from the compressor is fed into a summing amplifier. A fixed bias raises this signal so that a zero-level input signal will be encoded as level 64 and a maximum amplitude negative signal as level zero. A third input comes from the binary weighted network, whose resistors are successively connected to either a negative reference voltage or ground until the total contribution from the three sources, i.e., signal, bias and network, is zero. Fig. 12 illustrates the encoding of a PAM signal corresponding to level 91.

During the first bit time of the coding interval, the resistor of value R is connected to ground, all other resistors being connected to the negative reference voltage. If the resultant signal to the summing amplifier is positive, then the trigger circuit and output blocking oscillator fire, sending out the most significant PCM bit. A feedback signal to the logic and then to the memory element causes the most significant bit resistor to be reset; i.e., it will be connected to negative for each successive trial. If the resultant signal to the summing amplifier is negative, then a zero would be sent out for the PCM bit and the feedback would cause the

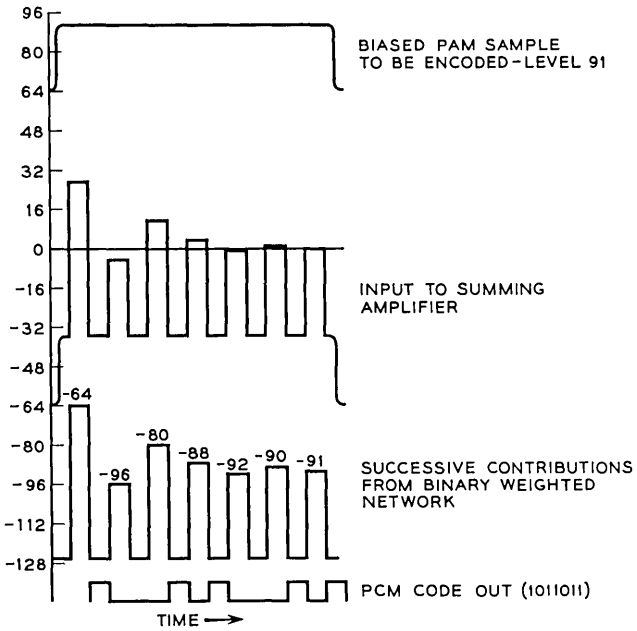

Fig. 12 — Waveforms in the encoder.

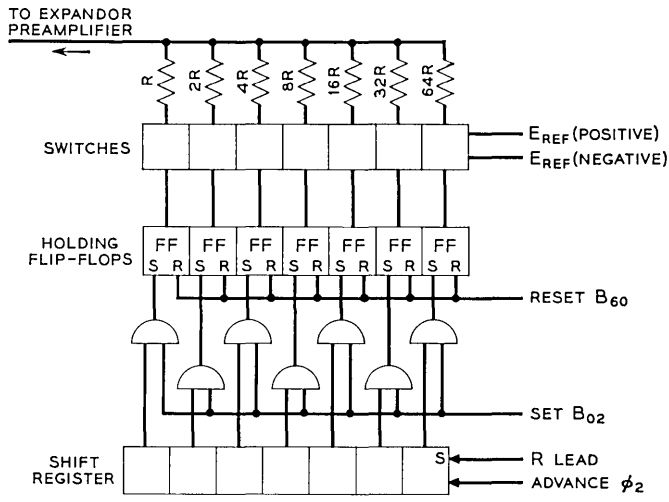

Fig. 13 — Decoder.

contribution from the most significant resistor to remain. In the successive bit periods, resistors of value $2R$, $4R$, through $64R$ are tried until the final code is reached and the resultant signal to the summing amplifier is zero.

The decoder is shown in Fig. 13. It consists of a shift register that converts incoming PCM code on the R lead from serial to parallel. When the code has completely stepped into the register, the bits are read out in parallel into holding flip-flops, which, in turn, operate transistor switches. The switches connect binary weighted resistors R, $2R$, through $64R$ to either negative or positive reference voltage according to the state of the holding flip-flops. The resultant decoded signal is then amplified by the expandor preamplifier, expanded and delivered to the hybrid. The holding flip-flops in the decoder are reset after a $5\frac{1}{2}$-bit time interval.

X. SCAN RESULT INJECTION

The S lead carries two types of signal; seven bits of PCM-coded speech in time slots 0 through 22 and the scan result in the bit 7 position of time slot 22 (Table I). The scan result is generated on the scan bus as soon as the line gate is operated. If a line is to be scanned, there will be a bit in the bit 7 position on the R lead. If this number corresponds to an active line, the bit is in the assigned time slot. If the number is inactive, both the address and the scan command, part of the framing signal, are in the last time slot. In either event, the result of scanning is gated to the "set" input of the scan flip-flop, Fig. 14, which stores the information until the bit

7 position of time slot 22 of the s lead words. At this time, the state of the scan flip-flop is interrogated and sent out over the s lead. The scan flip-flop is reset a few bits later under the control of the inhibit flip-flop signal. The two signals, seven-bit PCM plus the scan result, are amplified and pulse shaped, and drive the s lead cable pair.

In addition to scan result injection, one can add other signals in time slot 23 that would monitor the state of the remote line concentrator.

XI. LABORATORY MODEL

The laboratory model of the remote line concentrator, shown in Fig. 15, is contained in two 39-inch cabinets. The left-hand cabinet contains the subscriber line circuits and the 1-out-of-256 selector; the right-hand cabinet contains the remainder of the equipment. The division of the experimental model of the remote line concentrator into two parts was a result of the way in which the experiment was implemented. This implementation was divided into three parts, called Phases 1, 2 and 3.

In Phase 1, two cabinets were made, each of which contained subscriber line circuits, time-shared hybrids and selectors. The units were tested by connecting them with coaxial cables that carried the PAM signals. In Phase 2, the encoders, decoders, expandors, compressors, timing circuits, delay lines and a primitive form of central stage switch were added. Additions were made on a circuit-by-circuit basis — checking, testing and consolidating before moving on. The work on the remote line concentrator was essentially completed at the end of Phase 2. Phase 3 consisted of adding the concentrator controller, central stage switch and common control simulator.

This procedure permitted circuits and features to be added to the unit as required. The completed portion of the concentrator was therefore kept on a solid basis, which enabled it to be used as the test facility for

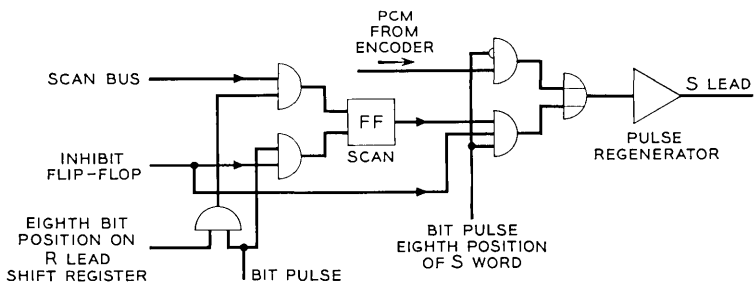

Fig. 14 — Scan result injection circuit.

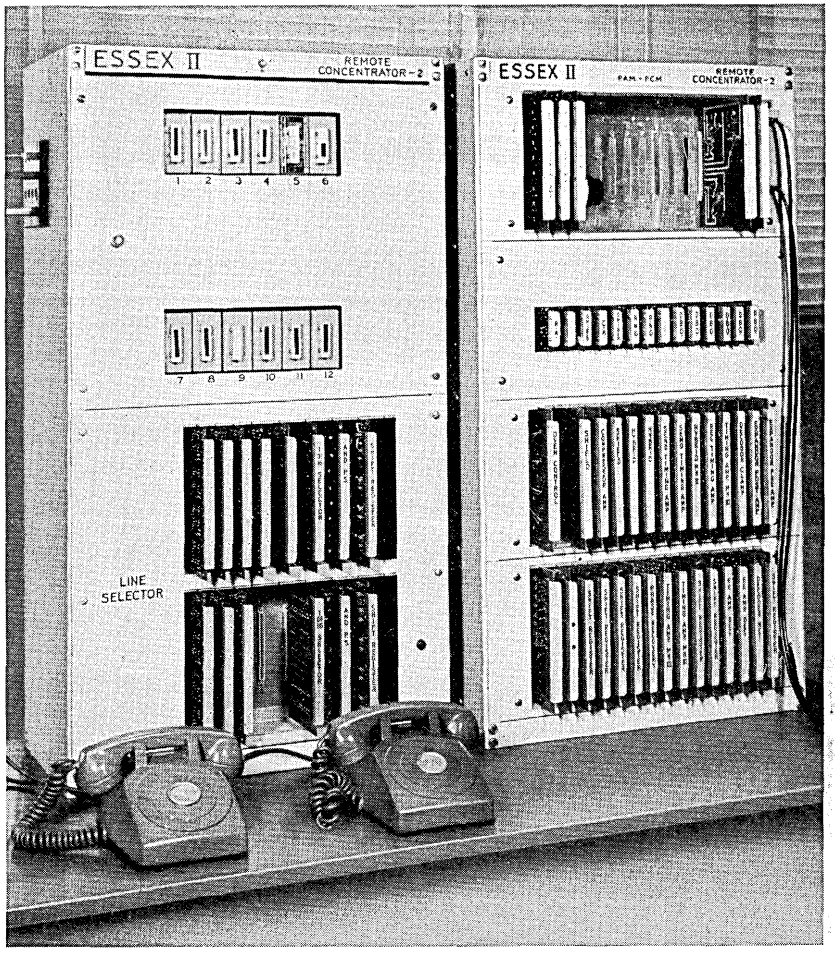

Fig. 15 — Remote line concentrator.

new circuitry and eliminated the need for synthesizing environmental conditions. One of the most attractive features of this procedure was the elimination of that most trying period that occurs when individually constructed circuits or groups of circuits are completed and then assembled and welded, hopefully, into a complete unit.

The laboratory model contains all the circuits and features required by the system, except for the number of subscriber line circuits installed. Twelve subscriber line circuits are installed in each remote line concen-

trator. Provision was made so that these could be connected to any number point on the 1-out-of-256 selector by means of a plugboard mounted in the back of the cabinet (Fig. 16).

The problems of equipment design were considered an integral part of the ESSEX experiment. Rigid standardization of layout and wiring was not employed, since it was recognized that new ideas and techniques would occur and become available during the course of the experiment. When any new ideas or techniques were proven satisfactory they were incorporated in the later units, no attempt being made to change or rework any of the finished units. The rear view, Fig. 16, of the remote line concentrator at an early stage of the experiment is an excellent example of this. The rack, on the right, which contains the subscriber

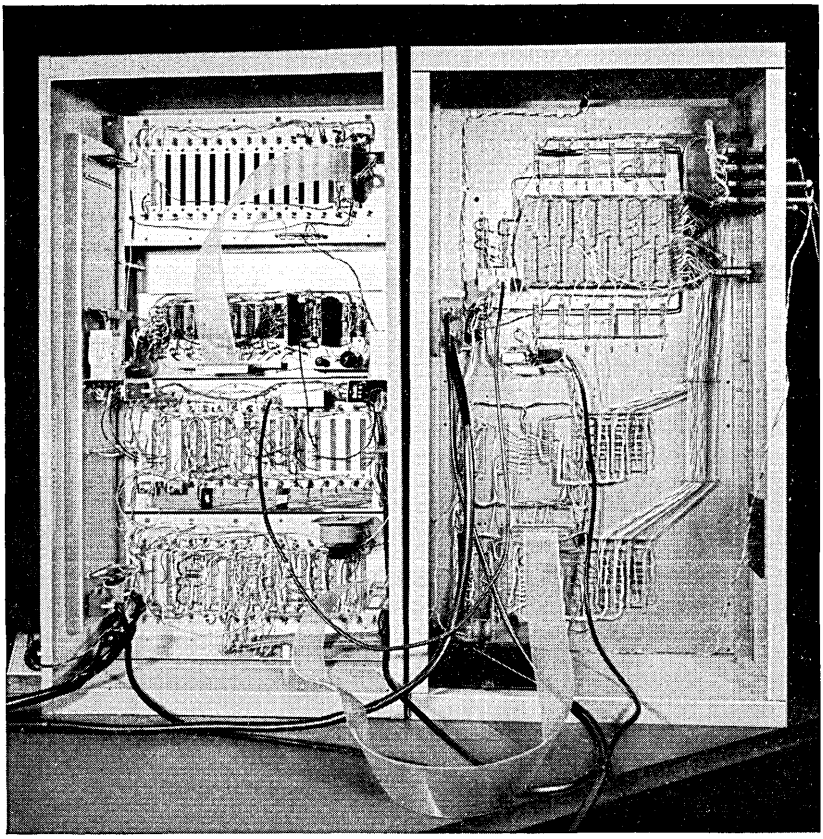

Fig. 16 — Rear view of remote line concentrator.

line equipment and the 1-out-of-256 selector was the first unit made. This single-panel construction was found to be inflexible, and the subpanel arrangement shown in the left-hand rack, in Fig. 16, was adopted. This arrangement was used in the concentrator controllers, the trunkor unit, the trunkor controller, the central stage switch and the common control simulator.

Each subpanel can accomodate 15 printed circuit cards, which plug into printed circuit connectors. Interconnections between printed circuit connectors on the same subpanel were made with color-coded plastic-covered stranded wire, no particular attempt being made to keep these interconnections as short as possible. Power and clock pulses were brought into the subpanels by means of printed circuit cards, which plugged into multiples located in troughs on the left and right walls of the cabinet. Interconnections between subpanels and between cabinets were made with flat multiconductor cable. Several types of cable were used in order to determine the advantages and disadvantages of each.

The individual circuits are mounted on 5- by 8-inch gold-plated printed circuit cards. Most of the analog circuits and some special purpose digital circuits, such as the encoder and 1-out-of-16 selector, are located on cards laid out for the specific circuit or groups of circuits involved. Special-purpose cards were not used for the digital circuits. Instead, three cards were provided, one containing five flip-flops, another holding eight pulse amplifiers and the third being laid out to accomodate AND and OR logic circuits and emitter followers. The digital circuits were formed by cross-connections on the printed circuit connectors on the rear of the panel. To make this scheme work effectively, each part of a digital circuit was given a number that located the card and the position location of the circuit on the card. Examples of typical printed circuit cards are given in Fig. 17. The card in the upper left-hand corner contains eight pulse amplifiers, the card in the upper right-hand corner contains the 1-out-of-16 selector. The lower card contains five flip-flops. Wherever possible, the circuit diagram, input and output connections and supply voltage busses have been clearly marked. This relatively minor detail has been a great aid and time saver in fault location.

The logic circuits used were mostly AND and OR circuits in conjunction with flip-flops and amplifiers. These circuits used the fastest transistors and diodes commercially available at the start of the experiment. Even so, in order to obtain the 50-millimicrosecond rise and fall times required, clipping and clamping techniques had to be used. This can be seen in Fig. 18, which shows the circuit diagrams of the pulse amplifier and flip-flop used in the experimental system.

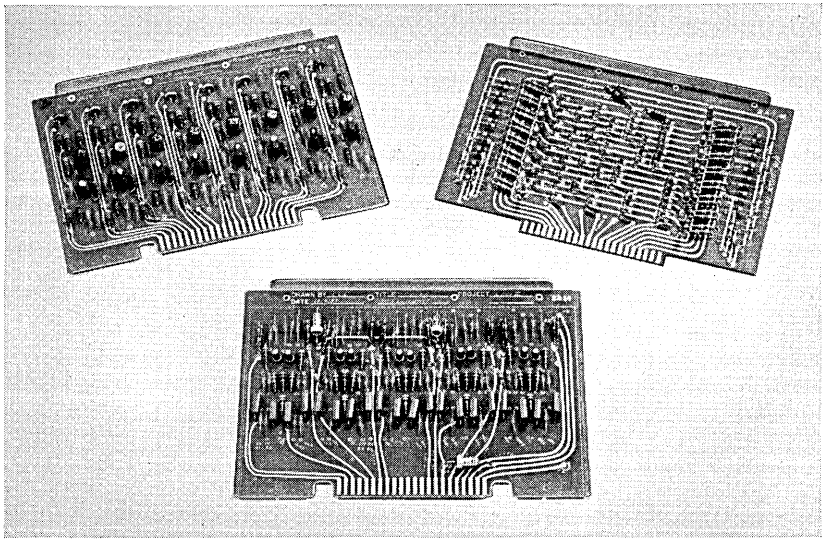

Fig. 17 — Printed circuit cards.

The power required by the remote line concentrator, exclusive of telephones, is 75 watts. The transistorized telephones used in the experiment required two-thirds of a watt each. Methods of powering the remote line concentrator have been considered, but not in any great detail. Power for the experimental system was obtained from a bank of commercially available transistor-regulated supplies.

The measured terminal-to-terminal transmission characteristic of the system is 6 $\pm$ 0.5 db from 100 to 3200 cps, and the 9-db points are at 70 and 3500 cps, with very small variability from channel to channel. The midband loss of the system was set at 6 db for convenience, and does not represent the minimum loss obtainable in a stable system. The minimum loss obtainable is twice the loss in a subscriber line circuit.

The seven-bit PCM with logarithmic compression gives a signal-to-noise ratio of about 30 db for a large dynamic range of signals. A simple demonstration of the noise introduced into the system by the PCM encoding and decoding has been incorporated into the experiment. This is done by operating a key, which drops out the PCM and connects the terminals of the filters on one channel by means of a physical pair. The difference is observable only in a very quiet room. Usually, however, the quantizing noise is almost completely masked by room noise.

The measured crosstalk level is 65 db down from an adjacent channel in the same remote concentrator. As explained earlier, the analog cir-

cuits and the high-level pulses present at the output of the 1-out-of-256 selector must be laid out very carefully in order to achieve this result.

At the time this paper was written, the remote line concentrator had been operating for a period of ten months. During the last two months it was operated over a 25-mile microwave link between Murray Hill and Holmdel, New Jersey. Since only one microwave channel was available from Murray Hill to Holmdel, the R and C leads were multiplexed onto the channel at Murray Hill and demultiplexed at Holmdel.

An artist's sketch of what an ESSEX remote line concentrator might look like when condensed into a single unit is shown in Fig. 19. In this conception, the subscriber line equipment is installed in the doors, with 128 subscriber line circuits in each door. Traffic consideration in an operating system would require that about 115 subscribers be connected to a remote line concentrator. In this case, one door could be replaced by a cover plate. The central portion contains all the shared circuits of

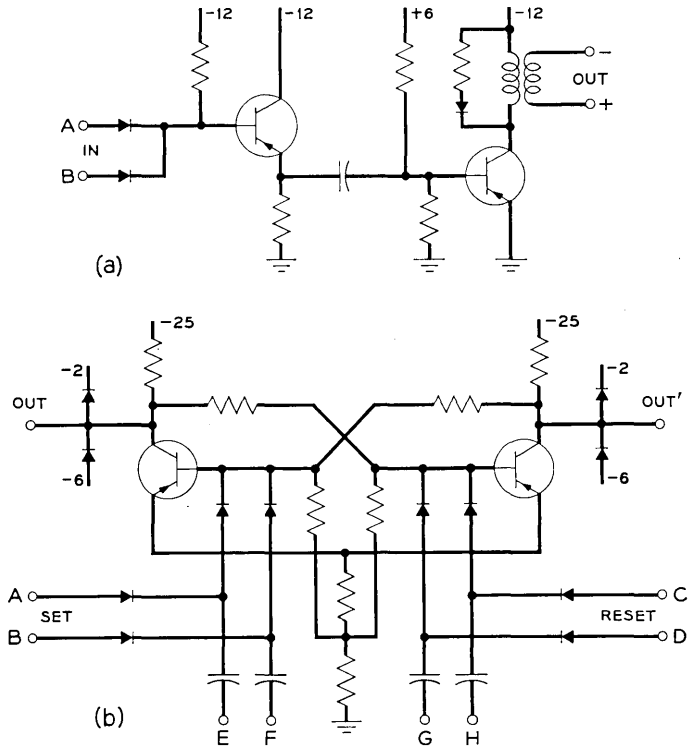

Fig. 18 — Pulse amplifier and flip-flop circuit diagrams.

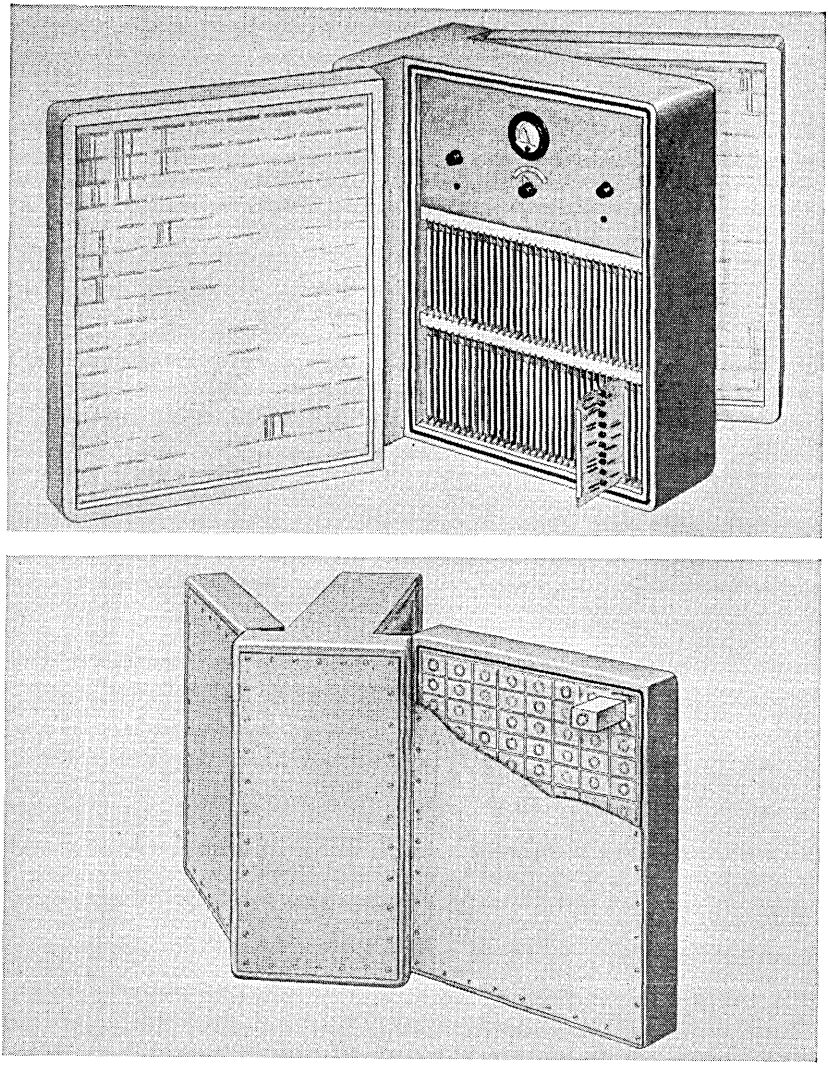

Fig. 19 — Artist's conception of remote line concentrator.

the unit and space for a remote power supply and four-hour battery standby if required. The unit, as pictured, would measure 16 × 27 × 30 inches with both doors closed, and would weigh approximately 250 pounds. The size and weight figures were obtained from the equipment used in the experimental version of the remote line concentrator.

XII. SUMMARY

Two remote line concentrators and a trunkor have been built and have operated successfully in the experimental environment. The units work at the required speeds. The design is straightforward. The implementation presented few problems of any magnitude and required no change in the basic system plan. This, in very large measure, is a result of the fundamental simplicity of time-division switching and digital communication systems.

XIII. ACKNOWLEDGMENTS

Many of the features and circuits of the concentrator are due to the efforts of T. H. Crowley, J. D. Foulkes, W. G. Hall, W. A. Malthaner and J. H. McGuigan. The suggestions and encouragement of E. B. Ferrell, W. D. Lewis and H. E. Vaughan contributed much to the work. A special note of thanks is due J. F. Muller and his group for the equipment design job, and our associates in the Systems Research Department for their many contributions.

We are indebted to other departments in Bell Telephone Laboratories for their help, and especially H. M. Straube and C. P. Villars of the Transmission Systems Development Department, who provided the basic designs for the coding and companding equipment.

REFERENCES

1. Vaughan, H. E., Research Model for Time-Separation Integrated Communication, B.S.T.J., **38,** July 1959, p. 909.
2. Malthaner, W. A. and Runyon, J. P., this issue, p. 59.
3. James, D. B., Johannesen, J. D. and Myers, P. B., A Two-Transistor Gate for Time-Division Switching, I.R.E.–A.I.E.E. Transistor and Solid State Circuits Conf., February 1958.
4. De Soer, C. A., A Network Containing a Periodically Operated Switch Solved by Successive Approximations, B.S.T.J., **36,** November 1957, p. 1403.
5. Crowley. T. H., private communication.
6. Meacham, L. A., Power, J. R. and West, F., Tone Ringing and Pushbutton Calling, B.S.T.J., **37,** March 1958, p. 339.
7. Crowley, T. H., private communication.
8. Smith, B., Instantaneous Companding of Quantized Signals, B.S.T.J., **36,** May 1957, p. 653.
9. Villars, C. P., Design of Transistorized 1.5 Megabit Analog to Digital Encoders, I.R.E.–A.I.E.E. Solid State Circuits Conf., February 1959.

Controller for a Remote Line Concentrator in a Time-Separation Switching Experiment

By W. A. MALTHANER and J. P. RUNYON

(Manuscript received August 26, 1959)

Remote line concentration, time-separation switching and PCM transmission are combined in a communication system experiment called ESSEX (Experimental Solid State Exchange). Organization and design details of the concentrator controller used in the research model are presented and discussed.

I. INTRODUCTION

An earlier paper[1] has described the general organization of an experiment in time-separation switching. Two principal functional parts were incorporated into this experiment: a line concentrator, which might be remote from a central switching office, and a concentrator controller, which would be located within the central office. In this paper we will describe in greater detail the organization and design of the concentrator controller. A companion paper[2] presents details of the remote line concentrator. In both papers it will be assumed that the reader is generally familiar with Ref. 1.

The experimental equipment provides 24 time-division channels between the remote concentrator and the controller. Since the controller would be located at a central switching point, it would be more accessible for maintenance than the concentrator; therefore, whenever there was a choice between locating equipment in the concentrator or in the controller, the controller was chosen.

Of the 24 time-division channels, which will be referred to as *time slots*, the first 23, numbered 0 through 22, are used for pulse-code-modulated speech. Time slot 23 is reserved for scanning and control functions, which will be described in the body of this paper. Each time slot recurs at an 8-kc rate, and each contains eight binary digits; the first seven are

pulse-code-modulated speech, and the eighth is reserved for control purposes. One cycle of time slots, 0 through 23, will be referred to as a *frame*. There are 8000 frames per second, 192,000 time slots per second, and 1.536×10^6 binary digits per second per repeatered transmission pair.

Each concentrator controller is interposed between its remote concentrator, the central-stage switch and a common control circuit that would serve a number of concentrators and trunkors.[1] The controller stores and delivers to the concentrator the information needed to control the gating of speech samples, delivers to the central-stage switch the information needed to gate PCM speech signals over the appropriate routes and maintains supervision of the subscribers' lines, which terminate in the concentrator. In performing this last duty, it scans each line every one-eighth second, and, upon two successive appearances of an "off-hook" signal from an idle line or an "on-hook" signal from a busy line, the controller sends an alerting signal to the common control. In the experimental arrangement, the common control is simulated by a manually operated console; hence, we will describe the steps taken in setting up and taking down connections in terms of the actions of an operator at the console.

II. SIGNALS TO AND FROM THE CONTROLLER

The environment of the controller is shown in Fig. 1. The concentrator and the switching center are connected by three repeatered pairs, one (shown as S), transmitting speech and line scan results toward the center, one (shown as R) transmitting speech and framing signals toward the concentrator and the third (shown as C) transmitting, toward the concentrator, control information identifying in real time the subscriber line that should next be sampled. The concentrator control also delivers, in real time, signals for the control of the gates in the central stage switch (see Section VII). The experiment employs a two-stage central switching network with a contemplated capacity of about 32 concentrators plus trunkors, whose S and R leads are interconnected by 32 junctors. In this case, five binary digits suffice to identify a junctor. Hence, the controller must deliver a five-binary-digit junctor gate number (JGN) per time slot to the network and an eight-binary-digit line gate number (LGN) per time slot to the C lead.

The gate numbers that are dispatched over lead C and toward the network are stored in the controller in circulating memories (see Section V). When a telephone connection is made or broken, gate numbers must be read into or erased from these memories. This information is not gener-

ated within the controller, but must be supplied from the manual control. Communication between the manual control and the concentrator control is provided by means of insert and dispatch circuits (see Sections VIII and IX).

The scanning of subscriber lines in the remote concentrator is carried out by the concentrator control with the assistance of an external scan number generator (see Section IV), which delivers subscriber line numbers to each concentrator controller. These line numbers are delivered serially, in cyclic order, and synchronously with the line numbers dispatched over the c lead. Each number is repeated once per time slot for four entire frames before the number generator advances to the next number. The signal delivered to the controller by the scan number generator is called the scan gate number (SGN).

Since there are eight bits per word, there are $2^8 = 256$ SGN numbers to be generated and, since the number generator advances every fourth frame, there is an interval of approximately one-eighth second between the periods during which a particular number is generated.

The use of a four-frame cycle for the scanning of a single line number

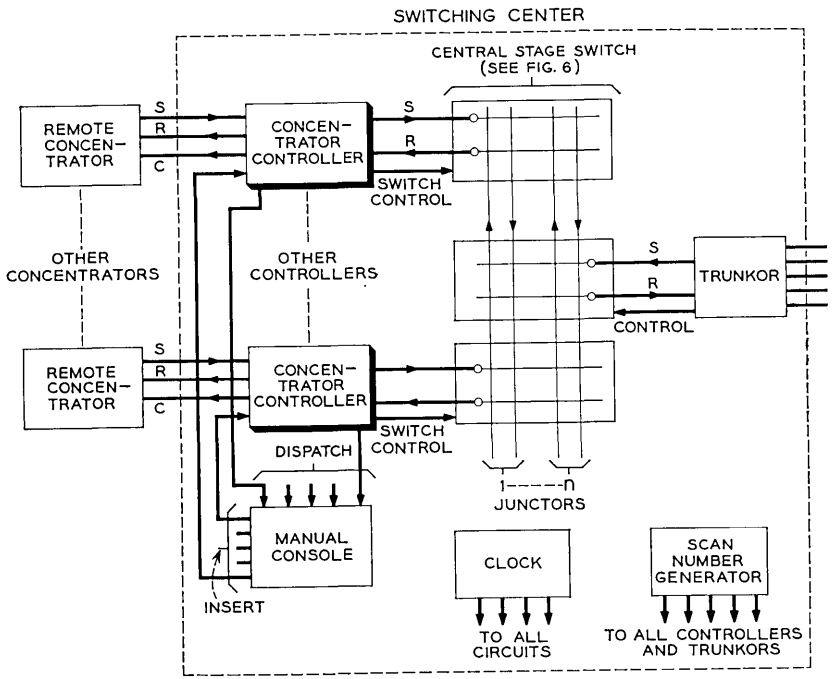

Fig. 1 — The environment of the controller.

gives rise to a natural numbering of frames when scanning is discussed. The scan number generator advances from one number to the next at the beginning of frame 0; delivers the same number throughout frames 0, 1, 2 and 3; and advances to the next number at the end of frame 3, which coincides with the beginning of frame 0 for the next scanning cycle.

We have now mentioned every category of signals that goes to or from the concentrator controller except one. The signals in this final category are timing signals that derive from a clock, used in common by all units at the switching center, which consists of a stable oscillator and attendant pulse counters (see Section XIII).

A list of the distinct signals that flow to or from the concentrator controller is given in Table I.

III. CALL PROGRESS WORDS

The signals listed in Table I describe the performance of the controller as seen from its terminals. The general arrangement of equipment that was chosen to meet these requirements is shown in Fig. 2. (Fig. 2 is a reproduction of Fig. 6 of Ref. 1.)

Digital operations within the controller are governed by the call progress memory, the call progress coder-decoder and the line scanning control. An understanding of the plan of operation of the controller is most easily obtained by considering the sequencing of call progress words in the call progress memory.

A *call progress word*[3] is a record of the state of a call to which a time slot has been assigned. The *call progress memory* is a circulating memory with a capacity of 192 binary digits, eight for each of the 24 time slots. In these locations are stored call progress words for the corresponding time slots.

Sequences of states through which a given call can progress, as reflected by the corresponding sequences of call progress words, are shown in Fig. 3. The call progress words themselves are indicated by the rectangular boxes of Fig. 3, the various admissible transitions from one call progress word to the next being indicated by the labeled arrows. Transitions that take place without the intervention of the operator are indicated by circles, and those that are caused by operator action by diamonds. Each automatic transition bears also a label made up of some or all of the letters H, M and D, with or without primes. These letter labels indicate, in a Boolean notation, the circumstances in which the transition occurs. The letter symbols stand for the following propositions:

H: "the scanner indicates off-hook";
M: "the SGN matches the LGN";
D: "the dispatch circuit is not occupied".

A prime indicates the negation of the proposition. Thus, the transfer from word P11 to P17, which is labeled $H'MD$, occurs when the scanner does not indicate off-hook, the SGN matches the LGN and the dispatch circuit is free.

Whenever the line scanner returns an "off-hook" indication for a hitherto idle line, the number of the line is entered in the line gate number memory in the first idle time slot encountered. This time slot is identified as idle by the appearance of P1 as its call progress word. At the time of entry of the line gate number, the call progress word is advanced to P2 ("suspect request for service"). No further action is taken until one-eighth second later, when the scan number generator next arrives at this line number. If a second "off-hook" is seen for this line at this time, the call progress word is advanced to P3, provided that the

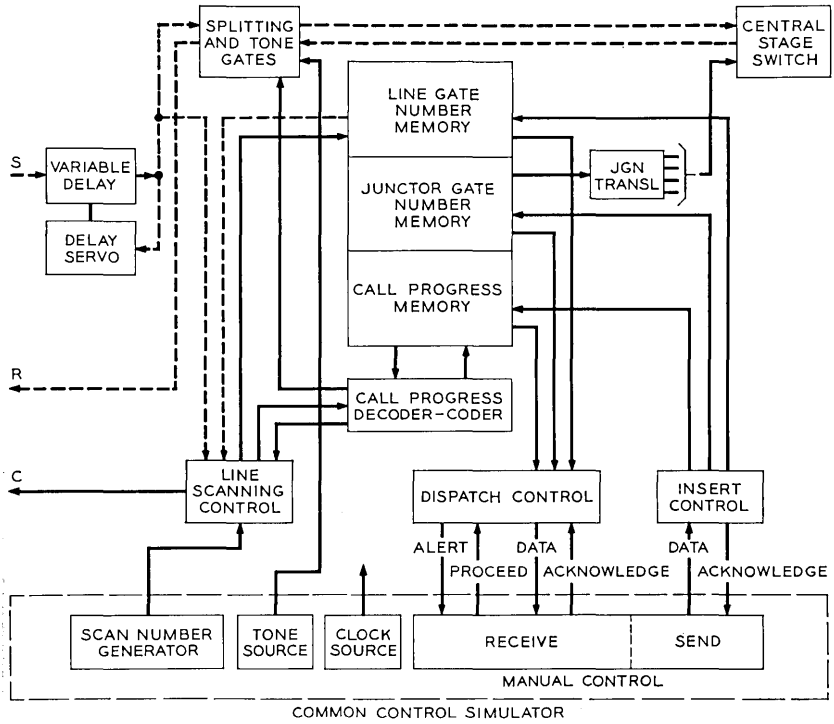

Fig. 2 — ESSEX concentrator controller.

TABLE I — SIGNALS THAT ENTER OR LEAVE THE CONCENTRATOR CONTROLLER

From or to	Name of signal	Sent via (See Fig. 3)	Time when sent or received by controller	Purpose
A. To concentrator	1. Line gate number (LGN)	C lead	Bits 0 through 7 of time slots 0 through 22 in every frame	Orders operation of line gate, thereby sampling
	2. Scan gate number (SGN)	C lead	Bits 0 through 7 of time slot 23 in frame 0	Scans idle lines to determine whether they are "off-hook"*
	3. PCM speech or tone	R lead	Bits 0 through 6 of time slots 0 through 22 in every frame	Delivers speech or tone signal to decoder
	4. Scan command	R lead	Bit 7 of time slots 0 through 22 in frame 0	Adds scan order to signal A1*
	5. Framing command	R lead	Bits 0 through 7 of time slot 23 in every frame	Orders reset of counters in slave clock
B. From concentrator	1. PCM speech	S lead	Bits 0 through 6 of time slots 0 through 22 in every frame	Transmits speech signal from encoder
	2. Scan result	S lead	Bit 7 of time slot 22 in frame 1 or 2	Indicates whether line scanned (A2 or A4) was "off-hook"
C. To central stage switch	1. PCM speech or tone	S lead	As in A3	Transmits speech or tone signal to other end of connection
	2. Junctor gate number (JGN)	Five parallel leads shown in Fig. 2	Bit 7 of time slots 23 through 21	Operates proper network gates for call in next-following time slot
D. To manual console	1. Request-for-service alert	Dispatch circuit	When dispatch circuit is free and two successive scans of an idle line have shown it to be "off-hook"	Informs operator that request for service has been detected and indicates time slot tentatively assigned the calling line

	2. Readout of controller memory data†	Dispatch circuit	After E1, memory contents in chosen time slot is read to manual console once per frame until receipt of E3	Informs operator of contents of controller memory in time slot interrogated
	3. Hang-up alert	Dispatch circuit	When dispatch circuit is free and two successive scans of a busy line have shown it to be "on-hook"	Informs operator that a hangup has occurred and indicates its time slot
	4. Acknowledge	Dispatch circuit	Response to E2	Informs operator that insert data was received with correct parity
E. From manual console	1. Proceed	Insert circuit	Under control of operator	Evokes D2; normally a response to D1 or D3
	2. Insert	Insert circuit	Time slot selected by operator	Permits operator to alter contents of controller memory
	3. Acknowledge	Insert circuit	Response to D2	Indicates D2 received with correct parity

F. Timing signals received regularly from clock and numbers for scan received regularly from scan number generator (see Sections IV and XIII).

* Either A2 or A4 is used, not both (see Section IV).

† One time slot may be read out per interrogation; this signal is a response to E1.

dispatch circuit is free. Advance of the call progress word to P3 initiates signal D1 of Table I, informing the manual console of line action. Had the dispatch circuit not been free when the second "off-hook" was seen, the call progress word would have remained P2 until either the calling subscriber returned the telephone to its hook, (in which case the word would be restored to P1), or the dispatch circuit became free, (in which case the word would be advanced to P3).

Advance from P3 to P4 takes place as follows: entry to P3 causes the signal D1 (Table I) to be sent to the manual console, informing the operator that a request for service exists, and indicating the time slot in which this request is being handled. In response, the operator returns signal E1, requesting transmission of the memory content in the indicated time slot. Response D2 automatically follows the receipt of E1, and, if the information dispatched by D2 is received with correct parity at the console, acknowledgment E3 is automatically returned to the controller. The receipt of E3 advances the call progress word from P3 to P4. This frees the dispatch circuit and, if another time slot is at P2, it may now advance to P3.

When the call has reached P4, the operator knows the identity of the calling line and the time slot that has been tentatively assigned to it. The operator now must consult her records to see what sort of service this line receives. Let us suppose the operator discovers that the line is to be connected to an operator trunk.* By means which need not concern us here, the operator must locate an idle operator trunk and match a path through the network between the trunkor in which the operator trunk is located and the concentrator controller, in a time slot in which both trunkor and controller are idle. For this purpose, the operator may consider the tentatively chosen time slot in which P4 is stored be idle, but the operator may find it impossible to match through the network in that time slot. Let us suppose that a match is finally found that uses a different time slot.

As soon as it has been determined that a new time slot is to be used, the new slot should be reserved immediately in the controller memory by advancing its call progress word from P1 to P26, in order to prevent its seizure by suspected requests for service in the interim between matching and the inserting of line gate and junctor gate data from the console. In a practical system, the matching would be done by an automatic circuit, not described in this paper, which would cause the transfer from P1 to P26 by means of a signal that is not listed in Table I.

* In most cases, the line should be connected to a dial register. There is no point in detailing here the process of accumulating the called number and passing it to common control or to the console operator, which this introduces.

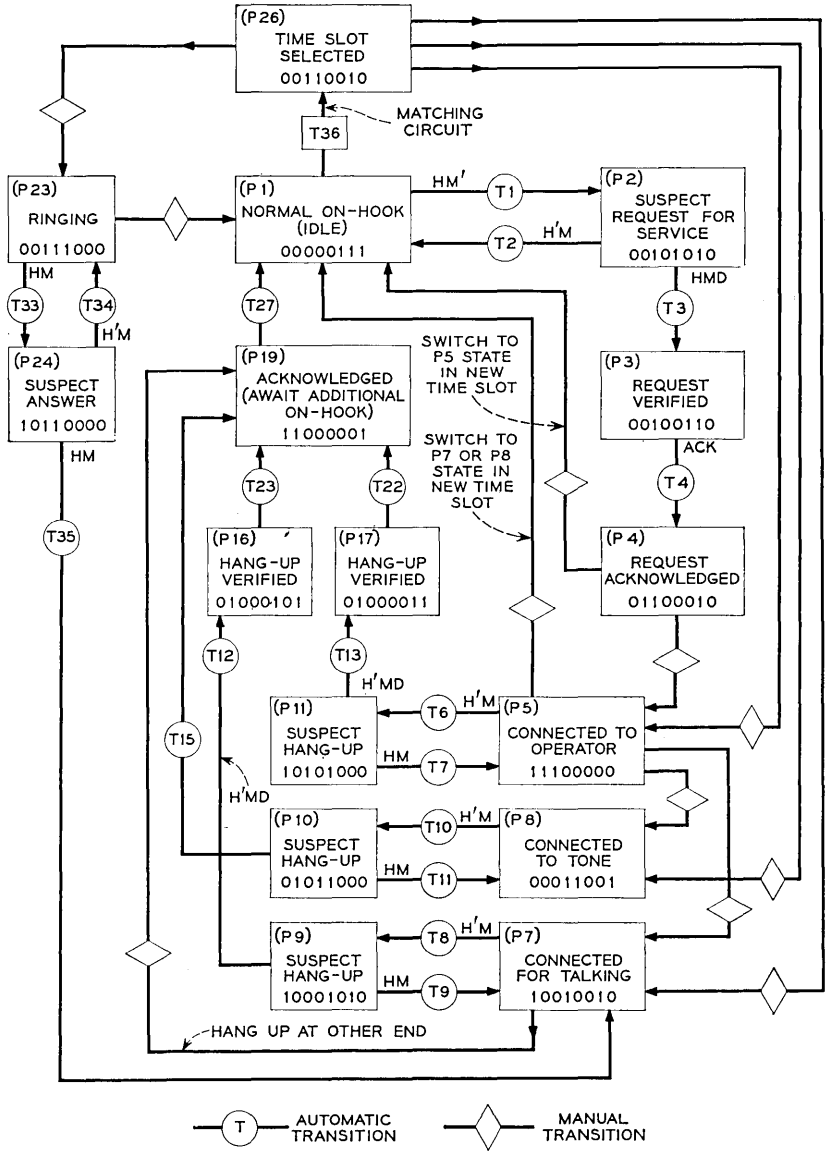

Fig. 3 — Call progress word state diagram.

To set up the connection, the operator first uses E2 to insert the line gate number, call progress word P5 and the junctor gate number yielded by the matching operation into the controller memory in the new time slot. This transmission is automatically checked for correct parity at the controller, and E4 acknowledges correct receipt (transition from P26 to P5). Signal E2 must then be used a second time to erase the information in the old time slot and restore its call progress word from P4 to P1.

From P3 and P4 there are no transitions that depend on the line scan information. If the calling subscriber should hang up while in P3 or P4, this fact would not be signaled to the operator until P5 was entered. This "filtering" of the scanner indications reduces the number of special sequences of events with which the common control must cope.

Once P5 has been reached, and the operator has made suitable entries in the trunkor memories to establish the other end of the connection, the subscriber is in voice communication with the operator. He may now indicate the number that he wishes to call. The operator now locates the concentrator in which the called line terminates and performs a busy test by means of 23 D2 interrogations, to determine whether the called line gate number appears in any one of the 23 working time slots of the terminating concentrator controller. If the called line is not thus found to be busy, the operator matches a path through the network in a time slot in which the two concentrator controllers are idle. Here again, the operator may consider the time slot in which she is connected to the calling subscriber to be idle; however, if the called concentrator is busy in this time slot, a new time slot may be chosen. If this is the case, the call progress word should be promptly advanced in both the originating and terminating controllers from P1 to P26 in the new time slot. The operator now inserts the line gate number and junctor gate number information in the chosen time slot in both controllers, inserting call progress word P7 in the originating controller and P23 in the terminating controller. Call progress word P23 causes operation of the splitting and tone gates (Fig. 3) in the terminating controller. These gates interrupt the S and R leads of the terminating concentrator in the chosen time slot and send a pulse-coded ringing signal on the R lead and a pulse-coded audible ringing tone on the S lead, which returns this tone through the network to the originating concentrator. If the called line is answered, the first "off-hook" signal from the scanning of that line causes transition from P23 to P24. In P24, the ringing signals continue. A second "off-hook" signal from the called line, one-eighth second later, causes advance of the call progress word to P7, which effects restoration of the

splitting gates to the normal condition, removing the ringing tone and establishing a talking connection.

If the called line is located in the same concentrator as the calling line, the equipment is so arranged that two time slots are required for the conversation. Both calling and called subscribers must be connected through the network, in different time slots, to the same "unterminated" terminal of a trunkor. Here both sets of PCM signals are decoded to analog samples that are stored in and read from a capacitor. This capacitor replaces the customary line filter and permits interchange of the samples between the two time slots. This arrangement, although exceedingly simple, is wasteful of transmission capacity. In a working system where such "intraconcentrator" calls occurred frequently, changes would be made to permit such calls to be handled in a single time slot.

If the called line is not answered, the terminating concentrator will remain at state P23 and the originating concentrator will remain at state P7 until the calling subscriber tires of waiting and hangs up. The first "on-hook" scan result for the calling line will cause an advance to P9, and the second will cause an advance to P16, provided that the dispatch circuit is not occupied. Entry to P16 will initiate a "hang-up alert" signaling sequence between the controller and operator console, which corresponds exactly to the "request-for-service alert" sequence between P3 and P4. When the controller finally receives a "parity correct" acknowledgment from the console, the call progress word advances from P16 to P19. This causes immediate erasure of the JGN entry in the controller memory, and will cause erasure of the LGN entry and transition to P1 as soon as an additional "on-hook" signal has resulted from scanning the calling line.

The operator still must disconnect the line being rung. To do this, she consults the information sent to her during the signaling sequence between P16 and P19. She knows the time slot in which the hang-up occurred and the number of the junctor in use for the call. The operator now must locate the controller or trunkor whose memory stores the same junctor number in the same time slot. To do this, she may, for example, interrogate in turn the memory of each controller or trunkor, in the time slot in question, until she finds the location at which the sought-for junctor number is stored. The operator now erases the memory contents at the point she has just located, and restores the call progress word from P23 to P1, thus ending the ringing of the called line.

This typical history of the progress of a call is intended to acquaint the reader with the general scheme of operation of the controller, as

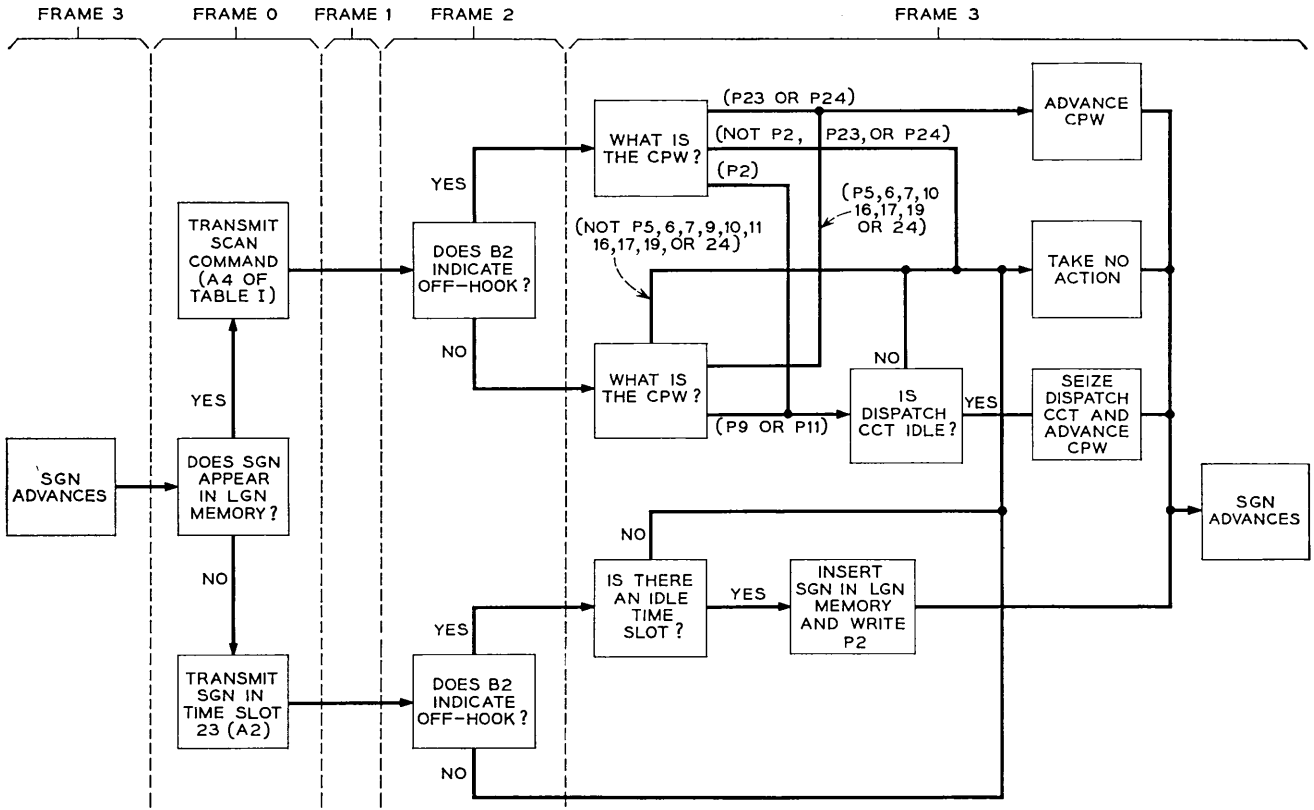

Fig. 4 — Sequence diagram of events and decisions in line scanning.

reflected in the sequence scheme of the call progress words and the manner in which they control the sequence of events in the system. From this description, a call progress word may be seen to be a sort of conditional "order," indicating the next operation that is to be performed as soon as the appropriate conditions have been established, rather than at some predetermined point in time.

We will leave to the reader the exercise of tracing out the call progress word sequences corresponding to other possible sequences of subscriber action, and will proceed to describe the circuits that have been provided in the experimental equipment.

IV. LINE SCANNING AND SUPERVISION

"Off-hook" and "on-hook" conditions at customer lines are detected by scanning sequentially all the lines at a remote concentrator. For each "off-hook" condition detected, a signal is transmitted to the concentrator central unit.

The scan number generator consists of an eight-bit shift register, called the SGN register, which is connected to recirculate its contents through an "add one" circuit. The number circulates through the SGN register every time slot (5.2 microseconds) and is available to the controllers for scanning and matching purposes. Each number circulates through the register without change 96 times, i.e., once each word time during four frames. After four frames, the "add one" circuit is enabled and the next higher binary number is entered into the SGN register. This four-frame scanning cycle allots one frame to a preliminary match of the scan gate number with the various line gate numbers in the concentrator memory, since the actions to be taken depend upon whether or not the line to be scanned was in an active state on the preceding scan. The round trip of the scan number from the concentrator controller to the concentrator and the return of the scan result may take one or two frames of the scanning cycle. The final frame is used for entering the line number into an idle time slot of the controller memory or for making any required alteration of the memory state if the line number was entered previously. Fig. 4 exhibits the various sequences of events encountered in line scanning.

The line scanning control circuit operates as follows. During frame 0 of the scanning cycle, the scan number is serially matched against the line numbers stored in time slots 0 through 22 in the controller memory. If no match is found, the line is idle and the scan number is transmitted on the c lead in the last time slot of this frame (time slot 23) to the concentrator (signal A2 of Table I). If the line is still on-hook, signal B2

is returned to the controller by the absence of a pulse, but if the line is "off-hook" B2 returns as a single pulse.

The "not-in-memory" and "off-hook" conditions are registered in flip-flops. Let us assume that B2 indicates "off-hook". During frame 3, the first idle time slot in the controller memory, as indicated by the call progress word memory, is assigned to the scanned line. The scanned number is gated from the scan number generator into the line gate number memory in this time slot and the call progress word is changed to indicate that an initial "off-hook" has been detected.

If the leakage of the remote line gates is not sufficiently low and an active line is gated to the scanner in time slot 23, and gated for speech in its assigned time slot, scanning noise in the form of "ticks of silence" may be introduced into the conversation. To prevent this, an active line is scanned in its assigned time slot.

If, during frame 0, the scan generator number is found to correspond to one of the line numbers in the controller memory, a "one" pulse (signal A4 of Table 1) is transmitted on the R lead in the eighth-bit position. To avoid interference with the framing operations (see Section XIV) at the concentrator the seventh, and least significant, speech bit in this and the next time slot are transmitted as "zero". The scan generator number in this case is not transmitted on the C lead in time slot 23. The presence of a pulse at bit 7 on the R lead causes the concentrator to scan the line in its assigned time slot. The signal indicating "off-hook", however, is still returned to the controller as bit 7 of time slot 22 of frame 3. During frame 3, the matching operation between the scan number and the line numbers in memory is repeated to identify the assigned time slot. The scan result is then gated into the call progress decoder-coder to alter as necessary the state of the scanned time slot.

By these processes, a line originating a call is identified, assigned to an available time slot and permitted, upon verification of the "off-hook", to alert the common control circuits of the office. In the same way, the answer of a called line is detected and the advance of the call from the ringing to the talking state is initiated. Similarly, the hang-up of a line at any time is detected, initiating the necessary disconnect operations, which may or may not require common control actions.

V. MEMORY STRUCTURE

The memory in each concentrator controller stores control orders for the remote and central switches. It also holds a record of the current status of each time slot. For each of the 24 time slots or channels, 24 bits of information are stored: eight bits for the remote line gate, five bits

for the central stage switch, eight bits for the call progress words and three bits for checking. Three recirculating serial memory units are used in each controller. Each concentrator memory circulates 192 bits—that is, eight bits for each of 24 channels.

Binary-coded pulses are transmitted through a magnetostrictive delay line that holds 177 of the memory loop bits. This line is a 3-mil supermendur wire with a solenoid around one end as an input transducer and a similar solenoid at the other end as an output transducer. Binary-coded electrical impulses are impressed magnetostrictively on the line at the input transducer and travel along the line in acoustic form. The magnetostrictive effect is used at the output solenoid to convert back to electrical impulses; thus, an electrical delay and information store is produced that is dependent upon the distance between the delay line transducers (see Fig. 12 of Ref. 1.)

The output of the delay line connects to the first stage of a shift register, and the last stage of the shift register feeds pulses back into the magnetostrictive line to complete the loop. A diagram of one of these circulating loops is shown in Fig. 5. Information appears in the central stages of the shift register during the particular time slot with which it is associated, recurring, of course, every 125 microseconds. The infor-

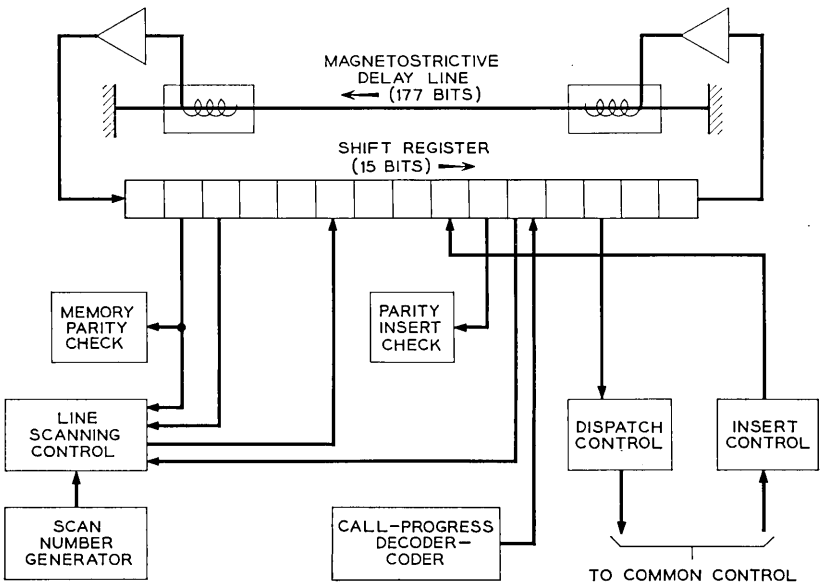

Fig. 5 — Line gate number memory unit.

mation may be examined and transmitted to external circuits in either serial or parallel form from the shift register. New information can be entered into the memory in serial or parallel form at the shift register at the half-bit time. A relative time shift in the information, causing it to lead or lag the occurrence of the master clock time slot with which it is associated, may be introduced during the design of any system function by operation at selected earlier or later stages of the shift register. This permits time coordination through the system without the use of auxiliary delay buffers or time-counting circuits.

In the model, line gate words are inserted serially from the line scanning circuit or common control and are transmitted serially to the line-scanning circuit, to the remote concentrator via the c lead and to common control. The junctor gate words are transmitted in parallel to the central stage switch selectors, transmitted serially to common control and inserted serially from common control. Call progress words are transmitted to and entered from common control serially, examined in parallel and altered in parallel by the call progress decoder-coder, as will be described in Section VI.

VI. MANIPULATION OF CALL PROGRESS WORDS

The call progress words are coded so that each word contains exactly three ones. There are 56 such words, but only about 30 of them are required in the model.

As each call progress word passes through the shift register of the memory loop, its eight bits are examined in parallel by a decoder-coder circuit, which consists principally of a diode matrix. The decoder energizes a single lead to indicate the state of the time slot, and actions such as the operation of ringing gates may result directly. Additional inputs from the line scan, dispatch or insert control circuits combine with certain indicated call progress states in the coder to produce a new call progress word. The word in the memory loop is immediately altered by parallel inputs to the shift register from the coder. The speed of the decoder-coder circuits is sufficiently fast that buffer storage between the shift register and the decoder is not required, and the word in the register can be altered by the coder before the occurrence of the next advance pulse. Certain control actions, such as the seizure of the dispatch control circuit, occur simultaneously with a coder alteration of the call progress word.

Call progress words are also examined serially for operations such as idle-time-slot selection during initial assignment of an originating call to a time slot, and for the path-matching operations of common control.

Call progress words may also be altered by serial transmission from common control via the insert control circuit to one of the later stages of the memory shift register.

VII. CENTRAL STAGE SWITCH

Each concentrator (or trunkor) connects to an associated section of central stage switching. The S and R leads of a concentrator are multipled to 32 concentrator-to-junctor gates, and the junctor outputs are multipled to the corresponding junctor outputs of all other concentrators (and trunkors) as shown in Fig. 6. Thus, each concentrator has access to the other concentrators and trunkors, and to junctors to other office modules[1] over a total of 32 space-separated paths in any of 23 time slots. For intramodule calls, the junctors are grouped in pairs. In a call between two concentrators the S lead of the first concentrator and the R lead of the second are connected to one junctor of the pair, while the R lead of the first concentrator and the S lead of the second are connected to the other junctor of the pair. The same time slot must be used in both concentrators.

The junctor gates are simple diode-and-transistor gates that pass binary signals unilaterally at low power. A five-bit transistor flip-flop register acts as a buffer between the circulating junctor gate memory of a concentrator and its junctor gate selector. The junctor gate selector is a two-stage diode AND matrix. Two of the input bits provide a one-out-of-four partial selection in one first stage, and the other three input bits provide a one-out-of-eight partial selection in a separate first stage. The one-out-of-four and one-out-of-eight partial selections are combined in the second selection stage to operate one of the 32 junctor gates, and the coded speech signals in both directions of a conversation pass each other simultaneously on the junctors.

VIII. DISPATCH CONTROL AND MANUAL CONSOLE

Each concentrator controller contains a dispatch control circuit through which it sends orders and data to common control. This communication is necessary at various stages in the progress of a call, such as on the detection of a request for service, answer, or hang-up condition at a customer's line. Common control correlates the information from the particular concentrator with information as to the rest of the office, and advances the call as required. In the experimental model, the dispatch circuit connects to the receiving circuit of a manual console. Information is displayed at the console to an operator who, with the

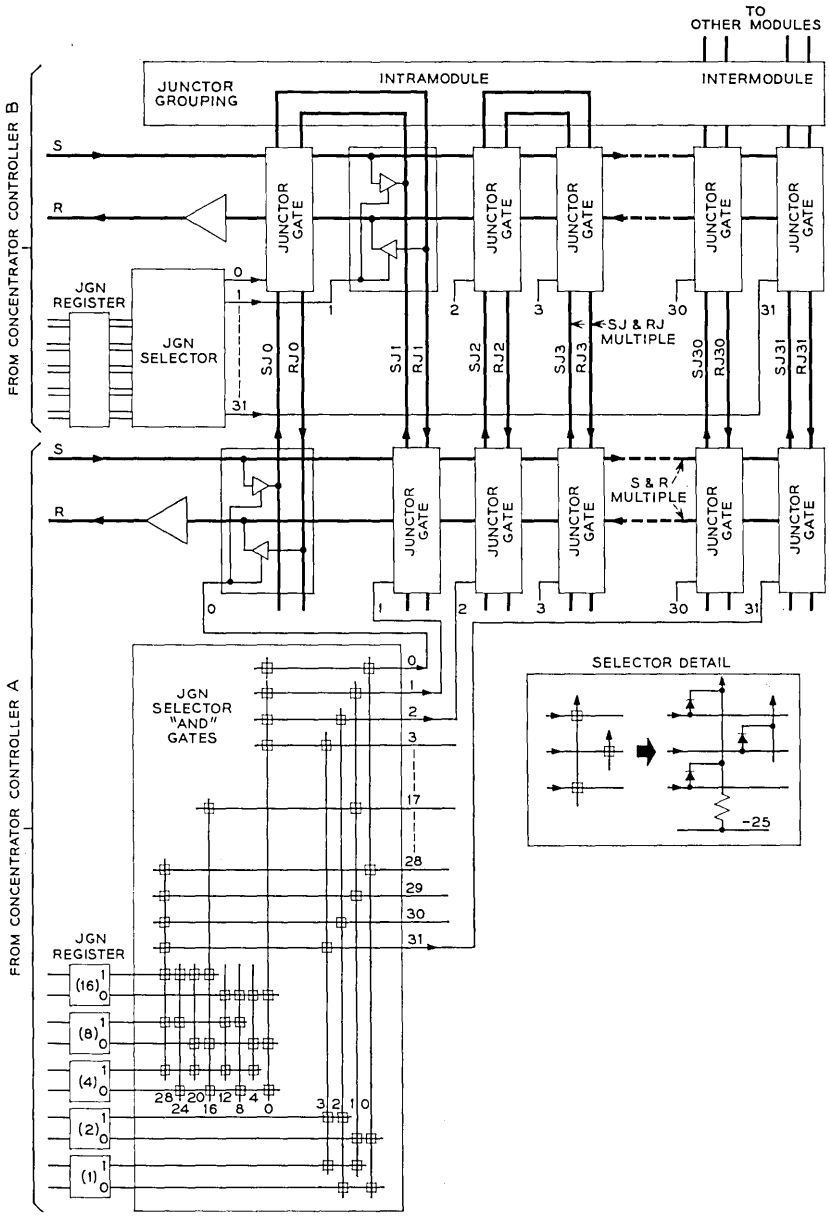

Fig. 6 — Central stage switch and selector.

assistance of manual aids to memory, makes the decisions and initiates the actions that would be performed automatically by common control in a fully developed system.

Functional relationships between the concentrator controller, its dispatch circuit and the receiving circuit of the manual console are outlined in Fig. 7. When the call progress word in a concentrator controller reaches a transition into a state that requires the use of the logic and memory functions of common control, the transition is made only if the dispatcher is idle, and the dispatcher is seized on the transition. This provides lockout between the various time slots of a concentrator in their access to the dispatcher. Lockout is required since the dispatcher may be used from any time slot and is held for a few frames on each usage. An alert register in the dispatch control is set from the call progress word decoder-coder upon seizure. The setting is identical to some of the bits in the new call progress word that the decoder-coder simultaneously writes in the associated time slot memory. An alert indicator in the console informs the operator that the particular concentrator requires attention and gives information as to the priority of the action required. The dispatch control on all subsequent frames indicates which time slot is in the dispatch state by means of a serial match between the alert register setting and the pertinent bits of each word in the call progress memory loop.

The console operator elects to serve the concentrator with the highest priority alert indication by setting the C/T selector switch to the corresponding position. To identify the alerting time slot, the setting of the console time-slot selector switch is varied until its output pulses coincide with pulses from the serial match detector. Readout keys enable the receiving sequence circuit, which in turn controls the dispatch sequence circuit, so that data from the alerting time slot of the concentrator memories are transmitted serially through dispatch data gates to data shift registers in the receiving circuit. The junctor gate number and line gate number information are read to the receiving registers consecutively in the frame following the operation of a readout key. If the order information in the alert indication is not sufficiently explicit, the complete call progress word may be transmitted, by operation of an alternate readout key, to a receiving shift register in the frame following the line and junctor number transmission. For maintenance purposes, provision has also been made for transmitting information from the memory loops to the data registers in nonalerting time slots by operation of special readout keys.

Each block of data transmitted to the receiving circuit is preceded

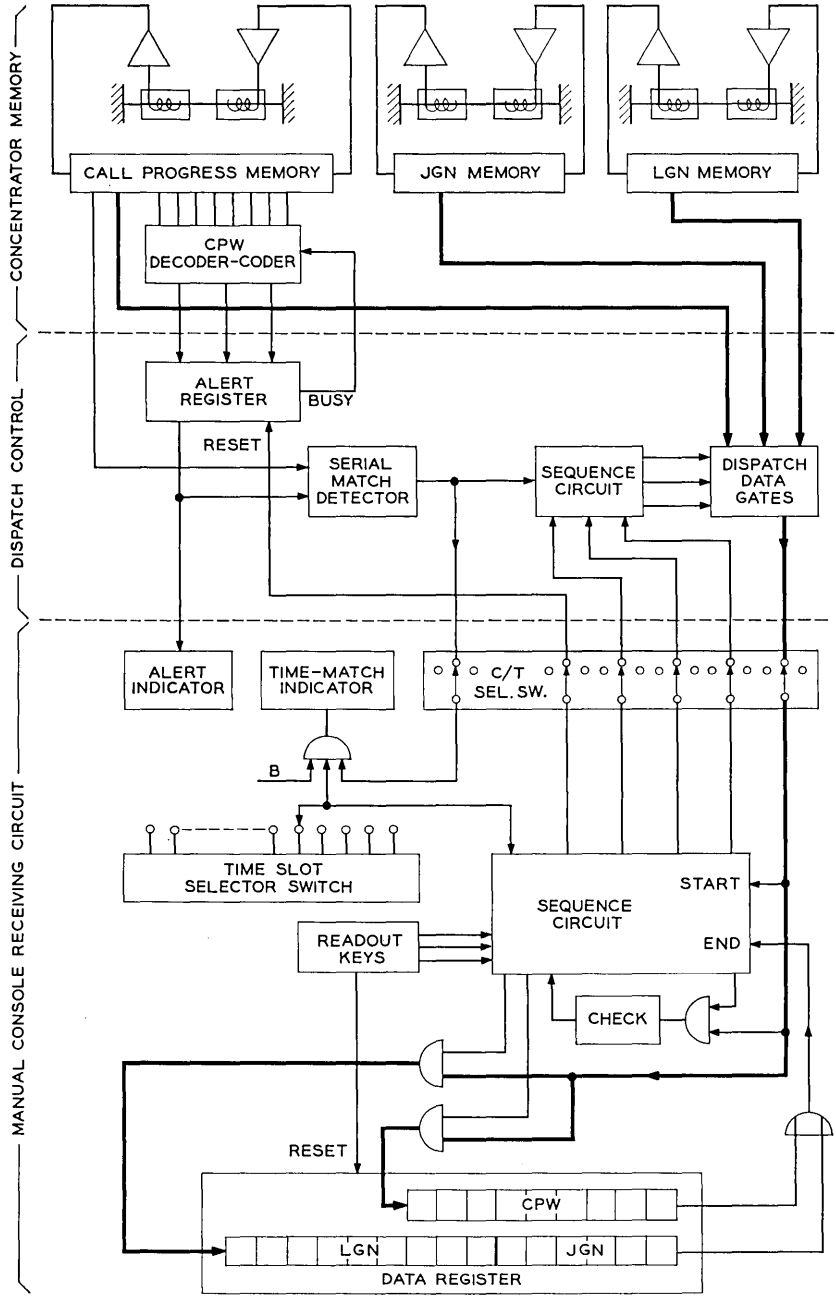

Fig. 7 — Diagram of dispatch control operations.

by a start bit. This causes the receiving sequence circuit to provide advance pulses to the receiving shift registers and enablement to an associated data-check circuit. The emergence of the start bit from the far end of a receiving shift register informs the sequence circuit that the block of data has been completely received. Advance pulses to the shift register are discontinued, the state of the check circuit is examined and the sequence circuit is primed for a second trial or its next function. The data registers are equipped with indicators for visual readout of the data to the operator, who resets them manually when the data are no longer needed.

IX. INSERT CONTROL AND MANUAL CONSOLE

Each concentrator controller contains an insert control circuit through which information from common control may be passed for storage in the controller memory. This communication is necessary at the various stages in the progress of a call in which common control actions for path selection, path tracing, billing, etc., are performed. In the experimental model, the logic and large-scale memory functions of common control were not implemented. Instead, a manual control console with visual display of current common control data was provided. An operator can select the concentrator or trunkor and the time slot into which information is to be inserted, preset the insert order and data items, and initiate the insert actions.

The relationship between the concentrator controller, its insert control circuit and the sending circuit of the console is outlined in Fig. 8. The operator selects the particular concentrator or trunkor memory to be altered by setting the C/T selector switch accordingly. The time slot selector switch also is set to the proper position, and the operator manually selects an insert order word, call progress word and, if necessary, line gate number and junctor gate number. Every insert operation includes at least an order word and a call progress word. The selected order and data words are registered in parallel in the data shift register stages by operation of a manual set key, which advances the sequence control circuit. The register setting is displayed to the operator for verification. Operation of a spill key enables the sequence circuit, so that advance pulses are applied to the data shift registers, and the send gates are operated in proper anticipation of the occurrence of the selected time slot at the concentrator memory input positions. A start bit in the order word, which precedes all other transmission, is detected by the sequence circuit of the insert control. The order word itself is then stored in an order word shift register under control of advance

pulses from the sequence circuit. Arrival of the start bit at the far end of the order word register advances the sequence circuit, so that the data-in gates to the concentrator memory are opened as required and in the proper sequence. Data flow directly into the shift registers of the memory loop serially in real time, so that no buffer storage is re-

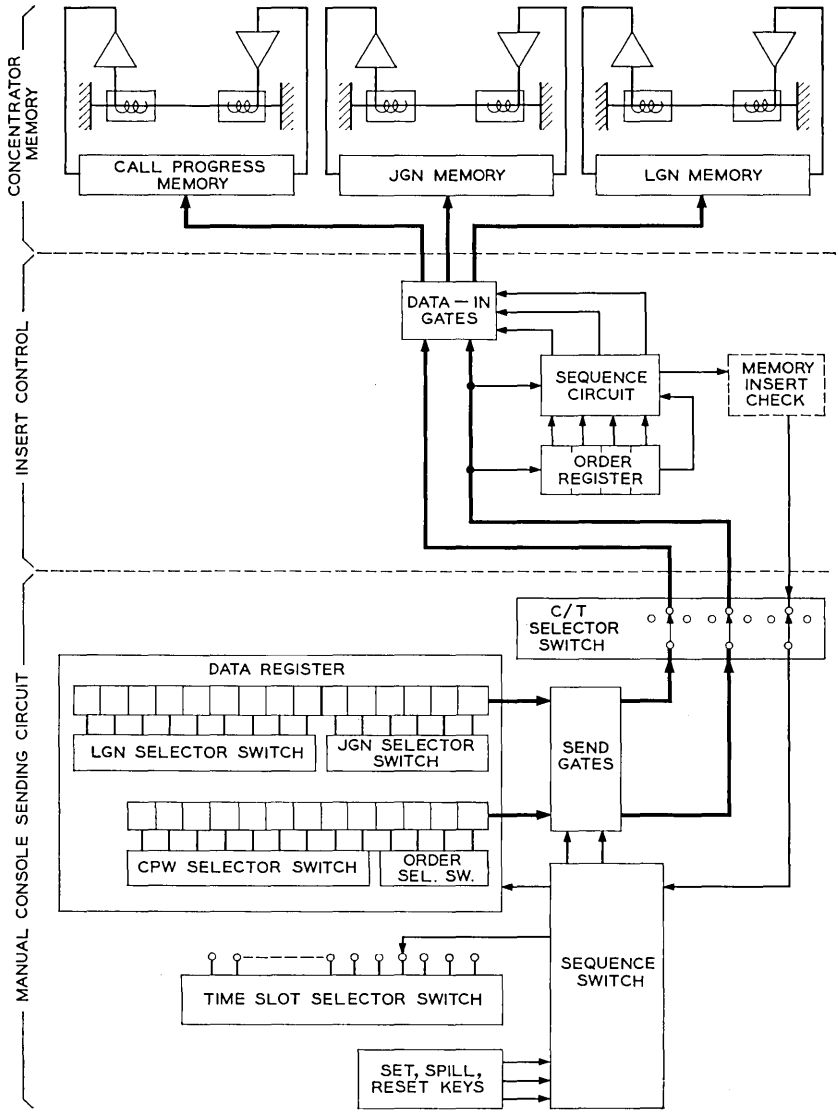

Fig. 8 — Diagram of insert control operations.

quired in the insert control. The sequence circuit in insert control also enables, in the proper time sequence, a memory-insert check circuit, which monitors the new data immediately after its insertion into the memory loop registers. Second trial and release signals are sent from this check circuit to the sending circuit of the console.

X. DELAY-STABILIZING SERVO

A delay pad is provided in the s lead, so that the round trip delay, from the switching center out to the remote concentrator and back to the switching center, is exactly equal to an integral multiple of 125 microseconds. This delay is required so that, for each conversation, the pulses representing speech in one direction may pass through the central stage switch at the same time as the pulses representing the other direction of the same conversation. The delay is provided by a magnetostrictive wire delay line of the same construction as the lines for the memory loops (Section V). The binary-coded electrical impulses of the s lead are impressed magnetostrictively on the line at the input transducer, and they travel along the line in acoustic form. The magnetostrictive effect is used at the output solenoid to convert back to electrical impulses; thus, an electrical delay is produced that is dependent upon the distance between the delay line transducers.

The delay can be set initially to the required value by manual positioning of the transducers. However, because the delay in transmission through cables varies with temperature, it is necessary to vary the delay of the pad in a compensating manner. This is done by a servo unit, shown in Fig. 9, that changes the physical position of the input solenoid.

Each information pulse from the delay line is transmitted to an integrating network through a "phase slicer" controlled by phase-timing pulses from the master clock. The portion of the information pulse that precedes the leading edge of the timing pulse drives the integrating network with one voltage polarity, and the portion of the information pulse that occurs after the leading edge of the timing pulse drives the integrating network with the opposite voltage polarity. An error voltage is developed unless the delay is such that the information pulses are centered about the leading edge of the timing pulses. If the error voltage exceeds a threshold value, a servo motor moves the input transducer along the delay line in the correcting direction. The reference phase-timing pulses are also used to regenerate the delay line output, so that standard-timed information is delivered to the central stage switch despite small variations in the delay line setting.

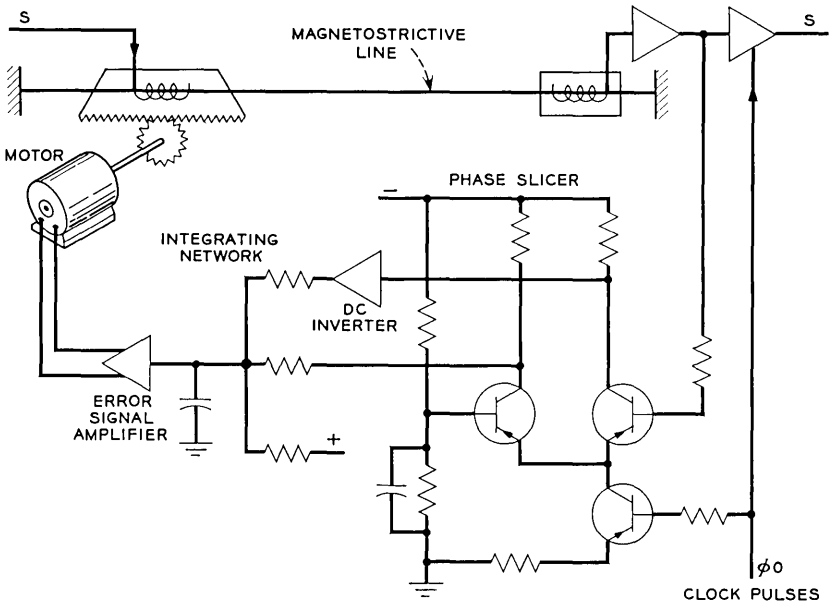

Fig. 9 — Delay pad and servo circuit.

XI. SPLITTING AND TONE GATES

The splitting and tone gate circuit, shown in Fig. 10, is inserted in the path between a remote concentrator and its appearance on the central stage switch connections. At this point, the normal send and receive paths may be opened and special coded signals may be connected to either lead under control of call progress words. Since the S and R lead signals are in pulse-code-modulated form, the signals inserted at these splitting gates must be in the same form. Time-varying signals for ringing, ringing tone and busy tone are generated in a circuit common to the office, as described in Section XII. Other special signals (see Table I), such as the framing command, the scan command and the zero-level speech code, may be obtained directly from the master clock. Simple diode and transistor gates, similar to those that make up the crosspoints of the central stage switch, are controlled from the call progress decoder-coder to connect these signals to the S and R leads as required. A buffer flip-flop register extends the call progress decoder order for the duration of each time slot.

This circuit provides full access to the concentrator speech paths, so that no blocking can occur in connecting these special signals, and reduced traffic load is placed on the junctor paths of the central state switch. In addition, some operations in the common control are avoided,

since transitions in the concentrator controller, such as from the ringing to the talking state upon the answer of the called customer or from the busy tone to the idle state upon hang-up of the customer hearing the tone, are now made locally within the controller.

XII. DIGITAL TONES

In order to pass through the speech transmission and gate circuits of an electronic system, ringing is accomplished in the voice frequency band and at the power levels satisfactory for speech. In addition, since the voice-frequency signals in the central office are in pulse-code-modulated form, the special tones must also be in this digital form.

The PCM codes to be generated include the maximum amplitude positive speech-sample code and the maximum amplitude negative

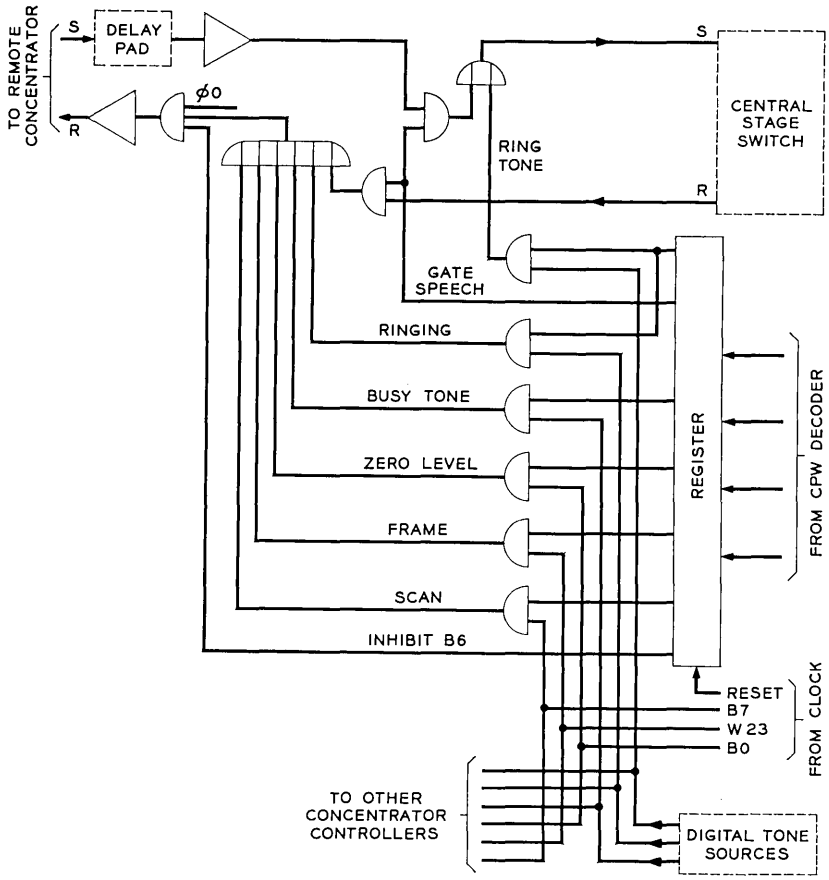

Fig. 10 — Splitting and tone gate circuit.

speech-sample code for ringing, the half-maximum amplitude codes for ring tone and busy tone, and the "zero level" speech-sample code for the silent intervals of the interruptions. These must be logically combined at the proper frequency and interruption rates.

The PCM tone signals generated are:

i. *ring*, which is a 1000 cps signal of maximum amplitude interrupted at both a 7.8 and 0.245 cps rate;

ii. *ring tone*, which is similar to ring but of lower amplitude;

iii. *busy*, which is a 250 cps signal of the same amplitude as the ring tone interrupted at both a 7.8 and 0.98 cps rate.

XIII. CLOCK

Timing of operations in the model is under the control of a central source of pulses, Fig. 11. A crystal oscillator operating at the basic bit rate of 1.536 mc acts as the master clock. Blocking oscillators driven through buffer and phase-shifting stages provide negative-going power pulses of 0.1-microsecond duration at the bit rate. These pulses are provided in two-phases, with the pulses of the second phase, ϕ_2, occurring midway in the interval between pulses of the first phase, ϕ_0. All shift registers in the system are advanced with ϕ_2 and may have their contents modified at ϕ_0 time. The phase pulses drive an eight position commutating circuit to provide eight-bit pulses, each of 0.65-microsecond duration and each spaced by 5.2 microseconds, the width of a time slot. In a similar manner, one of the bit pulses drives a commutator to form 24 "word" or time slot pulses, each 5.2 microseconds wide and repeated every 125 microseconds, which is the basic "frame" time of the system. In turn, one of the word pulses drives a commutator to furnish four frame pulses, each of which is 125 microseconds wide and is repeated every 500 microseconds.

Each commutating circuit is a shift-register, 8, 24 or 4 stages long, operated as a re-entrant ring with a single active stage. Each of the various time pulses occurs for the time interval in which the corresponding stage in its ring is active. These basic signals and logical combinations of them are distributed by power amplifiers to all units in the office as required. Similar pulses at the basic bit and time slot rate are generated at the remote units by timing recovery circuits dependent upon a received train of data bits.

XIV. FRAMING

Although the remote concentrator can recover the basic system bit rate from the PCM signals transmitted to it and can reconstruct pulses of time slot duration, it must be provided with a phase reference signal in order to associate a time slot pulse with the proper group of bit pulses.

This is the function of a framing command signal transmitted on the R lead. In time slots 0 to 22, this lead carries a seven-bit PCM speech code and an eighth bit that is normally a "zero". In time slot 23, eight "one" bits are transmitted, forming the unique framing signal. This signal originates in the splitting and tone gate circuit, Fig. 10. Occasionally, as described in Section IV on supervision, an eighth-bit "one" is inserted in one of the time slots 0 to 22 as a scan command, and speech code bits are suppressed to maintain the unique status of the framing code.

XV. SAFEGUARDS AND CHECKS

In an experimental model such as this, all of the checks, redundancy and automatic standby circuits necessary to guarantee continuity of service are seldom provided. However, some provision for their eventual incorporation must be contemplated. In this model, safeguards were included principally through the use of odd-parity coding for the line

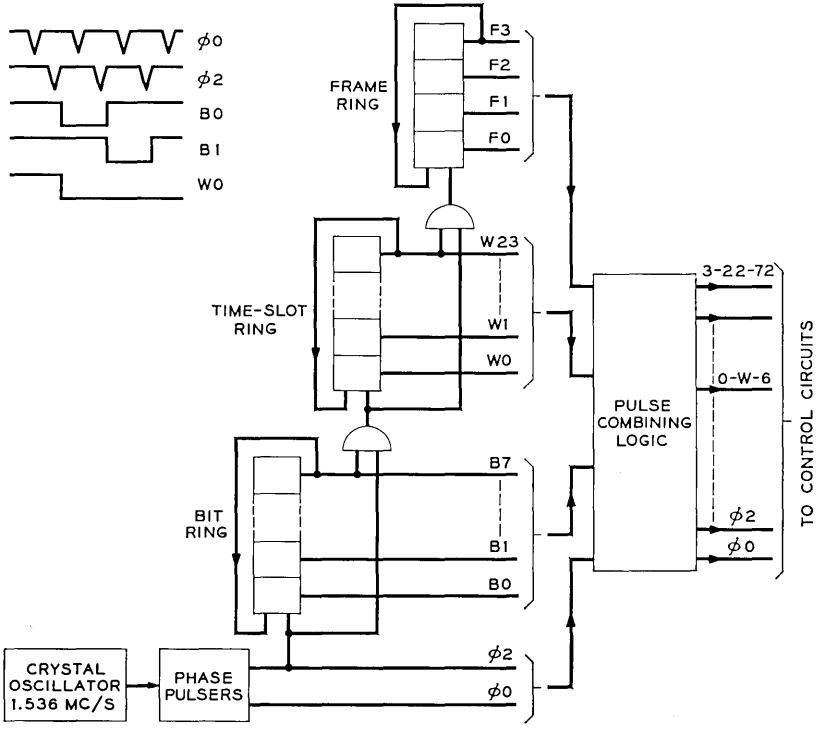

Fig. 11 — Clock structure.

and junctor numbers and the call progress words. The call progress words are chosen from the three-out-of-eight code group. A sixth bit is added to the five-bit junctor number and a ninth bit to the eight-bit line number, so that these complete binary number groups are converted to odd-parity groups. In the memory of the concentrator controller, the parity bit for a line number is stored in one of the spare bit positions of the junctor number memory. Parity check circuits associated with the controller memory continuously monitor the stored information as it circulates. A detected failure disables the actions of the call progress decoder-coder in the associated time slot to prevent compounding of the error, and an alarm indication is produced. The parity of information received by the insert circuit and transmitted to the console receive circuit is also checked, and automatic second trial is provided on the insert and dispatch operations. Failure on second trial results in an alarm indication.

XVI. CONCLUSIONS

Although this equipment is experimental, we have tried to assume realistic constraints as the basis for design. Thus, signaling between controller and common control has been kept to a reasonable minimum. Flexibility has been obtained by the use of call progress words, so that the particular sequences of control operations carried out within the controller can be altered, to accommodate changes in system requirements, by replacing a plug-in diode matrix in the call progress word decoder-coder. The use of call progress words to control the application and removal of tone signals within the controller reduces the usage of the common control and guarantees that tone connections can always be made when needed. The use of such tone connections also reduces traffic through the switching network and removes the necessity for tone trunks. Advantage has been taken, as in the control of line scanning, of the presence within the controller memory of stored information which specifies the state of the switching system.

We would like to acknowledge the support of our many colleagues in the Systems Research Department of Bell Telephone Laboratories, who contributed much to the work here described.

REFERENCES

1. Vaughan, H. E., Research Model for Time-Separation Integrated Communication, B. S.T.J, **38**, July 1959, p. 909.
2. James, D. B. and Johannesen, J. D., this issue, p. 31.
3. Malthaner, W. A. and Vaughan, H. E., An Automatic Telephone System Employing Magnetic Drum Memory, Proc. I.R.E., **41**, October 1953, p. 1341.

Electrical Properties of Gold-Doped Diffused Silicon Computer Diodes

By A. E. BAKANOWSKI and J. H. FORSTER

(Manuscript received August 17, 1959)

Planar diffused silicon junctions with storage times of one millimicrosecond or less are readily obtained by gold doping. The introduction of uniform gold concentrations (in the range from 1.2 $\times$ 10^{15} cm^{-3} to 8 $\times$ 10^{16} cm^{-3}) is conveniently done using solid state diffusion techniques. The gold diffusion technique allows relatively precise control of recombination center density, and, although applicable to almost any diffused silicon device, is particularly useful in control of storage time in small-area diffused silicon computer diodes. In this application, reverse recovery time of about one millimicrosecond may be obtained without substantial degradation of other electrical parameters. The process of gold doping by diffusion and its effect on electrical characteristics of diffused silicon computer diodes are discussed. Included are comparisons of first-order calculations and experimental results for variations of reverse recovery time, reverse current and forward current with gold atom density.

I. INTRODUCTION

In transistor circuits designed for high-speed switching or computing operations, computer diodes may perform useful functions. Often the use of computer diodes for such functions as pulse gating or shaping can lead to simplification of logic design, reduction of the number of transistors required and relaxation of circuit tolerances. With high-speed transistor circuitry, these benefits are most readily obtained when the switching times for the computer diodes are comparable to or less than those of the associated transistors. Diffused silicon transistors with good switching characteristics and switching times on the order of 20 millimicroseconds are available.[1] Accordingly, in logic or switching circuits using fast silicon transistors, computer diodes can be used most advantageously if their switching times can be reduced to the low millimicroseconds.

Where the use of silicon transistors is contemplated, comparable leak-

age currents, temperature characteristics and dc voltage drops strongly suggest the use of silicon junction diodes in associated circuitry. The major problem in design of a suitable silicon junction computer diode is meeting the requirement on switching time mentioned above. If the capacity of the diode is made sufficiently small by the use of a lightly graded junction and a small junction area,[2] the lower limit on the transient response is set by the storage time, or the time required to sweep out excess minority carriers from the region near the junction.

Planar diffused silicon junctions with storage times of one millimicrosecond or less can be readily obtained by doping the silicon with a sufficiently large concentration of recombination centers. This can be done most conveniently after the fabrication of a diffused silicon junction with otherwise suitable electrical characteristics. The solid-state diffusion of gold atoms into the silicon lattice provides a relatively simple and precise means of introducing recombination centers in densities as large as 8×10^{16} cm^{-3} in a prediffused silicon slice. Mesa diodes fabricated with this process (a convenient one for high-level production) can have storage times lower than one millimicrosecond.

In this paper, the process of gold doping by diffusion and its effect on the electrical characteristics of diffused silicon diodes are discussed. Emphasis is on those characteristics pertinent to computer diode applications. Relations between gold atom concentration and reverse recovery time, reverse current and forward current have been obtained experimentally. The reverse current and the forward current at small bias voltages increase directly with the gold atom density. At larger bias voltages, the forward current increases with the square root of the gold atom density. These relations and the approximate magnitudes of the currents can be obtained from first-order calculations that take into account diffusion currents[3] and the recombination and generation of carriers in the p-n transition region.[4] Theory and experiment indicate that the performance of a small-area diffused computer diode is not substantially degraded by gold doping to a level sufficient to decrease the recovery time to one millimicrosecond.

A brief discussion of the desirable electrical characteristics for computer diodes is given in the following section. In Section III the gold diffusion process and the diode structure used in the experiments are described. In Section IV the experimental results are quoted and compared with calculations.

II. ELECTRICAL CHARACTERISTICS OF COMPUTER DIODES

Desirable electrical properties have been briefly mentioned in the introduction. The relative importance of these can be best evaluated by

considering the equivalent circuit of Fig. 1, which is a reasonably good representation of a computer diode.[5]

The capacitance C_T is the depletion layer capacitance associated with the space-charge region of the junction. The resistance R_S is an ohmic series resistance dependent on the body resistivities and the diode geometry, including contacts. The remaining p-n junction symbol represents the action of an idealized p-n junction structure, in which diffusive, drift and generative current components may be important. Such a breakdown of the junction diode essentially follows the classical p-n junction treatment of Shockley.[3]

Since these elements, in effect, are all determined by the physical parameters of the diode, they are interdependent to some extent. It is therefore necessary to consider a physical configuration that produces an electrically favorable combination of the above elements. We will not attempt this optimization, but instead consider only the case of a planar diffused junction.

The value of R_S is influenced by body resistivity, the impurity gradient attained by diffusion and the diode geometry. Acceptable values of R_S can be achieved by suitable choice of body resistivity and geometry without introducing other undesirable compromises. The value of C_T is proportional to the junction area and, to a first approximation, varies directly with the cube root of the impurity gradient near the junction. In general, low values of C_T can be achieved by restriction of the junction area and use of a suitably small impurity gradient.[2] The lower limit on the area variation is largely determined by the maximum value permitted for R_S. Further restriction on the impurity gradient may be imposed by specific breakdown voltage requirements.

Assuming that the values of R_S and C_T have been reduced sufficiently, the circuit operation of the device is primarily dictated by the properties of the idealized junction diode element. The important properties of this element are its current–voltage characteristic and its transient response.

We will not consider here the problem of designing a diode with an optimum combination of R_S, C_T and junction element characteristics; instead, optimization of the junction element itself will be discussed. Such individual attention is possible because large variations in forward

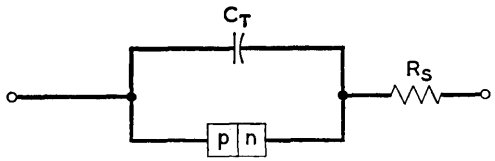

Fig. 1 — Equivalent circuit for a junction diode.

and reverse characteristics and, in transient response, can be obtained almost independently of R_S and C_T by variations in body lifetime.

We will discuss (in Section IV) the transient response, reverse current and forward current as functions of body lifetime, where the lifetime is varied by the addition of known densities of recombinations centers.

III. EXPERIMENTAL DIFFUSED JUNCTION DIODE

3.1 *Junction Diode Structure*

The same basic diode structure (illustrated in Fig. 2) was used in all experiments. The p-n junction is obtained by diffusing boron from high surface concentration into n-type silicon (0.12 to 0.15 ohm-cm). The impurity gradient at the junction is about 10^{21} cm^{-4}. Ohmic contact can

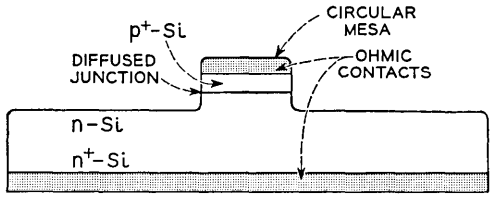

Fig. 2 — Test diode structure.

be made to the highly doped p-type (p^+) silicon near the surface with a suitable gold plating technique. An n^+ layer provided by a phosphorous diffusion allows similar ohmic contact to the n-type silicon.

To permit measurement of body lifetimes in the millimicrosecond range by the reverse recovery measurement technique,[6] a small transition region capacity is required. The junction area is the cross-sectional area of the circular mesa illustrated in the figure. For a mesa of 0.005 inch, values of $C_T < 3\mu\mu f$ are obtained. The value of R_S for this geometry is about 12 ohms.

3.2 *Control of Recombination Center Density with Gold Diffusion*

The transient response of a planar diffused diode (to be discussed in more detail in the following section) is determined largely by the carrier lifetimes at the edges of the transition region and in the regions near it. The carrier lifetimes are, in turn, related to the density and nature of the recombination centers in these regions. In our experiments, the recombination centers of interest are provided by gold atoms in the silicon

lattice.[7,8] Struthers[9] has measured the diffusion constant and the solid solubility of gold in silicon for the temperature range from 800° to 1250°C. The solubility of gold in silicon has been measured by Collins et al.[7] for the temperature range from 1000°C to 1380°C. Their values do not agree with those reported by Struthers. Later measurements[10] indicate a slightly steeper variation of the solid solubility with temperature than that given in Ref. 9, and values at higher temperatures closer to those reported in Ref. 7. The values of the diffusion constant given in Ref. 9 and the available solubility data indicate that solid state diffusion of gold in silicon should be a reproducible way of introducing gold atoms in silicon to density values at least as large as 8×10^{16} cm^{-3}.

The diffusion constant of gold is substantially larger than that of boron or phosphorus at temperatures up to 1300°C. Thus, it is possible to diffuse boron and phosphorus into a silicon slice, then plate the slice with a thin gold layer and diffuse in the gold without producing substantial changes in the boron and phosphorus distributions. If sufficient time is allowed for the gold diffusion, the concentration of gold in the slice can approach the solid solubility limit corresponding to the diffusion temperature, and the density of recombination centers is determined by the temperature chosen for the gold diffusion. Test diodes similar to the one illustrated in Fig. 2 can then be fabricated from the diffused silicon slice.

IV. EXPERIMENTAL RESULTS

4.1 *Transient Response*

4.1.1 *Reverse Recovery Time Measurement*

For switching applications, the transient response of a diode may be specified in terms of a forward and a reverse recovery time. The forward recovery time is usually substantially shorter than the reverse, and will therefore not be considered. The reverse recovery time is primarily dependent on the body lifetime, and can therefore vary with the density of recombination centers and their recombination cross section.

To clarify subsequent discussion of these variations, the definition and measurement of reverse recovery time will be briefly considered. We have measured reverse recovery time, t_r, under the following conditions: the diode is initially biased to a forward current I_f. A reverse pulse is applied, and the circuit is such that the diode initially conducts a current I_{r0} (equal in magnitude to I_f) in the reverse direction (see Fig. 3). The time dependence of the reverse current is then observed on a traveling-

wave oscilloscope. This method of test is essentially that discussed by Kingston.[6]

As shown in Fig. 3, the reverse diode current remains constant for a time t_I, while the rate of removal of carriers is determined primarily by the external circuit. When the number of carriers near the junction begins to decrease substantially, the current begins to decrease, approaching the steady-state reverse current for the diode. The time t_{II} is the additional time required for the current to reach a value arbitrarily specified as $0.1I_{r0}$. We will define t_r, the reverse recovery time, to be the sum $t_I + t_{II}$. According to Kingston's analysis (for a simple planar

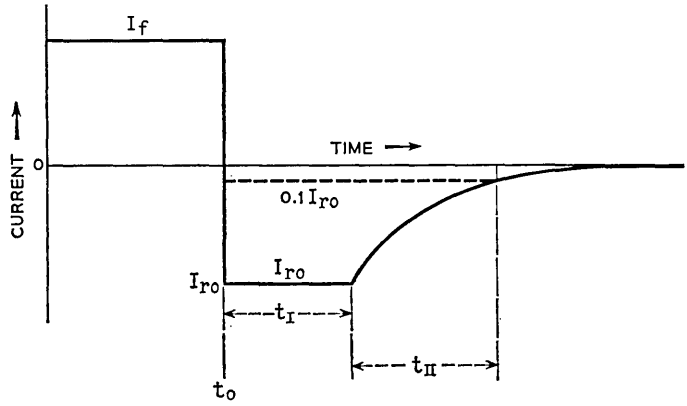

Fig. 3 — Diode reverse recovery pulse.

step-junction) the sum $t_r = t_I + t_{II} \cong 0.9\tau$, where τ is the body lifetime.

Our measurements have been made on diffused junctions. In this case, taking into account the retarding field resulting from the impurity gradient, a similar analysis indicates that t_I is about 0.4τ and t_{II} is about 0.1τ. Thus, we have the sum $t_r \cong 0.5\tau$.

4.1.2 *Dependence of Reverse Recovery Time on Gold Concentration*

The dependence of body lifetime on recombination center density has been considered by Hall,[11] and in detail by Shockley and Read.[12] Only the steady-state value of carrier lifetime will be used in interpretation of

recovery time measurements. For a single kind of recombination center, present in density N, with capture cross sections σ_{p0} and σ_{n0} for holes and electrons, at low level of minority carrier injection, the lifetime[12] is given by

$$\tau = \frac{n_0 + n_1}{n_0 + p_0}\tau_{p0} + \frac{p_0 + p_1}{n_0 + p_0}\tau_{n0}, \tag{1}$$

where n_0 and p_0 are equilibrium electron and hole densities,

$$n_1 = n_i e^{(E_t - E_i)/kT},$$

$$p_1 = n_i e^{(E_i - E_t)/kT},$$

($E_t - E_i$ = the difference between the energy level of the trap and the intrinsic Fermi level position) and

$$\tau_{p0} = \frac{1}{\sigma_{p0} v_p N}, \tag{2}$$

$$\tau_{n0} = \frac{1}{\sigma_{n0} v_n N}, \tag{3}$$

where v_n and v_p are thermal velocities for electrons and holes.

We will be concerned with the recombination centers that result from the introduction of gold atoms into the silicon lattice. According to Bemski,[8] recombination in gold-doped silicon is facilitated by two trapping levels associated with each gold atom. These are presumably the acceptor level, E_{t1}, located about 0.54 ev below the conduction band, and the donor level, E_{t2}, about 0.35 ev above the valence band (see Fig. 4)

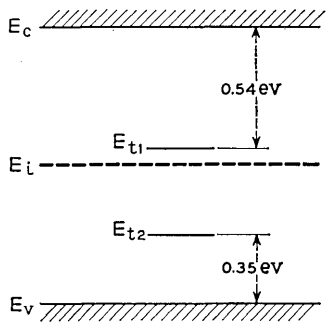

Fig. 4 — Energy levels for gold in silicon.

reported by Collins et al.[7] In n-type silicon, only one of these (E_{t1}) is effective, and in p-type silicon only the other (E_{t2}) is effective[8] (provided the silicon is sufficiently extrinsic). Thus the form of (1) is preserved, whether n- or p-type silicon is considered. However, in n-type silicon, n_1, p_1, τ_{n0} and τ_{p0} refer to the level E_{t1} and, in p-type silicon, these quantities are associated with the level E_{t2}.

In sufficiently extrinsic n-type silicon, the lifetime τ_1 given by (1) reduces to

$$\tau_1 \cong \tau_{p01} = \frac{1}{\sigma_{p01} v_p N}, \tag{4}$$

where τ_{p01} and σ_{p01} are associated with the level E_{t1}. In sufficiently extrinsic p-type silicon, the lifetime τ_2 is given by

$$\tau_2 \cong \tau_{n02} = \frac{1}{\sigma_{n02} v_n N}, \tag{5}$$

where τ_{n02} and σ_{n02} are associated with the level E_{t2}.

It would now appear possible to relate t_r (as measured on a p^+-n diode) and the trap density N, provided σ_{p01} is known, since, as mentioned in Section 4.1.1, $t_r \cong 0.9\tau_1$. In a similar way, with an n^+-p junction $t_r \cong 0.9\tau_2$, t_r can be related to N provided σ_{n02} is known. However, the diffused junctions described in Section 3.1 are more like linearly graded junctions in which minority carrier injection on both sides of the junction must be taken into account. While it is possible to carry out a solution for such a case, it seems intuitively evident that t_r must lie between $\tau_1/2$ and $\tau_2/2$. In this instance, $\sigma_{p01} v_p = 1.38 \times 10^{-8}$ cm^3-sec^{-1} and $\sigma_{n02} v_n = 3.5 \times 10^{-8}$, as given by Bemski.[8] Therefore, to a fair approximation,

$$t_r \cong \frac{2.53 \times 10^7}{N} \tag{6}$$

since τ_1 and τ_2 are almost equal.

Several groups of experimental diodes were fabricated as described in Section III. The density of recombination centers was varied from group to group by varying the gold diffusion temperature in the range from 800° to 1200°C. The recovery time was then measured as a function of the gold diffusion temperature. The results of the experiment are illustrated in Fig. 5. Each circle represents the average t_r for a group of diodes gold diffused at a temperature T_D represented by the corresponding abscissa. The solid bars indicate the range from maximum to minimum t_r as measured for each group, and the dotted bars indicate the

estimated precision of the measurement. The number of diodes in each group varies, the largest group (200 diodes) being gold diffused at 1100°C, and the other groups including from five to ten diodes. The dotted line represents the calculated value of t_r using (6) and the values of N taken from available solubility data.[7,10]

Accuracy in measuring the lowest values of t_r is poor. However, it

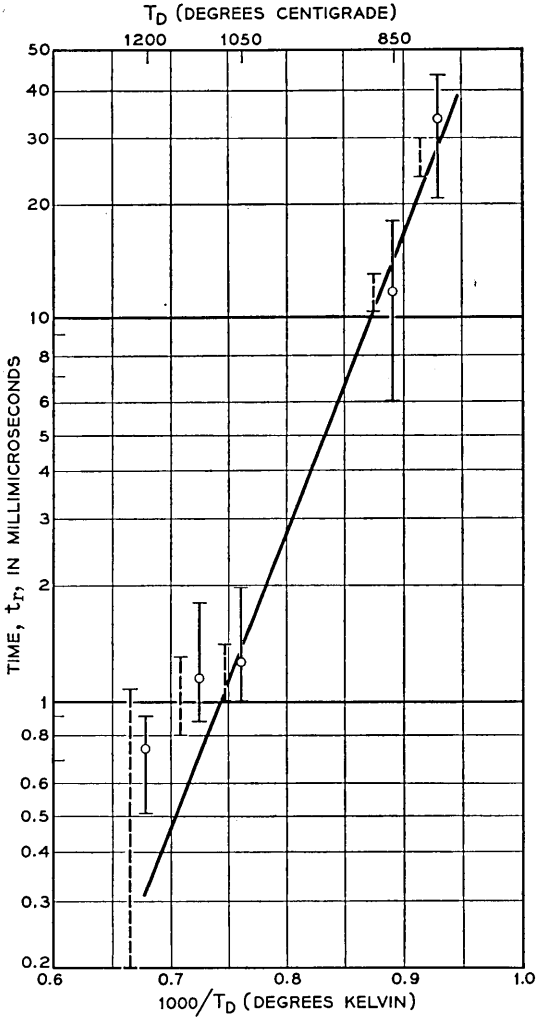

Fig. 5 — Reverse recovery time as a function of inverse gold diffusion temperature.

appears that t_r is inversely proportional to the recombination center density. The better-than-order-of-magnitude agreement between the experimental and calculated values of t_r indicates that the relation between the density of gold atoms and lifetime reported by Bemski[8] can probably be extended to gold concentrations as large as 8×10^{16} cm^{-3}.

4.2 *Reverse Current*

4.2.1 *Reverse Current Arising from Generation in the Space-Charge Region*

There is good experimental evidence that the reverse current in silicon p-n junctions largely arises from carrier generation in the space-charge region.[4,13] According to Shockley and Read,[12] the carrier generation rate, U, is given by

$$U = - \frac{np - n_i^2}{(n + n_1)\tau_{p0} + (p + p_1)\tau_{n0}}. \tag{7}$$

In the space-charge region, for the case of intermediately large reverse bias, it is assumed that $n \to 0$ and $p \to 0$, and therefore U_1 reduces to

$$U = \frac{n_i^2}{n_1\tau_{p0} + p_1\tau_{n0}} \tag{8}$$

and the magnitude of reverse current, I_r, is given by

$$I_r = qUWA, \tag{9}$$

where q is the electronic charge, W is the width of the space-charge region, and A is the junction area.

Equation (8) applies to the case of a single-energy level associated with the recombination centers in the swept-out region. For the case of gold-doped silicon, we must consider the recombination levels represented by E_{t1} and E_{t2}. The calculation of the generation rate in the space-charge region, taking into account the respective population of states associated with each gold atom, will not be attempted here. However, an upper limit for the current generated may be obtained. In Fig. 6, at the edge of the space-charge region on the p-type side, the number of effective recombination centers N_{t2} is close to N, and the energy associated with these centers is E_{t2} (substantially less than E_i). Similarly, at the edge of the space-charge region on the n-type side, the number of recombination centers N_{t1} is close to N, and the associated energy level is E_{t1} (close to E_i). Thus, the generation rates, U_1 and U_2, at these

edges of the space-charge region are given by

$$U_1 \cong \frac{n_i N}{\frac{1}{\sigma_{p01} v_p} + \frac{1}{\sigma_{n01} v_n}} \tag{10}$$

and

$$U_2 \cong \frac{n_i N \frac{n_i}{p_i}}{\frac{1}{\sigma_{n02} v_n} + \left(\frac{1}{\sigma_{p02} v_p}\right)\left(\frac{n_1}{p_1}\right)^2}. \tag{11}$$

Bemski[8] indicates that $\sigma_{p01} v_p$ differs from $\sigma_{n01} v_n$ only by a factor of two. Thus $\sigma_{p01} v_p$, $\sigma_{n01} v_n$ and $\sigma_{n02} v_n$ are of similar magnitudes. Further, $\sigma_{p02} v_p$ is not substantially less than $\sigma_{n02} v_n$. Then, since $p_1 \gg n_i$, inspection of (10) and (11) indicates that $U_1 \gg U_2$. We will therefore compute the maximum reverse current using the expression

$$I_r = qU_1 WA = q \frac{n_i NWA}{\frac{1}{\sigma_{p01} v_p} + \frac{1}{\sigma_{n01} v_n}} \tag{12}$$

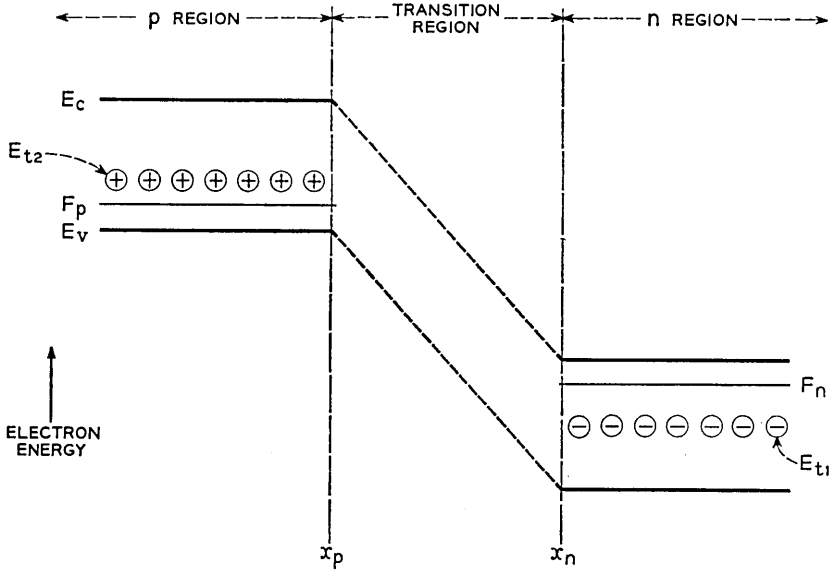

FIG. 6. — Effective recombination centers near the edges (x_n and x_p) of the transition region of a reverse-biased p-n junction; F_p and F_n are quasi-Fermi levels for holes and electrons.

for comparison with our experimental results. For this computation, W has a value of one-half the total space-charge width.

4.2.2. *Reverse Current as a Function of Gold Concentration*

Experimental data for a series of gold-doped diodes is indicated in Fig. 7. The bars represent maximum and minimum values of I_r (measured at a bias voltage of 10 volts) for groups of diodes fabricated from silicon with gold diffused at a temperature indicated on the abscissa. The value of I_r at 10 volts (well below body breakdown voltage) is chosen to minimize error from surface-dependent reverse current components, which often tend to increase with bias more rapidly than does body current. The dotted line is the reverse current calculated from (12), using the cross sections determined by Bemski[8] and the solubility data of Struthers.[10]

The magnitude of the observed reverse current is evidently proportional to N. The agreement between observed and computed values of current is good in view of the very approximate nature of (12).

It is significant from the viewpoint of device design that, for values of gold concentration approaching 8×10^{16} cm^{-3}, recovery time can be less than one millimicrosecond, although the reverse current is still substantially less than 5×10^{-8} ampere.

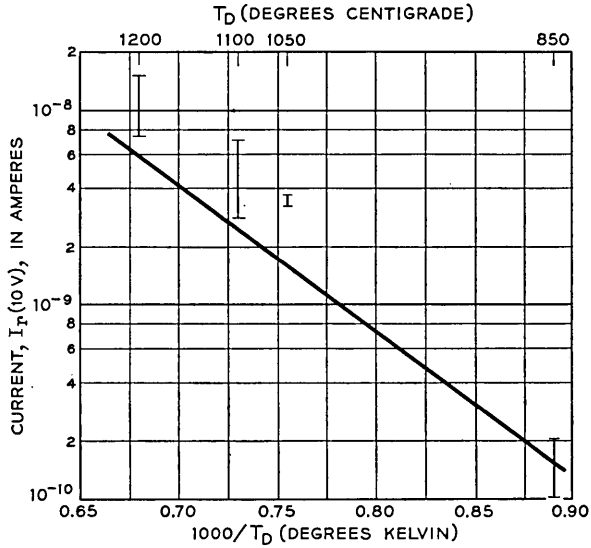

Fig. 7 — Reverse current as a function of inverse gold diffusion temperature.

According to (12), the magnitude of I_r should increase proportionally with W, and therefore for larger values of V vary as $V^{\frac{1}{2}}$, up to biases near breakdown. We have, in general, failed to find this voltage dependence of I_r at voltages greater than 15 volts, although body breakdown voltage for these diodes is between 50 and 60 volts. Both adverse surface conditions and flaws in the silicon body can introduce excess currents that are voltage-dependent, and this discrepancy at higher voltage may result in poor control of these variables. The reason for this discrepancy has not been definitely established; from the practical standpoint, however, the reverse currents are not strong functions of voltage, and are usually less than 0.1 microampere at two-thirds of breakdown voltage for a one-millimicrosecond diode.

4.3 *Forward Current*

4.3.1 *Nature of the Forward Characteristic*

Forward current flow in silicon p-n junctions can arise from diffusion of minority carriers in a region just outside the depletion layer and from carrier recombination in the depletion layer. It has been shown by Sah et al.[4] that the latter mechanism dominates at low current densities and the former dominates at higher current densities. The cross-over from one dominant mechanism to the other occurs in a region of current density that is determined by the nature and concentration of the recombination centers and by the conductivities in the vicinity of the junction. On the basis of analysis similar to that of Shockley[3] for a planar diffused junction, the diffusion current density, J_D, is given by

$$J_D = q\left[\sqrt{\frac{D_p}{\tau_p}}\, p_n(x_n)\cdot\theta_p + \sqrt{\frac{D_n}{\tau_n}}\, n_p(x_p)\cdot\theta_n\right](e^{\beta v} - 1), \qquad (12)$$

where D_p and D_n are diffusion constants for holes and electrons, τ_p and τ_n are hole and electron lifetimes in the n and p regions near the junction, p_n and n_p are equilibrium hole and electron concentrations at the depletion layer boundaries x_n and x_p, and V is the applied voltage.

In the linear grade approximation to the impurity distribution in a diffused junction, θ_p and θ_n are given by[14]

$$\theta_p = \frac{iH_0\left(i\dfrac{x_n}{\sqrt{D_p\tau_p}}\right)}{H_1\left(i\dfrac{x_n}{\sqrt{D_p\tau_p}}\right)} \qquad (13)$$

and

$$\theta_n = \frac{iH_0\left(i\frac{x_p}{\sqrt{D_n\tau_n}}\right)}{H_1\left(i\left(\frac{x_p}{\sqrt{D_n\tau_n}}\right)\right)}, \tag{14}$$

where $i^2 = -1$ and H_0 and H_1 are Hankel functions of the first kind.* For the case of gold-doped silicon, as indicated in Section 4.1.2 [(4) and (5)], $\tau_p \cong \tau_1$ varies with $1/N$, and $\tau_n \cong \tau_2$ varies with $1/N$. Since θ_p and θ_n are slowly varying functions of τ_p and τ_n,

$$J_D \propto \sqrt{N}. \tag{15}$$

The space-charge recombination current[4] is given by

$$J_R \cong \frac{qn_iW}{\sqrt{\tau_{p0}\tau_{n0}}}\frac{e^{\beta V/2}}{\beta(\psi_0 - V)}f(b), \tag{16}$$

where

$$b = e^{-\beta V/2}\cosh\left[\frac{E_t - E_i}{kT} + \tfrac{1}{2}\ln\left(\frac{\tau_{p0}}{\tau_{n0}}\right)\right] \tag{17}$$

and $f(b)$ is as given by Sah et al.[4] Thus:

$$\text{for } E_t \cong E_i \quad \text{and} \quad \frac{\tau_{p0}}{\tau_{n0}} \cong 1, \qquad f(b) \cong 1;$$

$$\text{for } \left|\frac{E_t - E_i}{kT}\right| \gg 1 \quad \text{and} \quad \frac{\tau_{p0}}{\tau_{n0}} \cong 1, \qquad f(b) \ll 1.$$

In the structure under consideration, recombination current arises from the presence of the two levels, E_{t1} and E_{t2}. Since exact analysis of this situation can be quite complicated, we will assume that the level E_{t1} dominates and calculate J_R from (16), ignoring the contribution from E_{t2}. To some extent this is justified, since $E_{t1} \cong E_i$, $\tau_{p01}/\tau_{n01} \cong \frac{1}{2}$, $(E_{t2} - E_i)/kT$ is about -9.6, and τ_{p02}/τ_{n02} is probably not substantially different from unity. We will therefore assume that J_R is approximately the current that would arise from recombination in that part of the space-charge region between $x = 0$ (at the junction), and x_n. Thus, we have assumed that the level E_{t1} is effective on the n side of the junction. In this case

* Strictly speaking, this analysis is inapplicable, since the assumption made that minority carrier injection does not appreciably alter the "built-in" field at the junction is not strictly valid for forward currents of interest. Consideration of higher injection level not only leads to modification of the field at the junction but also modifies the calculated value of β. However, in cases of interest here the functional dependence of current on doping level will be similar to that of (12).

$$J_R \cong \frac{qn_i x_n}{\sqrt{\tau_{p01}\tau_{n01}}} \frac{e^{\beta V/2}}{\beta(\psi_0 - V)}, \tag{18}$$

and therefore

$$J_R \propto \frac{1}{\sqrt{\tau_{p01}\tau_{n01}}} \propto N. \tag{19}$$

4.3.2 *Forward Current as a Function of Gold Concentration*

Shown in Fig. 8 is a plot of the forward current, I_f, as a function of bias voltage for a typical diode gold-diffused at 1100°C. At low voltages,

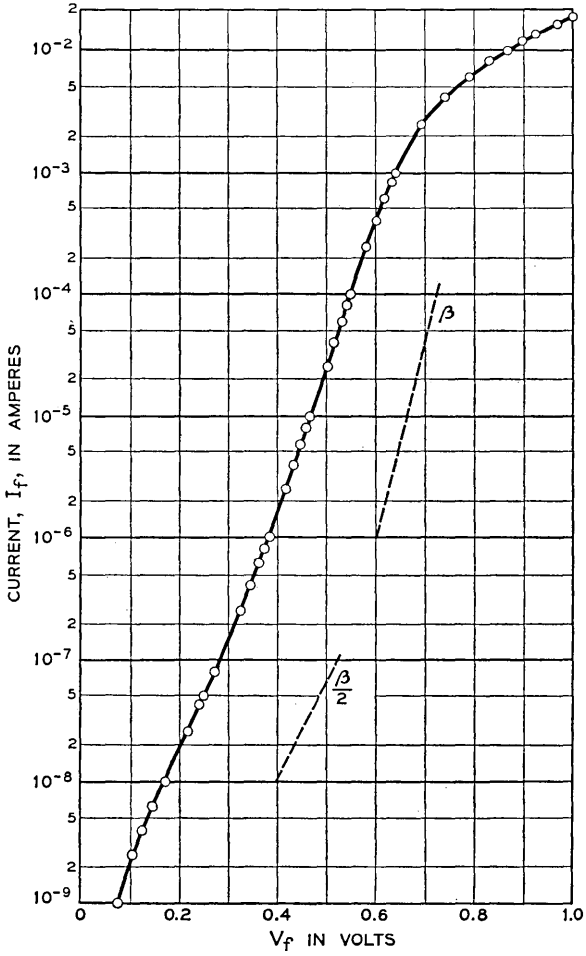

Fig. 8 — Forward current for a typical diode gold diffused at 1100°C.

the slope is close to $\beta/2$ [in agreement with (17)]; at higher voltages, the slope approaches β [in agreement with (12)] and, at still higher voltages, a leveling-off occurs, presumably resulting from series resistance effects. As the gold diffusion temperature is increased, the transition to higher values of β occurs at progressively higher values of voltage, and the magnitude of the current at a given voltage increases.

The effect of gold doping on the forward characteristic has been observed by studying the variation in foward current, $I_f(V)$, with gold diffusion temperature. The recombination current variation is studied by measuring $I_f(0.3)$, where it is expected that (19) should apply. The diffusion current variation is studied by measuring $I_f(0.7)$, where (15) should apply. To obtain the proper value of $I_f(0.7)$, a series resistance

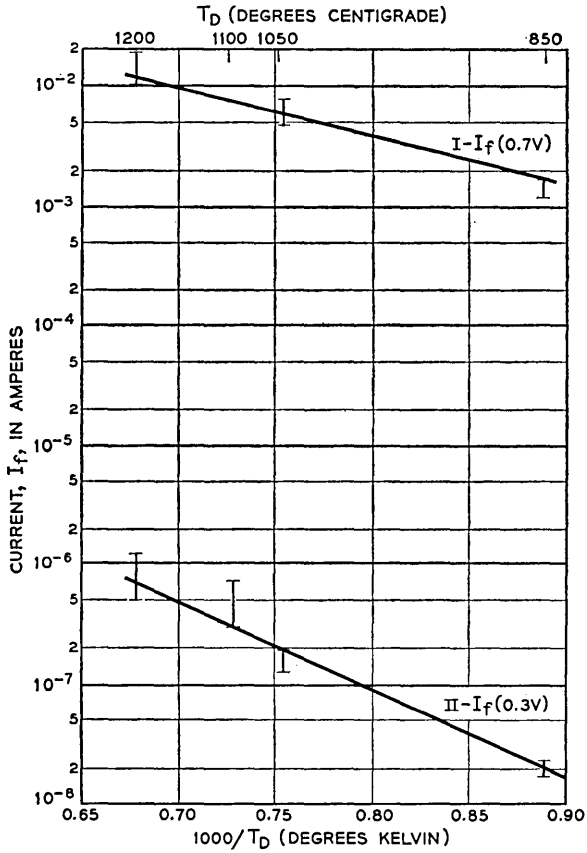

Fig. 9 — Forward current as a function of inverse gold diffusion temperature.

correction must be made. The results of such an experiment are indicated in Fig. 9.

In Fig. 9 (on the upper curve, I), the bars represent the maximum and minimum values of log $I_f(0.7)$ as measured for groups of diodes gold diffused at three different temperatures. These ranges are plotted against $1/T_D$, where T_D is the gold diffusion temperature. According to (15), log I_f is proportional to log $\sqrt{N}$. The solid line in the figure represents the dependence of log $\sqrt{N}$ on $1/T_D$, obtained from Struthers' data.[10] Reasonable agreement is indicated. The magnitudes of the currents calculated from (12) are reasonably close to experimental values, in view of the limited accuracy involved in the use of this equation.

On the lower curve, II, in Fig. 9, the bars represent the ranges of log $I_f(0.3)$, as measured for groups of diodes gold-diffused at various temperatures. According to (19), log $I_f(0.3)$ should vary linearly with $1/T_D$ in the same manner that log N varies with $1/T_D$. The solid line in the figure indicates the dependence of log N on $1/T_D$ from the solubility data. Again, the agreement is reasonably good. However, the magnitude of the currents calculated from (18) are, for this case, sizably less than the indicated experimental values, the discrepancy being larger than expected even when the inaccuracies of the calculation are considered.

V. CONCLUSIONS

The reverse recovery time of diffused silicon p-n junctions can be reduced to one millimicrosecond or less by gold doping. The decrease in recovery time results primarily from a decrease in minority carrier lifetimes in the p and n regions at the edges of the transition region.

It is found that, to a first approximation, the reverse recovery time is inversely proportional to the gold atom concentration, if the latter quantity is estimated from the measured solubility[7,10] of gold in silicon. Measured values of the recovery time are in agreement with calculated values using recombination cross sections measured by Bemski.[8]

Relatively simple diffusion techniques permit the introduction of desired gold concentrations within the range 1.2×10^{15} to 8×10^{16} cm^{-3}. If initial carrier lifetimes in the silicon diode are sufficiently high, then a desired recovery time value between 0.7 and 35 millimicroseconds may be reproduced to within ± 60 per cent or better. At higher levels of gold concentration, uncontrolled impurities introduced during diffusion and thermal cycling tend to be swamped out, and recovery time is probably more reproducible. However, accurate measurement of t_r in this range is difficult.

Reduction of recovery time to one millimicrosecond or less is accompanied by changes in the reverse and forward diode characteristics. The reverse current (at lower bias voltages), primarily generated in the space-charge region, increases in proportion to the gold atom density. The forward current in a gold-doped diode has space-charge recombination and diffusion components. The former component increases directly with the gold atom density, and the latter increases with the square root of the gold atom density.

These changes in characteristics are not sufficient to substantially degrade the performance of a diffused diode suitable for computer applications. Reverse current can be kept well below circuit requirements by minimizing the junction area. The relative increase in the space-charge recombination component of the forward current introduces only a small increase in dynamic forward resistance in a rather limited portion of the current voltage characteristic, and can serve to lower the dc drop at low forward bias. Thus, gold doping of a small-area diffused silicon diode can produce a relatively high-performance computer diode with transient response better than one millimicrosecond.

VI. ACKNOWLEDGMENT

The authors wish to acknowledge the help of R. L. Rulison and D. F. Ciccolella in the fabrication of gold-diffused silicon diodes. Acknowledgment is also due W. C. Meyer and Miss F. R. Lutchko for electrical measurements.

REFERENCES

1. Miller, L. E., I.R.E. Wescon Conv. Rec., Vol. 2, Part 3, 1958, p. 132.
2. Forster, J. H. and Zuk, P., I.R.E. Wescon Conv. Rec., Vol. 2, Part 3, 1958, p. 122.
3. Shockley, W., B.S.T.J., **28,** 1949, p. 435.
4. Sah, C. T., Noyce, R. N. and Shockley, W., Proc. I.R.E., **45,** 1957, p. 1228.
5. Uhlir, A., Jr., Proc. I.R.E., **44,** 1956, p. 1184.
6. Kingston, R. H., Proc. I.R.E., **42,** 1954, p. 829.
7. Collins, C. B., Carlson, R. D. and Gallagher, C. J., Phys. Rev., **105,** 1957, p. 1168.
8. Bemski, G., Phys. Rev., **111,** 1958, p. 1515.
9. Struthers, J. D., J. Appl. Phys., **27,** 1956, p. 1560.
10. Struthers, J. D., unpublished results.
11. Hall, R. N., Phys. Rev., **87,** 1952, p. 387.
12. Shockley, W. and Read, W. T., Jr., Phys. Rev., **87,** 1952, p. 835.
13. Veloric, H. S. and Prince, M. B., B.S.T.J., **36,** 1957, p. 975.
14. Bakanowski, A. E., unpublished work.

Analysis of Quality Factor of Annular Core Inductors

By V. E. LEGG

(Manuscript received August 25, 1959)

This papes summarizes and formally presents methods in use during the past 25 years at Bell Telephone Laboratories for the design of high-quality inductors. The quality factor, Q, of annular core inductors can be maximized by adapting core dimensions, winding details, etc., to the specific properties of the magnetic core. Commercial core materials include 2–81 Mo-permalloy powder and carbonyl iron powder. Analysis of the relationships of dimensions and magnetic properties is given first in terms of the bare fundamentals of dc copper resistance and magnetic core losses. The paper then derives expressions for, and considers the effects of, additional losses due to the windings, such as eddy currents in the copper, dielectric losses and the contribution to effective resistance due to parallel distributed capacitance. Finally, a graph is supplied showing the maximum Q and optimum frequency for many commercial sizes of annular cores.

I. INTRODUCTION

The analysis of annular magnetic core inductors in terms of the desired characteristics and the properties of their components entails adjustments of core and winding dimensions, and the selection of best arrangements for each set of requirements. The present paper summarizes the methods in use during the past 25 years at Bell Telephone Laboratories. Such analysis involves consideration of numerous properties of core and winding, some of which may occasionally be negligible. It will be illuminating first to make an analysis based upon the fewest basic properties. Subsequently the contributions and complications of secondary properties will be treated, as their need becomes apparent.

II. SIMPLE ANALYSIS

2.1 *Magnetic Core*

The *inductance* of a winding of N turns on an annular core is

$$L = \frac{4N^2\mu wh \times 10^{-9}}{D} \text{ henry,} \qquad (1)$$

where μ is the permeability of the ring core material, which is of mean diameter D, radial thickness w, and axial height h, all in centimeters.[1]

For simplicity, the "air inductance", due to the windings on the space outside of the core, will be neglected. It represents a very small contribution to the total inductance of coils with high permeability cores.

Energy losses in the core material show themselves as an *effective resistance* increase in the winding:

$$R_m = (aB_m + c + ef)\mu Lf \quad \text{ohm}, \tag{2}$$

where a is the hysteresis loop area constant, B_m is the peak induction in the magnetic material, c is the "residual" loss constant, e is the eddy current coefficient and f is the ac frequency at which the measurement is made.[1] If the hysteresis loss is not negligible, the induction B_m must be calculated in terms of the rms current I(in amperes) in the winding

$$B_m = 0.4\sqrt{2}\,\frac{\mu NI}{D},$$

or[2]

$$B_m = \frac{I}{5}\sqrt{\frac{2\mu L \times 10^9}{whD}} \quad \text{gauss}. \tag{3}$$

2.2 *Copper Winding*

The copper winding is assumed to be applied by means of a shuttle that passes through the center of the annulus, leaving a circular opening unfilled with wire, as illustrated in Fig. 1, The depth of the winding is d, on the inside diameter of the core. This represents the basic limitation on copper winding area. The available winding area is therefore

$$\pi d(D - w - d) = A_c\,.$$

On a per-turn basis, the copper area is $s\pi d(D - w - d)/N$, where s is the packing factor, or the fraction of the available area occupied by copper conductor.

The average length of turn in the winding can be calculated closely on the assumption that the winding depth is uniformly equal to d both inside and outside of the core ring.* It amounts to $2(w + h + 2d)$.

The *total copper resistance* is thus

$$R_c = \frac{2\rho(w + h + 2d)N^2}{s\pi d(D - w - d)},$$

* This assumption is conservative, since it yields a slightly larger value than is obtained by more rigorous analysis.

or

$$R_c = \frac{\rho LD(w + h + 2d) \times 10^9}{2\mu s\pi dwh(D - w - d)} \quad \text{ohm}, \tag{4}$$

where ρ is the resistivity of the copper, in ohm-centimeters. This is the dc resistance, which may be considered as the entire copper resistance at frequencies low enough to make copper eddy current losses negligible. It is assumed to depend solely upon core dimensions and to increase smoothly, without discontinuities due to wire size changes. This assumption is based on the proposal to keep the present analysis simple, reserving more complicated analyses to subsequent sections.

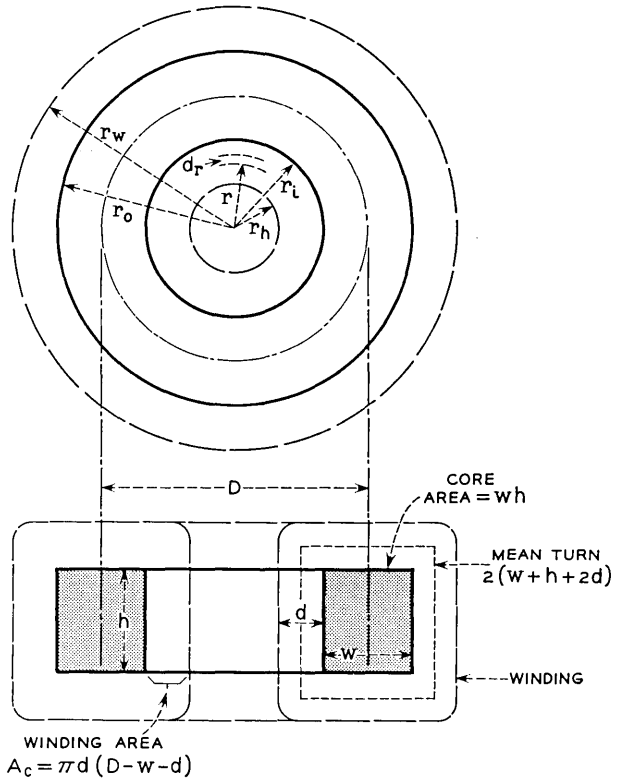

Fig. 1 — Idealized annular core and winding. Winding depth d is assumed constant for inside and outside of core ring. Core mean diameter is D, radial thickness w and axial height h.

2.3 *Coil Quality Factor, Q*

The quality factor of a coil, which enters into many important circuit computations, is $\omega L/R$, or

$$Q = \frac{2\pi fL}{R_c + R_m} = \frac{2\pi fL}{\frac{\zeta L}{\mu s} + \mu Lf(aB_m + c + ef)}, \tag{5}$$

where the dc resistance coefficient is

$$\zeta = \frac{\rho D(w + h + 2d) \times 10^9}{2\pi dwh(D - w - d)}. \tag{5a}$$

Simplifying (5) yields an expression that is independent of L:

$$Q = \frac{2\pi}{\frac{\zeta}{s\mu f} + \mu(aB_m + c + ef)}. \tag{6}$$

Inspection of this expression for Q shows that it is a maximizable function of the frequency f and the core permeability μ. Since frequency is the parameter that is always accessible for experimental control, it is chosen for initial analysis, with permeability optimization following, as explained below.* Inspection of the denominator of (6) shows that it will have a minimum when

$$\frac{\zeta}{\mu} = s\mu e f_0^2,$$

or

$$f_0 = \frac{1}{\mu}\sqrt{\frac{\zeta}{es}} \quad \text{cycles per second}, \tag{7}$$

where f_0 is the frequency at which the denominator of (6) will be a minimum and the coil Q will be a maximum. It is interesting to note in this expression that the optimum frequency is a function of the dc resistance coefficient and of the permeability and eddy current coefficient of the core, but that it is not explicitly a function of the hysteresis and residual losses. If a given core material is to be used at a specified optimum frequency, it is necessary to solve (7) for ζ, the dc resistance coefficient, which will fulfill the requirements.

Inserting this value of ζ into (6) gives the maximum Q of the coil, which is reached at the optimum frequency. Thus,

* The choice of frequency for initial consideration is consistent with the procedure of Arguimbau.[3]

$$Q_0 = \frac{\pi}{\mu e f_0 + \mu \dfrac{(aB_m + c)}{2}}. \tag{8}$$

It will appear in further analysis that coil design is facilitated by focusing attention on Q_0, the maximum quality factor. This shifts around to various frequencies, as indicated by (7), for cores having different values of μ, ζ and e.

The quality factor at frequencies other than f_0 will be smaller than Q_0 by a factor related to the difference in frequency from f_0. If we take the ratio Q/Q_0 from (6) and (8) and recall, from (7), that $\zeta/\mu = s\mu e f_0^2$ and, from (8), that $(aB_m + c) = 2\pi/\mu Q_0 - 2ef_0$, we may insert these values, rearrange terms, and obtain

$$\frac{Q}{Q_0} = \frac{1}{1 + \dfrac{e\mu Q_0}{2\pi f}(f_0 - f)^2}. \tag{9}$$

This equation shows that the shape of the Q versus frequency curve is roughly parabolic with vertex at Q_0, located at frequency f_0. As pointed out previously, it will generally be profitable to design coils based on Q_0, and resort to (9) when behavior at nonoptimum frequency is sought. Fig. 2 gives a log-log graph of a typical Q versus f curve.

Equation (8) gives an explicit relationship between core loss constants and coil Q at the optimum frequency f_0, and reveals that the maximum Q is inversely proportional to μ. Equation (8) also shows that Q_0 and f_0 are inversely related, for a given core permeability and eddy current coefficient. It is thus evident that Q_0 and f_0 are not open to arbitrary selection and that, if such selection is required, adjustment of core permeability and/or eddy current loss coefficient will be necessary. Such arbitrary adjustments of characteristics of compressed powder cores are not available to coil designers under usual circumstances. Core manufacturers have met their needs by standardizing on a line of core characteristics[2] stepping from low permeability to permeability as high as practicable, together with eddy current coefficient as small as possible.

It was noted in discussion of (7) that the optimum frequency f_0 depends on ζ, which includes geometric constants of the core and winding, and on the resistivity of the copper winding. It is instructive to analyze (8) on the basis of eliminating f_0, using the values from (7). Thus, the optimum Q can be expressed alternatively as

$$Q_0 = \frac{\pi}{\sqrt{\dfrac{\zeta e}{s}} + \dfrac{\mu(aB_m + c)}{2}}. \tag{10}$$

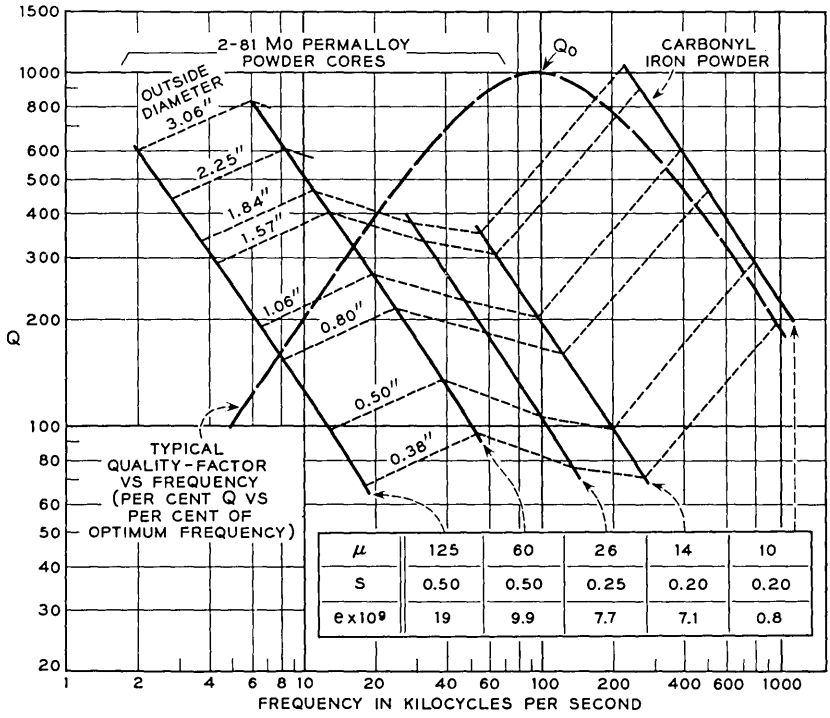

Fig. 2 — Maximum Q and optimum frequency for annular cores of conventional diameters and materials. Hysteresis and residual losses are assumed to be negligible. Dimension rations: $w' = \frac{1}{3}$, $h' = \frac{3}{8}$, $d' = \frac{1}{6}$. Commercial cores that differ from these ratios require specific calculation to determine possible further impairment of Q.

Equation (10) resembles (8), except that the copper resistance contribution to Q has become conspicuous, through its implied presence in f_0 from (7). In revealing the frequency at which Q reaches a maximum, (7) must hold, with either expression for Q_0. If hysteresis and residual core losses can be neglected in comparison with $\sqrt{\zeta e/s}$, (10) reduces to

$$Q_0 = \pi \sqrt{\frac{2s\pi dwh(D - w - h)}{\rho De(w + h + 2d) \times 10^9}}. \tag{11}$$

Equation (11) contains specific geometric dimensions of the core and winding. For clarification, it is reasonable to assume that the *relative* dimensions of any coil will be constant, regardless of the coil size. For example, the dimensions of the "archetype" coil may all be assumed proportional to the mean diameter of the core, D; i.e., $w = w'D$, $h = h'D$ and $d = d'D$, where the primed letters indicate fractions. The winding

resistance coefficient, from (5a), is

$$\zeta = \frac{\rho(w' + h' + 2d') \times 10^9}{2\pi D^2 d' w' h' (1 - w' - d')}. \tag{11a}$$

Thus (10) becomes

$$Q_0 = \frac{\pi D}{\sqrt{\frac{\rho e(w' + h' + 2d') \times 10^9}{2s\pi d' w' h' (1 - w' - d')} + \frac{\mu D(aB_m + c)}{2}}}. \tag{12}$$

It is apparent that the maximum Q of an annular coil is directly proportional to the diameter of the core, to a first approximation, when hysteresis and residual losses are small. Core manufacturers have met the need for Q versus diameter adjustments by standardizing on a line of core diameters ranging from heavy cores down to a size so small that special winding machine shuttles are required to pass through the hole in the winding. The choice of dimension ratios depends on wise apportionment of space to magnetic core, copper, wire insulation, mechanical strength of core, dimensions and shape of winding shuttle, etc. The relative dimensions, width, height and winding depth of commercial cores have been adjusted to fulfill specific design needs. It has been found practicable to maintain for the winding shuttle a hole size no less than $D/3$.

A type of core well suited to Formex-insulated solid-copper wire windings maintains the dimension ratios approximately $w' = \frac{1}{3}$, $h' = \frac{3}{8}$ and $d' = \frac{1}{6}$. It is instructive to compute f_0 and Q_0 for a series of cores of conventional materials based on these dimension ratios. Copper resistivity may be taken as 1.75×10^{-6} ohm-cm, and the winding packing factor as $s = 0.50$. The inefficiency of packing is due partly to the space occupied by insulation, but largely to the failure of circular cross-sectioned wire to occupy all of the available space. Upon inserting these values in (12) and neglecting hysteresis and residual losses, the maximum Q becomes

$$Q_0 = \frac{420D}{\sqrt{e \times 10^9}}. \tag{13}$$

The same substitutions in (7) yield the optimum frequency at which the maximum Q is obtained:

$$f_0 = \frac{7400}{\mu D\sqrt{e \times 10^9}} \quad \text{kc}. \tag{14}$$

Equations (13) and (14) apply to typical cores wound with Formex-

TABLE I — OPTIMUM FREQUENCY AND MAXIMUM Q OF IDEAL FORMEX WIRE-WOUND CORE RINGS

Dimension ratios: $w' = \frac{1}{3}$, $h' = \frac{3}{8}$, $d' = \frac{1}{6}$. Note: Core rings listed below differ more or less from these assumed dimension ratios; Q_0 and f_0 values will require specific calculation for accuracy.

Piece part number	O.D., inches	D, cm	f_0, kc	Q_0
Permalloy powder with $\mu = 125$, $e = 19 \times 10^9$, $a = 1.6 \times 10^6$, $c = 3.0 \times 10^5$				
P-11D719*	0.38	0.72	18.8	68
A-050056-2	0.50	1.01	13.3	97
A-206068-2	0.80	1.65	8.3	155
A-903157-2	1.06	2.06	6.8	190
A-254168-2	1.57	3.07	4.4	290
A-438281-2	1.84	3.53	3.8	335
A-109156-2	2.25	4.63	2.9	445
A-866142-2	3.06	6.35	2.1	600
Permalloy powder with $\mu = 60$, $e = 9.9 \times 10^9$, $a = 2.5 \times 10^6$, $c = 5.0 \times 10^5$				
P-11D719*	0.38	0.72	54.3	96
A-051027-2	0.50	1.01	38.8	134
A-848032-2	0.80	1.65	23.7	216
A-894075-2	1.06	2.06	19.0	270
A-083081-2	1.57	3.07	12.7	410
A-795135-2	1.84	3.53	11.1	460
A-488075-2	2.25	4.63	8.4	620
A-123068-2	3.06	6.35	6.2	830

* New size (Western Electric Company)
All "A" core designations follow Arnold Engineering Co. general catalog. These cores may be commercially available also from several other manufacturers. Cores of these dimensions were originally standardized by the Western Electric Company for telephone system applications.

insulated wire; they would be altered somewhat for cores of different geometrical ratios. They neglect hysteresis and residual losses, so that the actual quality factor will be correspondingly smaller than Q_0. In Table I computations of Q_0 and f_0 are shown for a series of commercial cores of various diameters, using permalloy powder cores of permeabilities 125 and 60. For any size of core, it will be noted from the table that Q_0 and f_0 are higher for 60-permeability cores.* This trend would be observed with other grades of cores having still lower permeability, but computations would not be rigorous enough to be profitable, based only on the analysis thus far presented. The reasons for this will be explained in the next section.

Thus far we have examined the coil quality factor Q in the relationships between core loss resistance and copper winding dc resistance. It

* Q_0 for 60-permeability cores is disproportionately higher, due to the decreased eddy current coefficient, as compared with 125-permeability cores.

has been found that Q reaches a maximum at a frequency f_0, defined by (7), and that the maximum value of Q_0 can be expressed in terms of core properties, (8), or of copper winding and core properties, (10). If core and winding dimensions are assumed to be proportional to core diameter, D, (12) gives a basic expression for Q_0, from which the effects of core diameter and permeability can be derived.

III. WINDING COMPLICATIONS

The analysis thus far is adequate for inductor designs for operating frequencies below about 30 kc. Greater rigor in analysis of coil design leads to a closer scrutiny of the windings. It is obvious that ac magnetic induction permeating a layer of copper wires will set up eddy currents, which result in energy loss. This correspondingly increases effective resistance of a coil, and decreases quality factor. Similarly, capacitance and leakance between the coil terminals increase losses and decrease Q. These winding complications are somewhat involved, and they will now be taken up one at a time.

3.1 *Copper Eddy Currents*

It has already been noted that eddy current losses in a magnetic core result in effective resistance contribution, $R_e = e\mu Lf^2$. The magnetic permeability of copper wire is unity, so that we may write a corresponding expression for copper eddy current resistance, $R_{ce} = R_c m f^2$, where the constant m is a function of the degree of subdivision of the copper. Calculation of m is given in the Appendix.

Use of stranded instead of solid wire involves a sacrifice of at least a further 25 to 30 per cent in winding-space efficiency, due to the large amount of insulation and the inefficiency of packing many strands in each turn of the winding. Thus, s will be reduced to 0.25 or less, in comparison with about 0.50 for nonstranded Formex-insulated wire. It is therefore evident that stranding of wire is not profitable at low frequencies, nor until the savings in eddy current loss at least offset the increase in dc resistance due to inefficient packing.

Analysis of the interaction of copper eddy current losses with other coil properties can be made by reference to (6). The contribution of copper eddy current loss will affect the coil Q both by the factor $(1 + mf^2)$ and by substitution of an inferior packing factor s. Thus, (6) becomes

$$Q = \frac{2\pi}{\dfrac{\zeta(1 + mf^2)}{s\mu f} + \mu(aB_m + c + ef)}. \tag{15}$$

Solving for conditions for maximum Q yields the necessary new optimum frequency,

$$f_{00} = \frac{1}{\mu}\sqrt{\frac{\zeta}{e\left(s + \frac{m\zeta}{\mu^2 e}\right)}}. \tag{16}$$

Comparing this with (7), it is evident that copper eddy currents shift the optimum frequency downward by the relation

$$f_{00} = \frac{f_0}{\sqrt{1 + \frac{m\zeta}{\mu^2 es}}}. \tag{17}$$

We can now proceed to calculate the new optimum Q, following a procedure similar to that of Section 2.3. Solving (16) for the copper resistance coefficient gives

$$\zeta = \frac{se\mu^2 f_{00}^{\,2}}{1 - mf_{00}^{\,2}}. \tag{18}$$

Substituting in (15) gives the new maximum quality factor,

$$Q_{00} = \frac{\pi}{\frac{\mu e f_{00}}{1 - mf_{00}^{\,2}} + \frac{\mu(aB_m + c)}{2}}. \tag{19}$$

This corresponds to (8), but shows the penalty exacted by eddy currents in the coil windings. If hysteresis and residual losses may be neglected, (19) can be rewritten

$$Q_{00} = \frac{\pi(1 - mf_{00}^{\,2})}{\mu e f_{00}}. \tag{20}$$

The reduction in Q due to copper winding eddy current losses is represented by the term $mf_{00}^{\,2}$, which should be as small as possible in relation to the first term, 1. Decision as to where stranding will be required rests on a further analysis of the term $mf_{00}^{\,2}$. From (50) of the Appendix, the critical term can be written

$$mf_{00}^{\,2} = \frac{932 \times 10^{-6} f_{00}^{\,2}}{nN}[Dsd'(1 - w' - d')]^3. \tag{21}$$

For the cores considered in Table I, which have $w' = \frac{1}{3}$ and $d' = \frac{1}{6}$, (21) becomes

$$mf_{00}^{\,2} = \frac{0.539 \times 10^{-6} f_{00}^{\,2} D^3 s^3}{nN}. \tag{21a}$$

The value of the packing factor, s, is linked to the number of strands, n, changing from about 0.50 when $n = 1$ to 0.25 and progressively lower for higher values of n. Designs of coils are usually arranged to make the winding turns, N, large enough to adjust the specified inductance to a precision of better than 2 per cent per turn; i.e., N must be 100 turns or more.

With these assumptions, Table II has been calculated. It is apparent that nonstranded wire ($n = 1$) is already very disadvantageous at a frequency of 50 kc. Multiple stranding of very high n is required at higher frequencies, and for large-size cores. For example, wire composed of 30 strands entails a decrease in Q of 9.8 per cent at 50 kc, on a 3-centimeter core. The table should be recognized as pessimistic, in that most windings will have more than the assumed 100 turns, and correspondingly will have less than the indicated value of eddy current loss coefficient mf_{00}^2.

For purpose of inductor computation, it may now be convenient to express Q_{00} in terms giving explicit recognition to the copper resistance coefficient ζ, instead of (20). Thus, inserting the value of f_{00} from (16) in (15) gives an alternative expression for the new maximum quality factor:

$$Q_{00} = \frac{\pi}{\sqrt{\frac{\zeta e}{s}\left(1 + \frac{m\zeta}{es\mu^2}\right)} + \frac{\mu(aB_m + c)}{2}}. \tag{22}$$

TABLE II — VALUES OF mf_{00}^2 = FRACTION DECREASE IN TOROIDAL COIL Q DUE TO STRANDED WIRE WINDING (100 TURNS)

Frequency f_{00}, kc	Core diameter, cm	Number of strands and winding efficiency			
		$n = 1$, $s = 0.5$	$n = 7$, $s = 0.25$	$n = 30$, $s = 0.20$	$n = 81$, $s = 0.20$
50	1	1.7	0.030	0.0036	0.00133
	3	—	0.81	0.098	0.029
	6	—	—	0.78	0.283
100	1	—	0.12	0.0144	0.0053
	3	—	—	0.39	0.0144
	6	—	—	—	0.116
200	1	—	0.48	0.058	0.021
	3	—	—	0.157	0.058
	6	—	—	—	0.467
400	1	—	—	0.232	0.086
	3	—	—	—	0.23
	6	—	—	—	—

This corresponds to (10), but shows the penalty exacted by copper eddy current losses. If hysteresis and residual losses may be neglected, (22) reduces to the form

$$Q_{00} = \frac{\pi}{\sqrt{\frac{\zeta e}{s}\left(1 + \frac{m\zeta}{es\mu^2}\right)}}. \tag{23}$$

This expression can be converted by inserting the values of ζ and m into an equation resembling (13):

$$Q_{00} = \frac{1.88 \times 10^{-2} \sqrt{\frac{s}{e}} D}{\sqrt{1 + \frac{0.015 D s^2}{N e \mu^2}}}. \tag{24}$$

This equation is inconvenient for general use, since it requires knowledge as to the specific values of winding turns and strand number involved in the term introduced by copper eddy current losses. Assumptions as to a tolerable value for this term may be made — for example, that it decrease the coil Q by, say, 10 per cent. The necessary stranding can then be calculated for any core material, diameter and number of winding turns. It should be remarked that these computations apply at the optimum frequency f_{00}, defined by (16).

3.2 *Winding Capacitance*

The ends of the coil winding have unavoidably some small capacitance and associated conductance with respect to each other. Similarly, adjacent turns have mutual capacitance, and they have capacitance to the magnetic core. These several capacitances and conductances may be taken effectively as lumps, C and G, shunted across the terminals of the coil, so as to be in parallel with the inductance L and resistance R of the coil, which are taken to be in series with each other. Computations of such a network at frequency corresponding to $\omega = 2\pi f$ show that it will be observed to have inductance and resistance

$$L_{\text{obs}} = \frac{L(1 - \omega^2 LC) - CR^2}{(1 - \omega^2 LC)^2 + 2GR + G^2(R^2 + \omega^2 L^2) + \omega^2 C^2 R^2} \tag{25}$$

and

$$R_{\text{obs}} = \frac{R + G(R^2 + \omega^2 L^2)}{(1 - \omega^2 LC)^2 + 2GR + G^2(R^2 + \omega^2 L^2) + \omega^2 C^2 R^2}. \tag{26}$$

These expressions simplify at frequencies well below resonance to the approximations

$$L_{\text{obs}} \doteq L(1 + \omega^2 LC), \tag{27}$$

$$R_{\text{obs}} \doteq (R + G\omega^2 L^2)(1 + 2\omega^2 LC). \tag{28}$$

It is evident that both inductance and resistance of a coil effectively increase above their initial values, with the addition term proportional to the square of the frequency. If a precise inductance is required at some operating frequency, the contribution due to distributed capacitance must be taken into consideration. Furthermore, this contribution is unfortunate in that it robs the coil of that constancy of inductance which may be desired for circuit performance over a broad band of frequencies.

Since distributed capacitance is undesirable, means are sought for reducing its value. Such means include spacing apart the end turns of the toroidal winding, bank-winding the entire coil and spacing the winding away from the magnetic core by means of material having low dielectric constant and low conductance (high dielectric Q_c). Such spacing is desirable in that it provides a controlled low capacitance, in series with the capacitance due to the core material, composed of insulated metallic particles, or perhaps ferrite, having high dielectric constant.

The dielectric quality factor, Q_c, for such composite systems of capacitance tends to be much lower than for the usual insulating materials — values around $Q_c = 30$ are not uncommon. This factor is very much dependent upon the moisture content of the dielectrics, and it can be improved or multiplied several fold by thorough drying of the coil and core structure. Proper design requires provision of permanent means for exclusion of moisture from a coil.

Spacing of windings away from core is costly, since this absorbs possible winding space and compels the use of smaller gage wire than was assumed in the computations of Section 2.3 and Table I. As already noted in Section 3.1, the impairment of copper winding space due to stranding shows up in a drastic decrease in the winding packing factor s and a corresponding increase in the dc resistance. The exact apportionment of available space to insulating spacing and copper tends to be a compromise, which is often resolved by the availability of insulating materials in convenient thicknesses, or of stranded wires.

The ultimate criterion of coil design is generally the coil quality factor Q achieved at the desired operating frequency. Since both inductance and resistance have been seen to depend upon the distributed capaci-

tance, the effective quality factor will reflect such dependence. The observed coil quality factor derived from (25) and (26) is

$$Q_{\text{obs}} = \frac{\omega L_{\text{obs}}}{R_{\text{obs}}} = \frac{\omega L\left(1 - \omega^2 LC - \dfrac{CR^2}{L}\right)}{R\left[1 + GR\left(1 + \dfrac{\omega^2 L^2}{R^2}\right)\right]}. \tag{29}$$

This expression can be simplified by inserting the intrinsic coil quality factor $Q = \omega L/R$ and the capacitance quality factor $Q_c = \omega C/G$. Making these substitutions and rearranging terms yields

$$Q_{\text{obs}} = Q\,\frac{1 - \omega^2 LC\left(1 + \dfrac{1}{Q^2}\right)}{1 + \omega^2 LC\left(1 + \dfrac{1}{Q^2}\right)\dfrac{Q}{Q_c}}. \tag{30}$$

Since intrinsic coil Q is generally much larger than one, the equation above can be rewritten to close approximation as

$$Q_{\text{obs}} = Q\,\frac{1 - \omega^2 LC}{1 + \omega^2 LC\,\dfrac{Q}{Q_c}}. \tag{31}$$

The effect of distributed capacitance on coil Q can be more clearly seen by expressing (31) in series form. Thus, performing the indicated division yields

$$Q_{\text{obs}} = Q\left[1 - \omega^2 LC\left(1 + \frac{Q}{Q_c}\right)\left(1 - \omega^2 LC\,\frac{Q}{Q_c} + \omega^4 L^2 C^2\,\frac{Q^2}{Q_c^2} - \cdots\right)\right], \tag{32}$$

which is convergent for $\omega^2 LCQ/Q_c < 1$. For small values of $\omega^2 LCQ/Q_c$ the series reduces practically to its first two terms.

The analysis of inductor design again has lost generality, in that impairment to Q at any frequency is dependent upon the specific values of L, C, Q and Q_c. Useful design advance can be made by setting some limit on the value of $\omega^2 LC$, such as, say, 0.02. Reference to (27) indicates that the inductance under this condition would appear to be 2 per cent higher than its low frequency value, L. Regarding the observed Q, (32) indicates that Q is decreased by the same percentage for an inductor having very high Q_c, and by a greater percentage for inductors in which Q_c is low in relationship to Q. Thus, for $Q/Q_c = 5$ and $\omega^2 LC = 0.02$, (32) indicates that the observed Q will be approximately 10 per

cent less than for the same inductor with very high Q_c. Similar calculation with $Q/Q_c = 1$ yields a loss of Q of some 4 per cent, corresponding to the assumed 2 per cent increase of inductance. Such illustrative calculations show the importance of maximizing Q_c — by use of low-loss dielectric materials in wire insulation, spacing and core insulation; by rigorous drying of coil and core structure before use; and by preservation of dryness throughout the lifetime of the inductor by means of hermetic seals.

The relationship of specific values of inductance and distributed capacitance can be derived, illustratively, from (27) on the above-assumed basis that $\omega^2 LC = 0.02$. The following table has been computed for the inductance which satisfies the assumption (2 per cent increase) at various frequencies and several values of the distributed capacitance C:

	Inductance limit (maximum), in millihenries			
C, in $\mu\mu$f:	5	10	20	50
Frequency, in kc				
30	113.0	56.3	28.2	11.3
50	40.6	20.3	10.15	4.06
100	10.1	5.06	2.58	1.12
150	4.54	2.27	1.13	0.454
200	2.54	1.27	0.63	0.254

This table show the limitations on inductance necessary to maintain less than 2 per cent apparent increase at any frequency. It is evident that low distributed capacitance is desirable, if large inductance values are to be employed. Experience shows that capacitances as low as those of the first two columns can be realized only at the expense of great pains in spacing of winding from the core, and in careful bank winding of wire.

Measurement of distributed capacitance may be conveniently made with the use of (27) and ac bridge measurements of inductance over a frequency range wide enough to make the term $\omega^2 LC$ vary from approximately zero up to say 0.02 or more. Q-meter measurements, in which the resonating capacitance is observed at a low frequency f_1 and again at a frequency twice f_1, are also convenient. The distributed capacitance is then obtained as $(C_1 - 4C_2)/3$, where C_1 and C_2 are the resonating capacitances at f_1 and $2f_1$, respectively. Both these methods assume constancy of intrinsic inductance over the frequency range, i.e., freedom from appreciable eddy current shielding in the core material.

IV. PRACTICAL DESIGN PROCEDURE

It was found in Section 2.1 that copper winding resistance and core loss resistance can be adjusted to yield the largest possible coil quality

factor Q_0 at some frequency f_0, with these parameters being functions of core size, permeability and loss characteristics. Complications enter at higher frequencies, due to eddy current losses in the copper wire winding and to capacitance and dielectric losses in the insulating materials adjacent to the winding. Copper eddy current losses impair the quality factor by the factor mf_{00}^2, as shown in (19) and more specifically in (21a), where the interplay of core size, winding turns and strand number is shown. For practical coils, the turns number is generally more than 100, in which case Table II gives the values of mf_{00}^2 for common stranding numbers. From this table it is possible to select the stranding number needed to avoid excessive impairment of Q at any frequency.

It is to be noted that the winding packing factor s declines markedly upon the adoption of stranded wire. Similarly, the steps taken to minimize distributed capacitance in the winding entail sacrifices in the possible s. The minimum winding hole size that preserves space for a shuttle is $D/3$. Experience shows that copper packing factor s in the available winding space ranges from about 0.50 for Formex-insulated solid wire to 0.25, 0.20, or lower for stranded wire windings with bank winding and anticapacitance spacing from the core. A further impairment to packing factor is introduced by the intervals between commercial wire sizes, which may prevent maximum filling of winding space with the number of turns needed to obtain the desired inductance. Theoretically, this impairment means that the packing factor will range downward by as much as 20 per cent of its highest attainable value before the next regular wire size may be used for producing the desired inductance in a regular series of inductances on a given core. Reference to (24) shows that a reduction of 20 per cent in packing factor is reflected in a reduction of Q by more than 10 per cent. This range of impairment must be considered in any practical design based upon the foregoing calculations, which have assumed step-free change of wire sizes.

Once the numerous sources of impairment to quality factor Q are recognized, it is necessary to embark on a practical design by suppressing the impairments as much as possible. Thus, (24) yields the maximum Q for any diameter of core, packing factor and core eddy current coefficient. Assuming that wire stranding is such as to eliminate copper eddy current losses, and that bank winding and spacing eliminate capacitance impairments, calculations of Q_0 can be made for other types of cores than those given in Table I. Such calculations are reported in Table III, for permalloy powder cores of lower permeabilities, and for carbonyl iron powder cores. A summary of core data from both Table I and Table III is shown in Fig. 2. This graph has been prepared to show maxi-

TABLE III — OPTIMUM FREQUENCY AND MAXIMUM Q OF STRANDED WIRE-WOUND CORE RINGS

Dimension ratios: $w' = \frac{1}{8}$, $h' = \frac{3}{8}$, $d' = \frac{1}{6}$; hysteresis and residual losses neglected.

Material	O.D., inches	D, cm	f_0, kc	Q_0
Permalloy powder with $\mu = 26$, $e = 7.7 \times 11^9$, $a = 6.9 \times 10^6$ $c = 9.6 \times 10^5$ $s = 0.25$	0.38	0.72	142	77
	0.50	1.01	102	109
	0.80	1.65	62	178
	1.06	2.06	49	221
	1.57	3.07	33	330
	1.84	3.53	29	380
Permalloy powder with $\mu = 14$, $e = 7.1 \times 11^9$, $a = 11.4 \times 10^6$, $c = 14.3 \times 10^5$, $s = 0.20$	0.38	0.72	275	72
	0.50	1.01	197	101
	0.80	1.65	121	165
	1.06	2.06	97	206
	1.57	3.07	65	307
	1.84	3.53	56	353
Carbonyl iron powder with $\mu = 10$, $e = 0.8 \times 11^9$, $a = 5.0 \times 10^6$, $c = 6.0 \times 10^5$, $s = 0.20$	0.38	0.72	1160	213
	0.50	1.01	830	300
	0.80	1.65	500	460
	1.06	2.06	400	620
	1.57	3.07	270	910
	1.84	3.53	235	1020

mum Q and optimum frequency for permalloy powder cores and carbonyl iron powder cores of typical dimensions. It is useful in selecting core size and permeability to fulfill any desired Q and frequency requirement. The graph must be recognized as optimistic, in that it does not show the reductions in Q due to copper eddy current losses, distributed capacitance losses and failure to fill the winding space to its maximum capability. These reductions in Q can be estimated by means of the formulae provided in the text.

APPENDIX

Copper Eddy Current Losses in Windings on Annular Cores

Theoretical analysis of copper eddy current loss in solenoidal windings has been attacked by Wien.[4] We now proceed to a similar calculation of the more complicated shape of an annular winding.

Thus a turn of copper wire t centimeters in diameter in a coil experiences a transverse alternating flux $B \sin \omega t$, which induces eddy currents in the copper very similar to those in a magnetic lamination. In the dia-

gram (Fig. 3) flux transverse to the wire induces emf in the "ribbon" elements dy thick in the wire, which drives eddy currents as indicated. Assuming a length of l centimeters of wire, the emf in a circuit of area $l \times 2y$ is

$$e = 2ly \frac{dB}{dt} = 2lyB_m\omega \cos \omega t,$$

and

$$E = 4\pi lyf \frac{B_m}{\sqrt{2}} \quad \text{abvolt, rms.} \tag{33}$$

The resistance around the periphery of this area is

$$dR = \frac{\rho l}{\sqrt{\frac{t^2}{4} - y^2}\, dy} \quad \text{abohm},$$

when ρ is expressed in abohm-centimeters.

The corresponding power consumption is

$$dP = \frac{E^2}{dR} = \frac{8l\pi^2 f^2 B_m{}^2}{\rho} y^2 \sqrt{\frac{t^2}{4} - y^2}\, dy.$$

Integrating from $y = 0$ to $y = t/2$ gives the total power consumption in the length l of a copper wire as

$$P_1 = \frac{l\pi^3 f^2 B_m{}^2 t^4}{32\rho} \quad \text{erg}. \tag{34}$$

It is now appropriate to compute the value of the induction B_m, which causes the copper eddy current loss. This value will vary from point to point throughout the winding depending upon the integrated magnetizing forces due to current in the nearby turns of the winding.

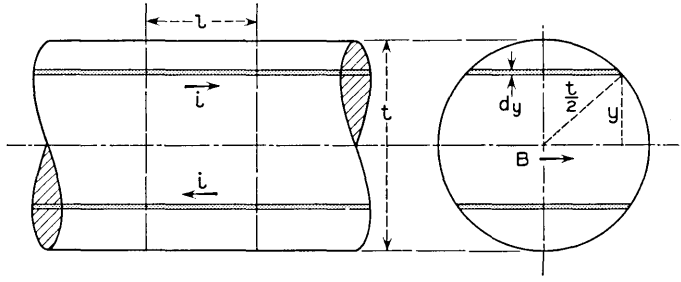

Fig. 3 — Sectional diagram of copper wire, showing eddy current sheets in sections of thickness dy. Magnetic induction B is in direction shown.

For example, if the winding comprises several layers upon an annular core (Fig. 1) the air induction in any layer through the hole in the core will be

$$B_r = H_r = 2\sqrt{2}\,\frac{N_r I}{r}, \tag{35}$$

where r is the distance of the layer from the center, and N_r is the number of turns in the winding, inside the radius r. This number of turns is

$$N_r = \frac{\pi\,(r^2 - r_h^2)}{a}, \tag{36}$$

where r_h is the radius of the hole left through the winding and a is the area of each turn, including copper, insulating material and packing inefficiency. It will be noticed that the analysis of the present section is based on the *radius* of winding layers, core, etc., in contrast with the earlier sections, which were based upon the mean *diameter* of the core. This modification is introduced in the interest of mathematical convenience and it will be converted later into terms of D.

Combining (35) and (36) yields the air induction transverse to any layer of wire through the hole in a core:

$$B_r = 2\sqrt{2}\pi\,\frac{(r^2 - r_h^2)I}{ar}. \tag{37}$$

Referring to (34), the power expended in a length l of a wire at radius r inside the core will be

$$P_{lr} = l\pi^5 f^2 t^4\,\frac{(r^2 - r_h^2)^2 I^2}{4\rho a^2 r^2}. \tag{38}$$

The entire power consumption in the layer requires multiplication of the single-wire loss, per (38), by the number of wires at radius r. Since the turns in the winding are presumed to be stranded, the wire diameter t applies to the individual strands, for this eddy current computation. The cross-sectional area occupied by one layer of turns is $2\pi r\sqrt{a}$, assuming square wire. The number of turns in the layer is then $2\pi r/\sqrt{a}$, and the number of strands is $2\pi rn/\sqrt{a}$, where n is the number of strands per turn. The number of strands per centimeter of radius will then be $2\pi rn/a$, and the power consumption in a differential thickness dr will be

$$dP_{lr} = \frac{nl\pi^6 f^2 t^4 I^2 (r^2 - r_h^2)^2\,dr}{2\rho a^3 r}. \tag{39}$$

The total loss for axial height l of winding inside the hole of the core is

obtained by integrating (39) between the limits r_i and r_h, where r_i is the radius of the hole in the core, as follows:

$$P = \frac{nl\pi^6 f^2 t^4 I^2}{2\rho a^3}\left[\frac{r_i^4 - r_h^4}{4} - r_h^2(r_i^2 - r_h^2) + r_h^4 \log \frac{r_i}{r_h}\right]. \quad (40)$$

The effective resistance due to this power loss is P/I^2, or

$$R_{ce} = \frac{nl\pi^6 f^2 t^4}{2\rho a^3}\left[\frac{r_i^4 - r_h^4}{4} - r_h^2(r_i^2 - r_h^2) + r_h^4 \log \frac{r_i}{r_h}\right]. \quad (41)$$

It is interesting to compare this resistance increase at frequency f with the dc resistance. Thus, the dc resistance of the N turns through the hole in the core, for an axial height l, is

$$R_c = \frac{4\rho l N}{n\pi t^2},$$

whence

$$\frac{R_{ce}}{R_c f^2} = \frac{n^2\pi^7 t^6}{8\rho^2 a^3 N}\left[(r_i^4 - r_h^4) - r_h^2(r_i^2 - r_h^2) + r_h^4 \log \frac{r_i}{r_h}\right]. \quad (42)$$

The winding area relationship gives

$$a = \frac{\pi(r_i^2 - r_h^2)}{N}, \quad (43)$$

whence

$$\begin{aligned} \frac{R_{ce}}{R_c f^2} &= \text{“}m\text{”} \\ &= \frac{n^2 N^2 \pi^4 t^6}{8\rho^2 (r_i^2 - r_h^2)^3}\left[(r_i^4 - r_h^4) - r_h^2(r_i^2 - r_h^2) + r_h^4 \log \frac{r_i}{r_h}\right]. \end{aligned} \quad (44)$$

Noting that the cross-sectional area of copper in any turn of the winding is $n\pi t^2/4$, (44), becomes

$$m = \frac{8\pi N^2 a_c^3}{n\rho^2} F(r), \quad (45)$$

where a_c is the cross-sectional area of all the copper strands in each turn of the winding, and $F(r)$ is a complicated function of the inside radii of winding hole and core hole.

The above calculations apply to that part of the winding in the hole through the core. Similar computations for the winding on the outer periphery of the core show that (44) becomes

$$\frac{R_{ce}}{R_c f^2} = m$$
$$= \frac{n^2 N^2 \pi^4 t^6}{8\rho^2 (r_w^2 - r_0^2)^3} \left[(r_0^4 - r_w^4) - r_w^2 (r_0^2 - r_w^2) + r_w^4 \log \frac{r_0}{r_w} \right], \tag{44'}$$

where the hole and inside core radii of (44) are replaced by the outside core radius r_0 and outside winding radius r_w, respectively. Obviously, the coefficient of (44) remains unchanged, for the inside and outside parts of the winding, and only the function $F(r)$ varies, to accommodate the different radii.

Computations for the radiating parts of the turns on the top and bottom of the core would lead to a complicated relationship, which essentially would yield some sort of average between the values of (44) and (44'). Rather than work through such expressions, we will simply postulate that the effective resistance of the entire copper winding will be a function of the mean diameter of the core supporting the winding. Thus,

$$m = \frac{8\pi N^2 a_c^3}{n\rho^2} F(D). \tag{46}$$

The functional relationship $F(D)$ in (46) is most conveniently found empirically, by resistance measurements at high frequencies of windings on nonmagnetic cores of various sizes. Using for the resistivity of copper 1750 abohm-centimeters, the equation above becomes

$$m = \frac{R_{ce}}{R_c f^2} = \frac{8.2 \times 10^{-6} N^2 a_c^3}{n} F(D). \tag{47}$$

Proceeding with construction and measurement of coils of important sizes has involved preparing nonmagnetic wooden or phenol fiber rings, and winding them as efficiently as possible in approximately bank distribution with stranded wire. A survey of data obtained over the past 20 years is given in Table IV. The data have been "boiled down" in terms of only basic geometrical facts of the core diameter, stranding, number and diameter of individual wires, number of turns in the winding and the value of $R_{ce}/(R_c f^2) = m$ measured over a sufficiently wide frequency range. The ratio of the observed value of m to the winding factors in (47) yields values for $F(D)$.

Inspection of Table IV shows that $F(D)$ is inversely proportional to D^3, over a considerable range of diameters. Thus, for D between 2 to 6 centimeters, the value of $F(D)$ averages $3.61/D^3$. Deviations of specific windings in this range amount to -36 and $+38$ per cent maximum. Considering that the average deviation is 15.7 per cent, the analysis is of sufficient accuracy to warrant its acceptance for a useful range of core

TABLE IV — DATA ON STRANDED WIRE-WOUND AIR CORE TOROIDS

D, cm	n, strands	B&S gage	t, cm $\times 10^3$	N, turns	a_c, cm^2 $\times 10^4$	$\frac{8.2\times10^{-6}N^2a_c^3}{n}\times10^{12}$	Observed $m=\frac{R_{ce}}{R_{cf}f^2}\times10^{12}$	$F(D)$	$F(D)\times d^3$
1.65	60	42	6.33	100	18.9	9.22	3.20	0.347	1.56
2.10	120	46	3.99	100	15.0	2.31	0.97	0.420	3.90
2.10	90	44	5.02	100	17.8	5.14	1.40	0.272	2.52
3.14	81	38	10.1	68	64.6	148	11.0	0.0743	2.30
3.7	7	40	8.00	400	3.507	8.09	0.66	0.0815	4.13
3.7	7	40	8.00	800	3.507	32.4	2.17	0.0670	3.40
3.7	7	40	8.00	1000	3.507	50.5	3.40	0.0673	3.42
3.7	7	40	8.00	1336	3.507	90.1	6.90	0.0765	3.88
3.7	19	40	8.00	208	9.52	16.1	1.23	0.0763	3.87
3.7	30	40	8.00	100	15.03	9.31	0.76	0.0816	4.13
3.7	30	40	8.00	180	15.03	30.2	2.22	0.0735	3.73
3.7	30	40	8.00	300	15.03	83.9	6.46	0.0769	3.90
3.7	30	40	8.00	400	15.03	149	11.0	0.0737	3.74
3.7	7	34	1.60	400	14.99	631	32.0	0.0507	2.57
3.7	1	25	4.55	400	16.24	5620	336	0.0598	3.03
6.20	30	36	12.7	600	38.0	5400	97.6	0.0181	4.33
6.20	30	40	8.00	1002	15.03	935	19.5	0.0209	4.99
9.05	1	21	72.3	390	41.05	8630	1825	0.0212	15.7
9.05	81	38	10.1	485	64.6	6430	116	0.0180	13.3
Average (omitting 1.65 cm and 9.05 cm cores)									3.61

sizes. Beyond 6.2 centimeters the coefficient deviates considerably. Thus (47) can be rewritten (for core diameters 2 to 6 centimeters):

$$m = 30 \times 10^{-6} \frac{N^2 a_c^3}{nD^3}. \tag{48}$$

Equation (48) is inconvenient, since it includes N, the number of turns in the winding, and a_c, the area of copper in each turn. These parameters can be combined by noting that Na_c is the total copper area in the winding, and N_{ac}/s is the entire available winding area $A_c = \pi d(D - w - d)$. Thus, for archetype cores, as discussed above, prior to (12), these relationships reduce to

$$Na_c = s\pi D^2 d'(1 - w' - d'). \tag{49}$$

Hence, (48) becomes

$$m = 30 \times 10^{-6} \frac{D^3}{nN} s^3\pi^3 d'^3(1 - w' - d')^3,$$

or

$$m = \frac{932 \times 10^{-6}}{nN} [Dsd'(1 - w' - d')]^3. \tag{50}$$

It appears, finally, that the copper eddy current loss coefficient is directly proportional to the cube of the mean diameter of the core, and inversely proportional to the number of turns and number of strands per turn of the coil winding.

REFERENCES

1. Legg, V. E., B.S.T.J., **15**, 1936, p. 39. 2. Legg, V. E. and Given, F. J., Elect. Engg., **59**, 1940, p. 414. 3. Arguimbau, L. B., General Radio Experimenter, **11**, No. 5, 1936. 4. Wien, M., Ann. d. Phys., **14**, 1, 1904, p. 1.

General Stochastic Processes in Traffic Systems with One Server

By V. E. BENEŠ

(Manuscript received September 1, 1959)

Congestion in systems with one server, exemplified by simple queues, individual telephone trunks and particle counters, is considered without any restrictions on the statistical character of the incoming load. Elementary methods establish formulas and equations describing probabilities of delay and loss. These methods deemphasize special statistical models and yield a general theory. In spite of this generality, it is attempted to give intuitive proofs and extensive explanations of the physical significance of formulas, as well as rigorous derivations. The theory is applied to specific models to obtain illustrative new results.

I. INTRODUCTION AND SUMMARY

Congestion theory is the study of mathematical models of service systems such as telephone central offices, waiting lines and trunk groups. It has two practical uses: to provide engineers with specific mathematical results — curves and tables — for designing actual systems, and to provide a general framework of concepts into which new problems can be fitted and in which current problems can be solved. Corresponding to these two uses are two kinds of results: *specific results* pertaining to special models and *general theorems* valid for many models.

Most of the present literature of congestion theory consists of specific results resting on particular statistical assumptions about the traffic in the service system under study. Indeed, few results are known in congestion theory that do not depend on special statistical assumptions, such as negative exponential distributions or independent random variables. In this paper we obtain some mathematical results that are free of statistical restrictions, but which apply only to a limited class of systems. These results concern general stochastic processes in traffic systems, such as counters and simple queues, that have only one server.

The limitation to a single server is severe, and may be difficult to

overcome. We hope that appropriate analogs of the methods used here, applied to stochastic processes in several dimensions, will yield statistically general theories for systems with many servers. Some systems, for example, queues with many servers in parallel, cannot be given a theory free of special assumptions without the use of multidimensional processes; other systems have a structure that permits us, in principle, to treat them by "iterating" results for one server. In Section VIII we suggest, as a conjecture, how the general theory we give for one telephone trunk might be applied several times to give a general loss formula for a group of trunks.

The classical problems of telephone traffic engineering usually involve many servers, and so theories dealing with single servers have been of peripheral interest to traffic engineering in the past. However, modern increases in the speed of components have made possible *electronic* telephone exchanges, now under study and development. By dint of their speed, some of these systems either have a single active server (the "common control"), or may reasonably be viewed as a set of (relatively independent) single servers. Therefore, the case of one server is in fact of immediate and practical interest.

The aims of this paper are three: (a) to describe a new general approach to certain congestion problems; (b) to show that this approach, although general, can nevertheless be presented in a relatively elementary way that makes it available to many persons; (c) to illustrate how the new approach yields specific results, both new and known. What follows is written only partly as a contribution to the mathematical analysis of congestion. It is also a frankly tutorial account, aimed at increasing the public understanding of congestion by first steering attention away from special statistical models and then obtaining a general theory. Such a point of view, it is hoped, will yield new methods in problems other than congestion.

When a general theory can be given, it should be useful in several ways: in increasing our understanding of complex systems; in obtaining new specific results, curves, tables, etc; in extending theory to cover interesting cases that are known to be inadequately described by existing results. At first acquaintance, the theorems of such a general theory may not resemble "results" at all; that is, they may not seem to be facts that one could obviously and easily use to solve a real problem. A general theory is really a tool or principle expressing the essence or structure of a system; properly explained and used, this tool will yield formulas and other specifics with which problems can be treated.

Let us give examples of questions we shall be able to answer and re-

sults we shall obtain. As a first example of increased understanding, we shall be able to answer, affirmatively, the question

> Is there a small number of specific features of the incoming traffic which suffices, quite generally, to determine delay distributions and loss probabilities?

We shall exhibit these features, and give delay and loss probabilities in terms of them.

As a second example, of both increased understanding and a new case, let us consider the matter of Markov stochastic processes. A Markov process has a particularly simple probabilistic structure, in that all information about the past history of the randomly moving point is irrelevant to its future development, if its present position is known. Such processes have been very useful in congestion theory, because of the convenient functional equations (such as "statistical equilibrium" equations, continuity equations, renewal equations) that are associated with them. However, it is also true that congestion theory has occupied itself almost exclusively with Markov processes, or with the generalizations of these called regenerative, and semi-Markov, processes. Thus the following question has doubtless been asked:

> What functional equations can be derived for non-Markov problems in delay, or in loss?

In answer, we show that the probability that a queue be empty at t satisfies, quite generally, a Volterra equation of the first kind. This equation coincides with a known equation in the Markov case, as it should; similar results hold for loss.

By way of specific results, we prove that, in many cases of practical interest described by Markov processes, the solution of the Volterra equation has the intuitive form

$$\Pr\{\text{queue is empty at } t\} =$$

$$\text{average of the greater of 0 and } \left(1 - \frac{\text{total load in } [0,t)}{t}\right)$$

$$+ \frac{\text{initial load}}{t} \text{ (chance that total load in } [0,t) \text{ is at most } t).$$

In these cases, it is no longer necessary to solve the Volterra equation; one merely computes the quantities on the right-hand side of the formula just given. The "initial load" is the time the queue would take to empty if no customers arrived after time zero; the "total load in $[0,t)$"

is the initial load plus the sum of all the service-times of customers arriving in $(0,t)$.

This result is illustrated by Fig. 1, which shows the probability that a queue is empty as a function of time for two well-known cases, negative exponential service times and constant service times. In each case, arrivals are in a Poisson process at the rate λ, and the mean service time is b. The curves suggest that the approach to equilibrium is considerably faster with constant service times than with exponential service times. The "exponential" curve was computed from (48), and the "constant" curve from the formula given above. Both curves approach the same limiting value, 0.750.

The physical background of intended applications is discussed briefly in Section II, while Section III describes the cumulative *load* or *traffic*. Sections IV and V are devoted, respectively, to equations characterizing delay and loss operation in terms of the offered load; these sections are purely descriptive, and no probability is involved. Probability first enters in Section VI, where we discuss the general nature of the problems we try to solve, the methods we use and the results we obtain. Formulas and equations are stated and explained for probabilities of delay in Section VII, and for probabilities arising in loss operation in Section

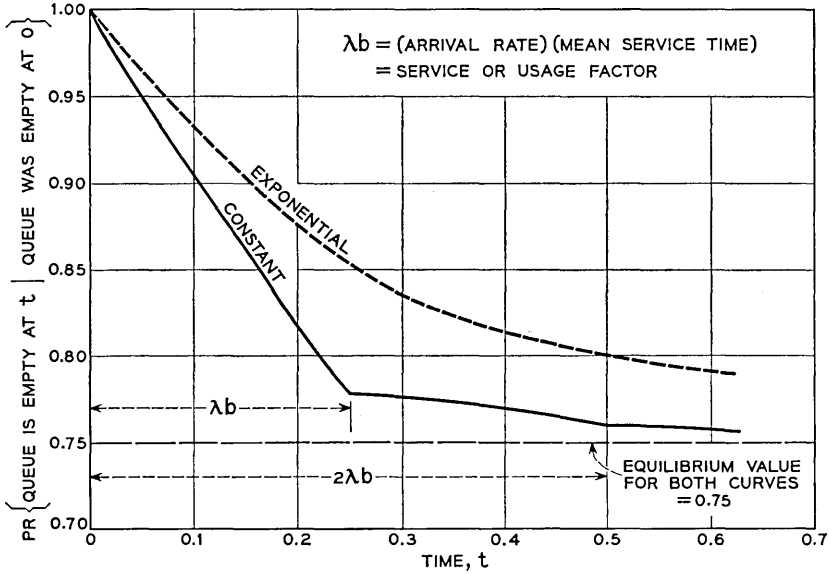

Fig. 1 — Probability that a queue, starting empty at 0, is empty at t, as a function of t for negative exponential and constant service times, with $\lambda = 1$, $b = 0.25$.

VIII. Proofs of, and additional comments on, the results for delay appear in Section IX, and a number of specific formulas are deduced in Section X. Precise derivations of the results for loss are in Section XI, with an additional limit theorem being presented in Section XII.

II. PHYSICAL BACKGROUND

We shall discuss results applicable to the following type of system: there is given a machine that is either idle or busy; every so often, at random instants, someone tries to use the machine; and the machine can be operated in two ways, called *loss* and *delay*. In delay operation, people who find the machine busy wait for their turns in order of arrival, and then make it busy (by using it) for the originally intended length of time; in loss operation, people who find the machine busy are sent away or "lost," while those who find the machine idle seize it and use it for a random time, the *service time*.

The system to be studied has been described in terms of people who try to use a machine, but the same structure appears in other applications. In telephony, for example, the machine to be used may be a marker, and the "people" arriving may be registers filled with dialed digits; telephone engineers usually refer to the service time as the *holding time*. In the study of radioactivity, the machine is an ionization counter, the "people" are impinging particles, and the service time (called the *dead time* of the counter) is a length of time after the registering of a particle during which the counter cannot count, so that it misses particles arriving during this interval.

The quantities of engineering interest in these models vary with the application, but the following ones seem to be of general importance: for *loss* operation, the chance of loss, i.e., the probability that an arriving customer find the machine busy, and the probability distribution of the busy period, i.e., the amount of time that must elapse until the machine next becomes idle; for *delay* operation, the distribution of waiting time.

III. THE CUMULATIVE LOAD, $K(t)$

Before we can study the two modes of operation, loss and delay, we must describe the offered load of work, or the arriving traffic. This is done most easily by using a step function, $K(t)$, which is nondecreasing and left-continuous. The locations of the jumps are the epochs of arrival of customers, and the magnitudes of the jumps are the service times or the lengths of time the machine is made busy. Equivalently, the offered load is completely determined by the arrival epoch t_k and the service-

time S_k of the kth arriving unit, $k = 1, 2, \cdots$. This situation is depicted in Fig. 2. We shall use $K(t)$ to describe the load for both loss and delay operation.

IV. INTEGRAL EQUATION DESCRIBING DELAY OPERATION

As a mathematical description of the delays to be encountered under delay operation, we use the virtual waiting time, $W(t)$, which can be defined as the time a unit would have to wait for "service" if it arrived at time t. At the epoch t_n of arrival of the nth unit, $W(t)$ jumps upward discontinuously an amount equal to S_n, the work or service time of the unit. Otherwise, $W(t)$ has slope -1 if it is positive; if it reaches zero, it stays equal to zero until the next jump of the load function $K(t)$. Corresponding graphs of $K(t)$ and $W(t)$ appear in Fig. 3.

If $K(t)$ is interpreted as the work offered in the interval $(0,t)$, then $W(t)$ can be thought of as the amount of work remaining to be done at time t. In terms of this interpretation, it can be seen that

work remaining at t = total work load offered up to t − elapsed time + total time during which machine was idle in $(0,t)$.

The machine is idle when, and only when, $W(u) = 0$. Then, formally, $W(t)$ is defined in terms of $K(t)$ by the integral equation

$$W(t) = K(t) - t + \int_0^t U[-W(u)]\,du, \qquad t \geqq 0, \qquad (1)$$

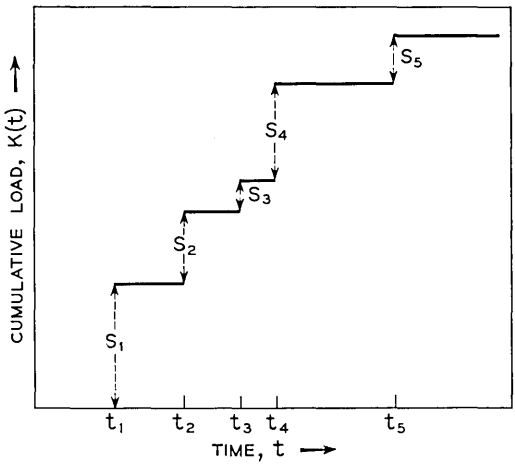

Fig. 2 — Cumulative load.

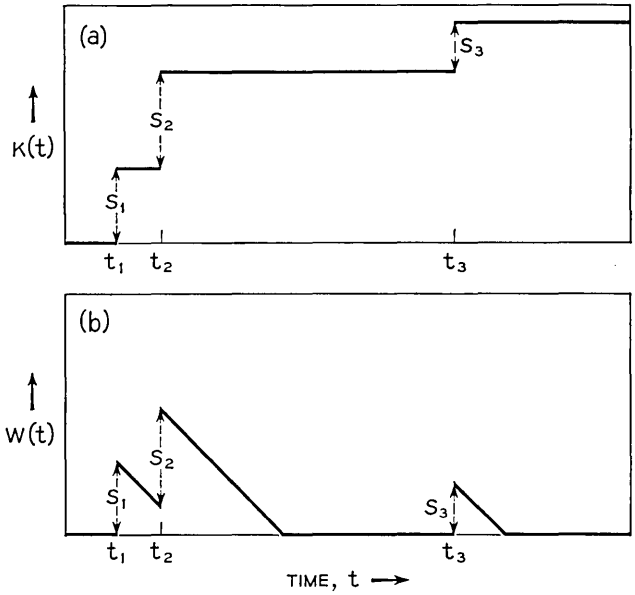

Fig. 3 — (a) Graph of $K(t)$ vs. t; (b) graph of $W(t)$ vs. t.

where $U(t)$ is the unit step function; i.e., $U(x) = 1$ for $x \geqq 0$, and $U(x) = 0$ otherwise. For simplicity, we have set $W(0) = K(0)$.

It has been shown by Reich[1] that the solution of (1) is, if $K(x) - x$ has a zero in $(0,t)$,

$$W(t) = \max_{0 \leqq x \leqq t} [K(t) - K(x) - t + x].$$

This fact may be interpreted physically as follows: the quantity in brackets, $[K(t) - K(x) - t + x]$, is, if positive, the excess of arriving load in the interval (x,t) over the elapsed time $t - x$; it is therefore the *overload* in (x,t). Reich's formula then says essentially that

$$\text{delay at } t = \max_{0 \leqq x \leqq t} [\text{overload in } (x,t)].$$

The relationship between the waiting time $W(t)$ and the offered traffic $K(t)$ can be further elucidated graphically by reference to Fig. 4. The light solid line shows $K(t) - t$, the traffic offered up to time t minus the traffic that could have been served if the server had been kept busy throughout the interval $(0,t)$. It is supposed in Fig. 4 that the server starts busy at $t = 0$. It is busy until $t = a$. At this point, the server becomes idle and $K(t) - t$ turns negative, and its negative value is the

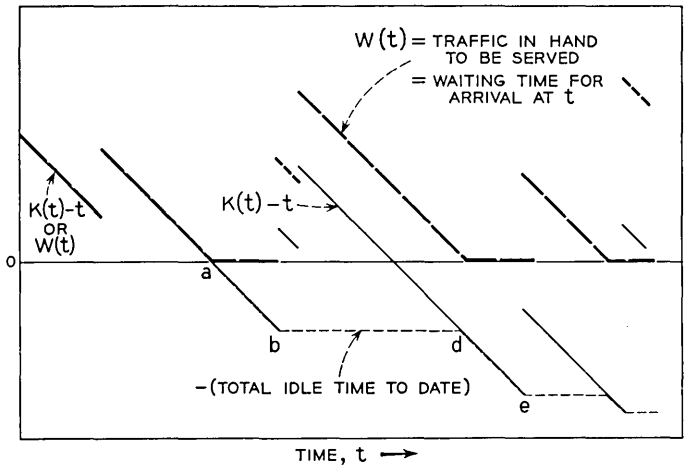

Fig. 4 — Relationship between waiting time and offered traffic.

negative of the idle time. At $t = b$, more traffic is offered and $K(t) - t$ jumps up.

The heavy dashed line represents the waiting time at t, $W(t)$; $W(t)$ can also be thought of as the traffic in hand and yet to be served. It can never go negative. It is equal to $K(t) - t$ before a, and is zero from a to b. At b, it jumps up, remaining above and parallel to $K(t) - t$ until $t = d$, when the server goes idle again. At b, $W(t)$ is above $K(t) - t$ by exactly the amount by which $K(t) - t$ was most negative. At d, when $W(t)$ reaches zero, $K(t) - t$ is just reaching its previous minimum, and $K(d) - d = K(b) - b$.

During the interval (d,e), $W(t)$ remains at zero, while $K(t) - t$ becomes more negative, establishing new minima as it goes and building up more idle time. At $t = e$, $K(t) - t$ and $W(t)$ both jump up; $W(t)$ is again parallel to $K(t) - t$, but it is now above it by an amount equal to the negative of the last minimum, $K(e) - e$.

In Fig. 4,

$$\min_{a \leqq x \leqq t} [K(x) - x], \quad t \geqq a,$$

is shown as a light dashed line. It is a monotone, nonincreasing function of t, and is the negative of the total idle time up to time t. To account for the period $t < a$, when $K(t) - t$ has not yet become negative and the server has not yet been idle, we write

$$W(t) = K(t) - t - \min \{0, \min_{0 \leqq x \leqq t} [K(x) - x]\},$$

and thus obtain another representation for the delay, i.e., a solution of (1).

In a manner similar to that of Fig. 4, Fig. 5 depicts, simultaneously, the offered load $K(t)$ in a light solid line; the waiting-time $W(t)$ in a heavy solid line; the negative of the accumulated idle time in a heavy dashed line; and the "load-time excess" $K(t) - t$, when it does not coincide with the negative of the idle time, in a light dashed line. The terminology in Fig. 5 has been purposely chosen to suggest an interpretation in terms of inventory or storage theory: $W(t)$ is the *real backlog* (of orders, say,), $K(t)$ is the *cumulative amount ordered* and $K(t) - t$ might be termed the load time excess or the *virtual backlog*. Then

$$\text{real backlog} = \text{virtual backlog} + \text{accumulated idle time.}$$

V. INTEGRAL EQUATION DESCRIBING LOSS OPERATION

For loss operation, we use $A(t)$, the service time remaining at t, as an indicator of the condition of the machine. We may define $A(t)$ as

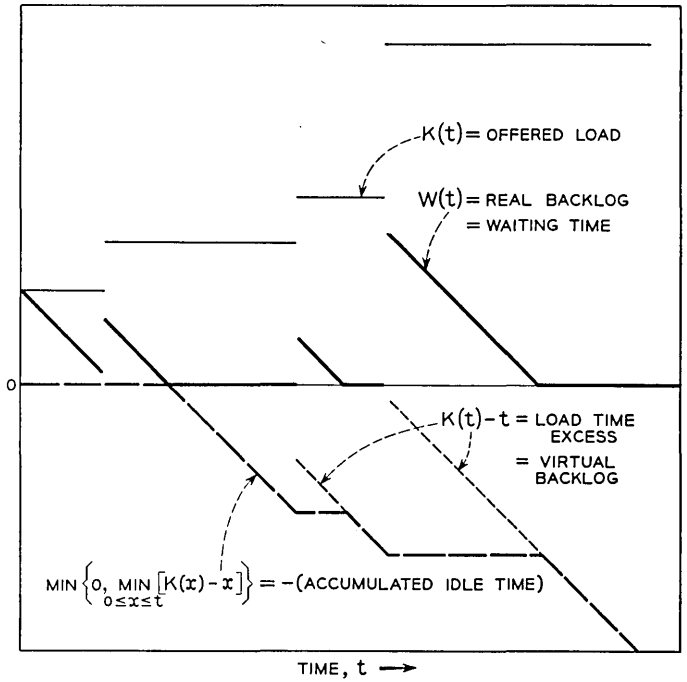

Fig. 5 — Relationships among offered load, waiting time, negative of accumulated idle time and "load time excess".

the amount of time remaining until the machine next becomes idle, with the convention that $A(t) = 0$ means that the machine is idle at t. At epochs of arrivals, $A(t)$ jumps upward discontinuously if the arrival in question finds the machine idle, the amount of jump being the service time of the arriving customer. Arrivals that find the machine busy do not affect $A(t)$. While $A(t)$ is positive, it has slope -1; when it reaches zero, it stays equal to zero until the next jump. A graph of $A(t)$ consists of isosceles "right triangles" laid on their sides; corresponding graphs of $K(t)$ and $A(t)$ appear in Fig. 6. For the study of $A(t)$, it is more convenient to define $K(t)$ to be right-continuous.

In terms of the interpretation of $K(t)$ as cumulative work load it is easy to see that

work remaining to be done at t = total work load offered up to t

- load missed in $(0,t)$ because machine was busy

- elapsed time + total time during which machine was idle in $(0,t)$.

This means that $A(t)$ is defined in terms of $K(t)$ by the equation

$$A(t) = K(t) - \int_{0+}^{t} \{1 - U[-A(u)]\}\, dK(u) - t + \int_{0}^{t} U[-A(u)]\, du, \tag{2}$$

where, as before, $U(t)$ is the unit step function, and we have set $A(0) = K(0)$.

VI. CHARACTER OF THE GENERAL RESULTS

Models of waiting lines and telephone traffic usually contain explicit assumptions about the statistical nature of the offered load $K(t)$. For instance, Erlang's original models[2] amount to assuming that the interarrival times $(t_n - t_{n-1})$ are all independent with the same negative exponential distribution; and similarly for the service times. These assumptions give a class of models whose parameters are the means of the negative exponential distributions.

A broader class of models is specified by retaining the assumptions that the interarrival times be independent and identically distributed (and similarly for service times), but allowing any distribution, not just the negative exponential. The interarrival and service time distributions may still be said to "parametrize" this broader class of models, since their choice determines a model in the class.

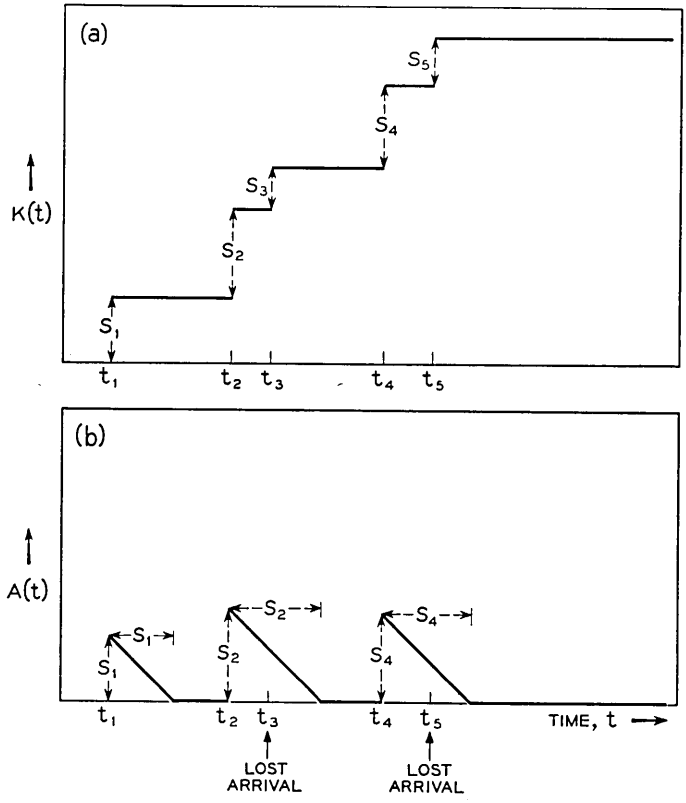

Fig. 6 — Graphs of (a) $K(t)$ vs. t and (b) $W(t)$ vs. t for loss operation.

As more general assumptions even than those of the last paragraph are considered, it becomes extremely laborious first to specify the model and then to compute interesting quantities such as distributions of delay, probabilities of loss, etc. So, instead of looking for ways of exactly characterizing the model, we can try to search directly for simple ways of expressing the quantities of interest in terms of the model. Since the probabilities

$$\Pr\{W(t) \leqq w\}, \tag{3}$$

$$\Pr\{A(t) \leqq w\} \tag{4}$$

are what we actually wish to compute from the model, the question arises whether this calculation can be made without first specifying the entire probabilistic structure of $K(t)$. The following intuitive argument

can be adduced for answering "yes". Both $W(t)$ and $A(t)$ are defined in terms of the load by very special relationships that are expressed in the integral equations we have given. Hence, no matter what are the statistical features of $K(t)$, it is likely that the distributions of $W(t)$ and $A(t)$ depend only on some very particular, physically interpretable statistical functions associated with $K(t)$. It is not obvious that such an economy can be made in the generality we desire.

Two principal results, described in the next sections, state that the probabilities (3) and (4) can in fact be given fairly simple expressions that are generally valid for any load process. These expressions depend only on certain special functions obtainable from the statistical structure of the load $K(t)$. Each function has a definite intuitive or physical significance, to be given later. These statistical functions achieve the desired economy of description, because we can state that the desired probabilities depend only on the features of $K(t)$ expressed in the functions. For the purposes of calculating (3) and (4), we do not need the entire probabilistic structure of $K(t)$, but only the small relevant part.

Another way of putting the problem that we have attempted to solve is: for general load processes $K(t)$, what small amount of information about the statistical nature of $K(t)$ will determine $\Pr\{W(t) \leqq w\}$ for all t and w, and how is this computation to be made? [A similar question arises for $A(t)$.] The statistical functions are then the *information about* $K(t)$; the formulas for the probabilities (3) and (4) indicate the *method of computation*. Specifically, (1) defines a transformation of a stochastic process, $K(t)$, of service times and arrival epochs into another stochastic process, $W(t)$, of waiting times. For each t, there is an operator or formula that gives the distribution of $W(t)$ in terms of suitable fundamental statistical functions associated with $K(u)$ for $u \leqq t$. The principal problem is to find the form of the operator and the character of the fundamental functions. The answer to this problem should depend only on the integral equation (1) and on the fact that $K(t)$ is a nondecreasing step function. It should depend on no special features of the probability measure for $K(t)$ except those implied by this last property. Accordingly, we shall assume at first that $K(t)$ is a random, nondecreasing step function; its *only* statistical peculiarity is that it is a nondecreasing step function.

There already exists an extensive literature dealing with probabilistic models of telephone trunks, electronic counters and similar machines; we shall attempt to relate our approach to this literature. The models used to date have usually included strong hypotheses of independence or else used special distribution functions, such as the negative exponen-

tial. The results obtained depend methodologically, almost without exception, on the possibility of writing Kolmogorov equations for a Markov process, or integral equations of renewal type for probabilities or expectations associated with a regenerative process (Smith[3]). The extensive work of Takács[4,5,6] falls entirely in this category (see also the references therein).

Our approach is, in a sense, an inversion of the usual method described above. The latter consists in first doing probability theory to set up Kolmogorov or renewal equations, and then doing analysis to solve the equations. We can, however, achieve greater generality by taking maximum advantage of the fact that the processes of interest, $W(t)$ and $A(t)$, already satisfy (1) and (2), respectively. This we do by careful analysis of $W(t)$ and $A(t)$ themselves, by in effect performing some of our analysis in the domain of random variables, and taking averages only at convenient points.

VII. PROBABILITIES OF DELAY

It has been shown[7] that $\Pr\{W(t) \leqq w\}$ can be expressed in terms of the two statistical functions

$$\Pr\{K(t) \leqq w\}, \tag{5}$$

$$R(t,u,w) = \Pr\{K(t) - K(u) - t + u \leqq w \mid W(u) = 0\}. \tag{6}$$

We first clarify the physical significance of the functions (5) and (6) entering into the representation of $\Pr\{W(t) \leqq w\}$. The first is simply the probability distribution of the cumulative load, $K(t)$. The second, written as

$$R(t,u,w) = \Pr\{[K(t) - t] - [K(u) - u] \leqq w \mid W(u) = 0\}$$

is the (conditional) probability distribution of the change in the "overload" that occurs in the interval (u,t), given that the server was idle at time u. Here we have used the term "overload" to mean the amount of work that would have been left still undone even if the traffic had been so spaced that there was no idle time. This overload is represented by $K(t) - t$. A negative value of the overload represents the negative of the idle time that would have occurred if the traffic spacing had permitted the server to finish all of the work by time t.

The significance of $R(t,u,w)$ can be made more apparent by describing how to measure it. We look at many identical copies of the system at a time u when the server is idle, and we measure the totaled service times

of all customers arriving in the next interval (u,t), for $t > u$; the fraction of these numbers that is in the range $(0,\ t - u + w)$ is $R(t,u,w)$.

The delay $W(t)$ is never negative, and so

$$\Pr\{W(t) \leqq w\} = 0, \qquad \text{if } w < 0. \tag{7}$$

For $w > 0$, $\Pr\{W(t) \leqq w\}$ is given by

$$\Pr\{W(t) \leqq w\} = \Pr\{K(t) - t \leqq w\} - \frac{\partial}{\partial w}\int_0^t R(t,u,w)\ \Pr\{W(u) = 0\}\ du. \tag{8}$$

The chance, $\Pr\{W(u) = 0\}$, that the server will be idle at time u satisfies the Volterra equation of the first kind, for $-t \leqq w \leqq 0$,

$$\text{average of } \max\,[0,w - K(t) + t] = \int_0^{t+w} R(t,u,w,)\ \Pr\{W(u) = 0\}\ du, \tag{9}$$

and the left side of (9) is expressible in terms of (5) as

$$E\{\max\,[0,w - K(t) + t]\} = \int_0^{t+w} \Pr\{K(t) \leqq u\}\ du,$$

where $E\{u\}$ is the expectation of u.

Once the "basic" functions (5) and (6) are known, the computation of $\Pr\{W(t) \leqq w\}$ proceeds by first solving the integral equation (9) for the chance $\Pr\{W(u) = 0\}$ that the server will be idle at time u. This probability can then be used in (8) to give $\Pr\{W(t) \leqq w\}$.

Proofs for, and extensive explanations of, these results for delay have been deferred until Section IX, while we continue by discussing results for loss operation. The proofs to be given are new, and much simpler than those of Ref. 7. They also make it easier to exhibit the physical significance of the formulas and functions arising.

VIII. FORMULAS FOR LOSS OPERATION

We shall show that $\Pr\{A(t) \leqq w\}$ can be expressed in terms of two kernels, $R(t,u)$ and $Q(t,u)$, as follows:

$$\Pr\{A(t) \leqq w\} = \Pr\{A(0) \leqq t + w\} - \int_0^t [1 - R(t + w,u)]\ dE\{S(u)\}, \tag{10}$$

where $E\{S(u)\}$ is the average number of units that arrive in $(0,u]$ and find the machine idle. It satisfies the integral equation

$$E\{S(t)\} = \Pr\{y_1 \leqq t\} + \int_0^t Q(t,u)\, dE\{S(u)\}, \qquad (11)$$

where y_1 is the epoch of the first successful arrival.

The kernel $R(t,u)$ may be interpreted as a rigorous version of

> Pr{service time of a successful unit arriving at u is $\leqq t - u$ | a successful unit has arrived at u},

and $Q(t,u)$ may be thought of as a rigorous version of

> Pr{next successful arrival after u occurs before t | a successful unit has arrived at u}.

The first term on the right in (10) is self-explanatory. Precise definitions of $R(\cdot,\cdot)$ and $Q(\cdot,\cdot)$ are given in Section XI.

To explain (10) itself, we observe that $A(0) > t + w$ implies $A(t) = A(0) - t > w$, and hence

$$\Pr\{A(t) \leqq w\} \leqq \Pr\{A(0) \leqq t + w\}.$$

The integral term in (10) is therefore a correction to the *overestimate*, $\Pr\{A(0) \leqq t + w\}$. From the interpretation of $R(t,u)$, and that of $dE\{S(u)\}$ as the "density" of successful arrivals at u, we see that (10) can be rendered in words as

$\Pr\{\text{work time left at } t \leqq w\} = \Pr\{\text{work time left at } 0 \leqq t + w\}$

$-\Pr\{\text{some successful arrival during } (0,t] \text{ stays beyond time } t + w\}$.

It is reasonable to suspect that, if the load process $K(t)$ has some weak stationarity properties, and if certain "averages" exist, then

$$\Pr\{A(t) \leqq w\}$$

has, for each $w \geqq 0$, a nonzero limit as $t \to \infty$. For $w = 0$, one expects intuitively that this limit will have the form

$$\Pr\{A(\infty) = 0\} = \frac{a}{a + b},$$

where a,b are constants which can be interpreted as follows: if we watch the process $A(t)$, we notice that periods of time during which $A(t) = 0$ alternate with those during which $A(t) > 0$; then a is the average length of a period during which $A(t) = 0$, and b is the average length of a period during which $A(t)$ exceeds 0.

These conjectures are justified in Section XII, where it is shown that, if the kernels $R(t,u)$ and $Q(t,u)$ used in (10) and (11) are functions of

$(t - u)$ only, and if

$$1 - R(x) \tag{12}$$

is integrable over $(0, \infty)$, then

$$\lim_{t \to \infty} \Pr\{A(t) \leqq w\}$$

exists and has the form

$$\frac{a + \int_0^w [1 - R(u)]\, du}{a + b},$$

with

$$b = \int_0^\infty [1 - R(u)]\, du,$$

$$a + b = \int_0^\infty [1 - Q(u)]\, du.$$

We have already pointed out that $R(t,u)$ can be interpreted as

$$\Pr\{\text{service time of a successful unit which arrives at } u \text{ is} \leqq t - u \mid \text{a successful unit arrived at } u\}. \tag{13}$$

Then $R(t,u) = R(t - u)$ states that (13) does not depend on the first and third occurrences of u therein; i.e., that service times have the same distribution no matter when they begin. Thus, the dependence of $R(t,u)$ on $(t - u)$ only is a weak sort of stationarity property, and $R(x)$ can be interpreted as the probability distribution of service times of successful units. Then

$$b = \int_0^\infty [1 - R(u)]\, du = \{\text{average service time of successful units}\},$$

since when the mean of a positive variate exists it equals the integral of the "tail" of its distribution.

In a similar way, the fact that $Q(t,u)$ is a difference kernel may be interpreted as a stationarity condition, and $Q(x)$ can be thought of as the distribution function of the intervals between successful arrivals, so that

$$a + b = \int_0^\infty [1 - Q(x)]\, dx$$

$$= \{\text{average interval between successful arrivals}\}$$

and

$$a = \text{average length of an idle period.}$$

The preceding discussion suggests that we use

$$\text{chance of loss} = \frac{b}{a + b} \tag{14}$$

as a natural measure of the probability of loss for the single telephone trunk or particle counter that we are considering. Of course, any engineer would have used (14) to describe loss, justifying it by intuitive arguments. This fact does not detract from our result, which gives some idea of the weak assumptions that are sufficient for *proving* (14).

Suppose that, instead of having only one machine, we had $N \geqq 1$ machines, and used them in a fixed serial order of preference. That is, arrivals finding the first n machines busy try the $(n + 1)$th. Such a situation arises in telephony, for instance: the "machines" are telephone trunk lines, the "arriving units" are attempts to place a call and the work or service times are the holding times of calls. Each trunk is then receiving the traffic overflowing the previous trunks in the ordering. Our theory then applies to each trunk considered by itself, and if the conditions for the validity of (14) obtain for each trunk, the chance of loss for the whole group must have the form

$$\prod_{n=1}^{N} \frac{b_n}{a_n + b_n},$$

where

$$b_n = \left\{\begin{matrix}\text{average service time of units that find the first (n} - 1)\\ \text{trunks busy, the } n\text{th idle}\end{matrix}\right\},$$

$$a_n + b_n = \left\{\begin{matrix}\text{average time interval between calls accommodated on}\\ \text{the } n\text{th trunk}\end{matrix}\right\}.$$

Formulas (10) and (11) have been essentially proved[8] by involved arguments using the integral equation (2) defining $A(t)$. We shall give a simple heuristic derivation in this section and a rigorous one in Section XI.

In order to explain the results, we first recall that the process $A(t)$ consists of alternating intervals during which $A(t)$ is first zero, then positive with slope -1, then zero again, and so on. The next arrival to find the machine idle always makes it busy again.

We are interested in expressing $\Pr\{A(t) \leqq w\}$, and so we search for

other ways of specifying the event $\{A(t) \leqq w\}$. We first consider those cases in which $A(0) = 0$; that is, the system starts empty, as in Fig. 7. Then it is not hard to see that $A(t)$ will be less than or equal to w only if the number $S(t)$ of successful arrivals during $(0,t]$ equals the number of successful arrivals in $(0,t]$ that have left the system by time $t + w$. In other words, if the machine is idle at $t = 0$, then the work $A(t)$ remaining at time t is less than or equal to w if and only if all people who arrived to find it idle in $(0,t]$ are finished with it by time $t + w$. We shall set

$$\theta(t,t + w) = \text{number of successful arrivals in } (0,t] \text{ who have left the system by time } t + w.$$

Then, if $A(0) = 0$, the events

$$\{A(t) \leqq w\} \quad \text{and} \quad \{S(t) = \theta(t,t + w)\} \tag{15}$$

are the same.

If $A(0) > 0$, the system starts busy, and the graph of $A(t)$ appears as in Fig. 8. Assume first that $A(0) > t$; then, of course, $A(t) = -t + A(0)$, because the machine has been busy since $t = 0$, and is not yet finished. If $t \geqq A(0)$, though, the machine became idle at $t = A(0)$. In the first instance, $S(t) = \theta(t,t + w) = 0$, because there have not yet been any successful arrivals, and $A(t) \leqq w$ if and only if

$$t < A(0) \leqq t + w.$$

In the second instance, the argument we used for the case $A(0) = 0$ applies, and (15) holds. We now average these cases according to their

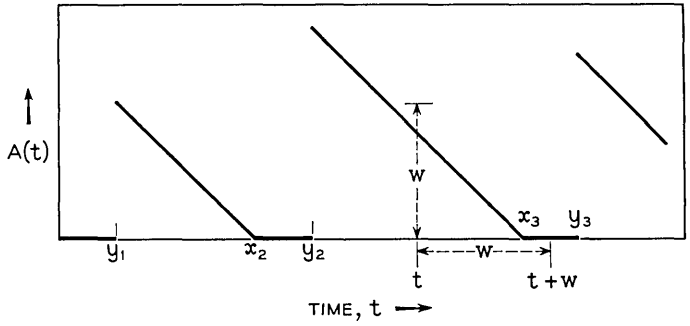

Fig. 7 — Graph of $A(t)$ vs. t when system starts empty.

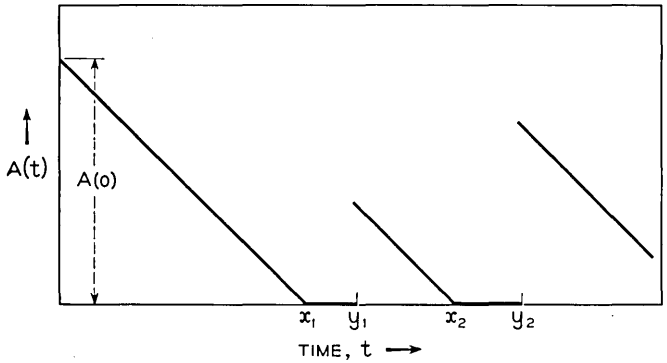

Fig. 8 — Graph of $A(t)$ vs. t when system starts busy.

probabilities; since either $S(t) - \theta(t,t+w) = 0$ or $S(t) - \theta(t, t+w) = 1$, we find

$$\Pr\{A(t) \leqq w\} = \Pr\{t < A(0) \leqq t+w \quad \text{and} \quad S(t) = \theta(t,t+w)\} + \Pr\{A(0) \leqq t \quad \text{and} \quad S(t) = \theta(t,t+w)\}. \tag{16}$$

Where $\{\cdots\}$ is an event, let $\chi\{\cdots\}$ be its *characteristic function*, defined as 1 if the event happens, and 0 otherwise. Then the first term on the right in (16) is

$$\text{average of } \big(\chi\{t < A(0) \leqq t+w\} \cdot [1 - S(t) + \theta(t,t+w)]\big),$$

and the second is

$$\text{average of } \big(\chi\{A(0) \leqq t\} \cdot [1 - S(t) + \theta(t,t+w)]\big),$$

because

$$1 - S(t) + \theta(t,t+w) = 0 \qquad \text{if} \quad S(t) \neq \theta(t,t+w).$$

Therefore

$$\Pr\{A(t) \leqq w\} = \Pr\{A(0) \leqq t+w\} - E\{S(t)\} + E\{\theta(t,t+w)\}. \tag{17}$$

Thus, we have expressed $\Pr\{A(t) \leqq w\}$ in terms of the initial distribution and the average values of $S(t)$ and $\theta(t,t+w)$.

To prove (10), it remains to express the average of $\theta(t,t+w)$ as the integral of a kernel with respect to the average of $S(t)$. To do this, let u be a point in $(0,t]$, and suppose that a successful arrival occurs at u. Such an arrival can only contribute to the average of $\theta(t,t+w)$ if it leaves before $t+w$. The proportion of such arrivals is just

$$\begin{aligned}&\Pr\{\text{service time of a successful unit which arrives at } u \text{ is}\\ &\leqq t + w - u \mid \text{a successful arrival occurred at } u\}\\ &= R(t + w,u).\end{aligned} \tag{18}$$

The "density" of successful arrivals at u is $dE\{S(u)\}$; therefore,

$$E\{\theta(t,t + w)\} = \int_0^t R(t + w,u)\, dE\{S(u)\},$$

which proves (10).

We continue with a heuristic derivation of the integral equation (11) for $E\{S(t)\}$. First we notice that the event $\{A(t) = 0\}$ can occur in two ways: either some successful arrival has occurred in $(0,t]$, or none has. If none has, then either $A(\cdot)$ started idle at 0 and is still idle at t, or else it started busy at 0, became idle at the point $A(0) < t$ and is still idle at t. If t_1 is the first arrival in $(0,\infty)$ and y_1 the first successful arrival, the chance that no successful arrivals occurred in $(0,t]$ and $A(t) = 0$ is

$$\Pr\{A(0) = 0 \quad \text{and} \quad t_1 > t\} + \Pr\{0 < A(0) \leqq t \quad \text{and} \quad y_1 > t\}.$$

Assuming some successful arrival did occur in $(0,t]$, suppose it occurred at u. Such an arrival is only relevant to the event $\{A(t) = 0\}$ if it is the last such arrival in $(0,t]$, and if the service time of the customer then arriving is at most $(t - u)$. The proportion of such "relevant" arrivals is

$$\begin{aligned}&\Pr\{\text{service time of successful arrival occurring at } u \text{ is} \leqq t - u\\ &\text{and no more customers arrive in the time interval between}\\ &\text{his departure and } t \mid \text{a successful arrival occurred at } u\}\\ &= G(t,u).\end{aligned} \tag{19}$$

As before, we now argue that the density of successful arrivals at u is $dE\{S(u)\}$ and so, using (19) and noting that $y_1 = t_1$ if $A(0) = 0$, we find

$$\Pr\{A(t) = 0\} = \Pr\{A(0) \leqq t, y_1 > t\} + \int_0^t G(t,u)\, dE\{S(u)\}. \tag{20}$$

By combining this result with (10) for $w = 0$, we obtain an integral equation for $E\{S(u)\}$:

$$E\{S(t)\} = \Pr\{y_1 \leqq t\} + \int_0^t Q(t,u)\, dE\{S(u)\}, \tag{21}$$

where the kernel $Q(t,u) = R(t,u) - G(t,u)$. By examining the interpretations (18) and (20) of $R(t,u)$ and $G(t,u)$, it can be seen that

$$Q(t,u) = \Pr\{\text{next successful arrival after } u \text{ occurs before } t \mid \text{a successful arrival occurred at } u\}.$$

Hence, $Q(t,u)$ should be a distribution function in t. Proof of this, together with discussions of (18) and (20), appears in Section XI.

IX. PROOF AND DISCUSSION OF THE RESULTS FOR DELAY

The proof and explanation of (8) and (9) depend on two simple preliminary results. The first of these is as follows: let x be any nonnegative random variable; then, for $y \geqq 0$,

$$\int_0^y \Pr\{x \leqq u\}\, du = E\{\max(0, y - x)\}. \tag{22}$$

This formula states that the area under the (cumulative) distribution of x to the left of y is just the average value of the greater of zero and $y - x$. This is easily seen from an integration by parts:

$$\int_0^y \Pr\{x \leqq u\}\, du = u \Pr\{x \leqq u\} \Big|_0^y - \int_0^y u\, d \Pr\{x \leqq u\}.$$

To begin the proof we note that the total idle time $T(t)$ represented by the term

$$T(t) = \int_0^t U[-W(u)]\, du$$

in the integral equation (1), is always nonnegative, so that

$$W(t) \geqq K(t) - t, \tag{23}$$

and also

$$\Pr\{W(t) \leqq w\} \leqq \Pr\{K(t) - t \leqq w\}.$$

Now in Fig. 9 the larger area represents the event $\{K(t) - t \leqq w\}$ and the smaller one inside it represents the event $\{W(t) \leqq w\}$. This latter event is included in the former because, if $W(t) \leqq w$, then

$$K(t) - t \leqq w,$$

by (23). The difference between the two areas represents the event

$$\{K(t) - t \leqq w < K(t) - t + T(t)\},$$

and so

$$\begin{aligned} \Pr\{W(t) \leqq w\} &= \Pr\{K(t) - t \leqq w\} \\ &\quad - \Pr\{K(t) - t \leqq w < K(t) - t + T(t)\}. \end{aligned} \tag{24}$$

We next observe that (8) is the derivative with respect to w of

$$\int_0^w \Pr\{W(t) \leqq u\}\, du = \int_0^w \Pr\{K(t) - t \leqq u\}\, du - \int_0^t R(t,u,w)\, \Pr\{W(u) = 0\}\, du, \tag{25}$$

which may be written, using (24) and taking the condition inside, as

$$E\{\max[0,w - K(t) + t - T(t)] - \max[0,w - K(t) + t]\} = -\int_0^t \Pr\{K(t) - K(u) - t + u \leqq w \quad \text{and} \quad W(u) = 0\}\, du. \tag{26}$$

Thus (8) and (9) are established if we can prove (26).

To establish (26) we need the second preliminary result, a general property of monotone continuous functions.

Lemma: If $F(t)$ is continuous and monotone increasing and $F(0) = 0$, then, for any $x \geqq 0$ and $t \geqq 0$,

$$\max[0,x - F(t)] = x - \int_0^t U[x - F(y)]\, dF(y), \tag{27}$$

where $U(y) = 1$ for $y \geqq 0$, and $U(y) = 0$ for $y < 0$.

Proof: We note that, as y increases, the integrand in (27) is unity until either $x = F(y)$ or $y = t$, whichever occurs first, and it is zero thereafter. If $x = F(y)$ occurs first, then the integral equals x; if $y = t$ occurs first,

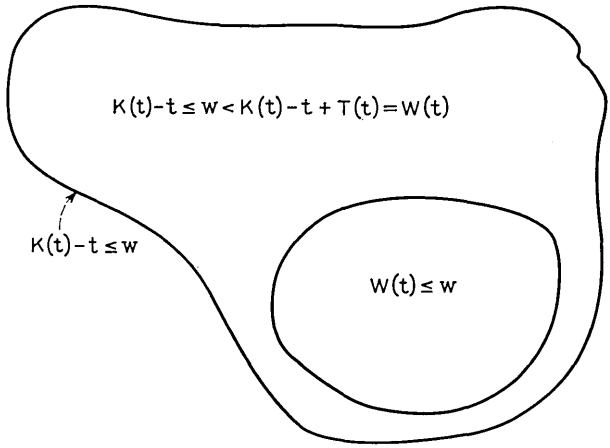

Fig. 9 — Graphical representation of events $\{K(t) - t \leqq w\}$ and $\{W(t) - t \leqq w\}$.

the integral equals $F(t)$. Hence,

$$\int_0^t U[x - F(y)]\, dF(y) = \min[x,F(t)]. \tag{28}$$

For $x \geqq 0$ it can be seen that

$$\min[x,F(t)] = x - \max[0,x - F(t)],$$

and this proves the lemma.

To use this result in proving (26) we interpret $F(t)$ in the lemma as $T(t)$, the total idle time during $(0,t)$, which is a continuous increasing function, with $T(0) = 0$. We next consider the expressions, for $w \geqq -t$,

$$A = w - K(t) + t - \int_0^t U[w - K(t) + t - T(u)]\, dT(u),$$

$$B = w - K(t) + t + T(t) - \int_0^t U[w - K(t) + t + T(t) - T(u)]\, dT(u).$$

If $w - K(t) + t \geqq 0$, then, by the lemma,

$$A - B = \max[0,w - K(t) + t - T(t)] - \max[0,w - K(t) + t]. \tag{29}$$

However, since $w - K(t) + t \geqq 0$, we see that the integrand in B is always unity, so that B equals $w - K(t) + t$, and

$$A - B = -\int_0^t U[w - K(t) + t - T(u)]\, dT(u).$$

Hence $w - K(t) + t \geqq 0$ implies

$$\max[0,w - K(t) + t - T(t)] - \max[0,w - K(t) + t] = -\int_0^t U[w - K(t) + t - T(u)]\, dT(u). \tag{30}$$

However, if $w - K(t) + t < 0$, then both sides of (30) vanish and so (30) holds generally, for any $w - K(t) + t$.

From (1) it is evident that, while $W(u)$ is zero, then

$$T(u) = u - K(u);$$

so the integral in (30) is

$$\int_0^t U[w - K(t) + t + K(u) - u] \frac{dT(u)}{du}\, du, \tag{31}$$

where $dT(u)/du$ is one if $W(u) = 0$, and is zero if not; i.e.,

$$\frac{dT(u)}{du} = U[-W(u)]. \tag{32}$$

We recall that, if x and y are random variables, then

$$\Pr\{x \leqq \lambda_1 \quad \text{and} \quad y \leqq \lambda_2\} = E\{U[\lambda_1 - x]U[\lambda_2 - y]\}; \tag{33}$$

i.e., the joint distribution is the average of the product

$$U[\lambda_1 - x] \cdot U[\lambda_2 - y].$$

If we now average (30) and bear in mind (31), (32) and (33), we obtain (26) for all $w \geqq -t$. For negative w,

$$E\{\max[0,w - K(t) + t - T(t)]\} = 0,$$

so that formula (26) takes the form, for $-t \leqq w \leqq 0$,

$$E\{\max[0,w - K(t) + t]\} = \int_0^{t+w} \Pr\{K(t) - K(u) - t + u \leqq w \quad \text{and} \quad W(u) = 0\}\, du. \tag{34}$$

This is (9), and we have completed the proof of the results stated at the beginning of this section.

It can be seen from (8) and (24), and from Fig. 9, that

$$\frac{\partial}{\partial w} \int_0^t \Pr\{K(t) - K(u) - t + u \leqq w \quad \text{and} \quad W(u) = 0\}\, du \tag{35}$$

is the correction term to the overestimate $\Pr\{K(t) - t \leqq w\}$ for $\Pr\{W(t) \leqq w\}$. [See (24).] It is the probability of the event

$$\{K(t) - t \leqq w < K(t) - t + T(t)\},$$

represented by the difference between the areas in Fig. 9. Now, the presence of the derivative $\partial/\partial w$ in (24) is explained by the fact that (24) is the derivative of (25) for $w \geqq 0$, and thus is due to our use of (22). However, it is not obvious intuitively why, in (35), the rest of the term after the $\partial/\partial w$, should be a *time integral.*

An explanation of this can be obtained from (26), which expresses the average of the random variable

$$\alpha = \max[0,w - K(t) + t - T(t)] - \max[0,w - K(t) + t].$$

It can be seen that

$$\alpha = \begin{cases} -T(t) & \text{if } w > K(t) - t + T(t) = W(t) \\ -w + K(t) - t & \text{if } K(t) - t < w < W(t) \\ 0 & \text{if } w < K(t) - t. \end{cases}$$

A graph of α as a function of w is shown in Fig. 10. The point $K(t) - t$ at which α starts downward may, of course, be negative, although it is positive in the figure.

Thus, α is a negative quantity whose magnitude is no greater than $T(t)$, the total idle time prior to t. Now, if α were in fact equal to $-T(t)$, we could write its average as

$$-E\{T(t)\} = -\int_0^t \Pr\{W(u) = 0\}\, du.$$

But α may be smaller in magnitude than $T(t)$; this fact explains the presence of the conditional probability in the integrand of

$$E\{\alpha\} = -\int_0^t \Pr\{K(t) - K(u) - t + u \leqq w \mid W(u) = 0\} \Pr\{W(u) = 0\}\, du.$$

The kernel in this expression is a probability, so it reduces the magnitude of the integrand whenever it is less than one.

X. DELAY EXAMPLE: POISSON ARRIVALS, GENERAL SERVICE TIMES

For a first example, we assume that customers arrive in a Poisson process of intensity λ, and that service times are mutually independent, with a general distribution function $B(x)$. Such a system has been treated before,[9,10,11] but few explicit formulas are known except for the

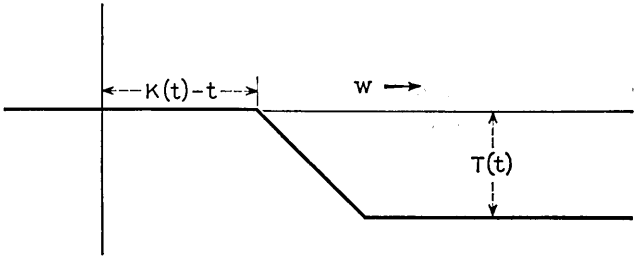

Fig. 10 — Graph of $\alpha(w)$.

exponential case

$$B(x) = \begin{cases} 1 - e^{-\mu x} & x \geqq 0, \\ 0 & x < 0, \end{cases}$$

considered by Ledermann and Reuter[11] and Bailey.[9] The author[10] has sketched a method for calculating $\Pr\{W(t) = 0 \mid W(0)\}$ for general distributions $B(x)$ of service time, but gave no explicit results. He proved that the Laplace transform of $\Pr\{W(t) = 0 \mid W(0)\}$ is given by

$$\int_0^\infty e^{-\tau t} \Pr\{W(t) = 0 \mid W(0)\}\, dt = \frac{e^{-\eta(\tau) W(0)}}{\eta(\tau)}, \qquad \mathrm{Re}(\tau) > 0, \quad (36)$$

where $\eta(\tau)$ is the unique root in the right half-plane of the equation

$$\tau - \eta + \lambda = \lambda B^*(\eta), \qquad \mathrm{Re}(\tau) > 0,$$

with

$$B^*(s) = \int_0^\infty e^{-st}\, dB(t).$$

It was also shown that any function of $\eta(\tau)$, analytic in the right half-plane, could be expanded in a Lagrange series. We shall now derive these results (by a quite different way) directly from the general integral equation (9), and then obtain some specific new results.

Since arrivals are Poisson and service times are independent with distribution $B(x)$, then the load process $K(t)$ is the compound Poisson process, and

$$E\{e^{-s[K(t)-K(u)]}\} = e^{\lambda(t-u)[B^*(s)-1]}.$$

The kernel $R(t,u,0)$ of (9) is

$$\sum_{n=0}^\infty \frac{e^{-\lambda(t-u)}\lambda^n (t-u)^n}{n!} B_n(t-u) = \Pr\{K(t) - K(u) \leqq t - u\}, \quad (37)$$

where $B_n(x)$ is the convolution of $B(x)$ with itself n times, i.e., the distribution of the sum of n service times, and $B_0(x)$ is 1 for $x \geqq 0$ and is 0 otherwise. In fact, the Poisson term in (37) is just the chance that n customers arrive in (u,t), and $B_n(t-u)$ is the chance that their combined service time is not more than $(t-u)$. Similarly, we find that

$$\int_0^t \Pr\{K(t) \leqq u\}\, du = E\{\max[0, t - K(t)]\}$$

$$= \sum_{n=0}^\infty \frac{e^{-\lambda t}(\lambda t)^n}{n!} \int_0^t \max[0, t - x - W(0)]\, dB_n(x). \quad (38)$$

Equation (9) for this example is therefore

$$\sum_{n=0}^{\infty} \frac{e^{-\lambda t}(\lambda t)^n}{n!} \int_0^t \max(0,t - x - W(0))\, dB_n(x) = \int_0^t \sum_{n=0}^{\infty} \frac{e^{-\lambda(t-u)}\lambda^n(t-u)^n}{n!} B_n(t-u)\Pr\{W(u) = 0 \mid W(0)\}\, du. \tag{39}$$

Since the right-hand side is a convolution, we take Laplace transforms. That of the kernel, (37), can be written as

$$\int_0^{\infty} e^{-\tau t} R(t)\, dt = \frac{1}{2\pi i} \int_0^{\infty} \int_{c-i\infty}^{c+i\infty} e^{-t[\tau - s + \lambda - \lambda B^*]} \frac{ds}{s}\, dt, \qquad c > 0$$
$$= \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} \frac{\frac{ds}{s}}{\tau - s + \lambda - \lambda B^*}, \tag{40}$$

since the order of integration can be interchanged.

Let S_R be the semicircle that is the right-hand half of the circle $|s - c| = R$. It can be seen that on this semicircle

$$\frac{s^{-1}}{\tau - s + \lambda - \lambda B^*} = O(R^{-2}),$$

so that

$$\lim_{R\to\infty} \frac{1}{2\pi i} \int_{S_R} \frac{s^{-1}\, ds}{\tau - s + \lambda - \lambda B^*} = 0$$

It has been shown[10] that, for $\mathrm{Re}(\tau) > 0$, the function $\tau - s + \lambda - \lambda B^*(s)$ has a unique zero, $\eta(\tau)$, in the right half-plane. Hence,

$$\frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} \frac{s^{-1}\, ds}{\tau - s + \lambda - \lambda B^*} = \frac{1}{\eta(\tau)} [\text{residue of } (\tau - s + \lambda - \lambda B^*)^{-1} \text{ at } s = \eta(\tau)], \tag{41}$$

when $c < \mathrm{Re}\ \eta(\tau)$.

The Laplace transform of $E\{\max[0,t - K(t)]\}$ can be written in the form

$$\frac{1}{2\pi i} \int_0^{\infty} \int_{c-i\infty}^{c+i\infty} e^{-t(\tau - s + \lambda)} \sum_{n=0}^{\infty} \frac{(\lambda t)^n}{n!} (B^*)^n \frac{e^{-sW(0)}}{s^2}\, ds\, dt. \tag{42}$$

Formula (42) simplifies to

$$\frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} \frac{s^{-2} e^{-sW(0)}\, ds}{\tau - s + \lambda - \lambda B^*},$$

which, by arguments like those already used for (40), can be shown to equal

$$[\eta(\tau)]^{-2}e^{-\eta(\tau)W(0)}[\text{residue of } (\tau - s + \lambda - \lambda B^*)^{-1} \text{ at } s = \eta(\tau)]. \quad (43)$$

It follows from (39), (40) and (43) that the Laplace transform of $\Pr\{W(t) = 0 \mid W(0)\}$ is given by

$$\int_0^\infty e^{-\tau t} \Pr\{W(t) = 0 \mid W(0)\}\, dt = \frac{e^{-\eta(\tau)W(0)}}{\eta(\tau)}, \quad (44)$$

where $\eta(\tau)$ satisfies $\tau - \eta + \lambda = \lambda B^*(\eta)$.

As shown,[10] any function F of $\eta(\tau)$, analytic in $\text{Re}(\tau) > 0$, may be expanded in the Lagrange series

$$F[\eta(\tau)] = F(\tau + \lambda) + \sum_{n=1}^{\infty} \frac{(-\lambda)^n}{n!} \frac{d^{n-1}}{ds^{n-1}} \left[\frac{dF}{ds} (B^*)^n\right]_{s=\tau+\lambda}.$$

We can use this expansion to invert the transform given by (44). In some cases, this inversion gives an explicit expression for

$$\Pr\{W(t) = 0 \mid W(0)\}$$

in terms of the kernel and forcing function of the integral equation (39). Setting $F(x) = x^{-1}e^{-xW(0)}$ in the expansion and inverting the resulting transform, we find

$$\Pr\{W(t) = 0 \mid W(0)\} = e^{-\lambda t}U[t - W(0)] + \sum_{n=1}^{\infty} \frac{e^{-\lambda t}\lambda^n t^{n-1}}{n!} \left\{\int_0^t B_n[u - W(0)]\, du + W(0)B_n[t - W(0)]\right\}. \quad (45)$$

By rearranging terms in (45), comparing with (37) and (38), and recalling that $W(0) = K(0)$ by convention, we can put (45) into the form

$$\Pr\{W(t) = 0 \mid W(0)\} = t^{-1}E\{\max[0,t - K(t)]\} + \frac{W(0)}{t} \Pr\{K(t) \leqq t\}. \quad (46)$$

Note that $E\{\max[0,t - K(t)]\}$ is the left-hand side (the forcing function) of the integral equation (39) and that, when $W(0) = 0$, (46) gives an explicit representation of the solution of (39) *in terms of the forcing function alone.* The intuitive meaning of (46) can be expressed as follows: the chance that the system is empty, conditional on the initial load $W(0)$ is equal to

average of greater of 0 and $1 - t^{-1}K(t)$

$$+ \frac{W(0)}{t} \text{ (chance that } W(0) \text{ plus load arriving in } (0,t) \text{ is at most } t).$$

It is easy to obtain new specific formulas from (46). For example, suppose that service times have the fixed length b. In this case of "constant" service times, for $t > W(0)$

$$E\{\max[0,t - K(t)]\} = \sum_{nb \leqq t-W(0)} \frac{e^{-\lambda t}(\lambda t)^n}{n!} [t - nb - W(0)],$$

$$\Pr\{K(t) \leqq t\} = \sum_{nb \leqq t-W(0)} \frac{e^{-\lambda t}(\lambda t)^n}{n!},$$

and hence

$$\Pr\{W(t) = 0 \mid W(0)\} = 1 - \lambda b - P(T,\lambda t) + \lambda b P(T - 1,\lambda t),$$

where $bT = t - W(0)$ and

$$P(c,a) = \sum_{n \geqq c} \frac{e^{-a}a^n}{n!}$$

is the cumulative term (the "tail") of the Poisson distribution with mean a. This formula for $\Pr\{W(t) = 0 \mid W(0)\}$ for constant service times was used to compute the curve of Fig. 1.

By rewriting (46) in terms of inversion integrals, we obtain another representation of $\Pr\{W(t) = 0 \mid W(0)\}$. This one is more useful because, from it, we can find explicit formulas for new cases (by evaluating the inversion integrals). From (42) we see that

$$E\{\max [0,t - K(t)]\} = \frac{e^{-\lambda t}}{2\pi i} \int_{c-i\infty}^{c+i\infty} e^{st - sW(0) + \lambda t B^*(s)} \frac{ds}{s^2},$$

and (45) yields

$$\Pr\{K(t) \leqq t\} = \frac{e^{-\lambda t}}{2\pi i} \int_{c-i\infty}^{c+i\infty} e^{st - sW(0) + \lambda t B^*(s)} \frac{ds}{s}.$$

Therefore, from (46),

$$\Pr\{W(t) = 0 \mid W(0)\} = \frac{e^{-\lambda t}}{2\pi i t} \int_{c-i\infty}^{c+i\infty} e^{st - sW(0) + \lambda t B^*(s)} \frac{sW(0) + 1}{s^2} ds. \quad (47)$$

As we shall see, this result is useful because it is often easier to evaluate the complex integral than to sum the series (37) or (38).

For example, if service times have the negative exponential distribution with mean u, then it has essentially been shown[9,11] that

$$\frac{d}{dt}\Pr\{W(t) = 0 \mid W(0) = 0\} = -\left(\frac{\lambda}{\mu}\right)^{\frac{1}{2}} e^{-(\mu+\lambda)t}\frac{I_1[2(\mu\lambda)^{\frac{1}{2}}t]}{t}. \dagger$$

Hence

$$\Pr\{W(t) = 0 \mid W(0) = 0\} = 1 - \rho\int_0^{\mu t} \rho^{-\frac{1}{2}} e^{-(1+\rho)x} I_1(2\rho^{\frac{1}{2}}x)\frac{dx}{x}, \quad (48)$$

where $\rho = \lambda/\mu$ = traffic intensity, and μt is the time measured in mean holding times. This result can be obtained directly from (44) by solving a quadratic equation, and it can be put into another form by using (47). Thus,

$$\Pr\{W(t) = 0 \mid W(0) = 0\} = \frac{e^{-(\lambda+\mu)t}}{2\pi i t}\int_{c+\mu-i\infty}^{c+\mu-i\infty} e^{ut+\lambda\mu t/u}(u-\mu)^2\,du.$$

Writing the integrand as $e^{ut}(u-\mu)^2(e^{\lambda\mu t/u} - 1 + 1)$ and using Ref. 12, p. 244, no. 31, we find

$$\Pr\{W(t) = 0 \mid W(0) = 0\} = e^{-\lambda t}\left[1 + \int_0^{t/(4\lambda\mu)} e^{-y^2/(4\lambda\mu t)}\left(1 - \frac{y^2}{4\lambda\mu t^2}\right) I_1(y)\,dy\right].$$

For another example, suppose that service times have the "gamma" probability density

$$\frac{t^{-\frac{1}{2}}\mu^{\frac{1}{2}}e^{-\mu t}}{\Gamma(\frac{1}{2})}$$

whose Laplace transform is $[\mu/(\mu+s)]^{\frac{1}{2}}$. Then (47), with a change of variable, gives

$$\Pr\{W(t) = 0 \mid W(0) = 0\} = \frac{e^{-(\lambda+\mu)t}}{2\pi i t}\int_{c-i\infty}^{c+i\infty} e^{ut+\lambda t(\mu/u)^{\frac{1}{2}}}(u-\mu)^2\,du,$$

and

$$\Pr\{W(t) = 0 \mid W(0) = 0\} = t^{-1}e^{-\lambda t}\int_0^t e^{-\mu u}G(u)(1+\mu u)\,du\},$$

---

† This formula is a simplification of that of Ref. 9 obtained by using standard Bessel function relations; $I_1(x)$ is the Bessel function of imaginary argument, of order one.

where

$$G(u) = u^{-5/2}\alpha^{-3/4}\int_0^\infty x^{3/2}e^{-x^2/\alpha u^2}I_1(x)\,dx$$

and $\alpha = 4\lambda\mu^{\frac{1}{2}}$.

XI. PRECISE DERIVATION OF LOSS RESULTS

The proof of (10) given in Section VIII is rigorous up to and including (17), so it suffices to give a precise construction of the kernels $R(t,u)$, $G(t,u)$ and $Q(t,u)$. Let x_n be the nth epoch at which $A(t)$ becomes equal to 0, and y_n be the arrival time of the nth successful unit. It is readily seen that

$$\begin{aligned} E\{\theta(t,t+w)\} &= \sum_{n=1}^{\infty} \Pr\{y_n \leqq t \quad \text{and} \quad x_{n+1} \leqq t+w\} \\ &= \sum_{n=1}^{\infty} \int_0^t \Pr\{x_{n+1} \leqq t+w \,|\, y_n = u\}\, d\Pr\{y_n \leqq u\}. \end{aligned}$$

Similarly,

$$E\{S(t)\} = \sum_{n=1}^{\infty} \Pr\{y_n \leqq t\}. \tag{49}$$

Thus, each $\Pr\{y_n \leqq \cdot\}$ measure, $n = 1, 2, \cdots$, is absolutely continuous with respect to (49). We can therefore express $E\{\theta(t,t+w)\}$ as the integral of a kernel $R(t+w,u)$ against $E\{S(\cdot)\}$ measure, as in

$$E\{\theta(t,t+w)\} = \int_0^t R(t+w,u)\, dE\{S(u)\},$$

where the kernel is defined in terms of the indicated Radon-Nikodym derivatives by

$$R(y,u) = \sum_{n=1}^{\infty} \Pr\{x_{n+1} \leqq y \,|\, y_n = u\} \frac{d\Pr\{y_n \leqq u\}}{dE\{S(u)\}}.$$

In a like manner, we can write

$$\Pr\{A(t) = 0 \text{ and some successful arrival occurred in } (0,t]\} = \sum_{n=1}^{\infty} \Pr\{x_{n+1} \leqq t < y_{n+1}\}. \tag{50}$$

Introducing the kernel

$$G(y,u) = \sum_{n=1}^{\infty} \Pr\{x_{n+1} \leqq y < y_{n+1} \,|\, y_n = u\} \frac{d\Pr\{y_n \leqq u\}}{dE\{S(u)\}},$$

we can render (50) as

$$\int_0^t G(t,u)\, dE\{S(u)\}\,.$$

The kernel $Q(t,u)$, finally, can be defined as $R(t,u) - G(t,u)$ or as

$$\sum_{n=1}^{\infty} \Pr\{y_{n+1} \leqq t \,|\, y_n = u\} \frac{d \Pr\{y_n \leqq u\}}{dE\{S(u)\}}\,.$$

The sense in which $Q(\cdot,u)$ is a "distribution function" is given by the following result: for almost all u with respect to $E\{S(\cdot)\}$ measure, $Q(\cdot,u)$ is a distribution function. We show first that the derivatives

$$\varphi_n(u) = \frac{d \Pr\{y_n \leqq y\}}{dE\{S(u)\}}$$

have the properties

$$\sum_{n=1}^{\infty} \varphi_n(u) = 1, \tag{51}$$

$$0 \leqq \varphi_n(u) \leqq 1, \tag{52}$$

almost everywhere in $E\{S(\cdot)\}$. Now (51) is true by definition. Suppose (52) failed on a set B of positive measure; then either $E\{S(\cdot)\}$ is not a positive measure, or else

$$\Pr\{y_n \epsilon B\} > \int_B dE\{S(u)\}\,,$$

both of which are impossible. It is readily seen, by an elementary decomposition of the events, that, for each $n \geqq 1$,

$$\Pr\{x_{n+1} \leqq t \,|\, y_n = u\} - \Pr\{x_{n+1} \leqq t < y_{n+1} \,|\, y_n = u\} = \Pr\{y_{n+1} \leqq t \,|\, y_n = u\}.$$

Except on a set C_n of $\Pr\{y_n \leqq \cdot\}$ measure zero, this is a distribution function in t, and the result follows from

$$\int_C dE\{S(u)\} = 0\,, \qquad \text{if} \quad C = \bigcup_{n=1}^{\infty} Cn.$$

XII. A LIMIT THEOREM

We shall prove that dependence of the kernels $R(t,u)$ and $Q(t,u)$ on $(t - u)$ only and existence of the "mean service time of successful arrivals" are sufficient to guarantee that

$$\Pr\{A(t) \leqq w\}$$

approaches a limit as $t \to \infty$. It is natural to study cases in which only difference kernels occur, because of the Volterra equation derived in Section IX. The result to be proved requires no restriction on the "arrival rate" of units — an upper bound is unnecessary because of the loss operation; i.e., if the arrival rate increases, the rate at which successful units leave the system can only increase to a limit.

Theorem: If the kernels $R(t,u)$ and $Q(t,u)$ only depend on $(t - u)$, if the average

$$b = \int_0^\infty [1 - R(x)]\, dx = \text{(average service time of a successful unit)}$$

exists, and if $Q(\cdot)$ is not a lattice distribution, then

$$\lim_{t\to\infty} \Pr\{A(t) \leqq w\} = \frac{a + \int_0^w [1 - R(u)]\, du}{a + b}, \tag{53}$$

where

$$(a + b) = \int_0^\infty [1 - Q(x)]\, dx,$$

the limit being one if $(a + b) = \infty$.

Proof: By the remarks at the end of Section XI, both $R(x)$ and $Q(x)$ may be taken to be the distribution functions of positive variates. Equation (11) becomes a renewal equation, and $E\{S(u)\}$ is essentially the renewal function $H(\cdot)$ of Smith.[13] The integrand of (10) is

$$1 - R(t + w - u),$$

and is nonincreasing and integrable. Also, $Q(\cdot)$ is not of lattice type. Hence, by Smith's Theorem 1:

$$\lim_{t\to\infty} \int_0^t [1 - R(t + w - u)]\, dE\{S(u)\} = \frac{\int_w^\infty [1 - R(u)]\, du}{\int_0^\infty x\, dQ(x)}, \tag{54}$$

which gives (53) upon rearrangement if (54) is taken to be zero if the mean of $Q(x)$ does not exist.

XIII. ACKNOWLEDGMENTS

The author is indebted to E. B. Ferrell for suggesting many improvements in exposition, and to J. Riordan and W. O. Turner for reading the draft.

REFERENCES

1. Reich, E., On the Integrodifferential Equations of Takács, I., Ann. Math. Stat., **29**, 1958, p. 563.
2. Jensen, A., An Elucidation of Erlang's Statistical Works Through the Theory of Stochastic Processes, in Brockmeyer, E., et al., *The Life and Works of A. K. Erlang*, Copenhagen Telephone Co., Copenhagen, 1948, p. 23.
3. Smith, W. L., Regenerative Stochastic Processes, Proc. Royal Soc. (London) A, **232**, 1955, p. 6.
4. Takács, L., On the Generalization of Erlang's Formula, Acta Math. Acad. Sci. Hung., **7**, 1956, p. 419.
5. Takács, L., On a Coincidence Problem Concerning Telephone Traffic, Acta Math. Acad. Sci. Hung., **9**, 1958, p. 45.
6. Takács, L., A Telefon-forgalom Elméletének Néhány Valószinüségszámitási Kérdéséröl, A. Magyar Tud. Akad. (Math. and Phys.), **8**, 1958, p. 151.
7. Beneš, V. E., Combinatory Methods and Stochastic Kolmogorov Equations in the Theory of Queues with One Server, to be published.
8. Beneš, V. E., General Stochastic Processes in the Theories of Counters and Telephone Traffic, to be published.
9. Bailey, N. T. J., A Continuous Time Treatment of a Simple Queue Using Generating Functions, J. Royal Stat. Soc. B, **15**, 1954, p. 288.
10. Beneš, V. E., On Queues with Poisson Arrivals, Ann. Math. Stat., **28**, 1957, p. 670.
11. Ledermann, W. and Reuter, G. E. H., Spectral Theory for the Differential Equations of Simple Birth and Death Processes, Phil. Trans. Royal Soc. (London) A, **246**, 1954, p. 321.
12. Erdelyi, A., et al., *Tables of Integral Transforms*, McGraw-Hill Book Co., New York, 1954.
13. Smith, W. L., Asymptotic Renewal Theorems, Proc. Royal Soc. (Edinburgh) A, **64**, 1954, p. 9.

Round Waveguide with Double Lining

By HANS-GEORG UNGER

(Manuscript received September 22, 1959)

Doubly lining the walls of round waveguide with a base layer of dissipative material and a top layer of low loss material provides mode filtering for TE_{01} transmission and reduces TE_{01} loss in bends. The effects of thin layers are calculated as perturbations of the empty waveguide characteristics. For best performance, the dissipative layer should have low permittivity and high loss factor. The layers should be only a few mils thick.

I. INTRODUCTION

Transmission of the TE_{01} wave in round waveguide is degraded by manufacturing and laying imperfections.[1] To reduce the effects of mode conversion at manufacturing imperfections, mode filters are required. To reduce the effects of laying curvature, the phase constant of the TM_{11} wave must be made different from the TE_{01} phase constant.

Instead of using mode filters, lining the waveguide wall with a thin layer of dissipative material has been suggested.[2,3] Lining the waveguide with a low-loss material reduces the bending effects.[4]

A waveguide with a double lining was proposed by S. E. Miller to solve both problems. A base layer of dissipative material mainly introduces mode filtering, and a top layer of low-loss material changes the TM_{11} phase.

II. PROPAGATION CHARACTERISTICS

For thin dielectric layers, which fill only a small part of the total cross section, the normal modes may be considered perturbed normal modes of the empty waveguide with ideally conducting walls.

A cavity resonating in a mode with field vectors $\mathbf{E}_0$ and $\mathbf{H}_0$ at frequency ω will change its resonance when a small body V_1 of relative permittivity ϵ is introduced:[5]

$$-\frac{\Delta\omega}{\omega} = \frac{\epsilon_0 \int_{V_1} (\epsilon - 1)\mathbf{E}_1\mathbf{E}_0^* \, dV}{\epsilon_0 \int_V \mathbf{E}_0\mathbf{E}_0^* \, dV + \mu_0 \int_V \mathbf{H}_0\mathbf{H}_0^* \, dV}, \tag{1}$$

where ϵ_0 and μ_0 are the permittivity and permeability of the unperturbed cavity, $\mathbf{E}_1$ is the resulting field vector within the volume V_1 and V is the total volume of the cavity. The asterisk denotes a conjugate complex value.

If the cavity is a section of a cylindrical waveguide and if V_1 is also cylindrical, so that ϵ is independent of the axial distance, then the change in resonance frequency (1) of the cavity may be related to a change in the propagation constant γ of the waveguide:

$$\frac{\Delta\gamma}{\gamma} = -\frac{v}{u}\frac{\Delta\omega}{\omega}, \tag{2}$$

where v and u are the phase and group velocities of the unperturbed waveguide.

Although the internal field $\mathbf{E}_1$ is unknown, it is often possible to determine it from elementary boundary conditions. For example, when $\mathbf{E}_0$ is perpendicular to the boundary of V_1,

$$\mathbf{E}_1 = \frac{1}{\epsilon}\mathbf{E}_0 \tag{3}$$

or when $\mathbf{E}_0$ is parallel to the boundary of V_1,

$$\mathbf{E}_1 = \mathbf{E}_0. \tag{4}$$

For the circular waveguide with a thin dielectric layer of permittivity $\epsilon(r)$ adjacent to the wall, either (3) or (4) will determine $\mathbf{E}_1$ from $\mathbf{E}_0$. Substituting the normal mode fields of circular waveguide for $\mathbf{E}_0$ and $\mathbf{H}_0$, these expressions for the change in propagation constant are obtained:

TM_{nm} waves:
$$\frac{\Delta\gamma}{\gamma_{nm}} = \frac{1}{a}\int_0^a \frac{\epsilon - 1}{\epsilon}\,dr, \tag{5}$$

TE_{nm} waves:
$$\frac{\Delta\gamma}{\gamma_{nm}} = \frac{1}{1 - \nu_{nm}^2}\frac{n^2}{k_{nm}^2 - n^2}\frac{1}{a}\int_0^a \frac{\epsilon - 1}{\epsilon}\,dr, \tag{6}$$

$n \neq 0$

TE_{0m} waves:
$$\frac{\Delta\gamma}{\gamma_{0m}} = \frac{1}{1 - \nu_{0m}^2}\frac{k_{0m}^2}{a^3}\int_0^a (\epsilon - 1)(a - r)^2\,dr, \tag{7}$$

where a is the radius of the waveguide, k_{nm} is the mth root of $J_n(x) = 0$ for TM_{nm} waves and the mth root of $J_n'(x) = 0$ for TE_{nm} waves and $\nu_{nm} = \omega_{cnm}/\omega$ with cutoff frequency ω_{cnm}. Complex permittivities will cause a complex $\Delta\gamma$ corresponding to a change in phase and attenuation constant.

Equations (5), (6) and (7) hold for waveguides with ideally conducting walls. For walls of finite conductivity the modes of the plain pipe suffer wall current losses. These losses are changed by the presence of a dielectric layer. This change, however, is of higher order in the thickness of the layer and can be neglected against the attentuation change caused by losses in the lining from (5) and (6), for most of the modes. For circular electric waves, however, the loss contribution from (7) is of higher order in layer thickness. Then, the change in wall current loss has to be taken into account.

Wall currents of circular electric waves are given by the axial magnetic field at the wall, H_z. The wall current losses are proportional to the square of the wall current amplitudes. Therefore the change in wall current losses $\Delta\alpha$ from its unperturbed value α_0 is:

$$\frac{\Delta\alpha}{\alpha_0} = 2R_e\left(\frac{\Delta H_z}{H_{z0}}\right). \tag{8}$$

From Maxwell's equations for circular electric waves:

$$\frac{\partial H_z}{\partial r} = -j\omega\epsilon\,\epsilon_0\mathbf{E}_\varphi\,. \tag{9}$$

Equation (9) can be integrated over a thin dielectric layer adjacent to the walls by using the circumferential electric field $\mathbf{E}_{\varphi 0}$ of the unperturbed mode:

$$\frac{\Delta H_z}{H_{z0}} = \omega^2\mu_0\epsilon_0\int_0^a(\epsilon - 1)(a - r)\,dr. \tag{10}$$

Then, from (8) and (10)

$$\frac{\Delta\alpha}{\alpha_0} = 2\omega^2\mu_0\epsilon_0\int_0^a(\epsilon' - 1)(a - r)\,dr, \tag{11}$$

with ϵ' from $\epsilon = \epsilon' - j\epsilon''$.

For a double lining, according to Fig. 1, the following expressions are obtained for the change in phase constant:

$$\mathrm{TM}_{nm}: \quad \frac{\Delta\beta}{\beta_{nm}} = \left(1 - \frac{\epsilon_1'}{|\epsilon_1|^2}\right)\delta_1 + \left(1 - \frac{\epsilon_2'}{|\epsilon_2|^2}\right)\delta_2\,, \tag{12}$$

$$\mathrm{TE}_{nm}: \quad \frac{\Delta\beta}{\beta_{nm}} = \frac{1}{1 - \nu_{nm}^2}\frac{n^2}{k_{nm}^2 - n^2}\left[\left(1 - \frac{\epsilon_1'}{|\epsilon_1|^2}\right)\delta_1 + \left(1 - \frac{\epsilon_2'}{|\epsilon_2|^2}\right)\delta_2\right], \tag{13}$$

$n \neq 0$

$$\mathrm{TE}_{0m}: \quad \frac{\Delta\beta}{\beta_{0m}} = \frac{k_{0m}^2}{3(1 - \nu_{0m}^2)}[(\epsilon_1' - 1)\delta_1^3 + (\epsilon_2' - 1)(\delta^3 - \delta_1^3)]; \tag{14}$$

and for the change in attenuation constant:

$$\mathrm{TM}_{nm}: \quad \frac{\Delta\alpha}{\beta_{nm}} = \frac{\epsilon_1''}{|\epsilon_1|^2}\delta_1 + \frac{\epsilon_2''}{|\epsilon_2|^2}\delta_2, \tag{15}$$

$$\mathrm{TE}_{nm}: \quad \frac{\Delta\alpha}{\beta_{nm}} = \frac{1}{1-\nu_{nm}^2}\frac{n^2}{k_{nm}^2 - n^2}\left(\frac{\epsilon_1''}{|\epsilon_1|^2}\delta_1 + \frac{\epsilon_2''}{|\epsilon_2|^2}\delta_2\right), \tag{16}$$

$n \neq 0$

$$\mathrm{TE}_{0m}: \quad \frac{\Delta\alpha}{\beta_{0m}} = \frac{k_{0m}^2}{3(1-\nu_{0m}^2)}[\epsilon_1''\delta_1^3 + \epsilon_2''(\delta^3 - \delta_1^3)] + \frac{\alpha_{0m}}{\beta_{0m}}\frac{k_{0m}^2}{\nu_{0m}^2}[(\epsilon_1' - 1)\delta_1^2 + (\epsilon_2' - 1)(\delta^2 - \delta_1^2)], \tag{17}$$

where $\delta = d/a$ is the relative thickness of a lining. Index 1 refers to the base layer and index 2 to the top layer.

Note that the change in propagation constants is of first order in δ for all TM_{nm} waves and for the TE_{nm} waves with $n \neq 0$. It is of third order only for circular electric waves. The second-order term in the expression for added TE_{0m} loss, representing increase in wall current loss, is only of significance for a single low-loss lining.

III. DESIGN OF DOUBLE LINING

It can easily be seen that the present placement of dissipative and low loss material is the best one. The present design objectives are: the change in TM_{11} phase should be as large as possible and the undesired mode loss should be as much as possible, while the added TE_{01} loss should remain small. For unwanted modes, it does not matter where the dissipative material is placed in the lining, since the integrand in (5) and (6)

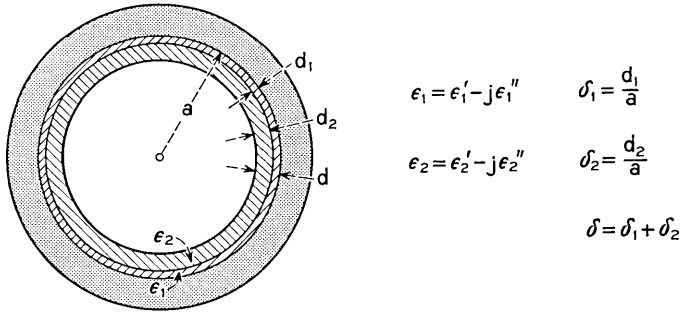

Fig. 1 — Round waveguide with double lining.

depends only on the permittivity. A dissipative material of certain thickness will cause the same perturbation no matter what distance from the wall it has. The relation is quite different for circular electric waves, since in (7) the integrand contains the square of the distance. In order to keep the perturbation small the dissipative material should be placed as close to the wall as possible, so that the dissipative material is being moved into the region of smallest electric field of the circular electric wave.

The same rules apply for dielectric constant and loss factor of the lossy material, as had previously been found for the single lossy lining.[3] The dielectric constant should be low and the loss factor high.

The selection of a suitable double lining is best demonstrated by a numerical example: $\epsilon_1' = 3$ and $\epsilon_1'' = 1.5$ is as close as present materials can be hoped to approach the above design rules for the dissipative layer;[3] $\epsilon_2' = 2.5$ and $\epsilon_2'' = 2.5 \times 10^{-3}$ are electrical properties of common low-loss materials. With $\Delta\alpha/\alpha_{01} = 0.2$, the added TE_{01} loss in the lined pipe is limited to 20 per cent of the TE_{01} loss in the plain pipe. Equation (17) is now reduced to a relation between δ_1 and δ_2. For a 2 inch I.D. copper pipe operated at 55.5 kmc, this relation is plotted in Fig. 2. On the left-hand border a dissipative layer of $\delta_1 = 0.00136$ alone adds 20 per cent to the TE_{01} loss. On the right-hand side it is a single low loss layer of $\delta_2 = 0.0090$.

To decide on a suitable combination of δ_1 and δ_2 two characteristic

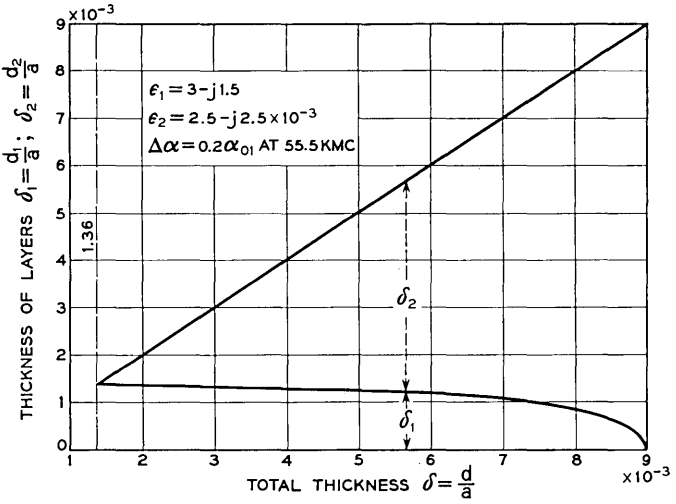

Fig. 2 — Double lining of 2-inch I.D. copper pipe.

quantities of the lined waveguide have been plotted over the same abscissa (Fig. 3). First, there is the added TE_{12} loss according to (16). The TE_{12} mode is most seriously coupled to TE_{01} at all kinds of manufacturing imperfections. To reduce the degrading effects on TE_{01} transmission of such imperfections it is most important to absorb TE_{12} power. Added TE_{12} loss is therefore a good measure for the mode filtering ability of the waveguide.

Secondly, a radius of curvature of a continuous bend is plotted that adds another 20 per cent of α_{01} to the TE_{01} loss. Such a radius characterizes laying imperfections that might be tolerated without excessive TE_{01} loss. The smaller this radius, the more freedom in laying is allowed.

The added TE_{12} loss is highest for a single lining of dissipative material, on the left-hand border. But when a low-loss layer is added, $\Delta\alpha$ of

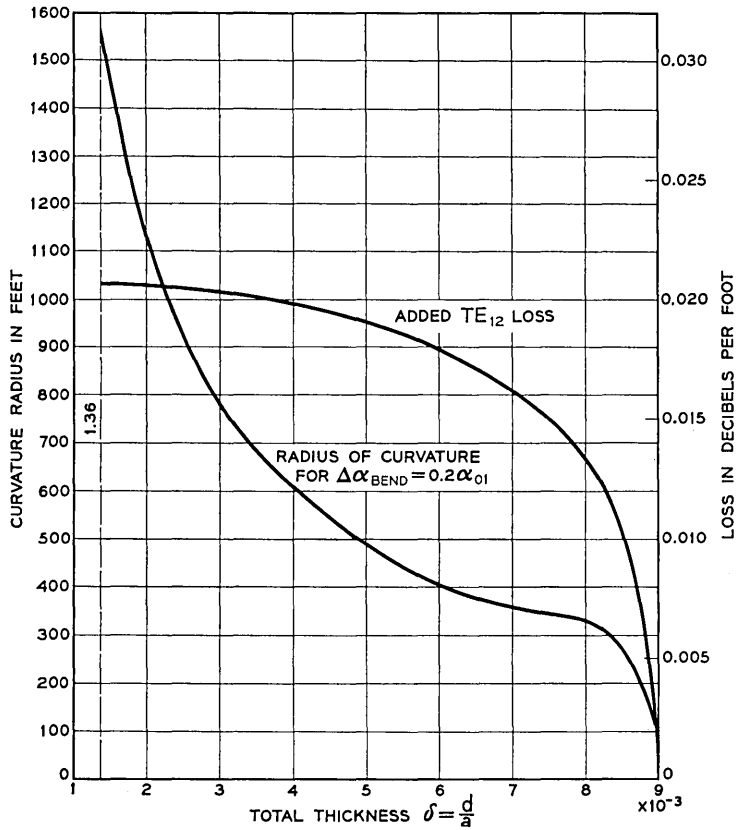

Fig. 3 — Mode filtering and bending in copper pipe with double lining of Fig. 2.

TE_{12} does not change very much at first. Even at $\delta = 0.005$, it decreases only by 9 per cent. On the other hand, the tolerable radius of curvature decreases quite rapidly from its high value for a single dissipative layer. At $\delta = 0.005$, R falls to less than one-third of its highest value. The curve is more level, however, for a substantial thickness of the second, low loss layer.

A good choice for a waveguide that has efficient unwanted mode absorption and allows much freedom in bending is $\delta = 0.005$ or, from Fig. 2,

$$\frac{\delta_2}{\delta_1} = 3.$$

The low-loss layer should be three times heavier than the dissipative layer.

IV. CONCLUSIONS

Applying two dielectric layers to the internal waveguide surface, a base layer of dissipative material and a top layer of low-loss material, is a useful modification for circular electric wave transmission. The lining changes attenuation and phase constant of circular electric waves very little. On the other hand, the dissipative layer effectively increases unwanted mode loss and the low-loss layer shifts their phase constant, in particular that of TM_{11}.

Thus, in combining the characteristics of a dissipative lining and a low-loss lining, this structure provides mode filtering and reduces TM_{11} conversion in bends. The mode filtering characteristics are nearly equivalent to those of a pipe with a single lossy lining. The additional low-loss lining gives much more freedom in bending.

REFERENCES

1. Warters, W. D. and Rowe, H. E., Transmission Deviations in Waveguide Due to Mode Conversion: Theory and Experiment, Conv. on Long-Distance Transmission by Waveguide, London, 1959; Proc. I.E.E., Suppl. 13.
2. Malin, V. V., The Influence of a Semiconductive Film on the Attenuation in a Waveguide of Circular Cross Section, Rad. Tech. i Elek., **1**, January 1956, p. 34.
3. Unger, H., Round Waveguide with Lossy Lining, Proc. of Symposium on Millimeter Waves, New York, 1959.
4. Unger, H., Circular Electric Wave Transmission in a Dielectric-Coated Waveguide, B.S.T.J., **36**, September 1957, p. 1253.
5. Müller, J., Untersuchungen über elektromagnetische Hohlräume. Hochfreq. Technik, **54**, 1939, p. 157.

Germanium and Silicon Liquidus Curves

By C. D. THURMOND and M. KOWALCHIK

(Manuscript received October 21, 1959)

New measurements are reported on the solubility of germanium in liquid gallium, thallium, tin, arsenic, bismuth, cadmium and zinc, and the solubility of silicon in liquid indium, tin, lead, antimony, bismuth and zinc. The measurements of other workers are reviewed, including those of the solubility of germanium and silicon in liquid copper, silver, gold and aluminum; of germanium in liquid indium, lead and antimony; and of silicon in liquid arsenic and nickel. All but two of the liquidus curves can be described within experimental error by a two-constant equation. The form of this equation suggests that the liquid solutions exhibit certain simple thermodynamic properties, and some evidence is cited indicating that the constants of this equation can be used to estimate the excess free energy of the solutions. Figures for the complete liquidus curves of these binary systems (T–x and log x–$1/T$) have not been included in this paper, but sets of these figures can be supplied upon request.

I. INTRODUCTION

Germanium and silicon have become very important elements in the last ten years. Great strides have been made in the exploitation of the electrical properties of these semiconductor elements, and many useful solid state electronic devices that employ these two elements are manufactured in increasing numbers every year. The electrical properties of pure germanium and silicon are of academic interest only, however. It is the modification of these properties by small amounts of added impurities that has made these elements technologically important.

A knowledge of the chemical reactivity of electrically active impurities with germanium or silicon is necessary in order to be able to control the impurity concentrations. Much of the chemistry of such reactions can be described in a simple way by the use of phase diagrams. These diagrams give the chemical compositions of phases at equilibrium at various temperatures. Of particular interest are the liquidus and solidus curves of germanium and silicon.

The liquidus curve gives the solubility of germanium or silicon in another element as a function of temperature. Each point on the curve corresponds to the temperature and composition at which a liquid phase is saturated with germanium or silicon respectively. The solidus curve gives the concentration of the other component which will be dissolved in germanium or silicon at various temperatures when the solid phase is in equilibrium with a saturated liquid phase.

In this paper, the liquidus curves of germanium and silicon will be discussed; the solidus curves are discussed in an accompanying paper.[1]

The liquidus curves formed by the dissolution of germanium and silicon in elements of Groups III, IV and V of the periodic system were reviewed a number of years ago.[2] It was found that these liquidus curves could be reasonably well approximated by a one-constant equation that had a form consistent with the thermodynamic properties expected of a "regular" liquid solution in equilibrium with a pure solid phase. Additional solubility measurements now show that the liquidus curves cannot be satisfactorily represented by such an equation. A simple two-constant equation has been found that describes the curves within the present limits of experimental error. This equation suggests that simple departures from ideal solution entropy occur, as well as a simple heat-of-mixing effect.

Our objectives have been: (a) to construct the best liquidus curves of germanium and silicon from measurements reported in the literature, supplemented by our own measurements and (b) to obtain the activity coefficients of germanium and silicon along the liquidus curves. This information will be used to estimate the thermodynamic properties of the liquid binary alloys.

One method of evaluating liquidus curves is to plot the solubility measurements on a temperature-composition (atom fraction) scale and draw the best curve through the points from the eutectic to the melting point. For many of the liquidus curves of interest here this is neither practical nor desirable. A number of the liquidus curves cover a very long temperature range (as much as 1380 degrees in the case of the silicon-gallium system) and long composition range (from an atom fraction of about 10^{-10} to 1 in the same system). There is only a limited amount of experimental data available for these systems and rather long extrapolations of the data, to both higher and lower temperatures and compositions, are required if an estimate of the complete curve is to be made.

It is possible, however, to make such extrapolations with greater confidence by taking advantage of all other available information that influences the position of a liquidus curve. The first and most important

bit of additional information is that the measured liquidus curves appear to represent equilibrium conditions between liquid and solid phases. Studies of solubility as a function of time lead to this conclusion, as do the agreement between measurements made by different methods. The disagreement that exists between the measurements of solubility made by different workers frequently arises, we believe, from experimental error, part of which may be attributable to the fact that thermodynamic equilibrium was not attained. However, in general, this problem is avoidable.

With the knowledge that thermodynamic equilibrium exists between the phases, additional information becomes pertinent: (a) germanium and silicon exhibit only one crystalline modification; (b) the solid solubility of the other component in germanium and silicon is small; (c) the melting points and heats of fusion of germanium and silicon are now better known; (d) the liquidus curves can be expected to exhibit certain simple properties in the region of the melting point of germanium and silicon; (e) the liquidus curves that extend to low temperatures and represent liquids that become very dilute in germanium or silicon can also be expected to exhibit certain simple properties.

The implications of this additional information are as follows: (a) the fact that germanium and silicon exhibit only one crystalline modification means that the liquidus curves will be smooth curves with no abrupt changes of slope; (b) the fact that the solid solubility of the second component is small means that the thermodynamic properties of the solid phase are essentially those of pure germanium or silicon; (c) the more accurate knowledge we now have of the melting points of germanium and silicon can be used to construct better curves, since curves now in the literature use melting points varying from 936 to 958°C for germanium and from 1400 to 1430°C for silicon; (d) a knowledge of the heats of fusion of germanium and silicon and the fact that the solid solubilities are small means that the liquidus curves can be expected to come into the melting point of germanium or silicon with a certain known limiting slope; (e) when the amount of solute in the liquid phase is small at low temperatures, it can be expected that the logarithm of the atom fraction of the solute will be well approximated by a straight line when it is plotted as a function of the reciprocal of the absolute temperature.

We have used a method of evaluating the liquidus curve measurements that takes advantage of all the above information related to these curves. In addition, this method of evaluation has led to the discovery of an apparent regularity that leads to an interesting suggestion about the thermodynamic properties of the liquid alloys. Specifically, we have

calculated the activity coefficients of germanium and silicon along a liquidus curve from the solubility measurements, and calculated the parameter $\alpha \equiv RT \ln \gamma_1/(1 - x)^2$, where R is the gas constant, T the absolute temperature, γ_1 the activity coefficient of germanium or silicon referred to the pure supercooled liquid state, and x the atom fraction of germanium or silicon in the saturated liquid. It has been found that α is a linear function of T. This leads to a two-constant liquidus curve equation that has all the desired features. Previously[2] it had been found that the liquidus curves of a number of these systems could be approximated by a one-constant equation, but more accurate solubility measurements now show that two constants are needed.

The liquidus curves formed by 14 different elements with germanium and silicon, for a total of 27 different curves, have been evaluated. The parameter α is found to be a linear function of temperature over a significant temperature range within experimental error for 17 of these binary systems (germanium-indium, -gallium, -aluminum, -lead, -tin, -bismuth, -cadmium, -silver, -copper, -gold; silicon-indium, -gallium, -tin, -bismuth, -antimony, -silver, -zinc). The solubility data for seven other systems (germanium-arsenic, -antimony; silicon-aluminum, -lead, -arsenic, -copper, -gold) are more limited or exhibit considerable scatter. An estimate of the best liquidus curve representing the available measurements is obtained by assuming that α is a linear function of T. The complete germanium-thallium liquidus curve has been estimated from very limited solubility measurements, and the silicon-thallium curve has been estimated, although no solubility measurements have been made. Two systems, germanium-zinc and silicon-nickel, have liquidus curves that appear to be qualitatively the same as the other curves when plotted as T versus x, but it is found that α is not a linear function of temperature. New solubility measurements are reported for 14 binary systems (germanium-thallium, -gallium, -tin, -bismuth, -cadmium, -arsenic, -zinc; silicon-tin, -lead, -bismuth, -antimony, -zinc and the previously unreported work of Hassion[3] on germanium-indium and germanium-lead).

After a description of our experimental procedure we will illustrate this method of evaluation of the solubility measurements by discussing the germanium-gallium system in some detail. Each of the other systerms will then be considered briefly.

II. EXPERIMENTAL PROCEDURE

The solubility measurements were made by a method similar to that used by Kleppa and Weil.[4] An excess of germanium or silicon was sealed

into an evacuated silica tube with a known weight of solvent. The germanium and silicon were in the form of single-crystal ingots of high purity. The temperature of the furnace containing the quartz tube was raised slowly to insure against overshooting, and then held constant for various periods of time before the furnace was tipped. A constriction in the tube permitted separation of the saturated melt from the excess germanium or silicon by tipping of the furnace. The equilibration time varied from one-half to one hour for germanium and up to 19 hours for silicon. During equilibration the temperature was held constant to within one degree centigrade, by manual control for the shorter times and by a Leeds and Northrup Speedomax controller for the longer heating times. The emf's of the calibrated platinum and platinum–10 per cent rhodium thermocouples used for temperature measurement were determined with a Leeds and Northrup portable precision potentiometer. Temperatures are considered to be accurate to within ±2 degrees.

The compositions of the saturated melts were obtained from the loss in weight of the germanium or silicon ingots, except for the germanium-arsenic system, in which case the poured-off melts were chemically analyzed. In order to obtain the loss in weight of the single-crystal ingots, the small amount of melt adhering to the crystal was dissolved in a solvent that would not attack the pure germanium.*

III. TREATMENT OF DATA

3.1 *Germanium-Gallium*

The solubility of germanium in liquid gallium has been measured by Klemm et al.,[5] Keck and Broder,[6] Greiner[7] and de Roche.[8] These measurements are shown in Fig. 1, along with our own measurements. The measurements of Klemm et al. are in substantial disagreement with the others. The disagreement between the measurements of de Roche and the others at the lower temperatures is quite evident in Fig. 2, where log x has been plotted as a function of the reciprocal of the absolute temperature, $1/T$. The measurements of Klemm et al. are not included in this figure.

An interesting feature of these measurements is that they fall close to the ideal liquidus curve. Included in Fig. 2 are two dashed curves, each an ideal liquidus curve corresponding to different assumptions about the heat capacity of supercooled liquid germanium. One of the curves

* In general, hot concentrated HCl was used. A mixture of 30 per cent H_2O_2, glacial HAc and H_2O in the volume ration 2:2:5 was used for lead on germanium.[3] Aqua regia was used for antimony on silicon.

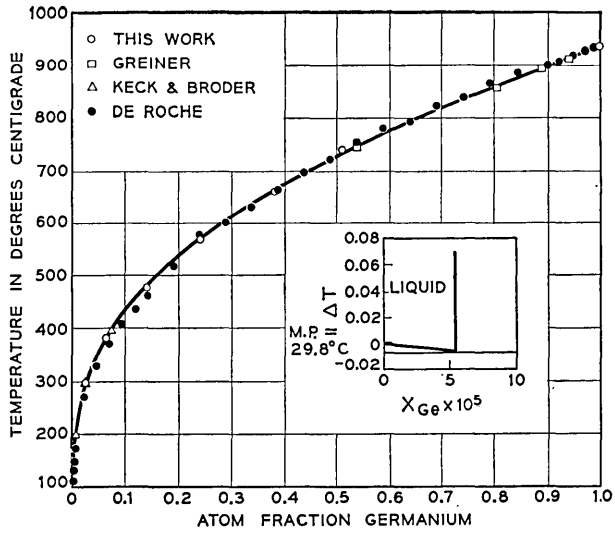

Fig. 1. — Germanium-gallium liquidus curves.

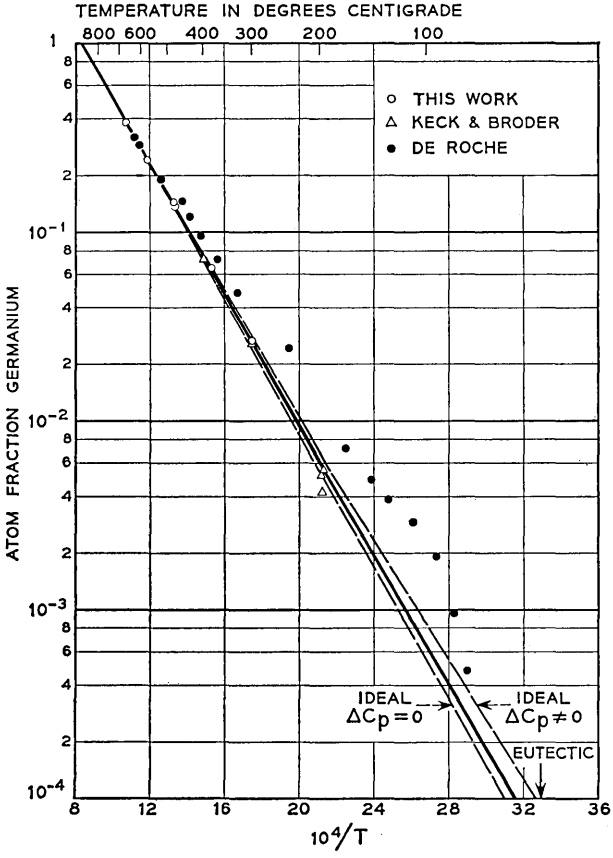

Fig. 2 — Germanium-gallium liquidus curves, log x vs. $1/T$.

results from the assumption that the heat capacity of liquid germanium is the same as solid germanium ($\Delta C_p = 0$); the other results from the assumption that the heat capacity of liquid germanium is constant ($\Delta C_p \neq 0$). If the liquid alloys are ideal solutions of atoms, if these liquids are at equilibrium with pure solid germanium and if the heat capacity of pure supercooled liquid germanium is the same as pure solid germanium, the liquidus curve will be given by the following equation (Ref. 9, pp. 357–358):

$$\ln x = \frac{\Delta H_1^F}{R}\left(\frac{1}{T_1^\circ} - \frac{1}{T}\right). \tag{1}$$

If, however, the heat capacity of pure supercooled liquid germanium is constant and the heat capacity of pure solid germanium has the form $a + bT$, the ideal liquidus curve will be given by

$$\ln x = \frac{\Delta H_1^F}{R}\left(\frac{1}{T_1^\circ} - \frac{1}{T}\right) + \frac{1}{R}\left\{(a - bT)\,\frac{T_1^\circ - T}{T} + a \ln \frac{T}{T_1^\circ} + \tfrac{1}{2}b\left[(T_1^\circ)^2 - T^2\right]\right\}. \tag{2}$$

The dashed curve of Fig. 2 designated $\Delta C_p = 0$ corresponds to (1) with ΔH_1^F, the latent heat of fusion of germanium, taken as 8100 calories per gram atom[10] and T_1°, the melting point of germanium, 937°C.[7,11] The dashed curve of Fig. 2 designated $\Delta C_p \neq 0$ is obtained from (2), using the specific heat equation listed by Kubaschewski and Evans[12] for solid germanium and assuming that the specific heat of liquid germanium is constant and equal to that of solid germanium at its melting point.

At temperatures near the melting point of germanium, de Roche reports that the lowering of the freezing point leads to a heat of fusion of 8100 ± 200 calories per gram atom. This is in agreement with Greiner and Breidt.[10] The agreement between the heat of fusion obtained by freezing point lowering measurements and the heat obtained by direct measurement is evidence that germanium and gallium form atomically dispersed liquid solutions.* If other elements form atomically dispersed binary liquid phases with germanium and also exhibit negligible solid solubility, their liquidus curves will all come into the melting point of germanium with the same limiting slope. In other words, (1) will be a good approximation to all the liquidus curves at temperatures near the melting point of germanium.

* The possibility cannot be ruled out that germanium and gallium form polymeric species $(Ge)_x$ and $(Ga)_x$, but this is believed to be unlikely.

At low temperatures, where the atom fraction of germanium in the saturated liquid phase is small, we expect that the form of (1) will be correct but that, in general, the slope and intercept (on a log x–$1/T$ plot) will be different for each element.

Let it now be assumed that (1) is the correct equation for the ideal liquidus curve (this corresponds to saying that the heat of fusion of germanium is not a function of temperature — i.e., the heat capacity of solid and supercooled liquid germanium are the same — and that the solid solubility of the component forming the ideal liquid solutions with germanium is zero). We can measure the departures of any experimental liquidus curve from the ideal liquidus curve in terms of the activity coefficient, γ_1, defined by the expression (Ref. 9, pp. 357–358):

$$\ln x\gamma_1 = \frac{\Delta H_1^F}{R}\left(\frac{1}{T^\circ} - \frac{1}{T}\right). \tag{3}$$

Consequently, a value of γ_1 can be calculated for every solubility measurement.

The activity coefficient defined in this way has the pure (supercooled) liquid germanium as the reference state. Thus, when x is unity, γ_1 is unity. Since x will be unity at the melting point of germanium, (1) becomes a good approximation to (3) when the temperature is near the melting point of germanium.

The relationships between γ_1 and other thermodynamic parameters are given below:

$$RT \ln \gamma_1 = \Delta \bar{F}_1^e, \tag{4}$$

$$\Delta \bar{F}_1^e = \Delta \bar{H}_1 - T\Delta \bar{S}_1^e. \tag{5}$$

The terms $\Delta \bar{F}_1^e$ and $\Delta \bar{S}_1^e$ are the relative partial molar excess free energy and entropy, respectively, and $\Delta \bar{H}_1$ is the relative partial molar enthalpy. From (4) and (5), and using the fact that the heat of fusion is equal to $T_1^\circ \Delta S_1^F$, (3) may be rewritten in the following form:

$$\ln x = -\frac{\Delta H_1^F + \Delta \bar{H}_1}{RT} + \frac{\Delta S_1^F + \Delta \bar{S}_1^e}{R}. \tag{6}$$

When the liquidus curve extends to low temperatures where x is small, $\Delta \bar{H}_1$ and $\Delta \bar{S}_1^e$ will no longer depend significantly on composition. Since the temperature dependences of $\Delta \bar{H}_1$ and $\Delta \bar{S}_1^e$ are expected to be small, (6) will give a linear relationship between $\ln x$ and $1/T$. In the germanium-gallium system, the values of $\Delta \bar{H}_1$ and $\Delta \bar{S}_1^e$ are close to zero.

Equation (6) can also be used as a general expression for the liquidus curves, where, in general, $\Delta\bar{H}$ and $\Delta\bar{S}_1^e$ are functions of composition and temperature. Since we expect the temperature dependence of $\Delta\bar{H}_1$ and $\Delta\bar{S}_1^e$ to be small, the variation in $\Delta\bar{H}_1$ and $\Delta\bar{S}_1^e$ will arise primarily from their composition dependence. The composition dependence of $\Delta\bar{H}_1$ and $\Delta\bar{S}_1^e$ can each be represented conveniently by a power series in $(1 - x)$:

$$\Delta\bar{H}_1 = \sum a_n(1 - x)^n, \tag{7}$$

$$\Delta\bar{S}_1^e = \sum b_n(1 - x)^n. \tag{8}$$

Since $\Delta\bar{H}_1$ and $\Delta\bar{S}_1^e$ are zero when $x = 1$, a_0 and b_0 are both zero. If long-range forces between the atoms are not present, a_1 and b_1 will be zero.[13] Consequently, a first approximation to the composition dependences of $\Delta\bar{H}_1$ and $\Delta\bar{S}_1^e$ can be expected to be

$$\Delta\bar{H}_1 = a_2(1 - x)^2 \tag{9}$$

and

$$\Delta\bar{S}_1^e = b_2(1 - x)^2. \tag{10}$$

Upon substitution of (9) and (10) in (6) and rearranging, we obtain

$$\alpha \equiv \frac{T\Delta S^F - \Delta H^F - RT \ln x}{(1 - x)^2} = a_2 - b_2T. \tag{11}$$

A value of the parameter α can be calculated from every solubility measurement. If $\Delta\bar{H}$ and $\Delta\bar{S}^e$ are independent of temperature and their composition dependences are given by (9) and (10), it follows that α will be a linear function of T.

We have evaluated α for a number of the liquidus curves of germanium and silicon and found that the experimental data can be satisfactorily approximated with α as a linear function of T. As will be discussed below, it does not follow that (9) and (10) must represent the thermodynamic properties of the liquid phases in order for α to be a linear function of T. For any given set of experimental points there are an infinite number of equations for $\Delta\bar{H}_1$ and $\Delta\bar{S}_1^e$ that would provide a fit to an α–T plot within experimental error. The discussion of the preceding paragraph describes the simplest thermodynamic properties the liquid alloys could have which would lead to the liquidus curve shapes we observe. A previous study[2] of a number of these liquidus curves suggested that α was independent of temperature along the liquidus curve. This property of α would be expected if the liquid alloys were regular solutions.*

* The term "regular solutions" was coined by Hildebrand to denote solutions for which "thermal agitation is sufficient to give practically complete random-

The solubility measurements of germanium and gallium evaluated in this way are shown in Fig. 3. The measurements of Greiner,[7] Keck and Broder[6] and our measurements can be represented by a line with $a = -150$ and $b = 0$. The measurements of de Roche[8] are included, except those at low temperatures and those near the melting point of germanium. The parameter α is quite sensitive to experimental error near the melting point, and, between about 1100 and 1211°K, $a = -150$ is a satisfactory representation of all the measurements when a T–x plot is used.

The solubility measurements we have made of germanium in gallium are given in Table I; the measurements of Keck and Broder, scaled from their published figure, and of Greiner are also given. These measurements have been used to obtain a and b.

3.2 *Germanium-Aluminum, Germanium-Indium*

The solubility of germanium in aluminum has been measured by Stöhr and Klemm.[16] Their results, scaled from a figure, are given in Table I.† The parameter α has been plotted as a function of temperature in Fig. 3. A line, $a = -5360$, $b = -3.16$ has been drawn through the points. Some curvature is suggested, but it is believed that the straight line fits the data within experimental error.

The solubility of germanium in indium has been measured by Keck and Broder,[6] whose data, scaled from a figure, are given in Table I, and by Hassion,[3] whose measurements are also given in the table. The measurements of Klemm et al.[5] are in sufficient disagreement with these measurements to justify not including them in Table I and Fig. 3. A line, $a = 1570$, $b = 0.56$, has been drawn through the measurements of Keck and Broder and Hassion, plotted as α versus T in Fig. 3.

3.3 *Germanium-Tin, Germanium-Lead*

The solubility of germanium in tin has been measured by Stöhr and Klemm.[16] However, we have also measured the solubility of germanium

---

ness.[14] This definition would, in general, include those solutions for which $\Delta\bar{S}_1^e$ is zero and for which (9) gives the partial molar heats of solution, but would not be restricted to them. It is useful to classify solutions empirically in terms of the mathematical functions used to describe their thermodynamic properties. The terms "regular solutions" and "strictly regular solutions" are sometimes used (Ref. 9, p. 246 and Ref. 15, p. 85) to describe solutions for which a_2 is constant and b_2 is zero in (9) and (10).

† The parameter α is very sensitive to error in T and x near the melting point. Consequently, solubility measurements in this range have not been weighted very heavily and frequently have not been included in the table of data.

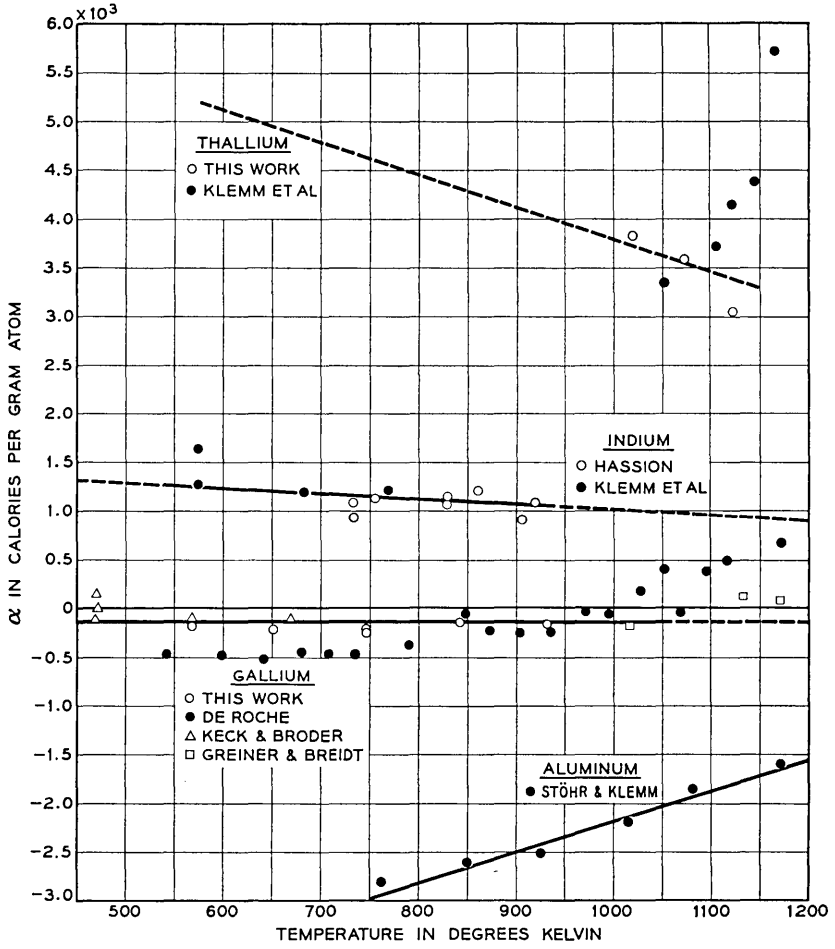

Fig. 3 — α vs. T for germanium-aluminum, germanium-gallium, germanium-indium and germanium-thallium.

in tin and are in significant disagreement. Our measurements are given in Table I and a plot of α versus T is shown in Fig. 4. A line, $a = 1680$, $b = 1.08$, has been drawn through the points. The measurements of Stöhr and Klemm when plotted as α versus T fall on a curve which crosses our data at about 800°K and is of opposite slope. Thus, around 800°K, our measurements agree with those of Stöhr and Klemm, but at higher temperatures we measure somewhat higher solubilities and at lower temperatures somewhat lower solubilities.

TABLE I — EXPERIMENTAL COMPOSITIONS (ATOM FRACTION OF GERMANIUM) AND TEMPERATURES OF GERMANIUM LIQUIDUS CURVES

Element	Source	Temperature, °K	x
Aluminum	Stöhr & Klemm (Ref. 16)	762	0.324
		849	0.410
		926	0.501
		1014	0.614
		1081	0.718
		1172	0.904
Gallium	This work	570	0.0268
		653	0.0655
		749	0.142
		749	0.144
		843	0.243
		933	0.382
	Keck & Broder (Ref. 6)	470	0.0042
		473	0.0052
		470	0.0055
		570	0.026
		670	0.071
	Greiner & Breidt (Ref. 7)	1018	0.541
		1134	0.796
		1171	0.894
		1191	0.944
Indium	Keck & Broder (Ref. 6)	575	0.0061
		575	0.0083
		683	0.033
		769	0.076
	Hassion (Ref. 3)	735	0.064
		735	0.068
		756	0.073
		830	0.138
		830	0.133
		862	0.160
		907	0.254
		921	0.249
Thallium	This work	1023	0.131
		1073	0.260
		1123	0.656
	Klemm et al. (Ref. 5)	1052	0.238
		1106	0.377
		1122	0.492
		1143	0.635
		1165	0.776
Tin	This work	601	0.014
		658	0.0305
		758	0.077
		810	0.125
		860	0.181
		908	0.257
		960	0.362
		999	0.459

TABLE I — *Continued*

Element	Source	Temperature, °K	x
Lead	Hassion (Ref. 3)	901	0.021
		953	0.040
		984	0.052
		1012	0.066
		1050	0.105
		1058	0.112
		1062	0.126
		1108	0.270
		1123	0.37
		1137	0.50
		1146	0.63
Antimony	Ruttewit & Masing (Ref. 18)	872	0.20
		907	0.25
		977	0.40
		1043	0.55
		1058	0.60
		1100	0.70
Arsenic	This work	1023	0.595
		1073	0.693
		1123	0.783
		1148	0.841
	Stöhr & Klemm (Ref. 16)	1023	0.605
		1102	0.738
		1128	0.798
Bismuth	This work	873	0.0273
		923	0.0440
		973	0.0656
		973	0.0728
		1023	0.118
		1073	0.201
Copper	Reynolds & Hume-Rothery (Ref. 21)	917	0.365
		948	0.407
		982	0.460
		1010	0.507
		1048	0.583
		1090	0.674
		1154	0.833
Silver	Maucher (Ref. 20)	925	0.241
		929	0.246
		951	0.271
		1028	0.389
		1104	0.598
		1131	0.690
		1151	0.774
	Briggs et al. (Ref. 22)	923	0.260
		948	0.270
		953	0.286
		978	0.308
		988	0.328
		1013	0.357
		1026	0.398

TABLE I — *Concluded*

Element	Source	Temperature, °K	x
		1070	0.497
		1103	0.618
		1128	0.708
		1149	0.803
		1186	0.923
Gold	Jaffee et al. (Ref. 23)	629	0.270
		919	0.537
		1070	0.731
		1164	0.891
Zinc	This work	723	0.0914
		773	0.148
		823	0.213
		873	0.288
		928	0.368
		973	0.440
		1023	0.530
		1073	0.636
Cadmium	This work	669	0.0076
		669	0.0076
		722	0.0144
		776	0.0284
		776	0.0282
		776	0.0292
		821	0.0469
		877	0.0717
		927	0.131
		978	0.205
		1015	0.306
		1015	0.298
		1015	0.279

The solubility of germanium in lead has been measured by Briggs and Benedict,[17] Ruttewit and Masing[18] and Hassion.[3] We have accepted the measurements of Hassion given in Table I, and have plotted these data as α versus T, shown in Fig. 4. A line, $a = 8780$, $b = 4.08$, has been drawn to represent these measurements. The measurements of Briggs and Benedict are in fair agreement with Hassion's measurements, but the low-temperature solubilities reported by Ruttewit and Masing are much lower than the values obtained from an extrapolation of Hassion's measurements using the α–T plot.

3.4 *Germanium-Arsenic, Germanium-Antimony, Germanium-Bismuth*

The measurements of Stöhr and Klemm[16] of the solubility of germanium in arsenic have been scaled from their published figure and are listed in Table I along with our measurements. The α–T plot is shown

in Fig. 5 and a line, $a = -5600$, $b = -4.16$, has been drawn to represent the data.

The solubility of germanium in antimony has been measured by Ruttewit and Masing[18] and Stöhr and Klemm.[16] The measurements of Ruttewit and Masing, scaled from a figure, are given in Table I and the α–T plot resulting from these measurements is shown in Fig. 5. The line, $a = 2640$, $b = 1.98$, has been drawn. Stöhr and Klemm found smaller solubilities at all temperatures than those found by Ruttewit and Masing. A plot of α versus T with their data leads to a curve lying 500 to 1,000 calories higher than the line drawn to represent the measurements of Ruttewit and Masing. We consider the antimony-germanium liquidus curve to be in some doubt.

The solubility of germanium in bismuth has also been measured by Ruttewit and Masing[18] and by Stöhr and Klemm.[16] We have also measured the solubility of germanium in bismuth, and our results are given in Table I and plotted in Fig. 5. The line, $a = 5505$, $b = 1.49$, has been used to represent the data. Neither the measurements of Ruttewit and Masing nor those of Stöhr and Klemm can be represented as a line on a plot of α versus T. Their measurements are in only fair agreement with those we have made.

3.5 *Germanium-Copper, Germanium-Silver, Germanium-Gold*

The solubility of germanium in copper has been measured by Schwarz and Elstner,[19] Maucher[20] and Reynolds and Hume-Rothery.[21] The meas-

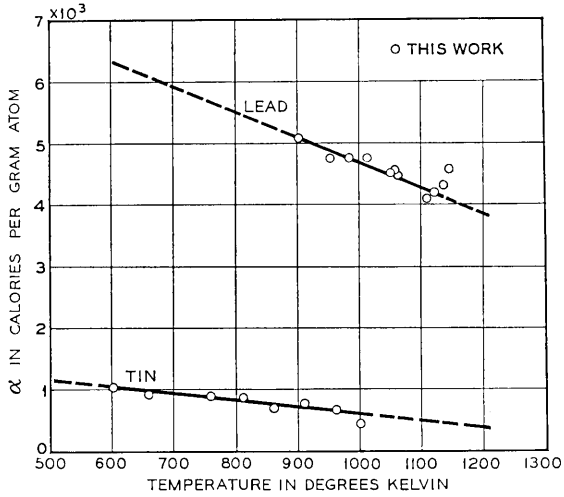

Fig. 4 — α vs. T for germanium-tin and germanium-lead.

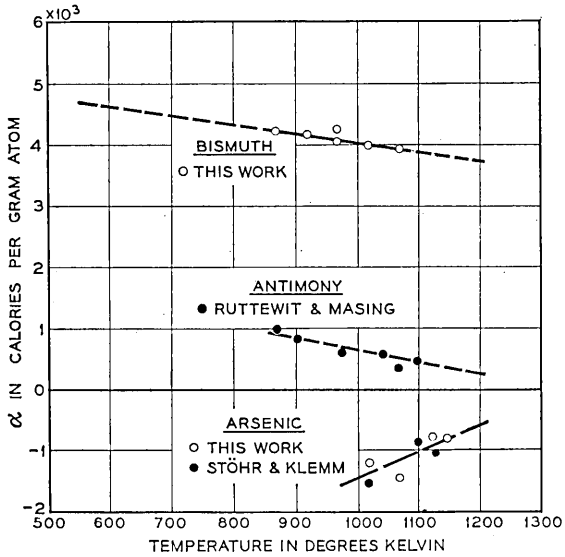

Fig. 5 — α vs. T for germanium-arsenic, germanium-antimony and germanium-bismuth.

urements of Reynolds and Hume-Rothery are given in Table I and the corresponding α–T plot is shown in Fig. 6. The measurements of the other workers scatter rather badly. They are in agreement with Reynolds and Hume-Rothery at low temperatures but show lower solubilities at higher temperatures. The line, $a = -7360$, $b = -7.67$, has been drawn in Fig. 6.

The solubility of germanium in silver, measured by Maucher,[20] and by Briggs, McDuffie and Willisford,[22] is given in Table I; the α–T plot of these data is shown in Fig. 6. The line, $a = -5500$, $b = -7.13$, has been drawn.

The solubility of germanium in gold has been measured by Jaffee, Smith and Gonser.[23] Their measurements are given in Table I and the α–T plot in Fig. 6. The line, $a = -4865$, $b = -1.02$, has been drawn.

3.6 *Germanium-Zinc, Germanium-Cadmium*

Gebhardt[24] and Kleppa and Thalmayer[25] have measured the solubility of germanium in zinc. The measurements of the latter are in fairly good agreement with our measurements, which are recorded in Table I. It can be seen in Fig. 7 that α versus T cannot be represented by a straight

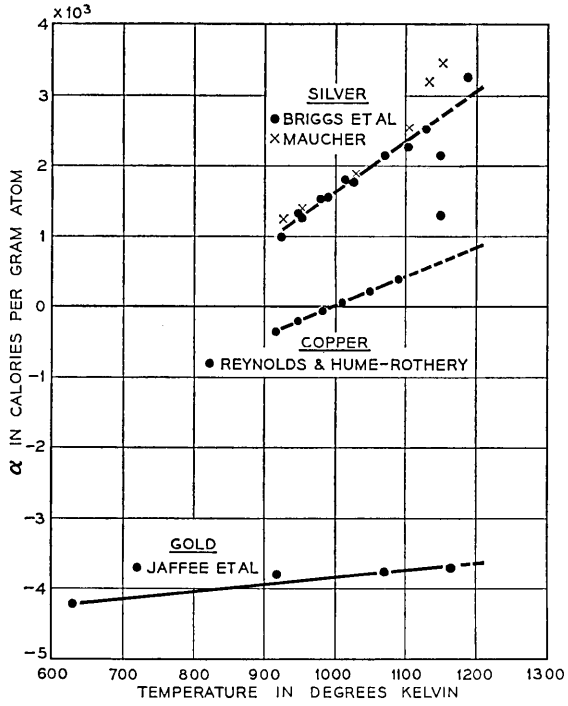

Fig. 6 — α vs. T for germanium-gold, germanium-copper and germanium-silver.

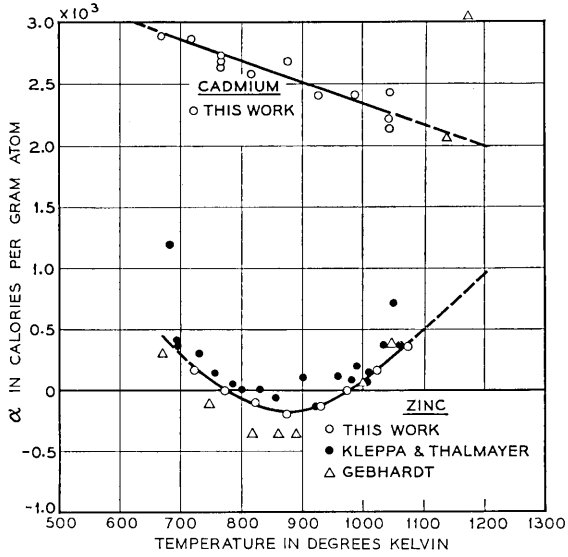

Fig. 7 — α vs. T for germanium-cadmium and germanium-zinc.

line. The measurements of Gebhardt and of Kleppa and Thalmayer are included in the plot.

We have measured the solubility of germanium in cadmium, and the results are given in Table I. The points of the α–T plot appear to be represented within experimental error by the line, $a = 4110$, $b = 1.75$. Spengler[26] has reported that the germanium-cadmium system is simple eutectic and has mentioned three solubility measurements. These three points are at much higher germanium concentrations than those we have found.

3.7 *Silicon-Aluminum, Silicon-Gallium, Silicon-Indium*

We have evaluated the measurements of the solubility of silicon in a number of elements of interest in a manner completely analogous to that used for germanium. We have used a heat of fusion of 12,100 calories per gram atom.[27] The melting point of silicon has been reported as 1415°C by Gayler[28] and 1408°C by Pell.[29] We have used a value of 1410°C.

The solubility of silicon in aluminum has been measured by Fraenkel,[30] Roberts,[31] Gwyer and Phillips,[32] Broniewski and Smailowski[33] and Craighead, Cawthorne and Jaffee.[34] The measurements of Fraenkel and Broniewski and Smailowski were scaled from the figure published by Hansen.[35] These measurements are listed in Table II and α versus T is plotted in Fig. 8.* The line, $a = -4140$, $b = -1.22$, has been drawn.

The silicon-gallium liquidus curve has been studied by Keck and Broder[6] and Klemm et al.[5] The measurements tabulated in Table II were scaled from figures. The points plotted in the α–T plot of Fig. 8 have been represented by the line, $a = 3250$, $b = 0.83$.

The solubility of silicon in indium has been measured by Keck and Broder[6] and by Klemm et al.[5] We have also studied this system. Our solubility measurements are given in Table II and the α–T plot is shown in Fig. 8, where the line, $a = 11{,}450$, $b = 3.37$, has been drawn. The measurements of Klemm et al., are in essential agreement with our measurements, but by themselves give no indication of a linear relation between α and T. The three measurements of Keck and Broder are all at higher solubilities than we have measured, their lowest temperature measurement being in greatest disagreement with the liquidus curve corresponding to the line of Fig. 8.

* The tabulation in Table II does not include the measurements made near the melting point of silicon. The measurements of Gwyer and Phillips near the eutectic, which are in agreement with the measurements of Craighead et al., are not plotted in Fig. 8.

TABLE II — EXPERIMENTAL COMPOSITIONS (ATOM FRACTION OF SILICON) AND TEMPERATURES OF SILICON LIQUIDUS CURVES

Element	Source	Temperature, °K	x
Aluminum	Fraenkel (Ref. 30)	957	0.194
		1232	0.391
		1439	0.597
		1503	0.717
	Roberts (Ref. 31)	969	0.179
		1116	0.319
		1232	0.391
		1307	0.482
		1503	0.698
		1526	0.731
		1607	0.850
	Broniewski & Smailowski (Ref. 33)	1131	0.337
		1333	0.510
	Gwyer & Phillips (Ref. 32)	850	0.113
		852	0.116
		852	0.119
		867	0.124
		875	0.134
		889	0.142
		950	0.183
	Craighead et al. (Ref. 34)	851	0.118
		854	0.123
		871	0.132
Gallium	Keck & Broder (Ref. 6)	548	0.000016
		769	0.00254
		1023	0.034
		1023	0.050
		1170	0.100
		1273	0.200
		1423	0.400
Indium	This work	1173	0.0088
		1273	0.0190
		1273	0.0205
		1373	0.0425
		1373	0.0494
		1373	0.0400
		1473	0.0955
		1473	0.101
		1473	0.101
Tin	This work	1025	0.00391
		1073	0.00607
		1075	0.00792
		1075	0.00760
		1099	0.00754
		1099	0.00778
		1150	0.0112
		1150	0.0113
		1173	0.0143
		1173	0.0136
		1202	0.0166
		1223	0.0258

TABLE II — *Continued*

Element	Source	Temperature, °K	x
		1235	0.0235
		1235	0.0232
		1247	0.0279
		1251	0.0251
		1251	0.0244
		1273	0.0280
		1273	0.0330
		1273	0.0292
		1296	0.0369
		1300	0.0358
		1300	0.0368
		1329	0.0454
		1373	0.0645
		1373	0.0597
Lead	This work	1323	0.0022
		1349	0.0025
		1373	0.0036
		1423	0.0046
		1474	0.0076
		1475	0.0072
		1523	0.0112
Arsenic	Klemm & Pirscher (Ref. 38)	1346	0.60
		1483	0.65
		1512	0.70
Antimony	This work	1073	0.0151
		1073	0.0143
		1170	0.0239
		1174	0.0243
		1272	0.0461
		1371	0.0835
		1473	0.150
Bismuth	This work	1273	0.00262
		1373	0.00544
		1423	0.00832
		1473	0.0113
		1548	0.0193
Zinc	This work	915	0.00845
		925	0.00962
		971	0.0131
		973	0.0168
		1017	0.0208
		1020	0.0200
		1024	0.0247
		1069	0.0321
		1071	0.0325
		1072	0.0394
		1117	0.0486
		1119	0.0526

TABLE II — *Concluded*

Element	Source	Temperature, °K	x
Copper	Hansen & Anderko (Ref. 41)	1075	0.300
	Rudolfi (Ref. 40)	1103	0.355
		1268	0.424
		1336	0.486
		1433	0.598
		1503	0.685
		1547	0.763
Silver	Rudolfi (Ref. 40)	1215	0.168
		1307	0.224
		1402	0.300
		1513	0.490
		1563	0.622
		1592	0.721
		1613	0.794
		1623	0.852
Gold	Hansen & Anderko (Ref. 41)	643	0.309
	di Capua (Ref. 43)	1063	0.439
		1303	0.553
		1403	0.637
		1448	0.701
		1538	0.775
		1568	0.824

3.8 *Silicon-Tin, Silicon-Lead*

We have measured the solubility of silicon in tin, and our results are given in Table II and Fig. 9. The line corresponds to $a = 8{,}145$, $b = 1.50$. The measurements of Tamaru,[36] who used 92.5 per cent silicon (the principle impurities were iron and aluminum), give higher solubilities than we have found.

Our measurements of the solubility of silicon in lead are also given in Table II, and the α–T plot (Fig. 9) shows the line, $a = 19{,}830$, $b = 4.58$, drawn through the points. Moissan and Siemens[37] reported a number of solubility measurements, all of which lie at appreciably lower silicon concentrations than those which we report.

3.9 *Silicon-Arsenic, Silicon-Antimony, Silicon-Bismuth*

Klemm and Pitscher[38] have obtained three points on the silicon-arsenic liquidus curve that are listed in Table II. These three points have lead us to estimate the complete liquidus curve with the line, $a = -49{,}990$, $b = -32.40$, in the α–T plot of Fig. 10.

Solubility measurements of silicon in antimony have been reported by Williams.[39] Our measurements, given in Table II, are in essential

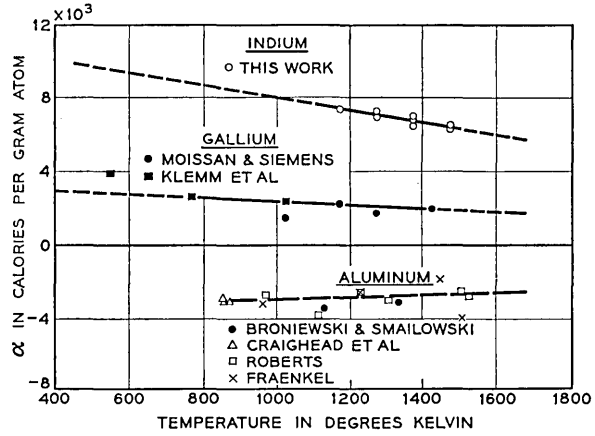

Fig. 8 — α vs. T for silicon-aluminum, silicon-gallium and silicon-indium.

agreement with these. However, we have used our measurements to establish the line that we believe to be a best approximation at present to the silicon-antimony liquidus curve. The line, $a = 3290$, $b = -1.61$, is shown with the experimental points in Fig. 10.

Our measurements of the solubility of silicon in bismuth are recorded in Table II and plotted in Fig. 10. The line, $a = 14{,}840$, $b = 2.06$, has been drawn.

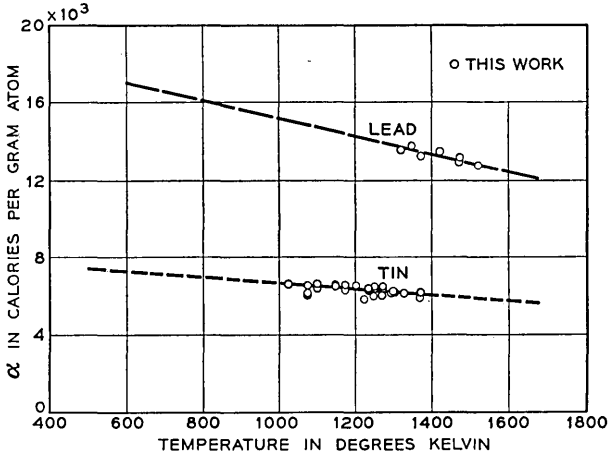

Fig. 9 — α vs. T for silicon-lead and silicon-tin.

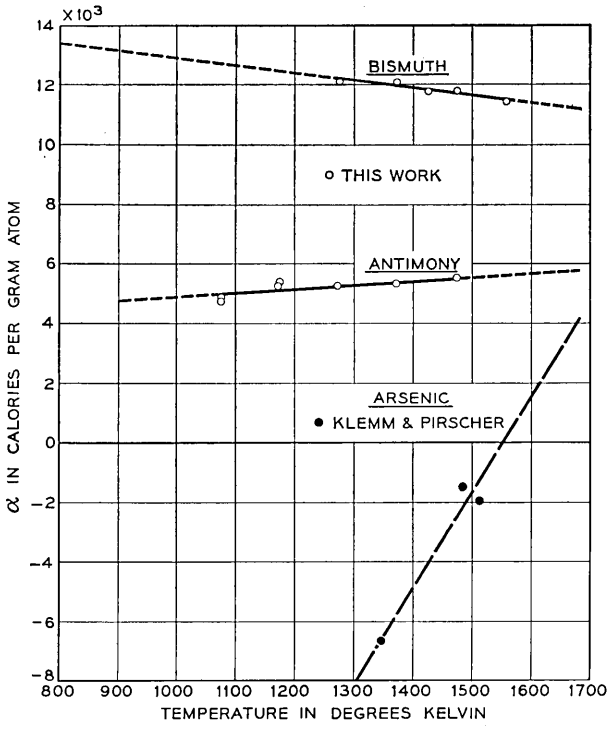

Fig. 10 — α vs. T for silicon-arsenic, silicon-antimony and silicon-bismuth.

3.10 *Silicon-Copper, Silicon-Silver, Silicon-Gold*

Rudolfi[40] has reported measurements of the solubility of silicon in copper. These are given in Table II, along with a value for the eutectic composition proposed by Hansen.[41] The α–T plot is shown in Fig. 11 and the line, $a = -11{,}910$, $b = -7.19$, has been drawn.

Arrivant[42] has measured the solubility of silicon in silver; his measurements are given in Table II. The line, $a = -7910$, $b = -7.63$, has been drawn through these data plotted as α versus T in Fig. 11.

The solubility of silicon in gold has been measured by di Capua,[43] and his measurements are listed in Table II. The plot of α versus T in Fig. 11 suggests that the data can be represented by the line, $a = -19{,}540$, $b = -10.28$. More weight has been attached to the four points at low temperatures than the high temperature points, where experimental error will lead to a greater scatter of points than at lower temperatures.

3.11 *Silicon-Zinc, Silicon-Nickel*

Our measurements of the solubility of silicon in zinc, given in Table II, lead to an α–T plot, Fig. 12, which suggests the line, $a = 4280$, $b = 1.14$. The measurements reported by Moissan and Siemens[37] lie at somewhat lower values of the solubilities than those we report.

The measurements of Iwase and Okamoto[44] of the solubility of silicon in nickel, scaled from Hansen's plot,[41] are shown in Fig. 12 as an example of a binary silicon system for which the α versus T plot is clearly nonlinear.

IV. DISCUSSION

4.1 *The Liquidus Curve Equation*

Equation (11) can be rearranged to give the following relationship:

$$T = \frac{\Delta H_1^F + a(1-x)^2}{\Delta S_1^F - R \ln x + b(1-x)^2}. \tag{12}$$

This will be called the liquidus equation since liquidus temperatures can be calculated as a function of x for each set of values of a and b. The values of a and b used to approximate the available experimental data have been summarized in Table III.

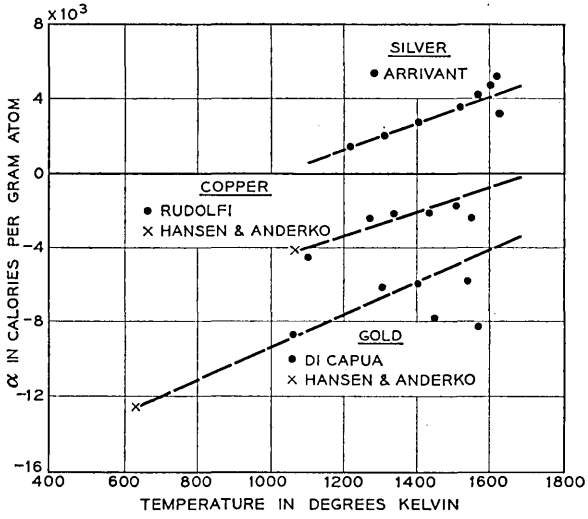

Fig. 11 — α vs. T for silicon-silver, silicon-copper and silicon-gold.

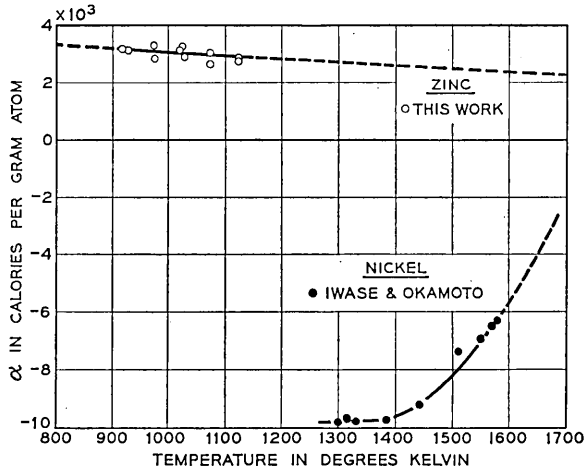

Fig. 12 — α vs. T for silicon-nickel and silicon-zinc.

The solid curves of Figs. 1 and 2 correspond to $a = -150$, $b = 0$ for the gemanium-gallium liquidus curve. The extrapolation of the α–T line to the melting point of germanium means that the liquidus curve approaches the melting point of germanium with the limiting slope determined by the heat of fusion and melting point of germanium. The extrapolation to lower temperatures of the α versus T line is in accord with the expected behavior of such liquidus curves — namely, that at low temperatures the log x versus $1/T$ plot will be linear.

TABLE III — COEFFICIENTS a AND b OF THE LIQUIDUS CURVE EQUATION FOR VARIOUS BINARY SYSTEMS

Element	Germanium		Silicon	
	a	b	a	b
Aluminum	−5360	−3.16	−4140	−1.22
Gallium	−150	0	3250	0.83
Indium	1570	0.56	11450	3.37
Thallium	(5700)	(1.90)	(16600)	(3.80)
Tin	1680	1.08	8145	1.50
Lead	8780	4.08	19830	4.58
Arsenic	−5600	−4.16	−49990	−32.40
Antimony	2640	1.98	3290	−1.61
Bismuth	5505	1.49	14840	2.06
Copper	−7360	−7.67	−11910	−7.19
Silver	−5500	−7.13	−7910	−7.63
Gold	−4865	−1.02	−19540	−10.28
Zinc	—	—	4280	1.14
Cadmium	4110	1.75	—	—

The lines in Figs. 3 through 11 have been drawn to indicate the complete liquidus curve and the amount of extrapolation of the experimental data required to give the full curve. The liquidus curves, plotted as T versus x or log x versus $1/T$, have not been included in this paper. These curves can be readily obtained from (12) and the constants of Table III. However, we have available complete sets of T versus x and log x versus $1/T$ plots which will be supplied to anyone upon request.

We have assumed that the distribution coefficients of the second components in germanium and silicon are small enough to have no influence on the liquidus curve. However, the solid solubility of gallium in germanium is high enough to cause us to expect that some influence should be expected. This is also true for aluminum-germanium. We will use the solid solubility measurement of Trumbore et al.[1,45] to estimate the magnitude of the error introduced in the germanium-gallium system by ignoring the solid solubility.

de Roche[8] has reported that the liquidus curve points he has measured that lie between 1.0 and 0.80 atom fraction of germanium lead to a heat of fusion of germanium of 8100 kilocalories. This calculation used (1), and no correction was made for the solidus curve. If we accept the statement that the experimental liquidus curve follows (1) over the range 1.0 to 0.8 with ΔH_1^F having the value 8,100, we can use the solid solubilities measured by Trumbore, et al.[1,45] and re-evaluate the heat of fusion of germanium.

Equation (1), modified to account for solid solubility, will be

$$\ln \frac{x_L}{x_S} = \frac{\Delta H_1^F}{R}\left(\frac{1}{T_1^\circ} - \frac{1}{T}\right), \tag{13}$$

where x_L is the atom fraction of germanium in the liquid phase and x_S the atom fraction in the solid phase. The value of x_L can be calculated using (1) with $\Delta H_1^F = 8100$ and $T_1^\circ = 1210°$K for various values of the temperature down to a value of x_L of 0.8. The measurements of Trumbore et al. give x_S at these temperatures. A plot of the logarithm of x_L/x_S versus $1/T$ leads to a value of ΔH_1^F of 7850 calories per gram atom. This value is essentially within the experimental error of 8100 $\pm$ 200 cals estimated by de Roche. Consequently, the neglect of the solid solubility introduces a negligible error.

4.2 *Eutectic Temperatures and Compositions*

The eutectic temperatures and compositions that are consistent with the a and b values summarized in Table III are given in Table IV. The

TABLE IV—EUTECTIC TEMPERATURES, T_{eu}°C, DIFFERENCES BETWEEN MELTING POINT OF SECOND COMPONENT AND EUTECTIC TEMPERATURE ($\Delta T = T_2° - T_{eu}$), AND EUTECTIC COMPOSITIONS IN ATOM FRACTIONS OF GERMANIUM AND SILICON

Element	Germanium			Silicon		
	T_{eu}°C	ΔT	x_{eu}*	T_{eu}°C	ΔT	x_{eu}*
Aluminum	424	236	0.28(0.30)	577	82	0.120(0.121)
Gallium	30	0.007	5×10^{-5}	30	6×10^{-8}	5×10^{-10}
Indium	157	0.2	5×10^{-4}	157	1×10^{-7}	2×10^{-10}
Thallium	304	(0.3)	(4×10^{-4})	304	(2×10^{-6})	(3×10^{-9})
Tin	232	0.9	0.003	232	4×10^{-5}	1×10^{-7}
Lead	327	0.1	2×10^{-4}	327	5×10^{-8}	9×10^{-10}
Arsenic†	736	78	0.58(0.59)	1073	−259	0.595(0.595)
Antimony	590	40	0.18(0.17)	630	1.1	0.003
Bismuth	271	0.05	2×10^{-4}	271	3×10^{-8}	1×10^{-10}
Copper	644	434	0.36(0.36)	802	281	0.32(0.30)
Silver	651	310	0.25(0.26)	830	131	0.125(0.154)
Gold	356	707	0.27(0.27)	370	693	0.31(0.31)
Zinc	398	22	0.044(0.055)	420	0.2	4×10^{-4}
Cadmium	320	1.0	0.002	—	—	—

* Compositions in parentheses given by Hansen.[41]
† Not simple eutectic systems.

eutectic temperatures have been taken from the literature and the corresponding composition calculated. There are a number of systems for which the eutectic temperatures are very near the melting point of the solvent. An estimate can be made of the eutectic temperatures and compositions by extrapolating the liquidus curve to the melting point of the solvent, then using this composition to calculate the freezing point depression, as was done in an earlier paper.[2] On the assumption that the solid solubility of germanium or silicon in the solvent element is negligible, the freezing point depression can be calculated from the following equation:

$$\Delta T = \frac{R(T_2°)^2}{\Delta H_2{}^F} x_{eu}, \tag{14}$$

where x_{eu} is the atom fraction of germanium (or silicon) in the eutectic liquid phase, $T_2°$ the melting point of the solvent, and $\Delta H_2{}^F$ its heat of fusion. Equation (14) follows from (1) applied to component 2. The freezing point depressions calculated in this manner, (see Ref. 14, Table B, pp. 284–305, for $\Delta H_2{}^F$), have been tabulated in Table IV, along with other eutectic compositions and temperatures. The germanium-gallium eutectic, calculated on the assumption of negligible solid solubility of germanium and gallium, is shown in the insert of Fig. 1.

4.3 *The Relationship Between a and b*

In Fig. 13 we have plotted the values of b against a. It can be seen that, in general, the more positive the value of a, the more positive b becomes, and the more negative the value of a, the more negative b becomes. We have used this relationship between a and b to estimate the germanium-thallium and silicon-thallium liquidus curves.

4.4 *The Germanium-Thallium and Silicon-Thallium Liquidus Curves*

The available experimental measurements[5] of the solubility of germanium in thallium are given in Table I, and the corresponding α–T points are plotted in Fig. 3. The position of these points leads to a value of α of about 3500 calories per gram atom at 1075°C. On the assumption that the α–T relationship is linear and passes through this point, and that the value of a and b must be related in such a way as to fall on the curve of Fig. 13, we obtain for our estimate of the complete germanium-thallium liquidus curve, $a = 5700$, $b = 1.90$.

From the position of the silicon-lead and silicon-bismuth α–T lines, we estimate a value for α of 12,500 at 1450°K for thallium. Then, proceeding as for the germanium-thallium system, we conclude that a reasonable estimate of the α–T line for thallium would be $a = 16{,}600$, $b = 3.80$. The values of a and b for the thallium systems have been included in Table III in parentheses, and the eutectic compositions and temperatures are shown in Table IV. The α–T line for germanium-thallium is shown in Figure 3.

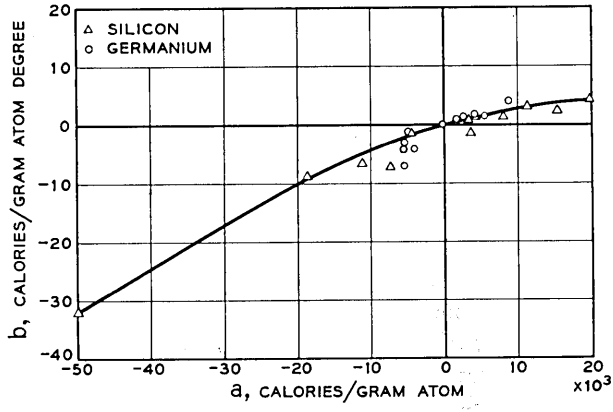

Fig. 13 — b plotted as a function of a.

4.5 *Implications of the $\alpha = a - bT$ Relationship*

In the preceding discussion we have used the α–T plots to estimate the complete liquidus curve from limited solubility measurements. We conclude that, within experimental error, the liquidus curves of a rather large number of binary systems for which germanium or silicon is the solid phase can be described by a two-parameter equation. Two systems, germanium-zinc and silicon-nickel, quite clearly fail to fall into the same class as the other liquidus curves.

In Section 3.1 it was shown that, if the thermodynamic properties of the liquid phases are given by (9) and (10), it would follow that α would be a linear function of T along a liquidus curve. One cannot draw the conclusion, however, that, because α appears to be a linear function of T, the thermodynamic properties of the liquid phase must be given by (9) and (10). The partial molar heat and excess entropy are related to the partial molar excess free energy by partial derivatives at constant composition. The liquidus curves have given partial molar excess free energies along a line in $\Delta\bar{F}_1^e$, x, T, space. It is not possible to obtain partial derivatives of $\Delta\bar{F}_1^e$ at constant composition from such limited information. The linear relationship between α and T only suggests that (9) and (10) describe the thermodynamic properties of these solutions. At present there is no information available which can be used as a direct check of this suggestion.

From (3), (4) and (11) we can conclude that the following relationship is true along the liquidus curve if α is a linear function of T:

$$\Delta\bar{F}_1^e = (a - bT)(1 - x)^2. \tag{15}$$

If this relationship applied off the liquidus curve as well, it would follow from the Gibbs-Duhem equation that, for the other component,

$$\Delta\bar{F}_2^e = (a - bT)x^2 \tag{16}$$

and, for the integral molar excess free energy of mixing,

$$\Delta F_m^e = (a - bT)x(1 - x). \tag{17}$$

It is possible for us to test the applicability of (16) to several of the systems studied by calculating the value of the activity coefficient of the second component at the eutectic temperature and composition. We can use the eutectic composition and temperature of a number of systems to obtain an experimental value of $\Delta\bar{F}_2^e$, which we can write in terms of the activity coefficient, γ_2, since

$$\Delta\bar{F}_2^e = RT \ln \gamma_2 . \tag{18}$$

TABLE V — ACTIVITY COEFFICIENT OF COMPONENT 2 AT THE EUTECTIC

System	γ_2	γ_2, calculated
Aluminum-Germanium	0.84	0.84
Antimony-Germanium	1.08	1.02
Gold-Germanium	0.36	0.78
Aluminum-Silicon	0.99	0.97
Silver-Silicon	1.01	1.00
Gold-Silicon	0.41	0.38

We have used (3), modified to account for a nonzero value of ΔCp (Cp of the liquid was assumed to be constant) as in (2), and applied it to the liquidus curve of the second component at the eutectic. The values of γ_2 have been obtained in this way for a number of systems, using the tabulated data of Kubashewski and Evans,[12] and are given in the first column of Table V. In the second column are values of γ_2 calculated from (16) and (18) using the eutectic temperatures and compositions in Table IV.

A comparison of the activity coefficients in Table V reveals a fairly good agreement between the experimental activity coefficients and those obtained using the appropriate values of a and b in Table III.* An additional bit of information is available for the silicon-silver system. Schadel et al.[46] have measured the vapor pressure of silver over silicon-silver melts at several temperatures in the range of 1000°C. The positive deviations from ideality found by Schadel et al. are in fair agreement with the values we calculate using the a's and b's found from the solubility of silicon in silver, but the temperature coefficient of the activity coefficient they found is opposite to the one we predict. However, it is probable that not much weight can be attached to the experimental temperature coefficients in view of the authors' comments on experimental errors.

4.6 *The Silicon-Thallium, Silicon-Lead and Silicon-Bismuth Binary Phase Diagrams*

We have used the α versus T linear relationship to fit the full liquidus curves of the silicon-lead and silicon-bismuth systems. When the values of a and b of Table III are put into (12), the liquidus curve equation, the curve rises from low temperatures to a maximum, passes through a point of inflection at $x = 0.5$ and then through a minimum before reach-

* The germanium-silver system was not included in Table V because the solid solubility of germanium in silver at the eutectic is large and the activity coefficient of silver in the solid solution is not known.

ing the melting point of silicon, with the appropriate limiting slope. This behavior is in keeping with the fact that a miscibility gap is known to exist in these systems. If we assume that the values of a and b obtained from the liquidus curve at low temperatures give a reasonable estimate of the free energy of the liquid phases of the silicon-lead and silicon-bismuth systems, and that our estimate of the α–T line for silicon-thallium is correct, we can calculate the complete phase diagram.

The thermodynamic conditions for two liquid phases to be in equilibrium lead to the following equation for the miscibility gap when the molar excess free energy of the liquid phases is given by (17):

$$T = \frac{a(1 - 2x')}{R \ln \frac{1 - x'}{x'} - b(1 - 2x')}. \tag{19}$$

The temperature at which this curve intersects the curve given by (12) is the so-called monotectic temperature, where two liquid phases, one rich in silicon and the other dilute in silicon, are in equilibrium with pure solid silicon. The maximum in the miscibility gap occurs at $x = 0.5$. The temperature of the maximum in the miscibility gap, T_c, the critical temperature, cannot be obtained directly from (19), but can be shown to be given by the equation

$$T_c = \frac{a}{2R + b}. \tag{20}$$

The critical temperatures predicted for the thallium, lead and bismuth systems are given in Table VI, along with the monotectic temperatures, T_m, and monotectic compositions, x_m' and x_m''.

4.7 *Liquidus Curves and Critical Temperatures*

The liquidus curves for three systems, silicon-indium, silicon-tin and silicon-antimony, are rather flat in the region of $x = 0.5$, when T is

TABLE VI — ESTIMATED CRITICAL TEMPERATURES, MONOTECTIC TEMPERATURES AND MONOTECTIC COMPOSITIONS FOR THREE SILICON BINARY SYSTEMS WITH LIQUID-LIQUID MISCIBILITY GAPS

Element	T_c, °C	T_m, °C	x'_m	x''_m
Thallium	1862	1387	0.06	0.94
Lead	2047	1397	0.03	0.97
Bismuth	2187	1393	0.04	0.96

plotted as a function of x. This indicates that the miscibility gap occurs at temperatures just below the liquidus curve in the supercooled liquids. If the free energies of the liquid phases were reasonably well approximated by (17), the calculated critical temperatures would all lie below the liquidus curve temperature at $x = 0.5$. The comparison is made in Table VII for these three systems. Included in the table are the critical temperatures and liquidus temperatures for the three germanium systems, germanium-thallium, germanium-lead and germanium-bismuth, which have the highest calculated miscibility gaps. In every case the liquidus curve lies above the calculated critical temperature of the miscibility gap. This shows that the use of (17) for the free energy of the liquid phases is consistent with the presently known properties of these systems.

The systems germanium-arsenic, -copper, -silver and silicon-arsenic, -copper, -silver, -gold all have critical temperatures calculated from (20) that are higher than the temperature of the liquidus curve at $x = 0.5$. It can be shown, however, that this critical temperature is a lower one; that is, it is predicted that, above this temperature, the liquid phase will split into two liquids and below it there will be homogeneity. In all other systems, no critical temperature is predicted.

Some indication of the degree of approximation to be expected from the above treatment of binary systems with miscibility gap can be obtained by applying this treatment to the copper-lead system, which has a miscibility gap that has been measured. We have evaluated α from the measurements of the solubility of copper in molten lead reported by Kleppa and Weil.[4] A plot of α versus T can be approximated by the line, $a = 6160$, $b = 0.85$, although the experimental points clearly depart from this line near the monotectic temperature. Using these values of a and b in (12) and (21), the complete phase diagram was calculated. A comparison between certain critical compositions and temperatures ob-

TABLE VII — COMPARISON OF LIQUIDUS CURVE TEMPERATURES AND CALCULATED CRITICAL TEMPERATURES

System	T (liquidus, $x = 0.5$)	T_c (calculated)
Germanium-Thallium	1114	972
Germanium-Lead	1132	1090
Germanium-Bismuth	1122	1007
Silicon-Indium	1590	1563
Silicon-Tin	1581	1487
Silicon-Antimony	1583	1213

TABLE VIII — COMPARISON OF CALCULATED AND EXPERIMENTAL PHASE DIAGRAM FEATURES FOR THE COPPER-LEAD SYSTEM

	Calculated	Experimental
Critical temperature, T_c	1263°K	1277°K
x_c	0.5	0.35
Monotectic temperature, T_m	1190	1227
x_m'	0.26	0.15
x_m''	0.74	0.67

tained by this calculation and those measured experimentally[41] can be made. This comparison is shown in Table VIII, where it can be seen that a fairly good approximation has resulted.

4.8 *The Thermodynamic Properties of Liquid Alloys*

Darken and Gurry[47] have used the α function $[RT \ln \gamma/(1 - x)^2]$ to describe the thermodynamic properties of liquid alloys at constant temperature. The temperature dependence of α was not considered. Guggenheim (Ref. 15, pp. 83–84) has pointed out that the excess molar free energy of the carbon tetrachloride-benzene and carbon tetrachloride-cyclohexane systems over the temperature range from 30 to 70°C can be quite satisfactorily accounted for by an interchange energy, w, which is a linear function of temperature. This is equivalent to finding that α is a linear function of temperature.

We have evaluated α along the liquidus curves of a number of other binary systems for which germanium or silicon is not one of the components. These systems have long liquidus curves and negligible solid solubility. The general features of the liquidus curves, on a T versus x plot, are the same as those of germanium and silicon binary systems. The α versus T plots of these liquidus curves* were nonlinear in every system.

Scatchard[48] has discussed the relationship between the temperature dependence of α (in essence) and the volume change on mixing. The conclusion has been reached[49] that a volume change upon mixing that is greater than additivity indicates a positive excess entropy of solution, while a volume contraction indicates a negative excess entropy. Prigogine[50] and his co-workers have discussed this relationship in some detail in terms of their average potential model. (See also Ref. 51.) Exceptions can be expected to the above generalization, and such have been observed, but, in general, a close relationship between the excess volume and excess entropy of mixing can be expected. The correlation found here

* Silver-tin, -lead; zinc-gallium, -indium, -tin, -cadmium.

between a and b (Fig. 13) suggests, then, that volume expansion upon mixing occurs for positive a and b while contraction occurs for negative values of a and b.

Kleppa[52,53,54] and Wittig[55] have measured the heats of mixing of quite a number of binary liquid alloys. Trends in heats of mixing with position of one of the components in the periodic system, keeping the other component fixed, were observed. Some of these trends can be satisfactorily accounted for[52] in terms of Friedel's alloy solution model, which, among other things, takes into account the difference in valence expected for the two components. The relative valence effect does not seem to appear, in any simple way, at least, in the system we have studied.

Wittig[55] has found some general trends in the heat of mixing that he has related to the position of the variable component in the periodic system. Germanium occupies a position as a variable component in the several series he has investigated. The heats of solution we have deduced from our analysis of liquidus curves do not appear to violate the trends Wittig has observed.

V. SUMMARY AND CONCLUSIONS

A two-constant equation, (12), has been used to fit the liquidus curves of a number of binary systems for which either germanium or silicon is one of the components. This equation gives a limiting slope of the liquidus curve at the melting point that is consistent with the known heats of fusion of germanium and silicon and the known solid solubilities. The form of the equation at low temperatures is also consistent with the expected thermodynamic properties of the two phases.

The general form of the equation suggests that the thermodynamic properties of the liquid alloys are simple functions of composition and temperature. It was pointed out that a first approximation to the relative partial molar heat of solution of germanium or silicon in the liquid alloys would be given by the equation

$$\Delta\bar{H}_1 = a(1 - x)^2 \tag{21}$$

and a first approximation to the relative partial molar excess entropy by the equation

$$\Delta\bar{S}_1^{\,e} = b(1 - x)^2, \tag{22}$$

where a and b are independent of temperature and composition. This means that the excess free energy would be given by

$$\Delta\bar{F}_1^{\,e} = RT \ln \gamma_1 = (a - bT)(1 - x)^2. \tag{23}$$

If solutions having the thermodynamic properties given by these three

equations were in equilibrium with pure solid germanium or silicon, the liquidus curve equation would have the form

$$T = \frac{\Delta H_1^F + a(1 - x)^2}{\Delta S_1^F - R \ln x + b(1 - x)^2}. \qquad (24)$$

This is the same equation found empirically.

The constants a and b of (24) were obtained from the straight line drawn through the points representing each solubility measurement when plotted as the function α versus T:

$$\alpha \equiv \frac{RT \ln \gamma_1}{(1 - x)^2} = \frac{T\Delta S_1^F - \Delta H_1^F - R \ln x_1}{(1 - x)^2} = a - bT. \qquad (25)$$

It cannot be concluded however, that the thermodynamic properties of the liquid alloys must be given by (21), (22) and (23) as a result of the observed linearity of α with T. The linearity of α as a function of T shows that (23) gives the activity coefficients of germanium or silicon along the respective liquidus curves. It was found, however, that the activity coefficient of the other component at the eutectic in several binary systems was fairly well approximated by the equation resulting when it was assumed that (23) applied at all temperatures and compositions, and the Gibbs-Duhem relationship was used. Based on this assumption, the extent of the miscibility gaps in the silicon-thallium, -lead, -bismuth binary systems were estimated.

The two-constant liquidus curve equation, (24), can be used to provide a best present estimate of the complete liquidus curves for 26 different binary systems for which germanium or silicon is one of the components. The constants for each system are given in Table III. Sets of the liquidus curves, plotted as T versus x or log x versus $1/T$, may be obtained from the authors upon request.

VI. ACKNOWLEDGMENTS

The suggestions and advice of F. A. Trumbore and M. Tanenbaum have been very helpful. We wish also to acknowledge the contributions to this paper made by F. X. Hassion.

REFERENCES

1. Trumbore, F. A., this issue, p. 205.
2. Thurmond, C. D., J. Phys. Chem., **57**, 1953, p. 827.
3. Hassion, F. X., unpublished work.
4. Kleppa, O. J. and Weil, J. A., J. Am. Chem. Soc., **73**, 1951, p. 4848.
5. Klemm, W., Klemm, L., Hohman, E., Volk, H., Örlamunder, E. and Klein, H. A., Z. anorg. allgem. Chem., **256**, 1948, p. 239.
6. Keck, P. H. and Broder, J., Phys. Rev., **90**, 1953, p. 521.
7. Greiner, E. S. and Breidt, P., Jr., J. Metals, **7**, 1955, p. 187.

8. de Roche, N., Z. Metall., **48,** 1957, p. 59.
9. Prigogine, I. and Defay, R., *Chemical Thermodynamics*, Longmans, Green & Co., New York, 1954.
10. Greiner, E. S. and Breidt, P., Jr., J. Metals, **4,** 1952, p. 1044.
11. Hassion, F. X., Thurmond, C. D. and Trumbore, F. A., J. Phys. Chem., **59,** 1955, p. 1076.
12. Kubaschewski, O. and Evans, E., *Metallurgical Thermochemistry*, 3rd Ed., Pergamon Press, London, 1958.
13. Buff, F. P. and Schindler, F. M., J. Chem. Phys., **29,** 1958, p. 1075.
14. Hildebrand, J. H., Farad. Soc. Disc., No. 14, 1953, p. 14.
15. Guggenheim, E. A., *Mixtures*, Clarendon Press, Oxford, 1952.
16. Stöhr, H. and Klemm, W., Z. anorg. allgem. Chem., **241,** 1939, p. 305.
17. Briggs, R. T. and Benedict, W. S., J. Phys. Chem., **34,** 1930, p. 173.
18. Ruttewit, K. and Masing, G., Z. Metall., **32,** 1940, p. 52.
19. Schwarz, R. and Elstner, G., Z. anorg. allgem. Chem., **217,** 1934, p. 289.
20. Maucher, H., Forschungsarbeiten über Metallkunde und Röntgen Metallographie, **20,** 1936.
21. Reynolds, J. and Hume-Rothery, W., J. Inst. Metals, **85,** 1956–7, p. 1731.
22. Briggs, T. R., McDuffie, R. O. and Willisford, C. W., J. Phys. Chem., **33,** 1929, p. 1080.
23. Jaffee, R. I., Smith, E. M. and Gonser, B. W., Trans. A.I.M.E., **161,** 1945, p. 366.
24. Gebhardt, E., Z. Metall., **34,** 1942, p. 255.
25. Kleppa, O. J. and Thalmayer, C. E., to be published.
26. Spengler, H., Metall., **8,** 1954, p. 937.
27. Olette, M., Compt. rend., **244,** 1957, p. 1033.
28. Gayler, M. L. U., Nature, **142,** 1938, p. 478.
29. Pell, E. M., J. Phys. Chem. Solids., **3,** 1957, p. 77.
30. Fraenkel, W., Z. anorg. allgem. Chem., **58,** 1908, p. 154.
31. Roberts, C. E., J. Chem. Soc., **105 II,** 1914, p. 1383.
32. Gwyer, A. G. C. and Phillips, H. W. L., J. Inst. Metals, **36,** 1926, p. 283.
33. Broniewski, W. and Smailowski, M., Rev. Met., **29,** 1932, p. 542.
34. Craighead, C. M., Cawthorne, E. W. and Jaffee, R. I., Trans. A.I.M.E., **81,** 1955, p. 203.
35. Hansen, M., *Der Aufbau der Zweistofflegierungen*, Springer, Berlin, 1936.
36. Tamaru, S., Z. anorg. allgem. Chem., **61,** 1909, p. 40.
37. Moissan, H. and Siemens, F., Compt. rend., **138,** 1904, p. 657.
38. Klemm, W. and Pirscher, P., Z. anorg. allgem. Chem., **247,** 1941, p. 211.
39. Williams, R. S., Z. anorg. allgem. Chem., **55,** 1907, p. 19.
40. Rudolfi, E., Z. anorg. allgem. Chem., **53,** 1907, p. 216.
41. Hansen, M. and Anderko, K., *Constitution of Binary Alloys*, McGraw-Hill Book Co., New York, 1958.
42. Arrivant, G., Z. anorg. allgem. Chem., **60,** 1908, p. 430.
43. di Capua, C., Rend. Accad. Lincci, **29,** 1920, p. 111.
44. Iwase, K. and Okamoto, M., Science Reports, Tohoku Imperial Univ., K. Hondu Anniversary Volume, 1936, p. 777.
45. Trumbore, F. A., Porbansky, E. M. and Tartaglia, A. A., to be published.
46. Schadel, H. M., Jr., Berge, G. and Birchenall, C. E., Trans. A.I.M.E., **188,** 1950, p. 1282.
47. Darken, L. S., Gurry, R. W., *Physical Chemistry of Metals*, McGraw-Hill Book Co., New York, 1953, p. 270.
48. Scatchard, G., Trans. Farad. Soc., **33,** 1937, p. 160.
49. Hildebrand, J. H. and Scott, R. L., *Solubility of Nonelectrolytes*, 3rd Ed., Reinhold Publishing Corp., New York, 1950, p. 136.
50. Prigogine, I., *The Molecular Theory of Solutions*, Interscience Publishers, New York, 1957.
51. Scott, R. L., J. Chem. Phys., **25,** 1956, p. 193.
52. Kleppa, O. J., Kgl. Norske Videnskab. Selskabs, Skifter (English), **2,** No. 6, 1957.
53. Kleppa, O. J., Acta Met., **6,** 1958, p. 225.
54. Kleppa, O. J., Acta Met., **6,** 1958, p. 233.
55. Wittig, F. E., Z. Electrochem., **63,** 1959, p. 327.

Solid Solubilities of Impurity Elements in Germanium and Silicon*

By F. A. TRUMBORE

(Manuscript received August 13, 1959)

The available data on solid solubilities of impurity elements in germanium and silicon are summarized in the form of solidus or solvus curves. New solubility data are presented for the lead-germanium, zinc-germanium, indium-germanium, antimony-silicon, gallium-silicon and aluminum-silicon systems. The correlation of the solid solubilities with the heats of sublimation and the atom sizes of the impurity elements is considered.

I. INTRODUCTION

In recent years a large amount of data has been obtained on the solid solubilities of impurity elements in germanium and silicon. Such data are of obvious practical importance in the semiconductor device field, where controlled impurity distributions are required. Of theoretical interest is the fact that, in favorable cases, these data can be used to provide information on heats and entropies of solution, binding energies and other thermodynamic properties of the solid solutions. In addition, one might hope that the attempts to interpret and correlate the relatively large amount of data on germanium and silicon will lead to a better understanding of the factors affecting solid solubility in other materials.

The purpose of this paper is to summarize and evaluate the experimental solid solubility data for germanium and silicon binary alloy systems. Included in this summary are some new data derived from crystal pulling and thermal gradient crystallization experiments. The use of a modified form of the distribution coefficient is illustrated by considering the empirical correlation of solid solubility with atom sizes and heats of sublimation of the impurity elements. A detailed consideration of the theoretical interpretation of the solid solubility data will be presented in a subsequent paper.[1]

* Presented in part at the meeting of the Electrochemical Society, Philadelphia, May 4, 1959.

II. SUMMARY OF SOLID SOLUBILITY DATA

For many systems the only available solid solubility data are values of $k°$, the distribution coefficient of the impurity element at the melting point of germanium or silicon. Table I summarizes what, in the author's opinion, are the best estimates of $k°$ presently available.* The remaining solid solubility data are summarized in Figs. 1 and 2 as plots of the solidus (solid-liquid equilibria) and, in some cases, the solvus (solid-solid equilibria) curves.† In plotting these curves the melting points of germanium[2] and silicon were taken as 937°C and 1410°C, respectively. The latter value is in agreement with Olette's value[3] of 1412 ± 2°C and Pell's value[4] of 1408 ± 2°C. In the following discussion the sources of the distribution coefficient and solid solubility data are discussed for each impurity element. While no attempt is made to give a complete bibliography of the work in this field, enough references are given to permit the interested reader to find further references to practically all of the work known to the author as of June 1959. Of particular value have been the papers of Burton,[5] Hall[6,7] and Tyler,[8] in which a considerable amount of data has been collected.

2.1 *Solid Solubilities in Germanium*

Lithium. The solidus curve in Fig. 1 is taken from Pell's solid solubility measurements[9] in the range from 593° to 899°C. Pell obtained a value of 1.6×10^{-3} for $k°$ by extrapolating these data to the melting point of germanium. A value of 0.002 is given in Table I because of the uncertainties involved in the 38° extrapolation and in the assumption made by Pell that the liquidus curve could be estimated by using a regular solution model. Pell's flame analyses appear more reliable than the electrical measurements of Reiss, Fuller and Morin[10] above the eutectic tem-

* In this paper the distribution coefficient, k, is defined as $k = x_S/x_L$, where x_S and x_L are the atom fractions of the impurity element in the solidus and liquidus alloys, respectively. Frequently, values of k are reported in terms of concentrations; i.e., $k_c = c_S/c_L$, where c represents the concentration of the impurity element. For small concentrations of impurity in the solid and liquid phases $k = k_c(d_L/d_S)$, where d_L and d_S are the densities of liquid and solid germanium or silicon, respectively. However, in many cases it is unclear whether account was taken of the density change of germanium or silicon on melting or whether k or k_c is reported. No attempt is made here to correct k_c values since the correction is only about 5 per cent for germanium and about 10 per cent for silicon. For most impurity elements this correction is less than the experimental uncertainty in k.

† In Figs. 1 and 2, the numbers following symbols refer to the references from which these points were taken. Where no number is given, the points are taken from the present work. Where a number and no symbol is given, the curve is based at least in part on the work in the reference quoted but no experimental points are given from that reference.

TABLE I — DISTRIBUTION COEFFICIENTS AT THE MELTING POINTS OF GERMANIUM AND SILICON

Element	Germanium	Silicon
Lithium	0.002	0.01
Copper	1.5×10^{-5}	4×10^{-4}
Silver	4×10^{-7}	—
Gold	1.3×10^{-5}	2.5×10^{-5}
Zinc	4×10^{-4}	$\sim 1 \times 10^{-5}$
Cadmium	$>1 \times 10^{-5}$	—
Boron	17	0.80
Aluminum	0.073	0.0020
Gallium	0.087	0.0080
Indium	0.001	4×10^{-4}
Thallium	4×10^{-5}	—
Silicon	5.5	1
Germanium	1	0.33
Tin	0.020	0.016
Lead	1.7×10^{-4}	—
Nitrogen	—	$<10^{-7}$(?)
Phosphorus	0.080	0.35
Arsenic	0.02	0.3
Antimony	0.0030	0.023
Bismuth	4.5×10^{-5}	7×10^{-4}
Oxygen	—	0.5
Sulfur	—	10^{-5}
Tellurium	$\sim 10^{-6}$	—
Vanadium	$<3 \times 10^{-7}$	—
Manganese	$\sim 10^{-6}$	$\sim 10^{-5}$
Iron	$\sim 3 \times 10^{-5}$	8×10^{-6}
Cobalt	$\sim 10^{-6}$	8×10^{-6}
Nickel	3×10^{-6}	—
Tantalum	—	10^{-7}
Platinum	$\sim 5 \times 10^{-6}$	—

perature. Reiss and Fuller[11] have calculated the solvus curve, plotted in Fig. 1, taking into account ion-pairing and hole-electron equilibria.

Copper. The solubility of copper in germanium has been studied quite extensively by a number of authors. The solidus curve in Fig. 1 is based in part on the Hall effect measurements of Woodbury and Tyler,[12] whose data above 650°C are in good agreement with the work of Finn[13] and Hodgkinson,[14] and especially with the equilibrium radiotracer measurements reported by Fuller et al.[15] However, more recent tracer and conductivity measurements of Wolfstirn and Fuller[16] yield somewhat higher solubilities, especially at the higher temperatures. The solidus curve in

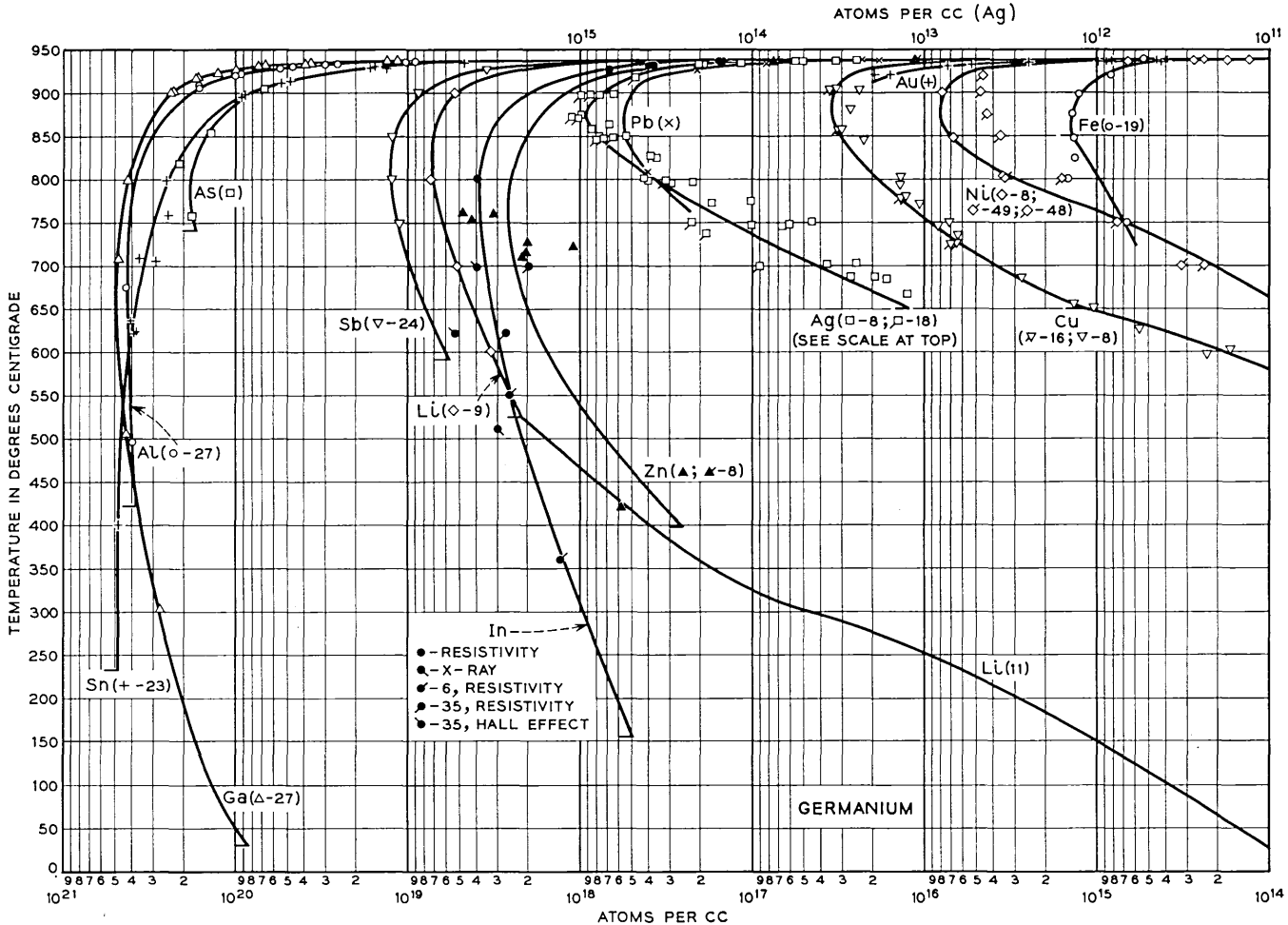
ATOMS PER CC (Ag)
10^{15} 10^{14} 10^{13} 10^{12} 10^{11}
TEMPERATURE IN DEGREES CENTIGRADE
950 900 850 800 750 700 650 600 550 500 450 400 350 300 250 200 150 100 50 0
AS(□)
Sb(▽-24)
Li(◇-9)
Al(○-27)
Sn(+-23)
Ga(△-27)
Pb(×)
Au(+)
Fe(○-19)
Ni(◇-8; ◇-49; ◇-48)
Ag(□-8; □-18) (SEE SCALE AT TOP)
Cu (▽-16; ▽-8)
Zn(▲; ▲-8)
Li(11)
In
●-RESISTIVITY
●-X-RAY
●-6, RESISTIVITY
●-35, RESISTIVITY
●-35, HALL EFFECT
GERMANIUM
10^{21} 10^{20} 10^{19} 10^{18} 10^{17} 10^{16} 10^{15} 10^{14}
ATOMS PER CC

Fig. 1 favors Wolfstirn and Fuller's data at the higher temperatures and Woodbury and Tyler's data at lower temperatures. Below the eutectic temperature the latter authors' solubilities are significantly lower than the previous values determined from electrical measurements,[14,15] probably due in part to the failure of the earlier workers to consider deeper energy levels in interpreting their data. The value for k° of 1.5×10^{-5} given in Table I was taken from the crystal pulling measurements of Burton et al.[17]

Silver. The value of 4×10^{-7} for k° was obtained by Tyler[8] from Hall effect measurements on crystals pulled at temperatures ranging from 0.7 to 5 degrees below the melting point of germanium. This value of k°, which is orders of magnitude lower than the value of 10^{-4} reported by Burton et al.[17], is, however, consistent with the tracer data of Bugay, Kosenko and Miseliuk[18] and with Tyler's Hall effect data at lower temperatures. These data[8,18] were used to construct the solidus and solvus curves in Fig. 1. The temperature data of Bugay et al. appear to be somewhat questionable since, in their work on silver and on iron,[19] measurements were reported extremely close to or slightly above the melting point of germanium, 937°C, used in this paper.

Gold. The value of 1.3×10^{-5} for k° was obtained by Tyler[8] from Hall effect measurements on crystals pulled at temperatures ranging from about 0.2 to 17 degrees below the melting point of germanium. This figure compares favorably with the value of 1.5×10^{-6} reported by Dunlap,[20] and is lower than the figure of 3×10^{-5} obtained by Burton et al.[17] Aside from Tyler's data near the melting point of germanium, no other solidus curve data are available.

Zinc. The value of 4×10^{-4} for k° was obtained by Tyler and Woodbury[8,21] from Hall effect measurements on crystals pulled at temperatures ranging from about 0.1 to 2 degrees below the germanium melting point. This value is in disagreement with the figure of 0.01 due to Burton et al.,[17] but is consistent with low-temperature solubility data obtained by the author. The latter data were obtained from spectrophotometric analyses of the zinc content of crystals grown from zinc-germanium melts in a thermal gradient using methods described previously.[22,23] These results, although not very precise, are somewhat lower than reported by the author in an earlier paper,[24] and are summarized in Table II, where x^{S}_{Zn} is the atom fraction of zinc in the solidus alloy at the temperature, T. It is apparent that the solidus curve in Fig. 1 is uncertain by at least a factor of two, and that further work is needed.

Cadmium. The value of $>1 \times 10^{-5}$ was taken from Woodbury and Tyler.[8,25] No other solid solubility data are available.

Boron. The value of 17 for k° was obtained by Bridgers and Kolb[26]

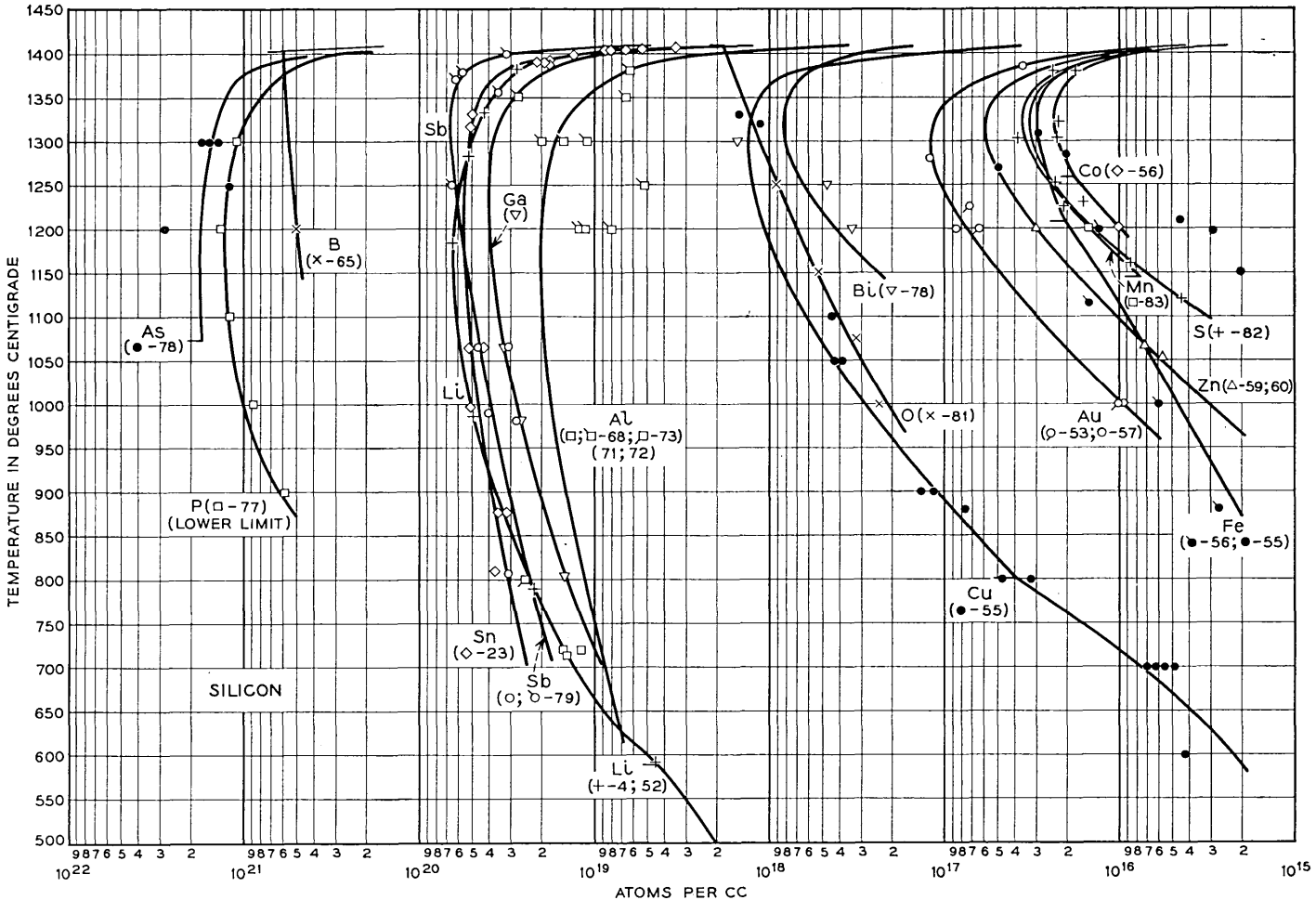
TEMPERATURE IN DEGREES CENTIGRADE
1450
1400
1350
1300
1250
1200
1150
1100
1050
1000
950
900
850
800
750
700
650
600
550
500
SILICON
As (●-78)
B (×-65)
P(□-77) (LOWER LIMIT)
Sb
Ga (▽)
Li
Al (□;▽-68;□-73) (71;72)
Sn (◇-23)
Sb (○;▽-79)
Li (+-4;52)
Bi(▽-78)
O(×-81)
Cu (●-55)
Co(◇-56)
Mn (□-83)
S(+-82)
Au (○-53;○-57)
Zn(△-59;60)
Fe (●-56;●-55)
9 8 7 6 5 4 3 2
10^{22}
10^{21}
10^{20}
10^{19}
10^{18}
10^{17}
10^{16}
10^{15}
ATOMS PER CC

TABLE II — SOLID SOLUBILITY OF ZINC IN GERMANIUM FROM THERMAL GRADIENT EXPERIMENTS

Temperature, in °C	x^{S}_{Zn}
420 ± 20	(1.3 ± 0.7) × 10^{-5}
709 ± 10	(4.9 ± 0.8) × 10^{-5}
714 ± 10	(4.7 ± 0.6) × 10^{-5}
720 ± 10	(2.4 ± 0.2) × 10^{-5}
725 ± 10	(4.6 ± 2.3) × 10^{-5}
752 ± 10	(9.8 ± 4.9) × 10^{-5}
759 ± 10	(7.2 ± 0.3) × 10^{-5}
760 ± 10	(1.1 ± 0.1) × 10^{-4}*

* This value was obtained by spectroscopic analysis while the other figures are the results of spectrophotometric analyses. The uncertainties quoted in this table represent the spread in the results of analyses on different portions of the same sample.

from a study of the effect of growth rate on k. No other solid solubility data are available.

Aluminum. The value of 0.073 for $k°$ and the solidus curve in Fig. 1 are taken from the work of Trumbore, Porbansky and Tartaglia[27] based on chemical analyses of crystals grown by pulling and thermal gradient techniques. Support for the validity of these results is found in the internal consistency of chemical and electrical measurements on the same crystals reported by Trumbore and Tartaglia.[28]

Gallium. The value of 0.087 for $k°$ and the solidus curve in Fig. 1 are also due to Trumbore, Porbansky and Tartaglia.[27] A discussion of most of the previous work on both aluminum and gallium is given in their paper. The value of $k°$ is in agreement with the work of Bridgers and Kolb,[29] who obtained a value between 0.085 and 0.01, and with the recent value of 0.085 due to Leverton.[30] Leverton's value of 0.085 is probably a value of k_c and if corrected to k using his density for liquid germanium would be ~0.095. This figure is probably slightly high because of his relatively large pull rates.

Indium. Burton et al.,[17] Hall,[31] Dowd and Rouse[32] and Leverton[30] have obtained values for $k°$ between 0.001 and 0.0013 from tracer and/or conductivity measurements on pulled crystals. In the author's opinion, the values 0.0012–0.0013 are probably high because of the relatively large pull rates used in these experiments.[30,32] Evidence that $k°$ *may* be as low as 7×10^{-4} has been obtained by the author from resistivity measurements on sections of 16 crystals pulled under a variety of growth conditions. The results of these experiments are summarized in Fig. 3(a), which is a plot of the resistivity as a function of the amount of indium

in the melt. All of these crystals were pulled at a rate of 0.5 cm per hour. The "small" crystals were between 5 and 8 mm in diameter, while the "large" crystals were between 19 and 28 mm in diameter. Unfortunately, chemical analyses of these crystals proved to be unreliable, so that no independent check of the impurity concentration could be made. However, if the resistivity versus concentration curve of Ref. 28 is used, distribution coefficients can be calculated. (Over most of the concentration range, these calculations are essentially based on the assumption in Ref. 28 that the degenerate Hall effect formula, $R_H = 1/pe$, is valid.) The results of these calculations are summarized in Fig. 3(b), where the calculated distribution coefficient is plotted as a function of the melt composition.

A very interesting feature is immediately apparent from Fig. 3. The set of "small" crystals pulled in the [100] direction at ~60 rpm gives

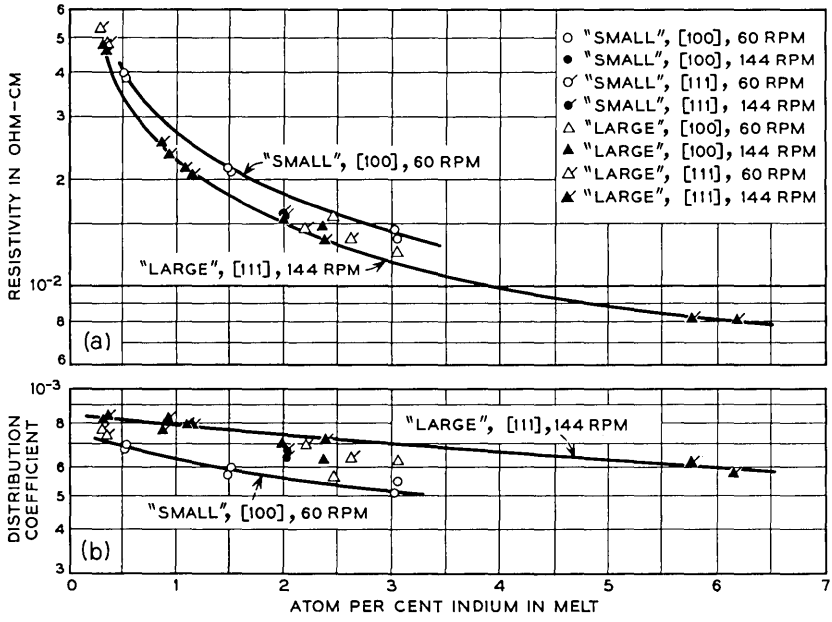

Fig. 3 — Distribution coefficient of indium in germanium as a function of melt composition in crystal pulling experiments. The values of k plotted in (b) were obtained from the resistivity measurements in (a) and the resistivity-concentration curves in Ref. 28. The points designated as corresponding to crystals rotated at 60 rpm correspond in some cases to rotation rates of 57 rpm and, for the point corresponding to 3.05 per cent indium in the melt, the rate was 50 rpm. With one exception all crystals were pulled at 0.5 cm/hr. One crystal (corresponding to the point at 2.03 per cent indium in the melt) was pulled at 0.1 cm/hr.

lower values of k than the set of "large" crystals pulled in the [111] direction at 144 rpm. Evidence that the size of the crystal is the most important factor determining the effective distribution coefficient is found in the fact that "large" crystals pulled in the [100] direction at either 144 or 57 rpm give effective k's significantly larger than the "small" [100] crystals. Also "small" [111] crystals rotated at 144 rpm give effective k's smaller than "large" crystals pulled under similar conditions. The cause of this size effect is not clear. Perhaps it is related to the thermal gradient effects discussed by Jillson and Sheckler[33] and by Goss.[34] Another possible explanation is that the larger true growth rate ($\sim$20 per cent larger) for the "large" crystals would lead to a larger effective k as discussed by Burton et al.[17] However, at the low growth rates of 0.5 and 0.6 cm per hour this would be expected to be a small effect. This statement is supported by the fact that two "small" crystal ([111], 144 rpm) gave the same effective k's, within experimental error, even though the pull rates were 0.5 and 1 cm per hour. Other obvious factors that could be playing a role are orientation, rotation rate, shape of interface, etc. The value of 0.001 for k° given in Table I represents a compromise between the apparent upper and lower limits.*

The solidus curve at lower temperatures is also not clearly established. Hall[6] presented evidence that the solidus curve obtained by Thurmond and Kowalchik[24] was in error. The latter authors analyzed crystals of germanium recovered from slowly cooled indium-germanium melts. The author has obtained some relatively crude data from thermal gradient experiments which qualitatively confirm Hall's criticism of the earlier data. However, Hall's data do not appear to be just upper limits of the solid solubility as he suggested. Rather, in the author's opinion, they seem to be reasonably good estimates of the true solid solubility. Support for this opinion is given also by the work of John,[35] whose liquidus and resistivity data can be reinterpreted in the light of the liquidus data of Thurmond and Kowalchik[36] and using the resistivity data in Ref. 28. These data are summarized in Table III. The curve plotted in Fig. 1 is drawn to give a reasonable fit to Hall's data and to the data in Table III. Near the melting point of germanium only a few of the lower points from the crystal pulling experiments in Fig. 3 are plotted.†

* Recently Thurmond[87] has proposed an interesting model for crystal growth. A consequence of this model is the possibility that an effective distribution coefficient *lower* than the equilibrium value might be obtained at finite growth rates.

† The X-ray measurements of Mack[88] on indium-germanium alloy p-n junctions indicate an indium concentration of about 6×10^{18} atoms per cc in the regrowth layer. The temperature corresponding to this concentration is of necessity rather uncertain, but it is apparently in the range of from 400 to 600°C.

TABLE III — LOW-TEMPERATURE SOLUBILITY DATA ON THE INDIUM-GERMANIUM SYSTEM

Temperature, in °C	n_{In}^{S} (atom/cc)	Source	Remarks
~620	5.4×10^{18}	John[35]	from Hall measurement
~620	2.7×10^{18}	John[35]	from ρ measurement and Ref. 28
~700	3.8×10^{18}	John[35]	from Hall measurement
~700	2.0×10^{18}	John[35]	from ρ measurement and Ref. 28
~800	4.0×10^{18}	this work	from ρ measurement and Ref. 28
~510	3×10^{18}	this work	from X-ray measurement of lattice constant assuming Vegard's law and Pauling's tetrahedral radius for indium[37]
~300	$<1 \times 10^{19}$	this work	from ρ measurement and chemical analysis

Thallium. The value of 4×10^{-5} for k° is taken from Burton et al.[17] No other data are available.

Silicon. The value of 5.5 was obtained by Thurmond[38] from an analysis of the germanium-silicon phase diagram and from crystal pulling experiments. The solidus curve was determined by Stöhr and Klemm,[39] whose work has been confirmed in part by Hassion, Goss and Trumbore.[40] This system differs from all other known germanium and silicon systems in that a series of homogeneous solid solutions is formed. The reader is referred to Thurmond[41] for a plot of the phase diagram and a discussion of this system.

Tin. All of the tin data were taken from the work of Trumbore, Isenberg and Porbansky.[23] Their value of 0.020 for k° is in agreement with the work of Struthers quoted by Burton.[5]

Lead. The value of 1.7×10^{-4} for k° was obtained by the author from spectrophotometric analyses on four crystals pulled at 0.5 cm per hour. The results of these experiments are summarized in Table IV. One might speculate that the lower k value for the [100] crystal is due to an orientation effect, although more data are obviously needed. In addition, an analysis was made on one crystal grown by the thermal gradient tech-

TABLE IV — RESULTS OF CRYSTAL PULLING EXPERIMENTS ON THE LEAD-GERMANIUM SYSTEM

Growth direction	Rotation rate, in rpm	x_{Pb}^{L}	k
[100]	144	0.00259	1.61×10^{-4}
[111]	144	0.00296	1.76×10^{-4}
[111]	144	0.0106	1.74×10^{-4}
[111]	144	0.0322	1.48×10^{-4}

nique, indicating that, at 805 ± 10°C, $x^{S}_{Pb} = 8.9 \times 10^{-6}$. The solidus curve in Fig. 1 was constructed from these data assuming a simple solution model for the solid solutions as will be discussed in a subsequent paper.[1]

Phosphorus. The value of 0.080 for $k°$ was obtained by Hall,[7] who apparently pulled crystals from melts containing GaP and InP. This value compares with the earlier value of 0.12 due to Burton et al.[17]

Arsenic. The figure of 0.02 is due to Jillson and Sheckler[33] and compares with the value of 0.04 due to Burton et al.[17] and Hall.[31] The solidus curve in Fig. 1 is taken from Thurmond et al.[24] and is based on data obtained by the author from spectroscopic and spectrophotometric analyses of the germanium remaining after the evaporation of arsenic out of arsenic-germanium liquid alloys. These data are summarized in Table V. It should be noted that the evaporation technique requires considerable care and is very conducive to the production of occluded material. While there is no evidence to indicate that the data in Table V are not valid results, it would certainly be desirable to check these results by an independent method, e.g. by thermal gradient crystallization.

Antimony. The value of 3.0×10^{-3} for $k°$ in Table I is taken from Hall[7] and is in good agreement with the work of Burton et al.[17] and with the value of 3.3×10^{-3} (probably a value of k_e) obtained by Leverton,[30] whose value is probably slightly high because of the relatively large pull rate. This value of 3.0×10^{-3} is also consistent with the solidus curve derived from the tracer diffusion measurements of Thurmond and Kowalchik reported by Thurmond et al.[24]

Bismuth. The figure of 4.5×10^{-5} given for $k°$ in Table I is a compromise between the crystal pulling work of Burton et al.[17] and of Mortimer,[42] who obtained a value of 5×10^{-5} for crystals grown in a horizontal boat at $1\frac{1}{2}$ inches per hour. However, Mortimer also gives a figure of 0.23 ohm-cm for a section of a crystal pulled at $\frac{3}{4}$ inch per hour, corresponding to a melt concentration of 1.3 per cent bismuth. Assuming the latter concentration to be weight per cent, the value of k using Prince's mobility data[43] would be 4×10^{-5}. At the germanium melting point, $k°$

TABLE V — RESULTS OF EVAPORATION EXPERIMENTS ON THE ARSENIC-GERMANIUM SYSTEM

Temperature, in °C	x^{S}_{As}
905 ± 6	0.00155
853 ± 4	0.0032
818 ± 4	0.0049
757 ± 2	0.0041

would probably be somewhat higher than 4×10^{-5} but probably not 5×10^{-5}. No solidus data are available.

Sulfur, Selenium and Tellurium. Tyler[8] quotes maximum solid solubilities of sulfur, selenium and tellurium as $>5 \times 10^{15}$, $>5 \times 10^{15}$ and $>2 \times 10^{15}$ atoms/cc, respectively. A value for tellurium for k° of $\sim 10^{-6}$ is given by Tyler.[8] No other data are available.

Vanadium. The value of $>3 \times 10^{-7}$ for k° is from Woodbury and Tyler.[44] No other solidus data are available.

Manganese. The value of $\sim 10^{-6}$ for k° is from Woodbury and Tyler[44] and Tyler.[8] No other solidus data are available.

Iron. The value of 3×10^{-5} for k° was obtained by Bugay, Kosenko and Miseliuk,[19] who also determined the solidus curve. It should be noted that the temperature measurements of these authors appear to be somewhat doubtful, since their curve is extrapolated above the value of the melting point of germanium accepted in the present paper.

Cobalt. The value of $\sim 10^{-6}$ for k° is from the resistivity measurements of Tyler, Newman and Woodbury,[45] whose results agree with the tracer work of Burton et al.[17] A maximum solid solubility of $\sim 2 \times 10^{15}$ atoms/cc is quoted by Tyler,[8] but no other details are given. A reasonable first approximation to the solid solubility at lower temperatures is probably given by the iron-germanium curve in Fig. 1.

Nickel. The value of 3×10^{-6} for k° is from Tyler and Woodbury[46] and compares favorably with the earlier values of 5×10^{-6} due to Burton et al.[17] and 2.3×10^{-6} due to Tyler, Newman and Woodbury.[47] Solidus data have been obtained by Tyler and Woodbury,[8,46] while both solidus and solvus data were reported by van der Maesen and Brenkman.[48] In addition, Wertheim[49] has recently obtained solidus and solvus data from lifetime and conductivity measurements. Wertheim's solidus results are in good agreement with Tyler and Woodbury's solidus data, which are about a factor of two higher than the data of van der Maesen and Brenkman. The solidus curve in Fig. 1 favors Tyler and Woodbury's and Wertheim's data, while the solvus curve is from Wertheim, who is in agreement with van der Maesen and Brenkman where the data overlap.

Platinum. The value of $\sim 5 \times 10^{-6}$ for k° is due to Dunlap.[50] No other solidus data are available.

2.2 *Solid Solubilities in Silicon*

Lithium. Pell[4] has determined the solidus curve from flame analyses of diffused crystals in the range from about 592° to 1382°C. Pell pointed out that the earlier solidus data of Reiss, Fuller and Pietruszkiewicz[51] were in error, as confirmed later by Fuller and Reiss.[52] Pell estimated a

value of 0.010 for $k°$ by extrapolating his solidus data to the melting point of silicon. Aside from the uncertainty in $k°$ due to the extrapolation, Pell had to assume regular solution behavior to estimate the liquidus curve. Hence, a value of 0.01 is given in Table I. The solvus curve in Fig. 2 is taken from Fuller and Reiss,[51,52] whose data appear reliable below the eutectic temperature.*

Copper. The value of 4×10^{-4} for $k°$ is taken from the work of Struthers,[53,54] who obtained a value of 4.5×10^{-4} from tracer analysis of crystals pulled at rates of about 10 or 20 cm per hour. Since a large pull rate was used, it is likely that Struthers' value is high. The solidus and solvus curves in Fig. 2 are based on the data of Struthers,[55] whose work has been confirmed by Collins and Carlson.[56] The shape of the solvus curve in Fig. 2 is slightly different from Struthers' curve, since an attempt has been made here to fit his data differently, in order to take into account the eutectic point.

Gold. The value of 2.5×10^{-5} for $k°$ is due to Collins, Carlson and Gallagher[57] and compares favorably with earlier values of 3×10^{-5} due to Taft and Horn[58] and to Struthers as quoted by Burton.[5] The solidus and solvus curves in Fig. 2 are based on the combined measurements of Collins et al.[57] and of Struthers.[53] The earlier work of Struthers[55] on gold is in error.[53]

Zinc. The value of $\sim 1 \times 10^{-5}$ for $k°$ is from Hall's treatment[6] of the data of Fuller and Morin[59], but their results are divided by a factor of two, as suggested by Carlson.[60] A similar treatment of Fuller and Morin's data was used to estimate the solidus curve which was plotted to pass through the data from one radiotracer and two electrical measurements of Carlson.[60]

Boron. The value of 0.80 for $k°$ is quoted by Hall,[7] who refers to his earlier work[31] and that of Theuerer.[61] A value of 0.80 has also been obtained by Gould.[62] The work of Pearson and Bardeen,[63] when modified by the X-ray and density measurements of Horn,[64] may be interpreted as indicating a eutectic at roughly 3 to 7 degrees below the melting point, so that the solidus curve probably covers this limited temperature range. The solvus curve is also a very rough estimate based on a single point at 1200°C obtained by Howard[65] from diffusion measurements. Support for this curve has recently been obtained by Holonyak,[66] who found solubilities of $> 10^{20}$ atoms per cc in the range 700 to 800°C.†

* The eutectic point for the lithium-silicon system shown in Fig. 2 is Pell's value of 590 ± 10°C. More recent work by Böhm[89] indicates a value of 635 ± 10°C.

† It now appears that the sheet resistivity curves used to interpret the boron diffusion data at 1200°C were in error. Although the correct curves are not yet available, the point given in Fig. 2 is probably low, perhaps by as much as a factor of 2 to 3.

Aluminum. The $k°$ value of 2.0×10^{-3} is taken from Hall,[7] who found this value to be consistent with the amount of aluminum required to compensate a given amount of antimony in the melt. Support for this figure, which compares with the value of $>4 \times 10^{-3}$ quoted by Burton,[5] is found in some zone leveling experiments of Kolb and Tanenbaum.[67] The solidus curve has been the subject of a number of conflicting studies, the discrepancies being as high as three to four orders of magnitude in solid solubility at certain temperatures. R. C. Miller and Savage[68] have discussed critically the earlier works of Spengler[69] and Goldstein,[70] which represent the extreme values obtained for the solid solubility. More recently, Navon and Chernyshov[71] obtained a solidus curve from temperature gradient zone melting experiments which agree with Miller and Savage's scattered data at the higher temperatures. Similar agreement with the high-temperature data of Miller and Savage was obtained by Gudmundsen and Maserjian[72] by extrapolating their data obtained at lower temperatures in a study of the properties of regrowth layers. At low temperatures, however, there are appreciable discrepancies. From spectrophotometric and spectroscopic analyses on crystals grown in two thermal gradient experiments, the author obtained the results given in Table VI. S. L. Miller,[73] from capacitance measurements on p-n junctions, calculated a solubility of about 2.5×10^{19} atoms/cc at 800°C, a result that is more consistent with the data in Table VI. These figures are more than an order of magnitude larger than Navon and Chernyshov's solubilities and about a factor of 2 or 3 larger than the values of Gudmundsen and Maserjian at these temperatures. The discrepancy between these sets of data may be partially resolved by the use of Backenstoss' mobility data[74] extrapolated to higher impurity concentrations to interpret the resistivity data.[71,72] In constructing the solidus curve in Fig. 2, Gudmundsen and Maserjian's low-temperature curve, reinterpreted in this manner, was favored to represent a good compromise between the conflicting sets of data.

Gallium. The value of 8.0×10^{-3} for $k°$ is from the work of Hall,[7] who found this value to be consistent with the amount of gallium needed to

TABLE VI — RESULTS OF THERMAL GRADIENT EXPERIMENTS ON THE ALUMINUM-SILICON SYSTEM

Temperature, in °C	x^S_{Al}
715 ± 10	0.00029 (spectrophotometric analysis)
720 ± 10	0.00024 (spectrophotometric analysis)
720 ± 10	0.00030 (spectroscopic analysis)

TABLE VII — RESULTS OF THERMAL GRADIENT EXPERIMENTS ON THE GALLIUM-SILICON SYSTEM

Temperature, in °C	x^S_{Ga}
805 ± 10	3.0×10^{-4}
982 ± 10	5.2×10^{-4}
1066 ± 10	6.4×10^{-4}

compensate a given amount of antimony in the melt. The solidus curve in Fig. 2 was constructed using three results obtained by the author from spectrophotometric analyses of crystals grown at relatively low temperatures by a thermal gradient technique. These results are summarized in Table VII.

Indium. The value of 4×10^{-4} for $k°$ is taken from Hall[7] and compares with the value of 5×10^{-4} quoted by Burton.[5] No solidus data are available, although Backenstoss[74] did obtain a solution of 4×10^{17} atoms/cc in pulled crystals, which would indicate a higher value for the maximum solubility.

Germanium. The value of 0.33 for $k°$ was obtained by Thurmond[38] from crystal pulling experiments and an analysis of the germanium-silicon phase diagram.

Tin. The value of 0.016 for $k°$ and the solidus curve in Fig. 2 are based on the work of Trumbore, Isenberg and Porbansky.[23] The value of $k°$ compares with a figure of 0.02 due to Struthers quoted by Burton.[5]

Nitrogen. The value of $<10^{-7}$ for $k°$ is from the work of Kaiser and Thurmond.[75] It should be emphasized that this figure represents only electrically active nitrogen and is not valid if nitrogen is electrically inactive in silicon. In the author's opinion, such a low value for $k°$ seems rather unlikely in view of the correlations discussed later.

Phosphorus. The value of 0.35 for $k°$ is taken from Burton[5] and Hall,[7] and is supported by the work of James and Richards.[76] No phase diagram is available,* but solid solubility data have been obtained from diffusion measurements reported by Mackintosh.[77] These solubilities, however, might not represent the true solid solubilities because of the possibility that, in the diffusion experiments involving P_2O_5, the phosphorus may have been dissolved in a glassy SiO_2 phase. Hence, the curve in Fig. 2 should probably be considered a lower limit.

Arsenic. The value of 0.3 for $k°$ was taken from Burton.[5] The solidus

* Giessen and Vogel[90] have recently published a partial phase diagram for the silicon-phosphorus system. The silicon-rich liquidus curve was determined, and a Si-SiP eutectic temperature of 1131°C was measured.

curve in Fig. 2 was estimated from the scattered data obtained from capacitance measurements by Hassion and Russo.[78]

Antimony. The value of 0.023 for k° is from Hall,[7] who found it to be consistent with the amount of antimony required to compensate given amounts of aluminum and gallium in the melt. The solidus curve was estimated from some diffusion measurements of Rohan, Pickering and Kennedy[79] and from data obtained by the author from spectrophotometric analyses of crystals grown in three thermal gradient experiments. The latter results are summarized in Table VIII. A resistivity of ~0.0016 ohm-cm was found for one section of a crystal grown at 1066°C. Using the mobility data of Backenstoss,[74] the expected donor concentration would be $\sim 5 \times 10^{19}$ atoms per cc, in reasonable agreement with the results in Table VIII.

Bismuth. The value for k° of 7×10^{-4} is taken from a patent issued to Christian.[80] A rough estimate of the solidus curve is given, based on capacitance measurements of Hassion and Russo.[78]

Oxygen. The value of 0.5 for k° was obtained by Thurmond[38] from a vacuum fusion gas analysis on a quenched silicon sample which was melted in a silica tube. The solvus curve is based on Hrostowski and Kaiser's work.[81]

Sulfur. The value of 10^{-5} for k° and the solid solubility curve were taken from Carlson, Hall and Pell.[82] Since no germanium-sulfur phase diagram is available, it is not known whether this is a solidus and/or a solvus curve.

Manganese. The value of $\sim 10^{-5}$ for k° is from Carlson.[83] One tracer measurement obtained by Carlson at 1200°C was used to obtain a rough estimate of the solidus curve.

Iron. The value of 8×10^{-6} for k° is taken from the work of Collins and Carlson,[56] who obtained a value of 6×10^{-6} from tracer measurements and of between 5 and 10×10^{-6} from electrical measurements. These results are in accord with the tracer measurements of Struthers,[53]

TABLE VIII — RESULTS OF THERMAL GRADIENT EXPERIMENTS ON THE ANTIMONY-SILICON SYSTEM

Temperature, in °C	x^S_{Sb}
807 ± 10	6.2×10^{-4}
980 ± 10	5.6×10^{-4}
991 ± 10	8.1×10^{-4}
1066 ± 10	6.2×10^{-4}
1066 ± 10	9.3×10^{-4}

who has obtained a value of $\sim 10^{-5}$ for $k°$. The only solubility measurements above the eutectic temperature appear to be the tracer data of Struthers.[53,55] Struthers' data at the higher temperatures appear to be consistent with an extrapolation of the data of Collins and Carlson below the eutectic temperature. However, Struthers' data below the eutectic temperature disagree with Collins' and Carlson's data by as much as 2 or 3 orders of magnitude. It would appear that Struther's experiments, which include the "saturation" of silicon with iron as well as the calculation of surface concentrations from short-time diffusion experiments, should rule out any complications due to slow and fast diffusing species, as suggested by Collins and Carlson. However, as pointed out by Collins and Carlson who checked Struthers' tracer measurements at 1200°C, the amount of iron used in the tracer experiments might have been insufficient in both sets of experiments to obtain the equilibrium solubility. Accordingly, the solvus curve in Fig. 2 has been arbitrarily drawn to favor Collins' and Carlson's data although, in the author's opinion, the discrepancy remains unresolved.

Cobalt. The value of 8×10^{-6} for $k°$ was obtained by Collins and Carlson,[56] who also obtained a solid solubility of 1×10^{16} atoms/cc at 1200°C using tracer techniques. The curve plotted in Fig. 2 is an estimate based on these two figures.

Tantalum. The value of 10^{-7} for $k°$ is taken from Burton.[5] No other data are available.

Silver, Cadmium, Palladium. Collins and Carlson[56] state that the solid solubilities of these elements at 1200°C are from 10^{15} to 3×10^{16} atoms/cc as determined from tracer measurements. No other data are available.

2.3 *General Comments*

In the above discussion the author has made no attempt to assess the absolute accuracy of the experimental data. The temptation to do this was tempered by the fact that over the years there has been a pronounced tendency for the "accepted" equilibrium solid solubilities to decrease, even by orders of magnitude in certain cases. This tendency is quite understandable, in view of the recent development of more refined electrical and chemical techniques for determining impurity concentrations and of better techniques for the growth of single crystals free from occlusions and other imperfections. At the present time, it is likely that, even for the most carefully investigated systems, the accuracy of the data is no better than ± 10 to 20 per cent. Indeed, agreement to within a factor of two for different investigations is often considered good. While

for most semiconductor device applications such accuracy is sufficient, for thermodynamic studies it would be highly desirable to obtain accuracies of even better than ± 10 per cent, if possible.

For further progress it seems that two general areas of experiment would be especially profitable. First, more work is needed on the direct correlation of chemical, tracer or other direct measurements of impurity concentration with resistivity and Hall effect measurements. Such work is particularly desirable in concentration ranges where the semiconductor is degenerate and where the relations between the Hall coefficient and the carrier concentration are in doubt. Second, more accurate work is needed on the effect of various crystal growth parameters on the solid solubility. The results for the indium-germanium system plotted in Fig. 3 appear to show an effect due to size, orientation, rotation rate or perhaps some other factor not immediately obvious. The very interesting work of Jillson and Sheckler[33] and of Goss[34] indicates possible effects due to thermal gradients, rotation, shape of interface, etc., on the effective distribution coefficient. Such studies, together with earlier works,[17,31] indicate that the accurate determination of equilibrium solid solubilities is subject to considerable complications. Along similar lines, it would be very worthwhile to carry out further experiments similar to those of John[35] and compare the crystal pulling and thermal gradient data at low temperatures.

III. CORRELATION OF SOLID SOLUBILITIES

The extent to which a solute element will dissolve in a solid solvent element is determined by the thermodynamic requirement that the compositions of the resulting solid solution and the coexisting solid, liquid or gaseous phase(s) must be such as to minimize the free energy of the system. Considerable work has been done on the correlation of the extent of primary solid solution with various properties of the solute and solvent elements, e.g. with atom size, valence, electronegativity, crystal structure, etc. (See, for example, Darken and Gurry.[84]) In such correlations a common practice is to compare maximum solid solubilities for various solutes as a function of the property in question. However, since the solid solubilities depend on the interactions between the solid solution and other phase(s), the maximum solid solubilities are not necessarily parameters that accurately indicate the relative compatibilities of the solute elements with the solvent element in solid solution.*

* For example, if the bonds between a solute element, A, and a solvent, B, are stronger than A–A or B–B bonds, one might expect a solid solution of A in B to be relatively stable with a resulting high solubility of A in B. However, the relative

In the case of germanium and silicon systems, we are mainly concerned with the equilibrium between solidus and liquidus alloys. Here, the distribution coefficient is a desirable parameter for comparing the relative tendencies of various impurities to dissolve in solid germanium or silicon, since the effect of concentration in the liquid phase is taken into account. Thus, at the melting point of germanium or silicon the values of k° can be thought of as giving the relative solid solubilities of the impurities, each at the same constant concentration (near infinite dilution) of impurity in the liquid phase and at essentially constant temperature.

Although the effect of liquid phase concentration is taken into account by using k as a solid solubility parameter, no account is taken of the effect of nonideal liquid solution behavior, i.e., of departures from Raoult's law. As discussed previously,[23] such departures can be taken into account by defining a new parameter

$$k' = \frac{x_S}{a_L} = \frac{k}{\gamma_L} = \frac{1}{\gamma_S} = \exp\left(\frac{\mu_L^\circ - \mu_S}{RT}\right),$$

where a_L, γ_L and γ_S are the activity and activity coefficients, respectively, of the impurity in the liquidus and solidus alloys, based on a standard state of the pure liquid impurity element, and μ_L° and μ_S are the chemical potentials of the impurity in the pure liquid impurity and in the solidus alloy, respectively. The parameter k' is used here instead of $1/\gamma_S$ because of its similarity in form to the distribution coefficient. The parameter $k^{\circ\prime}$ may also be considered as a solubility, in the sense that it represents the atom fraction of impurity in a hypothetical solid solution in equilibrium with the pure liquid impurity element.† As expected, k' is larger the more stable the solid solution, i.e., the smaller μ_S is compared to μ_L°.

Let us now examine the use of k and k' in correlating solid solubilities with atom sizes. It was pointed out by Burton et al.[17] that a rough correlation exists between the tetrahedral radius of an impurity atom

strength of the A–B bond might also stabilize an intermetallic compound of A and B of a different crystal structure and decrease the amount of primary solid solution.[84] In such a case, the solid solubility itself does not accurately represent the stability of the solid solution relative to the pure components or, perhaps, relative to the solid solution of another solute element where no intermetallic compound is formed.

† This is a hypothetical solid solution in the sense that it must be assumed that γ_S remains constant even when the solid solution is no longer infinitely dilute, i.e., Henry's law must be obeyed. (The parameter $k^{\circ\prime}$ may be considered a Henry's law constant.) Although this assumption may seldom be valid, $k^{\circ\prime}$ is still a good measure of the stability of the infinitely dilute solid solution relative to the pure liquid impurity. It is in the dilute solution range where one can best obtain information about the basic solute-solvent interaction.

and its distribution coefficient in germanium at the melting point of germanium. In recent years a considerable body of additional data, including some substantial revisions of the earlier data, has been accumulated. Hence, it seems worthwhile to reconsider the situation for both germanium and silicon. In Figs. 4 and 5 the distribution coefficient at the germanium or silicon melting point is plotted as a function of the tetrahedral radius[37] of the impurity element.

For the case of germanium it is seen that elements from various groups of the periodic table tend to lie on different smooth curves. As expected there is a trend toward lower solubility as the radius of the impurity atom increases, with a relatively rapid decrease in $k°$ in the vicinity of 1.35 to 1.5Å. This decrease is in the neighborhood of a 15 per cent size difference between solute and solvent, where solid solubility becomes restricted in other alloy systems.[84] In the case of silicon, where the data

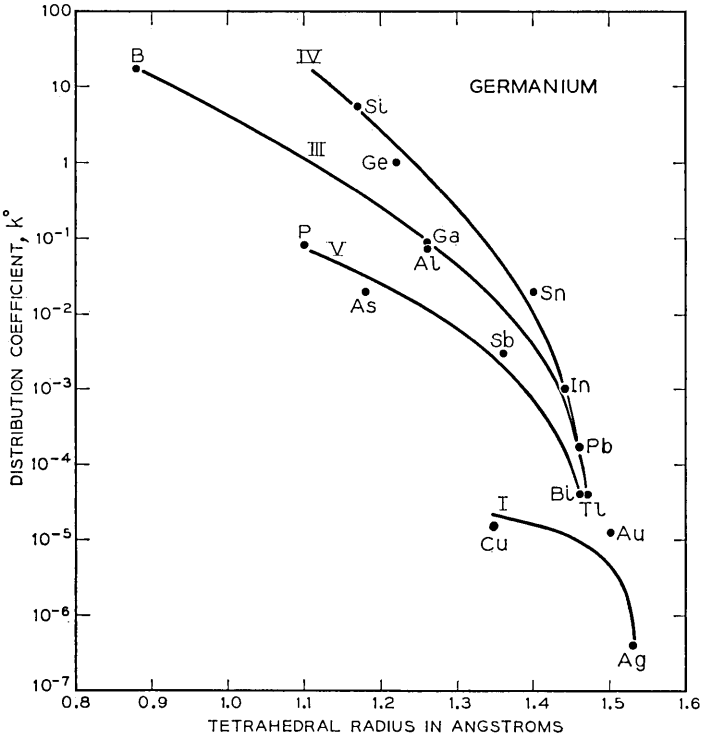

Fig. 4 — Distribution coefficients of impurities at the melting point of germanium as a function of the tetrahedral radii.

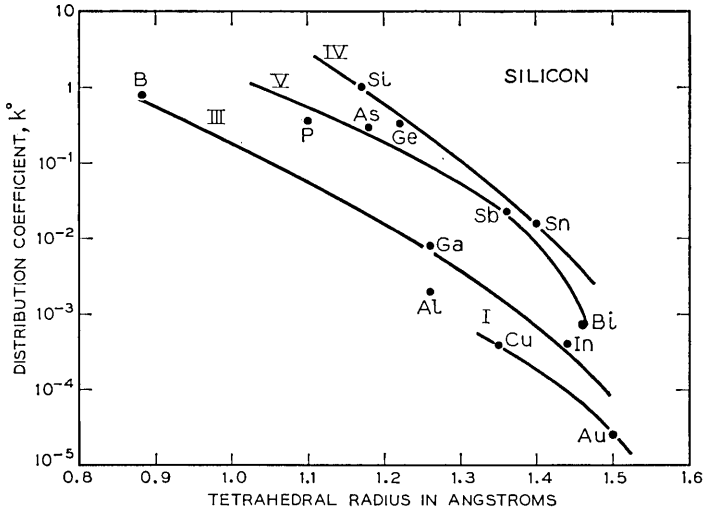

Fig. 5 — Distribution coefficients of impurities at the melting point of silicon as a function of the tetrahedral radii.

are fewer, similar trends are observed. A number of interesting differences should be noted, however. In germanium the relative order of $k°$ of elements from different groups is roughly IV > III > V, while for silicon the order is IV > V > III. Also, for the silicon case the discrepancy in size, i.e., the difference in tetrahedral radii of silicon and the impurity elements, for elements with a radius greater than 1.22Å, is greater than in the germanium systems. Yet there is little indication of the relatively sharp decrease in solubility found for germanium at 1.35 to 1.5Å. The elements gallium and aluminum are rather interesting in that they appear to have about the same tetrahedral radius and are in the same group of the periodic table. One might, therefore, expect them to have about the same distribution coefficients. While this is approximately true in the case of germanium, an appreciable difference is noted for the case of silicon.

Let us turn now to the consideration of correlations involving $k°'$. Unfortunately, there are no experimental data on values of γ_L° that must be evaluated to obtain $k°'$. However, the liquidus curve treatment discussed by Thurmond and Kowalchik[36] may be used to estimate the values of γ_L° to a first approximation. From these estimates of γ_L°, estimates of $k°'$ have been made for those elements treated by Thurmond. These values are plotted against the tetrahedral radii in Figs. 6 and 7 for germanium and silicon.

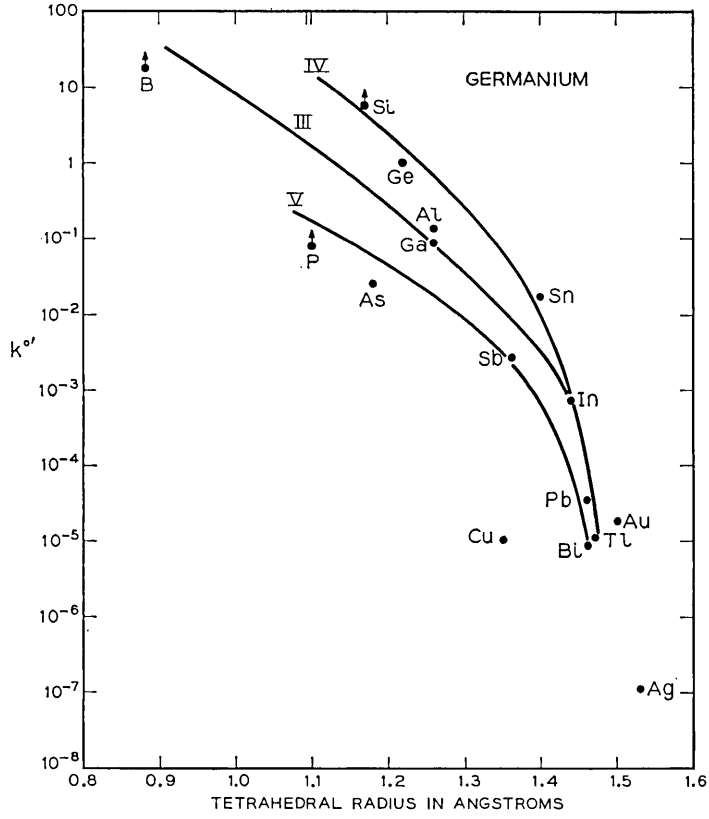

Fig. 6 — Plot of the parameter $k^{\circ\prime}$ versus the tetrahedral radii of the impurity elements for germanium. The arrows indicate the general direction of $k^{\circ\prime}$ where liquidus curve data are not available to estimate a value of $k^{\circ\prime}$.

For germanium there is not too much difference between the plots of k° and $k^{\circ\prime}$, due to the relatively small departures from ideality in the liquid phase. However, for silicon the shape of the curves is altered considerably by the use of $k^{\circ\prime}$. Although the data are relatively fewer, there is now evidence of a sharp decrease in solubility at the higher values of the tetrahedral radii, as in the case of germanium. It is also apparent that, in the case of silicon, the difference between gallium and aluminum has been altered significantly. A rather disturbing feature of some of the correlations is the anomalous behavior of gold, which does not fit smoothly into the pattern observed for the other elements. Aside from possible errors in k° or in the tetrahedral radius, it is possible that gold is substantially interstitial in the germanium or silicon lattice.

Another interesting feature of these correlations is that boron appears to fit relatively smoothly into the correlation, even though the difference in tetrahedral radii is very large, on the order of 25 to 30 per cent. Because of this difference, one might except a considerable reduction of solid solubility. If it is assumed that carbon and nitrogen behave similarly and fit on the curves for groups IV and V, respectively, the distribution coefficients of these elements should also be large, on the order of unity or perhaps significantly larger. However, in the case of nitrogen, Kaiser and Thurmond[75] suggest that k° may be less than 10^{-7} in silicon, at least as far as electrically active species is concerned.

In treating the solid solubilities of tin in germanium and silicon, the distribution coefficient or k' was used to calculate the binding energies of tin in the semiconductor.[23] The binding energies were then related to the bond energies in pure tin, since the bond energy for gray tin is simply half the heat of sublimation to the monomeric vapor species. It was thought that a simple correlation might be expected to exist between solid solubilities and heats of sublimation of impurity elements of group IV and perhaps other groups of the periodic table. A theoretical treatment relating the distribution coefficient to bond energies and strain

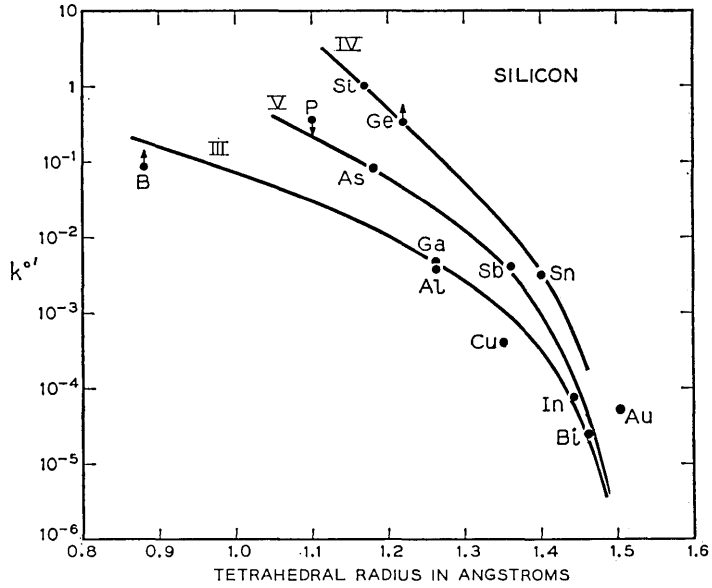

Fig. 7 — Plot of the parameter $k^{\circ\prime}$ versus the tetrahedral radii of the impurity elements for silicon. The arrows indicate the general direction of $k^{\circ\prime}$ where liquidus curve data are not available to estimate a value of $k^{\circ\prime}$.

energies has recently been given by Weiser.[85] This treatment would also lead one to expect some sort of correlation between $k°$, $k°'$ and heats of sublimation or atom sizes. A rather striking correlation was found, especially for germanium, as is evident from Figs. 8 through 11, which are plots of $k°$ and $k°'$ as functions of the heats of sublimation of the solute elements. The heats of sublimation were taken principally from Honig.[86] No attempt was made to correct these heats to the melting points of germanium and silicon, since the form of the correlation would

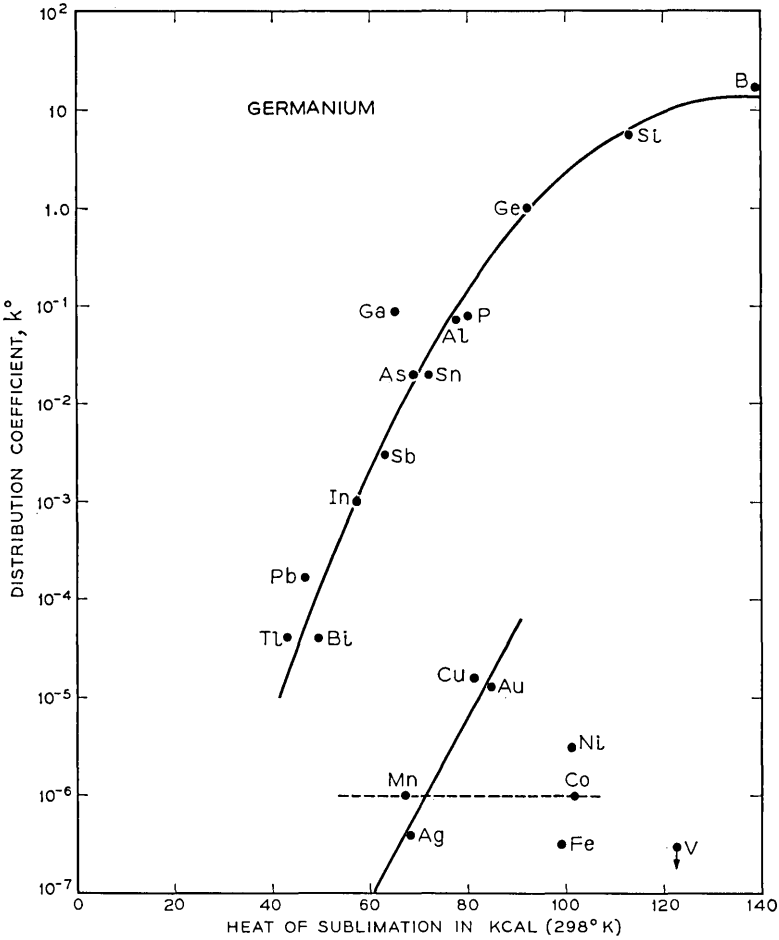

Fig. 8 — Distribution coefficients of impurity elements at the melting point of germanium as a function of the heats of sublimation of the impurities to the monomeric vapor species at 298°K.

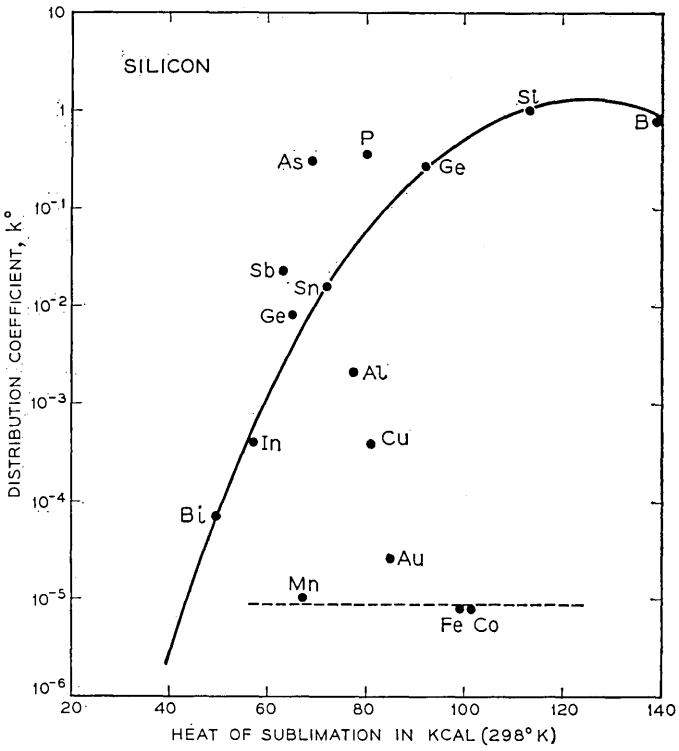

Fig. 9 — Distribution coefficients of impurity elements at the melting point of silicon as a function of the heats of sublimation of the impurities to the monomeric vapor species at 298°K.

not be affected significantly. It should be emphasized that these heats refer to the sublimation of the impurity element to the monomeric vapor species, not to the equilibrium vapor.

The most numerous and probably the most reliable data are on the germanium alloy systems. The outstanding feature of the correlation of both k° and $k^{\circ\prime}$ is the fact that, for germanium, elements of groups III, IV and V, with the exception of gallium, fall quite close to the same smooth curve. It is also seen that the copper-silver-gold and transition metal series fall into separate groups, with the latter group showing no evident trend of solubility with heat of sublimation.

In the case of silicon, where the data are fewer and probably less reliable, the plot of k° bears only a qualitative resemblance to the behavior found for germanium. Turning to $k^{\circ\prime}$, the situation improves somewhat, and one finds that the same general trends are present as in

germanium, although the fit of group III, IV and V elements to a single curve is questionable. The point for gold again seems anomalous, as was the case for the tetrahedral radius correlation.

The above correlations have involved the use of high-temperature data and theoretical assumptions about the nature of the liquid solu-

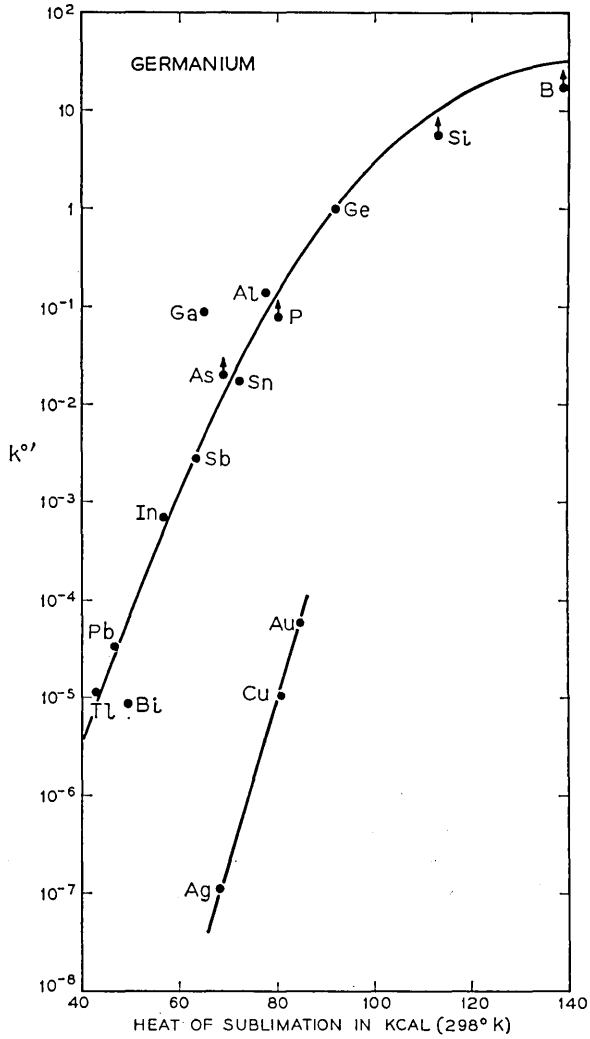

Fig. 10 — Plot of the parameter $k^{\circ\prime}$ for impurity elements in germanium versus the heats of sublimation of the impurities to the monomeric vapor species at 298°K. The arrows indicate the general direction of $k^{\circ\prime}$ where liquidus curve data are not available to estimate a value of $k^{\circ\prime}$.

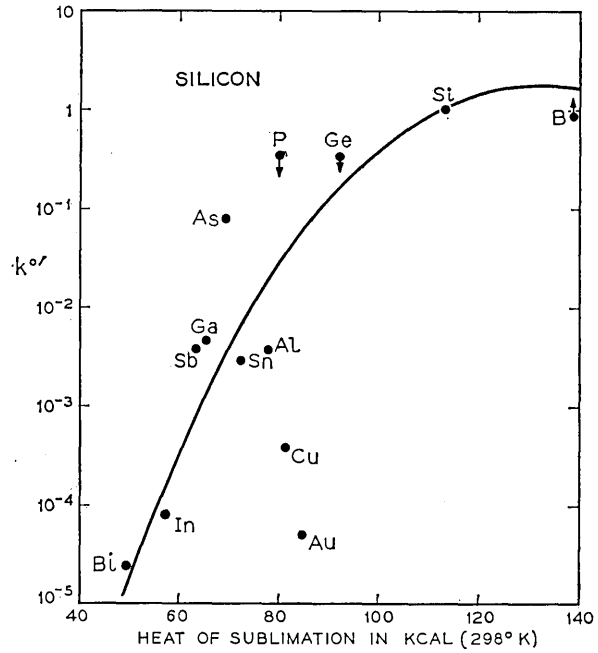

Fig. 11 — Plot of the parameter $k^{\circ\prime}$ for impurity elements in silicon versus the heats of sublimation of the impurities to the monomeric vapor species at 298°K. The arrows indicate the general direction of $k^{\circ\prime}$ where liquidus curve data are not available to estimate a value of $k^{\circ\prime}$.

tions. It is to be hoped that sufficient reliable data at lower temperatures can be obtained so that a comparison can be made in regions of the liquidus curves with smaller departures from ideality. Experimental work on activities of the liquidus alloys would be of great importance in interpreting the solid solubility data.

IV. ACKNOWLEDGMENTS

The author gratefully acknowledges the assistance of A. A. Tartaglia, E. M. Porbansky and M. Kowalchik, who carried out the thermal gradient and crystal pulling experiments. Special thanks are due the analytical chemistry department of Bell Telephone Laboratories, particularly Mrs. M. E. Bott and C. L. Luke for the spectrophotometric analyses and D. Babusci for the spectroscopic analyses. The X-ray measurement on the indium-germanium crystal was made by W. L. Bond. A number of workers at Bell Telephone Laboratories and elsewhere have been most helpful in permitting the use of unpublished data. The author also

wishes to acknowledge the helpful comments of C. D. Thurmond and M. Tanenbaum.

REFERENCES

1. Trumbore, F. A. and Thurmond, C. D., to be published.
2. Hassion, F. X., Thurmond, C. D. and Trumbore, F. A., J. Phys. Chem., **59,** 1955, p. 1076.
3. Olette, M. Compt. rend., **244,** 1957, p. 1033.
4. Pell, E. M., J. Phys. Chem. Solids., **3,** 1957, p. 77.
5. Burton, J. A., Physica, **20,** 1954, p. 845.
6. Hall, R. N., J. Phys. Chem. Solids., **3,** 1957, p. 63.
7. Hall, R. N., in *Fortschritte der Hochfrequenztechnik*, Akademische Verlagsgesellschaft, M.B.H., Frankfurt am Main (to be published); also General Electric Research Lab. Report No. 58-RL-1874.
8. Tyler, W. W., J. Phys. Chem. Solids, **8,** 1959, p. 59.
9. Pell, E. M., J. Phys. Chem. Solids, **3,** 1957, p. 74.
10. Reiss, H., Fuller, C. S. and Morin, F. J., B.S.T.J., **35,** 1956, p. 535.
11. Reiss, H. and Fuller, C. S., J. Phys. Chem. Solids, **4,** 1958, p. 58.
12. Woodbury, H. H. and Tyler, W. W., Phys. Rev., **105,** 1957, p. 84.
13. Finn, G., Phys. Rev., **91,** 1953, p. 754.
14. Hodgkinson, R. J., Phil. Mag., **46,** 1955, p. 410.
15. Fuller, C. S., Struthers, J. D., Ditzenberger, J. A. and Wolfstirn, K. B., Phys. Rev., **93,** 1954, p. 1182.
16. Wolfstirn, K. B. and Fuller, C. S., J. Phys. Chem. Solids, **7,** 1958, p. 141.
17. Burton, J. A., Kolb, E. D., Slichter, W. P. and Struthers, J. D., J. Chem. Phys., **21,** 1953, p. 1991.
18. Bugay, A. A., Kosenko, V. E. and Miseliuk, E. G., Zh. tekh. fiz. (Moscow), **27,** 1957, p. 1671.
19. Bugay, A. A., Kosenko, V. E. and Miseliuk, E. G., Zh. tekh. fiz. (Moscow), **27,** 1957, p. 210.
20. Dunlap, W. C., Jr., Phys. Rev., **97,** 1955, p. 614.
21. Tyler, W. W. and Woodbury, H. H., Phys. Rev., **102,** 1956, p. 647.
22. Trumbore, F. A., J. Electrochem. Soc., **103,** 1956, p. 597.
23. Trumbore, F. A., Isenberg, C. R. and Porbansky, E. M., J. Phys. Chem. Solids, **9,** 1959, p. 60.
24. Thurmond, C. D., Trumbore, F. A. and Kowalchik, M., J. Chem. Phys., **25,** 1956, p. 799.
25. Woodbury, H. H. and Tyler, W. W., Bull. Amer. Phys. Soc., **1,** 1956, p. 127.
26. Bridgers, H. E. and Kolb, E. D., J. Chem. Phys., **25,** 1956, p. 648.
27. Trumbore, F. A., Porbansky, E. M. and Tartaglia, A. A., to be published.
28. Trumbore, F. A. and Tartaglia, A. A., J. Appl. Phys., **29,** 1958, p. 1511.
29. Bridgers, H. E., private communication.
30. Leverton, W. F., J. Appl. Phys., **29,** 1958, p. 1241.
31. Hall, R. N., J. Phys. Chem., **57,** 1953, p. 836.
32. Dowd, J. J. and Rouse, R. L., Proc. Phys. Soc. (London), **66B,** 1953, p. 60.
33. Jillson, D. C. and Sheckler, A. C., Phys. Rev., **98,** 1955, p. 229.
34. Goss, A. J., Marconi Rev., **22,** 1959, p. 18.
35. John, H., J. Electrochem. Soc., **105,** 1958, p. 741.
36. Thurmond, C. D. and Kowalchik, M., this issue, p. 169.
37. Pauling, L., *The Nature of the Chemical Bond,* Cornell Univ. Press, Ithaca, N. Y., 1945, p. 179.
38. Thurmond, C. D., private communication.
39. Stöhr, H. and Klemm, W., Z. anorg. allgem. Chem., **241,** 1935, p. 305.
40. Hassion, F. X., Goss, A. J. and Trumbore, F. A., J. Phys. Chem., **59,** 1955, p. 1118.
41. Thurmond, C. D., J. Phys. Chem., **57,** 1953, p. 827.
42. Mortimer, G., J. Electrochem. Soc., **105,** 1958, p. 731.
43. Prince, M. B., Phys. Rev., **92,** 1953, p. 681.
44. Woodbury, H. H. and Tyler, W. W., Phys. Rev., **100,** 1955, p. 659.

45. Tyler, W. W., Newman, R. and Woodbury, H. H., Phys. Rev., **96,** 1954, p. 874.
46. Tyler, W. W. and Woodbury, H. H., Bull. Amer. Phys. Soc., **2,** 1957, p. 135.
47. Tyler, W. W., Newman, R. and Woodbury, H. H., Phys. Rev., **98,** 1955, p. 461.
48. van der Maesen, F. and Brenkman, J. A., Phillips Res. Rep., **9,** 1954, p. 225.
49. Wertheim, G. K., Phys. Rev., **115,** 1959, p. 37.
50. Dunlap, W. C., Phys. Rev., **96,** 1954, p. 40.
51. Reiss, H., Fuller, C. S. and Pietruszkiewicz, A. J., J. Chem. Phys., **25,** 1956, p. 650.
52. Fuller, C. S. and Reiss, H., J. Chem. Phys. **27,** 1957, p. 318.
53. Struthers, J. D., private communication.
54. Thurmond, C. D. and Struthers, J. D., J. Phys. Chem., **57,** 1953, p. 831.
55. Struthers, J. D., J. Appl. Phys. **27,** 1956, p. 1560.
56. Collins, C. B. and Carlson, R. O., Phys. Rev., **108,** 1957, p. 1409.
57. Collins, C. B., Carlson, R. O. and Gallagher, C. J., Phys. Rev., **105,** 1957, p. 1168.
58. Taft, E. A. and Horn, F. H., Phys. Rev., **93,** 1954, p. 64.
59. Fuller, C. S. and Morin, F. J., Phys. Rev., **105,** 1957, p. 379.
60. Carlson, R. O., Phys. Rev., **108,** 1957, p. 1390.
61. Theuerer, H. C., Trans. A.I.M.E., **206,** 1956, p. 1316.
62. Gould, J. R., unpublished data.
63. Pearson, G. L. and Bardeen, J., Phys. Rev., **75,** 1949, p. 865.
64. Horn, F. H., Phys. Rev., **97,** 1955, p. 1521.
65. Howard, B. T., private communication.
66. Holonyak, N., Jr., unpublished data.
67. Kolb, E. D. and Tananbaum, M., J. Electrochem. Soc., **106,** 1959, p. 597; also Kolb, E. D., private communication.
68. Miller, R. C. and Savage, A., J. Appl. Phys., **27,** 1956, p. 1430.
69. Spengler, H., Metall., **9,** 1955, p. 181.
70. Goldstein, B., Bull. Amer. Phys. Soc., **1,** 1956, p. 145.
71. Navon, D. and Chernyshov, V., J. Appl. Phys., **28,** 1957, p. 823.
72. Gudmundsen, R. A. and Maserjian, J., Jr., J. Appl. Phys., **28,** 1957, p. 1308.
73. Miller, S. L., private communication.
74. Backenstoss, G., Phys. Rev., **108,** 1957, p. 416.
75. Kaiser, W. and Thurmond, C. D., J. Appl. Phys., **30,** 1959, p. 427.
76. James, J. A. and Richards, D. H., J. Elect. & Cont., **3,** 1957, p. 500.
77. Mackintosh, I. M., to be published.
78. Hassion, F. X. and Russo, L. J., private communication.
79. Rohan, J. J., Pickering, N. E. and Kennedy, J., J. Electrochem. Soc., **106,** 1959, p. 705.
80. Christian, S. M., U. S. Patent No. 2,820,185.
81. Hrostowski, H. J. and Kaiser, R. H., J. Phys. Chem. Solids, **9,** 1959, p. 214.
82. Carlson, R. O., Hall, R. N. and Pell, E. M., J. Phys. Chem. Solids, **8,** 1959, p. 81.
83. Carlson, R. O., Phys. Rev., **104,** 1956, p. 937.
84. Darken, L. S. and Gurry, R. W., *Physical Chemistry of Metals,* McGraw-Hill Book Co., New York, 1953, Ch. 4.
85. Weiser, K., J. Phys. Chem. Solids, **7,** 1958, p. 118.
86. Honig, R. E., R.C.A. Rev., **28,** 1957, p. 195.
87. Thurmond, C. D., in *Semiconductors,* Hannay, N. B., ed., Reinhold Publishing Corp., New York, 1959, Ch. 4.
88. Mack, G., Z. Physik, **152,** 1958, p. 26.
89. Böhm, H., Z. Metall., **50,** 1959, p. 44.
90. Giessen, B. and Vogel, R., Z. Metall., **50,** 1959, p. 274.

Pushbutton Calling with a Two-Group Voice-Frequency Code

By L. SCHENKER

A customer voice-frequency pushbutton signaling scheme involving a new two-group signal code, termed "four-by-four", is described. It is shown that the use of this code, with judiciously chosen frequencies, permits the detection of bona fide signals after transmission over any ordinary voice connection and facilitates discrimination against false signals resulting from speech. Apparatus for generating the tones is described, and the principles of reception are discussed.

I. INTRODUCTION

In recent years, customer signaling from a telephone set by means of pushbuttons has appeared increasingly promising. An early plan has been reported[1] in which the operation of a pushbutton produces four effects: (a) the generation of a damped oscillatory wave at one of ten digit-identifying frequencies within the voice range; (b) the generation of a similar damped wave at one of eight party-identifying frequencies, also in the voice range; (c) a stepwise reduction of the direct current drawn by the set and (d) the temporary disablement of the speech transmitter. Actions (c) and (d) provide protection against talk-off, i.e., false signaling due to speech or noise at the transmitter.

This plan contemplated only the transmission of information to the central office, but in the long run it would be advantageous to be able to signal "end-to-end," over any established connection that will transmit speech. With this added objective several new considerations are introduced. Clearly, the signals should not contain an out-of-band component such as the dc step. Again, there may be a need for more than ten distinct signals, and these may be difficult to provide in a scheme based on a one-out-of-N code, for a new frequency must be added for each new signal. On the other hand, party identification would not be a feature of end-to-end signaling. Lastly, sustained rather than damped signals are strongly preferred for end-to-end signaling, in order to main-

tain adequate signal-to-noise margins despite wider ranges of transmission loss.

Voice-frequency signaling as such is, of course, already an established practice in the Bell System. One example is multifrequency key pulsing[2] (MFKP), which is widely used for toll signaling. However, minimization of talk-off was not a factor in its development, and tolerance of considerable carrier shift was one of the prime objectives. Another example is in-band single-frequency signaling,[3] which is used in the long-distance telephone plant for transmission of supervisory and dial signals.

This paper describes an all-voice-frequency signaling system that provides substantial protection against talk-off. It is based on a two-group arrangement commonly called the "four-by-four" code. Field and laboratory tests of the signaling system have been encouraging, and user reaction to pushbutton operation during a trial involving some 400 customers has been favorable.

II. CHOICE OF CODE

When only voice frequencies are employed, protection against talk-off must rely heavily on statistical tools. This protection is required only during interdigital intervals; speech interference with valid signals is conveniently avoided by transmitter disablement. Since signals with a simple structure are prone to frequent imitation by speech and music, some form of multifrequency code particularly difficult of imitation is indicated. If the signal frequencies are restricted in binary fashion to being either present or absent, the greatest economy in frequency space results from the use of all combinations of N frequencies, yielding $n = 2^N$ different signals. However, some of these are no more than single frequencies and are therefore undesirable from the standpoint of talk-off. Another drawback is that as many as N frequencies must be transmitted simultaneously; these involve an N-fold sharing of a restricted amplitude range, and may also be costly to generate. If $n > 10$, N would need to be at least four. At the expense of more frequency space we are led to a P-out-of-N code, yielding $n = N!/P!(N - P)!$ combinations. There is statistical advantage in knowing the number, P, of components in all valid signals.

In order to minimize the number of circuit elements,* as well as to reduce the sharing of amplitude range, P should be as small as possible, yet be larger than unity for the sake of talk-off protection. Let us then examine codes in which $P = 2$. If one can be found that is not readily

* It is here assumed that the P components are simultaneously produced in P resonant circuits, at least one of which can be tuned to each of the N possible frequencies.

imitated by speech or music, there is no merit in choosing P higher than two, provided that the total number of frequencies N needed for the required number of combinations can be accommodated in the available frequency spectrum. With $P = 2$, N must be at least five to provide ten combinations (for the ten digits). With $N = 6$, 15 combinations are available. The familiar MFKP signaling scheme makes use of a two-out-of-six code.

There are advantages, as we shall see, in imposing the further restriction that, with $P = 2$, the frequencies for each combination fall respectively in two mutually exclusive frequency bands. If, for example, 15 combinations are required, N must be at least eight (giving 16 combinations). In the four-by-four code, eight signal frequencies are divided into two groups: group A, the lower four frequencies, and group B, the upper four. Each signal is composed of one frequency from group A and one from group B.

III. BAND SEPARATION AND LIMITER ACTION

With a group arrangement, it is possible at the receiver to separate the two frequencies of a valid signal by band filtering before amplification or determination of the specific components. This separation of the two components of a signal renders reliable discrimination between valid signals and speech or noise simpler for two reasons: (a) each component can be regulated separately, thus compensating for "slope," the differential transmission loss incurred by the two components, and (b) an instantaneous extreme "limiting" can be applied to each component after band separation, and thus provide a substantial guard action.

It is a characteristic of extreme instantaneous limiters that they accentuate differences in levels between the components of a multifrequency signal. This may be used to provide guard action, that is, action to reduce the probability of false response to speech or other unwanted signals. Fig. 1(a) shows the input and output of such a limiter with an input signal composed of two frequencies. The lower frequency (dotted line) has the larger input amplitude. Assuming infinite gain, the limiter output may be constructed from the axis crossings of the input. The higher frequency produces some interference, but the low frequency dominates in the output. In this example, noise could be substituted for the higher frequency. In Fig. 1(b) the higher frequency has the larger amplitude and dominates in the output. Quantitative relationships are shown in Fig. 2, where the abscissa is the ratio at the limiter input of the power in the interference to the power at the wanted frequency, and the ordinate is the power at the wanted frequency at the limiter output relative to its value in the absence of any interference at the input. Curve 1

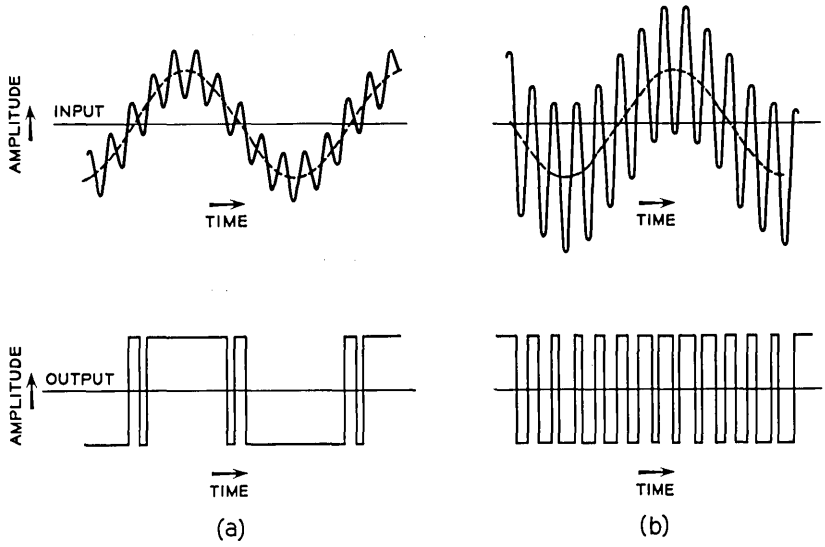

Fig. 1 — Instantaneous output for two-frequency input.

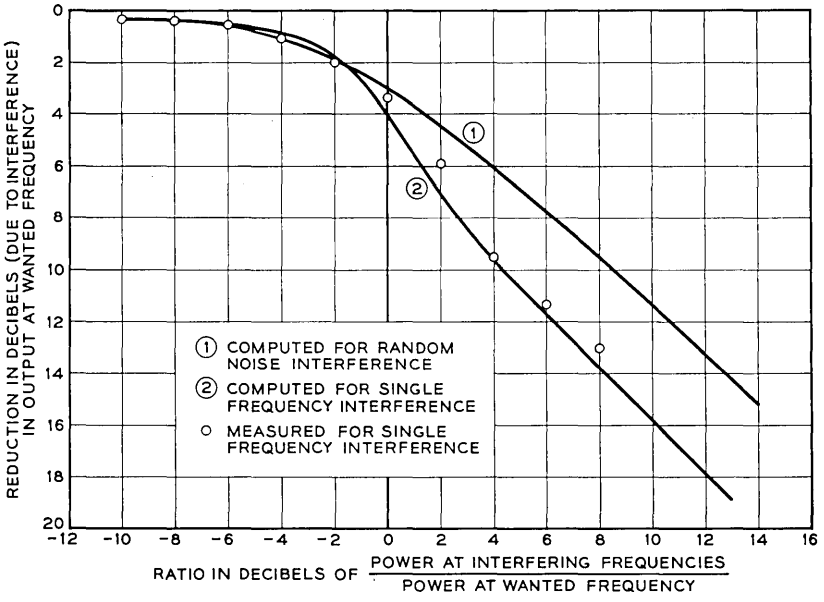

Fig. 2 — Effect of interference at limiter input on its output.

is for the case where the unwanted component is random noise; curve 2 applies where the unwanted component is a single frequency, not a harmonic of the wanted frequency. (The mathematical analysis leading to these results was made by S. O. Rice.) Experimental results are also shown for the case of a wanted frequency of 900 cps and an unwanted frequency of 500 cps.

What is the significance of the data shown in Fig. 2 with respect to guard action? Speech may contain components which simulate proper signals, but it is likely to include energy at other frequencies also. The selective circuitry that follows the limiter is designed to recognize a signal as bona fide only when it not only falls within a rather narrow passband, but also appears at an amplitude within about $2\frac{1}{2}$ db of the full output that the limiter is capable of delivering. Thus, when a burst of speech contains components at more than just the two signal frequencies, this fact is used to inhibit recognition of the signal frequencies in the burst. Inspection of Fig. 3, a simplified block diagram of the receiver, may be helpful at this point.

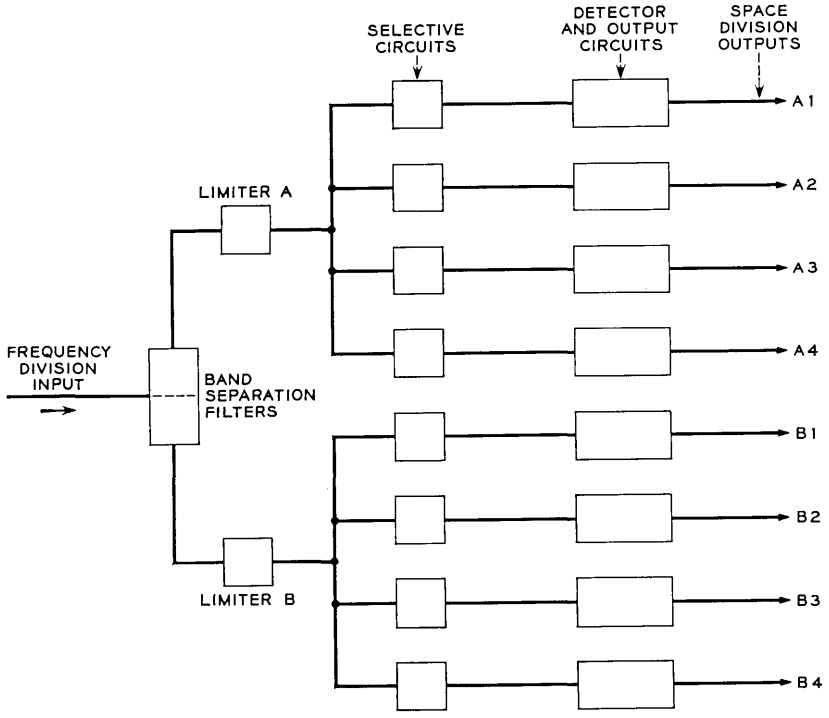

Fig. 3 — Simplified block diagram of four-by-four receiver.

Aside from considerations of guard action, it is interesting to note that, in divesting signal tones of amplitude variations, a limiter endows a selective circuit with an appreciably higher degree of selectivity than would normally be associated with the selective circuit's response characteristics. Thus, simple tuned circuits can adequately perform the function of frequency discrimination. An incidental by-product of this feature is that the response to tones with frequencies just outside the range recognized as valid is only moderately less than the response to valid tones. Consequently, sidebands associated with pulsing rates are only slightly attenuated and no specific allowance need be made for them in selecting the spacing between signal frequencies. G. C. Prins has taken advantage of this property for the control signaling for TASI,* in which a four-by-four-by-three code is used.

Guard action of the type that has been discussed requires that only one of the two tones making up a valid signal be admitted to each limiter. In order to derive the full benefit of limiter guard action, as much of the speech spectrum as possible should be given access to the limiter. Clearly, a bandpass filter preceding the limiter, to separate the two components of a valid signal, would defeat this objective, since it would permit competition for limiter capture between a signal frequency and only that portion of the speech burst lying in the same band (A or B). On the other hand an elimination filter allows competition with the whole speech spectrum except the attenuated band (B or A). Suitable characteristics for the band elimination filters are shown in Fig. 4. The loss in the

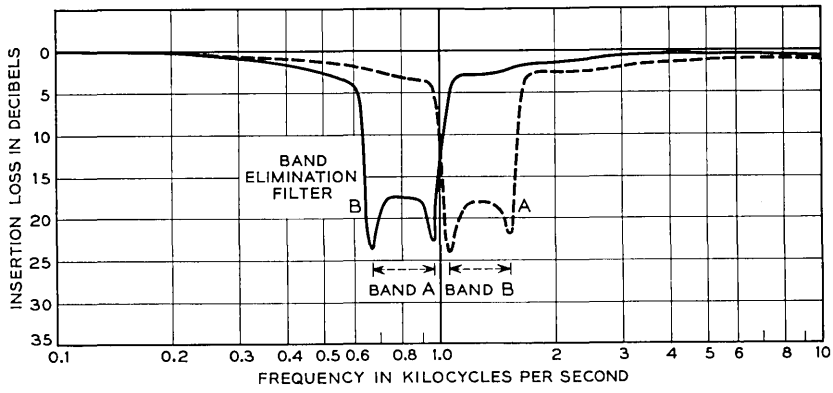

Fig. 4 — Insertion loss characteristics of band-elimination filters.

* The engineering aspects, but not the signaling aspects, of TASI (Time Assignment Speech Interpolation) are described in Ref. 4.

elimination band is more than 17 db. It will be shown that the two components of a bona fide signal differ at most by 5 db; consequently the interfering effect at the limiter input of the complementary component is never greater than that of an unwanted signal 12 db below that at the wanted frequency. It may be seen from Fig. 2 that the effect of noise 10 db or more below the signal is insignificant.

The fact that the band-elimination filters provide discrimination no greater than 17 db or so is advantageous in that guard action is still effective when an unwanted signal — speech or music — contains two major components at frequencies essentially the same as in a bona fide signal, but with amplitudes that differ by 17 db or more. Even in the absence of any other speech energy, the stronger component, although attenuated by the filter, would have at least the same amplitude as the weaker component at one of the limiter inputs, with the result that the limiter output would be reduced by 3 db, which would prevent recognition.

It may be desirable under some circumstances to increase the extent of guard action. This can be achieved by placing an equalizer ahead of the two band-elimination filters. Pre-emphasis of frequencies outside the signaling range enhances guard action. A speech burst containing two proper signaling frequencies might be registered as a valid signal if there is an insufficient aggregate of power at extraneous frequencies to invoke limiter guard action. Statistically, there are borderline cases where enhancement of these extraneous frequencies will tilt the balance. The practical limit to this procedure is set by system noise levels, which would also be enhanced and which would eventually tend to prevent the recognition of bona fide signals.

IV. CHOICE OF FREQUENCIES AND AMPLITUDES

4.1 *Frequencies*

Attenuation and delay distortion characteristics of typical combinations of transmission circuits are such that it is desirable to keep a multifrequency signal system within the 700 to 1,700 cps range.[5]

The choice of frequency spacing depends in part on the accuracies of the signal frequencies. It is expected that signals generated at the station set will be within 1.5 per cent of their nominal frequency values, and that the pass bands of the receiver selective circuits can be maintained within 0.5 per cent of their nominal ranges. (The basis for these estimates is discussed later.) Consequently, the selective circuits of central

office receivers need to have recognition bands of ±2 per cent about the nominal frequencies. In the case of receivers designed for end-to-end signaling, additional tolerance will be needed to accommodate the frequency changes introduced by carrier shift in some toll systems. Large frequency shifts are introduced only by the nonsynchronized systems (C, H and J), and even here the increasing use of voice-frequency telegraph and telephone signaling puts a premium on closer limits. Tentative requirements of ±10 cps of carrier shift were adopted for the signaling system described here.

The standardization of amplitude brought about by limiting permits an accurate definition of recognition bands in the receiver, irrespective of loop losses or slope. Judicious use of timing — namely, in delayed scanning of the selective circuit outputs — minimizes the effects of transient components upon adjacent channels. As a result, frequencies may be spaced quite closely, approaching, in fact, the recognition bandwidth of 4 per cent + 20 cps. If the lowest frequency is chosen as 700 cps, the next frequency must then be more than 748 cps, which is 7 per cent higher. Another 1 or 2 per cent increase in spacing makes the precise maintenance of the bandwidth less critical.

Another factor can profitably be taken into account in the selection of a frequency spacing. To reduce the probability of talk-off, the combinations of frequencies representing bona fide signals should be such that they are not readily imitated by the output from the speech transmitter. In a receiver with the guard action described, no sound composed of a multiplicity of frequencies at comparable levels is likely to produce talk-off. Thus, consonants present no problem. Vowels do, however, as do single-frequency sounds such as whistles that are large enough to encounter some harmonic distortion in a carbon transmitter. Fletcher[6] has shown that an electrical analog of the mechanism involved in the articulation of vowels is a buzzer (the vocal cords producing a fundamental and a long series of harmonics) followed by a selective network that shapes the harmonics into "formants" of the vowel sound. As a result, the spectrum of any sustained vowel contains a number of frequencies bearing harmonic relationships to each other. Hence, it is desirable that the pairs of frequencies representing valid signals avoid as many of these harmonic relationships as possible. The number of undesirable combinations is large but finite. Considering harmonics based on fundamental frequencies down to 100 cps, there are about 65 combinations that fall into the 700 to 1700 cps range, with one frequency below and one above the geometric mean of 700 and 1700 cps.

A family of frequencies that avoids a large proportion of troublesome combinations and also meets all the other requirements discussed so far

is as follows, the adjacent frequencies in each group being in the fixed ratio of 21:19, with one-and-a-half such an interval between the groups:

Group A	*Group B*
697 cps	1,094 cps
770	1,209
852	1,336
941	1,477

All frequencies are essentially within the 700 to 1,700 cps range, and the spacing is adequate to accommodate the recognition bands. The 16 pairs of frequencies representing valid signals avoid low-order ratios.

This is illustrated in Fig. 5 where frequencies are plotted on two logarithmic scales: those below 1,000 cps as abscissae and those above 1,000 cps as ordinates. Any valid signal is represented by a pair of coordinates — namely, its two component frequencies. The 16 square "windows" represent the ±2 per cent recognition bands required, with no

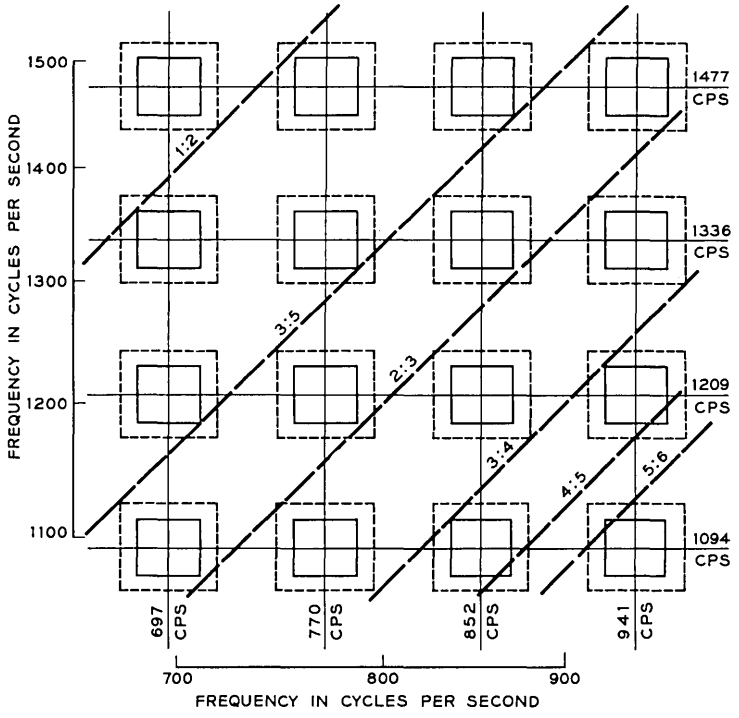

Fig. 5 — "Window" diagram for four-by-four code signal frequencies.

allowance for carrier shift. The rectangles with ±(2 per cent + 10 cps) sides represent the recognition bands for receivers involved in end-to-end signaling. The diagonal lines are the loci of pairs of frequencies having simple harmonic relationships, i.e., 1:2, 2:3, 3:4, 3:5, 4:5 and 5:6. The avoidance of these particular diagonals is beneficial, because they represent the effects of harmonics not only of the corresponding order but also of higher order, for example, the third and sixth, the ninth and 15th, etc. Not all applicable diagonals are shown in Fig. 5, and 18 of the 65 potentially troublesome combinations are not avoided. However, if two frequencies in speech representing two fairly high-order harmonics happen to fall at proper signal frequencies, their fundamental will be low in frequency, and the odds are then good that other harmonics within the telephone speech band will, together, have sufficient relative intensity to bring guard action into play.

Obviously, the larger windows needed in end-to-end receivers increase the probability of talk-off from harmonics in vowels.

Comparable benefits with respect to talk-off can be derived from other sets of frequencies with the same geometric spacing but displaced up or down on the frequency scale. The effect of such displacement is to shift all the windows in a direction parallel to the diagonals.* This specific set of frequencies was chosen because it is interleaved with the frequencies proposed for the tone ringer.[1]

It may be noted that, although a multigroup code requires a larger number of frequencies than a nonredundant P-out-of-N code, the closer frequency spacing made possible by group separation and limiting can actually result in economy of frequency spectrum.

4.2 *Amplitudes*

Since signaling information does not bear the redundancy of spoken words and sentences, and yet must be transmitted with a high degree of reliability, the signal power needs to be larger than the average speech power output from a telephone. A nominal combined signal power for the two frequencies of 1 db above 1 milliwatt at the telephone set terminals has been adopted as a realistic value, a ±3 db tolerance about this nominal value being acceptable.

For subscriber loops the maximum slope between 697 and 1,477 cps is about 4 db, the attenuation increasing with frequency. In two-subscriber loops (as in end-to-end signaling) the slope may be 8 db, although statistically this is unlikely. A reduction in the maximum amount

* Raising the frequencies reduces the areas in the rectangles representing carrier shift, but this effect is small enough to be ignored.

of slope can be achieved by transmitting the group B frequencies at a level 3 db higher than that of the group A frequencies. In this way the amplitude difference at the receiver input between the two components of a valid signal is never more than 5 db, and rarely more than 3 db. The nominal output powers are chosen as $-3\frac{1}{2}$ dbm and $-\frac{1}{2}$ dbm for groups A and B respectively, adding up to +1.3 dbm. In more than 99 per cent of station-to-central-office connections the 1,000 cps loop loss is expected to be less than 10 db (during the early 1960's). Similarly, the station-to-station loss at 1,000 cps is estimated to be less than 27 db in more than 99 per cent of all connections. Making some allowance for slope and variations in the generated power, the minimum signal power at a central office receiver is estimated to be −15.5 dbm. At a receiver involved in end-to-end signaling the minimum power is estimated to be −32.5 dbm.

V. GENERATION OF SIGNALING TONES

5.1 *Circuit Aspects*

A circuit devised by L. A. Meacham and F. West for the generation of the four-by-four code dial signals is shown in Fig. 6. Variations of this circuit are also under consideration, but the version in Fig. 6 is most easily described.

As a practical matter, numerous advantages arise out of the integration of the pushbutton dial circuit with the speech network of the 500-type telephone set. This may be done without any modification in the existing speech networks. The integrated circuit is shown in Fig. 7.

Operation of any pushbutton results in the generation of two tones, which last as long as the button is held down. While a button is depressed the speech circuit is disabled. Such performance is achieved as follows:

There are two tuned circuits, each consisting of a three-winding coil (A, A′, A″ and B, B′, B″) and a tuning capacitor (C_A and C_B). Windings A and B have a number of taps. The operation of a pushbutton results in the actuation of three switches, one of the K_A's and one of the K_B's (early in the button stroke) and, lastly, the common switch, K_1. In Fig. 6(a), K_1 is shown in the normal (speech) condition. In this condition, most of the direct current drawn by the speech circuit flows through a diffused junction silicon varistor RV_1 and some through windings A and B. The direct current through RV_1 is more than 15 milliamperes, and under this condition the ac resistance introduced in series with the loop is less than 3 ohms and has a negligible effect on the transmission of speech. The forward drop of the diode is stabilized at about 0.6 volt.

When a pushbutton is operated, the closure of one of the K_A and one of the K_B contacts connects each tuning capacitor to one of the taps on the associated tuning coil. Resonant frequencies are thus established for each tuned circuit, corresponding to the digit signal to be transmitted, but no signal is as yet generated. In the latter part of the travel of the

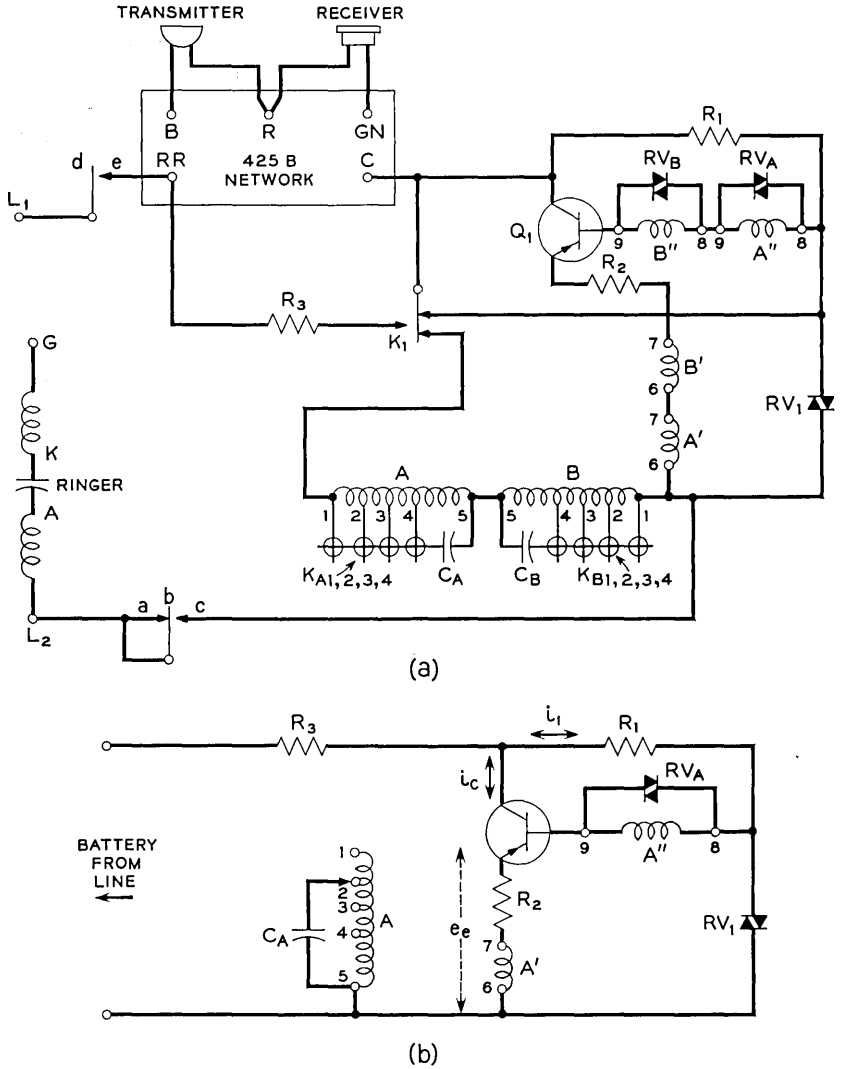

Fig. 6 — Pushbutton caller circuit: (a) shown as an appliqué to the 500-type telephone set circuit; (b) basic oscillator circuit.

button, switch K_1 is actuated through a mechanism common to all buttons. This produces three results: (a) The direct current through the main windings A and B is interrupted, causing shock excitation of the two tuned circuits. (b) The speech circuit is almost short-circuited (the resistance of R_3 is low, but so chosen that the subscriber hears the outgoing signal at a low level). There is, therefore, no appreciable interference from the speech transmitter while a signal is going out. (c) Voltage is made available to the transistor. Having been started at full amplitude, the two-frequency oscillation is sustained by feedback through transistor current multiplication and transformer action between the secondary and tertiary windings of each coil.

The operation of the oscillator is more readily apparent from Fig. 6(b), which shows the circuit in the signaling condition, but omits for clarity the elements of one of the two tuned circuits. It may be shown

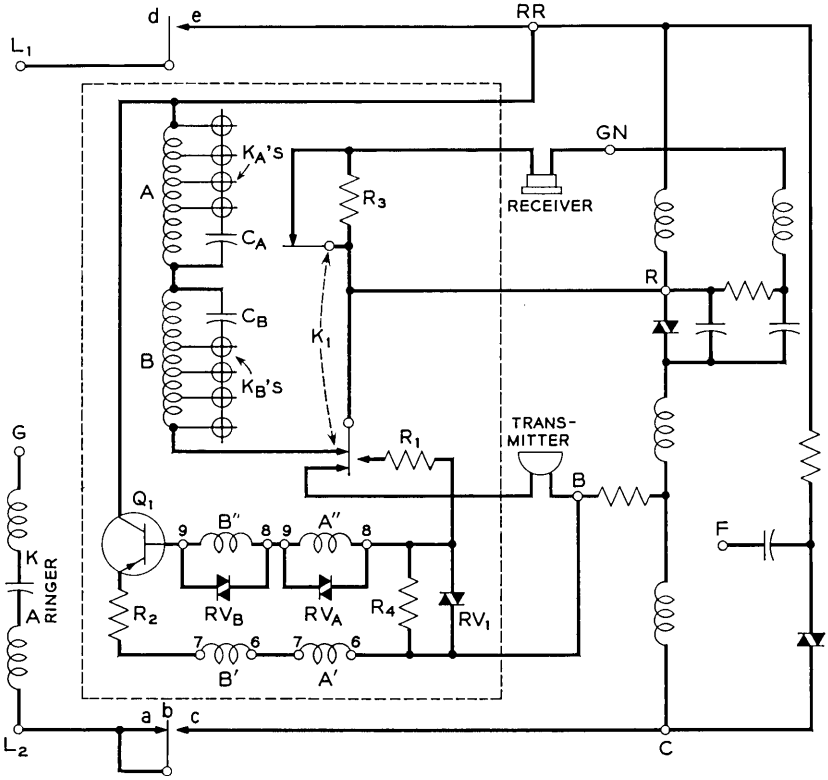

Fig. 7 — Pushbutton caller circuit integrated with 500-type set.

that this is a bridge-stabilized oscillator[7] employing a T network bridged by mutual inductance.

The bias for class A operation of the transistor, provided by the resistor R_1 and the silicon varistor RV_1, results in a constant direct emitter current. This current need only slightly exceed the peak alternating current, and is preferably kept small to minimize collector dissipation; there is no reason for allowing it to go up with collector voltage (on short loops) as it would if a resistance were used instead of the diode. Further, the diode represents a low ac resistance, obviating the need for a bypass capacitor.

The oscillator action is determined by two key equations, one ensuring that the feedback loop gain exceeds unity, and the other governing the amplitude of the oscillation. The first of these equations may be shown to be

$$R_{2\,\text{eff}} = k\,\frac{Q\omega L}{r_t^2}\,(r_b - 1)[1 - r_b(1 - \alpha)], \tag{1}$$

where

$R_{2\,\text{eff}}$ = R_2 + resistance of 6–7 windings of both coils + h_{ib} (transistor grounded base input impedance);
r_b = turns ratio of base winding (8–9) to emitter winding (6–7);
r_t = turns ratio of tuned winding (1–5) to emitter winding (6–7);
k = a factor less than unity (see below);
L = nominal inductance of tuned winding (1–5) and
Q, ω, and α have the customary meanings, (nominal values).

Equation (1) is a modification of the inequality

$$R_{2\,\text{eff}} \lessapprox \frac{Q\omega L}{r_t^2}\,(r_b - 1)[1 - r_b(1 - \alpha)], \tag{1a}$$

the significance of which is explained below. The coefficient k is less than unity by an amount sufficient to ensure that the inequality (1a) is satisfied in spite of all admissible combinations of variations in Q, ω, L and α. The physical significance of (1) and (1a) is easily recognized by consideration of the case of $r_b = 2$ and $\alpha = 1$. Then,

$$R_{2\,\text{eff}} \lessapprox \frac{Q\omega L}{r_t^2}; \tag{2}$$

i.e., $R_{2\,\text{eff}}$ must be less than the impedance seen across the 6–7 winding at resonance. If the peak value of the ac voltage from emitter to "ground" is e_e, more than $\frac{1}{2}e_e$ appears across the 6–7 winding, and more than e_e

appears across the 8–9 winding, one end of which is grounded through RV_1. Neglecting the small ac drop from base to emitter, more than e_e appears at the emitter, the assumed starting point. Thus, the loop gain exceeds unity and the amplitude of oscillation must grow until the loop gain stabilizes as a result of some nonlinearity.

Nonlinearity is introduced by means of the silicon varistor RV_A, which rapidly reduces Q as soon as the amplitude across its terminals exceeds its forward drop v_d, which is approximately 0.6 volt. Since the active element is thus not driven into its nonlinear region, several independent oscillations can be sustained at the same time by the single transistor. Let

r_d = turns ratio of winding bridged by RV_A to the emitter winding (6–7).

Then the peak value of the ac emitter current is given by

$$i_e = \frac{v_d}{r_d} \frac{r_b - 1}{R_{2\,\mathrm{eff}}}. \tag{3}$$

From this, r_d can be computed to give any desired i_e. More exactly, the actual ac voltage-current characteristic of the silicon varistor should be taken into account.

Neglecting leakage, the voltage-current relation for a silicon diode is quite accurately represented by

$$V_d = A + B \log_{10} I_d, \tag{4}$$

where

A and B = constants,
V_d = instantaneous voltage across diode,
I_d = instantaneous current through diode.

The ac resistance of a diode when it is conducting a direct current I_d is found by differentiation of (4). However, in the present application, the voltage-limiting diodes RV_A and RV_B do not conduct any direct current. F. T. Boesch has shown analytically as well as experimentally that, in this condition (assuming voltage to remain sinusoidal), the relation between the peaks of voltage and current is well approximated by

$$v_d = a + b \log_{10} i_d, \tag{5}$$

where

$$a = A + 0.33B$$
$$b = B.$$

Thus, the effective ac characteristics of the diode can be computed from a measurement of the dc characteristics. Substituting the value of v_d given by (5) into (3), one obtains

$$i_e = \frac{a + b \log_{10} i_d}{r_d} \frac{r_b - 1}{R_{2\,\text{eff}}}. \tag{6}$$

By making use of the various relationships arising out of the transformer and transistor action, i_d is eliminated from (6) and the second key equation is thus obtained, from which r_d can be computed to yield the desired amplitude i_e. This is

$$i_e \left[1 - r_b(1 - \alpha) - \frac{R_{2\,\text{eff}}\, r_t^2}{(r_b - 1)Q\omega L}\right] = r_d \exp\left[\frac{r_d i_e R_{2\,\text{eff}}}{b(r_b - 1)} - \frac{a}{b}\right]. \tag{7}$$

The equation is not explicit in terms of r_d, but yields sufficient accuracy with two or three successive approximations.

The design procedure then is: (a) Select r_b. The choice is not very critical. Values of r_b between 2 and 5 have been found to result in stable designs. (b) Compute r_t from (1). (c) Compute r_d from (7). These computations are made at the lowest of the frequencies in each group. The condition for oscillation, (1), is then also met at the higher frequencies of the group.

5.2 *Frequency Stability*

Owing to the bridge-stabilized nature of the oscillator, many potential sources of frequency instability have a relatively minor effect. Thus, the total change in frequency caused by variations in loop impedance, battery voltage and transistor properties is less than ± 0.15 per cent. There remain, however, two major sources of frequency instability: (a) variations in tuning elements and (b) frequency pulling.

Stability of the tuning elements is, of course, of prime importance. Ferrite cup core inductors offer a combination of advantages not available with other types of cores: good mechanical strength, moldability, simplicity of separately fabricated cylindrical winding, inherent enclosure of the winding, adjustability of inductance and good Q for a given size and stability. Temperature stability and Q can be traded by incorporation of an air gap in the magnetic path with the effect of diluting the inherently high permeability of ferrite, but also diluting the temperature variations in this property.

At audio frequencies the core losses are small compared to copper losses. It is readily derived that, in this range,

$$Q \propto \omega D^2 \frac{\mu S}{S_e}, \tag{8}$$

where

D = typical linear dimension of core,
μ = permeability of ferrite,
$1/S$ = fractional change in permeability μ per degree of temperature change,
$1/S_e$ = fractional change in effective permeability μ_e per degree of temperature change.

Thus, for a core of given size, D, and of material having given properties μ and S, Q is inversely proportional to temperature stability. For manganese-zinc ferrite, typical values are $\mu = 1{,}200$ and $1/S = 2{,}000$ ppm/°C. With a core about 1 inch in diameter and 0.5 inch in height, and with an air gap such as to achieve a stability corresponding to $1/S_e = 300$ ppm/°C, a typical value of Q is 30 at 700 cps. Capacitors with polystyrene as the dielectric have a negative temperature coefficient of about -100 ppm/°C, which compensates for part of the change in coil inductance. Frequency variations due to temperature changes over the range from -30°C to $+55$°C may thereby be held down to less than ± 0.5 per cent.

Adjustability of inductance is desirable so that requirements on initial values of the tuning capacitors may be reduced. This is brought about by varying the reluctance introduced by the air gap. In one design, it is achieved by axial motion of a central slug bridging an air gap between the center posts of the two cups. In another design, parts of the center posts are cut away in a manner such that rotation of one cup relative to the other produces a change in reluctance.

In a dual-frequency oscillator with a common active element there exists some "pulling" together of the two frequencies that are generated simultaneously. The presence of the tuned circuit of the second frequency introduces some reactance into the feedback loop at the first fequency. To compensate for the resulting phase shift, the first tuned circuit must go off frequency — the closer the two frequencies, the greater the shift. A mathematical study, confirmed by experiment, has led to the following expression for the frequency change:

$$\Delta f_A = \frac{f_A f_B}{4Q_A Q_B (f_B - f_A)}, \tag{9}$$

where

f_A, f_B = nominal frequencies being generated,
Q_A, Q_B = circuit Q's obtaining at the above frequencies.

In the four-by-four code, pulling is worst with the 941–1094 cps combination of frequencies. With practical circuit components, pulling for this

combination is about 3 cps for each of the two frequencies. By changing tap locations, this error can be halved, to become about ±0.15 per cent.

Except for manufacturing tolerances, there is no other single large source of frequency variations. Considering all sources, it appears practical to hold frequencies within ±1.5 per cent of nominal values; even a ±1 per cent tolerance may eventually be approached.

5.3 *Amplitude Stability*

The emitter current amplitude is given by (3), and the peak ac collector current is

$$i_c = \alpha \frac{v_d}{r_d} \frac{r_b - 1}{R_{e\,\mathrm{eff}}} . \tag{10}$$

Since r_b and r_d are turns ratios, the quantities α, v_d and $R_{2\,\mathrm{eff}}$ are mainly responsible for variations in i_c. As far as the power delivered to the loop is concerned, further variations are associated with the range of loop impedances.

Whereas it is essential to hold frequencies to values within the recognition band of the receiver, no severe requirements need be set on signal power at the set terminals. Using a precision resistor for R_2, but relatively inexpensive units for the amplitude controlling diodes RV_A and RV_B, power delivered to a 900-ohm loop may vary by up to ±2 db from the nominal values. Differences in loop impedance may introduce an additional variation, but the total is expected to be less than ±3 db. In the case of the circuit shown in Fig. 7, these values are modified by the effect of loop equalization inherent in the 500-type set's circuit.[8] This effect is beneficial since on short loops smaller signal amplitudes are not only acceptable but desirable.

5.4 *Pushbutton Mechanism*

Means must be provided to translate the customer's operation of pushbuttons into the switch operations described earlier, that is, closure of one each of the K_A and K_B contacts followed by the common switch K_1. A mechanism developed by C. E. Mitchell and R. E. Prescott to perform these functions is shown in Fig. 8.

The operation of a pushbutton causes the rotation of two rods, one associated with a row of buttons, the other with a column. The coordinates of a button uniquely determine the pair of rods that rotate when the button is pressed. In the mechanism in Fig. 8, there are four rods corresponding to rows and three to columns. Hence, there are 12 possible combinations of the seven rods; two of these are spares, and the cross-

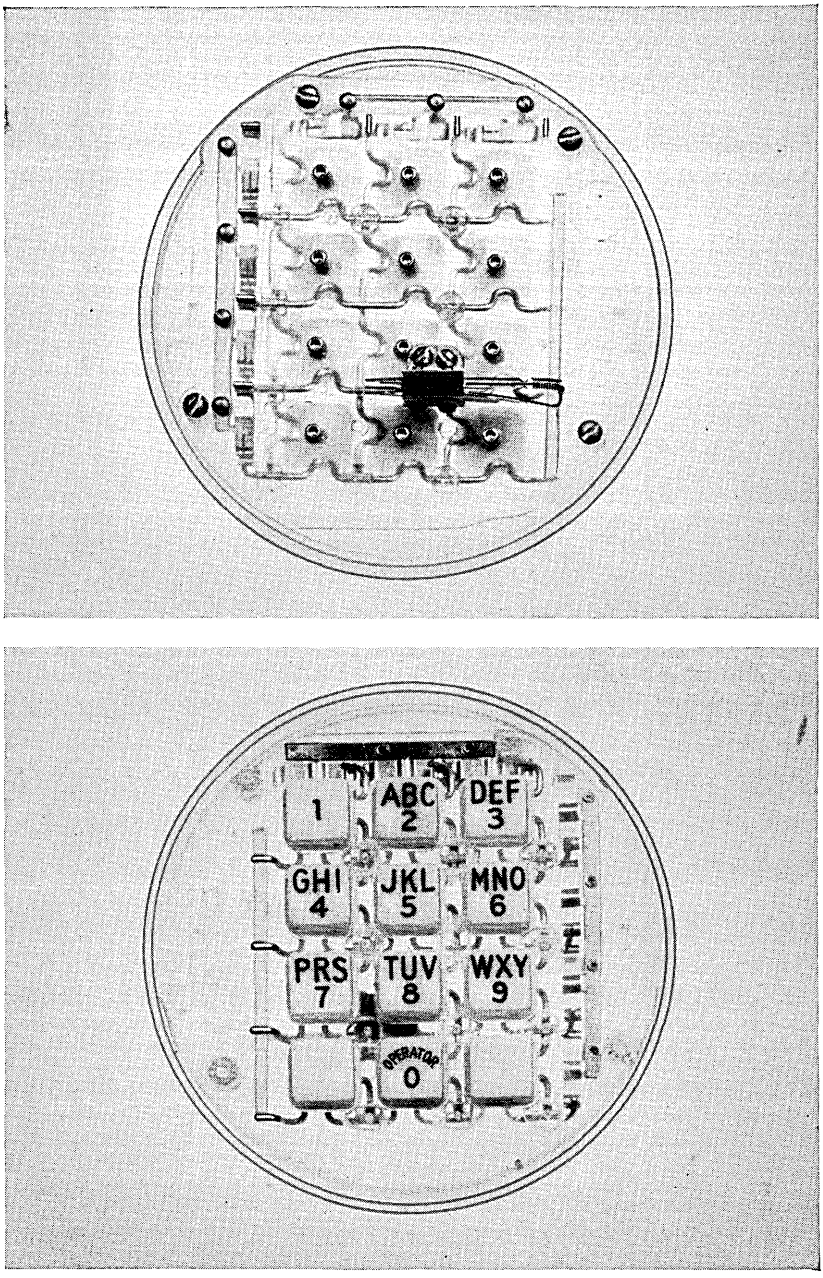

Fig. 8 — Transparent model of pushbutton mechanism.

points may not be equipped with buttons. The four-by-four permits an extension of the array, as may be desired, up to eight rods and 16 buttons.

The rotation of a rod closes one of the contacts K_A and K_B. Moreover, rotation of any of the four rods corresponding to the four rows produces linear motion of the link on the left side of the button array. This link is coupled to a common snap switch performing the function of K_1 in Fig. 6(a). Although not shown in Fig. 8, all signaling circuitry is mounted on the back of the mechanism, making the pushbutton caller a self-contained unit that can be substituted for the rotary dial.

VI. RECEIVER

A detailed discussion of receiver design for this two-group signaling system is beyond the scope of the present paper. However, it seems desirable to touch briefly upon the principles involved, emphasizing those that are directly related to the specific code and frequency allocation of the four-by-four system.

The general requirements to be met are:

i. To detect the presence of a bona fide pair of signaling tones in a sufficiently short time that essentially no restriction is placed on the customer's operations or on the pushbutton mechanism to provide delay in release.

ii. To provide the maximum possible talk-off protection consistent with adequate immunity to interference from circuit noise.

Earlier sections have discussed the means employed to optimize detection of valid signals and the guard action inherent in the concatenation of (a) band elimination filters for separation of the two signaling tones, (b) instantaneous extreme limiting of the tones individually, (c) selective circuits passing the fundamentals of these tones after limiting and (d) threshold detection of the response in the selective circuits.

A test for the presence of two and only two tones is a further means for distinguishing between valid signals and others. Part of such a test is provided by limiter guard action: no more than one selective circuit from each group can ever by energized at one time. However, logic must be incorporated to apply the criterion of simultaneous presence of at least one tone in each group.

Recognition time for a valid signal — that is, the time interval for which both components are required to be present without interruption — is an important parameter in receiver design. With the rapid changes that are characteristic of most speech, there is a substantial advantage in making this time as long as possible. However, an upper limit is set

by the requirement that the system accommodate the most rapid push-button operation that a customer is at all likely to attain. A compromise between these conflicting objectives is, of course, necessary. Work to date indicates that the recognition time should be of the order of 40 milliseconds.

VII. ACKNOWLEDGMENT

The names of some — but by no means all — of those who have contributed to the development of the signaling system described in this paper are indicated in appropriate places. The writer wishes to thank these individuals, and particularly L. A. Meacham, who not only made many contributions but also guided the whole project.

REFERENCES

1. Meacham, L. A., Power, J. R. and West, F., Tone Ringing and Pushbutton Calling, B.S.T.J., **37,** March 1958, p. 339.
2. Dahlbom, C. A., Horton, A. W., Jr. and Moody, D. L., Applications of Multifrequency Pulsing in Switching, Trans. A.I.E.E., **68,** 1949, p. 392.
3. Weaver, A. and Newell, N. A., In-Band Single-Frequency Signaling, B.S.T.J., **33,** November 1954, p. 1309.
4. Bullington, K. and Fraser, J. M., Engineering Aspects of TASI, B.S.T.J., **38,** March 1959, p. 353.
5. Horton, A. W., Jr. and Vaughan, H. E., Transmission of Digital Information Over Telephone Circuits, B.S.T.J., **34,** May 1955, p. 511.
6. Fletcher, H., *Speech and Hearing in Communication*, D. Van Nostrand Co., New York, 1953.
7. Meacham, L. A., The Bridge Stabilized Oscillator, Proc. I.R.E., **26,** October 1938, p. 1278.
8. Bennet, A. F., An Improved Circuit for the Telephone Set, B. S.T.J., **32,** May 1953, p. 611.

Recent Monographs of Bell System Technical Papers Not Published in This Journal*

ANDERSON, E. W., see McCall, D. W.

BALA, V. B., see Geller, S.

BATTERMAN, B. W.

X-ray Integrated Intensity of Germanium Effect of Dislocations and Impurities, Monograph 3357.

BURRUS, C. A.

Zeeman Effect in the 1- to 3-Millimeter Wave Region, Monograph 3316.

CHYNOWETH, A. G.

Effects of Space Charge Fields in Barium Titanate, Monograph 3317.

DOUGLASS, D. C., see McCall, D. W.

EISINGER, J.

Electrical Properties of Hydrogen Adsorbed on Silicon, Monograph 3358.

ELLIS, W. C., WILLIAMS, H. J. and SHERWOOD, R. C.

Growth of MnBi Crystals and Evidence for Subgrains from Domain Patterns, Monograph 3261.

FORSTER, J. H. and VELORIC, H. S.

Effect of Variations in Surface Potential on Junction Characteristics, Monograph 3377.

FROSCH, C. J., see Spitzer, W. G.

* Copies of these monographs may be obtained on request to the Publication Department, Bell Telephone Laboratories, Inc., 463 West Street, New York 14, N. Y. The numbers of the monographs should be given in all requests.

GELLER, S., GILLEO, M. A., WOOD, E. A. and BALA, V. B.

Crystallographic Studies of Perovskite-like Compounds Involving Rare-Earth Ions, Monograph 3260.

GELLER, S. and GILLEO, M. A.

Effect of Dispersion Corrections and Refinement of Yttrium-Iron Garnet Structure, Monograph 3361.

GESCHWIND, S. and WALKER, L. R.

Exchange Resonances in Gadolinium Iron Garnet, Monograph 3362.

GILLEO, M. A., see Geller, S.

GILLEO, M. A., see Geller, S.

GNAEDINGER, R. J., JR.

Some Aspects of Vacuum Deposition of Metals in Transistor Fabrication, Monograph 3379.

GYORGY, E. M., see Humphrey, F. B.

HOWARD, J. B.

A Review of Stress-Cracking in Polyethylene, Monograph 3349.

HOWARD, J. B.

Engineering Thermoplastics for Ocean Telephone Cables, Monograph 3344.

HROSTOWSKI, H. J. and KAISER, R. H.

The Solubility of Oxygen in Silicon, Monograph 3363.

HUMPHREY, F. B. and GYORGY, E. M.

Flux Reversal in Soft Ferromagnetics, Monograph 3364.

KAISER, R. H., see Hrostowski, H. J.

KAISER, W. and THURMOND, C. D.

Nitrogen in Silicon, Monograph 3322.

KLEINMAN, D., see Logan, R. A.

KLEINMAN, D., see Spitzer, W. G.

KNOX, K.

Structure of K_2CuF_4 — New Kind of Distortion for Octahedral Copper (II), Monograph 3351.

KOLB, E. D. AND TANENBAUM, M.

Uniform Resistivity p-Type Silicon by Zone Leveling, Monograph 3365.

LOGAN, R. A., PEARSON, G. L., and KLEINMAN, D.

Anisotropic Mobility in Plastically Deformed Germanium, Monograph 3366.

MCCALL, D. W., DOUGLASS, D. C. and ANDERSON, E. W.

Molecular Motion in Polyethylene. II, Monograph 3352.

MCSKIMIN, H. J.

Measurement of Ultrasonic Wave Velocities and Elastic Moduli for Small Solid Specimens, Monograph 3369.

MORRISON, J. A., see Pierce, J. R.

NELSON, L. S.

Flash Heating — A New Technique, Monograph 3371.

NIELSEN, J. W., see Williams, J. C.

PEARSON, G. L., see Logan, R. A.

PETER, M.

Millimeter-Wave Paramagnetic Resonance Spectrum of ^6S State Impurity (Fe^{+++}) in $MgWO_4$, Monograph 3372.

PIERCE, J. R. and MORRISON, J. A.

Disturbances in A Multi-velocity Plasma, Monograph 3328.

RALSTON, A.

A Family of Quadrature Formulas which Achieve High Accuracy in Composite Rules, Monograph 3373.

REISS, H.

Influence of Solutes on Self-Diffusion in the Face-Centered Cubic Lattice, Monograph 3353.

ROHN, W. B.

Reliability Prediction for Complex Systems, Monograph 3378.

SCHLABACH, T. D.

The Temperature Dependence of Electrical Resistivity of Laminated Thermoset Materials, Monograph 3374.

SCHAWLOW, A. L. and TOWNES, C. H.

Infrared and Optical Masers, Monograph 3345.

SHERWOOD, R. C., see Ellis, W. C.

SLICHTER, W. P.

Nuclear Resonance Studies of Polymer Chain Flexibility, Monograph 3348.

SPITZER, W. G., KLEINMAN, D., WALSH, D. and FROSCH, C. J.

Infrared Properties of Hexagonal and Cubic Silicon Carbide, Monograph 3375.

TANENBAUM, M., see Kolb, E. D.

THURMOND, C. D., see Kaiser, W.

TOWNES, C. H., see Schawlow, A. L.

VELORIC, H. S., see Forster, J. H.

WALKER, L. R., see Geschwind, S.

WALSH, D., see Spitzer, W. G.

WHITE, A. D.

New Hollow Cathode Glow Discharge, Monograph 3356.

WILLIAMS, H. J., see Ellis, W. C.

WILLIAMS, J. C. and NIELSEN, J. W.

Wetting of Original and Metallized High-Alumina Surfaces by Brazing Solders, Monographs 3376.

WOOD, E. A., see Geller, S.

Contributors to This Issue

A. E. BAKANOWSKI, B.S., 1943, Worcester Polytechnic Institute; S.M., 1948, and Ph.D., 1954, Brown University; Bell Telephone Laboratories, 1954—. He has specialized in development of improved semiconductor diodes for use as microwave frequencies as modulators, detectors and amplifiers and computer diodes. Member American Physical Society, American Institute of Physics, Sigma Xi, Tau Beta Pi.

VÁCLAV E. BENEŠ, A.B., 1950, Harvard College; M.A., Ph.D., 1953, Princeton University; Bell Telephone Laboratories, 1953—. Mr. Beneš has been engaged in mathematical research, on stochastic processes, traffic theory and servomechanisms. In 1959–60 he was visiting lecturer in mathematics at Dartmouth College. Member American Mathematical Society, Association for Symbolic Logic, Institute of Mathematical Statistics, Mind Association, Phi Beta Kappa.

ALEXANDER FEINER, Technische Hochschule, Vienna; M.S.E.E., 1952, Columbia University; Bell Telephone Laboratories, 1953—. He has been engaged in various phases of switching systems development. In 1959 he was appointed Switching Systems Development Engineer. Member Sigma Xi.

J. H. FORSTER, B. A., 1944, and M. A., 1946, University of British Columbia; Ph.D., 1953, Purdue University; Bell Telephone Laboratories, 1953—. He has been engaged in development and reliability studies of transistors and semiconductor surface studies. He served as instructor in semiconductor electronics in the Bell Laboratories' C.D.T. program. Member Sigma Pi Sigma, Sigma Xi.

DENNIS B. JAMES, B.Sc., 1948, University of Wales; Ph.D., 1953, Cambridge University; Telecommunication Research Establishment (England), 1944–46; Atomic Energy Research Establishment (England), 1946–47; University of British Columbia, 1952–54; Bell Telephone Laboratories, 1954—. His first work with Bell Laboratories was on magnetic core circuits. More recently he has been engaged in research in time division switching and pulse code modulation, especially with application to electronic switching systems. Member I.R.E.

JOHN D. JOHANNESEN, B.S.E.E., 1943, Bucknell University; M.S.E.E., 1948 and Ph.D., 1953, Case Institute of Technology; Bell Telephone Laboratories, 1954—. He has specialized in development of solid state switching systems with emphasis on time-division switching. Member I.R.E., Eta Kappa Nu, Sigma Xi.

MICHAEL KOWALCHIK, B.S., 1954, Seton Hall University; Bell Telephone Laboratories, 1951—. He has been engaged in research in semiconductors with special emphasis on the control of impurities.

V. E. LEGG, A.B., 1920 and M.S., 1922, University of Michigan; Western Electric Company Engineering Department, 1922–25; Bell Telephone Laboratories, 1925—. He has specialized in studies and development of magnetic materials for application to submarine telephone and telegraph cables, loading coils, magnetic detecting apparatus, and recently in studies of magnetic materials for telephone apparatus. Fellow American Physical Society; member A.I.E.E., American Society for Metals; American Society for Testing Materials, Phi Beta Kappa.

CLARENCE A. LOVELL, B.A., 1922, Mississippi College; M.A., 1928 and Ph.D., 1932, University of Pennsylvania; Bell Telephone Laboratories, 1929—. While taking graduate studies, before joining Bell Laboratories, he taught mathematics at Mississippi College and Drexel Institute of Technology. His early work was research on electroacoustical apparatus and mechanical filters. From 1934 to 1936 he worked on design of television terminal apparatus for the first experimental coaxial cable. He supervised acoustical research on recording and measuring instruments and tone synthesizers until World War II, when he worked on gun directors and gun data computers. Since World War II he has headed groups in switching research and is now Director of Switching Systems Development in charge of electronic switching systems. Awarded Presidential Medal for Merit and Howard N. Potts Medal of the Franklin Institute for work on Electrical Gun Director. Fellow Acoustical Society of America; member, A.I.E.E., American Mathematical Society, Franklin Institute.

TERRELL N. LOWRY, B.E.E., 1952 and M.S.E.E., 1955, Georgia Institute of Technology; Bell Telephone Laboratories, 1955—. He has been engaged in exploratory development of remote line concentrators for use in electronic switching and is now in charge of a group with that responsibility. Member I.R.E., Eta Kappa Nu, Phi Kappa Phi, Tau Beta Pi; associate member, Sigma Xi.

W. A. MALTHANER, B.E.E., 1937, Rensselaer Polytechnic Institute; Bell Telephone Laboratories, 1937—. His first work was in development and research on automatic telephone central offices. During World War II he worked on fire control and radar systems, and after the war returned to research on automatic telephone central offices, customer dialing and supervisory arrangements, interoffice signaling systems and data transmission systems. He was appointed Systems Research Engineer in 1958. Senior member I.R.E.; member A.I.E.E., American Association for the Advancement of Science, Sigma Xi, Tau Beta Pi.

PHILIP G. RIDINGER, B.S.E.E., 1950, Lehigh University; Bell Telephone Laboratories, 1950—. After completing rotational assignments in the C.D.T. program, he worked on centralized AMA for crossbar tandem switching centers. More recently, he has been engaged in the development of remote line concentrators for electronic switching systems. Member Eta Kappa Nu, Pi Mu Epsilon, Tau Beta Pi.

JOHN P. RUNYON, M.E., 1944, Stevens Institute of Technology; Diploma in Mathematics, 1950, Swiss Federal Institute of Technology; Bell Telephone Laboratories, 1950—. He has been engaged in research and development on switching systems and apparatus. Member I.R.E., American Association for the Advancement of Science.

L. SCHENKER, B.Sc., 1942, University of London; M.Sc., 1950, University of Toronto; Ph.D., 1954, University of Michigan; Bell Telephone Laboratories, 1954—. He has been studying various phases of pushbutton calling relating to station apparatus, central office receiving equipment and voice frequency transmission. Member American Society of Civil Engineers, Engineering Institute of Canada, Phi Kappa Phi, Sigma Xi.

CARL D. THURMOND, B.S., 1943, and Ph.D., 1949, University of California; instructor, research fellow, University of California, 1949–51; Bell Telephone Laboratories, 1951—. He has been engaged in research in the physical chemistry and thermodynamic properties of semiconductors. Member American Chemical Society, American Physical Society, Sigma Xi.

F. A. TRUMBORE, B.S., 1946, Dickinson College; Ph.D., 1950, University of Pittsburgh; National Advisory Committee for Aeronautics, 1950–52; Bell Telephone Laboratories, 1952—. He has specialized in studies of thermodynamic properties of germanium and silicon alloys and compounds. Member American Chemical Society, American Association for the Advancement of Science, Phi Beta Kappa, Sigma Xi, Phi Lambda Upsilon.

HANS-GEORG UNGER, Dipl. Ing., 1951, and Dr. Ing., 1954, Technische Hochschule, Braunschweig (Germany); Siemens and Halske (Germany), 1951–55; Bell Telephone Laboratories, 1956—. Mr. Unger's work at Bell Laboratories has been in research in waveguides, especially circular electric wave transmission. Senior member I.R.E.; member N.T.G. (German Communication Engineering Society).