THE BELL SYSTEM
TECHNICAL JOURNAL

VOLUME XXXII JANUARY NUMBER 1

Surface Properties of Germanium

By WALTER H. BRATTAIN* and JOHN BARDEEN†

(Manuscript received September 3, 1952)

The contact potential (c.p.) and the change of contact potential with illumination $(\Delta c.p.)_L$ *of several germanium surfaces have been measured. The reference electrode used was platinum. It was found that the c.p. could be cycled between two extremes about 0.5 volts apart by changing the gaseous ambient. Ozone or peroxide vapors gave the c.p. extreme corresponding to the largest dipole at the Ge surface. Vapors with OH radicals produced the other extreme. There is a one to one correlation between c.p. and* $(\Delta c.p.)_L$. *For 12-ohm cm n-type Ge* $(\Delta c.p.)_L$ *was large and positive when the surface dipole was largest, decreased to zero and became slightly negative as the surface dipole decreased to its smallest value. The variation for 12-ohm cm p-type Ge was just opposite as regards both sign and dependence on surface dipole. The surface recombination velocity was found to be independent of c.p. For a chemically prepared surface it was 50–70 cm/sec and 180–200 cm/sec for n and p-type surfaces respectively. A theory is given that explains the results in terms of surface traps,* N_a *per cm*^{2} *donor-type traps near the conduction band and* N_b *per cm*^{2} *acceptor-type traps near the filled band. A quantitative fit with experiment is obtained with only one free parameter. The results are direct evidence for the existence of a space charge layer at the free surface of a semiconductor.*

INTRODUCTION

Every one is familiar with the fact that it is necessary to expend energy to remove an electron from a conducting solid. This energy is

* Bell Telephone Laboratories.
† University of Illinois. The contributions of the second author to this work started while he was a Member of the Technical Staff of Bell Telephone Laboratories and continued at the University of Illinois.

called the work function. The work function is caused in part by a charged double layer or dipole at the solid surface. In metals this dipole extends over a distance of the order of 10^{-8} cm. In semiconductors however part of the dipole extends into the semiconductor to a distance of the order of 10^{-6} to 10^{-4} cm depending on the properties of the semiconductor. This part of the surface dipole is called the space-charge layer. The rest of the surface dipole has approximately the same extent as in metals.

In the space between any two conducting solids there is a contact potential caused by the difference between the work functions of the two surfaces. One can measure this contact potential by several methods. We have used an adaptation of the well known method of Kelvin. If one has a reference electrode whose work function remains constant, then by measuring the c.p. between this electrode and another surface one can measure any changes in work function or total dipole of this second surface.

The above method has been used to study the properties of the germanium surface in a gaseous ambient at atmospheric pressure. It has been found that the total dipole at the germanium surface can be changed by changing the ambient and further that by proper control of the ambient the surface can be cycled back and forth between two extremes of small or large dipole corresponding to a c.p. change or work function difference of the order of one-half volt.

If one upsets the thermal equilibrium in the germanium by creating excess electron-hole pairs near the surface, the potential of the surface will change until a steady state is reached. When the extra electron-hole pairs are introduced by illuminating the surface with light, the potential change shows up as a measurable change in contact potential between the reference electrode and the Ge surface.[1] It has been found that this contact potential change on illumination $(\Delta\text{c.p.})_L$ is large and positive on n-type Ge when the surface dipole is large, then decreases to zero and becomes slightly negative as the surface dipole decreases to the smaller extreme. For p-type Ge the $(\Delta\text{c.p.})_L$ is large and negative when the surface dipole is small and goes through zero and becomes slightly positive as the surface dipole increases to the larger extreme.

One can describe qualitatively what is going on as follows. The extra hole and electron pairs created by the light diffuse either to the interior or to the surface to recombine. The recombination in the interior is governed by the body life time τ. The surface recombination is characterized by a recombination velocity v_s. When the surface is illuminated its potential, with respect to the interior, changes until the combined

flow of holes and electrons to the surface and interior is just equal to the rate they are being created by the light. The sign and magnitude of the potential change for a given illumination depends on the body properties of the germanium and on the size of the space charge layer.

These experimental results are direct evidence for the existence of a space charge layer at the free surface of a semiconductor. They not only confirm the results obtained for silicon surfaces[1] but go much further in that they enable one to determine how the layer is changed by the gaseous ambient used.

It is known that the surface recombination velocity, v_s, can be changed, by large factors, by surface treatment.[2] For mechanically treated surfaces v_s approaches thermal velocities. Every hole or electron striking the surface recombines. For such a surface it is found that $(\Delta c.p.)_L$ is too small to be measured. On the other hand v_s can be as low as 100 cm/sec for chemically polished or etched surfaces such as those used in the experiments where $(\Delta c.p.)_L$ was measured. In this case one wishes to know how v_s depends on the gaseous ambient. This was measured for the same surface used in measuring $(\Delta c.p.)_L$ and it was found that, for the ambients used, v_s is approximately a constant and therefore independent of the other surface changes.

A quantitative theory, some details of which are in the Appendix, has been formulated to explain the results. It is proposed that there are two types of recombination traps at the surface: donor type, N_a per cm^2, with energies, E_a, near the conduction band and acceptor type, N_b per cm^2, with energies, E_b, near the filled band. Surface recombination takes place by electrons and holes successively going into one of the two types of traps. To account for the fact that v_s is unchanged by changes in ambient, it is assumed that the concentrations of these traps are independent of ambient. Changes in c.p. with ambient are assumed to result from adsorption and desorption of fixed ions which are at an effective distance $\ell \sim 2 \times 10^{-7}$ cm outward from the surface traps. A schematic energy level diagram is given in Fig. 13, to be discussed later.

The charge of the ions is compensated mainly by charges in the surface traps which, together with the ions, form a double layer. A large part of the change in c.p. with ambient results from changes in this double layer. There is also a change in barrier height, $-eV_B$, associated with the redistribution of electrons in the traps. An increase in negative ions on the surface requires a decrease in number of electrons in traps, and thus a higher barrier.

Part of the change with light, $(\Delta c.p.)_L$, occurs in the body of the semiconductor and part occurs across the barrier layer. Changes in V_B

and occupancy of the traps compensate in their effect on surface recombination, so that v_s is unchanged by ambient. This means that changes in concentrations of electrons and holes in the interior with illumination and thus the body contribution to $(\Delta c.p.)_L$ are independent of ambient. Changes in $(\Delta c.p.)_L$ with ambient result soley from changes in V_B and are in the directions we have described earlier.

The concentrations of carriers, n and p, and their change with light are obtained as follows. Drift mobilities, μ_n and μ_p, and the equilibrium product, np, are known from earlier experiments of G. L. Pearson, J. R. Haynes and W. Shockley.[3] From these, and the resistivity of the sample, which was measured, n and p can be determined. The light source is calibrated in terms of hole-electron pairs created per cm^2 per sec. The recombination rate is determined from the body life time τ and the surface recombination velocity. From these latter three measurements, one can calculate the steady state density of electrons and holes near the surface when it is illuminated.

Mention should be made here of the fact that oxygen has been found to play a definite role on the Ge surface. The large extreme in dipole is obtained when active oxygen (ozone) is introduced into the gaseous ambient. Peroxide vapors have the same effect. The other extreme is produced by vapors having an OH radical, water vapor, alcohol etc. A number of vapors not falling in either of the above classes have little or no effect on the surface dipole. Another result is that the difference in work function or dipole between n- and p-type Ge is small. This is to be expected from previous work.

We shall first discuss the experimental technique and then give the experimental results. The main conclusions of the theory will then be outlined and compared with these results.

EXPERIMENTAL METHOD

The Ge surface to be measured and the reference electrode are mounted under a bell jar. Oxygen or nitrogen as desired is allowed to flow through the bell jar at a rate of approximately 2 liters per min. The volume of the bell jar is 16 liters. The gas used flows over a drying column of silica-gel and then calcium chloride. Means are provided for bubbling this gas through any desired liquid before it enters the bell jar. A spark discharge can be run in the gas flow line. The reference electrode is placed parallel to the Ge surface about 1 mm away. It is mounted on a vibrating reed which is driven, electromagnetically, at its resonant frequency of about 90 cycles per second and at an amplitude of the order of 0.1 mm. This varies the capacity sinusoidally giving rise to an electrical signal when any potential,

contact or other, exists between the surfaces. When the proper dc potential is applied between the surfaces this signal goes to zero. If no other potentials are present this dc potential is equal and opposite to the contact potential. A phase reference method* is employed to determine this balance point with a relative accuracy of $\pm 5 \times 10^{-4}$ volts. A diagram of the Ge-reference electrode circuit is shown in Fig. 1. Care must be taken to shield this circuit. Stray capacity reduces sensitivity and should be minimized. Charged insulators inside the shield will produce an apparent c.p. All conducting surfaces other than the Ge should be relatively far from the moving reference electrode. The surface is illuminated through a compound lens system by focusing the filament image, of a suitable projection bulb, on the germanium surface. This light passes through the grid of the reference electrode which removes about 10 per cent of the light. The light can be modulated by a square wave chopper, so that $(\Delta c.p.)_L$ can be measured on an ac basis. Both Ta and Pt reference electrodes have been used. The Pt electrode appears to be somewhat more constant. If the Ge surface is replaced by a gold electrode the contact potential difference is practically independent of the changes in the gas ambient. The arrangement is such that two samples can be mounted in the bell jar and the reference electrode moved from one to the other without opening the bell jar. In this way two surfaces can be compared, without any question arising of long time drifts in the reference electrode.

EXPERIMENTAL RESULTS

I. Change of c.p. between Ge and Pt reference electrodes as a function of the gaseous ambient

When this work was started the object was to find some means of varying the c.p. It was thought that the actual values of the c.p. would

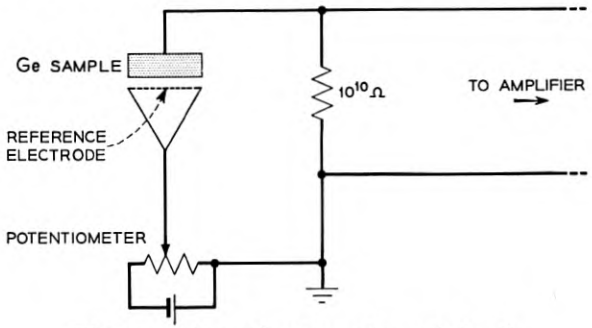

Fig. 1 — Schematic of experimental circuit.

* H. R. Moore designed and made the electronic equipment used to do this.

be highly dependent on past history of the Ge sample. If however one could measure one or more other properties, such as $(\Delta c.p.)_L$ and v_s, on the same surface at the same time, then one could look for correlations between these properties. In this manner one might be able to eliminate the past history as a factor. From previous experiments in the highly variable ambient of room air we knew the order of magnitude of the variation to be expected. At first it was impossible to produce this range of variation under the bell jar. The contact potential always drifted in the direction of a positive extreme, i.e., small total dipole at the Ge surface. The only way found to get a really large change in the opposite direction was to lift the bell jar and expose the sample to room air. These phenomena were finally traced to the presence of negative ions, possible salt ions, in the room air. These were not present in the oxygen and nitrogen supplies we used for creating the ambient in the bell jar.

It was found that the opposite c.p. extreme could be produced, under the bell jar, by running a spark discharge in dry oxygen as it was flowing into the system. The next step was to cycle the c.p. from one extreme to the other and back again. The procedure was to start with the spark discharge in dry oxygen, change to either wet O_2 or wet N_2 and to end with dry O_2. The development of this dependable and reproducible cycle was a great aid to the proposed study. Fig. 2 is a plot of contact potential versus time for a single crystal slice D of Ge cut from a melt that was p-type. The surface was prepared by removing some of the Ge with a silicon carbide (180 mesh) blast of approximately ten pounds air pressure. The Ge was then mounted in the bell jar within one-half minute of the "sandblast," and the dry O_2 flow started. The c.p. was followed for a few minutes to be sure everything was working properly.

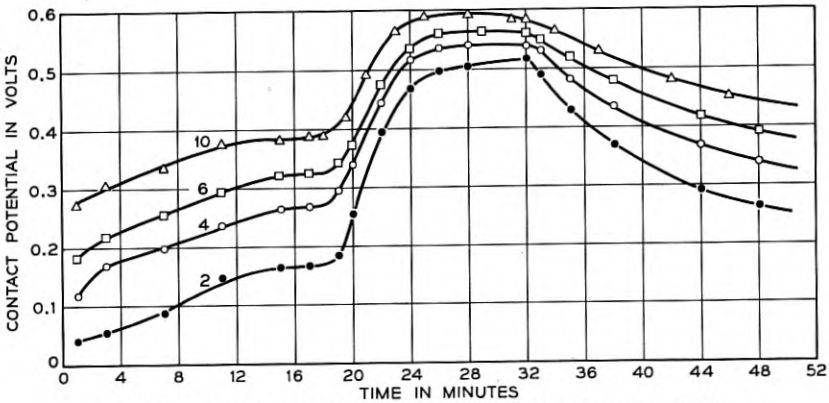

Fig. 2 — Contact potential cycles for sandblasted sample D.

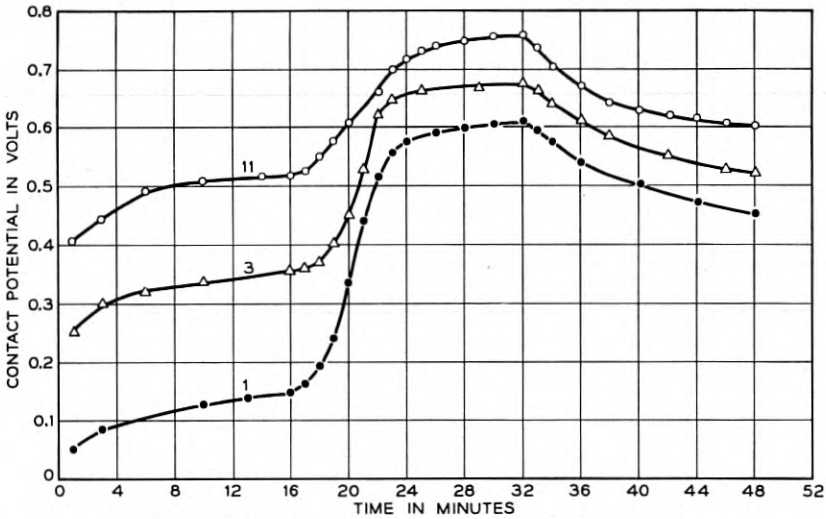

Fig. 3 — Contact potential cycles for etched sample A.

In Fig. 2 zero time is taken after the spark discharge was run in the O_2 flow line for 2 minutes. This started the first cycle. After approximately 17 minutes the O_2 was made to bubble through H_2O. Fifteen minutes or so later the O_2 was changed back to dry. At this time, 32 minutes, the flow rate was increased by a factor of three. The c.p. was followed for about 17 minutes, then the process was repeated. The results and the reason for the choice of time intervals, are all evident from a study of Fig. 2. The spark discharge in O_2 decreases the c.p. After this treatment the c.p. increases with time, most of the change occurring in the first 15 minutes. The wet O_2 then increases the c.p. to a maximum value which is reached in about 15 minutes. Finally the dry O_2 reduces the c.p. It is evident that there is a quasi-equilibrium value of c.p. in dry O_2, to which the c.p. returns after either extreme treatment. At first there is quite a large shift from cycle to cycle but this shift gradually disappears as the cycling is continued. In Fig. 2 cycles 2, 4, 6 and 10 are shown. Very little change takes place after cycle 10. Such results have been obtained many times over a period of two years. When allowance is made for shifts in work function of the reference electrode and for the fact that the experimental technique improved as the work progressed it is found that all the results for a given sample of Ge are very consistent.

In Fig. 3 are shown similar results for an n-type slice A. In this case the surface was first ground or sandblasted to remove any films, and

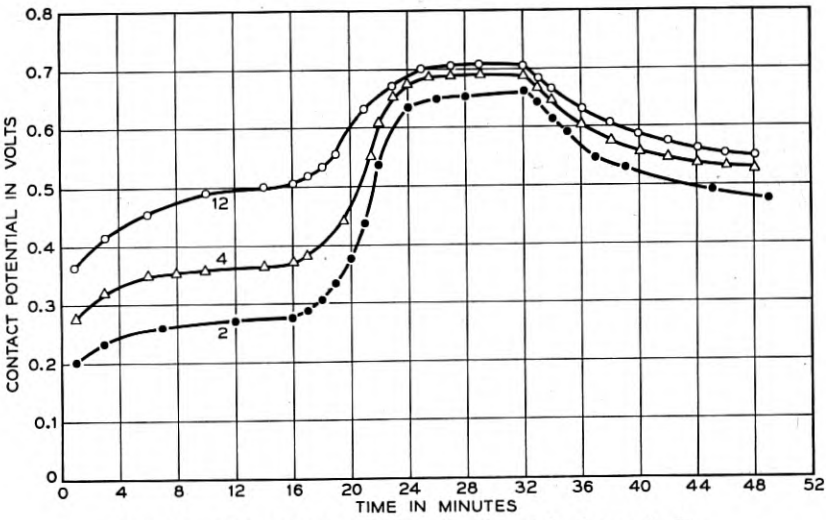

Fig. 4 — Contact potential cycles for etched sample D.

then given a polishing etch (CP-4). After the etch the surface was washed in running distilled water of reasonable quality. The surface was then dried with filter paper and kept covered until placed in the bell jar.[4] From here on the procedure was the same as before. In this case cycles 1, 3 and 11 are shown. The surprising result is the similarity between Fig. 2 and Fig. 3. To a first approximation one is tempted to say that the dipole of a Ge surface, in the bell jar atmosphere is independent of past history. Within certain limits this is approximately true. There are differences between Fig. 2 and Fig. 3 but they are small and probably due to differences in surface treatment. Fig. 4 shows the results for p-type slice, D, when this surface is etched as above. This slice and the slice A were placed in the bell jar at the same time. Cycle 1, Fig. 3, was taken on A, cycle 2, Fig. 4, on D and so on. Any differences between the results in these two figures are to be attributed to the differences in samples. They cannot very well be ascribed to the reference electrode and the initial surface treatments were as nearly the same as they could be made with reasonable care. By making runs of this type two samples at a time, different samples and different surface treatments can be intercompared. This method eliminates the shifts in the work function of the reference electrode that sometimes occur from run to run. Such results can be illustrated by plotting the data for different samples and different surface treatments for cycles 10 or greater where the results for successive cycles are the same. This has been done in Fig.

5 where one cycle each is plotted for both sample A and D, for each of the two surface treatments, sandblasting and etching. The curves are approximately all of the same shape so that the differences between them can be described by giving the shift in contact potential necessary to superimpose the curves. This treatment works very well except for the first part of the cycle after the spark coil where the shift necessary is sometimes more and sometimes less than that needed to make all of two curves superimpose well. Note in Fig. 5 that the contact potential for sample A etched is always greater than for sample D etched. In the case of the sandblasted surface just the reverse is true. Also the contact potential for the etched surface of either sample is always larger than for the sandblasted.

Using these methods comparable results were obtained on two other n-type samples, C and E, of increasingly lower specific resistance. Taking advantage of the relation between carrier concentration and the position of the Fermi level we have plotted in Fig. 6, the contact potential for both the etched and the sandblasted surfaces versus the position of the Fermi level ($E_F - E_i$) in electron volts. The contact potential values used were taken from the saturation values in wet O_2 but the shapes of the curves would be much the same for any other point in the cycle. The contact potentials are thought to be accurate to about 0.01 volts. Solid lines have been drawn through the points. Note that the contact potential for the etched surface is always greater than for the sandblasted surface, i.e., the work function is always less and that this difference is greater when $E_F = E_i$ or the germanium is nearly intrinsic.

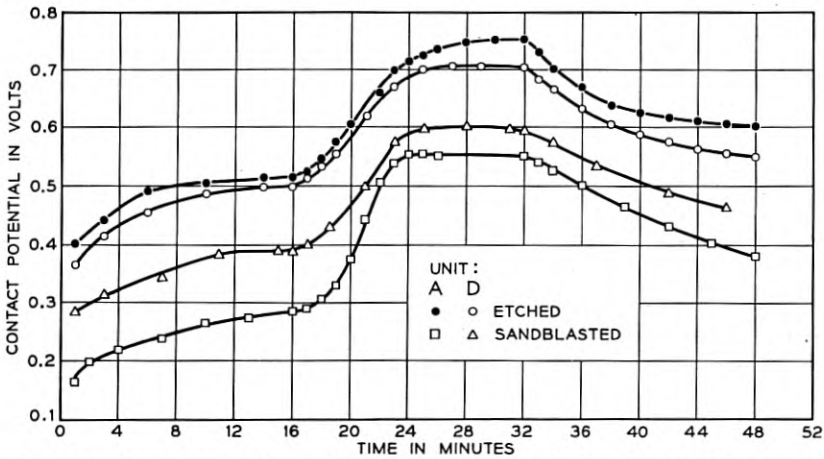

Fig. 5 — Comparison of cycles for samples A and D with sandblasted and etched surfaces.

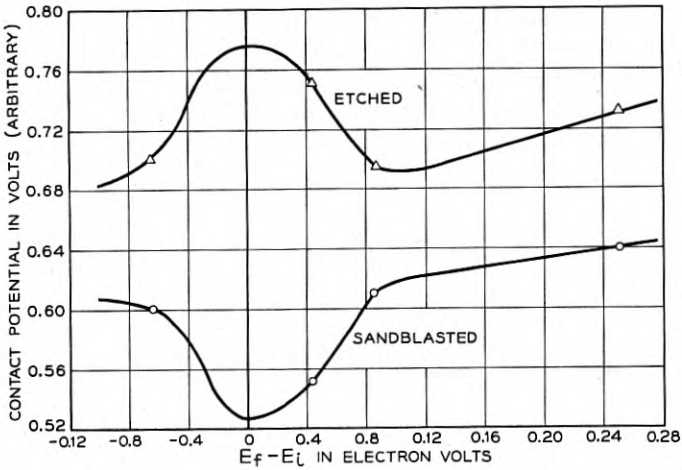

Fig. 6 — Contact potential dependence on position of the fermi level.

For variety Fig. 7 shows results on c.p. for a sandblasted polycrystalline sample of silicon. It is apparent that similar phenomena are taking place on the silicon surface, however the behavior is different. These preliminary results are shown to illustrate the generality of this method for investigating semiconductor surfaces in a controlled gaseous ambient.

The time between cycles was somewhat variable. The first and second cycles were always taken immediately after the specimen was placed in the bell jar. After this, successive cycles were taken one or two a day over a period of a week or more. The bell jar was not opened during a run and a small flow of dry gas was maintained between cycles. Little if anything occurred during these idle periods, showing that the changes that did take place were due to the cycling.

II. Change of contact potential with illumination

When the Ge surface is sandblasted the change of contact potential with illumination $(\Delta c.p.)_L$ is too small to be measured. If however the surface has been prepared by the polishing etch, the $(\Delta c.p.)_L$ is easily observed. This change if not too small can be measured by finding the balance on the potentiometer for light off and light on. When the contact potential is changing with time, or if the change is small this is difficult to do. A better procedure is to chop the light at a definite frequency and measure the amplified output on an ac meter. This gives a continuous reading that can be read easily at any given time. If a filter is used to pass only those frequencies near the fundamental of

the chopping frequency, the improvement in signal to noise enables one to read very small changes.

A plot of $(\Delta c.p.)_L$ in volts versus the contact potential for samples A n-type and D p-type is shown in Fig. 8. For the n-type sample the signal is large and positive when the contact potential is small and becomes small and negative as the contact potential increases. Except for the shift in the contact potential where $(\Delta c.p.)_L$ goes through zero, the results for the p-type sample are practically the opposite of those for the n-type sample. Similar curves have been obtained cycle after cycle in many complete runs. Within experimental error the curves have the same shape for a given sample for all cycles in all runs. Sometimes the curve for the first cycle in a run will differ in shape from the rest. However there are shifts such that the c.p. for which the light goes through zero $(c.p.)_0$ does vary from cycle to cycle throughout a run. Fig. 9 is a plot of $(c.p.)_0$ versus cycle number for p-type sample D and n-type sample A. The data are plotted for two distinct runs. Both of these units were in the bell jar when the measurements were taken. Consistent results of this kind give one confidence that the Pt reference electrode is staying constant.

These experimental results can be summarized as follows. All data for a given sample can be superimposed by shifts in contact potential scales for the different cycles. A plot of $(\Delta c.p.)_L$ versus c.p. $-$ $(c.p.)_0$ can be represented by a single curve for each unit. Moreover all the curves for all units both n- and p-type have quite similar shapes provided in the latter we plot $[-(\Delta c.p.)_L]$ versus $(c.p.)_0$ $-$ c.p. The shape of this curve is shown in Fig. 8 for a given sample and Fig. 9 shows how $(c.p.)_0$ varies from cycle to cycle. These plots adequately describe the results.

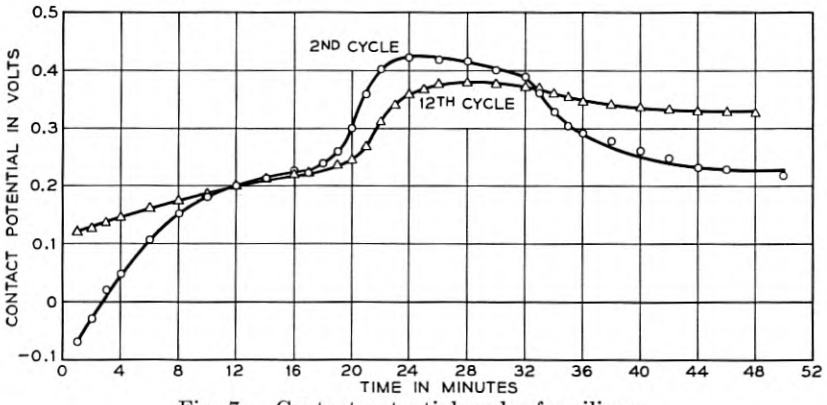

Fig. 7 — Contact potential cycles for silicon.

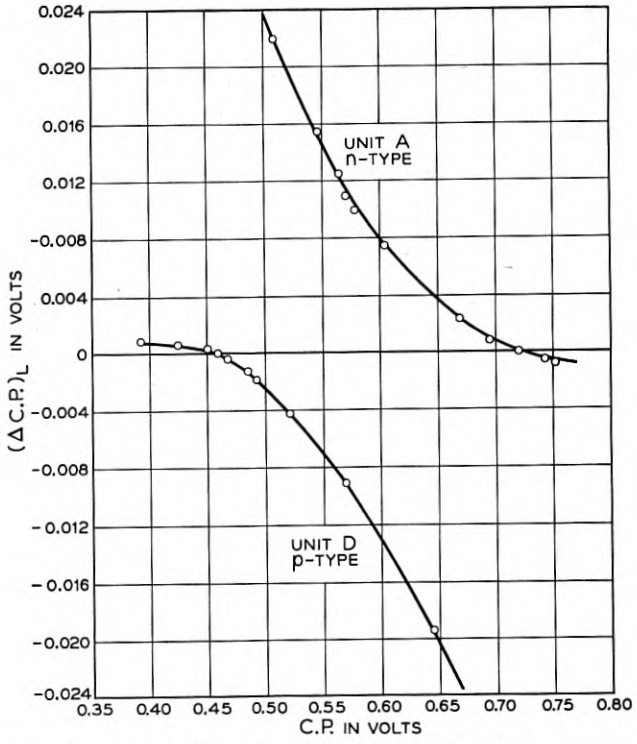

Fig. 8 — Change of contact potential with illumination versus contact potential, samples A and D etched.

Not shown in Fig. 8 because of the scale used is the result that $(\Delta c.p.)_L$ for n-type material does not increase indefinitely as c.p. decreases but approaches a maximum value. Likewise $-(\Delta c.p.)_L$ for p-type approaches a maximum as c.p. increases. Figures illustrating these results will be discussed after the theory is presented.

Some of the experimental details require discussion. It is necessary to calibrate the ac response in terms of absolute potential change. This can be done by comparing the ac reading with the dc reading when the light signal is large. One can also do this by introducing a known square-wave signal across the potentiometer in Fig. 1, and reading the ac signal out. Both methods agree when allowance is made for variation of the light signal with frequency. The latter response is almost flat from 25 to 300 cycles, but there is evidence for some very low frequency components in the dc measurements of $(\Delta c.p.)_L$. When the light signal goes through zero the signal is small and the dc value is difficult to read.

At times evidence has been obtained to indicate that the dc value changed sign. At other times the dc $(\Delta c.p.)_L$ behaved as if the place where it went through zero was shifted in c.p. from the point where the ac signal goes through zero. In view of this it was necessary to prove that the ac signal was changing phase at the zero point and not just going down in the noise and then increasing again without phase change. This was done by comparing the phase of the signal with a signal from a photocell placed in the same chopped light beam. By this means it was proved conclusively that the ac light signal was actually going through zero.

Some data were obtained on change of contact potential with light on *n*-type samples C and E having progressively smaller specific resistances. It was found that $(\Delta c.p.)_L$ decreased with specific resistance. It also decreased into the noise as the contact potential was increased, so that it could not be determined if it changed sign as for samples A and D. Because of the smaller signal $(\Delta c.p.)_L$ could not be measured easily except by the ac method and so far the signal has not been calibrated properly.

Some preliminary data on *p*-type silicon indicate that $(\Delta c.p.)_L$ for this sample was negative and that it decreased as c.p. was decreased. The magnitude of $(\Delta c.p.)_L$ for the same light intensity was much larger than for germanium and so far it has not been found to go through zero and change sign in the experimental range.

III. Other methods of varying the c.p.

In some cases N_2 was used in place of O_2. A spark discharge in the N_2 had very little effect on the c.p. On the other hand wet N_2 produced much the same effect as wet O_2. The positive extreme in c.p. was about 0.1 volt greater in the case of wet N_2. After the wet treatment dry N_2 was not nearly as effective as dry O_2 in reducing the c.p. to its intermediate value. This can hardly be due to a difference in dryness of the two gases since the same drying column was used in both cases. The results indicate that dry N_2 tended to leave the surface in whatever condition obtained before the dry N_2 flow was started, and that O_2 counteracts the effect of H_2O.

With dry N_2 as a carrier, other vapors were tried. A. N. Holden suggested trying a peroxide and picked out ditertiary butyl peroxide as being reasonably safe. Use of this vapor was found to produce the same changes as the spark coil in the O_2 flow. Other vapors having OH radicals such as methyl alcohol and acetic acid were found to act the same way

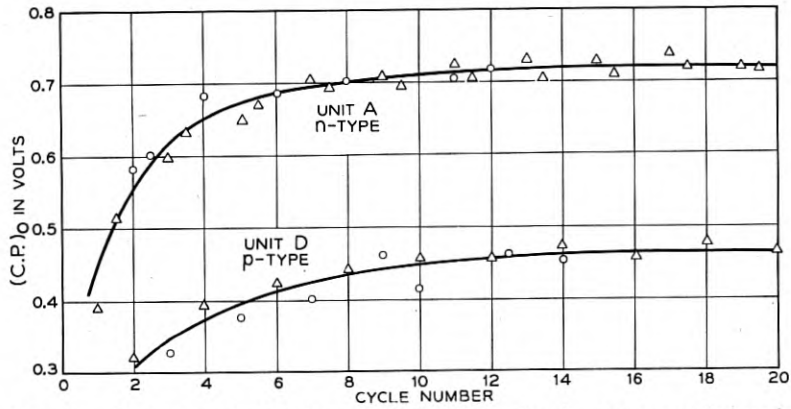

Fig. 9 — Contact potential for zero light effect, $(c.p.)_0$, versus cycle number, two runs for each sample.

as water vapor. To prove that this was not caused by traces of water in the alcohol or acid the N_2 was bubbled through water solutions of H_2SO_4. These results indicated that one needs appreciable amounts of water vapor to produce the effect, much more than could be present in the alcohol or acid. Other vapors, such as carbon tetrachloride, methyl-chloride, nitrobenzene and ether, were found to have no effect on either the contact potential or $(\Delta c.p.)_L$. Acetone has a small effect in the same direction as water. This is to be expected because this compound exists in part in a tautomeric form having an OH group. Vapor from 30 per cent H_2O_2, 70 per cent H_2O acted at first like a peroxide vapor and with a longer time of exposure behaved like water vapor. Small amounts of Cl_2 gas in N_2 produced the same change as the spark discharge in O_2 and after 14 minutes of flushing the bell jar with N_2 produced an additional effect when water vapor was introduced. On *n*-type samples before the usual increase in c.p. and decrease in $(\Delta c.p.)_L$ there was a large increase in $(\Delta c.p.)_L$. This was attributed to the reaction between the water vapor and the Cl left on the Ge surface, producing oxygen. The nature of the change was a rapid shift in $(\Delta c.p.)_0$ and thus a momentary increase in $(\Delta c.p.)_L$. This effect is only large in the first cycle. It indicates that the shifts in $(\Delta c.p.)_0$ plotted in Fig. 9 are probably due to an oxidation of the Ge surface as the cycling progresses. In all these experiments the relation between the c.p. and $(\Delta c.p.)_L$ was essentially the same as that obtained in the standard cycle.

One can detect the presence of thin surface films by electron diffraction techniques. R. D. Heidenreich took electron diffraction pictures of a germanium surface immediately after the polishing etch and water

wash. He found the surface to be quite clean, with a film thickness less than 10 Å. He also took pictures after the germanium surface had been cycled about fifteen times and found in this case definite evidence for a thin surface film. The film was either amorphorus or composed of very small crystals. He estimated the thickness to be between 20 and 50 Å.

In the discussion of the experimental work it has been assumed that all the changes in c.p. are to be attributed to the Ge surface and not to the Pt. It is not easy to give a definite proof that this is true. The fact that it is a reasonable assumption is suggested by the nature of the results themselves. Almost identical results were obtained using a Ta electrode. In Fig. 10 we show the results of two cycles when the Ge was replaced with gold. While there are some changes they are almost an order of magnitude smaller than the changes when Ge is present. It follows that one would expect much the same results with either Ta, Pt or Au reference electrodes and that most of the changes are due to the Ge. No change of c.p. with illumination is observed when both electrodes are metals. In one case a small amount of H_2 was added to the N_2 flow. Here the c.p. between Pt and Ge changed rapidly but the $(\Delta c.p.)_L$ did not change at all. Subsequent runs using the regular cycle indicated that the c.p. scale had been shifted corresponding to a reduction in work function of the Pt. Except for this shift the results were the same and the shift disappeared in about one day. The conclusion was that H_2 had little effect on Ge but reacted with the Pt decreasing its work function. This is a good illustration of the power of this method of measuring more than one property of a semiconductor surface at the same time. If only c.p. had been measured the conclusions would not have been so clear cut.

If one knew the work function of the Pt electrode then one would know the work function of the Ge. Work functions are all measured in high vacuum. We have not been able to think of a method of determining the work function of any electrode in a gaseous ambient unam-

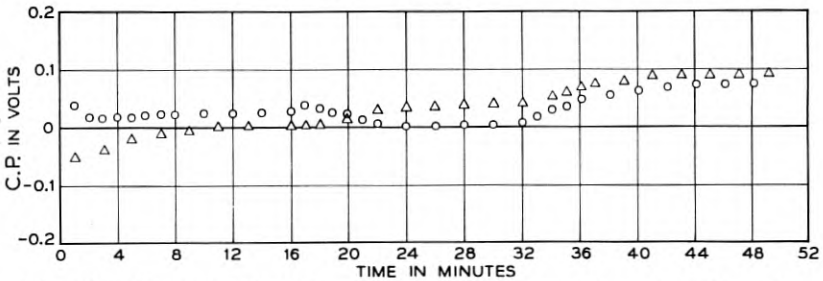

Fig. 10 — Contact potential cycles when germanium is replaced by gold.

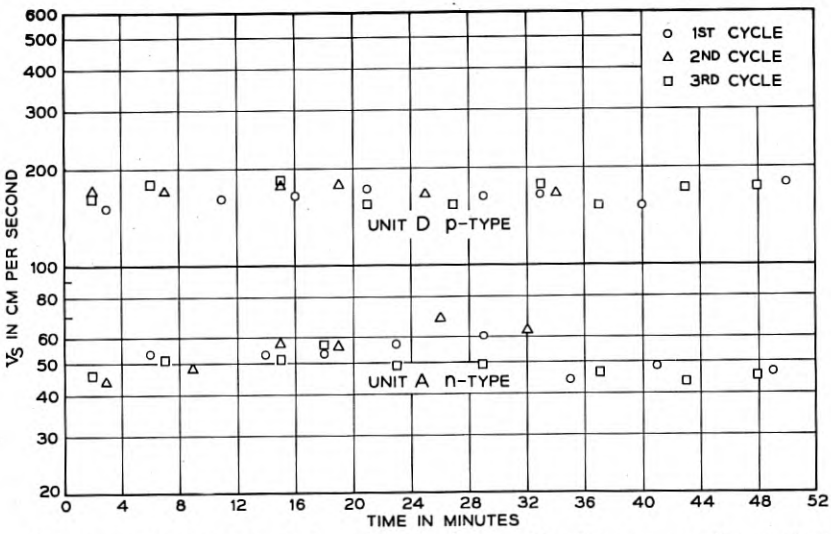

Fig. 11 — Dependence of the velocity of recombination, v_s, on the contact potential cycles.

biguously. It may well be impossible to do this in Bridgman's operational sense. Most physicists would agree that the Pt reference electrode probably has a work function of around 5 to 7 volts under these conditions, but this of course does not help much.

IV. Measurement of surface recombination velocity v_s

At first we tried various methods of measuring v_s on the Ge surface at the same time that c.p. and $(\Delta \text{c.p.})_L$ were measured. No method was found that could be trusted. As the study progressed we realized that the c.p. of a Ge surface could be cycled in a reproducible way. Since the proper geometry for measuring v_s was a rod or filament,[2] rods were prepared of samples A and D with the same chemical surface treatment. The decay of hole electron pairs, created by a point light source, was measured as a function of distance from the light. If the body life time τ and the dimensions of the filament are known one can then determine v_s.[2] These measurements were made while the gaseous ambient was cycled in the same manner as before. The results for filaments cut from samples A and D are plotted against time in Fig. 11. The ambient atmosphere was changed as a function of time just as it was in Figs. 2 to 4. The main conclusion is that within the accuracy of this experiment there is no evident dependence of v_s on the changes in

gas ambient and therefore no dependence on the corresponding changes in surface dipole. This result was somewhat unexpected and the first time the experiment was performed it was hard to believe that the previously measured changes in c.p. actually were taking place. In this case the gas ambient was experimented with to try to change v_s and it was found that v_s could be changed from the order of 10^2 cm/sec to greater than 10^5 cm/sec and back again by exposure of the filament to $(NH)_4OH$ fumes and then HCl fumes respectively.[5] While interesting, this has no direct bearing on the other experiments. The experiment was performed again with freshly prepared rods and this time the cycling used in the c.p. experiments was rigidly adhered to, giving v_s equal to 70 cm/sec and 200 cm/sec approximately for samples A and D respectively. The experiments were then repeated again using new filaments cut from the samples as close as possible to the surfaces used in the c.p. experiments leading to the results shown in Fig. 11, namely, v_s equal to 50 cm/sec and 170 cm/sec respectively. From these experiments it was concluded that v_s is approximately constant in the range involved and is determined by the nature of the sample and the surface treatment used. It was noted in some of these experiments that v_s for the first cycle was somewhat larger than for the subsequent cycles. This change when it occurred is probably to be correlated with the changes in the early cycles in the c.p. measurements. Similar measurements on sample C gave v_s equal to 1500 cm/sec. No measurements of v_s were made on samples B and E.

V. Other experimental measurements

The specific resistance of each sample was measured near the surface used in the experiments. It was approximately constant across the surface but did vary slowly with depth in some of the samples.

The body life times were measured on each sample. The thickness of the slices used was intentionally made large compared to their corresponding diffusion lengths, about 0.5 cm for A, B, C and E and about 2.0 cm for D. The mobilities were taken from J. R. Haynes[3] measurements: $\mu_n = 3600$ and $\mu_p = 1700$ cm^2/volt sec. There is some uncertainty as regards the exact value of the equilibrium product of holes and electrons, np, at 300°K. We have used the value 6.3×10^{26} obtained from some unpublished data of G. L. Pearson.[3]

The light source was calibrated by replacing the germanium sample with one of F. S. Goucher's n-p junctions.[6] The bell jar, with everything else including the reference electrode, was left in their normal

positions. An average was taken over the filament image, and the effective area of the p-n junction was determined and allowed for. This was done for all light intensities used. Most of the experiments were performed with a fixed intensity and the averaged result for this intensity was 6.0×10^{15}, hole electron pairs per cm^2 sec. The rate of pair production was found to be proportional to the light intensity.

Since practically all the light is absorbed in a depth (10^{-4} cm or less) that is small compared to the diffusion length, it is a simple matter to calculate the steady state increase δp in hole electron pairs due to the light. The relation is

$$\delta p = N/(v_s + v_d) \tag{1}$$

where N is the rate of pair production, v_s is the velocity of recombination at the surface and v_d is the diffusion velocity for the minority carrier. Since N is proportional to light intensity it follows that δp is too.

The magnitude of $(\Delta \text{c.p.})_L$ should depend on the light intensity. One might at first expect it to be proportional to light intensity. That this is not the case is shown by curve 1 in Fig. 12. Curve 1 is a plot of $(\Delta \text{c.p.})_L$ versus δp for unit D on a log-log scale. A smooth curve has been drawn through the experimental points. As we shall see later, theory predicts that if $(\Delta \text{c.p.})_L$ is large, it should be proportional to $\ell n(1 + \delta p/a)$ where a is the equilibrium density of the minority carrier,

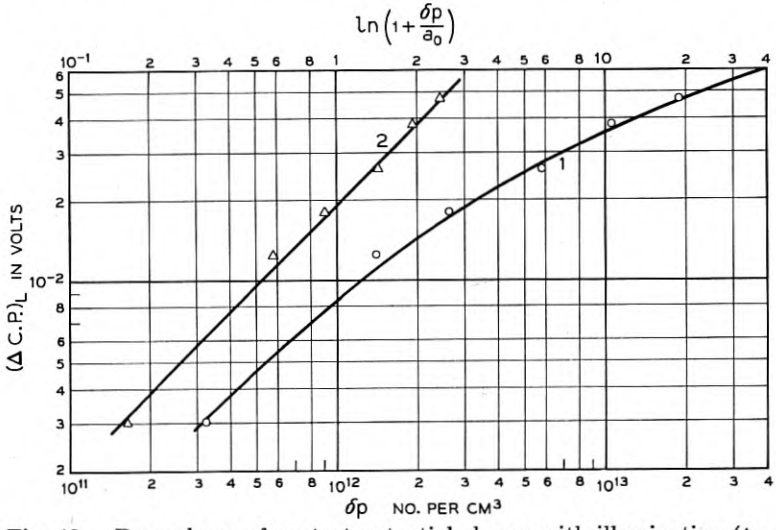

Fig. 12 — Dependence of contact potential change with illumination $(\Delta \text{c.p.})_L$ on light intensity.

n in p-type material and p in n-type. That this prediction is borne out is shown by how well the experimental points fit the straight line curve 2 in Fig. 12 where $(\Delta c.p.)_L$ is plotted versus this quantity. The scale for δp is shown along the bottom of Fig. 12 and that for $\ell n(1 + \delta p/a)$ along the top. Similar results were obtained for unit A.

In Table I the parameters, specific resistance ρ in ohm cm, life time τ in microseconds and the surface recombination velocity v_s in cm/sec are given for each unit used. Also given are some pertinent quantities derived therefrom, namely the equilibrium densities of electrons and holes, n and p in number per cm^3 and the increase in density at the surface δp in number per cm^3 when the rate of pair production due to the illumination was 6.0×10^{15} per cm^2 sec.

THEORY

The constancy of v_s throughout the range of surface dipole investigated puts rather stringent requirements on any theoretical model to be constructed. W. Shockley and W. T. Read[7] have investigated the theory of recombination via traps. It is evident from their work that if one assumes a trap density peaked near a single energy and very small elsewhere, then v_s will be constant over a range of surface dipole values provided that the peak energy is either near the conduction band or near the filled band. The experimental results make it appear very unlikely that the trapping mechanisms on the n and p-type surfaces are essentially different. It is assumed that both types of traps are present on the surface and that the traps are approximately the same for both n- and p-type samples. Further, it is assumed that the traps N_a of energy E_a near the conduction band are donor type, i.e., neutral when filled and positively charged when empty. Likewise the traps N_b of energy E_b near the filled band are assumed to be acceptor-type traps, i.e., neutral when empty and negatively charged when filled. The absolute charge on the traps is not important, however, because we are

TABLE I

Sample	Type	ρ	n_0	p_0	τ	v_s	δp	C
A	n	12.5	1.38×10^{14}	4.56×10^{12}	900	50–70	2.2×10^{13}	4.4×10^{-13}
B	n	15	1.14×10^{14}	5.6×10^{12}	600	(100)	1.9×10^{13}	—
D	p	12.0	2.1×10^{12}	3.0×10^{14}	4000	170–200	1.75×10^{13}	6.0×10^{-13}
C	n	2.5	7.0×10^{14}	0.91×10^{12}	48	1.5×10^{3}	1.67×10^{12}	20.0×10^{-13}
E	n	0.008	4.4×10^{17}	1.45×10^{9}	—	—	—	—

concerned only with differences between the occupied and unoccupied states.

The fraction of the traps which are occupied depends on the trap energy and on the position of the Fermi level at the surface. The latter in turn depends on the condition that the surface as a whole be neutral. A very obvious mechanism for changing the total surface dipole is the adsorption or desorption of ions on the surface. If this happens the position of the Fermi level at the surface shifts until the total charge on the surface, adsorbed ions, charge in the traps and charge in space-charge layer, adds up to zero. The consequences of this model have been carried through.

The basis for the theory is illustrated in the energy level diagram of Fig. 13. At the semiconductor surface there is a space-charge layer of thickness ℓ_B which gives a change in electrostatic potential of V_B, corresponding to a potential energy of an electron of $-eV_B$. Outside of the surface of the germanium proper, there is a surface film of thickness ℓ_D. A double layer giving a potential change V_D, is formed from a charge of ions, σ_I, on the outer surface of the film and charges in the surface traps of types a and b. Changes in c.p. with ambient result from changes in σ_I and consequent changes in V_B and V_D. It is assumed that the remainder of the work function is independent of ambi-

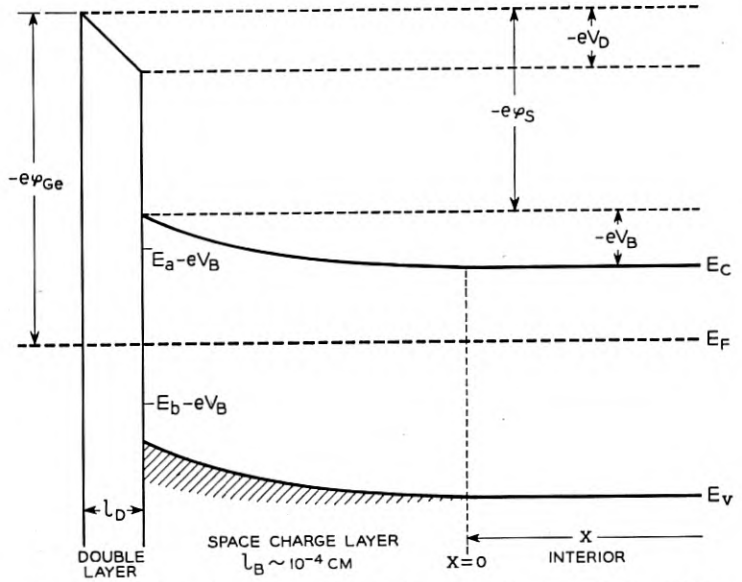

Fig. 13 — Schematic of energy level diagram at germanium surface.

ent. When light shines on the surface, V_B is changed to $V_B + \delta V_B$ and there is an additional potential drop, δV_i, in the body of the germanium resulting from the recombination current which flows to the interior. The change in contact potential with light is equal to $\delta V_B + \delta V_i$.

The film thickness ℓ_D is shown on an exaggerated scale. We expect $\ell_D \sim 10^{-6}$ cm and $\ell_B \sim 10^{-4}$ cm, so that $\ell_B \gg \ell_D$.

TABLE OF SYMBOLS

A. Energies:

E_a = E_a (true) + kT ln $(\omega_{unoc}/\omega_{oc})$, is the effective energy of the a-traps for $V_B = 0$. Here ω_{unoc} and ω_{oc} are the statistical weights of the unoccupied and occupied states, respectively.

E_b = effective energy of the b-traps for $V_B = 0$.

E_c = energy of lowest state of conduction band in interior of semiconductor just beyond the space-charge layer.

E_v = energy of highest state of valence band at the same position.

E_F = Fermi energy.

E_i = E_F when material is intrinsic.

V_D = potential drop across surface film.

V_B = potential drop across space-charge layer.

V_{B0} = value of V_B for which $n_a = p_b$, see below.

V_0 = value of V_{B0} for an intrinsic sample.

B. Concentrations:

n = $N_c \exp [(E_F - E_c)/kT]$ = equilibrium concentration (no./cm^3) of conduction electrons in interior of semiconductor just beyond the space-charge layer.

p = $N_v \exp [(E_v - E_F)/kT]$ = corresponding hole concentration.

n_i = intrinsic concentration.

n_s, p_s = equilibrium concentrations of electrons and holes, respectively, at the surface.

N_a, N_b = concentration (no./cm^2) of a- and b-traps, respectively.

n_a = equilibrium concentration (no./cm^2) of occupied a-traps.

p_b = $N_b - n_b$ = equilibrium concentration (no./cm^2) of unoccupied b-traps.

n_{a0}, p_{b0} = values of n_a and p_b for an intrinsic sample with $V_B = 0$.

n_1 = $n + \delta n$, $p_1 = p + \delta p$, and n_{s1}, p_{s1}, n_{a1}, p_{b1} = concentrations in presence of light. Electrical neutrality requires that $\delta n = \delta p$.

The theory is based on the following postulates:

I. Changes in c.p. with ambient result from changes in σ_I and the consequent changes in V_B and V_D.

$$\text{c.p.} = V_B + V_D + \text{const.} \tag{2}$$

The charge σ_I is largely compensated by charges in the surface traps. The barrier height, V_B, is determined by the requirement of electrical neutrality:

$$\sigma_I = e(n_a - p_b) + \text{const.} \tag{3}$$

Since V_B and V_D are of the same order, the net charge per unit area in the space-charge layer will be smaller than σ_I in the approximate ratio ℓ_D/ℓ_B and may be neglected.

II. Traps of type a have energies above E_F and of type b below E_F for all values of V_B attained in the different ambients used. More exactly

$$E_a - eV_B - E_F > kT, \tag{4}$$

$$E_F - E_b + eV_B > kT, \tag{5}$$

for all V_B. One may then use the Boltzmann approximations for n_a and p_b:

$$n_a = \frac{N_a}{1 + \exp\left[(E_a - eV_B - E_F)/kT\right]} \sim N_a \exp\left[(E_F - E_a + eV_B)/kT\right], \tag{6}$$

$$p_b = \frac{N_b}{1 + \exp\left[(E_F - E_b + eV_B)/kT\right]} \sim N_b \exp\left[(E_b - eV_B - E_F)/kT\right]. \tag{7}$$

It is not necessary for our arguments to assume that all traps of each type have the same energy. The only requirement is that the distributions of trap energies are such that the a-traps are always above and the b-traps always below the Fermi level for any ambient.

III. Creation of electron-hole pairs by absorption of light occurs near the surface in a distance that is small compared with the diffusion length. Optical constants of germanium indicate that practically all of the light with energy sufficient to create electron-hole pairs is absorbed within a distance of 10^{-4} cm of the surface. The diffusion length is of the order of 0.2 cm.

IV. In the presence of light, the concentration of electrons in a-traps

is in equilibrium with the concentration of electrons in the conduction band and holes in b-traps are in equilibrium with holes in the valence band. The barrier height is adjusted so that the total charge in both types of traps is unchanged by illumination. We shall show in the appendix that the resistance to flow of electrons from the conduction band across the space-charge layer and into a-traps is small compared with the resistance to flow of electrons from the valence band to the traps. Similar considerations apply to flow of holes to b-traps from the valence band as compared with flow from the conduction band.

V. Recombination is limited by holes going into traps of type a and electrons going into traps of type b. The two types of traps act in parallel for recombination. The contributions to the surface recombination velocity are proportional to $p_s n_a$ and $n_s p_b$, respectively. These products are independent of V_B and thus of ambient if postulates II and IV are satisfied. If other types of traps were important in recombination, one would expect v_s to depend on ambient, contrary to what is observed.

Postulates I and II are used to relate V_B and c.p. with changes in ambient. Postulates III, IV, and V are used to relate $(\Delta \text{c.p.})_L$ with V_B and the trap densities, and also to obtain an expression for the surface recombination velocity.

As V_B is made more positive, corresponding to a decrease in barrier height for electrons, n_a increases and p_b decreases. It will be convenient for the theoretical discussion to introduce the particular barrier potential V_{BO}, for which $n_a = p_b$. With use of the Boltzmann approximations in Equations (6) and (7), this gives

$$\begin{aligned} \exp [2\beta V_{BO}] &= (N_b/N_a) \exp [(E_a + E_b - 2E_F)/kT)] \\ &= (p_{b0}/n_{a0})(p/n), \end{aligned} \tag{8}$$

where $\beta = e/kT$. The last form follows from the definitions of p_{b0} and n_{a0} and by noting that $\exp [2(E_i - E_F)/kT] = p/n$. We then have

$$n_a/p_b = \exp [2\beta(V_B - V_{BO})]. \tag{9}$$

As so defined, V_{BO} depends on the Fermi level and thus on the conductivity of the specimen. We shall let V_0 be the value of V_{BO} for an intrinsic specimen. From (8)

$$n_{a0}/p_{b0} = \exp [-2\beta V_0]. \tag{10}$$

If $n_{a0} = p_{b0}$, then V_0 will be zero. Postulate II sets limits on V_0, but V_0 is otherwise undetermined in our experiments.

We shall first relate V_D and V_B. Since the electric field in the film is $4\pi\sigma_I/K_D$,

$$V_D = 4\pi\ell_D\sigma_I/K_D, \tag{11}$$

where K_D is the dielectric constant of the film. Substituting for σ_I from Equation (3) and making use of (8) and (9), we find

$$V_D = 2H \sinh \beta(V_B - V_{BO}) + \text{const}, \tag{12}$$

where

$$H = \frac{4\pi e\ell_D}{K_D}(N_aN_b \exp[(E_b - E_a)/kT])^{\frac{1}{2}} = (4\pi e\ell_D/K_D)(n_{a0}p_{b0})^{\frac{1}{2}}. \tag{13}$$

If V_D is expressed in volts and ℓ_D in cm.

$$H = 1.8 \times 10^{-6}(\ell_D/K_D)(n_{a0}p_{b0})^{\frac{1}{2}}. \tag{13a}$$

The contact potential Equation (2) may be expressed in the form

$$\text{c.p.} = V_B - V_{BO} + 2H \sinh \beta(V_B - V_{BO}) + \text{const}. \tag{14}$$

The change in contact potential with illumination results from a potential drop δV_i in the interior and a drop δV_B across the space-charge layer. The former comes from the recombination current of holes and electrons diffusing from the surface to the interior. Since electrical neutrality requires that $\delta n = \delta p$, the concentration gradients are equal. However, the mobility of electrons is greater than that of holes, so that the diffusion current of electrons is larger than that of holes by the mobility ratio "b." Since there can be no net current flow to the interior, an electric field is established which is in such a direction as to enhance the flow of holes and retard the flow of electrons. The net potential drop associated with this electric field is

$$\delta V_i = \frac{(b-1)}{\beta(b+1)} \ln\left\{1 + \frac{(b+1)\delta p}{bn + p}\right\} \tag{15}$$

where $\delta p = \delta n$ is the change in concentration in the interior just beyond the space-charge layer.[8]

This potential is positive for both n- and p-type material, and is independent of ambient, since δp is independent of ambient. The observed $(\Delta\text{c.p.})_L$ is generally much larger than given by (15) and is of opposite sign for n- and p-type material. To account for the observations, it is necessary to assume that the major part of the effect is associated with a space-charge layer at the free surface. We believe that the present experiments give the most convincing evidence obtained so far for the existence of such a space-charge layer.[1]

The value of δV_B, the change in potential across the space-charge layer due to light, is determined by the requirement that there be no net change in charge in the surface traps, or that

$$\delta n_a = \delta p_b . \tag{16}$$

The changes δn_a and δp_b come both from changes in δp and δn and from δV_B. According to postulate IV, $n_{a1} = n_a + \delta n_a$ is in equilibrium with the conduction band and $p_{b1} = p_b + \delta p_b$ with the valence band. We have then

$$n_{a1}/n_a = n_{s1}/n_s = (n_1/n) \exp [\beta \delta V_B], \tag{17}$$

$$p_{b1}/p_b = p_{s1}/p_s = (p_1/p) \exp [-\beta \delta V_B]. \tag{18}$$

from which it follows that

$$\delta n_a/n_a = (n_1/n) \exp [\beta \delta V_B] - 1, \tag{19}$$

$$\delta p_b/p_b = (p_1/p) \exp [-\beta \delta V_B] - 1, \tag{20}$$

$$\frac{n_a}{p_b} = \exp [2\beta(V_B - V_{BO})] = \frac{(p_1/p) \exp [-\beta \delta V_B] - 1}{(n_1/n) \exp [\beta \delta V_B] - 1} . \tag{21}$$

Equation (21) is a quadratic equation in $\exp [\beta \delta V_B]$ which may be solved to give an explicit expression for δV_B. The role of electrons and holes may be interchanged by changing the sign of δV_B and of $V_B - V_{BO}$. This accounts for the difference in behavior of n- and p-type samples. The total change with light is the sum of δV_i and δV_B:

$$(\Delta \text{c.p.})_L = \delta V_L = \delta V_B + \delta V_i . \tag{22}$$

In the analysis of the data, δV_i is calculated theoretically from (15) with $\delta p = \delta n$ determined from (1) and δV_B is obtained from the observed $(\Delta \text{c.p.})_L$ and δV_i using (22). Equation (21) is then used to find $V_B - V_{BO}$. A plot of c.p. $- (V_B - V_{BO})$ versus $\sinh \beta(V_B - V_{BO})$ should be a straight line with slope $2H$. An analysis of the observed data in this manner is given in the following section.

We turn finally to a discussion of the surface recombination velocity, v_s. According to postulate V, recombination is limited by flow of holes to a-traps and of conduction electrons to b-traps. The flow of holes from the valence band to a-traps (really electrons from a-traps drop into the vacant levels in the valence band corresponding to the holes) is proportional to the product of the hole concentration at the surface p_{s1}, and the concentration of electrons in a-traps, n_{a1}. The reverse flow is that of thermal generation of holes: electrons from the valence

band go into unoccupied a-traps. Since, according to postulate II, the number of unoccupied traps is nearly equal to the total number of traps, and is thus independent of δV_B and δp, the reverse flow will be practically equal to the thermal equilibrium value. If S_{pa} is the recombination cross-section for holes, the net flow of holes to a-traps is

$$U_{pa} = S_{pa} v_p (p_{s1} n_{a1} - p_s n_a) (\text{holes/cm}^2 \text{ sec}). \tag{23}$$

Here, v_p is the velocity factor which when multiplied by the concentration gives the number of holes crossing a unit area from one direction:

$$v_p = (kT/2\pi m_p)^{\frac{1}{2}}. \tag{24}$$

With use of the relations, $n_{a1}/n_a = n_{s1}/n_s$, $p_{b1}/p_b = p_{s1}/p_s$, from equations (17) and (18), which follow from the fact that the a-traps are in thermal equilibrium with the conduction band and the b-traps with the valence band, (23) becomes

$$U_{pa} = S_{pa} v_p (n_a/n_s)(p_{s1} n_{s1} - p_s n_s). \tag{25}$$

This expression may be simplified further. The ratio

$$\frac{n_a}{n_s} = \frac{N_a \exp\left[(E_F - E_a - eV_B)/kT\right]}{N_c \exp\left[(E_F - E_c - eV_B)/kT\right]} = \frac{N_a}{N_c} \exp\left(\frac{E_c - E_a}{kT}\right) \tag{26}$$

is independent of E_F and V_B. The ratio may be evaluated for an intrinsic specimen with $V_B = 0$, in which case $n_a = n_{a0}$ and $n_s = n_i$. Thus

$$n_a/n_s = n_{a0}/n_i. \tag{27}$$

We also have from postulate IV,

$$p_{s1} n_{s1} = p_1 n_1. \tag{28}$$

The equilibrium products are

$$p_s n_s = pn = n_i^2. \tag{29}$$

With use of (27), (28) and (29), (25) becomes

$$U_{pa} = S_{pa} v_p (n_{a0}/n_i)(p_1 n_1 - pn). \tag{30}$$

Similarly, it is found that net flow of electrons to b-traps is equal to:

$$U_{nb} = S_{nb} v_n (p_{b0}/n_i)(p_1 n_1 - pn). \tag{31}$$

The total rate of recombination is given by the sum of U_{nb} and U_{pa} and is given by an expression of the form:

$$U = U_{nb} + U_{pa} = C(p_1 n_1 - pn) = C(n + p)\delta p. \tag{32}$$

It should be noted that the coefficient C is independent of the Fermi level and thus of the conductivity of the specimen, whereas v_s is not. The relation between them is:

For n-type $$C = v_s/n, \tag{33}$$

For p-type $$C = v_s/p. \tag{34}$$

Values of C may be determined empirically from observed values of v_s. Referring to Table I, for sample A, $v_s = 60$ cm/sec and $n = 1.4 \times 10^{14}/\text{cm}^3$, so that $C = 4.3 \times 10^{-13}$ cm^4/sec. For sample D, $v_s = 180$ cm/sec and $p = 3.0 \times 10^{14}/\text{cm}^3$, so that $C = 6 \times 10^{-13}$ cm^4/sec. The values of C are approximately the same, indicating that the traps are not much different for the two specimens.

The theoretical value of C involves n_{a0} and p_{b0}. It is the product $n_a p_b = n_{a0} p_{b0}$, which is related to the parameter H and which can be estimated from empirical data. To obtain the concentrations themselves, a value must be assumed for V_0. We have

$$n_{a0} = (n_{a0} p_{b0})^{\frac{1}{2}} \exp [-\beta V_0], \tag{35}$$

$$p_{b0} = (n_{a0} p_{b0})^{\frac{1}{2}} \exp [\beta V_0]. \tag{36}$$

As we have mentioned previously, there is no way to determine V_0 from our experiments, although postulate II sets limits on its value.

Let us for simplicity assume that $S_{nb} v_n = S_{pa} v_p = S_t v$. Then, using (35) and (36) in (30) and (32), we have

$$C = 2 S_t v (n_{a0} p_{b0}/n_i^2)^{\frac{1}{2}} \cosh \beta V_0 . \tag{37}$$

This equation may be used to estimate the trapping cross-section S_t for an assumed V_0.

COMPARISON BETWEEN THEORY AND EXPERIMENT

In Fig. 14 we have plotted c.p. $-$ (c.p.)$_0$ $-$ $(V_B - V_{BO})$ versus 2 sinh $\beta(V_B - V_{BO})$ for p-type sample D, (see Table I). Each symbol represents experimental points for one cycle. Results are shown for five different cycles not all in the same run. These results are typical of all the data obtained for this sample. It is seen that these data can be fitted over most of the range by a straight line as drawn, giving a value of $H = 0.02$ e.v.

In Fig. 15 we have used the same experimental results to plot δV_L versus c.p. $-$ (c.p.)$_0$. Since the experimental values of δV_L cover a range of a factor of 20 or more, a logarithmic scale was used to plot

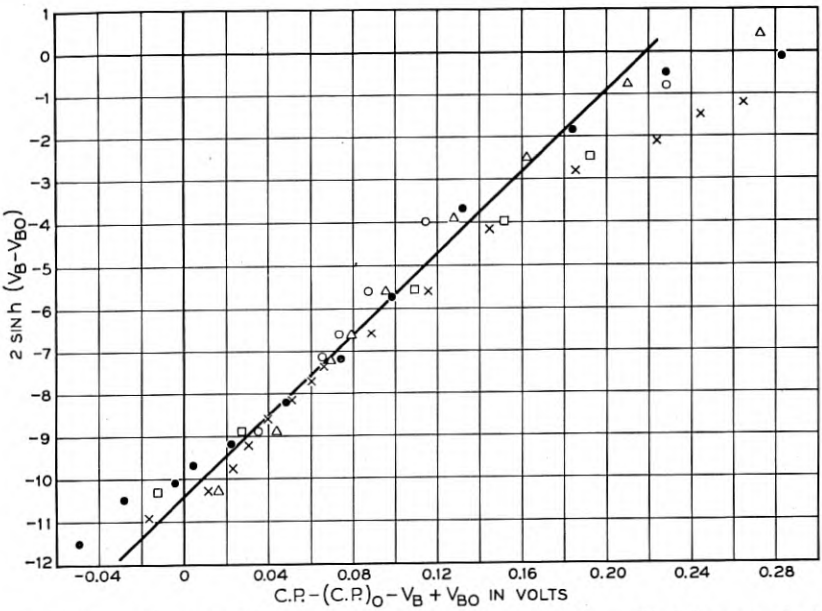

Fig. 14 — Plot of 2 sinh $\beta(V_B - V_{BO})$ versus c.p. $-(V_B - V_{BO})$ for sample D as suggested by theory.

δV_L. Both positive and negative branches of the curve are plotted on the same figure. The symbols used for the experimental points are consistent with Fig. 14. When c.p. $-$ $(\text{c.p.})_0$ is greater than zero, δV_L is negative. As c.p. increases it approaches a maximum negative value, this is the negative branch. When c.p. $-$ $(\text{c.p.})_0$ is less than zero, δV_L is positive and as c.p. decreases δV_L approaches a positive maximum value that is less in magnitude than the negative maximum. The solid curve represents the prediction of theory for $H = 0.02$ e.v. The agreement between theory and experiment is good. It should be emphasized that this fit is obtained with only one adjustable parameter.

The data for the other samples were analyzed in the same way. The results are shown by plotting δV_L versus c.p. $-$ $(\text{c.p.})_0$ as in Fig. 15. Fig. 16 is for n-type sample A and Fig. 17 for n-type sample B. The values obtained for H were 0.015 and 0.022 e.v. respectively. The fits obtained are about equally good in all cases with some deviation between theory and experiment near the extremes of contact potential. The values of H obtained are all of the same order as they should be if the surface trap structure is approximately the same from sample to sample. When $V_B - V_{BO}$ is large and positive and thus c.p., Eq. (14), is large, δV_L approaches the negative maximum

$$-\left[\left(\frac{1}{\beta}\ln\frac{n_1}{n}\right)-\delta V_i\right],$$

equation 21, and as $V_B - V_{BO}$ becomes small and negative, c.p. decreases and δV_L goes through zero and approaches the positive maximum

$$+\left[\left(\frac{1}{\beta}\ln\frac{p_1}{p}\right)+\delta V_i\right].$$

For p-type germanium n_1/n is greater than p_1/p so that the negative maximum for δV_L is greater than the positive maximum. Just the opposite is true for n-type germanium.

In this comparison of theory and experiment a tacit assumption has been made that the germanium surface is uniform in its properties. It might well be that this is not the case. The surface might be "patchy." To estimate what effect patches or non-uniformities might have the

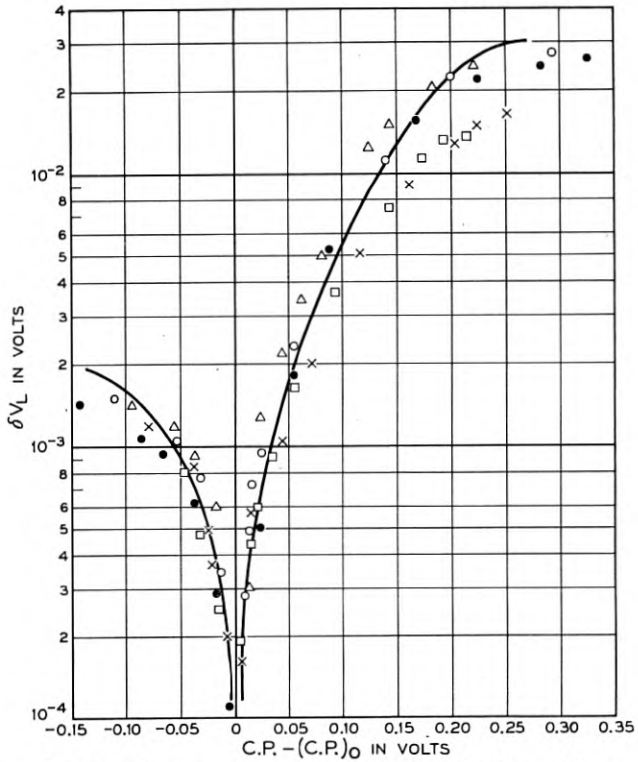

Fig. 15 — Change of contact potential with illumination, δV_L, versus c.p.; experiment and theory sample D, p-type.

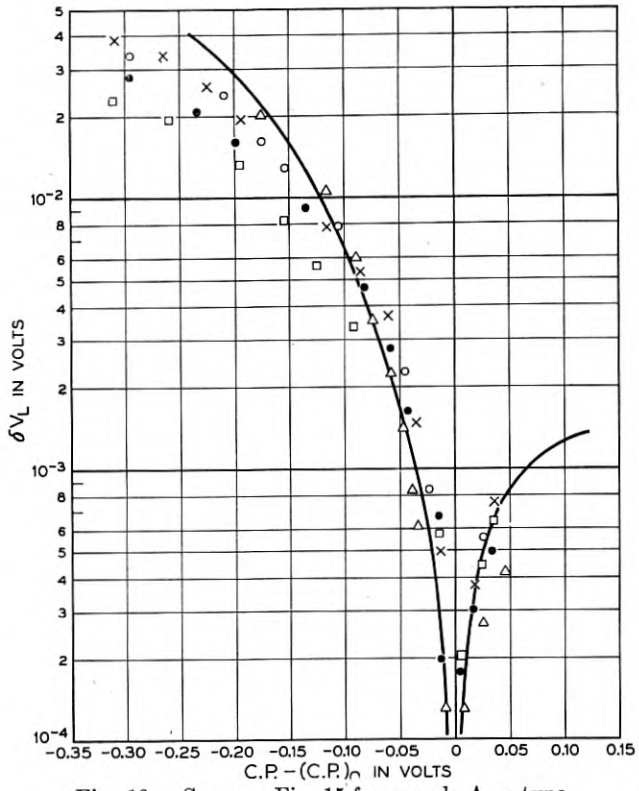

Fig. 16 — Same as Fig. 15 for sample A, n-type.

following was done. It was assumed that the surface was made up of two parts each having the theoretical dependence of δV_L on c.p. but with the contact potential where the light effect goes through zero differing by 0.05 volts. The values of δV_L at a given contact potential were averaged and plotted against c.p. The difference between this curve and the original one was almost entirely a simple shift in $(\text{c.p.})_0$. It is therefore unlikely that non-uniformities in c.p., over the surface, of the order of 0.1 volts or less would change the relation between δV_L and c.p. sufficiently to be detectable.

One can use equations (8), (10) and (21) to calculate values of $(V_B - V_0)$ for each experimental value of δV_L and then plot these values against the corresponding values of contact potential. These results are shown in Fig. 18 for n-type sample A and p-type sample D. $(V_B - V_0)$ changes with c.p. as one would expect and in an approxi-

mately linear fashion throughout most of the experimental range. The change in V_B is approximately one-fifth the change in contact potential. $(V_B - V_0)$ is positive for the p-type sample and negative for the n-type sample throughout most of the range. If the trap distribution on the surface were symmetrical about the intrinsic position of the Fermi level, then p_{b0} would be equal to n_{a0} and V_0 would be zero. In this case the space charge layers on n- and p-type germanium would be about equally developed. All the experimental evidence on germanium indicates that this is not the case but rather that $-V_{Bn}$ is much larger than V_{Bp}. This then means that V_0 is negative and that the trap distribution is unsymmetrical either in number or energy or both in such a way that $n_{a0} > p_{b0}$.

In Table II values are given for the differences $(V_B - V_{BO})$, etc.,

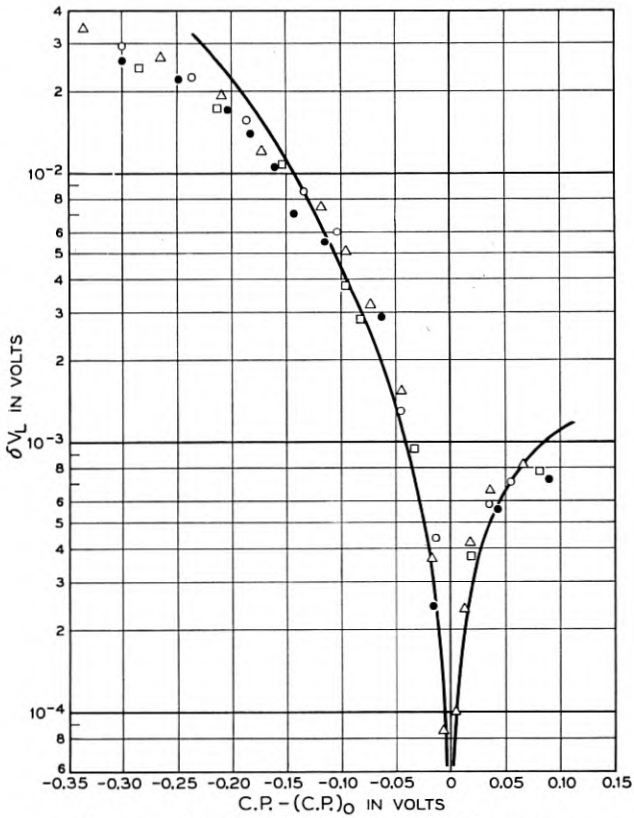

Fig. 17 — Same as Fig. 15 for sample B, n-type.

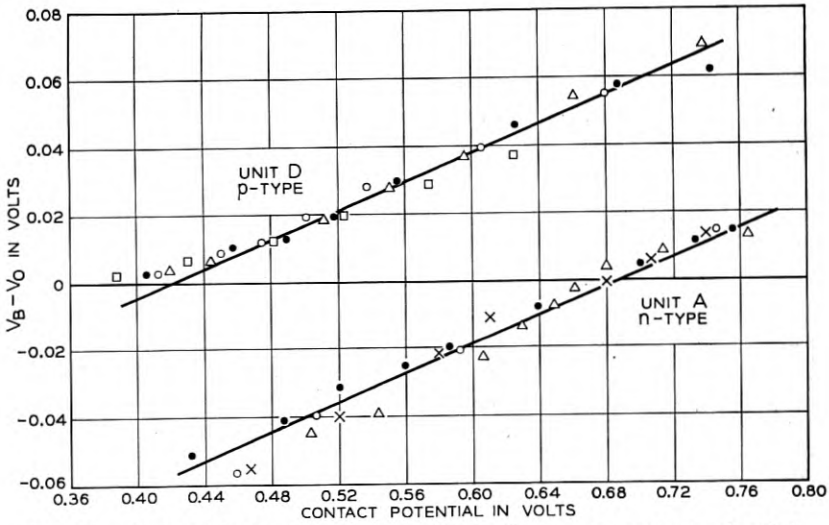

Fig. 18 — Potential across space charge layer, V_B, versus contact potential.

also δV's, for certain special cases for samples A and D as calculated from theory. From equations (8) and (10)

$$(V_{BO} - V_0) = \frac{1}{e}(E_i - E_F) = \frac{1}{2\beta}\ln(p/n). \tag{38}$$

It is a constant for any given sample as is δV_i (see equation 15). From equations (21) and (38) it follows that when $\delta V_B = 0$ then $(V_B - V_0) = 0$ for all samples. To say this another way, if the traps are symmetrically distributed about E_i and V_B is zero, then illuminating the germanium would produce no potential difference across the surface. One would not suspect offhand that $(V_B - V_0)$ for $\delta V_L = 0$ was also approximately independent of the material in the sample but it is. For the case where δp and the minority carrier are both small compared to the majority

TABLE II

	A (n-type)	D (p-type)
$(V_{BO} - V_0)$	−0.044	+0.064
δV_i	+0.0019	+0.0015
$(V_B - V_0)_{\delta V_B} = 0$	0	0
$(V_B - V_{BO})_{\delta V_B} = 0$	+0.044	−0.064
$(V_B - V_0)_{\delta V_L} = 0$	+0.010	+0.010
$(V_B - V_{BO})_{\delta V_L} = 0$	+0.054	−0.054
δV_L , $(V_B - V_{BO}) = 0$	+0.024	−0.026

carrier one can easily show that when $\delta V_L = 0$,

$$(V_B - V_0) = (1/2\beta) \ln b.$$

This follows from equations (15), (21), (22) and (38). The rest of the results hardly need comment.

It is interesting to compare differences in contact potential, etc., for units A and D. In equation (2) the Fermi energy has been included in the constant. This equation predicts that if the trap distributions on both surfaces are the same, then when the contact potentials of both surfaces are equal the values of V_D should also be equal and the difference between the V_B's should be equal and opposite to the differences between the Fermi energies in electron volts. This equation can be written

$$\text{c.p.} = (V_B - V_0) - (V_{BO} - V_0) + V_D + V_0 + \text{const}. \quad (40)$$

In comparing units we shall use Δ's to denote differences and these differences are always taken A–D.

For the case where Δc.p. $= 0$, $\Delta(V_B - V_O)$ can be read from Fig. 18 and $\Delta(V_{BO} - V_O)$ from Table II. We have then

$$\Delta(V_D + V_0 + \text{const}) = -0.05\,.$$

This indicates that our simple picture is not quite right. If ΔV_D were zero for this case it should follow from equation (12) that both $\Delta 2H \sinh \beta(V_B - V_{BO})$ and the difference in the const in this equation should be zero. It turns out, however, that $\Delta 2H \sinh \beta(V_B - V_{BO})$ is not zero but $+0.07$. Equation (12) can be substituted into equation (40) giving

$$\text{c.p.} = (V_B - V_0) - (V_{BO} - V_0) + 2H \sinh \beta(V_B - V_{BO}) + V_0 + \text{const}_2\,. \quad (41)$$

Where the constant now includes the constant part of V_D and is labeled const_2 to distinguish it from the constant in equation (40). We have then for Δc.p. $= 0$

$$\Delta(V_0 + \text{const}_2) = -0.12.$$

This difference $\Delta(V_0 + \text{const}_2)$ can be calculated three other ways, using the experiment results and theoretical values where necessary. These ways are

1) at the same time in the cycle

$$\Delta(V_0 + \text{const}_2) = -0.10$$

2) when $\delta V_L = 0$

$$\Delta(V_0 + \text{const}_2) = -0.13$$

3) when $\delta V_B = 0$

$$\Delta(V_0 + \text{const}_2) = -0.12\,.$$

All four results are consistent within the probable experimental accuracy and give an average result of $\Delta(V_0 + \text{const}_2) = -0.12$ instead of zero as one might expect from the simple picture. This indicates that the trap distributions are different on the two surfaces. The constant part of V_D is proportional to $-(N_a - N_b)$ and from equation (10) one would expect V_0 to decrease as N_a increases with respect to N_b, thus ΔV_0 and Δ const$_2$ should be of the same sign and additive so that Δ const$_2$ is less than 0.12 volts; assuming that ℓ_D/K_D is the order of 2×10^{-7} cm, this indicates that $\Delta(N_a - N_b)$ is the order of 3×10^{11} per cm^2 which is small compared with the probable trap density, N_a, N_d, of the order of 10^{14} per cm^2 as we shall see in the next paragraph.

Assuming that ℓ_D/K_D in equation (13a) is the order of 2×10^{-7} cm one can calculate $(n_{a0}p_{b0})^{\frac{1}{2}}$ obtaining 4.1×10^{10} and 5.5×10^{10} for samples A and D respectively. Using these values one can solve for S_t in equation (37) and obtain for the case of $V_0 = 0$ the value 5.0×10^{-17} cm^2 for the average capture cross section of the surface traps. The values of S_t, n_{a0} and p_{b0} depend on what one takes for V_0. The relations are

$$S_t = \frac{5 \times 10^{-17}}{\cosh \beta V_0}\,,$$

$$n_{a0} = 5 \times 10^{10} \exp\left[-\beta V_0\right],$$

$$p_{b0} = 5 \times 10^{10} \exp\left[\beta V_0\right].$$

This dependence is shown in the graph in Fig. 19. As already mentioned there are reasons for thinking that V_o is less than zero. If one takes -0.06 volts as a reasonable value then one gets $S_t = 10^{-17}$ cm^2, $n_{a0} = 5 \times 10^{11}$/cm^2 and $p_{a0} = 5 \times 10^9$/cm^2 respectively. One can push these calculations still further to estimate N_a, N_b, E_a and E_b. One knows that E_a and $-E_b$ must be greater than $1/kT$. Also N_a and N_b should be less than the number of germanium atoms per cm^2 of surface which is 1.4×10^{15}/cm^2. Values of N_a and N_b of the order of 1×10^{14}/cm^2 for the number of traps per cm^2 with energies E_a and $-E_b$ of the order of 0.2

e.v. measured from the midband energy are reasonable and not inconsistent with the original assumptions.

As mentioned in the experimental section, some data on $(\Delta c.p.)_L$ for samples C and E have been obtained. While no complete analysis of these results has been made, one can see that they are of the right order of magnitude. As the specific resistance of the sample decreases, the body life time τ decreases. This empirical result is to be expected.[7] Consequently δp for the same light intensity decreases and one would expect $(\Delta c.p.)_L$ to decrease as it does. For sample E of course one could not neglect the charge in the space layer so that the theory would be more involved.

The comparison between the contact potentials of samples A, C, D and E shown in Fig. 6 can be understood in part at least. Consider first the over-all result that at the same time in the cycle the c.p. for a sand-blasted surface is less than for an etched surface, i.e., the work function for the sand-blasted surface is greater. It is known that the surface recombination increases enormously when the surface is sand-blasted. This means that either the surface trap density has increased or that the distribution has changed or both in such a way as to increase surface

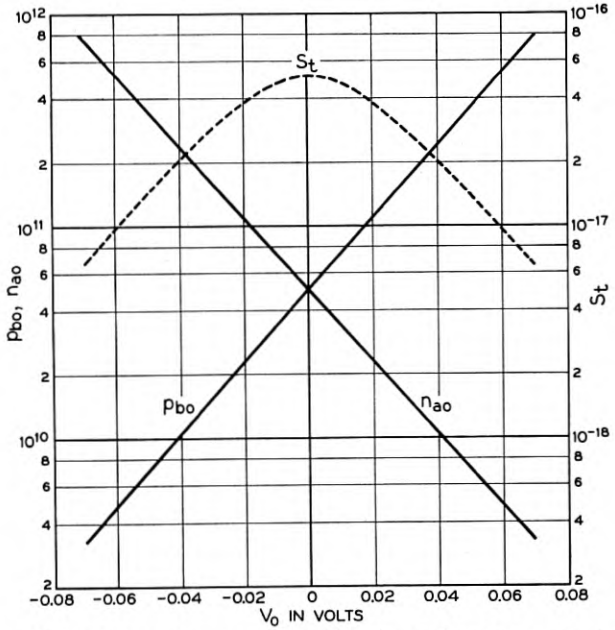

Fig. 19 — Dependence of carrier densities in surface traps, n_{a0} and P_{b0}, for $V_B = 0$, and cross section of trapping S_t on V_0.

recombination. The results of Fig. 6 indicate that this change in trap distribution also has been in a direction to increase $(-V_B)$. This is what one would expect if N_b has increased with respect to N_a, i.e., sand-blasting increases the effective number of acceptor traps. About all that can be said about the rest of the results in Fig. 6 is that if one knew the surface trap structure in number and energy distribution for the etched surface then one could deduce from the results in Fig. 6 the distribution of traps for the sand-blasted surface over part of the energy range at least.

The theory developed here for the germanium surface may not apply at all for a silicon surface. In a previous discussion of the silicon surface[1] it was assumed that the resistance to surface trapping occurred mainly in the flow across the space charge layer rather than from the surface to the traps. This may well be the case in silicon. An investigation of the silicon surface using these same techniques should clarify this point.

It should be emphasized that the ambient used for this study and the variations in it are special in that they do not necessarily correspond to the atmosphere of ordinary room air. It is very probable that constituents in room air other than oxygen ions and water vapor are important, such as salt ions for instance. It is quite probable that v_s for such a surface exposed to room air would increase considerably with time instead of staying relatively constant as it does under the bell jar.

CONCLUSIONS

A method has been developed for studying the surface properties of Ge in a gaseous ambient at atmospheric pressure. It has been found that the Ge surface interacts with this ambient. Two atmospheric constituents that are important in this interaction, oxygen and vapors with OH radicals, have been isolated. With the controlled use of these, the surface dipole of Ge can be cycled between two extremes. Thus the dependence of other properties of the Ge surface on surface dipole, such as change of contact potential with illumination and surface recombination, can be determined.

It is evident from the results that the method is very powerful. In order to complete the present study, it was necessary to stop trying new experiments since almost all directions of variation open up new and interesting phenomena. These results on Ge are really just a beginning, and the preliminary data on silicon indicate that the same method would also be fruitful on other semiconductors. The technique is of course not limited to atmospheric pressure.

A tentative theory of the Ge surface has been developed that is sufficient to explain the experimental results on a semi-quantitative basis. Theory and the experiment together predict approximately the

number, type and distribution in energy of the surface traps. One has therefore a tentative model of the Ge surface that should be very useful in any further investigation of its properties.

ACKNOWLEDGMENTS

We wish to acknowledge the help and assistance of all our colleagues who have contributed in many ways to make this investigation a success. We wish to mention in particular E. G. Dreher and R. E. Enz who took most of the experimental data, H. R. Moore who designed and made the electronic equipment used in making the measurements, and Conyers Herring for suggestions regarding the theory of large amplitude signals.

APPENDIX

We have assumed (Postulate IV) that traps of type a are in equilibrium with electrons in the conduction band and that traps of type b are in equilibrium with holes in the valence band. We wish to show that this is not really a separate assumption but follows as a consequence of Postulate II if the density of traps is not too high. For simplicity, we shall restrict the discussion in the appendix to the limiting case of small departures from equilibrium so that the equations are linear. The problem may then be discussed most conveniently[7] by means of quasi-Fermi levels* or imrefs, ϕ_n and ϕ_p, for the conduction electrons and holes, respectively.

Departures $\delta\phi_n$ and $\delta\phi_p$, of the imrefs from the Fermi level are a measure of the departures of the concentrations, δn and δp, from their equilibrium values:

$$n_1 = n + \delta n = n \exp(-\beta\delta\phi_n), \tag{A.1}$$

$$p_1 = p + \delta p = p \exp(\beta\delta\phi_p). \tag{A.2}$$

The imrefs in the interior are then

$$\delta\phi_n = \frac{-1}{\beta} \ln\left(1 + \frac{\delta n}{n}\right), \tag{A.3}$$

$$\delta\phi_p = \frac{1}{\beta} \ln\left(1 + \frac{\delta p}{p}\right). \tag{A.4}$$

Correspondingly, the imrefs at the surface are defined by:

$$n_{s1} = n_s + \delta n_s = n \exp[\beta(V_B + \delta V_B - \delta\phi_{ns})], \tag{A.5}$$

$$p_{s1} = p_s + \delta p_s = p \exp[-\beta(V_B + \delta V_B - \delta\phi_{ps})]. \tag{A.6}$$

* Reference 2, pages 302–308.

Changes in imrefs of the traps are defined by:

$$n_{t1} = n_t + \delta n_t = \frac{N_t}{1 + \exp\left[(E_t - E_F - e(V_B + \delta V_B) + e\delta\phi_t)/kT\right]}, \quad (A.7)$$

$$p_{t1} = p_t + \delta p_t = \frac{N_t}{1 + \exp\left[(E_F - E_t + e(V_B + \delta V_B) - e\delta\phi_t)/kT\right]}. \quad (A.8)$$

In these last two equations, t may refer to either type of trap (a or b) and $\delta p_t = -\delta n_t$.

For small signals, the recombination current via a given set of surface traps may be considered as a flow produced by differences in the imrefs, $\delta\phi_n$ and $\delta\phi_p$, through four effective resistances in series, as shown in Fig. 20. Here R_{nB} is an effective resistance for flow of electrons across the barrier layer, R_{nt} for flow of electrons from the conduction band at the surface to the traps, R_{pt} for flow of holes from the filled band to the traps, and R_{pB} for flow of holes across the barrier layer. Under steady state conditions, the net flow of conduction electrons to the surface is balanced by an equal flow of holes. The recombination current may be thought of as a flow of electrons from the conduction band via the traps to the valence band.

The recombination current per unit area is

$$I = -eU = (\delta\phi_n - \delta\phi_p)/R_t, \quad (A.9)$$

where R_t is the sum of the four resistances in series. Here U is the particle current and I is the corresponding electric current. If $\delta\phi_n$ and $\delta\phi_p$ are expressed in volts, R_t is in ohms/cm^2. We may define a recombination constant C_t by an equation corresponding to (34):

$$U = C_t(p_1 n_1 - n_i^2). \quad (A.10)$$

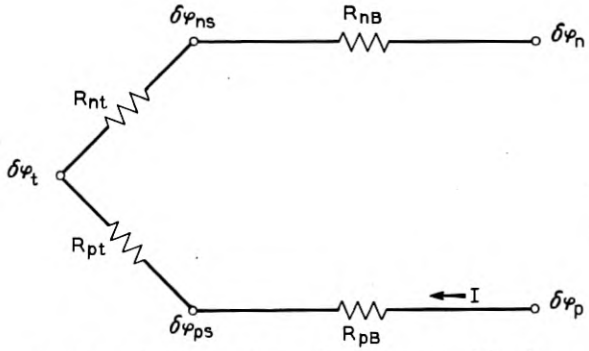

Fig. 20 — Circuit analogy of surface recombination.

The relation between C_t and R_t is obtained by use of (A.1) and (A.2) in (A.10). Since (A.9) is valid only to terms of the first order in $\delta\phi_n$ and $\delta\phi_p$, we make the corresponding approximation in (A.10) and find

$$1/R_t = e\beta n_i^2 C_t . \tag{A.11}$$

If there is more than one type of trap, the net recombination resistance, R, is that of the various traps in parallel. Values obtained for R for specimens A and D from the empirical values of C are about 500 ohms.

We shall show that for a-traps, R_{pa} is much larger than the other resistances in series, and that for b-traps, R_{nb} is the dominant resistance. This implies that $\delta\phi_a = \delta\phi_n$ and $\delta\phi_b = \delta\phi_p$, or in other words that a-traps are in equilibrium with the conduction band and b-traps with the valence band.

First consider the flow across the space-charge layer. The resistances R_{nt} and R_{bt} depend on the sign of V_B. When V_B is negative, the net electron current across the space-charge layer is:

$$\begin{aligned} I &= ev_n(n_{s1} - n_1) \exp [\beta(V_B + \delta V_B)] \\ &\cong e\beta v_n n_s(\delta\phi_n - \delta\phi_{ns}), \end{aligned} \tag{A.12}$$

where v_n is defined by an equation similar to (24). The second form is the linear approximation. Thus we have

$$1/R_{nB} = e\beta v_n n_s . \tag{A.13}$$

If V_B is positive, n_s is replaced by n.

Correspondingly, for V_B negative, the hole current is:

$$\begin{aligned} I &= ev_p(p_1 - p_{s1}) \exp [\beta(V_B + \delta V_B)] \\ &\cong e\beta v_p p(\delta\phi_{ps} - \delta\phi_p), \end{aligned}$$

and

$$1/R_{pB} = e\beta v_p p. \tag{A.14}$$

For V_B positive, p is replaced by p_s.

The maximum value of R_{nB} is obtained for the ambient which makes V_B most negative and R_{pB} is a maximum when V_B is most positive.

The current from the conduction band to the traps is calculated as in (23):

$$I = e(v_n S_{nt} n_{s1} p_{t1} - g_t n_{t1}) \tag{A.15}$$

$$\cong e\beta v_n S_{nt} n_s p_t(\delta\phi_{ns} - \delta\phi_t). \tag{A.16}$$

Here g_t is the rate of thermal generation of electrons from occupied traps to the conduction band. Since, for equilibrium conditions, $I = 0$, we have:

$$g_t = v_n S_{nt} n_s p_t / n_t . \tag{A.17}$$

It may be verified easily that g_t is independent of E_F and V_B. The ratio,

$$n_{ct} = n_s p_t / n_t = N_c \exp [(E_t - E_c)/kT], \tag{A.18}$$

is the equilibrium concentration of electrons in the conduction band when the Fermi level is at the level of the traps. The expression (A.16) is again the linear approximation. Thus

$$1/R_{nt} = e\beta v_n S_{nt} n_s p_t . \tag{A.19}$$

Similarly, we have

$$1/R_{pt} = e\beta v_p S_{pt} p_s n_t . \tag{A.20}$$

We shall show that the ratio

$$R_{nt}/R_{pt} = (v_n S_{pt}/v_p S_{nt})(p_s n_t/n_s p_t), \tag{A.21}$$

may be expected to be small compared with unity for a-traps and large compared with unity for b-traps. First consider a-traps. If anything, $S_{pa} < S_{na}$, because holes must give up a larger energy than conduction electrons in going to a-traps. The second factor is small compared with unity if

$$p_s \ll n_{ct} , \tag{A.22}$$

with n_{ct} defined by (A.18). This will be the case if the a-traps are closer to the conduction band than the top of the valence band at the surface is to the Fermi level. According to Postulate II, this should always be the case.

Similarly, for b-traps

$$R_{pt}/R_{nt} \ll 1 \tag{A.23}$$

if

$$n_s \ll p_{vt} , \tag{A.24}$$

where

$$p_{vt} = p_s n_t / p_t = N_v \exp [-(E_t - E_v)/kT]. \tag{A.25}$$

is the concentration of holes in the valence band when the Fermi level is at the level of the traps.

The values of R_{nB} and R_{pB} relative to R_{nt} and R_{pt} are:

$$R_{nB}/R_{nt} = S_{nt}p_t \tag{A.26}$$

$$R_{pB}/R_{pt} = S_{pt}n_t. \tag{A.27}$$

In our experiments these are always small compared with unity, since the trapping cross-sections are of the order of 10^{-17} cm^2 and p_t and n_t are always less than N_t, which is of the order of 10^{14}/cm^2.

Barrier resistances may be important for surfaces with a large number of surface traps. Analysis of earlier data on the change of the contact potential of silicon with light[1] was made on the assumption that the probability that an electron or hole reaching the surface be trapped is relatively large, and that consequently the barrier resistances are large compared with the trapping resistances. The present experiments on germanium throw doubt on this interpretation, but further experiments are required to clarify the situation.

REFERENCES

1. Bardeen, J., Phys. Rev., **71**, p. 717, 1947. Brattain, W. H., Phys. Rev., **72**, p. 345, 1947 and Semi-Conducting Materials Butterworths Scientific Publications Ltd., pp. 37–46, 1951.
2. Shockley, W., *Electrons and Holes in Semiconductors*. D. Van Nostrand, pp. 318–325, 1950.
3. Pearson, G. L., unpublished data. J. R. Haynes and W. Shockley, Phys. Rev. **81**, p. 835, 1951.
4. The CP-4 etch is due to R. D. Heidenreich. The formula is given in Phys. Rev., **81**, p. 838, 1951. This method of treating a Ge surface is due to C. S. Fuller.
5. C. S. Fuller suggested the use of these fumes.
6. Goucher, F. S., G. L. Pearson, M. Sparks, G. K. Teal and W. Shockley, Phys. Rev. **81**, p. 637, 1951.
7. Shockley, W. and W. T. Read, Phys. Rev., **87**, p. 835, 1952.
8. van Roosbroeck, W., Bell Sys. Tech. Jl., **29**, p. 560, 1950.

Frequency Economy in Mobile Radio Bands

By KENNETH BULLINGTON

(Manuscript received August 20, 1952)

The various factors affecting the usability of mobile radio channels are discussed, and estimates are obtained for the number of usable channels per megacycle for several present and proposed methods of operation. The lack of radio-frequency selectivity is the principal barrier to maximum frequency economy, but this difficulty can be avoided by sufficient geographical and operational coordination.

The increasing demand for all types of radio services emphasizes the need for efficient use of the radio frequency spectrum. In mobile radio operation the number of usable channels that can be obtained in the VHF and UHF mobile bands depends not only on the width of the individual channels, but also on how and where each channel is to be used. Activity on the same frequency at neighboring locations, and on neighboring frequencies at the same location both affect the usefulness of a channel. Halving the channel spacing doubles the number of potential assignments, but it does not double, and in some cases it does not appreciably increase the number of usable channels.

The usefulness of a single isolated channel is determined by the intensity of its signal above the noise level. Because of the very wide variation in received signal strength caused by distance, terrain, building shadowing, etc., the coverage area of a channel can be discussed only in statistical terms. There are likely to be islands of poor signal-to-noise ratio even close to the transmitter, and the coverage gradually fades out into more spotty conditions at greater distance.

If the same frequency is used at a neighboring location, the familiar problem of co-channel interference arises. There will now be locations where the desired signal is above noise, but the undesired signal is still stronger. Thus, the coverage area of a channel is reduced by the existence of the co-channel transmitter; again, it is possible to discuss this reduction only in statistical terms.

When two channels are being operated on different frequencies in the same general area, the coverage area of each is limited by signal-to-

noise considerations. In addition, each channel may affect the other because of spurious radiation from transmitters, insufficient receiver selectivity, receiver oscillator radiation, etc. The recent trend toward receivers with greatly improved IF selectivity is worthwhile, but even infinite IF selectivity cannot solve many of the present interference problems.

When three or more channels are operating in the same general area, another type of interference occurs because of intermodulation in transmitters or receivers. If it were technically feasible to build into the equipment sufficient radio frequency selectivity to separate the working channels, this interference could be removed. In fact, this is not feasible, and it is necessary to consider possible modulation products from channels falling within a frequency band several percent wide. The number of possible interference conditions that result from intermodulation (third order) rises from 9 for 3 working channels to 50 for 5 channels, to 450 for 10 channels, and to 495,000 for 100 working channels. Some of these interference combinations overlap and fall on the same channel; but even considering all possible duplication, intermodulation interference rapidly becomes controlling as the number of closely spaced channels working in the same area is increased.

It is not technically feasible to achieve enough radio frequency selectivity to permit unrestricted and uncoordinated use of many channels in a given area, unless the channels are, on the average, separated by about 1 per cent of the operating frequency. For any kind of efficiency of frequency utilization, it is necessary to have some coordination in the location of fixed transmitters and in the use of channels. The maximum efficiency of utilization requires the maximum coordination.

The technical factors that determine channel width, channel spacing, and the number of usable channels are described and tabulated below. The first section discusses the principal factors that affect the usefulness of channels equipped with transmitters and receivers with perfect filtering. This is followed by a consideration of the limitations imposed by insufficient total filtering and by insufficient radio frequency filtering. The next section shows the reduced requirements that are possible by coordination between systems. Finally, the quantitative data are used to illustrate the capabilities and efficiencies of various present and proposed methods of mobile system operation.

CHANNELS WITH PERFECT FILTERING

It has been found by experiment that the radio path loss between antennas in a mobile radio system can be ascribed to three principal fac-

tors: distance, shadowing and standing wave patterns. The variation with distance from the base station follows the theoretical free space loss up to 500 feet or more, as long as the points are within line of sight. Typical values of the free-space loss are shown in Table I. Beyond about one-half mile the median path loss over plane earth increases about 12 db each time the distance is doubled out to distances of 20–30 miles.[1, 2]

In addition to the increase in path loss with distance, which is accounted for reasonably well by the theory of radio propagation over plane earth, bold features of geography such as mountains and large buildings cause shadow losses that result in irregular coverage patterns. For example the median loss at street level for random locations in New York City is about 25 db greater than the plane earth values computed for the distance and antenna heights involved; the corresponding 10 per cent and 90 per cent losses are about 15 and 35 db respectively.[3]

Superimposed on the above effects which vary relatively slowly with location are standing wave patterns whose effect on path loss can change substantially within a foot or so. The standing waves are the result of random additions of multiple reflections from nearby buildings or terrain, and the variation in path loss follows the Rayleigh distribution for small changes in distance in urban areas. In other words, there is no theoretical limit on the deviation from the median but in 1 per cent of the possible locations the signal is likely to be more than 8 db above the median value and in 99 per cent of the possible locations the signal level is not expected to be more than 18 db below the median value.

The motion of the mobile unit through the standing wave patterns causes signal fluctuations or flutter in the received signal. The flutter

TABLE I — FREE SPACE LOSS BETWEEN DIPOLES

Separation Between Transmitting and Receiving Antennas	Free-Space Loss	
	150 mc	450 mc
5 ft.	16 db	26 db
50 ft.	36	46
500 ft.	56	66
½ mile	70	80

[1] Young, W. R., Jr., Comparison of Mobile Radio Transmission at 150, 450, 900 and 3700 Mc. Bell Sys. Tech. Jl., **30**, pp. 1068–1085, Nov., 1952.

[2] Aikens, A. J., and L. Y. Lacy, A Test of 450-Megacycle. Urban Transmission to a Mobile Receiver. I.R.E., Proc., pp. 1317–1319, Nov., 1950.

[3] Bullington, K., Radio Propagation Variations at VHF and UHF. I.R.E., Proc., pp. 27–32, Jan., 1950.

rate at 150 mc may be as much as 15 cycles per second for a speed of 30 mph and increases as either the radio frequency or the speed of the mobile unit is increased. The fast acting gain control needed to minimize the flutter effects is obtained automatically with frequency modulation but is more difficult to obtain with amplitude modulation. This factor is one of the principal advantages of the use of FM instead of AM for mobile radio systems.

The co-channel interference to be expected between stations having equal transmitter powers depends on the path loss statistics for both the desired and undesired signals. At the edge of the desired coverage area there must be a high probability that the desired signal will be strong enough to be useful and only a small probability that the undesired signal will be strong enough to be troublesome. The geographical separation needed between co-channel stations varies from about four to six times the desired coverage radius when FM is used and from six to eight times when AM is used.[4] If the needs for mobile channels were uniformly distributed geographically only a small part of the potential channel assignments would ever be used in a given area. However, the needs for mobile channels are usually concentrated in areas of high population density so that a large percentage of the channel assignments may be needed in the same area.

The above estimates on co-channel spacing depend somewhat on the antenna heights and the type of terrain, and assume that the same frequency is used in both directions of transmission. When the two-frequency method is used with adequate separation between the transmitting and receiving frequencies, the co-channel spacings can be reduced to about three to five times the coverage radius for FM and to about four to six times for AM. This reduction of approximately 30 per cent is possible because the most troublesome interfering path in the single frequency method (from base transmitter to base receiver) can be eliminated in the two-frequency method by sufficient selectivity.

The principal reason for using the single frequency method is to provide communication between two mobile units when they are relatively near each other but are beyond the range of the base station. When transmission of all messages through the base station is desirable, or at least not objectionable, the two-frequency method is preferable. It is shown in a later section that close geographical and operational coordination is needed to achieve maximum efficiency in the use of frequency space and that this coordination can be obtained only with the two-frequency method.

[4] See reference in Footnote 3.

The bandwidth needed to pass the desired signal depends on the frequency stability that can be maintained as well as on the type of modulation. The allowance for frequency drift includes the variations in both transmitters and receivers. The importance of these figures is indicated in Table II which shows the tolerances needed for frequency instability. For example, with an overall frequency stability of ±0.002 per cent the channel width at 450 mc needs to be 18 kc wider than the minimum bandwidth required to pass the modulated signals.

The use of frequency modulation has several important advantages that cannot be readily obtained with AM. The instantaneous gain control and the closer co-channel spacings have already been mentioned. In addition, for the same radiated power, FM with a frequency swing greater than about ±3 kc has the well known advantage of providing a higher output signal-to-noise ratio throughout most of the coverage area than is possible with double sideband amplitude modulation; this FM advantage is substantially reduced when the IF bandwidth is large compared with the bandwidth required to pass the desired sidebands.

The bandwidth required for frequency modulation of a 3 kc voice band must be at least ±3 kc. For reasonable FM signal-to-noise advantage, particularly in the presence of impulse noise, the frequency swing should be at least ±5 kc which requires a bandwidth of ±8 kc for good quality. The corresponding bandwidth for amplitude modulation is ±3 kc; the use of single sideband AM transmission does not seem feasible, at least not for single channel operation.

LIMITATIONS IMPOSED BY INSUFFICIENT (TOTAL) FILTERING

The frequency separation between carrier frequencies must be greater than the bandwidth required to pass the desired signal because additional frequency space or guard bands are needed to build up receiver selectivity against undesired signals and to avoid the extra band radiation from transmitters. The power of a 100 watt transmitter is about

TABLE II — TOLERANCE NEEDED FOR OVERALL FREQUENCY DRIFT

Frequency Stability	Allowance for Frequency Drift	
	150 mc Band	450 mc Band
±0.001%	±1.5 kc	± 4.5 kc
±0.002	±3	± 9
±0.005	±7.5	±22.5

TABLE III — REQUIRED SUPPRESSION VERSUS DISTANCE BETWEEN ANTENNAS

Distance Between Transmitting and Receiving Antennas	Total Selectivity or Filtering Required	
	150 mc	450 mc
0 ft.	160 db	160 db
50 ft.	124	114
500 ft.	104	94
½ mile	90	80

160 db greater than the minimum signal that is useful in the receiver (140 db below one watt), so ideally no appreciable interference would result if the overall selectivity of the receiver and the suppression of extra band radiation in the transmitter could be in excess of 160 db. This amount of isolation is difficult to obtain by filtering. The interaction between transmitter and receiver of the same system is frequently avoided by the use of "push-to-talk" operation, but the potential interference between different systems requires the full 160 db (based on 100 watt transmitters). Fortunately, a substantial part of the desired isolation can be obtained by modest geographical separation. The net requirements for either receiver selectivity or transmitter filtering are less than 160 db by the losses shown in Table I and are summarized in Table III.

Receiver selectivities of 90–100 db or more are feasible except on nearby channels and possibly on certain image channels. Typical values of the guard bands that are required between the edge of the desired pass band and the frequency at which the desired attenuation to interfering signals can be obtained are estimated in Table IV.

Even if the guard band, shown in Table IV, required to provide adequate selectivity in the receiver could be reduced to zero by providing infinitely steep sides on the IF selectivity curve, there would still remain the guard band needed to avoid the extra band radiation from the transmitter. The amount of suppression of extra band radiation needed

TABLE IV — GUARD BAND VERSUS IF SELECTIVITY

Desired IF Selectivity	Required Guard Band
40 db	12 kc
60	15
80	20
100	25
120	30

for unrestricted operation is equal to the required receiver selectivity given in Table III and can be translated into frequency space in the following manner.

Both AM and FM transmitters radiate some noise and distortion products outside of the ideal modulation bandwidth. In addition, some of the sideband energy in FM falls outside the desired modulation bandwidth. The magnitude of the undesired FM sideband radiation is higher than the noise immediately outside of the desired band, but it decreases more rapidly with the result that the noise is usually controlling in the region where the extra band radiation is more than 60 to 70 db down.

The guard bands that are required between the edge of the desired transmitted band and the frequency at which the necessary suppression of extra band radiation can be obtained are estimated in Table V. These values depend on the width of the voice band and are relatively independent of the radio frequency since r.f. selectivity is not possible.

Measurements on present day transmitters correspond to the above estimates for values of suppression less than 80 db, but a frequency separation of nearly one megacycle or more is needed for suppressions of 100 and 120 db. This limitation is not expected to be inherent so more optimistic estimates are indicated in Table V. If the present characteristics cannot be improved, that is, if suppressions greater than about 80 db cannot be obtained, Table III indicates that some interference may be expected within about one-half mile of an unwanted transmitter.

A comparison of the information given in Tables III, IV and V indicates that the guard bands required for unrestricted operation are approximately 100, 50 and 25 kc for minimum separations between transmitter and receiver of 50 feet, 500 feet and one-half mile, respectively. These values together with the bandwidth needed for modulation and for frequency instability determine the frequency separation required between channels operating in the same area and are summarized in Table VI.

TABLE V — GUARD BANDS REQUIRED TO AVOID EXTRA BAND RADIATION

Suppression of Extra Band Radiation	Guard Bands Required	
	AM	FM
40 db	3 kc	9 kc
60	10	15
80	25	25
100	50	50
120	100	100

TABLE VI — CHANNEL SPACING REQUIRED FOR UNRESTRICTED OPERATION OF TWO FM CHANNELS IN SAME AREA VERSUS ANTENNA SEPARATION

Minimum Separation Between Transmitting and Receiving Antenna	Channel Spacing, Neglecting Intermodulation			
	150 mc		450 mc	
	±0.002%	±0.005%	±0.002%	±0.005%
50 ft.	112–122 kc	121–131 kc	124–134 kc	151–161 kc
500 ft.	62– 72	71– 81	74– 84	101–111
½ mile	37– 47	46– 56	49– 59	76– 86

The above table shows that if interference of the types so far considered is to be kept below the minimum usable signal at all distances greater than about 500 feet from undesired transmitters, the channel spacing needs to be at least 62 to 75 kc in the 150 mc band and 74 to 105 kc in the 450 mc band. The channel spacings for AM are equal to the minimum shown above, while the higher figure is for FM with ±5 kc swing (a modulation bandwidth of ±8 kc).

Since the above channel spacings are considerably greater than the necessary IF bandwidth, it should be possible to use intermediate channels in adjacent non-overlapping areas. This geographical limitation does not appreciably decrease the overall efficiency in the use of frequency space as long as the needs for mobile channels are more or less uniformly distributed within a large region, but it becomes important where a large percentage of the available channels are needed in the same metropolitan area.

In a later section it is shown that channel spacings less than the values given in Table VI are feasible in the same area providing sufficient coordination is achieved in both geographical spacings and operating methods.

The estimated channel spacings shown in Table VI do not take into account the effect of intermodulation interference which is discussed in the following section. Intermodulation interference may limit the number of usable one-way channels to only 1 or 2 per megacycle instead of the above 6 to 20 per megacycle, unless further restrictions are placed on the selection of frequencies and on the method of operation.

LIMITATIONS IMPOSED BY INSUFFICIENT RF FILTERING

When a strong unwanted signal on a frequency within the RF bandwidth is present at the input to a receiver, overloading occurs and the receiver

TABLE VII — REQUIRED RF RECEIVER SELECTIVITY VERSUS ANTENNA SEPARATION

Minimum Separation Between Transmitter and Receiver	RF Selectivity	
	150 mc	450 mc
0 ft. (common antenna)	95 db	95 db
50 ft.	59	49
500 ft.	39	29
½ mile	25	15

is said to be desensitized. When two or more strong unwanted signals are present desensitization also occurs, but in addition, extraneous frequencies are generated by intermodulation in the receiver itself. As the levels of the unwanted signals become greater than about 75 db below one watt (1 or 2 millivolts across a typical receiver) the intensity of the modulation products rises rapidly above the set noise. The resulting interference can be 60 db or more above set noise and the number of the modulation products increases by at least the cube power of the number of operating channels.

Ideally, the intermodulation interference in the receiver caused by 100-watt transmitters (20 db above one watt) can be eliminated by $20 + 75 = 95$ db RF selectivity even when the receiver and the unwanted transmitters are connected to the same antenna. In practice, the effect of geographical separation assuming the free space loss given in Table I reduces the RF selectivity requirement to the values given in Table VII.

The RF selectivity requirements given in Table VII cannot be obtained on nearby channels. The approximate RF bandwidths associated with various amounts of RF selectivity in mobile receivers is shown in Table VIII. For example, in mobile receivers it seems feasible to provide 40 db of RF selectivity at frequencies removed from the desired channel by about 3 mc in the 150-mc band and by about 10 mc in the 450-mc band. At fixed stations the RF bandwidth required for a given selec-

TABLE VIII — FREQUENCY SPACING FROM MIDBAND VERSUS RF SELECTIVITY

Desired RF Selectivity	Frequency Spacing from Midband	
	150 mc	450 mc
20 db	±1.5 mc	± 5 mc
40	±3	±10
60	±6	±20

TABLE IX — SIGNIFICANT RF BAND VERSUS ANTENNA SPACING

Minimum Separation Between Receiver and Unwanted Transmitters	RF Band	
	150 mc	450 mc
50 ft	±6 mc	±14 mc
500 ft.	±3	± 7
½ mile	±2	± 4

tivity can be reduced to one-third and possibly to one-fourth of the above values by the use of bulky and expensive filters.

The critical frequency band that needs to be considered in determining the usefulness of any given channel can be obtained by combining the information given in the two preceding tables with the results shown in Table IX. For example, if it be desired to work mobile receivers unrestricted to within 500 feet of two or more unwanted transmitters, all frequency assignments within ±3 mc in the 150-mc band (or within ±7 mc in the 450-mc band) must be carefully chosen if intermodulation interference is to be avoided.

When the ±3-mc band is divided into 100 potential channel assignments of 60 kc each and when the channels assigned to a given area are chosen at random, 7 channels working 50 per cent of the time (or 37 channels working 10 per cent of the time) will, on the average, cause third order intermodulation interference about 10 per cent of the time on each channel within the band. The interference is expected to be above the minimum usable signal level in all receivers located less than about a mile from the unwanted transmitters. Even if the operating frequencies are selected carefully instead of at random, no more than 11 channels out of 100 can be found that are free of third order intermodulation when used simultaneously in the same general area. These results are discussed more completely in a companion paper.[5] When the number of potential channel assignments is greater or less than 100, the corresponding number of usable channels limited by third order modulation alone is shown in Table X. The numbers of usable channels shown above are further reduced when fifth and higher order intermodulation products are considered.

A reduction in the nominal channel spacing from 60 kc to 20 kc means a three-fold increase in the potential channel assignments, but Table X shows that the number of usable channels increases much more slowly.

[5] Babcock, W. C., Intermodulation Interference in Radio Systems. Page 63 of this issue.

TABLE X — NUMBER OF USABLE CHANNELS VERSUS NUMBER OF POTENTIAL CHANNELS

No. of Potential Channel Assignments in RF Band Shown in Table IX	No. of Usable Channels			
	Careful Selection No Interference	Random Selection 10% Chance of Interference		
		% of Time Transmitter Is On		
		50%	25%	10%
20	7	5	10	25
50	9	6	12	30
100	11	7	15	37
200	12	9	18	45
500	14	12	24	60

Thus far, only the intermodulation interference generated in the receivers has been considered. Intermodulation also occurs at the same frequencies in the transmitters, but it usually can be made less important than the corresponding interference in the receivers. Ideally, the intermodulation products generated in the transmitters should not be stronger than 140 db below one watt (about 1 microvolt at the input to the receiver) which requires about 75 db RF filtering in each transmitter output. This ideal requirement is based on 100 watt transmitters with both the transmitters and receiver working on the same antenna. In practice, the RF filtering requirement is less than 75 db because of physical separation between transmitters and receivers, and typical values based on free space transmission are shown in Table XI.

A comparison of the filter requirements on 100 watt transmitters with the corresponding receiver selectivity requirements given in Table VII shows that the receiver requirements are greater as long as the effective

TABLE XI — RF TRANSMITTER FILTERING VERSUS ANTENNA SEPARATION

Distance Between Receiver and Unwanted Transmitters	RF Filtering Needed in Each Transmitter in db							
	150 mc Distance Between Transmitters				450 mc Distance Between Transmitters			
	0 ft	10 ft	50 ft	500 ft	0 ft	10 ft	50 ft	500 ft
0 ft.*	75	—	—	—	75	—	—	—
50 ft.	57	46	39	—	52	36	29	—
500 ft.	47	36	29	19	42	26	19	9
½ mile	40	29	22	12	35	19	12	2

* Common antenna.

separation between transmitters is greater than about 50 feet. For example, with a 500-foot separation between the transmitting and receiving antennas, Table VII shows that the 150 mc requirement on r.f. selectivity is 39 db. The bandwidth between the 39 db points on the receiver selectivity characteristic determines the number of potential channel assignments to be used in Table X.

INCREASED EFFICIENCY OBTAINED BY COORDINATION

The preceding selectivity and filtering requirements are severe and in some cases virtually unattainable except at considerable sacrifice in frequency space. The principal reason for these exacting requirements is that the assumed unrestricted and independent operation results in large differences in field intensities among closely spaced frequencies. In order to pick out the weak signals from among the strong, sufficient selectivity must be provided to suppress the potential interference to below the minimum usable signal.

An alternative is to reduce the level differences and hence the filtering requirements by geographical and operational coordination. This means that the level of the potential interference can be permitted to be many db above set noise as long as it is always at least 10–20 db below the desired signal at all possible locations. By proper coordination the troublesome RF filtering problems can be eliminated within the coordinated system and the remaining IF selectivity problems can be minimized.

The first step is to use the two frequency method of operation with adequate separation between the frequencies used for the opposite directions of transmission. In this way substantial RF filtering can be obtained to eliminate the interference between one or more base transmitters and a base receiver. This type of interference is particularly troublesome between single frequency systems because of the relatively high base transmitter power and because the high antennas at both locations reduce the radio path loss to a minimum. The corresponding possible interference between transmitters and receivers on different mobile units is also reduced by the two frequency method but interference between mobile units is much less important because of the lower power and much lower antenna heights.

The potential interference between base transmitters and mobile receivers caused by insufficient total filtering can be reduced by locating all base transmitters at or near a common point so the level differences between the desired and undesired signals will never be excessive. When all transmitters radiate from a common antenna, a selec-

tivity or filtering requirement of about 40 db (instead of the values shown in Table III) is sufficient for a reasonable signal to interference ratio plus an allowance for differential path losses resulting from standing wave effects.

The RF selectivity or intermodulation problem in the mobile receiver can be eliminated by reducing the power level at the first converter to about 75 db below one watt. This can be done by providing a simple automatic gain control in the RF stage of the mobile receiver. In regions where the desired and undesired signals are weak the receiver has full sensitivity, while at locations near the transmitters both the desired and undesired signals are reduced in level before reaching the first converter. The result is that the intermodulation products generated in the receiver are reduced about 3 db for every db that the desired signal is lowered and the distortion becomes negligible before the output signal-to-noise ratio is reduced appreciably. In order that the a.g.c. circuit can be fully effective it is necessary that the transmitters be grouped together and that the desired carrier be transmitted to control the gain of the receiver.

Grouping the base transmitters at or near a common point together with the associated measures of transmitting the carriers and using a.g.c. greatly reduces the requirements on the mobile receiver, but these measures complicate the design of the base transmitter. The intermodulation products generated in the closely associated transmitters result in potential interference both within and outside of the desired transmitting band. The intermodulation that falls on the mobile receiver frequencies needs to be suppressed by at least 25 db below the carrier on any channel to prevent mutual interference within the coordinated system. The intermodulation that appears as extra band radiation outside the frequency range of the coordinated system must be suppressed by RF filters. The guard band needed to prevent mutual interference between the coordinated system and its neighbors is small compared with the frequency space that is saved by the close spacing of the channels within the coordinated system.

In the direction of transmission from the mobile transmitters to the base receivers, the above coordinating methods cannot be used but equally effective ones are available. The RF selectivity requirements shown in Table VII can be reduced 20 db by using 20 db less power in the mobile transmitter than in the base transmitter. This measure is somewhat analogous to the use of a.g.c. in the opposite direction of transmission; a further step would be automatic control of the radiated power but this complication does not appear to be necessary.

In order to regain the full coverage area, multiple base receivers at

different locations are needed and this use of space diversity techniques provides an opportunity to pick the receiver having the best signal-to-noise ratio. Moreover, the low power in the mobile transmitter together with the better RF filters that are possible in fixed locations reduces the critical bandwidth within which intermodulation interference can arise to about ±0.4 mc at 150 mc and to about ±0.6 mc in the 450 mc range. In these bandwidths approximately 20–25 channels can be obtained which with random location of the mobile units would be divided more or less uniformly among five or more base receiving stations. Since no more than 4 or 5 channels would be operating within the critical RF bandwidth at any one receiving location, the possibility of intermodulation interference is almost negligible. Finally, an off-channel squelch circuit is provided which disables the base receiver at a location where serious adjacent channel interference is most likely to occur and forces the choice of another base receiver in a different location. Another effect of the off-channel squelch circuit is that it keeps the base receiver quiet during idle times, and in this respect it is analagous to the advantage gained in the mobile receiver by continuous transmission of the desired carrier at the base transmitter.

Most of the above coordinating methods tend to emphasize and to increase the characteristic differences between the two directions of transmission. The net effects are that greater frequency economy is obtained and that the electrical requirements are reduced on the mobile equipment where size, weight and power are critical and where cost savings are important because of the large number involved. An increase in complexity occurs at the multi-channel base station but this seems economically justified because the cost can be divided among many working channels.

When the above methods of coordination are fully utilized, the RF requirements are eliminated in the mobile equipment and can be met in the base station equipment. In addition, the IF selectivity requirement on nearby channels is reduced to about 40 db in the mobile receiver and to about 60 db in the base receiver. The extra band radiation requirement on nearby channels is reduced to about 25–40 db in the base transmitter, depending on whether one or more than one antenna is used; and to about 60 db in the mobile transmitter.

These requirements coupled with the data given in Tables II, IV and V lead to the frequency separation between coordinated channels operating in the same area as given in Table XII. The channel spacings are shown for AM and for FM with a frequency swing of ±5 kc (which requires a bandwidth of ±8 kc for good quality).

The spacings shown in Table XII assume that each channel is trans-

TABLE XII — CHANNEL SPACING VERSUS SYSTEM STABILITY— COORDINATED SYSTEMS IN SAME AREAS

Stability	Channel Spacing							
	150 mc				450 mc			
	Mobile Receiver		Base Receiver		Mobile Receiver		Bass Receiver	
	AM	FM	AM	FM	AM	FM	AM	FM
±0.001%	21	31	25	35	27	37	31	41
±0.002%	24	34	28	38	36	46	40	50
±0.005%	33	43	37	47	63	73	67	77

mitted on an individual carrier. Single-channel operation seems to be the only practical arrangement for transmission from the mobile transmitter to the base receiver. In the other direction of transmission, from base transmitter to mobile receiver, the question naturally arises whether additional frequency economy could be achieved by multichannel methods. In this case individual carrier operation is also indicated for transmission and economic reasons. The multiple echoes that exist at street level in urban areas limit the number of usable channels that can be transmitted on a single carrier.[6] While the exact number is somewhat indefinite, it appears to be less than about 20 and perhaps less than 10 channels. In addition the selectivity and linearity requirements on multi-channel receivers (even for two channels) are much more severe than for single channel equipment. From these considerations it appears that the use of more expensive receivers and channel separation equipment in each mobile unit is not economically feasible.

FREQUENCY ECONOMY IN PRESENT AND PROPOSED MOBILE SYSTEMS

The technical factors given above provide a basis for estimating the number of usable mobile channels that can be obtained in a given bandwidth. This bandwidth must be sufficiently large to be isolated by RF filtering if the results are to be well defined.

The following examples assume two different geographical distributions: (1) the number of usable channels with overlapping coverage areas that can be obtained within a city or metropolitan area, and (2) the number of usable channels that can be obtained when the channels are distributed more or less uniformly over a state or other large area. The examples are based on the use of frequency modulation with a

[6] Young, W. R., Jr., and L. Y. Lacy, Echoes in Transmission at 450 Megacycles from Land-to-Car Radio Units. I.R.E., Proc., pp. 255–258, March, 1950.

modulation bandwidth of ±8 kc and a frequency stability of ±.002 per cent; with these assumptions, the IF passband should be at least 22 kc in the 150 mc band and 34 kc in the 450 mc band. Narrower bandwidths could be used but this would result in a substantial sacrifice in coverage under impulse noise conditions.

Five cases are considered:

(1) *Single Frequency Semi-Coordinated*—In this case, substantially no interference is expected from third order modulation problems, which are avoided by careful selection of operating frequencies, but higher order modulation products may be important. Base station locations are unrelated geographically to other systems in same general area, except that a minimum spacing of 500 feet between receiver and interfering transmitter is assumed.

(2) *Single Frequency with Interference*—In this case, the choice of frequencies is unrestricted, but a 10 per cent chance of third order intermodulation interference is accepted within 500 feet of unwanted transmitters, when transmitters are in operation 25 per cent of time.

(3) *Two Frequency Semi-coordinated*—This is the same as (1), except with two-frequency operation.

(4) *Two Frequency with Interference*—Same as (2) except with two frequency operation.

(5) *Fully Coordinated Broad-band*—This case assumes: (a) two frequency operation with the land transmitters coordinated in location, power, antenna height and emission of protective carriers; (b) low power mobile transmitters; (c) multiple land receivers; (d) no interference from third or higher order intermodulation; and (e) guard bands to protect mobile and neighboring services from mutual interference.

The number of usable channels that can be obtained in the same area is estimated in Table XIII for frequencies near 150 mc.

The minimum channel spacing shown in the first column of Table XIII is calculated as follows: in cases (1), (2), (3) and (4), the extra band radiation from the base transmitter is controlling. As shown in Tables III and V, to avoid interference for distances greater than 500 feet from the interfering transmitter requires a guard band of about 50 kc. This is added to the 22 kc required IF pass-band of which ±8 kc is allowed for the FM signal, and ±3 kc for 0.002 per cent system instability. In (5), the adjacent channel receiver selectivity is controlling: Table IV shows the required 60 db can be obtained in 15 kc, which added to the required 22 kc IF band gives approximately 40 kc.

It will be noted from Table VII that the assumption of a separation of 500 feet between the receiver and the interfering transmitter requires

TABLE XIII — USABLE CHANNELS IN CITY AT 150 MC

Method of Operation	Minimum (Not Average) Channel Spacing in Same Area	Number of Usable Channels in 6 mc
(1) Single frequency semi-coordinated	75 kc	10
(2) Single frequency with interference	75	14
(3) Two frequency semi-coordinated	75	5
(4) Two frequency with interference	75	7
(5) Fully coordinated broad-band*	40	45

* Includes three guard bands of 0.8 mc each to protect mobile and neighboring services from mutual interference.

about 40 db RF selectivity, and from Table VIII that the 40 db selectivity requires that all frequencies within ±3 mc need to be considered. With 75 kc channel spacing, there are 80 potential assignments in 6 mc. Table X indicates that 10 one-way channels can be found that are free of mutual third order intermodulation interference. If the available bandwidth were 12 mc the number of interference-free channels would be doubled.

By the same process from Table X, we derive the number of usable channels shown for case (2).

For cases (3) and (4), the methods are the same, but the number of usable channels is reduced to one-half that shown for the single frequency cases.

In the fully coordinated broad-band system (case 5) a usable one-way channel can be obtained every 40 kc. However, three guard bands totaling 2.4 mc are provided to protect both the mobile and neighboring systems from mutual interference. If the available bandwidth were 12 mc the number of interference-free channels would be increased from 45 to 120 since no additional guard bands would be required.

The comparison between various methods of operation given in Table XIII applies to 150-mc channels operating in the same city. When the channels are distributed more or less uniformly over a large area, the number of usable channels is increased by several factors. The separation between carrier frequencies in non-overlapping areas needs to be only slightly greater than the IF pass-band of the receiver, say, 30 kc at 150 mc. The guard bands needed in one location can be used in other areas at geographical separations less than co-channel spacings. Finally the required geographical separation between co-channel stations is less for the two frequency method than for the single frequency method and is less for FM than it would be for AM.

An estimate of the maximum number of usable channels within a large

TABLE XIV — USABLE CHANNELS IN STATE OR LARGE AREA AT 150 MC

Method of Operation	Minimum Channel Spacing*	Number of Usable Channels in 6 mc
(1) Single frequency semi-coordinated	25 kc	108
(2) Single frequency with interference	25	171
(3) Two frequency semi-coordinated	25	108
(4) Two frequency with interference	25	171
(5) Fully coordinated broad-band	25	240

* Assumes adjacent channels are not assigned in same area.

area can be obtained by considering an area whose radius is about six times the coverage radius of the individual transmitter. A larger area is unnecessary because single frequency FM channel assignments can be repeated at this distance, while a smaller area would tend to approach the common area concept used above. The large area can be divided into 9 subareas, each of which can be treated in the manner used in Table XIII. The results are shown in Table XIV, which again assumes an FM modulation bandwidth of ± 8 kc and ± 0.002 per cent overall system frequency stability.

The entries in Table XIV are calculated as follows: Once again, the smallest band to be considered is limited by the RF selectivity in mobile receivers to 6 mc; with 25 kc as the minimum channel spacing, there are $6000/25 = 240$ potential assignments. From Table X, only 12 can be found to be free of third order intermodulation. With 12 channels in each of 9 subareas, there is a grand total of 108 channels usable in the state or large area. With more frequency space, the usable number is increased in proportion.

By the same process, from Table X we derive the number shown for case (2).

In the two frequency cases, the co-channel separation can be made smaller than in the single frequency cases, since the most troublesome case of interference (that between base transmitters and base receivers) is eased by RF selectivity. Thus, the co-channel separation needs to be only about 0.7 that for single frequency operation, which means that there are now effectively 18 instead of 9 subareas. It follows that the grand total of usable channels is the same in cases (1) and (3) and cases (2) and (4).

In considering case (5), we note from Table XIII that 40 kc is the minimum channel spacing usable in a single subarea. However, the largest grand total of channels is found by using 50 kc spacing in the subareas, and assigning the adjacent 25 kc channels to other subareas.

TABLE XV — USABLE CHANNELS IN CITY AT 450 MC

Method of Operation	Minimum (Not Average) Channel Spacing	Number of Usable Channels in 14 mc
(1) Single frequency semi-coordinated	85 kc	12
(2) Single frequency with interference	85	17
(3) Two frequency semi-coordinated	85	6
(4) Two frequency with interference	85	9
(5) Fully coordinated broad-band*	50	68

* Includes three guard bands of 2.4 mc each to protect mobile and neighboring services from mutual interference.

Similarly, the guard bands of one subarea can be used for channels elsewhere so all of the available 240 channels can be used.

The examples given in Tables XIII and XIV represent the two extreme conditions and the practical situation lies in between the two.

By similar reasoning it is possible to estimate the number of usable channels that can be obtained at frequencies around 450 mc. The number of usable channels shown in Table XV is for overlapping coverage areas in a city or metropolitan area and the estimates given in Table XVI are based on a uniform distribution over a state or other large section of the country.

Again, FM modulation with a bandwidth of ± 8 kc and a system frequency stability of 0.002 per cent are assumed.

For a bandwidth of 28 mc instead of 14 mc the number of usable channels is doubled for the first four cases and is increased from 68 to 208 for the fifth case. The corresponding estimates for bandwidths less than 14 mc are indefinite because of insufficient r.f. selectivity.

CONCLUSIONS

The principal conclusions that result from Tables XIII, XIV, XV and XVI, and from the preceding discussion can be summarized as fol-

TABLE XVI — USABLE CHANNELS IN STATE OR LARGE AREA AT 450 MC

Method of Operation	Minimum Channel Spacing*	Number of Usable Channels in 14 mc
(1) Single frequency semi-coordinated	35 kc	117
(2) Single frequency with interference	35	198
(3) Two frequency semi-coordinated	35	117
(4) Two frequency with interference	35	198
(5) Fully coordinated broad-band	35	400

* Assumes adjacent channels are not assigned in same area.

lows:

1. A fully coordinated system requires a band of several megacycles that can be treated as a unit, but it offers substantial overall frequency economy and freedom from interference that can be obtained in no other way. This is particularly true in large metropolitan areas where the demand is greatest. With the same equipment and the same standards of quality and reliability, coordinated channels can always be spaced much closer in frequency than uncoordinated systems.

2. The advantages of coordination increase rapidly as the number of channels per unit area is increased. However, in areas where only three or four channels are required, the advantages of complete coordination are sufficiently small that only the semi-coordination of careful frequency allocation is required to preserve overall frequency economy.

3. For maximum economy, where full coordination is not used, the channels should be assigned as in FM and TV broadcasting first to areas and then to users within areas. The allocation of a block of channels to a particular service with a minimum of operational and geographical restriction frequently results in an ever-increasing interference problem as each additional station is placed in operation.

4. Single-frequency operation is most suitable where the operational need for single channel communication between mobile units (as contrasted with fixed-to-mobile) is more important than frequency economy.

5. A frequency separation between potential channel assignments of 25 kc in the 150 mc range, and 35 kc in the 450-mc range seems technically feasible; but adjacent channels with these minimum spacings cannot be assigned in the same area. These values may be reduced to about 20 and 30 kc, respectively, at the sacrifice of an appreciable reduction in coverage under impulse noise conditions. A further reduction in channel spacing would not appreciably increase the total number of usable channels, since the controlling factors are RF selectivity and extra band radiation, rather than IF selectivity or the total number of potential channel assignments.

6. The *average* spacing needed between channels operating in the same area varies from about 40 to 500 kc or more, depending on the method of operation and the criterion of usability.

7. The need is for a certain small number of channels in all areas, plus a peaked demand in centers of population. In the semi-coordinated cases, the maximum number of channels that can be allocated to the peak area is a small fraction of the total number of channels available. In the fully coordinated, broad-band case, there is much more flexibil-

ity and the peak area can be allocated a large fraction of the total available.

8. FM is preferable to AM for land mobile service because its instantaneous gain control feature minimizes the flutter caused by the motion of the mobile unit through standing wave patterns. This advantage increases in importance as the carrier frequency increases. In addition, FM with an adequate frequency swing provides an increased signal-to-noise advantage over most of the coverage area. The somewhat greater channel width required by FM is more than offset on an area coverage basis by the closer co-channel spacing.

Intermodulation Interference in Radio Systems

Frequency of Occurrence and Control by Channel Selection

By WALLACE C. BABCOCK

(Manuscript received August 25, 1952)

Intermodulation interference becomes a serious factor in frequency usage when a block of consecutive channels is provided for a given type of radio service in a confined area. Formulas are presented which show the number of potentially interfering 3rd and 5th order intermodulation products that can be formed in a band of n consecutive channels. The probability of encountering interference when a number of operating channels are picked at random from this band of n channels is developed and the number of interference free operating channels that can be obtained by careful selection in this same band is also derived.

When a block of consecutive radio channels is used in a confined area to provide a given type of service, interference becomes a serious problem. The situation is aggravated by the fact that whenever energy at two or more radio frequencies combines in a nonlinear circuit, as in transmitter output stages or in receiver input stages, products at other than the original frequencies are created. These are called intermodulation products, and they are capable of causing serious interference within the block of channels assigned to a given type of service as well as in other bands assigned to other types of service.

It is important in engineering a service to know something about the nature of these products in order to evaluate their interference potentialities and to study means of controlling or minimizing that interference. The numbers and locations of various types of intermodulation products are susceptible to mathematical computation. Whether or not all of these products would produce actual interference depends on the geographical locations of transmitters and receivers, and on their specific

electrical characteristics. This paper discusses the number of potential interferences, and in effect, envisages a situation where potential interferences are strong enough to be actual interferences.*

GENERAL

Intermodulation products are commonly referred to as 2nd, 3rd, 4th $\cdots$ nth order products depending on the order of nonlinearity which gives rise to the products. Interference within a system is not generally experienced from the even order products because the frequency separation between the channels involved and the product formed by them is so great that the selectivity of the transmitter and receiver radio frequency circuits is sufficient to reduce it to a negligible amount. Some of the odd order products can be discounted also for the same reason. There are odd-order products, however, involving both sums and differences of operating frequencies in such fashion that the frequencies of the products formed are very close to those which generated them. These products are those referred to throughout the remainder of this paper since they are the most likely to cause interference. The most general form of 3rd order interference occurs when three frequencies, A, B and C, intermodulate in such fashion as to produce interference on a channel operating at frequency D. In this case

$$A + B - C = D$$

Another form of 3rd order interference occurs when the second harmonic of A intermodulates with B to produce interference on a channel operating at frequency C. In this case

$$2A - B = C$$

In like fashion the following forms of 5th order interference may occur.

$$A + B + C - D - E = F$$

$$2A + B - C - D = E$$

$$A + B + C - 2D = E$$

$$2A + B - 2C = D$$

$$3A - B - C = D$$

$$3A - 2B = C$$

* Bullington, K., Frequency Economy in Mobile Radio Bands. Page 42 of this issue.

NUMBER OF INTERMODULATION PRODUCTS AND FREQUENCY BAND AFFECTED

It is of interest to know how many intermodulation products can be produced by a block of n uniformly spaced channels and where they will fall with respect to the frequency band occupied by the n channels.

Third Order Products

When all possible combinations of n uniformly spaced operating channels are activated three at a time, $n^2(n-1)/2$ products will be formed and they will lie between $A-(n-1)$ and $A+2(n-1)$ where the operating band lies between channels A and $A+(n-1)$. This means that the bandwidth of the intermodulation products is very nearly three times that of the operating channels. The products are symmetrically distributed with respect to the midpoint of the operating band. There will be

$$\frac{2n^3+3n^2-2n-6}{24}$$

third order products that will fall in the $n-1$ channels immediately below the operating band and a like number of products that will fall in the $n-1$ channels immediately above the operating band. The remaining products,

$$\frac{4n^3-9n^2+2n+6}{12}$$

in number, will fall in the n channels that constitute the operating band. Not all of these products, however, are capable of producing interference since some products of the $A+B-C$ type can fall on the very transmitting channels that combine to produce them. These products are harmless since they do not fall on receiving channels and are generally of much lower level than the carrier on the channels in which they do fall. If such products are not counted, there remain

$$\frac{n}{3}(n-1)(n-2)$$

products which fall within the operating band. The formulas presented here and in Table I are empirical and were derived for values of n up to 10. However, it is believed that they are reasonably accurate for much larger values of n.

Fifth Order Products

When all possible combinations of n operating channels are activated five at a time

$$\frac{n^2(n-1)(n^2-n+4)}{12}$$

fifth order products will be formed and they will lie between $A - 2(n-1)$ and $A + 3(n-1)$ where the operating band lies between channels A and $A + (n-1)$. This means that the bandwidth of the intermodulation products is very nearly five times that of the operating channels. The products are symmetrically distributed with respect to the midpoint of the operating band. It has been found empirically that if all products known to fall on transmitting channels are excluded, there can still be formed

$$\tfrac{1}{6}[6n^4 - 87n^3 + 575n^2 - 2214n + 3922]$$

potentially interfering products which fall within the operating band. This is strictly true only for values of n that exceed 8 and are not multiples of 3. There will remain

$$\tfrac{1}{12}[n^5 - 14n^4 + 179n^3 - 1154n^2 + 4428n - 7844]$$

products that will fall either outside the band or on transmitting channels within the band. Since relatively few of these products fall on transmitting channels within the band the above expression gives to a high degree of approximation the number of products that will fall outside the operating band.

Numbers of Potentially Interfering 3rd Order and 5th Order Products

The formulas given in the two preceding sections were developed by considering individually the various types of 3rd order and 5th order products. Table I shows the general formulas which apply to these individual types of 3rd and 5th order products from which the summation formulas given in the preceding sections were obtained. The number of potentially interfering 3rd and 5th order products is shown in Table II for specific transmission bands containing 10 and 20 consecutive channels.

PROBABILITY OF INTERMODULATION INTERFERENCE WHEN OPERATING CHANNELS ARE PICKED AT RANDOM

In an uncoordinated radio communication system it may be assumed that p operating channels are assigned on a random basis in a band con-

TABLE I — NUMBER OF POTENTIALLY INTERFERING INTERMODULATION PRODUCTS

n = Number of consecutive channels in available band

Type of Product	Number of Products	
$2A - B$	$\frac{(n-1)^2}{2}$ if n is odd	$\frac{n(n-2)}{6}$ if n is even
$A + B - C$	$\frac{(n-1)(n-3)(2n-1)}{6}$ if n is odd	$\frac{n(n-2)(2n-5)}{6}$ if n is even
$3A - 2B$	$\frac{n}{3}(n-3)$ if n is a multiple of 3	$\frac{(n-1)(n-2)}{3}$ other n's
$3A - B - C$	$3(n^2 - 11n + 32)$	
$2A + B - 2C$	$\frac{1}{2}(n^3 - 9n^2 + 34n - 56)$	
$2A + B - C - D$	$3(n-3)(n-5)^2$ for $n > 6$	
$A + B + C - 2D$	$4(n-3)(n-5)^2$	
$A + B + C - D - E$	$(n-5)(n-6)(n^2 - 11n + 37)$ for $n \neq 8$	

taining n consecutive channels. Let us suppose further that the average busy time of the channels is such that T represents the portion of time that an average channel is activated by a transmitter and R represents the portion of time that an average channel is connected to a receiver. The probability of interference, I, may be defined as the probability that one or more of the intermodulation products that are formed when pT channels are transmitting will fall on a specific channel in the operating band. The method used to determine I is based on the assumption that the distribution of intermodulation products is uniform over the operating band. It is further assumed that the magnitudes of the intermodulation products as encountered at the receiver input, are always strong enough to cause interference. Table III shows formulas for I that have been developed for each type of third order and fifth order product. Fig. 1 shows plots of p versus I when only 3rd order products are considered and Fig. 2 shows plots of p versus I when both 3rd and 5th order products are considered with n and T as independent variables.

1. Fig. 2a shows that a band in excess of 500 adjacent channels is required to limit the probability of 3rd and 5th order interference to 10 per cent ($I = 0.1$) when 10 operating channels are picked at random from that band if traffic is such as to fully use these 10 channels, 5 for

TABLE II — TYPE AND NUMBER OF INTERMODULATION PRODUCTS

Type of Product	Number of Products	
	10 Channels	20 Channels
$2A - B$	40	180
$A + B - C$	200	2,100
$3A - 2B$	24	114
$3A - B - C$	66	636
$2A + B - 2C$	192	2,512
$2A + B - C - D$	525	11,475
$A + B + C - 2D$	700	15,300
$A + B + C - D - E$	540	45,570

TABLE III — FORMULAS FOR I, PROBABILITY OF INTERFERENCE

n = Number of consecutive channels in available band.
p = Number of operating channels picked at random from those in band.
T = Portion of time average channel is activated by transmitter.
m = Number of intermodulation products of a specific type expected to fall within band of n channels when pT channels are activated.

Type of Intermodulation	Formula for m
$2A - B$	$m_1 = \dfrac{pT(pT-1)(n-2)}{2(n-1)}$
$A + B - C$	$m_2 = \dfrac{pT(pT-1)(pT-2)(2n-5)}{6(n-1)}$
$3A - 2B$	$m_3 = \dfrac{pT(pT-1)(n-2)}{3n}$
$3A - B - C$	$m_4 = \dfrac{3pT(pT-1)(pT-2)(n^2-11n+32)}{n(n-1)(n-2)}$
$2A + B - 2C$	$m_5 = \dfrac{pT(pT-1)(pT-2)(n^3-9n^2+34n-56)}{2n(n-1)(n-2)}$
$2A + B - C - D$	$m_6 = \dfrac{3pT(pT-1)(pT-2)(pT-3)(n-5)^2}{n(n-1)(n-2)}$
$A + B + C - 2D$	$m_7 = \dfrac{4pT(pT-1)(pT-2)(pT-3)(n-5)^2}{n(n-1)(n-2)}$
$A + B + C - D - E$	$m_8 = \dfrac{pT(pT-1)(pT-2)(pT-3)(pT-4)(n-5)(n-6)(n^2-11n+37)}{n(n-1)(n-2)(n-3)(n-4)}$

$I = 1 - \left(1 - \frac{1}{n}\right)^{m_1 + m_2}$ when only third order products are considered.

$I = 1 - \left(1 - \frac{1}{n}\right)^{m_1 + m_2 + \cdots + m_8}$ when both third and fifth order products are considered.

transmitting and 5 for receiving. ($T = 0.50 = R$) In this case it is the $2A + B - 2C$ type of product that requires the use of such a wide band.

2. Fig. 2a shows that there is practical certainty of interference in a band of 500 adjacent channels when 30 operating channels, picked at random from that band, are fully used. In this case it is the $A + B + C - D - E$ type of product that requires the use of such a wide band.

3. If the same total traffic as was assumed in (1) is handled by a greater number of operating channels, the number of available consecutive channels required for the same probability of interference remains the same. Thus, Fig. 2b shows that a band in excess of 500 consecutive channels is still required to limit the chance of interference to ten per cent when the traffic is distributed among 20 randomly selected operating channels; ($T = 0.25 = R$) similarly Fig. 2c shows that the required number of available consecutive channels remains the same when the traffic is distributed among 40 operating channels. ($T = 0.125 = R$).

CHANNEL SELECTION FOR THE ELIMINATION OF INTERMODULATION INTERFERENCE

Discounting the effect of selectivity in the radio equipment, it was shown in the preceding section that only a very limited number of channels can operate together without some degree of mutual interference when these channels are picked at random from a very considerable number of available channels. This is of course extravagant of frequency space. In this section, it is proposed to determine whether frequency space can be conserved by carefully selecting the operating channels in such fashion that the various types of intermodulation products that are formed will all fall on other than operating channels. This is readily accomplished in the case of 3rd order products by selecting the operating channels in such fashion that the frequency difference between any pair of these channels is unlike that between any other pair of channels. Many other rules inherently more complicated and more cumbersome to apply than the one stated above must be obeyed if 5th order as well as 3rd order products are to be controlled in this way. Table IV presents p operating channels selected from a band of n adjacent channels (numbered sequentially in order of ascending frequency) in such fashion as to avoid 3rd order interference within the system. Considerable effort has been spent in selecting these channels to insure that the number n associated with each value of p is the lowest

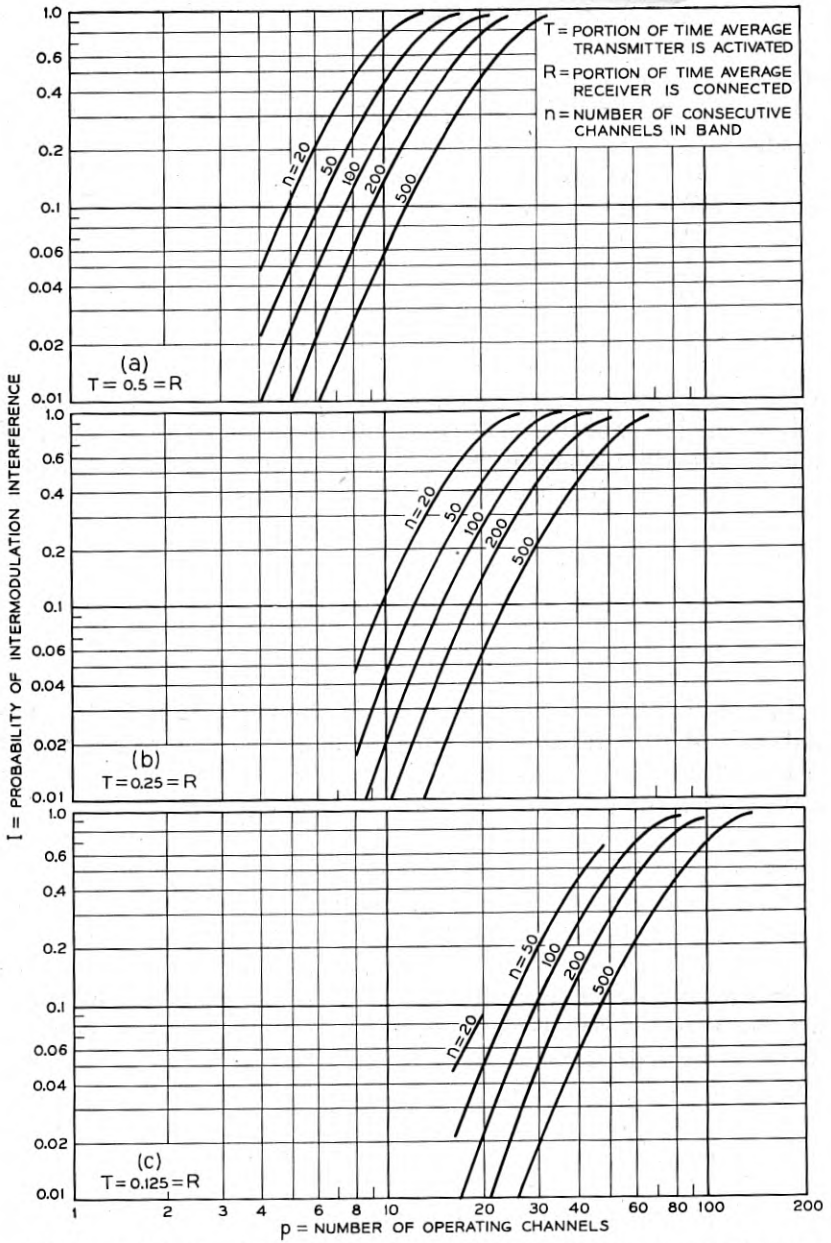

Fig. 1—Probability of intermodulation interference I versus number of operating channels p when only 3rd order products are considered. (a) T = Portion of time average transmitter is activated = 0.50; R = Portion of time average receiver is connected = 0.50. (b) Same as above except $T = 0.25 = R$. (c) Same as above except $T = 0.125 = R$.

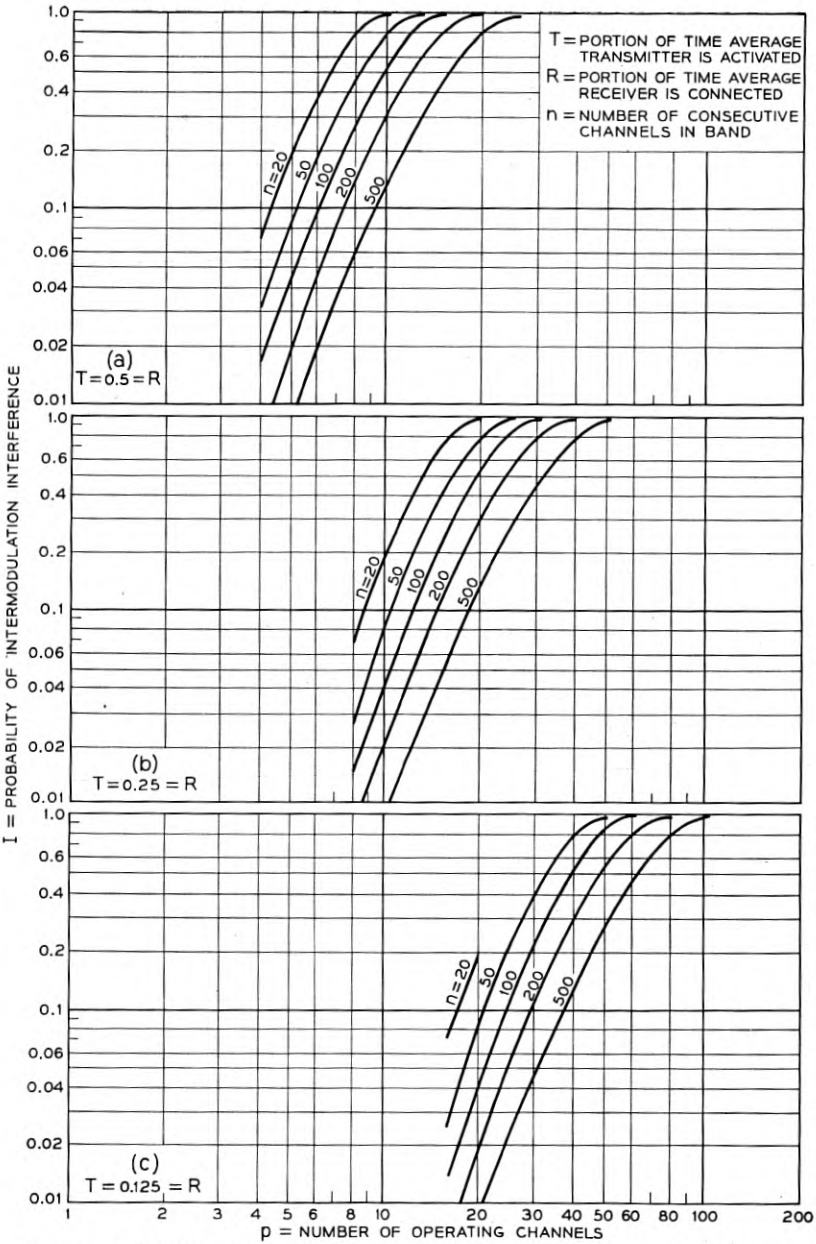

Fig. 2—Probability of intermodulation interference I versus number of operating channels p when both 3rd and 5th order products are considered. (a) T = Portion of time average transmitter is activated = 0.50; R = Portion of time average receiver is connected = 0.50. (b) Same as above except $T = 0.25 = R$. (c) Same as above except $T = 0.125 = R$.

TABLE IV — SPECIFIC OPERATING CHANNELS HAVING NO THIRD ORDER INTERMODULATION INTERFERENCE

p is the number of interference free operating channels which can be obtained from n consecutive channels.

p	n	Operating Channels Having No 3rd Order Interference
3	4	1, 2, 4
4	7	1, 2, 5, 7
5	12	1, 2, 5, 10, 12
6	18	1, 2, 5, 11, 13, 18
7	26	1, 2, 5, 11, 19, 24, 26
8	35	1, 2, 5, 10, 16, 23, 33, 35
9	46	1, 2, 5, 14, 25, 31, 39, 41, 46
10	62	1, 2, 8, 12, 27, 40, 48, 57, 60, 62
8*	137	1, 2, 8, 12, 27, 50, 78, 137

* Neither 3rd nor 5th order interference exists with this selection of eight operating channels.

possible number from which p channels having no 3rd order interference can be selected.

A plot of p versus n based on the above table is shown in Fig. 3. The curve which includes both 3rd and 5th order intermodulation products shows that 300 consecutive channels must be made available to provide for the careful selection of 10 operating channels which are interference-free at all times, regardless of the traffic load. For comparison, it was shown earlier (Fig. 2) that more than 500 consecutive channels must be available to permit picking at random 10 operating channels

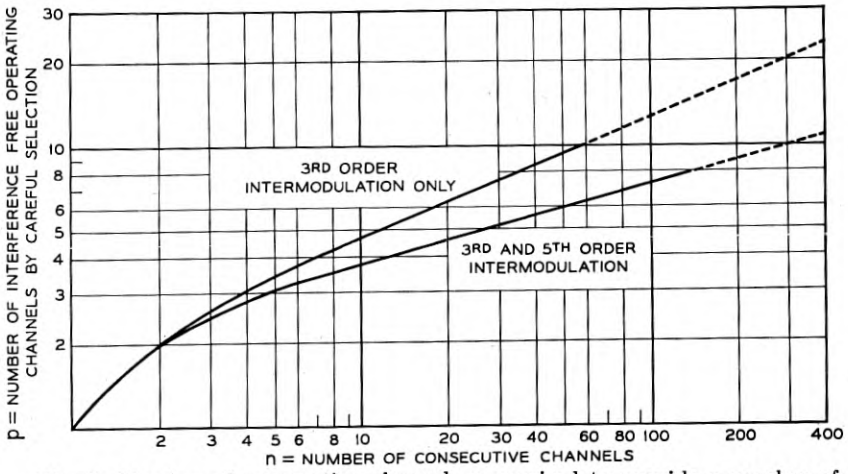

Fig. 3—Number of consecutive channels n required to provide a number of interference free operating channels p.

which are subject to interference 10 per cent of the time when fully loaded with traffic.

Careful channel selection is therefore a step toward better frequency usage. It is, however, only a small step, and a more effective solution must be found if efficient frequency usage is to be achieved in areas where sizeable numbers of operating channels are required. Such a solution for the mobile radio service is a "coordinated system," proposed in a companion paper previously referred to, wherein additional measures are described for reducing the probability of interference by proper geographical system layout and the use of certain practicable operational features.

Magnetic Resonance

PART I—NUCLEAR MAGNETIC RESONANCE

By KARL K. DARROW

(Manuscript received September 3, 1952)

Magnetic resonance is the name of a phenomenon discovered less than sixteen years ago, which from the start has had a high theoretical importance and is now attaining a notable practical value. Nuclear magnetic resonance occurs when a substance containing magnetic nuclei is exposed to crossed magnetic fields, one being steady and the other oscillating, and the strength of the former field and the frequency of the latter are matched in a particular way. When these are properly matched, the nuclei are turned over in the steady field, and energy is absorbed from the oscillating field. Another way of describing the effect is to say that resonance occurs when the applied frequency is equal to the frequency of precession of the nuclei in the steady field. This phenomenon illustrates very clearly some of the fundamental laws of Nature. For the purposes of nuclear physics it is used to determine the magnetic moments of nuclei and their relaxation-times in the substance that contains them. It is also used for chemical analysis, for measurement of magnetic fields, for analysis of crystal structure and for locating changes of phase of the substance containing the nuclei. Magnetic resonance of electrons is similar, but for a fundamental reason is confined almost exclusively to free atoms of certain kinds, to ferromagnetic substances and to certain strongly paramagnetic salts. For these last it serves to throw light on the fields prevailing within the crystals.

"Magnetic" is an ancient word in physics and so is "resonance," but "magnetic resonance" is something new. It is the name of a phenomenon which is sharp and clearcut and easy to evoke, which springs directly from the ultimate magnetic particles of matter, which illustrates the fundamental laws of these, and which has found and still is finding uses of importance. There are two types of it, the nuclear and the electronic. Nuclear magnetic resonance is the theme of the first part of this article: it will recur from time to time in the second part (to appear in a later issue of this JOURNAL) but the main topic of that second part will be

electronic resonance. It is fitting that they should be treated in this order, for the nuclear type of resonance is less distorted by complexities than is the other. Perhaps it is not premature to say that while nuclear magnetic resonance always goes by that name, the electronic type is usually called "paramagnetic resonance" or "ferromagnetic resonance."

Nuclear magnetic resonance was realized in 1937, in molecular-beam experiments. The war distracted physicists, and the next great step was not made until after — but very soon after — the armistices. In the winter months of 1945–46 the phenomenon was produced in liquids and in solids. The news burst upon the world from the pages of *The Physical Review* in the early weeks of 1946, causing among physicists an immediate and an immense sensation. Of some discoveries one wonders how they came to be made at all, of others one wonders afterward why they were not made earlier. Nuclear magnetic resonance is of the latter class. But this is a discovery that could not have been made *much* sooner than it was, for it required the apparatus and techniques of short-wave radio and microwaves, and these are recent.

The work of 1945 was done by two independent groups three thousand miles apart, using somewhat different experimental methods and expounding the theory in somewhat different ways. The differences are really superficial, and in the course of time will probably be minimized; but the two streams of later work that rose from those two sources are still distinguishable. The methods are called the nuclear resonance absorption method and the nuclear induction method: I treat them in this order. A sketchy account of the molecular-beam method will follow upon these, and then several of the applications — which of these are major and which are minor must be left for history to decide.

On the first few pages, and on many thereafter, the talk will be of protons. Protons are the commonest material particles in Nature, electrons excepted (neutrons are also an exception but not an important one here, as they are seldom found free). Protons also have the happy attribute called "spin $\frac{1}{2}$" soon to be explained, which simplifies the exposition greatly. This is one of the rare fields of physics in which the simplest case, the commonest case, and the most useful case, are all three of them one and the same.

PROTON RESONANCE ABSORPTION

To begin with, there must be a sample containing hydrogen, protons being the nuclei of ordinary (as distinguished from heavy) hydrogen atoms. It may be pure hydrogen in gaseous, liquid or solid form, or any

one of countless compounds of hydrogen. I first take water for the sample, and enumerate the other particles in water. There are the nuclei of the common isotope of oxygen, oxygen 16: they are non-magnetic and produce no resonance. There are the nuclei of rarer isotopes of oxygen and hydrogen: they will be mentioned later. There are the electrons: they are reserved for Part II of this article. We are now left with the protons.

The sample is placed between the poles of a magnet, Fig. 1, so that it is in a magnetic field which should be homogeneous and is usually strong. The strength of the field is denoted by H, and its direction is always that of the z-axis (and usually vertical). I should like to call it "the steady field," but usually it is modulated during the experiments, so I shall call it "the big field". Actually it can be very small, but nearly always it is between 8,000 and 15,000 gauss, and 10,000 gauss is a good figure to keep in mind.

The big field must not be the sole magnetic field applied to the sample. There must also be an alternating or oscillating field — stationary electromagnetic waves, formed in a solenoid (or sometimes in a resonant cavity). Such waves comprise, as Maxwell taught us long ago, an alternating electric field and an alternating magnetic field. In most of the uses of electromagnetic waves it is the electric field that counts, and the magnetic field is remembered only as something demanded by Maxwell's equations to keep the electric field going. In this application the electric field takes a back seat, and it is the magnetic field that counts. This oscillating magnetic field must be at right angles to the big field; we lay the x-direction along it. Its amplitude, to be denoted

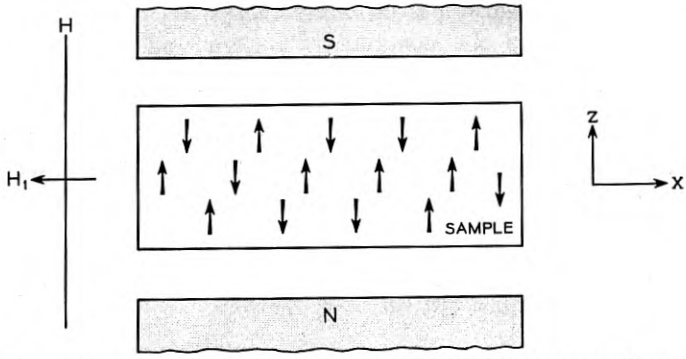

Fig. 1 — Scheme of the apparatus for observing nuclear magnetic resonance. The detecting circuits are omitted. The nuclei indicated by the arrows are of "spin ½," protons for example.

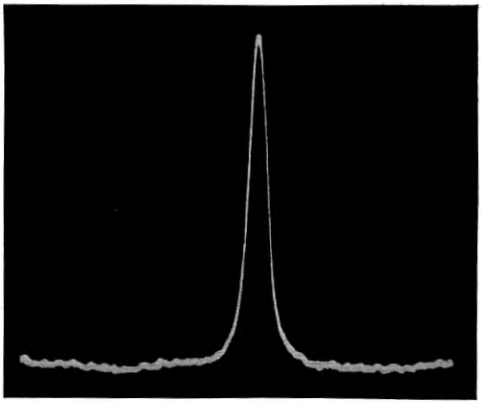

Fig. 2 — Peak of nuclear resonance absorption. This is the first peak to be published other than those obtained with molecular beams. It pertains to protons in water. (Courtesy of E. M. Purcell).

by H_1, is of the order of a small fraction of one gauss up to several gauss. Its frequency must be of the order of tens of megacycles. To be more specific, the effect that is sought with protons is located at 42.6 mc when H is 10,000 gauss.

Finally there must be circuits and detectors for measuring the absorption of the electromagnetic wave-energy in the sample. These are well known to those proficient in the art: we pass them over.

Now of the two quantities H and ν either is to be varied while the other is to be kept constant, and the absorption is to be measured. Usually H is varied while ν is kept constant, and the data consist of a plot of absorption against H for a set value of ν.

When such a curve is plotted it proves to be, in the main, a smoothly-sloping curve, of no interest in the present connection. What *is* of interest is that it is interrupted by a magnificent peak of extraordinary sharpness, deserving to be called a needle. Probably there is nothing that can please an experimenter more than a curve with a fine sharp peak: here he has it. Fig. 2 exhibits the first such peak on record. But neither Fig. 2 nor any other picture can convey an adequate idea of the sharpness of the peak, for the distance from this imposing feature to the axis of the ordinates at field-strength zero may be, and often is, tens of thousands of times as great as the breadth of the peak. (With the induction-method, peaks have been distinguished from each other that are separate by only a millionth of the value of H at which they are found). This needle has the narrowness that is characteristic of fine lines in optical spectra; and this is as it should be, for a spectrum-line

is just what it is, even though it lies in the radiofrequency range and for technical reasons appears on a scope instead of on a photographic film.

The peak is the phenomenon of magnetic resonance. We shall now interpret it in terms of theory somewhat oversimplified, for it is not the office of these opening paragraphs to introduce all of the complexities of quantum mechanics.

In Fig. 1, inside the rectangle which represents the sample, appear a number of arrows. These are symbols of the protons. In other circumstances we might imagine the protons as solid balls, in still others we might imagine them as centres of force; but for the present purpose we are regarding them as tiny bar-magnets, and the arrows symbolize the directions in which they are pointing. It is necessary to label these directions with perfect clearness. The figure has been drawn with the south pole-piece of the big magnet (the one responsible for the big field H) above. The point of each arrow represents the *north* pole of the protonic magnet.

Thus the arrows pointing upward represent protons in the orientation into which the big field would like to turn the protonic magnets, and would indeed succeed in turning them if they were literal compass-needles in literal compasses. The arrows pointing downward represent protons in the opposite orientation. I will call these, for shortness, the "up" orientation and the "down" orientation. Evidently if the physicist could reach into the sample with fingers or with forceps and turn a proton from the up orientation into the down one, he would be doing work upon the proton at the expense of energy from his muscles. Well, he cannot reach into the sample with fingers or with forceps and grasp and turn a proton. But he can reach into the sample with the oscillating field and turn the protons, and this is the experiment we are considering. Magnetic resonance is the turning of protons from the up orientation into the down one, from the orientation or "level" of lesser energy into the orientation or level of greater energy.

But why does the effect occur at one frequency only? And what determines that frequency? To cope with this problem we shall have to introduce symbols, equations, and quantitative reasoning.

The first step is to evaluate the work required to turn the proton, or, in other words, the energy-difference between the two orientations or levels. It shall be denoted by W, and the magnetic moment of the proton by μ_p. We proceed by strictly classical reasoning. The torque exerted on the proton by the magnetic field H is $-\mu_p H \sin\theta$. Here θ stands for the angle between the direction of the steady field and the direction in which the magnetic moment of the proton is pointing. We have

admitted the existence of only two values of θ, viz. the values 0° and 180°; more will be said about this later; but for the duration of this particular argument we shall have to admit all values of θ from 0° to 180°. The value of W which we are seeking is the integral of $\mu_p H \sin\theta$ from 0° to 180°, from the up orientation to the down one. It is easily obtained:

$$W = -\int_{0°}^{180°} \mu_p H \sin\theta d\theta = 2\mu_p H \tag{1}$$

Having arrived at equation (1) by strictly classical reasoning, we must now approach equation (2) by a starkly quantal argument. Immense amounts of evidence have shown that when energy is absorbed from electromagnetic waves of frequencies ν in the optical range of the spectrum and in the X-ray range, not to speak of other ranges, it is invariably absorbed in parcels or quanta equal to $h\nu$, h standing as always for Planck's constant. If this doctrine is sometimes difficult to assimilate when applied to the optical spectrum, how much more difficult it is to accept when applied to waves of radio frequencies! Yet here also it is to be accepted, so we put:

$$W = h\nu \tag{2}$$

Now we transfer the value of W from equation (1) to equation (2), and arrive at the destination:

$$H = \tfrac{1}{2}h\nu/\mu_p \tag{3}$$

In this equation h is known with very great accuracy, and μ_p had also been measured when the first experiments upon magnetic resonance were made, though not with nearly the accuracy that physicists now claim for it. It remains only for the experimenter to insert for ν the value of the frequency in his experiment and for H the value of the fieldstrength at which the peak appears. The test is whether the two sides of the equation agree. Needless to say, the test has been brilliantly passed.

Quantum-theory has entered into this argument in more ways than the one which led to equation (2). I return now to the fact that we have arrived at equation (3) by postulating two, and only two, "permitted" orientations of the protonic magnets in the steady field. This is illustrated by the presence, in Fig. 1, of arrows pointing up and arrows pointing down but no arrows pointing slantwise. We might have assumed that there are protons, and therefore arrows, pointing in every direction. We might have assumed that there is a proton pointing, say, at angle

76°13′ to the vertical, and that it can absorb a quantum $h\nu$ of just the right energy to turn it to the angle 118°36′. This would have led to the inference that instead of absorption confined to the fieldstrength $h\nu/2\mu_p$ corresponding to the actual peak, there would be absorption at every fieldstrength from $h\nu/2\mu_p$ on upwards toward infinity. The experiment frustrates this inference, and so declares for the two and the only two permitted orientations. One did not have to wait for this experiment to learn this fact: it has been known for thirty years, both as a consequence of quantum mechanics and as a fact of experience. However this is a very pretty proof of it.

We now must generalize equation (3) so as to make it take care of all nuclei and not the proton only; and in the course of this process we shall meet the actor behind the scenes who determines the permitted orientations. His names are *spin* and *angular momentum.*

THE GENERAL EQUATION FOR NUCLEAR MAGNETIC RESONANCE

We are now en route to the general equation of which (3) is the special case appropriate to the proton. Our first step takes us to the deuteron or nucleus of heavy hydrogen. Its magnetic moment differs from that of the proton, so we must write μ_d instead of μ_p. More significant is the fact that the deuteron has three permitted orientations in the big field instead of two. The orientations of proton and deuteron are shown in the first and third columns of Fig. 3; beside them are horizontal lines depicting their energy-values, energy being measured vertically upward from an arbitrary zero.

One guesses from the aspect of Fig. 3 that the deuteron will show three peaks of magnetic resonance; for it seems possible for the deuteron to be turned from orientation a to orientation b, from b to z and from a all the way to z. But of these three conceivable "transitions" the third

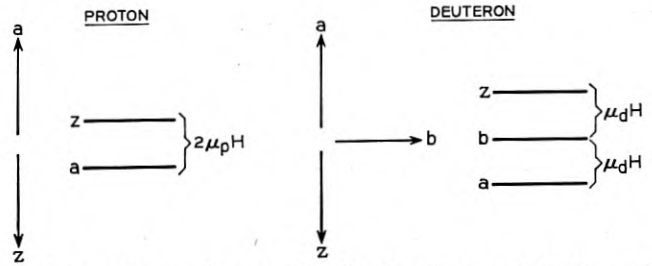

Fig. 3 — Orientations and energy-levels of protons and deuterons in a magnetic field, according to the "old" quantum-theory.

does not occur at all: the theorists know the reason why, and call it a "forbidden" transition. As for the other two, the mere symmetry of the picture shows that they involve equal absorptions of energy and therefore contribute coincident peaks. There is therefore only one distinguishable peak, and we have to find the value of H at which it appears. This is easily done. Going back to equation (1), we integrate the integrand which there appears from 0° to 90° (or from 90° to 180°); and so we come to the analogue of equation (3) which applies to the deuteron:

$$H = h\nu/\mu_d \tag{4}$$

In the course of this argument we have met with an example of two general rules: *no matter how many permitted orientations there are, transitions occur only between consecutive ones, and these permitted transitions always agree in energy-absorption, so that there is never more than one peak.* Yet equation (4) differs from equation (3), because in the right-hand member $h\nu/\mu$ — and now I am using μ as the general symbol for magnetic moment — is multiplied by ½ for the proton and by one for the deuteron. Now, ½ is the value of the spin of the proton and one is the value of the spin of the deuteron. We generalize from these two instances: we use I as the general symbol for the spin; and we arrive at the following:

$$H = (I/\mu)h\nu \tag{5}$$

The generalization is sound; and *equation (5) is the fundamental equation of nuclear magnetic resonance.*

I now have to interpret the word "spin." Spin is a particular measure of the angular momentum of the nucleus. That a magnetic nucleus has angular momentum is surely not surprising. We are trained to ascribe magnetism to the motion of charged bodies: an electric current flowing in a loop has the same magnetic field as a bar-magnet. When a nucleus is observed to have a magnetic moment and an angular momentum, it is natural to correlate one property with the other: one does not quite know how far the analogy may safely be pressed, but at least it is helpful.

But what sort of a measure of the nuclear angular momentum is the quantity I? The answer to this question is confused by the fact that in our times there have been two forms of quantum theory: the "new" quantum mechanics which is undoubtedly more competent in general, and the "old" quantum theory of the nineteen-twenties which is certainly more simple in the present case. Desire to be clear has led me to employ

the older theory up to now, but conscience obliges me to introduce the new one.

In the old theory, I is the nuclear angular momentum in terms of the unit $h/2\pi$. Two of the permitted orientations, which I will call the "extreme" ones, are straight along and straight against the field-direction. For these, the projections of the angular momentum upon the field-direction are $+Ih/2\pi$ and $-Ih/2\pi$. For the proton $I = \frac{1}{2}$, and the two extreme orientations are the only ones.

In the new theory, the nuclear angular momentum in terms of the unit $h/2\pi$ is $\sqrt{I(I+1)}$. For the two extreme orientations, the projections of the angular momentum on the field-direction are $+Ih/2\pi$ and $-Ih/2\pi$, just as they were in the old theory. But now these orientations are no longer straight along and straight against the field-direction. They must be inclined, one to the up direction and the other to the down direction, at the angle of which the cosine is $I/\sqrt{I/(I+1)}$.

Thus I has partly changed its meaning: it is still the maximum permissible projection, upon the field-direction, of the nuclear angular momentum in terms of the unit $h/2\pi$, and this is what it was before; but it is no longer the magnitude of the nuclear angular momentum. So also has μ changed a part of its meaning. It is the maximum permissible projection, upon the field-direction, of the magnetic moment of the nucleus, and this it was before; but it is not the magnitude of the nuclear magnetic moment. The language of this subject has not been well adjusted to this change. Fortunately I is called the "spin," which does not necessarily convey the impression that it is quite the same thing as angular momentum; but μ is still called the "magnetic moment," and in the new quantum mechanics this is a mistake.

Fig. 4 is Fig. 3 redrawn in the spirit of quantum mechanics. The arrows now represent angular momentum and magnetic moment jointly,

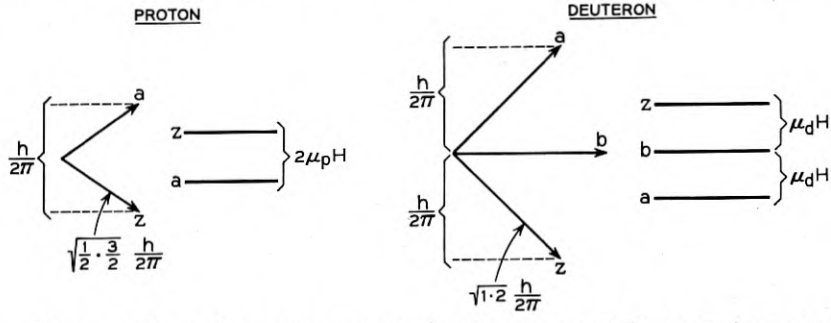

Fig. 4 — Orientations and energy-levels of protons and deuterons in a magnetic field, according to quantum mechanics.

but the numbers affixed to them are the values of angular momentum. The energy-levels in the second and fourth columns are often known as "Zeeman levels." I take this occasion to complete the statement about the allowed orientations, which in recent paragraphs has been made for the extreme orientations only. The projections of the nuclear angular momentum upon the field-direction are

$$+Ih/2\pi, +(I-1)h/2\pi, \cdots, -(I-1)h/2\pi, -Ih/2\pi.$$

From this principle combined with the fact that transitions occur only between consecutive levels, follows rigorously equation (5), which I derived in a looser way.

Spins are ascertained in various ways, usually from their influence on the electrons surrounding the nuclei, which manifests itself in details of optical spectra and in cleverly-designed molecular-beam experiments. They are always integer multiples of $\frac{1}{2}$. Important instances of nuclei of spin $\frac{1}{2}$ are the proton and the nucleus F^{19}. The neutron and the electron also belong in this category, as we shall see later on. The deuteron has already provided us with an important instance of a nucleus of spin one. Spins as high as $\frac{9}{2}$ are certainly known, and this is probably not the limit. Nuclei of spin zero are common: I have already mentioned one of them, oxygen 16. Such nuclei do not produce magnetic resonance; we shall have nothing to do with them.

A brief table shall conclude this section. To what has already been stated it adds the number of permitted orientations corresponding to each value of spin.

Spin	$\frac{1}{2}$	1	$\frac{3}{2}$	I
Number of orientations	2	3	4	$2I+1$
H for peak	$(\frac{1}{2})h\nu/\mu$	$h\nu/\mu$	$(\frac{3}{2})h\nu/\mu$	$(I/\mu)h\nu$

THE LARMOR PRECESSION AND NUCLEAR INDUCTION

Now we go back to first principles, make a fresh start, and arrive by a different route at the equation for magnetic resonance. On this route we meet with a vivid justification of the use of the name "resonance."

Resonance implies a tuning or a matching between an applied frequency and a frequency either actually or potentially present in the substance in question. A piano-wire, a membrane, the air in an organ-pipe, an electrical circuit comprising capacity and inductance, all resonate to the frequency which is that of their own natural vibrations. No mention has yet been made of a frequency peculiar to the nucleus which is matched by the applied frequency when magnetic resonance

occurs. There is indeed such a natural frequency, not however a frequency of vibration; it is a frequency of *precession*. Precession is a concept well known to astronomers and to such physicists as have to do with gyroscopes, perhaps not so well known as it should be to others.

In Fig. 5, the vertical is again the direction of the big magnetic field. The arrow represents the angular momentum of the nucleus, which I now denote by p. The magnetic field H exerts a torque on the nucleus. I have already given an expression for this torque, but I gave it in the language of the "old" quantum-theory. To employ this expression with as little apparent change as possible, I introduce the symbol μ_0 for the magnetic moment of the nucleus, and reserve μ for

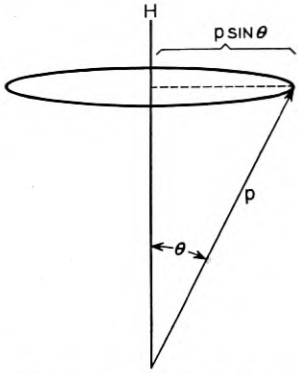

Fig. 5 — Illustrating the Larmor precession.

the maximum permissible projection of μ_0 on the field-direction. The torque now appears on the right-hand side of the following, purely classical, equation:

$$dp/dt = \mu_0 H \sin \theta \tag{6}$$

This is a vectorial equation, but I will endeavor to express its vectorial content by words instead of symbols. Fix the attention on the tip of the arrow. The torque makes it describe a circle of radius $p \sin \theta$ in the horizontal plane, with a frequency which I denote by ν. Its peripheral speed in this circle is ν multiplied by the circumference of the circle, therefore $\nu \cdot 2\pi p \sin \theta$. This speed is dp/dt. Putting its value into (6), we observe with pleasure that θ vanishes from the scene: the result is going to be the same for all orientations of the magnet: this is it:

$$H = 2\pi p \nu / \mu_0 \tag{7}$$

Making the substitutions that have already been describe, we get:

$$H = (I/\mu)h\nu \tag{8}$$

and this is none other than formula (5), the fundamental equation of magnetic resonance. The precession-frequency is the resonance-frequency.

This precession is often called the "Larmor precession," and the frequency given by (5) or (8) is called the "Larmor frequency." The name is a posthumous honor; Larmor died before magnetic resonance was discovered; his theory was applied to the Zeeman effect, the effect of magnetic fields upon optical spectra.

It is not hard to believe that when the applied frequency coincides with the Larmor frequency, something drastic must happen to the precession. The theory has been worked out on a classical basis. I will not pursue it into its details; but at least the first step should be taken.

I have said that the alternating field is perpendicular to the big field. We take the x-direction as its direction. The magnetic field, or magnetic vector as I will henceforth call it, has then $H_1 \cos 2\pi\omega t$ for its x-component (I use ω for the frequency so as to distinguish it from the Larmor frequency) and zero for its y-component. Now imagine a vector, of constant magnitude $(\frac{1}{2})H_1$, lying in the xy-plane, pointing away from the z-axis and revolving around this axis *clockwise* with frequency ω. Its x-component will be $(\frac{1}{2})H_1 \cos 2\pi\omega t$, its y-component will be $-(\frac{1}{2})H_1 \sin 2\pi\omega t$. Imagine another such vector revolving *counterclockwise*. Its x-component will be $(\frac{1}{2})H_1 \cos 2\pi\omega t$, its y-component will be $(\frac{1}{2})H_1 \sin 2\pi\omega t$. (It is evident that we have chosen their phases so as to bring about this result). The sum of these two vectors has $H_1 \cos 2\pi\omega t$ for its x-component and zero for its y-component. But this is the vector that we started out with. In the language of optics, we have resolved a plane-polarized wave into two circularly-polarized ones.

The foregoing is pure mathematics. Now comes the physics. Of these two revolving vectors, one is whirling in exact unison with the precessing magnet when ω is exactly equal to the Larmor frequency, the other is rushing round and round in the opposite direction. Our intuition tells us that the former may be expected to produce a great effect on the precession, the latter a small one. The latter is not always negligible, but may be neglected here. Thus in this artful way we have substituted a circularly-polarized field for the actual plane-polarized one.

The theory further leads to the prediction that when resonance exists, the precession will be exaggerated in such a way as to produce

an alternating magnetic flux across the xz plane. Now I describe an actual experiment, the first of its type.

The sample is water (or something else) in a spherical container. Around the container are wrapped two coils at right angles to one another. The coil of which the axis is parallel to the x-axis produces the alternating field. The coil of which the axis is parallel to the y-axis is connected with a rectifier and a detector. At resonance there is an alternating magnetic flux through the latter coil, and by the operation of the rectifier this is converted into a signal on the scope. The signal locates the resonance-frequency as accurately as does the peak in the absorption-method. This is the phenomenon called "nuclear induction."

I terminate this section by mentioning a paradox resulting from precession. Everyone has seen a compass-needle turning to point to the north: it is natural to infer that when a magnetic field is applied to a piece of matter, the elementary magnetic particles of which the nuclei (and also the electrons) are examples will automatically turn to point along the field. Yet the analogy fails and the inference is false: the nuclei do not turn to point along the field, but each of them maintains a constant angle with the field while it precesses. It seems to follow that matter cannot be magnetized by a magnetic field, but again the inference is false. Animistically speaking, the field makes the nuclei want to turn into its direction, but they cannot fulfill their desire without assistance from something other than the field. This something-other is not absent, and in the section on "relaxation" we shall meet with it.

THE MOLECULAR-BEAM EXPERIMENT

There are three methods for detecting and locating nuclear magnetic resonance, and we have now considered two of them. In one of these, the resonating nucleus makes itself manifest by absorbing energy; in the other, that of nuclear induction, by radiating energy; in the one which is to come, by simply failing to turn up at the scene of the measurement. This singular attribute is that of the molecular-beam experiment, which (I repeat) was done before the others and so receives the credit of revealing nuclear magnetic resonance. Molecular-beam experiments are so remarkable that it is hard to speak of them without yielding to temptation to say more than is essential to the purpose, but here the temptation must be withstood.

Conceive a narrow stream of hydrogen-containing molecules coming along the (horizontal) axis of y, and cutting across a big magnetic field parallel to the (vertical) axis of z. This big field differs from that of

Fig. 1 in one important way: it is *non-uniform*, increasing in strength from (say) the bottom to the top. In one respect the protons behave just as they do in a sample in a uniform field: roughly half of them are pointing up and the other half pointing down. But in the non-uniform field the "up" protons experience a net force pushing them upward and the "down" protons a net force pushing them down. (Visualizing each of the protons as a tiny bar-magnet, one sees that the fieldstrength is bigger where the upper pole of the magnet is than where the lower pole is). The beam is parted into two diverging pencils, the one containing the "up" protons only and the other the "down" protons only; I call the first the "up" pencil and disregard the second.

The "up" pencil now passes through a region just like that implied in Fig. 1: a magnetic field which is big and vertical and *uniform*—H will stand for its strength—and an oscillating field with the magnetic vector parallel to the x-direction. If in this second region some of the protons are turned by the oscillating field into the "down" orientations, that will make no difference to their course across the remainder of the second region where H is uniform. But beyond the second region lies a third where again there is a big field that is non-uniform. In this third region the "up" protons go one way and the "down" protons go another. The detector lies athwart the first way; the "down" protons will miss it.

The detector-reading is plotted against H for a set value of ν. One might think that two curves would be plotted, one with the alternating field off and the other with it on, and that the latter would be systematically lower than the former. But the latter will be lower than the

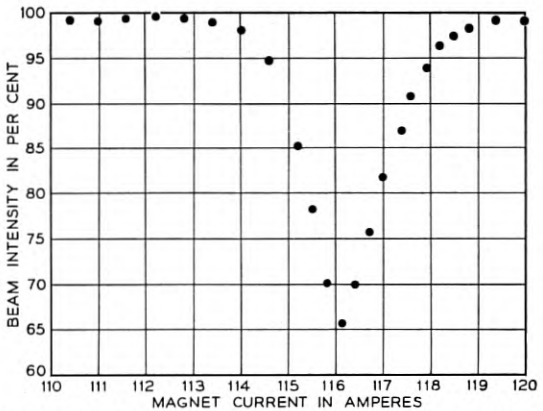

Fig. 6 — Negative peak or valley of nuclear resonance absorption obtained by the molecular-beam method. It pertains to lithium nuclei in lithium chloride molecules. This was the first experimental evidence of nuclear magnetic resonance. (I. I. Rabi, J. R. Zacharias, S. Millman and P. Kusch).

former only in the immediate vicinity of the value of H which conforms to equation (3); for the protons are turned over only when the Larmor frequency agrees very nearly with the applied frequency. Accordingly one keeps the alternating field on all the time and plots a single curve; and this is marked by a fine sharp peak, but this time a peak that points *downward*, Fig. 6, for it testifies to the absence of the overturned protons that have missed the detector.

The first experiment of this kind was done on molecules of lithium chloride. The reader may have been puzzled that I spoke of a beam of *molecules* and then of the deflection of *protons:* the protons, or whatever other magnetic nuclei are being studied, carry the molecules with them. In the experiments on LiCl, the peaks of lithium and of chlorine were found in different parts of the curve. Later the proton-resonance was discovered by using molecules of KOH and NaOH, and confirmed with molecules of H_2 and HD (the latter being a hydrogen molecule of which one nucleus is a proton and the other a deuteron). It is from this molecule of HD that the proton-resonance, and for that matter the deuteron-resonance also, stand out most clearly and sharply. In H_2 and in D_2 the resonances are perturbed and multiplied, but for reasons which are well understood so that the theory is strengthened instead of being weakened; but to describe these pretty things would be confusing unless they were explained, and to explain them would take us far afield.

SOME APPLICATIONS OF NUCLEAR MAGNETIC RESONANCE

The first of the uses of nuclear magnetic resonance is of interest mainly to the nuclear theorist. He wants to know (I/μ) for as many nuclei as possible; and this knowledge may be found by locating the resonance-peaks, and applying to their values of H and ν the equation (3) or (5) which I repeat:

$$H = (I/\mu)h\nu \tag{9}$$

Anyone who is going to burrow into the literature of this subject must be apprised beforehand, or else find out the hard way, that this simple statement is variously expressed. Here is a sad case of the ruination of a beautiful terminology by carelessness. The terms which have been ruined are "gyromagnetic ratio" and "magneto-mechanical ratio." The former ought to mean, as originally it did mean, the ratio of angular momentum to magnetic moment. The latter ought to mean the ratio of magnetic moment to angular momentum. Both have by

now been used in both these senses, and there are variants within each sense, depending on the unit that is preferred by the user. The appearance of either "gyromagnetic ratio" or "magneto-mechanical ratio" in a paper is a red light warning the reader to make sure just what the author means by it. In this paper both of these terms are discarded with regret.

An experimenter may give his value of (μ/I) directly, or may give his value of g, which is (μ/I) expressed in terms of a peculiar unit. The peculiar unit is $eh/4\pi m_p c$, in which m_p stands for the rest-mass of the proton and the other symbols have their normal meanings. This unit got into the picture because there was a doctrine that (μ/I) for the proton ought to be just two of it. This was based on an analogy with the electron which, to the consternation of theorists and the complication of Nature, proved to be fallacious. Reported values of g range from nearly 6 to 0.143; the proton has one of the two highest values, the triton or nucleus of hydrogen 3 has the other. Since all these values may be described without much extravagance as being "of the order of 2," the use of g remains convenient.*

Many people say that they have measured μ. Formally this is all wrong, but practically it is usually all right, for in most if not all cases I is known from experiments of other kinds. Most of these people give the value of μ in "nuclear magnetons." This means that they are giving the value of gI, as is seen from the following equation which resumes in notation what I just said in words, and provides the definition of g:

$$\mu = gI(eh/4\pi m_p c) \tag{10}$$

The quantity in brackets is called "nuclear magneton."

Now that this tiresome but necessary passage is behind us, we can review the results.

Values of gI — or of some other of the quantities catalogued above — have been published for about forty nuclei. The values of gI available toward the end of 1950 were gathered together and published in an article to which I give the reference in a footnote.† The largest is about twenty-five times the smallest: this is a wide range of variation, yet not so wide as that of the nuclear charges or the nuclear masses. Isotopes of one another may have values nearly the same or considerably different; the same is true of isobars. Most of the values are positive: this

* It is perhaps not premature to mention that in optical spectroscopy and in electronic magnetic resonance, the symbol g is used with a similar but not an identical meaning.

† Pake, G. E., American Journal of Physics, **18,** pp. 438–52, pp. 473–86, 1950. The table is on p. 440 of the October issue.

means that the angular momentum and the magnetic moment are parallel. A few are negative: for these (one of which is the neutron) the angular momentum and the magnetic moment are anti-parallel.

These values of μ (for, I repeat, gI is μ expressed in terms of a particular unit) are useful as challenges and as aids to the nuclear theorists. They are challenges, because the μ-value of a given nucleus is something to be explained; they may be aids, because a theory may be fortified by giving the right value of μ or confuted by leading to a wrong one. Now, nuclear theory is difficult, and by and large it is not so far advanced that it can demand experimental values accurate to let us say, the legendary "sixth place of decimals." This is a piece of temporary good fortune, for two reasons.

First, the strength of the big field H may not be known with adequate accuracy at the place where the nuclei are. It can however be ignored if one is concerned only to measure the ratio of the (μ/I) values of two nuclei. The experimenter has then to put into his apparatus successively samples containing the two kinds of nuclei, or a single sample containing them both: the ratio of the frequencies at which the resonance-peaks appear is the ratio of the (μ/I) values, and H vanishes in the division. Often the comparison-nucleus is the proton, so that many published values of gI come ultimately from ratios in which gI for the proton stands in the denominator. Such ratios are frequently adequate for the testing of theories, and their accuracies may be very good indeed, even attaining the sixth significant figure. (The basic determination of μ for the proton itself will be mentioned in Part II.)

Second, the true field which the nuclei experience may be slightly different from the big field H, because of local fields within the substance. This is of course an admission that our fundamental equation, (5) or (9) in this article, can be wrong. So it can be, and this is a development that may be thought distressing. But such developments are almost the rule in physics, whenever the art of measurement is bettered; and in the present case the errors in equation (9) must be regarded as felicitous, for they lead to some of the most fascinating applications of nuclear resonance.

Thus when ammonium nitrate, NH_4NO_3, is put into the apparatus, there are two peaks of nitrogen instead of one. They are not far apart — if for the frequency in use one is at $H = 10{,}000$ gauss the other is at 9,997. The formula NH_4NO_3 suggests, and the diagram of the molecule would confirm if we had it here, that the two nitrogen nuclei are differently placed in the molecule: one may say that they have different atomic surroundings. Thus the position of either of the peaks is dis-

tinctive not of nitrogen alone, but of nitrogen in its particular surroundings. These same surroundings might recur in several different types of molecule, or might be confined to one. The formula of ethyl alcohol may be written as $CH_3 \cdot CH_2 \cdot OH$. This compound presents three proton-peaks, Fig. 7, separated by a few per cent of one gauss when the big field is of the order of 10,000: they have been ascribed to protons in the three "groups" CH_3 and CH_2 and OH. To identify a group is to perform a process of chemical analysis, and this is a nascent application of nuclear magnetic resonance.

This is a good place to speak of the efficacy of nuclear resonance in revealing the presence of chemical elements or of individual isotopes. The proton is one of the easiest nuclei to discern in this way, largely owing to its relatively high magnetic moment. It has been calculated that $2 \cdot 10^{16}$ protons suffice to give a detectable "signal" by the induc-

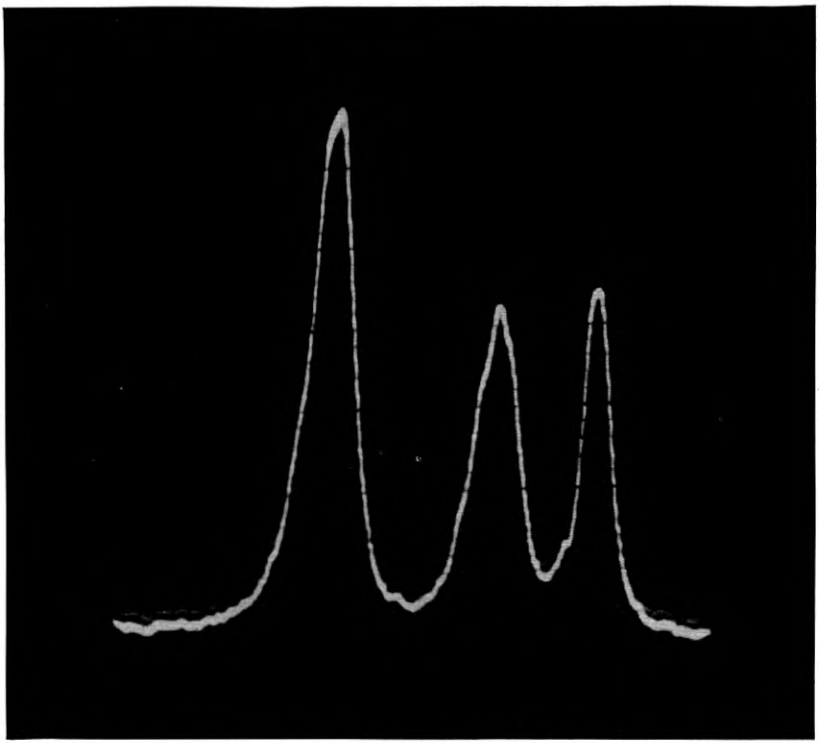

Fig. 7 — Breakup of the proton resonance peak of ethyl alcohol into three peaks, each believed to arise from protons in distinctive 'groups" within the molecule. (Courtesy of M. E. Packard).

tion-method; a small capsule of gaseous hydrogen at a pressure of only one atmosphere will show the resonance of protons. The second isotope of hydrogen is normally present in that substance in an abundance of only 1.5 parts in ten thousand, the second isotope of oxygen exists only in an abundance of four parts in ten thousand; neither was discovered for more than a decade after the search for isotopes was well under way; but both of them have been detected by nuclear resonance.

Another application is to crystallography. In the crystal called gypsum, each proton is exposed to a magnetic field of the order of ten gauss from its neighboring protons. The resonance-peak is split into two or three or even four, depending on the inclination of the big field to the crystal axes. It would take many pages to describe this effect in detail, but it is so intelligible that one may deduce from it the positions of the protons in the crystal lattice. Nuclear resonance in fact seemed called to play a great role in crystallography, since the principal tool of the crystallographer has been the diffraction of X-rays, and this will not disclose the presence nor *a fortiori* the locations of protons in a crystal lattice. However this promising child of resonance has apparently been throttled in its cradle, for the still newer art of neutron-diffraction has proved itself adequate for finding the protons in a lattice.

Another application is to the measurement of magnetic field strengths. One sees that if proton-resonance is produced at a measured frequency in a steady field of which the magnitude H is unknown, H may be determined by equation (3) with an accuracy contingent on the accuracy with which μ_p is known, and this is pretty high. This has become a common method of measuring magnetic field strengths.

RELAXATION

If anyone were asked to guess the most important use of nuclear magnetic resonance, he would have two good reasons for choosing the study of relaxation. More pages of the scientific journals have been devoted to it than to any other application. Moreover, the discoverers spoke of it almost as soon as they spoke of the discovery; one has the feeling that they were so confident of the discovery, that as soon as it was made they considered it much less important for its own sake than as a tool.

"Relaxation" is a word that entered long ago into physics. Its general meaning is the gradual self-adjustment of a system to a sudden change in conditions. In the immediate instance the system is our sample in the big magnetic field; the sudden change in conditions is the starting or

the stopping of the oscillating field; and what gradually adjusts itself is the distribution of the protons between the up and the down orientations. Now I will describe an experiment such as has been performed on protons in water.

Let the sample be placed in the big field some time — several hours will always be ample — before the experiment is to begin. The experimenter should know in advance the frequency of the Larmor precession, so that he can apply the oscillating field of proper frequency as soon as he and the sample are ready. The sample then enters into what I will call "the state of resonance." The experimenter is to measure the height of the peak as soon as the oscillating field is switched on: I call this initial stature A_0. The big field and the alternating field are now *both* to be kept on. The height of the peak, A, is to be measured from time to time, say once every tenth of a second (this has been done with movie techniques). It is found that A is a declining function of time; the peak is shrinking.

After a while, let the alternating field be switched off while the big field continues to be on. The state of resonance is now suspended. Again the height of the peak is to be recorded every tenth of a second. Needless to say, the alternating field must be on while the record is being made, but it shall be off all of the rest of the time, which is most of the time. It will be found that the peak is growing again. It is, in fact, trending back to its initial stature A_0, and the law of its rise is the exponential law:

$$A = A_0 [1 - \exp(-t/T_1)]^* \tag{11}$$

The constant T_1, which this experiment determines, is called the "spin-lattice relaxation-time." "Lattice" will be recognized as a term appropriate to crystals: in the literature of this subject it is however applied to all solids and liquids. Its meaning in this field may be put as follows: the "lattice" is all of the sample except the nuclear spins.

The actual experiment is not usually done quite as I just described it. The alternating field does not have to be switched on or off, because if its frequency is far from the Larmor frequency it is practically ineffectual. If the observer wants to end the state of resonance, he displaces H or ν away from the resonance-value; if he wants to restore it he brings H or ν back to the resonance-value. By modulating the big field with say a 60-cycle frequency, he may pass the system briefly

* This formula implies that $A = 0$ when $t = 0$; the reader can recast it to cover the general case in which $0 < A < A_0$ at $t = 0$.

through the state of resonance 60 times in a second; and it is possible to record and measure A on every such passage.

The fall and the rise of A are due to a cause so obvious that the reader has probably guessed it already. The peak, we recall, is a peak of absorption due to the turning of "up" protons into the "down" direction. When an up proton is turned into the down direction it goes out of business as an absorber, and continues out of business so long as it remains in the down direction. Since there is a fall and since there is a rise, the sojourn in the down direction must be neither zero nor infinite. If it were zero there would be no fall, and if it were infinite there would be no rise. The shrinkage of the peak in the presence of magnetic resonance, and the growth of the peak after resonance is discontinued, are signs that the sojourn of a proton in the down direction is finite but not zero. We divine already that T_1 is a measure of the average of this sojourn.

The foregoing may seem to imply that the height A of the peak is proportional to the number of up protons in the sample; but this is not so, and A is proportional to something else which I call the "margin." To present it I use N_u for the number of upward-pointing protons, N_d for the number of downward-pointing protons, N_0 for their constant sum, μ in the sense defined before, k for Boltzmann's constant; and I write down the fundamental theorem of Boltzmann:

$$N_u/N_d = \exp(2\mu H/kT) \tag{12}$$

The "margin" is $(N_u - N_d)$. We find:

$$\begin{aligned} N_u - N_d &= N_d\,[\exp(2\mu H/kT) - 1] \\ &= N_0\,(\mu H/kT) \text{ approximately} \end{aligned} \tag{13}$$

We have approximated by supposing μH to be very small compared with kT, which it is indeed; and by supposing N_u and N_d each to be nearly half of N_0—this second approximation is retroactively verified, for on substituting (for instance) 20,000 oersteds for H and room-temperature for T, one finds that out of two million protons selected at random a million plus seven are pointed up and a million minus seven are pointed down. The margin is thus 14 in two millions; but it may also be regarded as seven in two millions, since if seven protons out of two million should be turned down the margin would vanish and the peak would vanish with it.

Why is the stature of the peak proportional to the margin and not to N_u? The point is, that in addition to turning protons from the up

direction to the down direction, the alternating field also helps protons to turn from the down to the up direction. Processes of the first kind involve absorption, as we already know; processes of the second kind involve release of energy. What the detector receives, and what the peak makes manifest, is the net of the absorptions over the releases. (This effect of the alternating field in helping protons from the down to the up direction is called "stimulated emission").

Now we must scrutinize equation (13) more closely. It is evident that T stands for an absolute temperature: the question is, what is it the temperature of?

One supposes perhaps that T is the temperature of the sample — that is to say, the temperature which would be shown by a thermometer stuck into the sample or possibly into a surrounding bath. And this is indeed what is supposed when the peak has the stature A_0, signifying that the sample has stood long enough in the big field undisturbed by resonance or anything else. When A is A_0 and T is the temperature of the sample, (13) is right. But when A is less than A_0 because the peak is falling, has fallen or is rising, we must choose between saying that (13) is not right, and saying that (13) defines a temperature which is to be called the temperature of the spinning nuclei, or the "spin temperature" for short.

The second choice is made; and this is the most vivid language in which to describe the situation. In this language we say that the resonance elevates the spin-temperature, or heats up the spins; and that after resonance ceases, the spins cool down to the temperature of the lattice. Thus the study of relaxation becomes the study of the heating and the cooling of the spins with respect to the lattice — "lattice" being defined, I recall, as everything in the sample except the spins.

Recorded values of T_1 range from times of the order of hours down to times of the order of ten-thousandths of a second. The highest are exhibited by protons in ice at extremely low temperatures; protons in water have $T_1 = 2.33$ seconds; the lowest values are found in the presence of "magnetic impurities." The typical dependence of T_1 on temperature is represented by a curve with a single minimum.

The importance of "magnetic impurities" derives from the agent of relaxation. Relaxation is operated normally by the varying magnetic fields whereby the nuclei act on one another; these vary, as I shall presently say more fully, because the nuclei are wiggling in thermal agitation. But the magnetic fields of nuclei are comparatively small, and therefore normal relaxation is comparatively slow. Much bigger is the magnetic field of an electron, for the magnetic moment of an electron

is 660 times as great as that of a proton. Now, it is indeed true that all atoms and molecules contain electrons; and one may properly wonder why they do not always dominate the relaxation. The reason is that in most atoms and most molecules the electrons are paired "anti-parallel" so that their magnetic moments neutralize each other. (We are to see in Part II that this confines the electronic type of resonance to certain very specialized types of substances). One can however introduce into a substance, water for example, atoms or ions for which the neutralization is incomplete. These have much bigger magnetic moments and magnetic fields than any nucleus, and they speed up the spin-lattice relaxation. There is one special case which I treat in more detail, because of its relevance in this connection and its importance in solid-state physics.

There are crystals, of fluorite for instance, which occur colorless in Nature. These may be colored by exposing them to X-rays, or in other ways which we pass over here; and colored examples may also be found in Nature. Solid-state physicists have long been acquainted with these colorations, which they ascribe to what they call "F-centers." Various lines of reasoning have converged on the conclusion that an F-center is a cavity in the lattice (now I am using "lattice" in the normal sense, that of the crystallographers) in which a free electron is batting around like a wild animal in a cage. If this is so, then coloration of a colorless crystal by X-rays or otherwise should reduce its relaxation-time, and naturally-colored crystals should have lesser values of T_1 than those that are colorless. Experiment has ratified these inferences, and thus nuclear magnetic resonance has come to confirm the theory of the F-centres. So also has electronic resonance, since the F-centres display it with extreme clarity; but this is a topic for Part II.

The cause of relaxation has now been identified as the thermal agitation of the substance, working through the variation-in-time of the magnetic fields which act on every nucleus, weakly from its neighbor nuclei and strongly from any uncompensated electron that happens to be in the vicinity. In gases and liquids the nuclei cruise around, and so do the "magnetic impurities" if there are any; fieldstrengths change swiftly and relaxation tends to be rapid. In solids the atoms and their nuclei vibrate around fixed positions, and thermal agitation has come to be interpreted in the following way.

Nowadays one thinks of the solid sample as quivering with compressional waves, and perhaps torsional waves as well. These constitute the thermal agitation of the sample, and their various frequencies form its elastic spectrum. From this broad band of frequencies we isolate, in

mind, the one which agrees with the frequency of the Larmor Precession. Think now of any two adjacent protons. The distance r between them will fluctuate in a complicated way as time goes on, but in this complicated motion we distinguish, again in mind, the component which has the frequency of the Larmor Precession. Well, according to earlier theory it is this component of the vibrations — be they interpreted as vibrations of the whole lattice or of two neighboring protons relative to each other — which helps the protons to turn from down to up, or for that matter from up to down. This is the channel by which energy passes to and fro between the lattice and the spins.

On submitting this idea to calculation it was found to give values of T_1 that are far too long. The next recourse was to take into account the "beat tones." Choose any frequency whatever in the elastic spectrum, and then another frequency differing from the first one by the frequency of the Larmor Precession. The sum of the two will present a beat frequency equal to that of the precession. This is a mathematical statement which seems as empty of physical meaning as — well, as seemed in its turn the assertion that the alternating magnetic vector H_1 directed along the x-axis is the sum of two circularly-polarized vectors. But the force between two neighboring protons is not linear (it is proportional to the inverse third-power of the distance r), and this gives physical meaning to the statement: the two frequencies conjointly act as if the beat-frequency were present. When these channels of communication between the spins and the lattice are added to the one first thought of, the calculated relaxation times come down into the order of magnitude of the real ones. More in the way of precise agreement can scarcely be hoped for, because of the effect of impurities on T_1.

Two other topics in the field of spin-lattice relaxation must at least be mentioned.

Some of the known values of T_1 are too low to be measured by observing the rise and fall of the resonance-peak. To indicate how these are measured, I recall that the energy of a wave-train is proportional to the square of its amplitude, H_1^2 in the present case. To speak of protons for simplicity: if T_1 were zero the number of protons in the "up" orientation would always be the same, and hence the height A of the peak would be proportional to H_1^2. But since T_1 is not zero the number of protons capable of absorbing goes down as H_1^2 goes up; and the curve of A against H_1^2 starts off tangent to the ideal straight line for $T_1 = 0$, but is concave-downward, drops away from the straight line and eventually will cease to rise.

We may pursue the argument one step farther. Here is the equation

for the rate of change of N_u, the number of "up" protons:

$$dN_u/dt = (1/T_1)(N_{0u} - N_u) - bH_1^2N_u \qquad (14)$$

Here N_{0u} stands for the number of "up" protons in the condition of equilibrium between the spin-temperature and the lattice-temperature. If we were dealing with the rise of the peak after the alternating field is shut off, the second term on the right would vanish, and we should be back at equation (11). We are however dealing here with the fall of the peak when the alternating field is on. One sees that eventually N_u will reach a constant value — it is said to "saturate" — and the peak a constant stature. If this saturation-value is measured and the value of b is known, T_1 can be computed. The saturation-value is measured, and b is determined from quantum mechanics.

Though I have tried to avoid giving the impression that the stature of the peak necessarily has its equilibrium-value A_0 before the oscillating field is first applied, I may not have quite succeeded. It would be a miracle if A were equal to A_0 at the moment when the sample is first put into the big field. Time must be allowed for the nuclei to adjust themselves or "relax" to the big field: it was because of this that I said at the beginning that the sample was to be placed in the big field several hours (a generous allowance of time, by the way) before the application of the alternating field. One would expect in general to find A much smaller than A_0, when the sample has just been exposed to the big field; on the other hand it could be greater than A_0 if the sample had previously been exposed to a field of greater strength than the field of the experiment.

Conceivably one might miss the peak altogether by looking for it too soon, if the relaxation-time were long; or by looking for it too late, after it had been reduced to its "saturation" stature. It may be that early attempts to find magnetic resonance were frustrated in these ways. Such dangers are now avoided by mixing the sample deliberately with magnetic impurities in order to diminish the value of T_1 : the peak of Figure 2 was obtained with water mixed with ferric nitrate. There is some reason also to conjecture that nuclear magnetic resonance might have been sought and found some years earlier than it was, but for an imperfect theory which indicated that the spin-lattice relaxation-time would be so long as to make it hopeless to look for the peak.

Students of the literature will find many allusions to another type of relaxation — the "spin-spin" relaxation, with a relaxation-time denoted by T_2. Except for the bare statement that the breadth of the peak varies inversely as the spin-spin relaxation-time, this topic must be left for some other occasion. It may be mentioned here, even though not

explained, that the line-breadth diminishes suddenly when the substance melts, and may also decline at one or more temperatures where the substance is still a solid. Such temperatures are considered to be those at which some special type of molecular motion begins.

References and acknowledgements are to be appended to the second part of this article. Two names will however be mentioned in this place, those of Felix Bloch and Edward M. Purcell; for these are the names of the physicists to whom, on November 6, 1952, was awarded the Nobel Prize for the discovery of nuclear magnetic resonance by the techniques here respectively denoted as those of "nuclear induction" and "nuclear resonance absorption."

Delay Curves for Calls Served at Random

By JOHN RIORDAN

(Manuscript received March 11, 1952)

This paper presents curves and tables for the probability of delay of calls served by a simple trunk group with assignment of delayed calls to the trunks at random and with pure chance call input. These are contrasted with the classic results of Erlang ("Erlang C") which are based on service in order of arrival. Trunk holding times for both have an exponential distribution. The theoretical development for computation of the curves is directed to the determination of the moments, which seem to be a natural means of simplification.

1. INTRODUCTION

One of the classic results in the study of telephone traffic is the formula for delay given by the Danish engineer A. K. Erlang[1] in 1917. This is for random call input to a fully accessible simple trunk group with the trunk holding time exponential and calls served in the order of arrival. A proof for this formula and a set of curves for its use have been given by E. C. Molina.[2]

In many switching systems it is not feasible to fully realize this ethical ideal of first come, first served, and it has long been of interest to determine delays on another basis. The contrasting assumption is of calls picked at random, which is again an idealization but in large offices appears to be called for, as a bound for the service actually given.

The first attempt to formulate the last seems to be that of J. W. Mellor.[3] While his basic formulation is incomplete, it offers a useful approximation to the complete results, particularly in the most interesting region of heavy traffic, and will be referred to here as the "Mellor approximation." A complete formulation due to E. Vaulot[4] appeared in 1946 and included both the fundamental differential recurrence relation and formulas for delay probabilities for small delays. For completeness, these are repeated below. F. Pollaczek[5] has given a development of Vaulot's work directed toward determining an asymptotic delay formula.

Considerable further theoretical work has been necessary to obtain the results given here. Vaulot's differential recurrence relation, which formulates the probability of delay at least t of a call which arrives when n other calls are waiting, has no simple solution. By approximate methods, it was possible to use a differential analyser to determine these probabilities for small values of n. But it was not feasible in this way to cover the whole range of interest, and these results were supplemented by approximations for large n, which are described below. Finally the delay for an arbitrary call was obtained by summing on n.*

These results are not reported here, because the attempt to verify the accuracy attained led to formulation of the moments of the delay curves and this in turn to the representation of the curves as sums of exponential curves, with great simplification of the calculations required. As will appear, two exponentials furnish a sufficient approximation except for heavy traffic.

2. DELAY CURVES

The delay distribution on calls delayed for occupancy levels (defined below) from 0.1 to 0.9 in steps of 0.1 is shown in Fig. 1. The abscissae are derived time units which seem to be natural to the problem: $u = ct/h$, with c the number of trunks and h the average holding time. The ordinates, on a logarithmic scale, are conditional probabilities that a call delayed will be delayed at least u, that is, values of a function $F(u)$; the logarithmic scale is chosen to emphasize the dominantly exponential character of the curves. The occupancy level α is the ratio a/c where a is the average call input in average holding time h.

Fig. 1 is a master curve for all eventualities and may be changed to working curves for various sizes of trunk groups. For the construction of these curves Table I, from which Fig. 1 was made, and which also compares present results with those for calls served in order of arrival, is convenient. A more elaborate table will be given later. For the convenience of the reader, it may be noticed that for order of arrival service $F(u) = e^{-u(1-\alpha)}$.

The striking feature of Table I is the increase in delay time for random service, which becomes more pronounced with decreasing $F(u)$ and increasing occupancy (or traffic) level, α. The increase throughout the table is an effect of the limitation to small values of $F(u)$. For given

* Thanks are due to George W. Abrams for directing this work, to Dr. Richard W. Hamming for transforming the equations into forms suitable for the differential analyser and for supervising its operation, and to Miss Catherine Lennon for a great deal of calculation.

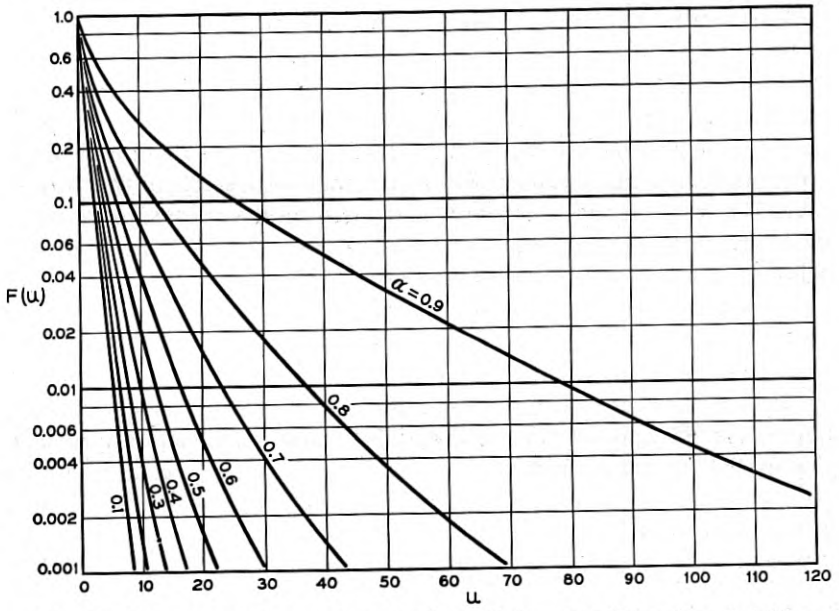

Fig. 1—Delay curves for random service. $F(u)$ = conditional probability of delay at least u; $u = ct/h$, c = no. trunks, h = av. holding time, a = call input in time h, $\alpha = a/c$.

α, the delay curves for order of arrival and random service include the same area, which is in fact equal to the mean delay (of calls delayed), $(1 - \alpha)^{-1}$. Since $F(0) = 1$ for both, and the random service curve decreases more slowly for large u, the curves must intersect at some point, say for $u = u_0$; for $u < u_0$, the o.a. curve must be above the ran-

TABLE I — DELAY-TIME AND RANDOM SERVICE

Delay Times, u, for given $F(u)$ and α and for order of arrival (o.a.) and random service.

α	$F = 0.1$		$F = 0.01$		$F = 0.001$	
	o.a.	Random	o.a.	Random	o.a.	Random
0.1	2.56	2.58	5.12	5.47	7.68	8.60
0.2	2.88	2.91	5.76	6.57	8.63	10.68
0.3	3.29	3.34	6.58	8.05	9.87	13.35
0.4	3.84	3.91	7.68	10.04	11.52	16.95
0.5	4.61	4.68	9.21	12.89	13.82	22.09
0.6	5.76	5.82	11.51	17.25	17.27	29.97
0.7	7.68	8.28	15.36	23.14	23.03	43.33
0.8	11.51	12.57	23.03	36.80	34.54	70.29
0.9	23.03	25.99	46.05	77.24	69.08	156.63

dom curve. This is shown in Fig. 2 for $\alpha = 0.9$, but the logarithmic scale for $F(u)$ obscures the equality of area.

The character of the comparison may be clearer if the picture is changed. Consider a department store counter with c clerks (corresponding to c trunks) in attendance. The time for a sale corresponds to the trunk holding time, and the rate of arrival of customers is like that of call input. For service in order of arrival customers are given serially numbered tickets on arrival; for random service, these tickets may be supposed drawn from a hat, or numbered from a series of random numbers, or since aggressiveness and the clerks' attention are subject to devious rule, it may be that no attention at all to order of service is equivalent to random service.

The fact that the average delay is independent of the order of service may be explained roughly by saying that the average rate at which waiting lines are removed depends only on the average rate of arrival of customers and the rate at which they are served. Notice however that service at random causes more variable delays (the second and all higher moments are larger than for order of arrival service). Thus with random service the proportion of waiting customers receiving quick

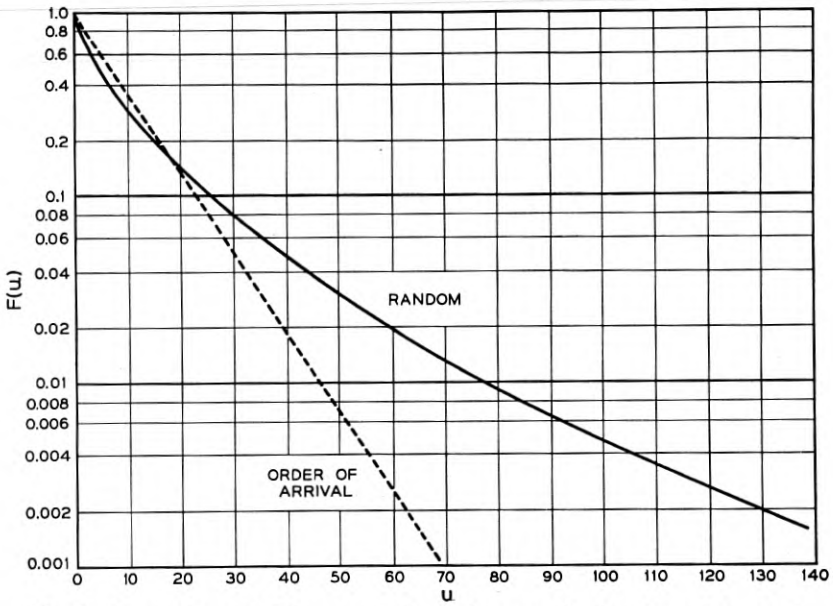

Fig. 2—Comparison of delay curves for order of arrival and random service; $\alpha = 0.9$.

service is increased (over order of arrival) but this is achieved at the cost of making other customers wait much longer.

Service in order of arrival has the advantage to the customer that his delay is independent of all who come after him, and this is particularly appreciated in times of heavy crowding when long delays are possible for random service. In Table I, these crowded conditions correspond to small values of $F(u)$ or large values of α, or both. In this picture it seems intuitively clear that much longer delays are possible for random service, for those unlucky customers who keep missing their turn. (Of course, a more realistic model would also include the effects of customers leaving before service, a factor of considerable telephone interest also.)

As noted at the start of this section, $F(u)$ is a conditional probability, the probability of delay at least u of a call that is surely delayed. To obtain unconditional probabilities of delay, $F(u)$ is multiplied by the probability that all trunks are busy, which is the probability that a call is delayed. This probability is given by a well-known formula due to Erlang and customarily written as

$$C(c, a) = \frac{a^c}{(c-1)\,!\,(c-a)}\left[1 + \frac{a}{1!} + \frac{a^2}{2!} + \cdots + \frac{a^{c-1}}{(c-1)!} + \frac{a^c}{(c-1)\,!\,(c-a)}\right]^{-1}$$

Tables of this function are available*.

Finally it may be noticed here that for random service and light traffic (roughly, α less than 0.7), with sufficient approximation

$$F(u) = \tfrac{1}{2}(y_1 e^{-u(1-\alpha)y_1} + y_2 e^{-u(1-\alpha)y_2})$$

with $y_1 = 1 - \sqrt{\alpha/2}$, $y_2 = 1 + \sqrt{\alpha/2}$.

* But there seems to be no extensive tabulation. However, the table for the Erlang B function made by Conny Palm (Stockholm, 1947) may be used with the relations

$$\frac{1}{C(c, a)} = \frac{1}{B(c, a)} - \frac{1}{B(c-1, a)}$$

$$= \frac{a}{c} + \frac{1 - (a/c)}{B(c, a)}$$

Notice that $C(c, a)$ also has the recurrence relation

$$\frac{1}{C(c, a)} = \frac{-1}{c - 1 - a} + \frac{(c-a)(c-1)}{a(c-1-a)C(c-1, a)}$$

3. BASIC FORMULATION

As noted above, the following notation is used: c is the number of trunks, h is the average holding time (the distribution of holding times is exponential) and a is the average number of calls arriving in time interval h. Then, if $F_n(t)$ is the probability of delay at least t of a call arriving when n other calls are waiting, the differential recurrence relation given by Vaulot is

$$\frac{dF_n(t)}{dt} = \frac{n}{n+1}\frac{c}{h}F_{n-1}(t) - \frac{c+a}{h}F_n(t) + \frac{a}{h}F_{n+1}(t) \qquad (1)$$

This may be derived as follows. Consider the interval dt after the epoch of arrival of the call in question. In this interval three events may occur: (i) a call may arrive, (ii) a trunk may be released, or (iii) neither of these. The probability of a call arrival is $(a/h)dt$ and if a call arrives the delay function is $F_{n+1}(t - dt)$. The probability of a trunk release, because of the assumption of exponential holding time, is $(c/h)dt$, and if a trunk is released the number of waiting calls is reduced by one; the probability that the call seizing the waiting trunk will not be the call in question is $n/(n+1)$. Finally the probability of the third event is $1 - (c + a)dt/h$. All this is summarized in the differential relation

$$F_n(t) = \frac{a}{h}dt\,F_{n+1}(t - dt) + \frac{n}{n+1}\frac{c}{h}dt\,F_{n-1}(t - dt) + \left(1 - \frac{c+a}{h}dt\right)F_n(t - dt)$$

Passing to the limit gives equation (1).

Using new variables: $u = ct/h$, $\alpha = a/c$, equation (1) may be written more simply as

$$\frac{dF_n(u)}{du} = \frac{n}{n+1}F_{n-1}(u) - (1+\alpha)F_n(u) + \alpha F_{n+1}(u) \qquad (1a)$$

This equation is a mixed differential-difference equation of the first order as a differential equation and of the second order as a difference equation; hence three boundary relations are required. For the differential part, it is clear that $F_n(0)$, which is the probability of some delay of the test call, is unity for all n in question, that is, for all integral non-negative n. Also $F_n(u) \equiv 0$ for all negative n, is an obvious necessity, and, since F_n is a distribution function $F_n(\infty) = 1$. Finally the third

condition may be stated as

$$\lim_{n=\infty} F_n(u) = 1, \qquad \text{all } u$$

The probability of delay at least u of an arbitrary call is the sum on n of the product of the probability that n calls are waiting when the call arrives and the probability, $F_n(u)$, that for this condition the call is delayed at least u. The first probability (for statistical equilibrium) is known to be

$$(1 - \alpha)\, C(c, a)\alpha^n$$

where $C(c, a)$, as stated above, is the probability that all trunks are busy; $(1 - \alpha)C(c, a)$ is the probability that all trunks are busy and no calls are waiting. Hence the probability in question, say $f(u)$, is given by

$$f(u) = (1 - \alpha)C(c, a) \sum_0^\infty \alpha^n F_n(u)$$

or by

$$f(u) = C(c, a)F(u)$$

if

$$F(u) = (1 - \alpha) \sum_0^\infty \alpha^n F_n(u) \tag{2}$$

$F(u)$, like $F_n(u)$, is then a conditional probability, the probability at least u of a delayed call. Notice that, consistent with this, $F(0) = 1$.

It is interesting to notice that Mellor's basic equation, which in present notation may be written as

$$\frac{dG_n(u)}{du} = -\frac{1}{n+1} G_n(u), \tag{3}$$

follows from (1) if first it is supposed that $F_{n-1}(u) = F_n(u) = F_{n+1}(u)$ and then, for clarity, G_n replaces F_n. Hence, as indicated by the third boundary condition, it may be expected to be useful for large values of n. Its solution is

$$G_n(u) = e^{-u/(n+1)} \tag{4}$$

A somewhat better approximation may be determined by the Mac-Laurin series obtained by repeated differentiation of (1a) and evaluation

at $u = 0$; this is as follows

$$F_n(u) \approx 1 - \frac{u}{n+1} + \frac{\alpha}{2}\frac{(u)^2}{(n+1)} - \frac{\alpha(2\alpha-1)}{3!}\frac{(u)^3}{(n+1)} + \frac{\alpha(2\alpha-1)(3\alpha-2)}{4!}\frac{(u)^4}{(n+1)} - \cdots \tag{5}$$

But this is the same* as:

$$F_n(u) \approx [1 - (1-\alpha)u/(n+1)]^{1/(1-\alpha)} \tag{5a}$$

As α approaches unity, (5a) approaches (4). Equation (5a) has been used, for large values of α, in the direct computations mentioned above.

It may also be noted that for $\alpha = 0$, equation (1a) has the solution (now writing $F_n(u, \alpha)$ for $F_n(u)$)

$$F_n(u, 0) = \phi(u, n) - \frac{u}{n+1}\phi(u, n-1) \tag{6}$$

where $\phi(u, n)$ is the Poisson sum

$$e^{-u}\left(1 + u + \frac{u^2}{2!} + \cdots + \frac{u^n}{n!}\right)$$

Finally, for completeness, note that for small values of u, the Mac-Laurin series for $F(u)$ is

$$F(u) = 1 - u\frac{1-\alpha}{\alpha}\log\frac{1}{1-\alpha} + \frac{u^2}{2}(1-\alpha)\left[2 - \frac{1-\alpha}{\alpha}\log\frac{1}{1-\alpha}\right] - \frac{u^3}{6}(1-\alpha)\left[1 + 3\alpha - (1-\alpha)\log\frac{1}{1-\alpha} - \sum_1^\infty \frac{\alpha^n}{n^2}\right] \tag{7}$$

4. MOMENTS

The k'th moment (about the origin) of the delay density function which is $-F'(u)$ ($F(u)$ itself is a distribution function is defined as

$$M_k = \int_0 u^k[-F'(u)]\,du, \qquad = k\int_0 u^{k-1}F(u)\,du, \quad k > 0, \tag{8}$$

the last by integration by parts.

* G. W. Abrams is due credit for noticing this.

Following (2), this may also be written as

$$M_k = (1 - \alpha) \sum_0^\infty \alpha^n m_{n,k}, \tag{9}$$

with

$$\begin{aligned} m_{n,k} &= \int_0^\infty u^k[-F'_n(u)]\, du, \\ &= k \int_0^\infty u^{k-1} F_n(u)\, du, \qquad k > 0. \end{aligned} \tag{10}$$

First, notice that

$$m_{n,0} = -\int_0^\infty F'_n(u)\, du = F_n(0) = 1;$$

hence

$$M_0 = (1 - \alpha) \sum_0^\infty \alpha^n = 1,$$

showing that $F(u)$ is properly normalized.

Next, by integrating both sides of (1a) with respect to u from 0 to ∞, and using the second form of (10) (with $k = 1$)

$$-(n + 1) = n m_{n-1,1} - (n + 1)(1 + \alpha) m_{n,1} + (n + 1)\alpha m_{n+1,1} \tag{11}$$

In the same way, after first multiplying (1a) throughout by u^{k-1}, it is found that

$$\begin{aligned} &-k(n + 1) m_{n,k-1} \\ &\qquad = n m_{n-1,k} - (n + 1)(1 + \alpha) m_{n,k} + (n + 1)\alpha m_{n+1,k} \end{aligned} \tag{12}$$

Unfortunately, neither (11) nor any other instances of (12) have simple solutions; nevertheless they may be used to determine M_k.

Consider first the simplest case, M_1. If (11) is multiplied throughout by α^n and summed on n, the result may be written

$$\begin{aligned} -L_{10} &= \alpha L_{11} - (1 + \alpha) L_{11} + L_{11} - L_{01} \\ &= -L_{01} \end{aligned} \tag{13}$$

where for convenience in writing and of later notation

$$\begin{aligned} L_{01} &= \sum \alpha^n m_{n,1} = (1 - \alpha)^{-1} M_1 \\ L_{11} &= \sum (n + 1)\alpha^n m_{n,1} \\ L_{10} &= \sum (n + 1)\alpha^n = D \sum \alpha^{n+1} = (1 - \alpha)^{-2} \end{aligned}$$

and $D = d/d\alpha$. Hence

$$M_1 = (1 - \alpha)^{-1}$$

This is the mean delay of calls delayed and as mentioned above is the same as for service in order of arrival.

In the general case*, the following notation is convenient

$$L_{0k} = \sum \alpha^n m_{n,k} = (1 - \alpha)^{-1} M_k$$
$$L_{jk} = \sum (n + 1)(n + 2) \cdots (n + j)\alpha^n m_{n,k}$$

Using the relations

$$\begin{aligned} n(n + 2) \cdots (n + j) &= n(n + 1) \cdots (n + j - 1) + (j - 1)n(n + 1) \cdots (n + j - 2) \\ &+ \cdots + (j - 1)_i n(n + 1) \cdots (n + j - i - 1) + \cdots \\ &+ (j - 1)!n, \end{aligned}$$

$$\begin{aligned} n(n + 1) \cdots (n + j - 1) &= (n + 1)(n + 2) \cdots (n + j) - j(n + 1)(n + 2)(n + j - 1) \end{aligned}$$

with

$$(j - 1)_i = (j - 1)(j - 2) \cdots (j - i),$$

the summing of (12) is found to result in

$$\begin{aligned} kL_{j,k-1} &= [j - (j - 1)\alpha]L_{j-1,k} - \alpha[(j - 1)_2 L_{j-2,k} \\ &+ (j - 1)_3 L_{j-3,k} + \cdots + (j - 1)_i L_{j-i,k} + \cdots + (j - 1)!L_{1,k}] \end{aligned} \tag{14}$$

But this may be simplified by multiplying through by j and subtracting from the same equation with j replaced by $j + 1$; the result is

$$(j + 1 - j\alpha)L_{jk} - j^2 L_{j-1,k} = kL_{j+1,k-1} - jkL_{j,k-1} \tag{15}$$

Notice that for $j = 0$, $k = 1$, $L_{01} = L_{10}$, as in (13). Notice also that

$$\begin{aligned} \alpha^{j-1} L_{j0} &= \sum (n + 1) \cdots (n + j)\alpha^{n+j-1} \\ &= D \sum (n + 1) \cdots (n + j - 1)\alpha^{n+j} \\ &= D(\alpha^j L_{j-1,0}) \end{aligned}$$

so that

$$L_{j0} = jL_{j-1,0} + \alpha D L_{j-1,0} = j!(1 - \alpha)^{-j-1}$$

* This procedure is the development of a suggestion made by S. O. Rice.

Then the ratio

$$\begin{aligned} L_{0k}(L_{k0})^{-1} &= (1-\alpha)^{k+1}L_{0k}/k! \\ &= (1-\alpha)^k M_k/k! = R_k \end{aligned} \tag{16}$$

is the ratio of these moments to those for order of arrival service; the last relation is a definition. In the same way the ratio

$$L_{jk}(L_{j+k,0})^{-1}$$

might be considered, but to avoid fractions the following somewhat odd change of variables seems convenient:

$$\binom{j+k}{k} g_k L_{jk} = p_{jk} L_{j+k,0} \tag{17}$$

where $g_0 = g_1 = 1$ and

$$\begin{aligned} g_{2k} &= (2-\alpha)^{2k-1}(3-2\alpha)^{2k-3}\cdots(k+1-k\alpha) \\ g_{2k+1} &= (2-\alpha)^{2k}(3-2\alpha)^{2k-2}\cdots(k+1-k\alpha)^2 \end{aligned}$$

Notice that

$$g_{2k+1}(g_{2k})^{-1} = g_{2k}(g_{2k-1})^{-1} = (2-\alpha)(3-2\alpha)\cdots(k+1-k\alpha) = D_{k+1}$$

the last being a definition, again.

Since

$$(1-\alpha)L_{k,0} = kL_{k-1,0}$$

it follows from (15) that

$$\begin{aligned} &(j+1-j\alpha)p_{jk} - j(1-\alpha)p_{j-1,k} \\ &\qquad = (g_k/g_{k-1})[(j+1)p_{j+1,k-1} - j(1-\alpha)p_{j,k-1}] \end{aligned} \tag{18}$$

By taking differences of this equation and writing

$$\begin{aligned} q_{0k} &= p_{0k} \\ q_{1k} &= p_{1k} - p_{0k} = \Delta p_{0k} \\ q_{2k} &= p_{2k} - 2p_{1k} + p_{0k} = \Delta^2 p_{0k} \\ q_{jk} &= \Delta q_{j-1,k} = \Delta^j p_{0k} \end{aligned}$$

a somewhat simpler recurrence relation is found to be as follows

$$\begin{aligned} (j+1-j\alpha)q_{jk} &= (g_k/g_{k-1}) \\ &\quad [j\alpha q_{j-1,k-1} + (j+1+j\alpha)q_{j,k-1} + (j+1)q_{j+1,k-1}] \end{aligned} \tag{19}$$

Since $p_{j0} = 1$, all j, $q_{00} = 1$, and $q_{j0} = 0$, $j \neq 0$. From these boundary conditions, it follows at once from (19) that

$$q_{jk} = 0, \qquad j > k.$$

By comparison of (17) and (16)

$$g_k R_k = p_{0k} = q_{0k}$$

A short table of the q's is as follows:

j/k	0	1	2	3
0	1	1	2	$2(2 + \alpha)$
1	0	$\alpha(2 - \alpha)^{-1}$	$4\alpha(2 - \alpha)^{-1}$	$2\alpha(18 - 5\alpha - 4\alpha^2)D_{\frac{3}{4}}^{-1}$
2	0	0	$2\alpha^2(3 - 2\alpha)^{-1}$	$2\alpha^2(18 - 7\alpha - 2\alpha^2)(3 - 2\alpha)^{-2}$
3	0	0	0	$6\alpha^3 D_{\frac{2}{4}}^2 D_{\frac{1}{4}}^{-1}$

Continuation of this leads to the values of R_k listed in Table II. Notice that for $\alpha = 1$, by (18)

$$\begin{aligned} p_{jk}(1) &= (j + 1)p_{j+1,k-1}(1) \\ &= (j + 1)(j + 2)p_{j+2,k-2}(1) \\ &= (j + 1)(j + 2) \cdots (j + k) \end{aligned}$$

since $g_k = 1$ for $\alpha = 1$ and $p_{j0} = 1$, all j. From this

$$p_{0k}(1) = g_k(1)R_k(1) = R_k(1) = k!$$

On the other hand, for $\alpha = 0$, $q_{jk} = 0$, $j > 0$ and, by (19)

$$q_{0k}(0)/g_k(0) = q_{0,k-1}(0)/g_{k-1}(0)$$

so that

$$R_k(0) = R_{k-1}(0) = R_1(0) = 1$$

5. MELLOR APPROXIMATION

It is useful to have the moments of the distribution corresponding to the Mellor approximation, since they serve as a guide. Here, following equation (4)

$$F(u) = (1 - \alpha) \sum_0^\infty \alpha^n e^{-u(n+1)^{-1}} \tag{20}$$

and

$$\exp - xM = \sum_0^\infty M_k(-x)^k/k!$$

$$= (1 - \alpha) \int_0^\infty du\, e^{-xu} \sum_0^\infty (n + 1)^{-1}\alpha^n e^{-u(n+1)^{-1}} \quad (21)$$

$$= (1 - \alpha) \sum_0^\infty \alpha^n[1 + x(n + 1)]^{-1}$$

Hence

$$M_k = k!(1 - \alpha) \sum_0^\infty (n + 1)^k\alpha^n \quad (22)$$

These moments are expressible in terms of polynomials associated with the distribution of permutations into classes according to the number of readings left to right necessary to find the elements in standard order.[6] Indeed the ratio

$$r_k(\alpha) = M_k\,(1 - \alpha)^k/k!$$

has the recurrence relation

$$r_{k+1}(\alpha) = (k\alpha + 1)r_k(\alpha) + \alpha(1 - \alpha)r'_k(\alpha) \quad (23)$$

and the first few values are as follows

$$r_1 = 1 \qquad r_3 = 1 + 4\alpha + \alpha^2$$

$$r_2 = 1 + \alpha \qquad r_4 = 1 + 11\alpha + 11\alpha^2 + \alpha^3$$

$$r_5 = 1 + 26\alpha + 66\alpha^2 + 26\alpha^3 + \alpha^4$$

Notice that $r_k(0) = 1$, $r_k(1) = k!$, just as for the precise results.

6. EXPONENTIAL SUMS

The shape of the delay curves, from direct calculation, and also from Mellor's results, suggests representation in exponential sums. If

$$F(u) = A_1\, e^{-(1-\alpha)u/x_1} + A_2\, e^{-(1-\alpha)u/x_2} + \cdots \quad (24)$$

then

$$M_k \frac{(1 - \alpha)^k}{k!} = A_1x_1^k + A_2x_2^k + \cdots \quad (25)$$

by a simple calculation. For k exponentials, $2k$ moments (including

M_0) may be fitted exactly by solution of $2k$ equations of form (25), as will be shown.

The first approximation ($k = 1$) is the order of arrival curve, say

$$F_1(u) = e^{-(1-\alpha)u}$$

which has $A_1 = x_1 = 1$, $A_k = x_k = 0$, $k > 1$, and matches M_0 and M_1.

The next approximation ($k = 2$) is determined from equations

$$\begin{aligned} A_1 + A_2 &= 1 \\ A_1x_1 + A_2x_2 &= 1 \\ A_1x_1^2 + A_2x_2^2 &= R_2 \\ A_1x_1^3 + A_2x_2^3 &= R_3 \end{aligned}$$

Eliminating A_2 from successive pairs,

$$\begin{aligned} A_1(x_1 - x_2) &= 1 - x_2 \\ A_1x_1(x_1 - x_2) &= R_2 - x_2 \\ A_1x_1^2(x_1 - x_2) &= R_3 - R_2x_2 \end{aligned}$$

Eliminating A_1 from these,

$$\begin{aligned} x_1 + x_2 - x_1x_2 &= R_2 \\ (x_1 + x_2)R_2 - x_1x_2 &= R_3 \end{aligned} \tag{26}$$

or, writing $a_1 = x_1 + x_2$, $a_2 = x_1x_2$, so that $x^2 - a_1x + a_2 = (x - x_1)(x - x_2)$

$$\begin{aligned} a_1 - a_2 &= R_2 \\ a_1R_2 - a_2 &= R_3 \end{aligned} \tag{26a}$$

From the first of the second set of equations, and from symmetry (or from $A_1 + A_2 = 1$)

$$\begin{aligned} A_1 &= \frac{1 - x_2}{x_1 - x_2} \\ A_2 &= \frac{1 - x_1}{x_2 - x_1} \end{aligned} \tag{27}$$

Taking R_2 and R_3 from Table II, it turns out that

$$\begin{aligned} x_1^{-1} &= 1 - \sqrt{\alpha/2} = 2A_1 \\ x_2^{-1} &= 1 + \sqrt{\alpha/2} = 2A_2 \end{aligned} \tag{28}$$

TABLE II – MOMENT RATIOS, CALLS SERVED AT RANDOM

$$R_k(\alpha) = M_k(1-\alpha)^k/k!$$

$$R_1 = 1$$

$$R_2 = \frac{2}{2-\alpha}$$

$$R_3 = \frac{2(2+\alpha)}{(2-\alpha)^2}$$

$$R_4 = \frac{4(6+5\alpha-4\alpha^2-\alpha^3)}{(2-\alpha)^3(3-2\alpha)}$$

$$R_5 = \frac{4(36+60\alpha-59\alpha^2-24\alpha^3+15\alpha^4+2\alpha^5)}{(2-\alpha)^4(3-2\alpha)^2}$$

$$R_6 = \frac{8f_6(\alpha)}{(2-\alpha)^5(3-2\alpha)^3(4-3\alpha)}$$

$$R_7 = \frac{8f_7(\alpha)}{(2-\alpha)^6(3-2\alpha)^4(4-3\alpha)^2}$$

$$f_6(\alpha) = 432+972\alpha-2016\alpha^2-437\alpha^3+1790\alpha^4-528\alpha^5-196\alpha^6+67\alpha^7+6\alpha^8$$

$$f_7(\alpha) = 10368+34560\alpha-89208\alpha^2-32772\alpha^3+177926\alpha^4-104287\alpha^5-29260\alpha^6+43876\alpha^7-9158\alpha^8-2039\alpha^9+588\alpha^{10}+36\alpha^{11}$$

and the second approximation is

$$2F_2(u) = (1-\sqrt{\alpha/2})\, e^{-u(1-\alpha)(1-\sqrt{\alpha/2})} + (1+\sqrt{\alpha/2})\, e^{-u(1-\alpha)(1+\sqrt{\alpha/2})} \tag{29}$$

which turns out to be a good fit for α roughly less than 0.7. Curiously the corresponding Mellor approximation has a more complicated expression.

Following the same procedure for three exponentials, it turns out that the correspondent to the set of equations (26a) is

$$\begin{aligned} a_1R_2 - a_2 + a_3 &= R_3 \\ a_1R_3 - a_2R_2 + a_3 &= R_4 \\ a_1R_4 - a_2R_3 + a_3R_2 &= R_5 \end{aligned} \tag{30}$$

with $a_1 = x_1 + x_2 + x_3$, $a_2 = x_1x_2 + x_1x_3 + x_2x_3$, $a_3 = x_1x_2x_3$, that is, the symmetric functions.

Using Table II for values of the R's, it is found that

$$\begin{aligned} a_1 &= (18 - 7\alpha - 2\alpha^2)(2 - \alpha)^{-1}(3 - 2\alpha)^{-1} \\ a_2 &= 18 \qquad\qquad (2 - \alpha)^{-1}(3 - 2\alpha)^{-1} \\ a_3 &= 6 \qquad\qquad (2 - \alpha)^{-1}(3 - 2\alpha)^{-1} \end{aligned} \tag{31}$$

x_1, x_2 and x_3 are then the roots of the cubic equation

$$x^3 - a_1x^2 + a_2x - a_3 = 0$$

The coefficients A_i, $i = 1, 2, 3$ are determined from equations like

$$A_1 = \frac{R_2 - (x_2 + x_3) + x_2x_3}{(x_1 - x_2)(x_1 - x_3)} \tag{32}$$

For the fourth approximation, matching 8 moments, the equations for the symmetric functions are

$$\begin{aligned} a_1R_3 - a_2R_2 + a_3 - a_4 &= R_4 \\ a_1R_4 - a_2R_3 + a_3R_2 - a_4 &= R_5 \\ a_1R_5 - a_2R_4 + a_3R_3 - a_4R_2 &= R_6 \\ a_1R_6 - a_2R_5 + a_3R_4 - a_4R_3 &= R_7 \end{aligned} \tag{33}$$

and x_1, x_2, x_3 and x_4 are roots of the quartic equation

$$x^4 - a_1x^3 + a_2x^2 - a_3x + a_4 = 0$$

Coefficients A_i are determined from equations like

$$A_1 = \frac{R_3 - (x_2 + x_3 + x_4)R_2 + (x_2x_3 + x_2x_4 + x_3x_4) - x_2x_3x_4}{(x_1 - x_2)(x_1 - x_3)(x_1 - x_4)} \tag{34}$$

It may be noted that

$$\begin{aligned} x_2 + x_3 + x_4 &= a_1 - x_1 \\ x_2x_3 + x_2x_4 + x_3x_4 &= a_2 - x_1(a_1 - x_1) \\ x_2x_3x_4 &= a_3 - x_1[a_2 - x_1(a - x_1)] = a_4x_1^{-1} \end{aligned}$$

which gives the general structure.

It is worth noting that equations (33) may be used to determine the R's if the a's may be determined otherwise. As a matter of fact, they have led to the determination of R_6 and R_7 in the following way. The

results for $k = 2$ and 3 suggest that

$$a_4 = 4!(2 - \alpha)^{-1}(3 - 2\alpha)^{-1}(4 - 3\alpha)^{-1}$$

$$a_3 = 4a_4$$

Then by the first two of equations (33)

$$a_1R_3 - a_2R_2 = R_4 - a_3 + a_4$$

$$a_1R_4 - a_2R_3 = R_5 - a_3R_2 + a_4$$

the solutions of which are

$$a_1 = 4(24 - 23\alpha + 3\alpha^3) \qquad [(2 - \alpha)(3 - 2\alpha)(4 - 3\alpha)]^{-1}$$

$$a_2 = 2(72 - 23\alpha - 10\alpha^2 - 3\alpha^3)[(2 - \alpha)(3 - 2\alpha)(4 - 3\alpha)]^{-1}$$

By the last two of equations (33), R_6 and R_7 are determined to be the values given in Table II, which have been verified independently. Note that for $\alpha = 0$, both R_6 and R_7 are 1, and for $\alpha = 1$, $R_6 = 6!$, $R_7 = 7!$

Table III tabulates, for $k = 2$ to 5, for convenience in avoiding fractions the symmetric functions b_{kj} related to those above by

$$b_{kj} = D_k a_j$$

with, as before,

$$D_k = (2 - \alpha)(3 - 2\alpha) \cdots [k - (k - 1)\alpha]$$

and $a_0 = 1$. The functions for $k = 5$ were obtained by a process like

TABLE III — SYMMETRIC FUNCTIONS FOR EXPONENTIAL SUMS OF CALLS SERVED AT RANDOM

$k = 2$	$b_{20} = 2 - \alpha$	$k = 3$	$b_{30} = 6 - 7\alpha + 2\alpha^2$
	$b_{21} = 4$		$b_{31} = 18 - 7\alpha - 2\alpha^2$
	$b_{22} = 2$		$b_{32} = 18$
			$b_{33} = 6$

$k = 4$	$b_{40} = 24$	$- 46\alpha$	$+ 29\alpha^2$	$- 6\alpha^3$	
	$b_{41} = 96$	$- 92\alpha$		$+ 12\alpha^3$	
	$b_{42} = 144$	$- 46\alpha$	$- 20\alpha^2$	$- 6\alpha^3$	
	$b_{43} = 96$				
	$b_{44} = 24$				

$k = 5$	$b_{50} = 120$	$- 326\alpha$	$+ 329\alpha^2$	$- 146\alpha^3$	$+ 24\alpha^4$	
	$b_{51} = 600$	$- 978\alpha$	$+ 329\alpha^2$	$+ 146\alpha^3$	$- 72\alpha^4$	
	$b_{52} = 1200$	$- 978\alpha$	$- 172\alpha^2$	$+ 78\alpha^3$	$+ 72\alpha^4$	
	$b_{53} = 1200$	$- 326\alpha$	$- 172\alpha^2$	$- 78\alpha^3$	$- 24\alpha^4$	
	$b_{54} = 600$					
	$b_{55} = 120$					

TABLE IV — SYMMETRIC FUNCTIONS FOR EXPONENTIAL SUMS, MELLOR APPROXIMATION

$k = 2$ $a_1 = 3+\alpha$, $a_2 = 2$ $\quad$ $k = 3$ $a_1 = 6 + 3\alpha$, $a_2 = 11 + 5\alpha + 2\alpha^2$, $a_3 = 6$

$k = 4$
$a_1 = 10 + 6\alpha$
$a_2 = 35 + 26\alpha + 11\alpha^2$
$a_3 = 50 + 26\alpha + 14\alpha^2 + 6\alpha^3$
$a_4 = 24$

$k = 5$
$a_1 = 15 + 10\alpha$
$a_2 = 85 + 80\alpha + 35\alpha^2$
$a_3 = 225 + 200\alpha + 125\alpha^2 + 50\alpha^3$
$a_4 = 274 + 154\alpha + 94\alpha^2 + 54\alpha^3 + 24\alpha^4$
$a_5 = 120$

$k = 6$
$a_1 = 21 + 15\alpha$
$a_2 = 175 + 190\alpha + 85\alpha^2$
$a_3 = 735 + 855\alpha + 585\alpha^2 + 225\alpha^3$
$a_4 = 1624 + 1604\alpha + 1194\alpha^2 + 704\alpha^3 + 274\alpha^4$
$a_5 = 1764 + 1044\alpha + 684\alpha^2 + 444\alpha^3 + 264\alpha^4 + 120\alpha^5$
$a_6 = 720$

that sketched above, and without determining R_8 and R_9. Notice that

$$b_{kj} = k!\binom{k}{j}, \qquad \alpha = 0$$

$$= j!\binom{k}{j}^2, \qquad \alpha = 1$$

which may be proved independently. All values in Table III satisfy the recurrence relation

$$b_{kj} = [k - (k-1)\alpha]b_{k-1,j} + [k + (k-1)\alpha]b_{k-1,j-1} - (k-1)^2\alpha b_{k-2,j-2} \tag{35}$$

which also satisfies the boundary relations for $\alpha = 0$ and 1 given above for all values of k.

The corresponding symmetric functions for the Mellor approximation are given in Table IV. These have the recurrence relation

$$a_{kj} = a_{k-1,j} + [k + (k-1)\alpha]a_{k-1,j-1} - (k-1)^2\alpha a_{k-2,j-2} \tag{36}$$

For $\alpha = 0$, the values are the signless Stirling numbers of the first kind, that is, the numbers given by the expansion of

$$(1 + x)(1 + 2x) \cdots (1 + kx).$$

For $\alpha = 1$, the results are the same as for the exact case, as given above.

TABLE V — APPROXIMATIONS TO DELAY FUNCTION $F(u)$ FOR RANDOM SERVICE

α	v							
	1	2	4	6	8	10	12	14
Two Exponentials								
0.1	.3590	.1351	.0220	.0041	.0008	.0002		
0.2	.3490	.1344	.0256	.0058	.0015	.0004	.0001	
0.3	.3392	.1332	.0291	.0079	.0023	.0007	.0002	.0001
0.4	.3292	.1315	.0325	.0101	.0033	.0011	.0004	.0002
0.5	.3190	.1293	.0357	.0125	.0046	.0017	.0006	.0003
0.6	.3085	.1265	.0386	.0151	.0061	.0025	.0010	.0004
0.7	.2978	.1232	.0412	.0177	.0078	.0035	.0015	.0007
0.8	.2868	.1193	.0434	.0203	.0097	.0047	.0022	.0011
0.9	.2756	.1148	.0451	.0229	.0118	.0061	.0031	.0016
Three Exponentials								
0.1	.3586	.1354	.0219	.0040	.0008	.0002		
0.2	.3491	.1356	.0254	.0057	.0014	.0004	.0001	
0.3	.3393	.1358	.0288	.0074	.0022	.0007	.0002	.0001
0.4	.3291	.1360	.0322	.0092	.0030	.0011	.0004	.0002
0.5	.3186	.1363	.0358	.0112	.0040	.0016	.0007	.0003
0.6	.3071	.1359	.0392	.0133	.0050	.0022	.0010	.0005
0.7	.2951	.1354	.0428	.0156	.0063	.0028	.0014	.0007
0.8	.2822	.1344	.0466	.0181	.0077	.0036	.0018	.0010
0.9	.2683	.1325	.0504	.0210	.0094	.0045	.0023	.0013

7. NUMERICAL RESULTS

Table V gives both two-exponential and three-exponential 4 decimal approximations to the delay function $F(u)$ for

$$\alpha = 0.1(0.1)0.9(0.1 \text{ to } 0.9 \text{ in steps of } 0.1)$$

and for

$$u(1 - \alpha) = 1(1)2(2)14,$$

in the same abbreviated notation.* The variable $v = u(1 - \alpha)$ is introduced to reduce the spread of these tables. It will be noticed that, as expected, the two orders of approximation agree closely for small values of α; indeed, only for the three largest values of α are the differences appreciable from the engineering standpoint.

* The results for two exponentials, some of those for three-exponentials, and all special results given below, have been obtained by Miss Marian Darville, whom I also thank for her careful drawing of the curves. The entire three-exponential table has been computed independently by Miss Lennon.

For $\alpha = 0.9$, results for four exponentials have also been obtained and compare with those of Table V as follows (k = number of exponentials):

k	v							
	1	2	4	6	8	10	12	14
2	.2756	.1148	.0451	.0229	.0118	.0061	.0037	.0016
3	.2683	.1325	.0504	.0210	.0094	.0045	.0023	.0013
4	.2748	.1402	.0483	.0195	.0091	.0047	.0026	.0015

It is somewhat surprising that two exponentials should do as well as they do for large values of v (in fact for $v = 12$ and 14 better than three); a similar behavior appears in the following comparison of approximations on the Mellor basis, again for $\alpha = 0.9$

k	v							
	1	2	4	6	8	10	12	14
2	.2725	.1115	.0446	.0237	.0129	.0070	.0038	.0021
4	.2671	.1379	.0502	.0207	.0097	.0051	.0029	.0018
6	.2777	.1408	.0477	.0205	.0102	.0054	.0031	.0018

From these comparisons, it appears a relatively small number of exponentials is sufficient for engineering purposes. The curves of Fig. 1 are those for three exponentials, for uniformity.

BIBLIOGRAPHY

1. Erlang, A. K., Løsning af nogle Problemer fra Sandsynlighedsregningen af Betydning for de automatiske Telefoncentraler. Elektroteknikeren **13,** p. 5, 1917; *The Life and Works of A. K. Erlang.* Copenhagen, pp. 138–155, 1948.
2. Molina, E. C., Application of the Theory of Probabilities to Telephone Trunking Problems, Bell System Tech. Jl., **6,** pp. 461–494, 1927.
3. Mellor, J. W., Delayed Call Formulae when Calls Are Served in a Random Order. P.O.E.E.J. **25,** pp. 53–56, 1942.
4. Vaulot, E., Delais d'attente des appels téléphoniques traités an hazard. Comptes Rend. Acad. Sci. Paris **222,** pp. 268–269, 1946.
5. Pollaczek, F., La loi d'attente des appels téléphoniques, Comptes Rend. Acad. Sci. Paris **222,** pp. 353–355, 1946.
6 Riordan, J., Triangular permutation numbers. Am. Math. Soc., Proc., **2,** pp. 429–432, 1951.

The Evaluation of Wood Preservatives Part I

Interpretation and Correlation of the Results of Laboratory Soil-Block Tests and Outdoor Test Plot Experience, with Special Reference to Oil-Type Materials

By REGINALD H. COLLEY

(Manuscript received September 22, 1952)

This paper offers a review and interpretation of laboratory and field experiments aimed at determining the necessary protective threshold quantities of wood preservatives. It details the procedure followed in the soil-block tests at the Bell Telephone Laboratories, Incorporated. Discussion of specific criticisms of the techniques involved and replies to these criticisms are included. The paper also presents for the first time a correlation of the results obtained from soil-block culture tests, outdoor exposure tests on stakes and on pole-diameter posts as well as pole line experience. It demonstrates that the same levels for toxicity-permanence requirements (thresholds) are obtained from the three different types of accelerated experimental evaluations. There is every reason to believe that the same limits apply for the outer inch of sapwood in pine poles in line.

TABLE OF CONTENTS

SUBJECTS TO BE COVERED IN PART II

INTRODUCTION

In discussing the problems involved in the evaluation of wood preservatives over the years, it has generally been found necessary to orient the audience — in this case the readers of this JOURNAL — in the field of biology, and particularly in the field of biological tests involving wood-

destroying fungi. It is impractical to expect from such tests the degree of accuracy in results that one would look for as a matter of course in certain types of well conducted physical or chemical experiments. One can, however, look for high reliability in the biological sense. In the half science, half art of wood preservation there is as yet no generally acceptable laboratory technique for measuring the preservative value of a given material. Although much development work has been done, both here and abroad, in an effort to promote standard laboratory procedures, their proponents have had very little success in bringing into line the techniques used in the various areas. The interest of the Bell System in establishing a standard bioassay test will become convincingly evident as this story unfolds.

When the first American telephone lines were built there was an adequate supply of naturally durable pole timber in northern cedar and chestnut forests. The chestnut trees have been killed by a fungus disease, the chestnut blight; and the chestnut supply failed completely about twenty years ago. Northern cedar trees are not straight enough nor large enough nor plentiful enough to meet the demands of the power and communication utilities, but they are still used to some extent in the Lake States area. Usually they are incised at the ground line by toothed machines; and they are then given a preservative treatment with creosote or with pentachlorophenol in petroleum to prolong the life of the butt and ground line section.

In the northern and western states the increasing demands for poles 35 feet and longer brought in western red cedar, a straight and nearly perfectly shaped pole tree. The present Bell System use of the species is relatively small, about 4 per cent of the total annual production. Butt treatment of western red as well as northern cedar began in earnest about thirty years ago. This procedure protects the ground section. Many western cedars are now full length treated because, although the species is durable, the tops and sapwood layers are subject to infection and decay, sometimes after a relatively short service life.

In the South and Southeast the great favorite is naturally the southern pine pole, full length pressure-treated with creosote. Such poles made their way in the Bell System as far north as Memphis and Washington by the turn of the century. Their use increased rapidly after World War I, and they moved into virtually all parts of the country. They now make up about 73 per cent of the telephone pole plant. New treatment procedures for southern pine employing pentachlorophenol in petroleum applied by pressure processes are now under way at a number of plants.

Pressure-treated Douglas fir and butt-treated western red cedar dom-

inate other species on the West Coast and in the Pacific Northwest, while pressure and non-pressure treated lodgepole pine poles are favored in the Mountain States area. Pressure-treated jack pine and ponderosa pine move into telephone plant in small quantities in the Lake States and in the California areas, respectively.

To render telephone service the Bell System has some 20,000,000 wood poles carrying its wires and cables. Many of these poles are used jointly with the power companies. Since poles of the joint use sizes are not available in sufficient quantities in the southern pine forest to meet all the demands of the utilities all of the time it is inevitable that western cedar, Douglas fir, lodgepole pine, red pine and western larch should move into various parts of the System, either for the direct and sole use of the Operating Companies or for joint use.

The pole plant is continually changing. Pole species from the Northwest vary greatly in their treatability and they are generally harder to treat than southern pine. It is not possible to use traditional creosote pressure treatments for some of these species without running the risk of objectionable exudation, or bleeding, of the creosote.

The development of practical specifications for the application of new preservatives such as pentachlorophenol and greensalt, as well as the various types of creosote, to all of the pole species now used in Bell System plant calls for setting as exactly as possible necessary protective quantities of the various preservative materials. This is particularly true in view of the fact that for normal telephone use as well as for joint use it is absolutely essential to deliver to the Operating Companies poles that are clean and satisfactory for use in all types of telephone lines, without compromising on the question of adequate physical life for the treated units. This purpose is back of the Laboratories' efforts to develop bioassay tests that come as close as practicable to measuring the necessary protective amount of any given preservative, and to predicting its relative permanence in poles and crossarms in plant.

It has been pointed out in earlier papers[30, 31, 76] that Bell Laboratories' concept of preservative evaluation involves (a) laboratory evaluation tests, (b) test plot experiments with small stakes, (c) similar tests of pole size specimens, and (d) test lines selected for long time observation. The latter are chosen with the cooperation of the Operating Companies. Lumsden[76] has recently presented a summary of a quarter century of experience with pole-diameter posts in one test plot located at Gulfport, Mississippi. The principal aims of the present paper are to interpret the results of various laboratory methods of preservative evaluation, and to

indicate how these results may possibly be correlated with test plot and field experience.

A SHORT HISTORY OF THE DEVELOPMENT OF LABORATORY EVALUATION PROCEDURES

The practice of laboratory evaluation of wood preservatives developed along different lines in Europe and in the United States. Here the Petri dish method was the early favorite.[61, 100] The basic scheme of this test is to use agar culture media containing gradient concentrations of the preservative material to be tested, and to employ various easily grown test fungi as indicators of inhibiting or lethal doses. The same scheme is employed with stoppered Erlenmeyer flasks. The fungus now known as Madison 517, formerly referred to as *Fomes annosus*, has been used most frequently as the standard test organism although other fungi were also used.[101]

In the culture phase of the European standard agar-block method[33] the test fungi are grown in Kolle flasks on a malt agar medium. The impregnated wood test blocks are supported on glass "benches" just above the surface of the agar and the growing test fungus. Wood pulp or paper boards saturated with malt extract are used by some investigators[9, 54 (1)] in place of the agar medium alone. Generally an untreated block and a treated block are placed together in the same flask.

The concept of a test for wood preservatives that motivated the proponents of the German agar-block method was broad enough to include selection of the test blocks and test fungi, treatment and handling procedures except weathering tests, culture technique, determination of the protection boundary, and directions for reporting the results. Differences in the behavior of water solutions of single chemical compounds such as sodium fluoride, and of volatile oily preservatives such as the creosotes were recognized; and provision was made for dealing with both types of materials.

The formalizing of both the Petri dish agar method in the United States and of the agar-block method in Europe developed as a result of conferences called by Dr. Hermann von Schrenk, the first in St. Louis in 1929, and the second in Berlin in 1930. The Laboratories' representatives at the St. Louis conference were the writer and R. E. Waterman. The action taken at St. Louis was published by Schmitz in 1930.[100]

In a previous paper[99] about a year earlier, Schmitz had discussed various laboratory test procedures, and had offered an "improvement" in the Petri dish technique based on the idea of preventing evaporation of volatile materials. Some of his statements at that time now seem by

hindsight to have something of the character of a judgment before trial; but their bearing on the questions under discussion and their possible effect in retarding the development of more realistic methods appear to be important enough to warrant quoting at this time. For example, with special reference to Petri dish agar tests he says:

"The determination of the toxicity of relatively volatile substances, such as coal tar creosote, is particularly difficult, owing to the control of the loss of preservative during the sterilization process. In order to prevent this loss, it is proposed to place the preservative in small sealed glass ampules, which are later broken to liberate the preservative to form preservative-agar mixtures of any desired concentration."

He considers laboratory tests of toxicity of preservatives to have little or no application in commercial practice, and his opinions are definitely stated as follows:

"Toxicity studies deal only with the poisonous properties of a wood preservative, and therefore they do not give a complete picture of the value of any particular substance as a wood preservative. . . .

"*For commercial work, however, it is of interest to know the amount of material that must be initially injected into the wood to maintain the desired amount of preservative for a definite period of time.* (Author's italics). Laboratory studies of the toxicity of wood preservatives do not give this information. Attempts to calculate the amount of material which must be injected into the wood from laboratory studies of toxicity are, therefore, based upon an erroneous conception of the value of such studies."

Writing about the laboratory use of impregnated blocks of wood in testing wood preservatives, which was already well under way in Europe, he says that by using wood one may obtain conditions more or less closely resembling but not identical with conditions in actual service; but one would not only have to use a solvent in treating to low retentions, but there would be difficulties in obtaining an even distribution of the preservative in the wood. Furthermore:

"Getting rid of the solvent would require considerable time, during which a considerable loss of creosote would occur. . . .

"The composition of the creosote in the impregnated wood after the solvent has evaporated may be quite different from that of the original sample. More important still, the movement of the solvent in the wood during drying would cause an uneven distribution of the creosote."

The reader will bear in mind that these opinions were expressed in advance of the St. Louis and Berlin conferences on laboratory evaluation methods. Schmitz repeated them essentially in his 1930 paper, saying, for instance:

"Toximetric values are not in themselves an index of the wood preserv-

ing value of the substance tested. Other factors, such as leaching, volatility, chemical stability, penetrability, cost, cleanliness, etc., must all be considered in the final evaluation of a wood preservative."

With respect to the European wood block test, he felt that

" . . . until more confidence can be placed in the even distribution of the preservative in the test block(s) their use will be greatly limited."

He has maintained his arguments with a high degree of consistency in later papers, and they have unquestionably influenced American thought on laboratory procedures and their practical application.

The Petri dish method adopted as a possible American standard procedure at the 1929 St. Louis meeting followed closely the techniques that had been developed and published by Humphrey et al.,[61] Batemen[8] and Richards.[94] Bell Telephone Laboratories made an intensive study of the Petri dish method during this period. The data obtained were never organized for publication since it was felt that the required evaluation of toxicity and permanence of toxicity of preservatives could not be obtained by the Petri dish test.

European workers would accept neither the Petri dish test method nor Madison 517 as the test fungus. In 1931, about a year after the Berlin conference, and four years before Liese et al[71] reported on the task force development of the agar-block method, A. Rabanus of the I. G. Farbenindustrie Aktiengesellschaft, Germany, published his "Die toximetrische Prüfung von Holzkonservierungsmitteln" (Toximetric testing of wood preservatives).[86] A somewhat expurgated and amended translation of this paper was presented to the American Wood-Preservers' Association in 1933. In the writer's opinion much of the force of the Rabanus argument was lost in the translation. The emphasis on the relative merits of the agar toximetric test and of the agar-block test was considerably diluted; and the cautiously guarded but nonetheless positive philosophy on the possibilities of using the results of agar-block tests in actual wood preserving practice was made water thin.

Apparently there was an understanding that subsequent to the 1930 conference in Berlin[71] tests by the agar-block method would be run in the United States. For this project samples of creosotes as well as Scotch pine wood blocks were sent to a number of workers; but to the writer's knowledge no treated blocks were ever tested, or if they were no results were ever published. At Bell Telephone Laboratories some of the untreated Scotch pine blocks were put through preliminary trials with the Kolle flask technique,[125] and also a considerable number of plate and flask agar toxicity tests were run with the two sample creosotes. The inconsistency of the results — as far as translation to practical wood

preservation was concerned — was a strong stimulant toward the Laboratories' development of a block test, referred to later. It was more or less general information at the time that agar toximetric tests with native and European strains of test fungi were being run in other laboratories in this country; but again — as far as the writer knows — the results were not published.

In the meantime, and, as in the case of the Rabanus article cited above, before the publication of the Liese[71] report, Flerov and Popov[48] published in 1933 in German the basic general principles of a soil-block test. The significance of the article by these two Russian investigators was apparently completely lost on American workers until the publication in England in 1946 of Cartwright and Findlay's "Decay of Timber and its Prevention."[26] Findlay had been a member of the Berlin conference. Flerov and Popov were familiar with the discussions and result of the conference, and decided in favor of the soil base for their cultures after a critical review of the various methods then in use. Their proposals to all intents and purposes were unknown here.

Van den Berge's comprehensive thesis[116] on "Testing the Suitability of Fungicides for Wood Preservation" appeared in Dutch in 1934. A mimeographed English translation was made available soon after for limited distribution. European workers were about ready to confine the use of the agar toximetric test to determining relative toxicities only of various preservatives *in an agar medium*. Liese and his colleagues[71] summarized the arguments and experiments on the agar-block method in 1935, and launched it into a status of general acceptance in Europe and Great Britain. The British[24, 45] and German[33] editions of the standard were issued in 1939. The Rumanian version[110] — closely following the German — came out in 1950. Jacquiot,[64] Lutz[77] and Alliot[2] worked out proposals for standard procedures that would be more comprehensive and in their opinion better applicable to wood preservation research in France.

The Petri dish — and later the stoppered Erlenmeyer flask — agar methods continued to be used by many American investigators for testing wood preservatives, and there is no denying a certain utility in these methods for developing information about fungus poisons. The persistency of the agar techniques can be traced through publications by Richards,[94] Schmitz,[99, 101] Snell and Shipley,[108] Schmitz, Buckman and von Schrenk,[102] Schmitz, von Schrenk and Kammerer,[103] Bland[14] and Hatfield.[53] Baechler still uses the closed flask-agar method and the fungus called Madison 517 for determining basic toximetric values;[4, 5] and Finholt[46] has recently been bold enough to state that "Fungitoxic materials

can be evaluated as wood preservatives by mixing the toxic substances with a malt extract agar solution and then testing the mix against standard fungi."

Flerov and Popov had used sand in their preliminary experiments, but they were by no means the first to do so (see Falck[44]). Rabanus[86] had reported his experiments with sand-block cultures two years earlier. He placed a pair of wood blocks — one treated and one untreated — on glass rods on wet sand in Erlenmeyer flasks; and after sterilization he inoculated the blocks directly with his test fungi. He points out that in this procedure the conditions were less favorable for the fungi than when the treated wood is placed above or on a vigorously growing culture, as in the agar-block test.

Since the papers by Rabanus and by Flerov and Popov appeared in the same journal, one can assume that the latter knew of Rabanus' work. How much any of them knew of still earlier work by Breazzano is uncertain. His work in Italy,[20, 21, 22, 23] begun in the first decade of the century, is evidence of the intense interest of the management of the Italian railroads in some practical laboratory means for testing wood preservatives that would provide results sooner and with more definiteness than the traditional service tests. Parts of Breazzano's report of Oct. 9, 1913 are worth quoting in full from the English translation as historical background information. He reviews the situation as he sees it, and says:

"New systems and various substances for injection into wood are constantly being put on the market by industrial concerns, so that the Railway Administration finds itself confronted by an ever increasing number of processes to be examined and tested for efficiency."

By 1910 the Railway Experimental Institute

" . . . is well on the way toward testing the efficacy of a system of wood preservation by a method which gives dependable results even after a few months of observation.

" . . . after making use also of the advice on the subject received directly from Prof. Tubeuf and from Netzsch's laboratory . . . positive results were obtained with the following technique:

"On the bottom of an Erlenmeyer flask of 200 ml capacity was placed a thin layer of sand." After sterilization in dry heat at 180°C "sterilized water was poured on the sand to moisten it well. Then there was placed on the sand the sample of wood, of dimensions about 9 x 2 x 1 cm., with one end resting on the damp sand and the other on the inside wall of the flask."

The whole setup was sterilized in an autoclave at 120°C for about 20

minutes. The wood was inoculated by placing a piece of a culture of *Coniophora cerebella*, grown on agar medium, directly on the wood. Breazzano states that the wood was kept moist enough because of the water in the sand, that the fungus grew luxuriantly, and that "the development of the fungus was evidently at the expense of the wood, since no other nutritive substance was at its disposal."

He used blocks cut from treated beech ties. The fungus grew readily and he concludes that the treatment was not effective. He ends this early report with the statement:

" . . . If the experiment is carried on under carefully defined conditions the various methods proposed for immunizing woods can be judged all by the same standard."

Breazzano presented his method at Pisa in 1919, and in 1922[21] the principles of the sand-block culture were proposed as standard procedure (for Italy) for evaluating wood preservatives. Precise directions were given for the whole test technique, with important modification of the cultures, as indicated in the steps outlined below:

1. Sterilize by dry heat, at 180°C, "soyka" boxes 8 cm in diameter and 4 cm in height "in which is first placed a layer of sand 1 cm deep".
2. Prepare blocks of wood — treated and untreated — 4 x 4 x 2 cm, cutting them so that the broader faces will be transverse sections; and place these test blocks broad face down on the sand.
3. Sterilize at 100°C for one hour.
4. After sterilizing and cooling add sterile water in an amount that will be slightly in excess of what the sand can absorb.
5. After the wood blocks become moist plant *Coniophora cerebella* — without carrying over any agar medium with the transplant.
6. Incubate the "soyka" box cultures in a covered crystallizing dish in a dark place for one month at 20–25°C; and "Take care that in this time the water which the sand absorbs does not evaporate completely, and add sterile water when necessary."

At the end of the test the wood blocks were to be examined for decay; and if there was any doubt the wood was to be sectioned and examined microscopically for the presence of wood-destroying fungus hyphae (threads).

In retrospect the subsequent changes involving the use of soil instead of sand, and in the testing of blocks specially treated for the experiment, seem like refinements of Breazzano's methods. He later shifted to the use of very thin pieces of treated wood for his test specimens,[22] [23] severely criticizing the agar-block method that grew out of the Berlin conference as time consuming and inaccurate (loc. cit.).[72]

In Bell Telephone Laboratories, R. E. Waterman and his colleagues started work on a wood-over-water block method for testing wood preservatives soon after the St. Louis conference, and they published their early results in 1937 and 1938.[67, 126, 127] Their block was a ¾-inch cube with a hole drilled through it in the approximate center of a transverse face. The ¾-inch cubes simply represented sections of the ¾-inch square stakes that had been substituted for round saplings[69] in the small specimen test plot experiments. The hole served a double purpose — it facilitated handling the blocks during drying and sorting operations[126] and it served as a point of entrance of moisture, which was purposely provided for the block by means of a wood wick.

Leutritz[70] formalized a soil-block test completely independently of Flerov and Popov, and published his method in this JOURNAL (Vol. 25) in 1946, following an earlier short article in 1939[68] suggesting soil as a culture medium.

Beginning in the summer of 1944 and continuing until June 30, 1951, Bell Telephone Laboratories subsidized in part a series of studies by the Madison Branch of the Division of Forest Pathology, of the United States Department of Agriculture, Bureau of Plant Industry, in cooperation with the Forest Products Laboratory at Madison, Wisconsin. The results of these studies and of parallel investigations have appeared in eight papers[16, 35, 36, 37, 38, 39, 40, 95] from 1947 to date. The differences between the agar-block and the soil-block techniques, and the results obtained in comparable test series by the two methods are of fundamental importance. They are presented and discussed at length in a paper by Duncan.[41] Already some 40,000 blocks have been tested by the soil-block method at Madison, with 75 oil-type preservatives. Both at Madison and at Bell Telephone Laboratories, Murray Hill, additional work aimed at further refining of the soil-block technique is under way.

Subsequent to discussions of the new soil-block techniques between representatives of Bell Telephone Laboratories and of the Forest Products Laboratories of Canada, Sedziak[106] has developed a soil-block test involving burying the block in the soil all but one corner; and instead of placing it on a fungus culture growing on feeder blocks, he inoculates a corner of the test block directly.

For a general review of laboratory and test plot methods for evaluating wood preservatives the interested reader should have available, in addition to Cartwright and Findlay's book,[26] at least two more recent books, namely "Wood Preservation During the Last 50 Years" by van Groenou, Rischen and van den Berge,[118] and the third edition of Holzkonservierung by Mahlke-Troschel-Liese.[78] Hunt and Garratt[62] survey wood preservation

with particular reference to the American scene. The works of Boyce,[19] Baxter[10] and Hubert[58] should be consulted for general information on wood-destroying fungi and the pathology of timber products. Kaufert[65] prepared a concise bibliography of pertinent articles in 1949. For a fuller coverage the book by van Groenou, Rischen and van den Berge will be found most stimulating.

Much of the European work on the testing and application of wood preservatives has been summarized in challenging form by the investigators at the Berlin-Dahlem testing station.[54] In this memorial volume, the first paper, by Schulze, Theden and Starfinger, is a compilation of the results of comparative laboratory tests of wood preservatives by the agar-block method. So much work has been done that the ingenious graphical summary table is about 12 feet long; and even then the authors have omitted many results because the conditions of the standard test[33] were not observed. Becker[54(2)] brings up to date the results of testing insecticides in the second article; Becker,[54(3)] in the next paper, summarizes tests for termite control; and Becker and Schulze[54(4)] in the fourth article cover laboratory tests of preservative materials for the control of marine borers. Six additional articles on subjects directly related to wood preservation complete an excellent supplement to the Mahlke-Troschel-Liese book already cited. The emphasis is, somewhat naturally, centered on the work of the Berlin station.

Rennerfelt and his colleagues[42, 43, 88] are conducting a series of laboratory, decay chamber and test plot experiments in Sweden, aimed at evaluating wood preservatives for use in that country, and at possible correlation of experimental results with actual experience.

Bienfait and Hof[13] are working in Holland on what appear to be the broadest test post experiments in Europe at the present time, under both land and water exposure conditions. Their tests of 10 preservatives and some 3350 posts of Douglas fir, Scotch pine, European larch, Sitka spruce, poplar and willow rival Bell Telephone Laboratories' installations in four test plots at Gulfport, Miss., Orange Park, Fla., Chester, N. J., and Limon, Colo.[69, 75, 76] and the Forest Products Laboratory installations in Mississippi.[17] Bienfait and Hof, like Rennerfelt, have been using the standard European agar-block test in their plan for correlation of laboratory and field results. No report on the Holland tests has appeared since 1948.

Narayanamurti and his associates[82] in their first interim report on laboratory and field tests of creosotes of Indian origin present the results of some fifteen years work at the Forest Research Institute at Dehra Dun, indicating from still another quarter the compelling force that is

leading to the development of preservative evaluation methods to supplement or partly displace long and uncertain service tests. The authors present a mass of information on six different creosotes, on four creosote fractions, and on mixtures of the creosote with fuel oil of Persian origin. Sal (*Shorea robusta*) railway ties were used for the field trials. Many of the data are condensed into graphs that are small and difficult to read. The findings in general are favorable to the creosote-petroleum blends. The writer, on the basis of personal experience, is dubious about either the theoretical or practical significance, in experiments of the type reported, of the values given for standard deviations and standard errors. The scope of the work entitles it to more complete review than is practicable at this particular time and place.

Bell Telephone Laboratories are represented in a group carrying on comprehensive cooperative investigations of pedigreed creosotes on which four papers have already been published.[5, 12, 39, 93]

The results of outdoor tests of small stakes and fence posts are issued periodically by the Forest Products Laboratory at Madison, Wis.[15, 17, 18, 63] In this connection, the Proceedings of the American Wood-Preservers' Association are in the class of required reading. Additional references will be cited at appropriate points in the succeeding paragraphs.

Data will first be presented on some of the experience of Bell Telephone Laboratories and others with laboratory soil-block tests, with outdoor tests of small stakes, of pole-diameter posts, and with pole test lines in evaluating wood preservatives. Through analysis and discussion an attempt will be made to interpret the significance of the results obtained by the various evaluation procedures and to correlate the evidence. Emphasis will be placed naturally on creosote and pentachlorophenol because of their great importance to the Bell System pole plant. The writer intends to support his interpretations with experimental data wherever possible, reserving the privilege in some cases to make suggestions as to possible significance, even though complete technical proof may be lacking at present.

EVALUATION BY SOIL-BLOCK TESTS

General Procedures

Soil-block cultures have been described in a number of papers[95, 96, 52, 38, 39] since Leutritz presented his method in this JOURNAL in 1946.[70] Some of the following statements, therefore, will be repetition; but the intent is to outline the technique employed at Bell Telephone Laboratories as a base for later discussion.

The culture jars are wide mouth cylindrical 8-ounce bottles, provided with screw caps. The moisture content of the soil is predetermined on a representative sample, and enough distilled water is placed in the bottle so that when the soil is added its moisture content will be somewhat above 40 per cent by weight. The bottles are filled approximately half-full of screened field top soil — which means about 140 grams of an oven-dried sandy loam. The soil handles better if it is reasonably dry so that it can be poured through a suitable funnel; and putting the water in bottles before one puts in the soil results in a practically clean glass surface on the inside of the bottle above the soil level.

Two southern pine sapwood feeder blocks, measuring $1\frac{3}{8}$ inches in the direction of the grain by $\frac{3}{4}$ inch by approximately $\frac{5}{32}$ inch (35 x 20 x 4.0–4.5 mm) are placed carefully on the flattened soil surface, as shown in Fig. 1(a). The soil and feeder block setups are then sterilized for one-half hour at a pressure of 15 pounds per square inch, after which they are allowed to cool in the autoclave.

Inoculation is accomplished by carefully placing a piece of inoculum, cut from a fresh Petri dish culture of what may be called a standard test organism, at or near the middle of the feeder block surfaces. Under

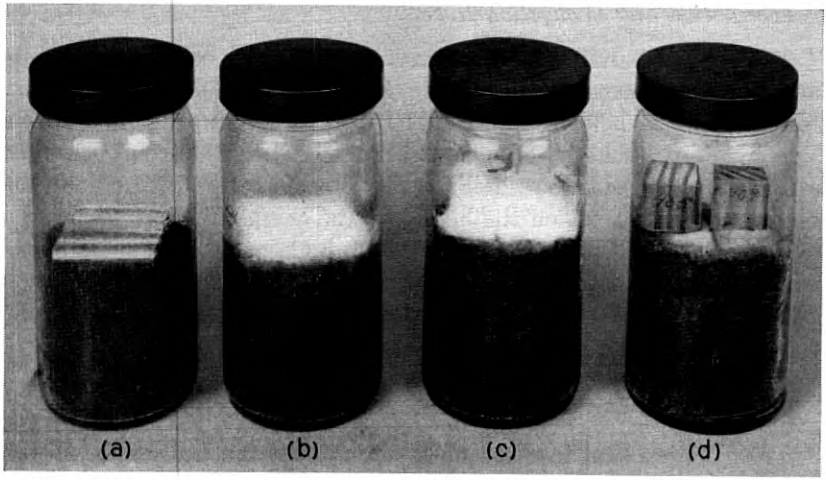

Fig 1—Four eight-ounce cylindrical bottles illustrating the soil-block cultures: (a) Bottle half-full of top soil, containing 40 per cent moisture on an oven-dry soil basis, with two flat feeder blocks of southern pine sapwood on top. (b) A sixteen-day old culture ready to receive the impregnated southern pine sapwood test blocks. (c) Two laboratory weathered test blocks from the same series treated to a below threshold concentration of pentachlorophenol 0.051 lb dry penta/cu ft, attacked by the test fungus, *Lenzites trabea*. (d) Two test blocks, laboratory weathered, treated to a retention of 0.194 lb dry penta/cu ft, near the threshold retention, showing resistance to fungus attack.

the temperature and humidity conditions of the incubation room, the growth of the fungus mycelium covers the feeder blocks in about two weeks and the fungus threads are then well started downward into the soil.

Treated test blocks are weathered and then conditioned under controlled temperature and humidity to approximate constant weight. They are then sterilized, along with untreated control blocks, in an autoclave for 15 minutes at 100°C, atmospheric pressure.

As a rule two treated blocks having approximately the same retention of preservative are placed together in a single test bottle. The incubation period is three months, in an incubation room held at a temperature of 80 ± 2°F and at a relative humidity of 70 ± 2 per cent. At the end of this period the cultures are taken down. This means that the blocks are removed from the bottles, brushed free of fungus mycelium, and weighed immediately. They are given a preliminary examination for decay evidence, and then reconditioned, under the same temperature and humidity conditions as before sterilization, to approximate constant weight. Fungus attack is determined by observation and by weight losses. The general setup of the cultures is illustrated in Fig. 1 (a–d).

Inoculation and Incubation Rooms

To facilitate handling the soil-block cultures, an inoculation room and an incubation room have been built (Fig. 2) at the Murray Hill Laboratories. Both are held at approximately the same temperature and relative humidity, that is, 80°F and 70 per cent. The inoculation room serves as a lock chamber, and passage from it to the incubation room has a negligible effect on the humidity and temperature of the incubation room. The latter is provided with an illuminated double plate glass window (Fig. 3), so that the interior can be exhibited without the necessity of entering the room. This window is fitted with a heavy roller shade, and the room ordinarily is kept dark.

Soil Characteristics and Moisture Content

The question that is asked most often about the cultures is whether a standard soil is used. European and American criticism has been definitely directed[78, 122] at the fact that the use of different soils might have so much effect upon the growth and the reaction of the test fungi in the cultures that quite different results would be obtained by investigators in different laboratories. This possibility is recognized; but the evidence to date seems to point to the general conclusion that perhaps the prin-

Fig. 2—Soil-block cultures on the shelves in the incubation room. The unpainted wood shelves come to equilibrium with the temperature and relative humidity and thus are a factor in keeping the conditions stable. The back edges of the shelves are set away from the wall to provide spaces for air circulation. The front edges of the shelves are provided with metal labeling strips.

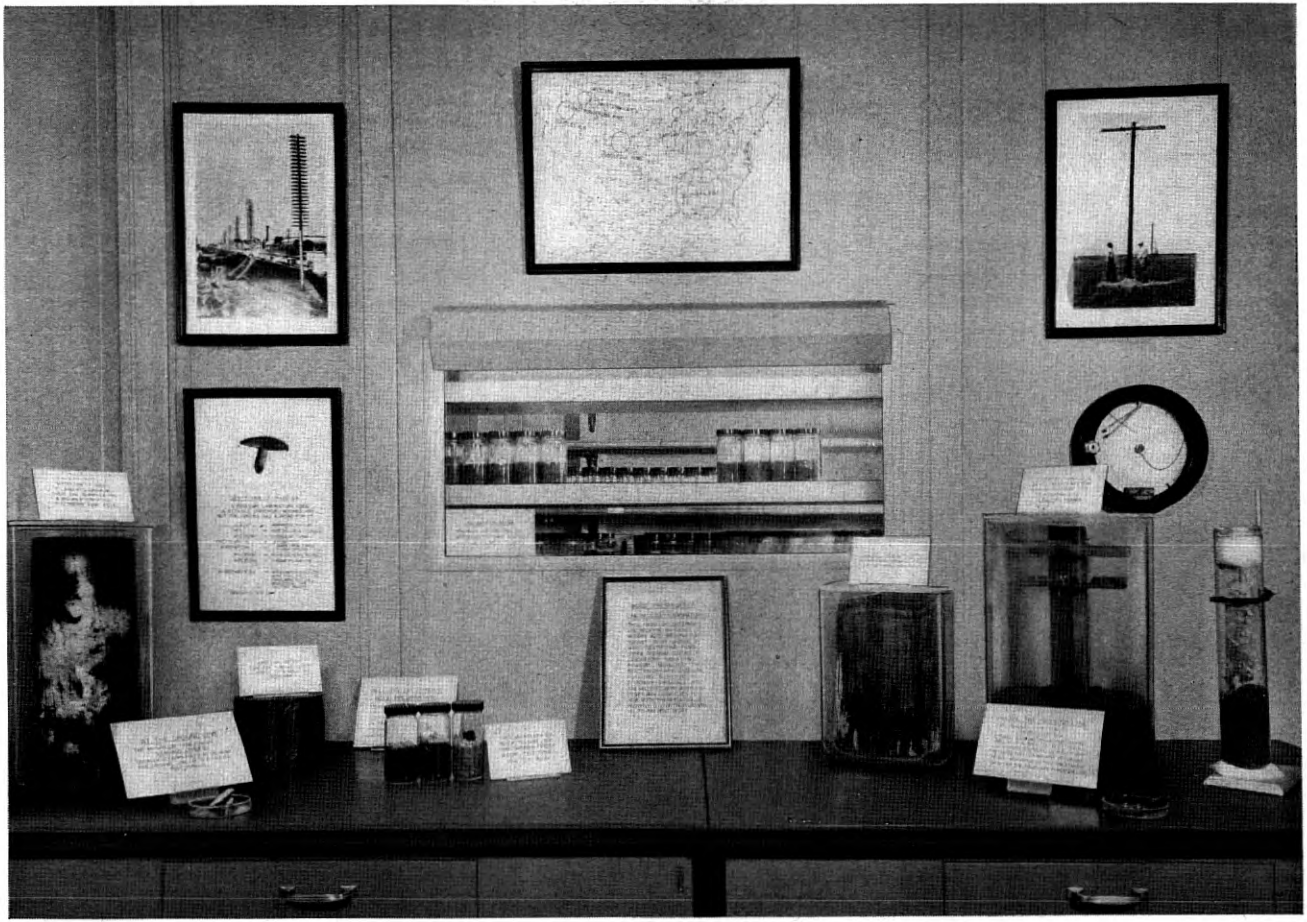

Fig. 3—The "Biossay Corner" at Bell Telephone Laboratories, Murray Hill. Centered in the illustration of the display

cipal and most important factor in the soil-block culture is the moisture holding capacity and content of the soil, rather than its nutrient function. If continued experimentation supports this conclusion it would not be necessary to limit the type of soils used except within rather broad limits. It also appears that the size and thickness of the feeder block now employed introduces enough wood into the culture bottle to mask any minor variations in the soil itself. The all important thing is to have enough water in the soil *throughout the test period* to keep the air above the soil essentially at 100 per cent humidity and the blocks at about fiber saturation — say about 27 per cent, oven-dry weight basis.

The soil in use at Bell Laboratories at present is obtained from a plot that has been set aside at the Chester (N. J.) Test Station. This plot has been fallow for twenty-five years. It supports a general grassy flora. The soil is a sandy loam with the following general description:

pH	4.9–5.0
Available magnesium	37.5 lb/acre
Available phosphorus	4.5 lb/acre
Available potassium	70.0 lb/acre
Organic matter	3.0 per cent

The cultures at the Forest Products Laboratory[41] have been made with a silt loam having a pH between 5.5 and 6.0. Bell Telephone Laboratories' tests have indicated the desirability of avoiding soils of either very sandy or very heavy clay types. The soil from the Chester Test Station described above is being used in all cultures, and there is a sufficient layer of top soil on the reserved plot to make parallel cultures for a good many years. Until such time as more definite and positive information on the effect of minor variations in the soil type are determined it is generally agreed that all of the comparative tests in any given series at least should be run on the same soil. Experimental work is now under way to determine the possible advantage of the addition of Krilium* to the soil in the culture bottles to maintain porosity and an even, high moisture holding capacity.

After the test blocks are placed in the culture bottles and during the course of the ninety-day incubation period the screw caps are left *loose*. The general technique followed in making up the soil cultures, as far as moisture is concerned, parallels that used at Madison. The moisture content is close to that recommended by Flerov and Popov,[48] namely 40–50 per cent of the weight of the soil plus the feeder block, with distilled water added during the test period, if necessary to maintain good

* An acrylonitrile product of the Monsanto Chemical Company.

growth of the test fungus. Breazzano[21] thoroughly saturated the sand base in his test cultures. Leutritz[70] and Harrow[51, 52] working with tightly closed culture jars found a 25 per cent level in the soil to be satisfactory.

Flerov and Popov state after special control tests "that replacement of (the) sand by soil had no effect on the results of the tests and only shortened their duration." Duncan[41] has found from her tests that variations in moisture content and soil type affect the degree of fungus attack only and that they do not change the determination of the treatment threshold concentration in any given set of test blocks.

Even-Aged Cultures

The thickness of the feeder blocks has been gradually increased to about $\frac{3}{16}$ inch, or between 4 and 4.5 millimeters. This provides food for the fungus to establish itself in the bottle. The inoculum pieces are roughly 1 cm square, cut from Petri dish cultures that are 15 ± 1 days old. The planting routine is carefully scheduled so that even-aged soil-block cultures — 13–15 days — are ready to receive the treated blocks when the latter are ready to be placed in test. This principle of using even-aged cultures has been stressed by the Madison investigators, and it is considered to be a factor of major importance in the proper culture technique.

Standard Test Organisms

There have been continuous discussions since the beginning of laboratory tests in Europe, as well as in this country, about what test organisms should be used. Conforming to the experience and practice at Madison the following three numbered strains of wood-destroying fungi are recognized as the "standard" strains for the testing of oil type preservatives in coniferous wood:

Lentinus lepideus, Madison 534
Lenzites trabea, Madison 617
Poria monticola, Madison 698

All three are known to be associated with the decay of treated timber. *Lentinus lepideus* is particularly tolerant of creosote,[41, 66] and relatively susceptible to pentachlorophenol. It has frequently been isolated from decaying creosoted southern pine poles and other creosoted coniferous timber in contact with the ground. *Lenzites trabea* is generally an "above ground" fungus. It also has been isolated from decaying creosoted timber; and it is the principal cause of "shell rot" in the above ground sap-

wood of western red cedar poles. It is relatively susceptible to creosote and quite tolerant of pentachlorophenol in the block tests. *Poria monticola* is relatively tolerant of pentachlorophenol and of copper compounds under laboratory test conditions, and relatively susceptible to creosote. It is of special interest also because it may be identical with some of the fungi tested in Europe under the name of *Poria vaporaria*, and thus its use may facilitate comparisons of a sort with results obtained by other investigators. For instance, information has reached the Division of Forest Pathology at Madison, from Findlay at the Princes Risborough laboratory in England, that Harrow's *Poria vaporaria*[51] is the same as Liese's,[71] and that it has been identified as a strain of *Poria monticola* by Miss M. Nobles of Canada.

Within the last few years another fungus, characterized by the formation of conspicuous saffron yellow strands, has been found associated with decayed specimens of creosoted pine poles.[60] The writer has seen the tell-tale strands in old cull dumps only. It has been identified as *Poria radiculosa*. Whether it is truly a primary attacker or a secondary organism is not yet clear. Soil-block tests are under way at Madison to determine its significance as a possible species to supplement *Lentinus lepideus* in the evaluation of creosote.

In connection with the use of the three numbered "standard" strains listed above, there may always be some reasonable doubt as to whether the cultures employed in different laboratories have the same virulence. To answer this question precisely involves a lot of careful biological check testing, and such tests are already being made in the Division of Forest Pathology at the Plant Industry Station, Beltsville, Md. It is assumed for the time being that the numbered strains are virulent and satisfactory test organisms for such preservatives as creosotes and pentachlorophenol-petroleum solutions.

The Scope of the Soil-Block Evaluation Test

For a complete understanding of the scope of the soil-block evaluation test it is necessary to consider this test as having two functions. The first function involves the use of the soil-block test per se (without weathering) to measure the reaction of the test organisms to various quantities of a given preservative, and to compare these reactions against different preservatives. In this function the test has been used in lieu of the agar Petri dish test,[94] and is considered to be much more satisfactory as a screening test by workers at the Laboratories. It has been employed at Madison for testing the natural durability of wood, plywood, fiber board, etc.

The second — and more important — function of the soil-block evaluation is that incorporating a weathering or aging procedure. This puts the test in the more practical category of testing the wood preservative properties, viz., *toxicity and permanence*. In this respect it has something in common with the German Standard DIN DVM 2176[33] for short time mycological testing of wood preservatives by the block method and covers a broad concept from the treating through partial aging of the blocks. Separate German standards cover procedures for leaching[34] and volatility tests,[78] (p. 264 and Fig. 42 of Reference 78).

The Laboratories' concept of the scope of the soil-block test including the weathering procedure is definite. It must be appreciated that this method which employs manipulative procedures involving both toxicity and permanency yields significant data in a period of only a few months. The data derived must be correlated subsequently, of course, with the data covering the results of tests of ¾ inch stakes six to seven years later, with the data on test posts some ten to fifteen years later, and with data on poles in line some twenty-five years later.

It is only being realistic to say that the Bell System cannot afford to wait for physical life tests of new materials under natural conditions of exposure before recommending them where techniques and extensive experience permit acceptable estimates to be made from accelerated evaluation in relatively short periods of time.

Preparation of the Test Blocks — Manufacture

Southern pine sapwood, free from stain or decay, is used as a base material for the test blocks. The process of manufacture begins at the saw mill, where freshly cut logs selected for the purpose are carefully sawed into one inch boards. Straight grain material is most desirable. The boards are kiln-dried immediately and shipped as soon as practicable to the Laboratories. It has been the practice to store the boards in a steam heated basement where the humidity is low enough to hold the moisture content of the boards down to about 5 to 7 per cent. The sapwood only is used, which means that any small heartwood portions must be marked out for rejection. The blocks are accurately cut ¾-inch cubes. A ⅛-inch hole is drilled through the center of the tangential surfaces of each block. It has been found that drilling the hole through the transverse surface, which was the early practice with Waterman, Leutritz and Hill[126] is a difficult procedure; and sometimes it amounts to an impossibility because the harder summerwood layers deflect and break the drills. In any event, drilling through the tangential surface opens up more paths for longitudinal absorption and penetration, as well as

evaporation, of the preservative. The feeder blocks and the ¾-inch test blocks are usually made at the same time, from parts of the same boards. The blocks are kept clean, and reserve stocks are carefully stored in a dry room. The blocks in storage reach an approximate moisture equilibrium of 6 to 7 per cent, on an oven-dry weight basis.

Test Block Selection for Density

Random samples of the blocks are weighed and segregated into groups at 0.1-gram intervals, 4.10 to 4.19 grams and 4.20 to 4.29 grams, for example. Blocks of practically equal weight can be chosen for the comparison within any given series of different concentrations of a preservative. The weighed groups of blocks are kept in convenient lot sizes in a dry place. Since the blocks are accurately cut the segregation by weight amounts to a segregation by density.

It has not been found necessary or practicable to separate the blocks into groups with the same numbers of annual rings, although in some instances an approximation to this ideal has been attempted. Furthermore, it has not been found practicable to separate the blocks on the basis of the direction in which the rings run across their transverse faces. From experience to date it does not appear that either ring direction or ring count has any material effect upon the behavior of the blocks in the culture as far as determination of preservative thresholds are concerned; but experiments are under way at Madison to determine the effect of density on the relative degree of decay. Inasmuch as all of the blocks are placed in culture with the transverse surface down, so that alternate spring- and summerwood layers are exposed directly to the test organism, the latter can enter either springwood or summerwood in accordance with its ability to resist the concentration of the preservative present in these two parts of the annual ring.

Average Block Volume

The average volume of the oven-dried blocks, determined from random samples by a mercury displacement technique, was found to be 6.484 cc, with a standard deviation of 0.0831. This represents a coefficient of variability of 1.28 per cent. The minimum-maximum range of volumes ran from 5.93 cc to 6.87 cc. These extreme deviations are normally detected in handling the blocks and both high and low volume blocks are rejected. The variation in density and volume of the test blocks will be discussed separately in the paragraphs dealing with the treatment of the blocks.

Treatment of the Test Blocks

The blocks selected for any given treatment are numbered serially with India ink on the upper half of one of the radial faces. All blocks are then oven-dried for 24 hours at 105°C to an approximate constant weight. The blocks are removed from the oven, and placed in a desiccator over P_2O_5. Check tests of blocks held under these conditions show that they do not change weight by more than one hundredth gram within the period they are held for weighing. The cooled, oven-dried blocks are weighed to the nearest hundredth gram.

The weighed blocks are placed in beakers and arranged with a tangential face down so that the transverse surfaces do not touch and the holes are vertical. This refinement in placing the blocks may not be necessary to obtain satisfactory absorption, but the procedure has worked out well, and it has been followed consistently. For any given concentration a sufficient amount of creosote, for example, and toluene are combined by weight to leave in the blocks, after treatment, the desired retention of preservative. Experience has indicated the concentration required, which depends to a certain extent upon the type of vacuum equipment that is available as well as upon the density of the blocks to be treated and the nature of the treating solution. Actually the process of treatment is simple. The beaker containing a given lot of weighed oven-dried blocks is placed within a bell jar and subjected to a vacuum of 3 to 4 millimeters of mercury. When this vacuum has been reached the line to the vacuum pump is shut off, and the preservative is run into the beaker from a separatory funnel[126] fitted into a rubber stopper on the top of the vacuum chamber, the blocks being weighted down below the level of the preservative.

The absorption and distribution of the oil within the blocks seems to take place very rapidly. Generally speaking, the beaker containing the blocks and the preservative are removed from the vacuum chamber as soon as practicable to permit continuing the treatment of another group of blocks. However, the blocks are usually held in the preservative solution for an hour or two, which apparently is long enough to bring about essentially complete saturation. When all the treatments for a given group of concentrations have been finished (a) the treated blocks are wiped to remove the excess oil, and (b) they are weighed immediately to 0.01 gram. The retentions of creosote or pentachlorophenol, for example, are determined on a gain in weight basis by calculations from the amount of material picked up during the treatment and the concentration of the preservative in the treating solution.

Retention Gradients

In the hope of setting at rest some of the doubts and criticisms that have arisen about the accuracy of the treatments and the retention of preservative in the blocks, some results of the treatment process just described will be presented in rather elaborate detail.

The success of the treatments depends upon experience, as indicated previously, with the particular type of vacuum equipment available. However, once the level of performance to be expected from the vacuum equipment is learned, one has to take into account the variations that are introduced by the density of the blocks and by the specific gravity of the treating solution. It is the intent in all of the treatments at the Laboratories to arrive at a series of gradient retentions, on as accurate a line as possible, and as nearly as possible equal gradients, so that the fairest comparison can be made of the behavior of the different preservatives. Fig. 4 shows the gradient obtained by plotting the data shown in Table I for retention of creosote and retention of pentachlorophenol solution over the concentration of these preservatives in the treating solution. The analysis of the creosote — BTL 5340 — is shown in Table II. The slopes of the two gradients are considered to be about as close as the experimental procedure will permit. Fig. 5 shows the gradient

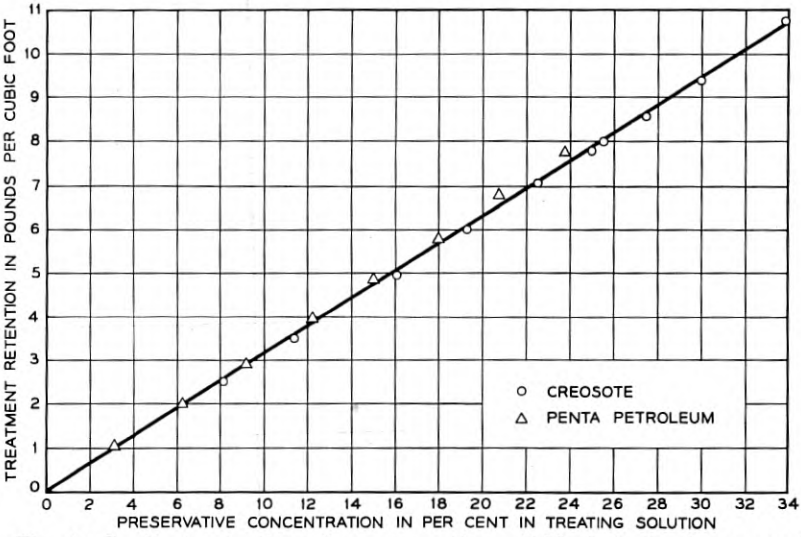

Fig. 4—Gradient retentions for comparative soil-block tests of a creosote (BTL No. 5340) and a penta-petroleum solution (4.92 per cent pentachlorophenol in Standard Oil Company of New Jersey No. 2105 Process Oil). The preservatives were used in toluene solution.

TABLE I — FULL-CELL TREATMENT

Soil block tests with creosote (No. 5340, see Table II) and with pentachlorophenol-petroleum (4.92 per cent in Standard of New Jersey No. 2105 Process oil) in toluene; absorption and retention of preservative data for parallel comparative tests.

Charge No.	Average Oven-dry		n	Average Absorption*		C†	Average retention						
							Whole preservative			Creosote		Penta	
							Creosote	Penta					
	Weight	Density		Total	Per cc			Sol.	Penta	Total	Per cc	Total	Per cc
	(gms)			(gms)		(per cent)	(lb/cu ft)			(gms)		(gms)	
9	3.77	.584	30	3.28	.509	8.18	2.49	—	—	.28	.043	—	—
10	3.78	.586	30	3.40	.527	11.45	3.48	—	—	.39	.061	—	—
6	3.80	.589	30	3.44	.533	16.11	4.96	—	—	.55	.085	—	—
5	3.83	.594	30	3.46	.536	19.33	5.99	—	—	.67	.104	—	—
8	3.80	.589	30	3.51	.544	22.55	7.08	—	—	.79	.123	—	—
1	3.74	.580	30	3.49	.541	25.00	7.81	—	—	.87	.135	—	—
4	3.77	.584	30	3.51	.544	25.55	8.03	—	—	.90	.140	—	—
3	3.80	.589	30	3.49	.541	27.50	8.58	—	—	.96	.149	—	—
2	3.78	.586	30	3.50	.543	30.00	9.40	—	—	1.05	.163	—	—
11	3.70	.574	30	3.33	.516	3.25		1.05	.052	—	—	.005	.0008
12	3.79	.588	30	3.33	.516	6.25		2.01	.099	—	—	.010	.0016
13	3.79	.588	30	3.29	.510	9.25		2.95	.145	—	—	.015	.0023
14	3.75	.581	30	3.31	.513	12.25		3.96	.193	—	—	.020	.0031
15	3.71	.575	30	3.34	.518	15.00		4.84	.238	—	—	.025	.0039
16	3.70	.574	30	3.34	.518	18.00		5.81	.286	—	—	.030	.0047
17	3.72	.577	30	3.38	.524	23.75		6.79	.334	—	—	.035	.0054
18	3.73	.578	30	3.38	.524	23.75		7.77	.382	—	—	.040	.0062

* Absorption is the total amount of the treating solution picked up at treatment, that is, the gain in weight, including both preservative and the toluene carrier.

† C is the concentration of the preservative, e.g., creosote or penta petroleum, in the treating solution, in grams per 100 ml.

for pentachlorophenol alone, without regard to the petroleum carrier, also plotted from data in Table I. The scale on the abscissa represents the concentration of either the creosote or the pentachlorophenol solution. The ordinate represents pounds per cubic foot retained by the blocks, calculated from the pickup during treatment and from the concentration of the creosote or pentachlorophenol in the treating solution.

The Amount of Preservative in the Blocks

The use of these gradient concentrations is a continuation of the procedure worked out in the earlier stages[39, 41] of the Madison tests.

TABLE II — ANALYSES OF CREOSOTE BTL No. 5340, WATER-FREE BASIS

	1. Fall, 1946	2. Spring, 1952*
Specific gravity 38/15.5°C	1.088	1.102
Distillation, per cent, cumulative		
to 210°C	0.00	0.00
210–235	0.80	0.00
235–270	12.87	13.59
270–300	42.12	
300–315	54.30	52.03
315–355	79.10	78.05
Residue above 355°C	20.90	21.64
Total	100.00	99.69
Sulph. res., gm/100 ml	0.51	0.59
Tar acids, gm/100 ml	4.10	4.44
Benzol insol., per cent	0.07	0.59
Specific gravity (38°C) 235–315°C		1.053
315–355°C		1.118

* Average of 2 analyses.

It should be noted that the retentions are calculated as averages for the respective charges. Attention is called to the small quantity of preservative material involved. Even in calculated retentions of creosote, for example, 9.40 pounds per cubic foot (Table I), the retention means 1.05 grams in the whole block, or 0.163 grams *in each cc of block volume*. In

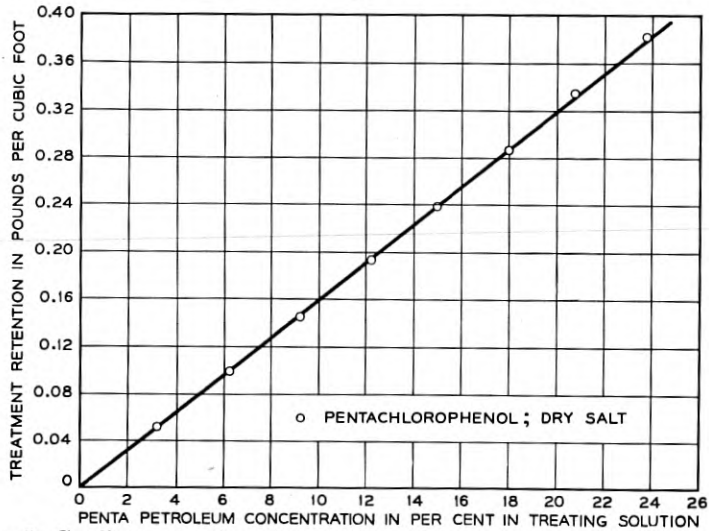

Fig. 5—Gradient retention of pentachlorophenol, calculated for the material alone, without the oil carrier. See Fig. 4.

the highest retention employed for the penta petroleum solution the calculated net retention averaged 7.77 pounds of the penta solution per cubic foot, or 0.382 pounds of pentachlorophenol per cubic foot; and these figures represent, respectively, 40 milligrams of pentachlorophenol in the average block, or 6.2 milligrams *per cc of block volume*. Exact data on treatment are discussed in the following paragraphs. The use of carefully calculated gradient retentions in each case makes it possible to detect any wide variation in the normal behavior of the blocks either with respect to pickup during treatment or in the reaction of the test fungus to the preservative.

Data are included in Table I on average oven dry weight of the blocks,

TABLE III — FULL-CELL TREATMENTS

Soil block tests; treating solution components, per cent by weight. (See Table I).

Charge No.	Penta-petroleum	Creosote	Toluene
19	4.65	4.65	90.70
20	10.50	10.50	79.00
21	15.00	15.00	70.00
71	—	23.00	77.00
72	—	23.00	77.00
73	—	24.75	75.25
74	—	25.50	74.50
75	—	30.00	70.00
76	—	32.00	68.00

average density on an oven dry weight and volume basis, average pickup of creosote or penta solution in pounds per cubic foot and in grams per block, the concentration of the preservative materials in the toluene preservative solution, and the average grams of preservative per cc of block volume. All of the blocks in these two groups of charges were chosen within a narrow density range.

Block Density and Preservative Absorption

It will be noted that in the charges in Table I there is a general trend upward in the grams absorbed at treatment per cc of block volume. as the specific gravity of the treating solution increases. This is, of course, one of the results of increasing the concentration of creosote, for example; and furthermore, as would be expected, the higher gravity solutions represented by the creosote treatments show a higher pickup in terms of total grams as well as in grams per cc. The make-up of the

TABLE IV — FULL-CELL TREATMENT WITH PENTA-PETROLEUM-CREOSOTE IN TOLUENE

Relation of variable block density to absorption of treating solution, grams per cc of block volume; density and volume on oven-dry basis. (See Tables V and VI).

Charge 19 Av. retention 2.99 lb/cu ft		Charge 20 Av. retention 6.69 lb/cu ft		Charge 21 Av. retention 9.59 lb/cu ft	
Density oven-dry	Absorption gms/cc of block vol.	Density oven-dry	Absorption gms/cc of block vol.	Density oven-dry	Absorption gms/cc of block vol.
.440	.599	.468	.589	.484	.581
.459	.583	.483	.584	.489	.576
.459	.586	.533	.549	.505	.564
.475	.575	.543	.542	.520	.550
.507	.561	.544	.527	.529	.555
.517	.545	.557	.537	.541	.541
.533	.544	.564	.529	.549	.538
.535	.525	.567	.530	.554	.532
.543	.530	.569	.569	.555	.526
.543	.537	.604	.510	.573	.538
.571	.508	.606	.511	.582	.528
.574	.524	.611	.504	.582	.530
.608	.488	.616	.499	.592	.517
.609	.495	.618	.491	.602	.512
.611	.500	.618	.497	.610	.497
.611	.533	.624	.503	.612	.507
.623	.483	.625	.493	.614	.502
.624	.493	.627	.492	.614	.508
.637	.476	.628	.488	.615	.540
.638	.477	.644	.481	.617	.495
.641	.474	.645	.488	.617	.504
.645	.478	.646	.482	.627	.495
.646	.473	.651	.482	.633	.497
.657	.462	.662	.433	.647	.488
.660	.453	.674	.461	.650	.487
.663	.466	.676	.465	.664	.416
.663	.468	.674	.452	.672	.473
.668	.462	.690	.456	.684	.470
.691	.447	.703	.445	.687	.464
.695	.450	.738	.427	.723	.442
Average					
.592	.507	.614	.501	.598	.512
Standard deviation					
.0726	.0435	.0616	.0405	.0591	.0370
Coefficient of variability—per cent					
12.26	8.58	10.03	8.08	9.88	7.23

treating solutions for Charges 19–21 and 71–76, inclusive, are shown in Table III. The relation of the density of the blocks to the pickup, i.e., the absorption at time of treatment, is illustrated in Tables IV, V and VI and in Fig. 6. The data have been split up to facilitate reference.

Table IV shows the complete data for oven dry density and for absorption in grams per cc of block volume for Charges 19, 20 and 21, with values for the average, for the standard deviation, and for the coefficient of variability. The pickup varies inversely as the density, which is to be expected when random blocks instead of selected density blocks are employed. The coefficient of variability in the density figures is evidently greater than it is in the pickup figures; and a lower figure for the latter is related to a lower figure for the former.

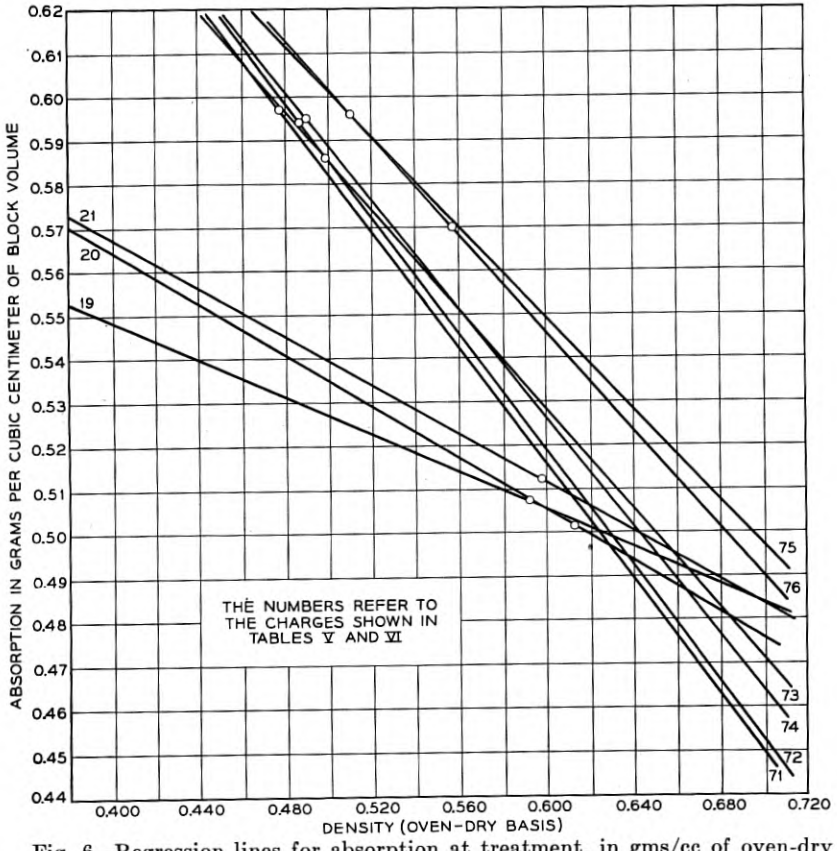

Fig. 6—Regression lines for absorption at treatment, in gms/cc of oven-dry block volume, on oven-dry density of the ¾-inch cube test blocks.

These same values for these three charges, 19, 20 and 21, and similar values for charges 71–76, inclusive, are incorporated along with average and range of retention data in Table V, and with statistical data for regression lines in Table VI. The data serve to illustrate the degree of variability in treatment results that may occur when random blocks are used. The best indices of these variations are in the columns showing

TABLE V — RETENTION DATA FOR FULL-CELL TREATMENT

Soil block tests; average and range of preservative retention, by charges, at treatment.

Charge No.	n	Penta petroleum lb/cu ft	Pentachlorophenol						Creosote					
			lb/cu ft			gms			lb/cu ft			gms		
			Av.	Min.	Max.	Av.	Min.	Max.	Av.	Min.	Max.	Av.	Min.	Max.
19	30	1.51	.093	.064	.097	.008	.007	.010	1.48	1.29	1.97	0.154	0.134	0.205
20	30	3.35	.163	.136	.219	.017	.014	.023	3.34	2.77	4.44	0.348	0.288	0.462
21	30	4.79	.237	.192	.269	.024	.020	.028	4.80	3.90	5.46	0.500	0.407	0.569
71	30	—	—	—	—	—	—	—	8.58	7.65	9.04	0.892	0.817	0.941
72	30	—	—	—	—	—	—	—	8.52	7.67	9.20	0.886	0.798	0.957
73	30	—	—	—	—	—	—	—	9.06	8.43	9.57	0.943	0.876	1.025
74	30	—	—	—	—	—	—	—	9.46	8.61	10.27	0.984	0.895	1.068
75	30	—	—	—	—	—	—	—	11.12	10.07	12.06	1.158	1.047	1.254
76	30	—	—	—	—	—	—	—	11.37	11.05	11.94	1.184	1.149	1.242

TABLE VI — FULL-CELL TREATMENT

Soil-block tests with (a) penta-petroleum creosote, and with (b) creosote, in toluene; relation of variable block density to absorption of treating solution.

Charge No.	n	Av. retention lb/cu ft		Density oven-dry			Absorption gms/cc vol.			Correl. coeff. r	Regression of absorption Y on density X
		Creosote	Penta-petro Creosote	Av.	σ*	c.v.†	Av.	σ*	c.v.†		
19	30	—	2.99	.592	.0726	12.26	.507	.0435	8.58	−.3444	.6286−.2066X
20	30	—	6.69	.614	.0616	10.03	.501	.0405	8.08	−.4282	.6736−.2820X
21	30	—	9.59	.598	.0591	9.88	.512	.0370	7.23	−.4718	.6893−.2957X
71	30	8.58	—	.477	.0428	8.97	.597	.0293	4.91	−.9589	.9109−.6580X
72	30	8.52	—	.486	.0457	9.40	.594	.0326	5.49	−.9233	.9143−.6589X
73	30	9.06	—	.499	.0456	9.14	.586	.0275	4.69	−.9426	.8701−.5692X
74	30	9.46	—	.489	.0402	8.22	.595	.0279	4.69	−.8916	.8969−.6181X
75	30	11.12	—	.510	.0538	10.55	.596	.0302	5.07	−.9206	.8595−.5170X
76	30	11.37	—	.557	.0095	1.71	.570	.0114	2.00	−.4583	.8781−.5543X

* σ = standard deviation.
† c.v. = coefficient of variability, per cent.

the average and total spread in grams of preservative absorbed. The effect of selection for density is shown clearly — within the particular treatment groups — by the figures for charges 75 and 76. In the latter the use of selected even density blocks reduced the spread to below half of that in charge 75, and reduced the coefficient of variability by two-thirds.

Regression lines for pickup in grams per cc of block volume on oven dry density are shown for the two groups of charges in Fig. 6. The flatter slope of the lines for charges 19, 20 and 21 seems to reflect the difference in specific gravity and viscosity of the treating solutions. Higher densities have greater effect on absorption of higher gravity solutions. The fact remains that there is considerable uncertainty as to the significance of strict selection of the blocks for density if one uses a series of closely spaced gradient retentions.

Data for 8-pound charges for comparative soil-block tests of two of the cooperative creosotes, Nos. 7 and 9, a low residue domestic oil, low in tar acids and naphthalene, and a British vertical retort tar creosote, respectively, are condensed in Table VII. The differences in the treatment results are not considered to be significant.

Weathering

The blocks remain on the racks on the laboratory tables for about one week, which is long enough to permit the evaporation of most if not all of the toluene. Experiments have shown that when blocks are treated with toluene alone the toluene is all lost, on a weight basis, within 24 hours. When treating solutions of creosote in toluene are used the evaporation rate, also determined by weight loss, is slower; but it is believed that all of the toluene is gone before the blocks are ready for test. In any event, after the above-mentioned preliminary drying period the blocks are handled in a manner that differs somewhat from the procedure that

TABLE VII — RETENTION FOR FULL-CELL 8-POUND TREATMENT DATA

For parallel comparative soil-block tests of cooperative creosotes No. 7 and 9 against *Lentinus lepideus*, Mad. 534: toluene-creosote treating solution.

Creosote No.	n	Density oven-dry	Creosote lb/cu ft				Creosote grams				Standard deviation	Coefficient of variability, per cent
			Av.	Min.	Max.	Spread	Av.	Min.	Max.	Spread		
7	30	.606	8.11	7.75	8.51	0.76	.848	.809	.889	.080	.0201	2.37
9	30	.601	7.90	7.49	8.56	1.09	.825	.780	.893	.113	.0237	2.87

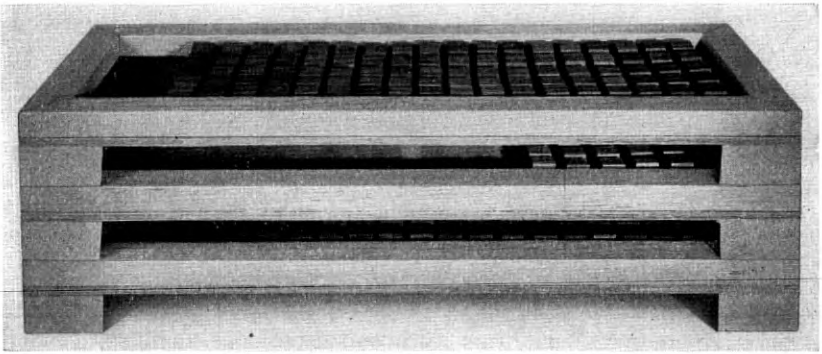

Fig. 7—Three unpainted stacked handling trays. The trays have plastic screen bottoms on which the blocks can be arranged with free air space all around, to promote even drying conditions.

has been followed up to this time at Madison. The principal difference is the omission of any tests of unweathered blocks. Instead, the emphasis is placed on the development of a weathering or aging cycle that will bring about total overall preservative losses like those that occur in ¾-inch stake specimens or in pole-diameter posts in the Gulfport test plot. The character and extent of such losses will be discussed later.

Two systems of weathering have been employed up to the time of this writing. The first consists in soaking the blocks over the week in water that is changed morning and night and drying them at room temperature over the weekend, in accordance with German standard for leaching;[34] and the second is the same method that is employed in the Madison tests[39, 41] in which the blocks are strung on nylon thread, separated by glass beads, and exposed to outdoor weathering under natural conditions for sixty days. The duration of this outdoor test has been limited to sixty days, regardless of the season or month of the year. The effectiveness of the climatic conditions at the Chester Field Station during the period from October, 1951, to April 1, 1952, compared with the roof weathering as conducted at Madison with the same creosote sample remains to be seen.

As for the German standard leaching procedure,[34] experience up to April 1, 1952, indicates that the method does not result in the removal of creosote, for example, in the same degree or manner as preservative materials of this type are removed by outdoor weathering conditions. It definitely is not comparable with the latter in its effect. The German leaching procedure simply uses too much water and not enough air and heat; and Bavendam[9] quotes Falck as saying that creosote is insoluble in water and that it cannot be *washed* out of wood. The failure of the

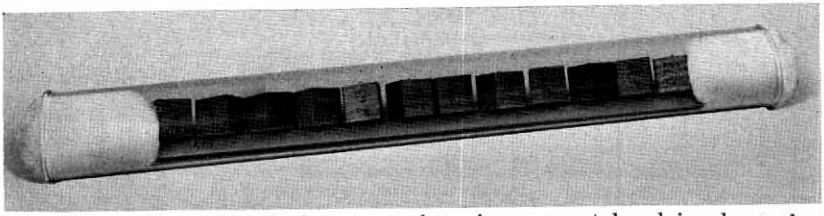

Fig. 8—Creosoted test blocks, arranged on pins on a metal rack in a large glass tube to facilitate handling during sterilization and the subsequent operation of planting in the soil cultures.

leached creosoted blocks to decay in the reported soil-block tests on the cooperative creosotes[39] confirms European experience. Schulze, Theden and Starfinger[56 (1)] (p. 15, Tab. 13 of Reference 54) indicate that there is little if any reduction in the preservative value of creosote as a result of their standard leaching tests. Therefore, in order to accelerate the weathering process new techniques are being worked out at the Laboratories in which the wet cycle is shortened (Cf. Rhodes et al,[50, 89]) and in which, without the use of a wheel, controlled artificial heat is used to speed up evaporation during the rest of the cycle.

Conditioning

Convenient unpainted wood trays (Fig. 7) with plastic screen bottoms are used for handling the blocks in groups at any time during their processing schedule. At the end of the weathering cycle, indoor or outdoor, the blocks are arranged on such trays and conditioned to an approximate constant weight and about 12 per cent moisture content on shelves in the 80°F and 70 per cent relative humidity of the incubation room. Weights before test, to the nearest 0.01 gram, are taken at the end of the conditioning period. The relative amounts of wood, water and creosote in the blocks are determined by weight from test blocks, and by weight and extraction from control blocks after sterilization.

Sterilizing

The test blocks are arranged for sterilization on metal racks in large glass tubes (Fig. 8). The autoclave temperature is held at 100°C for 15 minutes.

Flow Chart for the Bioassay Test

The various steps in the whole evaluation procedure are indicated in the flow chart shown in Fig. 9. Rhodes[89] used a similar idea to illustrate his procedure.

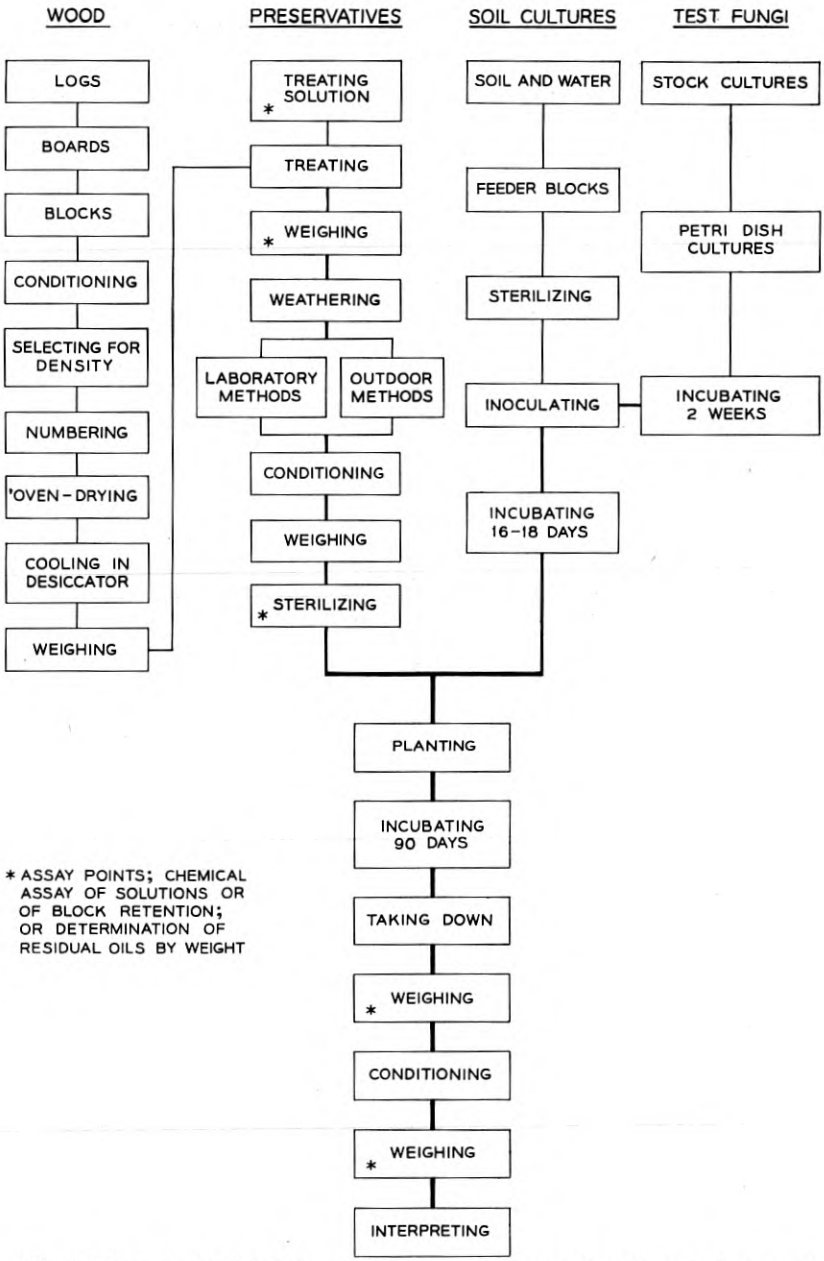

Fig. 9—Flow chart of the laboratory biossay test procedure. Steps in the overall processing must be carefully scheduled to provide adequate time for manipulation and to assure the requisite quantity of even-aged cultures at the time the treated and weathered test blocks are ready to go into test.

Some Madison Test Results

Inasmuch as results from recent Bell Laboratories' bioassay tests will not be immediately available, data from the Madison experiments[41] are used in order to illustrate the results that one may hope to secure from

TABLE VIII — BIOASSAY BY SOIL-BLOCK TESTS ON OUTDOOR WEATHERED BLOCKS

The relation of retention at treatment to block weight loss; Madison data.

Preservative†											
6			7			8			11		
n	R*	Weight loss per cent	n	R	Weight loss per cent	n	R	Weight loss per cent	n	R	Weight loss per cent
Test fungus, Lentinus lepideus, *Mad. 534*											
6	17.3	1.83	6	17.0	1.77	6	17.4	1.60	6	15.2	2.22
6	14.4	1.94	6	14.1	1.82	6	14.7	1.58	6	14.0	2.22
6	11.5	1.62	6	11.7	1.73	6	11.9	1.57	6	11.7	2.17
6	8.9	2.33	6	8.6	2.92	6	8.9	3.31	6	9.0	1.80
7	6.5	5.73	6	6.2	8.96	6	6.1	7.18	6	6.2	2.57
5	4.3	9.89	6	4.4	11.73	6	4.6	10.61	6	4.2	9.93
6	3.1	13.07	7	3.1	15.77	6	3.1	14.95	6	3.1	15.79
6	2.3	15.44	5	2.3	19.19	6	2.3	16.28	6	2.3	16.84

Preservative								
A			B			E		
n	R	Weight loss per cent	n	R	Weight loss per cent	n	R	Weight loss per cent
Test fungus, Lentinus lepideus, *Mad. 534*								
2	11.70	2.70	2	12.25	3.10	2	11.15	2.55
1	10.10	2.90	2	10.70	2.40	1	9.50	2.60
4	7.20	3.94	3	7.80	2.63	4	8.38	2.55
3	5.43	7.96	4	4.88	2.71	4	5.53	2.56
4	3.60	15.20	3	2.97	3.10	3	3.70	2.63
10	2.15	22.18	3	2.13	3.49	3	2.76	6.04
6	1.00	23.84	4	1.55	6.69	4	2.05	9.49
5	.58	17.35	3	.97	12.34	4	1.50	16.53
			3	.70	25.53	10	.77	29.70
			8	.40	35.43			

TABLE VIII—*Continued*

Test fungus, Lenzites trabea, *Mad. 617*

Preservative								
A			B			E		
3	9.14	2.57	2	12.35	2.80	3	10.93	2.70
4	7.58	2.42	2	10.30	2.60	3	9.17	2.40
3	5.53	2.73	3	7.80	2.80	2	6.95	2.60
2	3.70	2.81	4	4.95	3.24	3	5.30	2.49
3	3.20	3.97	3	3.33	14.21	4	3.95	3.14
6	2.17	8.57	11	1.51	35.76	3	2.47	8.00
7	1.14	38.96	10	.46	46.63	4	1.85	10.79
7	.60	55.44				6	1.08	52.90
						7	.64	55.06

* R = retention at treatment in lb/cu ft.
† Preservatives 6, 7, 8 and 11 are the numbered coop. creosotes (12).
A = BTL 5340 creosote,
B = 5 per cent penta in petroleum,
E = 50/50 by volume mixture of A and B.
All preservatives were applied in a toluene solution. The heavy lines represent approximate threshold levels.

carefully following the soil-block technique. The data, representing the writer's interpretation of the relation between average weight loss and average treatment retention, are shown in Table VIII and represented by the graphs in Figs. 10, 11 and 12. The preservatives in Figs. 11 and 12 labeled (A), (B) and (E) on the graphs, were respectively a domestic creosote (BTL No. 5340, Table II); a 5 per cent solution of pentachlorophenol in Standard Oil Company of New Jersey No. 2105 Process oil, and a 50/50 by volume mixture of these two.[41] Fig. 11 represents the results obtained with cooperative creosotes Nos. 6, 7, 8 and 11[39] and with BTL No. 5340.

In all three figures there are weight losses that can evidently be classed as operational losses, that is, losses by evaporation of some of the volatile materials still remaining in the blocks during the time they were in test and in the subsequent conditioning period.[41] The general areas in which the amount of preservative with which the blocks were treated failed to protect the wood against attack by the different fungi are shown by the rise in the weight loss lines. Perhaps the most interesting set of comparative results are revealed by Fig. 10. The test fungus was *Lentinus lepideus*, Mad. 534. From these graphs the threshold for creosote for this organism

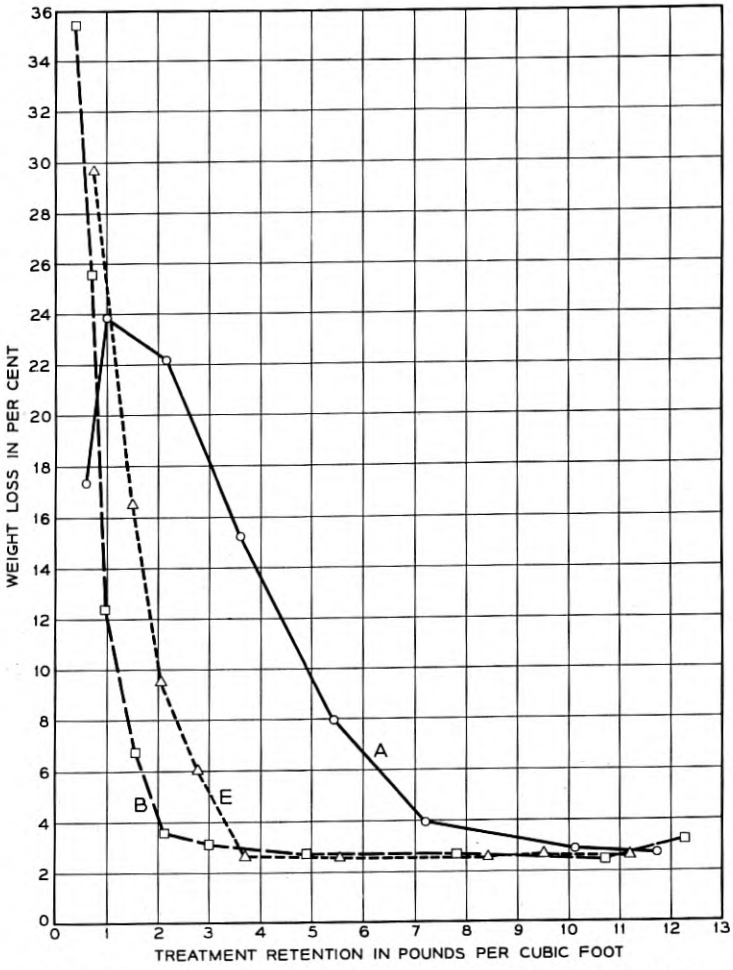

Fig. 10—Soil-block tests against *Lentinus lepideus*; weathered southern pine sapwood blocks; comparison of (A) creosote No. 5340, (B) a 5 per cent solution of pentachlorophenol in Standard Oil Company of New Jersey No. 2105 Process Oil and (E) a 50/50 by volume blend of the two; the relation of operational weight losses, losses by decay, and retention at treatment, lb/cu ft; based on Madison data. The Madison treatment thresholds for these 3 preservative solutions were set at 7.5, 2.4 and 3.4 lb/cu ft, respectively. See text, Table VIII, companion Figs. 11 & 12, Bibliography, Reference 41.

appears to be somewhat in excess of 7.5 pounds per cubic foot, whereas the thresholds for the pentachlorophenol solution are somewhere between 2 and 3 pounds, and the threshold for the mixtures of the creosote and the penta solution about 3.7 pounds per cubic foot. Duncan[41] gives these respective thresholds as 7.5, 2.4 and 3.4 pounds per cubic foot.

Fig. 12 shows that when decay does occur as a result of attack by *Lenzites trabea*, Mad. 617, the loss of weight in the wood is considerably greater than in the case of attack by *Lentinus lepideus*. In the case of *Lenzites trabea*, creosote appears as the best of the three preserva-

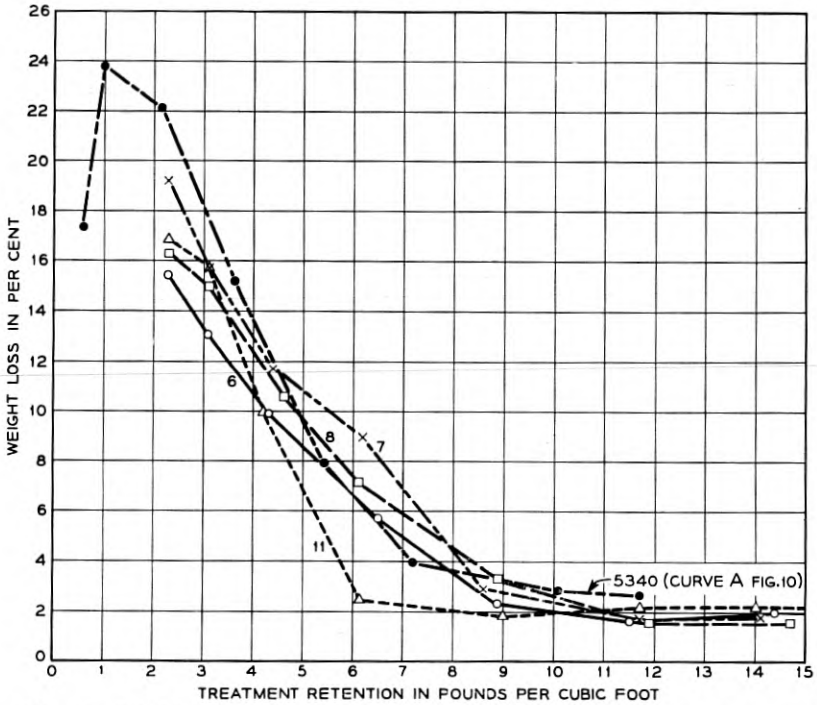

Fig. 11—Soil-block tests against *Lentinus lepideus*; weathered blocks; comparison of cooperative creosotes Nos. 6, 7, 8 and 11, and BTL No. 5340; based on Madison data. The Madison treatment thresholds for these five creosotes were set at 9.0, 9.0, 9.4, 6.5 and 7.5 lb/cu ft, respectively. See Table VIII and Bibliography, References 39 and 41.

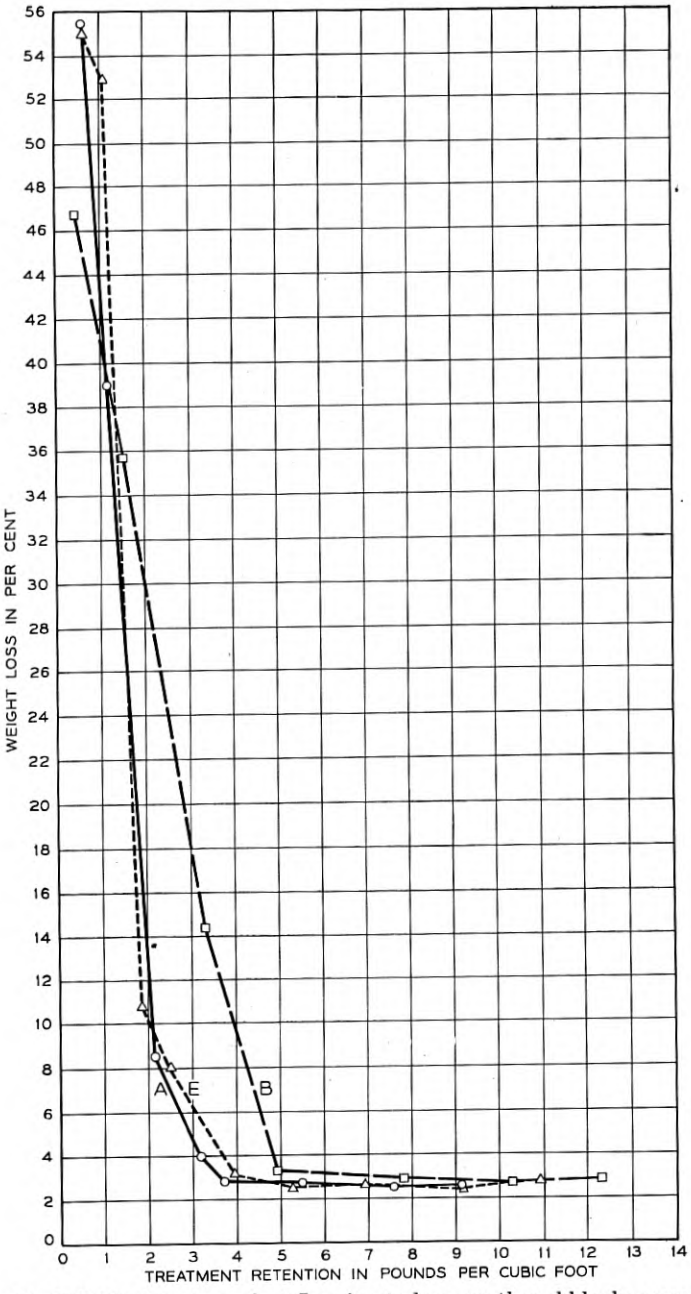

Fig. 12—Soil-block tests against *Lenzites trabea*; weathered blocks; comparison of preservatives A, B and E (Fig. 10); based on Madison data. The Madison treatment thresholds for these 3 preservatives were set at 3.2, 4.8 and 4.0 lb/cu ft, respectively. In the poorly protected blocks note the higher per cent weight losses caused by *Lenzites trabea* in comparison with weight losses caused by *Lentinus lepideus* (Figs. 10 & 11). See Table VIII and Bibliography, Reference 41.

tives; and the mixture of creosote and penta solution is somewhat better than the penta solution alone, although the difference is not great.

The results obtained in testing the different creosotes represented in Fig. 11 are similar in general character for the four domestic oils, but oil No. 11, the mixture of British vertical retort tar creosote and British coke oven tar creosote, appears to behave differently. The thresholds for all of these creosotes, as determined by the Madison investigators, are shown in Table XXXV. The figures in this table correspond very closely to thresholds determined by visual observation of the test blocks.

Check Tests at the Murray Hill Laboratories

It will be noted that in Fig. 11 the points used for locating the graphs are rather far apart in the general region of the estimated thresholds. The values shown in Table XXXV were obtained at Madison by the intersection of regression lines drawn through the points representing

TABLE XXXV — SUMMARY AND INTERPRETATION OF SOIL-BLOCK TESTS

Weathered, creosoted southern pine sapwood blocks; creosote losses; amounts and gross characteristics of residual oils at threshold retentions for *Lentinus lepideus*.

1	2	3	4	5	6	7	8	9	10	11
Item	Creosote No.*	Specific gravity 38/15.5°C	Residue above 355°C per cent	Threshold lb/cu ft	Per cent loss	Creosote loss lb/cu ft	Residual creosote lb/cu ft	Calculated residue above 355°C	Residual creosote	
									>355°C lb/cu ft	<355°C lb/cu ft
1	1	1.065	18.5	9.8	53.1	5.2	4.6	39.4	1.81	2.79
2	7	1.077	20.5	9.0	47.8	4.3	4.7	39.3	1.85	2.85
3	2	1.081	30.6	10.2	47.1	4.8	5.4	57.8	3.12	2.28
4	6	1.093	34.2	9.0	37.8	3.4	5.6	55.0	3.08	2.52
5	3	1.108	50.4	12.2	30.3	3.7	8.5	72.3	6.15	2.35
6	8	1.115	53.2	9.4	25.5	2.4	7.0	71.4	5.00	2.00
7	9a		21.2	5.7	40.4	2.3	3.4	35.6	1.21	2.19
8	9	1.001	20.0	5.8	43.1	2.5	3.3	35.1	1.16	2.14
9	10a		14.4	6.7	50.9	3.4	3.3	29.4	0.97	2.23
10	10	1.068	15.2	6.9	52.2	3.6	3.3	31.8	1.05	2.25
11	11	1.038	18.0	6.5	47.7	3.1	3.4	34.4	1.17	2.23
12	M1	1.107	41.9	8.0	33.8	2.7	5.3	63.3	3.35	1.95
13	M2	1.070	18.1	8.3	50.6	4.2	4.1	36.6	1.51	2.59
14	BTL 5340	1.088	20.9	7.5	46.6	3.5	4.0	39.1	1.57	2.43

* Creosotes 1, 2, 3, 6, 7, 8, 9, 10 and 11 are those in use in the Cooperative Creosote Tests (see Bibliography, References 12 and 39. Oils 9a and 10a are samples from the same lots as numbers 9 and 10. (See Bibliography, Reference 36.) For oils M1 and M2 see Bibliography, References 37 and 38. Creosote 5340 is shown in Table II.

operational losses and through the points representing weight losses. Data from a repetition of these tests is desirable in order to establish the thresholds more definitely from actual weight loss or observational data taken close to the assumed threshold points. At Bell Telephone Laboratories a check series of tests is now under way on cooperative creosotes 6, 7 and 8, domestic oils, and creosotes 9, 10 and 11, British oils; and comparison tests are also being run on creosote BTL-5340 and on 5 per cent pentachlorophenol in the 2105 process oil. The aim has been to treat the blocks to a series of retentions that vary narrowly around the thresholds set by the Madison investigators.

Across the Threshold

Fig. 13 is an illustration of representative blocks from the creosote series, line A (creosote BTL-5340) in Fig. 10, just at and below the threshold. Fig. 14 shows the character of the attack by *Lenzites trabea* on blocks treated with a 4.92 per cent solution of pentachlorophenol in Standard Oil Company of New Jersey's 2105 Process Oil in toluene. The blocks are represented at twice their original linear dimensions. The exact nature of the decay is difficult to show. The experimenter has to learn a system of diagnosis that involves both visual observation and the "feel" of the blocks for distortion and firmness that supplement weight loss data. For example, in Fig. 14, a threshold between 0.20 and 0.25 pound of penta per cubic foot (Blocks C and D) is indicated and this conforms closely to the results with the same penta-petroleum solution at Madison.[41]

The Significance of the Results of Laboratory Soil-Block Tests on Oil-Type Preservatives

The main conclusions from this discussion of the results of soil-block tests on weathered creosoted wood conducted at Madison are (a) that in general, under the test conditions, at least 8 and sometimes 9 pounds or more of creosote per cubic foot is a necessary treatment to prevent attack by *Lentinus lepideus* on ¾-inch cube blocks of southern pine sapwood; and (b) that a penta petroleum solution is much more effective than creosote against this same organism. As will be emphasized later, this general conclusion about *Lentinus lepideus* and creosote corresponds with the conclusions to be drawn from the interpretation of results of the small stake tests and from the test of pole-diameter posts in the Gulfport test plot.

The creosote tested is a better preservative against *Lenzites trabea* than the penta-petroleum, but the creosote threshold for this organism

is below what one would have to use commercially in order to insure protection against *Lentinus lepideus* and to preserve wood exposed to ground contact.

As far as *Lentinus lepideus* is concerned, the best overall preservative combination from the laboratory test would appear to be a 50/50 by volume blend of the creosote and the penta-petroleum solution, since such a blend appears to contain the best in both components. However, there are certain practical considerations, principally relating to the incompatibility of some creosotes and petroleums, which make the combined preservative a difficult one to operate with commercially.

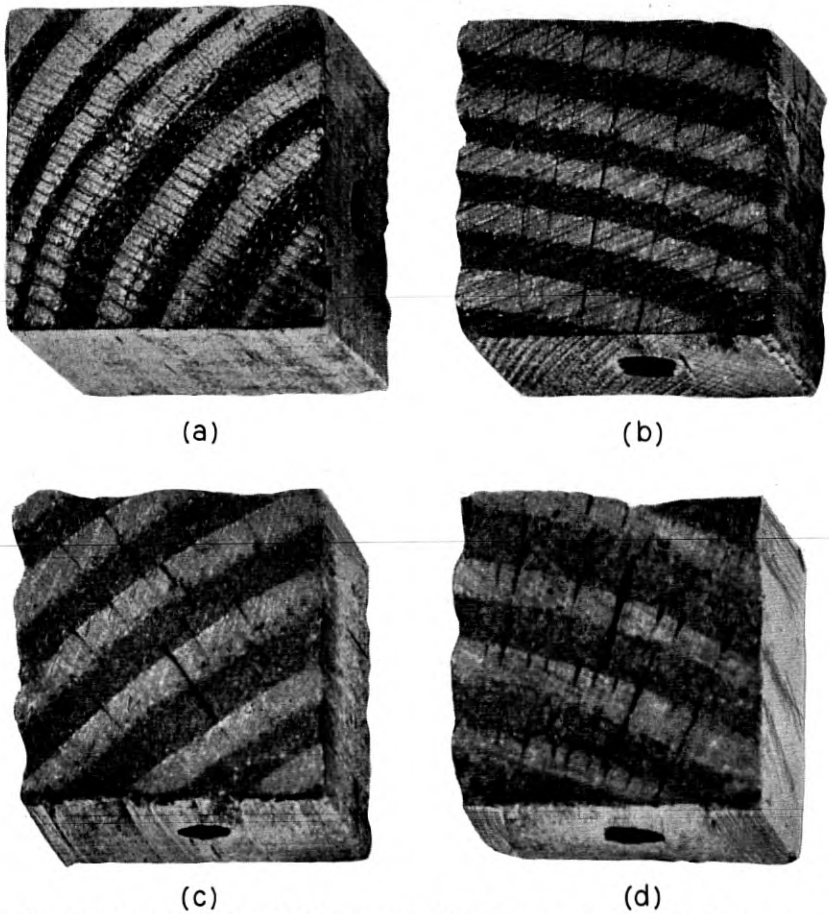

Fig. 13—Across the threshold—creosote. Test fungus, *Lentinus lepideus*; creosote concentration at treatment: (a) 11.70; (b) 6.92; (c) 5.45; and (d) 4.15 lb/cu ft.

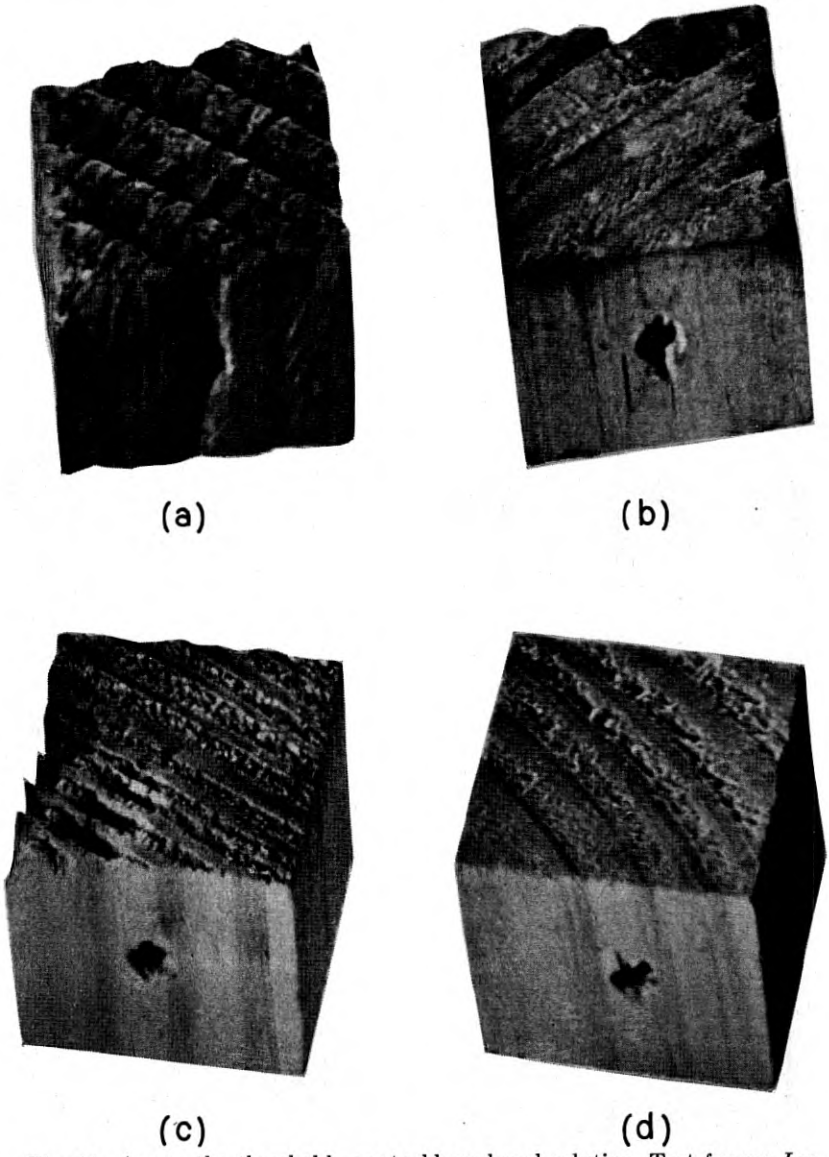

Fig. 14—Across the threshold—pentachlorophenol solution. Test fungus *Lenzites trabea*; dry penta concentration at treatment: (a) 0.052, (b) 0.099, (c) 0.193; and (d) 0.238 lb/cu ft. *Photo by A. H. Hearn.*

Relatively speaking the soil-block test procedure is much more rapid than the test plot experiments that are to be discussed next, but since the inferences with respect to retention requirements for creosote appear to be the same for the laboratory and the field tests, the former have a direct and immediate application in practical pole preservation.

BIBLIOGRAPHY

1. Alleman, G., Quantity and Character of Creosote in Well-Preserved Timbers. Circ. 98, Forest Service, U. S. Dept. Agric., May 9, 1907.
2. Alliot, H., Mèthode d'Essais des Produits Anticryptogamiques. Inst. Nat. du Bois Bull. Techn. 1, 1945.
3. Amadon, C. H., Recent Observations on the Relation between Penetration, Infection and Decay in Creosoted Southern Pine Poles in Line. Am. Wood Preservers' Assoc., Proc., **35**, pp. 187–197, 1939.
4. Baechler, R. H., Relations Between the Chemical Constitution and Toxicity of Aliphatic Compounds. Am. Wood Preservers' Assoc., Proc., **43**, pp. 94–111, 1947.
5. Baechler, R. H., The Toxicity of Preservative Oils Before and After Artificial Aging. Am. Wood Preservers' Assoc., Proc., **45**, pp. 90–95, 1949.
6. Bateman, E., Quantity and Quality of Creosote Found in Two Treated Piles After Long Service. Circ. 199, Forest Service, Forest Products Laboratory Series, U. S. Dept. Agric., May 22, 1912.
7. Bateman, E., Coal-Tar and Water-Gas Tar Creosotes: Their Properties and Methods of Testing. U. S. Dept. Agric., Bul. No. 1036, pp. 57–65, Oct. 20, 1922.
8. Bateman, E., A Theory on the Mechanism of the Protection of Wood by Preservatives. Am. Wood Preservers' Assoc., Proc., 1920, p. 251; 1921, p. 506; 1922, p. 70; 1923, p. 136; 1924, p. 33; 1925, p. 22; 1927, p. 41.
9. Bavendam, W., Die pilzwidrige Wirkung der im Holzschutz benutzten Chemikalen. Mitteilungen d. Reichsinst. f. Forst und Holzwirtschaft. No. 7, 20 pp. Aug., 1948.
10. Baxter, D. V., *Pathology in Forest Practice.* 2nd Ed. XI + 601 pp. John Wiley and Sons, New York, 1952.
11. Bertleff, V., Prüfung der Fungiziden Eigenschaften Ostrauer Steinkohlenteer Imprägnieröle und ihrer Bestandteile. Chem. Zeitung, **63**, p. 438, 1939.
12. Bescher, R. H., and others, Cooperative Creosote Tests. Am. Wood Preservers' Assoc. ,Proc., **46**, pp. 68–79, 1950.
13. Bienfait, J. L., and T. Hof, Buitenproeven met geconserveerde palen. I ste mededeling. Circ. 8 Ser. III, Conservering en Veredeling No. 3, Central Instituut voor Materiaal Onderzoek. Delft, Holland, Dec., 1948.
14. Bland, D. E., A Study of Toxicity of Australian Vertical Retort Creosote Oils to *Lentinus Lepideus* Fr., *Polystictus Versicolor* (L) Fr., and Madison 517. Australian Council Sci. and Ind. Jl. **15**, pp. 135–146, 1942.
15. Blew, J. O., Comparison of Wood Preservatives in Stake Tests. Am. Wood Preservers' Assoc., Proc., **44**, pp. 88–119, 1948.
16. Blew, J. O., C. A. Richards, and R. H. Baechler, Evaluating Wood Preservatives. For. Prod. Res. Society, Proc., **5**, pp. 230–238, (Wood-block tests, pp. 235–236), 1951.
17. Blew, J. O., Comparison of Wood Preservatives in Mississippi Post Study (1952 Progress Report). Report No. R1757, Forest Products Laboratory, U. S. Forest Service, Madison, Wis.
18. Blew, J. O., Comparison of Wood Preservatives in Stake Tests (1952 Progress Report). Report No. D1761, Forest Products Laboratory, U. S. Forest Service, Madison, Wis.
19. Boyce, J. S., *Forest Pathology.* 550 p. McGraw-Hill Book Company, Inc., New York, 1948.

20. Breazzano, A., Metodo Biologico di Controllo dei Sistemi di Preservazione, dei Legnami Adottato dall' Instituto Sperimentale delle FF. SS. (English translation: Biological Method of Testing Systems of Wood Preservation Adopted by the State Railway Experiment Institute.) Tech. Rev. Italian Railways, **4,** No. 5, pp. 3–8, Nov., 1913. Forest Products Laboratory, Madison, Wis.
21. Breazzano, A., Metodi Normali di Prova sulla Putrescibilitá dei Legnani. (English translation: Standard Methods of Testing the Putrescibility of Wood.) Extract from Report of the Ninth Meeting of the Italian Association for the Study of Building Materials; Turin; April, 1922. Forest Products Laboratory, Madison, Wis.
22. Breazzano, A., Osservazioni sul Metodo dei Blocchetti di Legno in Usa nell Analisi Tossimetrica delle Sostanze Conservatrici del Legno. Revista Technica Delle Ferrovie Italiane. **47,** No. 6, June 15, 1935. (Observations on the Wood Block Method in the Toximetric Analysis of Wood Preservatives. Excerpt from Technical Review of Italian Railways. English translation; Forest Products Laboratory.)
23. Breazzano, A., Methods for Determining the Fungicidal Power of Wood Preservatives. International Assn. for Testing Materials, Proc., London Congress, Group C—Organic Materials, Sub. group 3, pp. 484–486, April, 1937.
24. British Standard Method of Test for the Toxicity of Wood Preservatives to Fungi. British Standards Institution. British Standard No. 838, 17 pp., April, 1939.
25. Broekhuizen, S., Onderzoekingen over de Conserverende Waarde van een Aantal Houtconserveermiddelen. Rapp. Comm. Gebruikw. inh. hout. Deel II, Bijlage II, pp. 89–122, The Hague, 1937.
26. Cartwright, K. St. G., and W. P. K. Findlay, *Decay of Timber and its Prevention.* VI plus 294 p. London: His Majesty's Stationery Office, 1946. Reprinted 1948.
27. Colley, R. H., The Effect of Incipient Decay on the Mechanical Properties of Airplane Timber. (Abstract) Phytopathology, **11,** p. 45, 1921.
28. Colley, R. H., T. R. C. Wilson, and R. F. Luxford, The Effect of *Polyporus Schweinitzii* and *Trametes Pini* on the Shock-resistance, Compression Parallel to Grain Strength, and Specific Gravity of Sitka Spruce. Forest Products Laboratory, Project L-243-J1, Typewritten Report, 29 p., 32 figs., 4 plates. July 3, 1925.
29. Colley, R. H., and C. H. Amadon, Relation between Penetration and Decay in Creosoted Southern Pine Poles. Bell Sys. Tech. Jl., **15,** pp. 363–379, July, 1936.
30. Colley, R. H., Some Observations on the Selection and Use of Modern Wood Preservatives. Reports, Twenty-fourth Session, Communications Section, Association of American Railroads, October, 1947, pp. 17–25.
31. Colley, R. H., Wood preservation and Timber Economy. Forest Products Institute of Canada. Papers presented at the First Annual Convention, Ottawa, Oct. 30–31, 1950.
32. Curtin, L. P., B. L. Kline, and W. Thordarson, V—Weathering Tests on Treated Wood. Ind. and Eng. Chem. **10,** No. 12, pp. 1340–1343, Dec., 1927.
33. DIN (Deutsche Normen) DVM 2176, Blatt 1. Prüfung von Holzschutzmitteln. Mykologische Kurzprüfung (Klötzchen Verfahren). Berlin, Aug. 1939. (New Edition DIN 52176).
34. DIN DVM 52176, Blatt 2. Prüfung von Holzschutzmitteln. Bestimmung der Auslaugbarkeit. Berlin, May 1941. (Reprinted Oct., 1948).
35. Duncan, C. G., and C. A. Richards, Methods of Evaluating Wood-Preservatives: Weathered Impregnated Wood Blocks. Am. Wood Preservers' Assoc., Proc., **44,** pp. 259–264, 1948.
36. Duncan, C. G., A comparison of Two English Creosotes Produced from Coke-oven Coal Tar and Vertical-retort Coal Tar. Mss. Office Report, Division of Forest Pathology, Bur. Pl. Ind., Forest Prod. Lab., Madison, Wis., Jan. 21, 1949.
37. Duncan, C. G., and C. A. Richards, Evaluating Wood Preservatives by Soil-block Tests: 1. Effect of Carrier on Pentachlorophenol Solutions; 2. Com-

parison of a Coal Tar Creosote, a Petroleum Containing Pentachlorophenol or Copper Naphthenate and Mixtures of Them. Amer. Wood Preservers Assoc., Proc., **46,** pp. 131–145, 1950.
38. Duncan, C. G., and C. A. Richards, Evaluating Wood Preservatives by Soil-Block Tests: 3. The Effect of Mixing a Coal Tar Creosote and a Pentachlorophenol Solution with a Petroleum; a Creosote with a Coke Oven Tar or Pentachlorophenol Solution. Amer. Wood Preservers' Assoc., Proc., **47,** pp. 264–274, 1951.
39. Duncan, C. G., and C. A. Richards, Evaluating Wood Preservatives by Soil-Block Tests: 4. Creosotes. Amer. Wood Preservers' Assoc., Proc., **47,** pp. 275–287, 1951.
40. Duncan, C. G., Evaluating Wood Preservatives by Soil-Block Tests: 5. Lignite-Tar and Oil-Tar Creosotes. Amer. Wood Preservers' Assoc., Proc., **48,** 1952.
41. Duncan, C. G., Soil-Block versus Agar-Block Techniques for Evaluation of Oil-Type Preservatives: Creosote, Copper Naphthenate and Pentachlorophenol. Mss. Division of Forest Pathology, Bur. Pl. Ind., Forest Products Lab; Madison, Wis., April 29, 1952.
42. Eden, Johan, och Erik Rennerfelt, Undersökninar enligt klotsmetoden av nägra fraimpregneringsmedel. (Studies on Wood Preservatives, According to the Block Method) Meddelanden från Statens Skogsforskningsinstitut, Bd. **35,** No. 10, 1946.
43. Eden, Johan, och Erik Rennerfelt, Fält och rötkammarförsök avsedda att utröna skyddsverkan hos olika fräimpregneringsmedel. (Field and Decay-chamber Experiments to Ascertain the Protective Effect of Various Wood Preservatives.) Meddelanden från Statens Skogsforskningsinstitut. Bd. **38,** No. 4, 1949.
44. Falck, R., Die wichtigsten reinen Holzschutzmittel, die Methoden ihre Wertzahlen, Eingenschaften und Anwendung. Hausschwammforschungen, **8,** pp. 18–20, 1927.
45. Findlay, W. P. K., A Standard Laboratory Test for Wood Preservatives. British Wood Preserving Assoc., Jl., **5,** pp. 89–93, 1935.
46. Finholt, R. W., Improved Toximetric Agar-dish Test for Evaluation of Wood Preservatives. Anal. Chem., **23,** No. 7, pp. 1039–1039, July, 1951.
47. Finholt, R. W., M. Weeks, and C. Hathaway, New Theory on Wood Preservation. Ind. and Eng. Chem. **44,** No. 1, pp. 101–105, Jan., 1952.
48. Flerov, B. C., und C. A. Popov., Methode zur Untersuchung der Wirkung von antiseptische Mitteln auf holzzerstörende Pilze. Angew. Bot. **15,** pp. 386–406, 1933.
49. Frosch, C. J., V—The Correlation of Distillation Range with Viscosity of Creosote. Physics, **6,** pp. 165–170, May, 1935.
50. Gillander, H. E., C. G. King, E. O. Rhodes, and J. N. Roche, The Weathering of Creosote. Ind. and Eng. Chem., **26,** No. 2, pp. 175–183, Feb., 1934.
51. Harrow, K. M., Toxicity of Water-Soluble Wood-Preservatives to Wood-Destroying Fungi. New Zealand Jl., Sec. B., **31,** No. 5, pp. 14–19, Mar., 1950.
52. Harrow, K. M., Note on the Soil Moisture Content Used with the Leutritz Technique for Testing Toxicity of Wood Preservatives Against Fungi. New Zealand Jl. Sci. and Tech., **4,** pp. 39–40, Jan., 1951.
53. Hatfield, I., Information on Pentachlorophenol as a Wood Preserving Chemical. Am. Wood Preservers' Assoc., Proc., **40,** pp. 47–65, 1944.
54. Holzschutzmittel Prüfung und Forschung III. Wissenschaftliche Abhandlungen der Deutschen Materialprüfungsanstalten: Berlin-Dahlem. II Folge, Heft 7, 132 p. Springer-Verlag, Berlin, Göttingen, Heidelberg, 1950.
    1. Schulze, B., G. Theden u K. Starfinger, Ergebnisse einer vergleichenden Prüfung der Pilzwidrigen Wirksamkeit von Holzschutzmitteln. pp. 1–40.
    2. Becker, G., Ergebnisse einer vergleichenden Prüfung der insektötenden Wirkung von Holzschutzmitteln. II Teil. pp. 40–62.
    3. Becker, G., Prüfung der "Tropeneignung" von Holzschutzmitteln gegen Termiten. pp. 62–76.
    4. Becker, G. u B. Schulze, Laboratoriumsprüfung von Holzschutzmitteln gegen Meerwasserschädlinge. pp. 76–83.

55. Hopkins, C. Y., and B. B. Coldwell, Surface Coatings for Rotproofing Wood. Canadian Chemistry and Process Industries. N. R. C. No. 1256, Dec., 1944.
56. Howe, P. J., Weathering and Field Tests on Treated Wood. Amer. Wood Preservers' Assoc., Proc., 1928, p. 192.
57. Hubert, E. E., A study of Laboratory Methods Used in Testing the Relative Resistance of Wood to Decay. Univ. of Idaho Bulletin, **34,** No. 15, July, 1929.
58. Hubert, E, E., *An Outline of Forest Pathology.* 543 pp. John Wiley and Sons, New York, 1931.
59. Hudson, M. S., and R. H. Baechler, The Oxidation of Creosote—Its Significance in Timber Treating Operations. Amer. Wood Preservers' Assoc., Proc., **36,** pp. 74–112, 1940.
60. Hudson, M. S., Poria radiculosa, a Creosote Tolerant Organism. For. Prod. Res. Soc., Jl., **2,** No. 2, pp. 73–74, June, 1952.
61. Humphrey, C. J., and R. M. Flemming, The Toxicity to Fungi of Various Oils and Salts, Particularly Those Used in Wood Preservation. Bul. No. 227, U. S. Dept. Agric., Aug. 23, 1915.
62. Hunt, G. M., and G. A. Garratt, *Wood Preservation.* VIII, 457 p. McGraw-Hill, New York, 1938. (In course of revision.)
63. Hunt, G. M., and T. E. Snyder, An International Termite Exposure Test—Twenty-First Progress Report. Amer. Wood Preservers' Assoc., Proc., **48,** 1952.
64. Jacquiot, C., Controle de l'Efficacité des Fongicides Utilisés pour l'Imprègnation des Bois. Étude Critique de la Technique Standard Anglaise et de la Norme Allemande DIN DVM 2176. Principes pour l'Établissement d'une Norme Française. Extr. d'Ann. l'Ecole Eaux Forêts, **8,** pp. 185–206, 1942.
65. Kaufert, F. H., A Survey of Laboratory Methods Used in the Evaluation of Wood Preservatives. Report of Committee P-6, Appendix A. Amer. Wood Preservers' Assoc., Proc., **45,** pp. 55–59, 1949.
66. Leutritz, J., Jr., The Toxic Action of Various Compounds on The Fungus *Lentinus Lepideus* Fr.). Unpublished Thesis, Columbia University, Nov., 1933.
67. Leutritz, J., Jr., Laboratory Tests of Wood Preservatives. Bell Lab. Record, **16,** No. 9, pp. 324–328, May, 1938.
68. Leutritz, J., Jr., Acceleration of Toximetric Tests of Wood Preservatives by the Use of Soil as a Medium. Phytopathology, **39,** No. 10, pp. 901–903, Oct., 1939.
69. Leutritz, J., Jr., Outdoor Tests of Wood Preservatives. Bell Lab. Record, **22,** No. 4, pp. 179–182, Dec., 1943.
70. Leutritz, J., Jr., A Wood-Soil Contact Culture Technique for Laboratory Study of Wood-Destroying Fungi, Wood Decay and Wood Preservation. Bell Sys. Tech. Jl., **25,** No. 1, pp. 102–135, Jan., 1946.
71. Liese, Nowak, Peters, Rabanus, Krieg, and Pflug., Toximetrische Bestimmung von Holzkonservierungsmitteln. Angew. Botanik, pp. 484–488, Nov.–Dec., 1935.
72. Liese, J. et al., Toximetrische Bestimmung von Holzkonservierungsmitteln. Angew. Chemie, 48, Beihefte 11, 1935.
73. Loseby, P. J. A., and P. M. D. Krog, The Persistence and Termite Resistance of Creosote and Its Constituent Fractions. Jour. South African Forestry Assoc., Jl., No. 11, pp. 26–32, June, 1944.
74. Lumsden, G. Q., and A. H. Hearn, Greensalt Treatment of Poles. Amer. Wood Preservers' Assoc., Proc., **38,** pp. 349–361, 1942.
75. Lumsden, G. Q., Proving Grounds for Telephone Poles. Bell Lab. Record, **22,** pp. 12–14, Sept., 1943.
76. Lumsden, G. Q., A Quarter Century of Evaluation of Wood Preservatives in Poles and Posts at the Gulfport Test Plot. Amer. Wood Preservers' Assoc., Proc., **48,** 1952.
77. Lutz, M. L., Méthodes Permittant de Déterminer la Résistivité des Bois Bruts ou Immunisés Soumis a l'Attaque par les Champignons Lignicoles. Ann. l'Ecole Nat. Eaux Forêts 5, pp. 317–327, 1935.
78. Mahlke-Troschel-Liese, Holzkonservierung (Wood Preservation), 3rd Ed. XII + 571 p. Springerverlag, Berlin/Göttingen/Heidelberg, 1950.

79. McMahon, W., C. M. Hill, and F. C. Koch, Greensalt—A New Preservative for Wood. Amer. Wood Preservers' Assoc., Proc., **38,** pp. 334–348, 1942.
80. Martin, S. W., Characterization of Creosote Oils. Amer. Wood Preservers' Assoc., Proc., **45,** pp. 100–130, 1949.
81. Mayfield, P. B., The Toxic Elements of High Temperature Coal Tar Creosote. Amer. Wood Preservers' Assoc., Proc., **47,** pp. 62–85, 1951.
82. Narayanamurti, D., V. Ranganathan, Ragbir Singh, T. R. Chandrasekhar, and A. Banerjee, Studies on Coal Tar Creosote as a Wood Preservative, Part II. Indian Forest Bulletin, No. 144, 1948, 7 + 43 pp., Jl. of India Press, Calcutta, 1950.
83. Peters, F., W. Krieg and H. Pflug, Toximetrische Prüfung von Steinkohlenteeröl. Chem. Zeit., **61,** pp. 275–285, 1937. (English Edition, Pub. Int. Adv. Off. Wood Pres. The Hague. 1937.)
84. Preservative Treatment of Poles. (Condensed from report by American Telephone and Telegraph Company of Aug. 3, 1931.) Am. Wood Preservers' Assoc., Proc., p. 237, 1932.
85. Preservatives Committee; Report of Committee P-6. Am. Wood Preservers' Assoc., Proc., **48,** 1952.
86. Rabanus, Ad., Die Toximetrische Prüfung von Holzkonservierungsmitteln. Angew. Bot. 13, p. 352–371, 1931. (Partial translation in English. Am. Wood Preservers' Assoc., Proc., pp. 34–43, 1933.)
87. Reeve, C. S., Comment on Creosote-Permanence Toxicity Relationships. Am. Wood Preservers' Assoc., Proc., p. 78–79, 1934.
88. Rennerfelt, Erik, och Bo Starkenberg., Träskyddskomittens fält- och rötkammarförsök. (Field and Decay-Chamber Experiments with Wood Preservatives.) Meddelanden från Statens Skogsforskningsinstitut, Bd., **40,** No. 4, 1951.
89. Rhodes, E. O., J. N. Roche, and H. E. Gillander, Creosote Permanence-Toxicity Relationships. Am. Wood Preservers' Assoc., Proc., pp. 65–78, 1934.
90. Rhodes, E. O., History of Changes in Chemical Composition of Creosote. Am. Wood Preservers' Assoc., Proc., **47,** pp. 40–61, 1951.
91. Rhodes, F. H., and F. T. Gardner, Comparative Efficiencies of the Components of Creosote Oil as Preservatives for Timber. Ind. and Eng. Chem., **22,** No. 2, pp. 167–171, Feb., 1930.
92. Rhodes, F. H., and I. Erickson, Efficiencies of Tar Oil Components as Preservative for Timber. Ind. and Eng. Chem. **25,** pp. 989–991, Sept., 1933.
93. Richards, A. P., Cooperative Creosote Program; Preliminary Progress Report on Marine Exposure Panels. Am. Wood Preservers' Assoc., Proc., **48,** 1952.
94. Richards, C. A., Methods of Testing Relative Toxicity of Wood Preservatives. Am. Wood Preservers' Assoc., Proc., **19,** pp. 127–135, 1923.
95. Richards, C. A., and R. M. Addoms, Laboratory Methods for Evaluating Wood Preservatives: Preliminary Comparison of Agar and Soil Culture Techniques Using Impregnated Wood Blocks. Am. Wood Preservers' Assoc., Proc., **43,** pp. 41–56, 1947.
96. Richards, C. A., Laboratory Decay Resistance Tests—Soil-block Method. (In Cooperative Creosote Tests by R. H. Bescher et al.) Am. Wood Preservers' Assoc., Proc., **46,** pp. 71–76, 1950.
97. Scheffer, T. C., Progressive Effects of Polyporus Versicolor on the Physical and Chemical Properties of Red Gum Sapwood. U. S. Dept. Agric., Tech. Bul., No. 527, Sept., 1936.
98. Scheffer, T. C., T. R. C. Wilson, R. F. Luxford, and Carl Hartley, The Effect of Certain Heart Rot Fungi on the Specifid Gravity and Strength of Sitka Spruce and Douglas-Fir. U. S. Dept. Agric., Tech. Bul., No. 779, 24 pp., May, 1941.
99. Schmitz, H., Laboratory Methods of Testing the Toxicity of Wood Preservaservatives. Ind. and Eng. Chem., Anal. Ed., **1,** No. 7, pp. 76–79, April, 1929.
100. Schmitz, H., and Others A Suggested Toximetric Method for Wood Preservatives. Ind. and Eng. Chem., Anal. Ed., **2,** p. 361, 1930.
101. Schmitz, H., and S. J. Buckman, Toxic Action of Coal-Tar Creosote. Ind. and Eng. Chem., **24,** No. 7, pp. 772–777, 1932.
102. Schmitz, H., W. J. Buckman and H. von Schrenk, Studies of the Biological

Environment in Treated Wood in Relation to Service Life. Changes in the Character and Amount of 60/40 Creosote-Coal Tar Solution and Coal Tar and Decay Resistance of the Wood of Red Oak Crossties after Five Years Service. Am. Wood Preservers' Assoc., Proc., **37,** pp. 248–297, 1941.

103. Schmitz, H., H. von Schrenk, and A. L. Kammerer, Studies of the Biological Environment in Treated Wood in Relation to Service Life, III. Am. Wood Preservers' Assoc., Proc., **41,** pp. 153–179, 1945.
104. Schulze, B., und G. Becker, Untersuchungen über die pilzwidrige und insektentötende Wirkung von Fraktionen und Einzelstoffen des Steinkohlenteeröls. Holzforschung., **2,** No. 4, pp. 95–127, 1948.
105. Sedziak, H. P., The Wood-Block Soil Method of Accelerated Testing of Wood Preservatives. Report of Committee P-6, Appendix B. Am. Wood Preservers' Assoc., Proc., **46,** pp. 55–58, 1950.
106. Sedziak, H. P., The Evaluation of Two Modern Wood Preservatives. For. Prod. Res. Soc., Proc., 1952.
107. Snell, W. H., The Use of Wood Discs as a Substrate in Toxicity Tests of Wood Preservatives. Am. Wood Preservers' Assoc., Proc., **25,** pp. 126–129, 1929.
108. Snell, W. H., and L. B. Shipley, Creosotes—Their Toxicity, Permanence and Permanence of Toxicity. Am. Wood Preservers' Assoc., Proc., **32,** pp. 32–114, 1936.
109. Standard of N. W. M. A., Method for Testing the Preservative Property of Oil-soluble Wood-Preservatives by Using Wood Specimens Uniformly Impregnated. Nat. Wood Mfg. Assoc., M-1-51, April 27, 1951.
110. STAS 650-49, Incercarea Toxicitătii Substantelor de Impregnat Contra Ciupercilor. Comisiunea de Standardizare (Rumania) April 1, 1950.
111. Suolahti, Osmo, Über Eine das Wachstum von Fäulnispilzen Beschleunigende Chemischen Fernwirkung von Holz. Statens Tekniska Forskningsanstalt. 95 p. Helsinki, Finland, 1951.
112. Tamura, T., New Methods of Test on the Toxicity and Preservative Value of Wood Preservatives. Phytopathologische Zeitschrift, **3,** No. 4, pp. 421–437, 1931.
113. Teesdale, C. H., Volatilization of Various Fractions of Creosote After Their Injection into Wood. Circ. 188, Forest Service, Forest Products Laboratory Series, U. S. Dept. Agric., Oct. 17, 1911.
114. Tippo, O., J. M. Walter, S. J. Smucker, and W. Spackman, Jr., The Effectiveness of Certain Wood Preservatives in Preventing the Spread of Decay in Wooden Ships. Lloydia, **10,** pp. 175–208, Sept., 1947.
115. Trendelenburg, R., Über die Abkurzung der Zeitdauer von Pilzversuchen an Holz mit Hilfe der Schlagbiegeprüfung. Holz als Roh- und Werkstoff. **3.** No. 12, s. 397–407, Dec., 1940.
116. van den Berge, J. Beoordeeling van de Waarde van Fungicide Stoffen voor Houtconserveering. 183 p. N. V. Technische Boekhandel, J. Waltman, Delft, Holland, 1934.
117. van Groenou, H. Broese, Weatheringsproeven met Houtconserveermiddelen, (Weathering Tests with Wood Preservatives). Materiaalenkennis, **7,** No. 10, pp. 63–65, Oct., 1940.
118. van Groenou, H. Broese, H. W. L. Rischen and J. van den Berge, Wood Preservation During the Last 50 Years. XII + 318 p. A. W. Sijthoff, Leiden, Holland, 1951.
119. von Pechmann, H., u. O. Schaile, Über die Änderung der dynamischen Festigkeit und der Chemischen Zusammensetzung des Holzes durch den Angriff Holzzerstörender Pilze. Forstwissenschaftliches Zentralblatt, **69,** No. 8, s. 441–465, 1950.
120. Verrall, A. F., Progress Report on Tests of Soak and Brush Preservative Treatments for Use on Wood off the Ground. Southern Lumberman, Aug. 1, 1946.
121. Verrall, A. F., Decay Protection for Exterior Woodwork. Southern Lumberman, June 15, 1949.
122. Warner, R. W., and R. L. Krause, Agar-Block and Soil-Block Methods for Testing Wood Preservatives. Ind. and Eng. Chem., **43,** No. 5, pp. 1102–1107, May, 1951.

123. Waterman, R. E., Forecasting the Behavior of Wood Preservatives. Bell Lab. Record, **11,** No. 3, pp. 67–72, Nov., 1932.
124. Waterman, R. E., F. C. Koch, and W. McMahon, Chemical Studies of Wood Preservation. III. Analysis of Preserved Timber. Ind. and Eng. Chem., Anal. Ed., **6,** No. 6, pp. 409–413, Nov. 15, 1934.
125. Waterman, R. E. and R. R. Williams, Chemical Studies of Wood Preservation. IV. Small Sapling Method of Evaluating Wood Preservatives. Ind. and Eng. Chem., Anal. Ed., **6,** No. 6, pp. 413–418, Nov. 15, 1934.
126. Waterman, R. E., J. Leutritz, and C. M. Hill, A Laboratory Evaluation of Wood Preservatives. Bell System Tech. Jl., **16,** pp. 194–211, April, 1937.
127. Waterman, R. E., J. Leutritz and C. M. Hill, Chemical Studies of Wood Preservation: The Wood-block Method of Toxicity Assay. Ind. and Eng. Chem., Anal. Ed., **10,** No. 6, pp. 306–314, June, 1938.
128. Weiss, J. M., The Antiseptic Effect of Creosote Oil and Other Oils Used for Preserving Timber. Soc. Chem. Ind. Jl., **30,** No. 23, Dec. 15, 1911.
129. Williams, R. R., Chemical Studies of Wood Preservation. I. The Problem and Plan of Attack. Ind. and Eng. Chem., Anal. Ed., **6,** No. 5, pp. 308–310, Sept. 15, 1934.

Motion of Gaseous Ions in Strong Electric Fields

By GREGORY H. WANNIER

(Manuscript received August 20, 1952)

This paper applies the Boltzmann method of gaseous kinetics to the problem of charged particles moving through a gas under the influence of a static, uniform electric field. The particle density is assumed to be vanishing low, and the ion-atom collisions are assumed elastic, but the field is taken to be strong; that is the energy which it imparts to the charges is not assumed negligible in comparison to thermal energy. In Part I, the formal framework of such a theory is built up; the motion in the field is describable by the drift velocity concept, and the smoothing out of density variations as an anisotropic diffusion process. In Part II, the "high field" case is treated in detail; this is the case, for which thermal motion of the gas molecules is negligible; the equation is solved completely for the case that the mean free time between collisions may be treated as independent of speed; complete solutions are also presented for extreme mass ratios of the ions and the molecules; special attention is given to the case of equal masses, which has to be handled by numerical methods. In Part III, information about the "intermediate field" case is collected; with the help of a convolution theorem the case of constant mean free time is solved; beyond this, only the case of small ion mass (electrons) is available. In Part IV, the diffusion process, whose existence was proved in Part I, is pushed through to numerical results. Part V discusses the scope of the results achieved and demonstrates the possibility of extending them semiquantitatively beyond their original range.

Part I — General Theory of Strong Field Motion

1a. Qualitative discussion

It is well known that if we consider a mixture of gases under no external forces the steady velocity distribution which establishes itself in the mixture does not depend on the interactions between the gas molecules; we have always a Maxwellian distribution for each species

with a temperature common to all. This result arises from statistical mechanics; the derivation of it is simple and requires few assumptions, yet it enjoys a wide degree of generality. As soon, however, as a non-equilibrium feature is imposed upon the system this simplicity vanishes, and the subject acquires ramifications. Results must now be derived by kinetic theory. The amount of labor required increases, while, at the same time, the result achieved becomes less general.

A mixture of charged particles (ions or electrons; in the following often simply referred to as ions) and gas molecules can in principle never be in equilibrium since the presence of the former in itself represents an instability. However, one might expect, that equilibrium exists in a restricted sense, for instance, as regards motion. Even this is rarely the case under actual conditions of observation. The non-equilibrium features of greatest importance for analyzing ion motion are a constant force (electric field) acting upon one species but not the other (mobility theory), and a concentration gradient for one particular species (diffusion theory). It is the purpose of this paper to apply kinetic theory to these problems, and to compute with its help the most important properties which such a gas of charged particles possesses. The work will be distinguished from similar ones in that the electric field will not be supposed weak; velocity distributions which have no resemblance to the Maxwellian distribution will thus make their appearance. Furthermore, the mass of the charged particles will not be assumed small, which means the possibility of getting results for gaseous ions as well as electrons. Magnetic fields, plasma and A.C. phenomena will, however, be excluded. The quantities of interest under those conditions are the drift velocity of the ions, their energy, energy partition and diffusion constants. These quantities will be calculated by assuming plausible mechanical models. The work just outlined has been published in part in abbreviated form in the Physical Review;[1] the exposition to follow will, however, proceed independently from these articles.

Much of the work which concerns itself with transport processes in gases makes use of perturbation theory. This method permits us to predict the behavior of a gaseous assembly under an electric field or a concentration gradient in the limit when the field or the gradient are vanishingly small. The result of so perturbing a Maxwellian distribution can be expressed through certain constants, such as the mobility or the diffusion coefficient, which involve the Maxwellian distribution *and* the internal interactions, but not the perturbation itself.

---

[1] Wannier, G. H., Phys. Rev., **83**, p. 281, 1951 and Phys. Rev., **87**, p. 795, 1952.

The limits of such a procedure can easily be estimated. In the case of an electric field, perturbation techniques apply if the kinetic energy acquired by the ion from the field is small compared to thermal energy. This means at least that the energy acquired in one mean free path be small, i.e.,

$$eE\lambda \ll kT$$

where e is the electronic charge, E the electric field, k Boltzmann's constant, T the absolute temperature, and λ the mean free path. Actually the situation is not even that favorable. If the mass of the ions and the molecules is very different, the energy transferred upon collision is small, and hence the ions possess the ability to store the acquired energy through many collisions; for this reason, the inequality reads more properly

$$\left(\frac{M}{m}+\frac{m}{M}\right) eE\lambda \ll kT,$$

where m is the mass of the ions and M the mass of the gas molecules. After some substitutions this estimate becomes

$$\left(\frac{M}{m}+\frac{m}{M}\right) eE \ll p\sigma, \tag{1}$$

where p is the true gas pressure and σ the collision cross-section. Taking as an example an ion travelling in the parent gas we find

$$\frac{E}{p} \ll 2\,\frac{\sigma}{e} \sim 2\cdot\frac{4\pi\cdot 10^{-16}}{5\cdot 10^{-10}} = 5\cdot 10^{-6} \text{ e.s.u.}$$

or in commonly employed units

$$\frac{E}{p} \ll 2 \text{ volt/cm (mm Hg)}.$$

It is clear that this limit is often surpassed in experimental situations.

The cases in which the limit (1) is applicable are of no further interest here because they are well covered in the literature.[2] A field will be called "low" when it satisfies the criterion (1) and "high" when the inequality is reversed. It is important to notice that a fixed field at a fixed gas density may shift from "low" to "high" through a drop in temperature.

All calculations to follow will contain the assumption of "low ion concentration" which is often made in studies of this sort. It means that

[2] See for instance: A. M. Tyndall, *The Mobility of Positive Ions in Gases*, Cambridge University Press, 1938, Chapter IV.

all effects which ions exert upon each other are neglected. The equation for the distribution function of ionic velocities is then linear instead of quadratic. It is clear that this simplification presents great advantages from the point of view of calculation.

In deriving a criterion for the validity of this assumption we must distinguish two types of effects of the ions upon each other. The first is the space charge effect. In this effect the ions at large distances make the major contribution. Its magnitude depends on apparatus dimensions. The criterion for no space charge distortion of the field E is

$$n \ll \frac{E}{4\pi e X} \tag{2}$$

where n is the number density of the ions and X a suitable length chosen from apparatus dimensions. Inequality (2) is quite stringent because it predicts field distortions at values of n of the order of 10^8 cm^{-3}. This is the value at which it will become impossible, or at least difficult, to make significant experimental measurements. But from the point of view of theory this criterion is not relevant. Space charge does not change the character of the velocity distribution of the ions because the type of ion-ion interaction producing the space charge field is long range and creates only a smooth modification of the electric field which we may presume to have been included in the original field. What we are concerned with here are ion-ion interactions which have a random character and thus are apt to upset a velocity distribution derived from the "low concentration" theory. From this point of view neighboring ions are most effective because their relative location fluctuates rapidly, and hence, the Coulomb force between them will induce mutual scattering. The magnitude of this force is of the order $e^2n^{2/3}$ where n is the number density of the ions. It is known from theory[3] that the effect of a Coulomb force is preferably not represented by discrete "collisons" but by a continuous bending of the entire path. Thus we come to the conclusion that random ion-ion forces have no effect if the force given above cannot produce a significant deflection in one mean free path. This means

$$e^2n^{2/3}\lambda \ll \text{mean ion energy} \tag{3}$$

According to whether we are in the high or low field region we get different criteria from this. At low field the thermal energy predominates and we get

$$e^2n^{2/3} \ll p\sigma \tag{3a}$$

[3] Mott and Massey, *The Theory of Atomic Collision*, Oxford Press 1933, Chapter III.

At high field the "field" energy predominates and we get

$$e^2 n^{2/3} \ll eE\left(\frac{M}{m} + \frac{m}{M}\right) \tag{3b}$$

A rough evaluation of inequality (3a) for one mm Hg pressure gives

$$n^{2/3} \ll \frac{10^3 \cdot 4\pi \cdot 10^{-16}}{25 \cdot 10^{-20}} = \tfrac{1}{2} \cdot 10^7 \text{ cm}^{-2}$$

$$n \ll 10^{10} \text{ particles/cm}^3$$

This corresponds to a current of about 10^{15} particles/cm^2 sec or 200 μamps/cm^2. At lower pressure the criterion becomes more stringent. Equation (3b) gives similar results.

It is appropriate to survey at this point the past theoretical work treating the "low concentration" theory of ionic motion for arbitrary fields. A rather complete body of work exists for electrons where the following three assumptions seem appropriate: (a) that the mass of an "ion" is very small compared to the mass of a molecule, (b) that the total kinetic energy is conserved in each encounter, and (c) that the angular distribution is isotropic in the center of mass system.

These three assumptions lead to a distribution law given by Chapman and Cowling.[4] The law has considerable flexibility because it permits the substitution of an arbitrary relationship connecting mean free path and speed of encounter. In addition it contains no assumption as to whether we have low or high field. A more specialized and explicit distribution law is obtained if we assume in addition: (d) that the collision cross-section is independent of the speed of encounter (hard sphere approximation); and (e) that we deal with the high field case only. The special law resulting in this case is the distribution law of Druyvesteyn.

If an improvement over the Chapman-Cowling distribution for electrons is desired account should be taken of inelastic collisions, that is assumption (b) should be discarded. Work in that direction has been carried out by Smit, Allen[5] and others.

The assumption to be discarded first in theory of ionic motion is, of course, assumption (a). In order to understand what this implies we must understand what advantages assumption (a) has in a calculation. In the limit when the ionic mass is very small the encounters with gas

[4] Chapman-Cowling, *The Mathematical Theory of Non-uniform Gases*, Cambridge University Press 1939, Sections 18.7–18.74. Other references are found there.

[5] Smit, J. A., Physica, **3**, p. 543, 1937 and H. W. Allen, Phys. Rev., **50**, p. 707, 1937.

molecules become such that momentum is lost quickly, but energy is accumulated in the form of random motion. As a result of this we end up with a distribution function which is very nearly spherically symmetrical in velocity space. Such a situation permits obvious procedures through which the entire calculation is simplified. These procedures will not longer be available when assumption (a) is dropped.

Knowledge concerning the structure of the velocity distribution function for gaseous ions is practically nonexistent at this time. Hershey, who deals with the motion of ions in the high field case, simply substitutes for it a Maxwellian distribution with an unknown offset of the origin and unknown temperature parameter,[6] shown in Fig. 1(a). He then computes these two parameters by applying the laws of conservation of momentum and kinetic energy. It is to be expected that this procedure should give reasonable values for the mobility and the mean energy of the ion; indeed, if we consider the polarization force only, we get *exactly* the right values; the reason for this is that one may evaluate velocity averages for inverse fifth power forces ignoring the distribution function[7] and that he did this in effect for the drift velocity and the

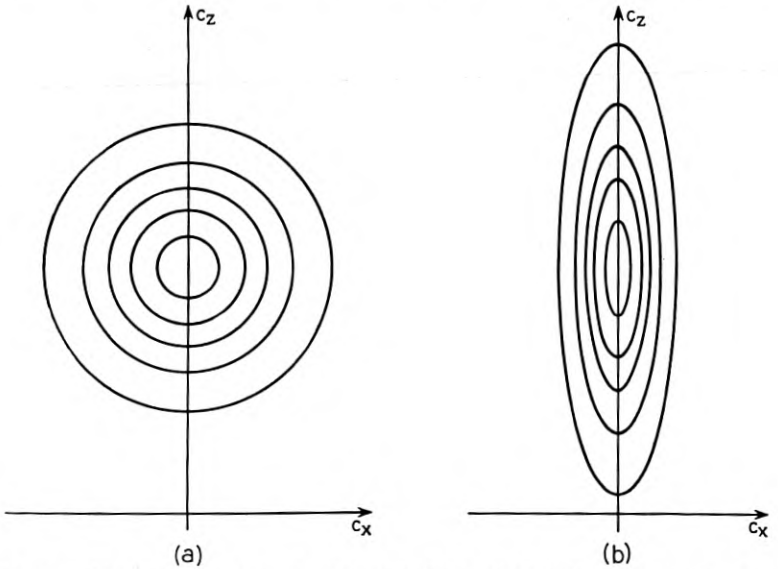

Fig. 1 — Simplified pictures for the high field velocity distribution of gaseous ions. (a) Hershey's assumption. (b) Modification with correct second moments.

[6] Hershey, A. V., Phys. Rev., **56**, p. 916, 1939.

[7] This will be shown in Section IIB.

total energy. In order to test whether an offset Maxwellian distribution is a satisfactory approximation we have to go one step further and examine the partition of the energy among the three degrees of freedom. There we find Hershey's distribution in error, for he assumes equipartition for the random motion, while, in reality, the random energy parallel to the field is much higher than at right angles,[7] giving the distribution a decided "ridge" structure. This discrepancy could be taken into account by the use of an elliptically distorted Maxwellian distribution, shown in Fig. 1(b), and this may prove to be convenient in some applications.

For a detailed knowledge of the distribution function it is necessary to specify the interaction between an ion and a molecule. This interaction can be, broadly speaking, summarized under three headings: (a) the polarization force, (b) the short distance repulsion, and (c) symmetry effects. The polarization force arises because an ion, when passing close to a molecule, induces on it a dipole moment; this moment is then attracted by the charge of the ion. The attractive force F resulting from this is

$$F = \frac{2e^2P}{\rho^5} \tag{4}$$

where P is the polarizability of a gas molecule and e the charge of the ion. The force varies inversely as the fifth power of the distance ρ; for such a force the cross section σ varies inversely as the speed of encounter γ. Whenever the cross section shows this type of variation it is advantageous to define a mean free time τ rather than a mean free path λ. The formula is

$$\tau = \frac{1}{N\sigma\gamma} \tag{5}$$

There is a standard difficulty which arises when one tries to make use of a formula of the type (5). For most force laws, a total cross-section σ cannot be defined; a differential cross section per unit solid angle always exists, but it becomes infinite in the forward direction because of small deflections suffered by particles passing by each other at a large distance. Thus equation (5) is, strictly speaking, meaningless. This is actually never a difficulty in the computation of a physical quantity. However, equation (5) is convenient for order-of-magnitude thinking and the question arises how it can be reasonably interpreted. The general method of salvaging (5) — excluding a small forward cone from consideration — is of little value for this purpose. An analysis of the inverse fourth power

attractive potential shows a better way out. The potential gives rise to two kinds of orbits; orbits of large angular momentum which look somewhat like hyperbolas, shown in Fig. 2(a), and orbits of small angular momentum for which the particles are "sucked" toward each other in a spiralling movement until a repulsive force reverses the trend, as shown in Fig. 2(b).[8] A calculation of Hassé[9] shows that the latter type of motion is much more efficient in scattering than the former and one gets therefor a picture which is semiquantitatively correct if one substitutes into (5) the cross-section for spiralling collisions and assumes isotropic scattering.[10] This cross section equals

$$\sigma = 2\pi \sqrt{\frac{1}{m} + \frac{1}{M}} \frac{\sqrt{P}\, e}{\gamma} \tag{6}$$

A numerical estimate of the cross section (6) automatically leads one to compare it with the short distance repulsion familiar from the kinetic theory of gases. The two are of the same order, but for the usual gaseous speeds (which enter into (6) through γ) and small molecules the cross

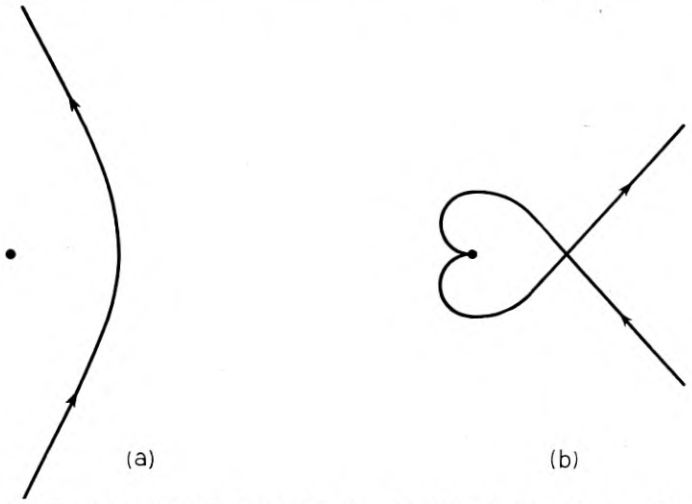

Fig. 2 — Sample orbits (schematic) showing the motion of a particle in the polarization force field. (a) Hyperbolic orbit (large angular momentum). (b) Spiralling orbit (small angular momentum).

[8] There are quantum mechanical analogues to these classical ideas; they should lead to practically identical answers unless the angular momentum quantum number is small.

[9] Hassé, H. R., Phil. Mag., **1**, p. 139, 1926.

[10] This will be discussed more fully in Section IIIB.

section (6) is bigger. This situation is accentuated in an actual scattering calculation which shows an attractive force to be generally more efficient than a repulsive force of equal range.

A detailed numerical discussion of these questions is found in Massey and Mohr[11] for the case of He^+ ions moving through He gas. Their interest is in the low field mobility. They show that for this problem the repulsive force makes so little difference that it could be neglected entirely without much affecting the results. It does finally come out that the polarization force gives a mobility which is too big by a factor of two. But the additional scattering is due to an effect which we listed above under (c): namely a resonance attraction between the He atom and the He^+ ion for which the cross section is abnormally large. It should be possible to eliminate this effect by increasing the cross section (6) until it masks even this special effect. Lowering the field is not sufficient to achieve this because of the temperature motion; it would be necessary in addition to reduce the absolute temperature by a sizeable factor and so to decrease the value of γ in (6). Thus we are led to the prediction that if the temperature of He is reduced the mobility of He^+ ions in He should gradually rise from its "anomalous" value of 12 cm^2/volt sec to the "normal" value of 22 cm^2/volt sec, which one gets by taking account of polarization forces only.

IB. GLOSSARY

The complicated appearance of equations in gaseous kinetics suggests special care in the use of symbols and a convenient arrangement for the reader to find their meaning. It is hoped that the glossary to follow will accomplish this purpose. It explains all symbols except those used at one location only.

Generally, Latin capital letters will refer to the gas molecules and Latin lower case letters to the ions; Greek letters will have no special relationship; exceptions will be made for generally recognized symbols. Thus we define

E, $\mathbf{E}$ = electric field.
x, y = cartesian coordinates at right angles to the field direction.
z = cartesian coordinate along the field direction.
$\mathbf{r}$ = position vector with components x, y, z.
t = time.
m = ionic mass.
e = ionic charge.

[11] Massey, H. S. W., C. B. O. Mohr, Proc. Roy. Soc., **144A**, p. 554, 1931.

a, $\mathbf{a} = \dfrac{e\mathbf{E}}{m}$ = ionic acceleration.

b = impact parameter.

b_{lim} = limiting value of the impact parameter separating hyperbolic and spiralling orbits (equation (125)).

c, $\mathbf{c}$, $\mathbf{c}'$, c_x, c_f, c_i = various ionic velocities or components.

e_x, e_y, e_z = energies of ionic motion (or "high field" parts thereof) along x, y, z.

e_z^* = random part of the above energy,

$\mathbf{j}_t$ = total particle current density of the ions (may be a function of $\mathbf{r}$ and t).

$\mathbf{j}$ = partial current density induced by the concentration gradient; see equation (22).

k = Boltzmann's constant (only when followed directly by T).

k, $\mathbf{k}$ = relative concentration gradient of the ions (a different use is made of k in Section IIE).

n = number density of the ions (may be a function of $\mathbf{r}$ and t).

p, q, r, s = undetermined constants; used three times independently (equations (82), (90) and (160)).

$p^{(0)}$, $p^{(1)}$ = various approximations to these numbers.

$\mathbf{u}$, $\mathbf{u}'$, $\mathbf{v}$ = ionic velocities.

$\mathbf{w}$ = ionic velocity rendered dimensionless (see eq. (75) or (85)).

g = the inner integral in the double integral eq. (69).

$\mathbf{C}$, $\mathbf{C}'$ = molecular velocities.

$\mathfrak{D}$ = ionic diffusion tensor.

$D_{||}$, $D_\perp$ = components of above tensor parallel and perpendicular to the field.

M = molecular mass.

N = number density of the molecules.

P = molecular polarizability.

T = gas temperature.

$\mathbf{U}$, $\mathbf{U}'$ = molecular velocities.

X, Y = left and right hand sides of equation (111a).

$\beta = \dfrac{1}{2kT}$ = temperature parameter (A different use of β is made in Section IIIB where it is the relative impact parameter b/b_{lim}).

γ, γ', η, η' = relative velocities of ion and molecule.

ξ, η, ζ = cartesian coordinates oriented on $\mathbf{c}$.

ρ = distance between ion and molecule.

σ = collision cross section of ions and molecules (may be a function of γ).

$\lambda = \dfrac{1}{N\sigma}$ = mean free path of ion between collisions with molecules (may be a function of γ).

$\tau = \dfrac{1}{N\sigma\gamma}$ = mean free time for the ion between collisions with molecules.

τ_s = same parameter for "spiralling" collisions.

$\alpha = \dfrac{d\ell n\tau(\gamma)}{d\ell n\gamma} + 1$. It is assumed constant in Section ID.

χ, χ_c, χ_u = angle of scattering of ion and molecule in the center of mass system.

κ = angle of scattering of the ion by a molecule in the laboratory system.

ϵ = scattering azimuth of ion and molecule in the center of mass system.

ω = scattering azimuth in the laboratory system (azimuth of the initial ion velocity about the final ion velocity).

ϑ, ϑ' = angle between velocity vector and field direction.

$\psi, \varphi, \theta, \phi, \delta$ = other angles (these angles are defined on spherical triangles which are exhibited in Figs. 8 and 15.

$d(\mathbf{c}, \mathbf{r}, t)$ = density function of ions in phase space.

$m(\mathbf{c}) = \left(\dfrac{\beta m}{\pi}\right)^{3/2} \exp(-\beta mc^2)$ = Maxwellian velocity distribution function for ionic mass.

$M(\mathbf{C}) = \left(\dfrac{\beta M}{\pi}\right)^{3/2} \exp(-\beta MC^2)$ = Maxwellian velocity distribution function for molecular mass.

$h(\mathbf{c})$ = "high field" distribution function of the ions for the case that the spatial distribution is uniform (the exact meaning of this term is to be explained in the text).

$f(\mathbf{c})$ = true velocity distribution of the ions for the case that the spatial distribution is uniform.

$g(\mathbf{c})$ = correction to $f(\mathbf{c})$ or $h(\mathbf{c})$ for the case of a constant relative concentration gradient $\mathbf{k}$.

$\delta(\mathbf{c})$ = vectorial δ-function in velocity space.

$Ei(x) = \displaystyle\int_x^\infty \frac{e^{-\xi}}{\xi}\, d\xi$ (suppression of two minus signs).

$I_0(x)$ = modified Bessel function of order 0.

$K_0(x), K_1(x)$ = Modified Hankel functions of order 0, 1. (Alteration of Macdonald function by a factor $\dfrac{2}{\pi}$).

$P_\nu(x)$ = Legendre Polynomials.

$h_\nu(c)$, $g_\nu(c)$ = expansion coefficients which result when $h(\mathbf{c})$, $g(\mathbf{c})$ are expanded in Legendre Polynomials about the field direction.
$I_{s,\nu}(\chi)$ = A set of functions of the scattering angle defined in (48).
$\langle \quad \rangle$ = the quantity in pointed brackets is to be averaged.
$\langle s, \nu \rangle$ = abbreviation for $\langle w^s P_\nu (\cos \vartheta) \rangle$; the average is taken over $h(\mathbf{w})$.
$\{s, \nu\}$ = A normalized correction to $\langle s, \nu \rangle$ contributed by $g(\mathbf{w})$; see equation (155).

A special convention will be adopted to distinguish velocities before and after a collision:

$\mathbf{c}'$, $\mathbf{C}'$ = velocities before the collision.

$\mathbf{c}$, $\mathbf{C}$ = velocities after the collision.

When used in this fashion the twelve components of the four vectors above satisfy the four identities:

$$m\mathbf{c}' + M\mathbf{C}' = m\mathbf{c} + M\mathbf{C} \tag{7}$$

$$mc'^2 + MC'^2 = mc^2 + MC^2 \tag{8}$$

The same convention is to apply to other vector quadruples, such as

$$\mathbf{u}, \mathbf{U}, \mathbf{u}', \mathbf{U}'$$

For the velocities in the center of mass system we use

$\boldsymbol{\gamma}' = \mathbf{c}' - \mathbf{C}'$ = relative velocity before the collision.

$\boldsymbol{\gamma} = \mathbf{c} - \mathbf{C}$ = relative velocity after the collision.

In consequence of (7) and (8) the γ's obey the relation

$$\gamma'^2 = \gamma^2 \tag{9}$$

The multiple integrations occurring in the theory are of the following two types. Either they are over the three components of a velocity in a Cartesian velocity space; we shall denote such integrations by $d\mathbf{c}$, $d\mathbf{u}$, $d\mathbf{U}'$, etc. Or they are proper "collision" integrations which classically have the form

$$\gamma\, b\, db\, d\epsilon$$

where b is an impact parameter and ϵ an azimuth. In most cases these integrals depend on extraneous factors for their convergence but this fact is usually disregarded for convenience; we shall follow this habit by

writing the above differential in the form

$$\frac{1}{4\pi}\gamma\sigma(\gamma)\Pi(\chi)\sin\chi\,d\chi\,d\epsilon = \frac{1}{4\pi}\gamma\sigma(\gamma)\Pi(\chi)\,d\Omega_\gamma$$

Here $d\Omega$ is meant to represent an integration over a solid angle and the subscript γ, that it is over the solid angle swept out by the vector $\boldsymbol{\gamma}$. The notation makes use of the fact that the choice of the polar axis is arbitrary in such an integration. The function $\Pi(\chi)$ is the probability of scattering which equals unity for isotropic scattering. In cases where small angle scattering is infinitely probable the above expression becomes meaningless, strictly speaking, $\Pi(\chi)$ being a δ-function at $\chi = 0$ and σ being infinite. However if a quantity such as $1 - \cos\chi$ is multiplied in, which removes the δ-function then the integration gives a finite number which may be denoted by $\langle\sigma\cdot(1 - \cos\chi)\rangle$.

IC. FORMAL SURVEY OF THE THEORY

Under the assumptions stated in Part IA we may describe the motion of ions in a gas by their density in phase space. The change in time of this function is described by a Boltzmann equation[12] which, in our notation, reads

$$\frac{\partial d(\mathbf{c}, \mathbf{r}, t)}{\partial t} + \mathbf{a}\cdot\frac{\partial d(\mathbf{c}, \mathbf{r}, t)}{\partial \mathbf{c}} + \mathbf{c}\cdot\frac{\partial d(\mathbf{c}, \mathbf{r}, t)}{\partial \mathbf{r}}$$
$$= \frac{N}{4\pi}\iint \{M(\mathbf{C}')d(\mathbf{c}, \mathbf{r}, t) - M(\mathbf{C})d(\mathbf{c}, \mathbf{r}, t)\}\gamma\sigma(\gamma)\Pi(\chi)\,d\Omega_{\gamma'}\,d\mathbf{C} \tag{10}$$

The equation is linear in the unknown function $d(\mathbf{c}, \mathbf{r}, t)$; this is due to neglect of ion-ion collisions, as stated earlier. The negative term on the right hand side actually reduces to a known function of $\mathbf{c}$ multiplying $d(\mathbf{c}, \mathbf{r}, t)$. The positive term is a genuine integral term; it has been shown by Pidduck[13] that the number of integrations in it can be brought down from five to three; this reduction will not be made use of in the following.

If there were no terms on the left hand side of equation (10) then the solution of it would have the equilibrium form

$$d(\mathbf{c}, \mathbf{r}, t) = nm(\mathbf{c}) \tag{11}$$

where n is a constant. This result is a direct consequence of equation (8) which makes the curly bracket in (10) vanish identically when Maxwellian functions are inserted.

---

[12] See Reference 4.

[13] Pidduck, F. B., Proc. Lond. Math. Soc., **15**, p. 89, 1915.

The function $m(\mathbf{c})$ is not the solution of our problem because of the presence of the second and third term on the left which arise from an electric field and a density variation respectively. These disturbances will be assumed of different relative importance. The density variation will be assumed sufficiently small so that the third term can be treated by perturbation theory; the field term, on the other hand will be taken so large that the equilibrium distribution (11) no longer represents a first approximation to the solution. In consequence, the equation is solved in two stages. In the first, only the second term on the left is retained, and the resultant equation is treated rigorously; in the second, the full equation (10) is used, but the new terms are taken as perturbations.

The first stage describes those properties of the ion gas which it possesses when assumed of uniform density. Since the field is also assumed uniform and not changing in time, the dependence on $\mathbf{r}$ and t drops out. We may then write

$$d(\mathbf{c}, \mathbf{r}, t) = nf(\mathbf{c}) \tag{12}$$

where n is a constant and $f(\mathbf{c})$ is a velocity distribution function. The equation for f reads

$$\mathbf{a}\cdot\frac{\partial f}{\partial \mathbf{c}} = \frac{N}{4\pi}\iint \{M(\mathbf{C}')f(\mathbf{c}') - M(\mathbf{C})f(\mathbf{c})\}\gamma\sigma(\gamma)\Pi(\chi)\, d\Omega_{\gamma'}\, d\mathbf{C} \tag{13}$$

with the side condition

$$\int f(\mathbf{c})d\mathbf{c} = 1 \tag{14}$$

As a result of solving (13) we shall obtain the distribution function $f(\mathbf{c})$ as a function of the electric field contained in a. This distribution differs essentially from the Maxwellian one in that it is not symmetric about the origin. The vectorial mean of the velocity is therefore not zero

$$\langle \mathbf{c} \rangle = \int f(\mathbf{c})\mathbf{c}\, d\mathbf{c} \neq 0 \tag{15}$$

This is the drift velocity of the ion in the field which is reached as a compromise between the acceleration $\mathbf{a}$ and the frictional losses caused by the ion-atom collisions. From the structure of equation (13) there is one general prediction that can be made concerning this velocity, namely that it depends on the gas density and the field only through a/N; this is the well known E/p_0 of the experimental analysis. This type of

dependence does not only hold for $\langle \mathbf{c} \rangle$, but for all averages derivable from $f(\mathbf{c})$, notably the mean energy.

A more important formal prediction can be made about the second stage of the contemplated calculation. For it will be shown now that the diffusion concept is still applicable in the presence of a strong electric field. It is true, that if we have a variable density in space the primary motion observed is not diffusive but a displacement of the entire density pattern with the drift velocity $\langle \mathbf{c} \rangle$. However, once this dominant component is subtracted out, then a supplementary current proportional to the density gradient is identified. The constant of proportionality is anisotropic, that is, we have a diffusion tensor rather than a diffusion coefficient. The tensor is axially symmetric about the field direction, yielding a longitudinal and a transverse diffusion coefficient.

To demonstrate these features it is convenient to assume a special type of variation of ion density in space. As we shall see the velocity distribution is primarily sensitive to the relative density gradient $\mathbf{k}$; we shall therefore assume it to be a constant. In other words we set

$$n(\mathbf{r}, t) = n_0 \exp [\mathbf{k} \cdot (\mathbf{r} - \langle \mathbf{c} \rangle t)] \tag{16}$$

The relation can of course not hold everywhere since n increases beyond all bounds in one direction, but we must remember here that we are doing perturbation theory, that is $\mathbf{k}$ is assumed small. The inconsistencies in the assumption (16) can then be pushed as far away as we please. Furthermore there is no inconsistency at all in the half space where n decreases. It is to be observed that according to the assumption (16) the spatial distribution is moving unchanged through space with the drift velocity $\langle \mathbf{c} \rangle$. This seems to contradict the program of finding the effect of diffusion upon $n(\mathbf{r}, t)$. However, we follow in this simply conventional steady state computational methods in which a gradient is assumed maintained from an infinitely strong source; the modification appears then as a change in the velocity distribution function, which, in turn, yields a steady diffusion current. We set therefore

$$d(\mathbf{c}, \mathbf{r}, t) = n(\mathbf{r}, t)[f(\mathbf{c}) + g(\mathbf{c})] \tag{17}$$

where $g(\mathbf{c})$ is a correction to the solution of (13) which arises from the assumption (16). It follows from the definition of $n(\mathbf{r}, t)$ and (14) that

$$\int g(\mathbf{c}) d\mathbf{c} = 0 \tag{18}$$

The consistency of the assumptions (16) and (17) with equation (10) becomes evident when they are substituted into this equation. We

find, after simplification with (13)

$$\mathbf{a}\cdot\frac{\partial g(\mathbf{c})}{\partial \mathbf{c}} + \frac{N}{4\pi}\iint \{M(\mathbf{C})g(\mathbf{c}) - M(\mathbf{C}')g(\mathbf{c}')\}\gamma\sigma(\gamma)\Pi(\chi)\, d\Omega_{\gamma'}\, d\mathbf{C} = -\mathbf{k}\cdot(\mathbf{c} - \langle\mathbf{c}\rangle)\{f(\mathbf{c}) + g(\mathbf{c})\} \tag{19}$$

This is an equation in velocity space only, $\mathbf{r}$ and t having disappeared completely; this justifies the assumptions. In solving the equation we observe that our interest is only in diffusion, that is, the current resulting from a concentration gradient when treated in first order perturbation. In this case both $\mathbf{k}$ and $g(\mathbf{c})$ are to be treated as small and their product in (19) is to be neglected. The equation then becomes

$$\mathbf{a}\cdot\frac{\partial g(\mathbf{c})}{\partial \mathbf{c}} + \frac{N}{4\pi}\iint \{M(\mathbf{C})g(\mathbf{c}) - M(\mathbf{C}')g(\mathbf{c}')\}\gamma\sigma(\gamma)\Pi(\chi)\, d\Omega_{\gamma'}\, d\mathbf{C} = -\mathbf{k}\cdot(\mathbf{c} - \langle\mathbf{c}\rangle)f(\mathbf{c}) \tag{20}$$

The homogeneous prototype of this inhomogeneous equation is (13); an arbitrary amount of $f(\mathbf{c})$ could thus be added to a particular solution of (20) were it not for the orthogonality condition (18) which makes the solution definite.

The existence of the diffusion phenomenon follows easily from equation (20). The total current $\mathbf{j}_t$ is given by

$$\mathbf{j}_t(\mathbf{r}, t) = \int d(\mathbf{c}, \mathbf{r}, t)\mathbf{c}\, d\mathbf{c} \tag{21}$$

Upon substitution of (17) into this expression two terms result

$$\mathbf{j}_t(\mathbf{r}, t) = n(\mathbf{r}, t)\langle\mathbf{c}\rangle + \mathbf{j}(\mathbf{r}, t) \tag{22}$$

with

$$\mathbf{j}(\mathbf{r}, t) = n(\mathbf{r}, t)\int g(\mathbf{c})\mathbf{c}\, d\mathbf{c} \tag{23}$$

The first term in (22) is seen from (15) to just equal the product of the density and the drift velocity; this is the expected drift current. The new current $\mathbf{j}(\mathbf{r}, t)$ induced by the density gradient is thus given by (23). From (20) it follows that $g(\mathbf{c})$ is a linear function of the three components k_x, k_y, k_z with coefficients which do not depend on the density or its gradient, but only on the unperturbed velocity distribution $f(\mathbf{c})$; furthermore, the first two of these coefficients are equal. Hence, from (23) $\mathbf{j}$ comes out as a linear function of the three quantities $n(\mathbf{r}, t)\cdot k_x$, $n(\mathbf{r}, t)\cdot k_y$, $n(\mathbf{r}, t)\cdot k_z$; these are the components of the density gradient

as is evident from (16); in addition the multipliers of the first two components are equal. We may write therefore

$$\mathbf{j}(\mathbf{r}, t) = -(\mathcal{D}) \frac{\partial n}{\partial \mathbf{r}} \tag{24}$$

where $(\mathcal{D})$ is a tensor which is axially symmetric about the field direction; its two components which we shall call the longitudinal diffusion coefficient $D_{||}$ and the transverse coefficient $D_{\perp}$ are computed entirely from the unperturbed velocity distribution $f(\mathbf{c})$. This makes $D_{||}$ and $D_{\perp}$ independent of the density or its gradient; it is to be noted, however, that they do depend on the electric field as a parameter because this quantity enters several times in the course of the computation.

ID. DIMENSIONAL ANALYSIS

Dimensional analysis is a convenient tool in a qualitative discussion of (13) and (20). In order to get results the situation has to be schematized somewhat, but not so much as to impair its usefulness. In the first place it is convenient to keep in mind the two limiting cases of high and low field, as discussed in the introduction. In addition some assumption must be made about $\sigma(\gamma)$ and $\Pi(\chi)$ occurring under the integral sign. The most convenient way to dispose of $\Pi(\chi)$ is to take it as independent of γ. This happens to be true for the two models treated in detail later, the polarization force model, and the hard sphere model. Actually $\Pi(\chi)$ can be taken as approximately independent of γ in a wider sense. The forces which produce scattering are either repulsive or short range attractive, that is, long range attractice forces are absent. As long as this is the case the scattering is roughly isotropic and hence can change but little with γ.[14]

A more drastic assumption is needed to dispose of $\sigma(\gamma)$. We must assume

$$\sigma(\gamma) \cdot \gamma^{\alpha} = \Gamma \tag{25}$$

where α and Γ are taken to be constants. This assumption contains two important special cases in it. They arise respectively by taking $\alpha = 0$ and $\alpha = 1$. The case $\alpha = 0$ is the case of a constant mean free path as exemplified by the hard sphere model. The case $\alpha = 1$ is the case of constant mean free time; it is applicable to the polarization force as dis-

[14] This statement is checked in detail in Section IIIB for the polarization force. This is the attractive force with the longest range which can arise in this field.

cussed in Section IA. When (25) is inserted into (13) it is seen that a, N and Γ enter only in the combination $a/N\Gamma$. The quantity

$$\left(\frac{a}{N\Gamma}\right)^{\frac{1}{2-\alpha}}$$

has the dimension of a velocity. A second such quantity is

$$\left(\frac{kT}{M}\right)^{1/2}$$

which arises from the Maxwellian functions under the integral sign.[15] In the high field case, this quantity does not enter, that is, the velocity distribution functions for the molecules could be replaced by δ-functions at the origin. Hence the first combination controls all velocity averages. For the mean drift velocity, we can thus write

$$\langle c_z \rangle = \text{const} \cdot \left(\frac{a}{N\Gamma}\right)^{\frac{1}{2-\alpha}} \tag{26a}$$

This formula gives the variation of the drift velocity with the electric field. It is worth while writing the result out explicitly for the two special cases discussed above. The first is the case of constant mean free path, $\alpha = 0$, for which

$$\langle c_z \rangle = \text{const} \cdot a^{1/2} \lambda^{1/2} \tag{26b}$$

This is a drift velocity varying as the square root of the field or a mobility varying inversely as the square root of the field. The second case is the one of constant mean free time $\alpha = 1$, for which

$$\langle c_z \rangle = \text{const} \cdot a\tau \tag{26c}$$

This means a drift velocity proportional to the field or a constant mobility.

In the low field case we cannot disregard one of the two velocity parameters constructed above; but now equation (13) is to be solved by perturbation theory only, it then yields a drift velocity varying with the first power of $a/N\Gamma$. Dimensional analysis then yields the dependence of the mobility on the temperature. We find

$$\langle c_z \rangle = \text{const} \cdot \frac{a}{N\Gamma} \left(\frac{kT}{M}\right)^{\frac{\alpha-1}{2}} \tag{27a}$$

[15] Dimensional analysis is incapable of distinguishing between m and M; this means that we cannot master dependence on mass by the method of this section; all our "pure numbers" are actually unknown functions of m/M.

with the special cases

$$\langle c_z \rangle = \text{const} \cdot a\lambda \left(\frac{kT}{M}\right)^{-1/2} \tag{27b}$$

for constant mean free path and

$$\langle c_z \rangle = \text{const} \cdot a\tau \tag{27c}$$

for constant mean free time. Comparison of (26c) and (27c) might lead one to surmise that we have here twice the same formula. This is indeed the case, as will be shown in Section IIIA.

Proceeding now to the diffusion problem, we observe from (20), (23) and (24) that we must add the quantity

$$\frac{nk}{N\Gamma}$$

to the previous list of parameters when computing the diffusion current. However, the current is always linear in this quantity, which means that the diffusion coefficients contain the factor

$$\frac{1}{N\Gamma}$$

and beyond this factor depend on the same variables as previously. This gives in the high field case

$$D = \text{const} \cdot \frac{1}{N\Gamma} \cdot \left(\frac{a}{N\Gamma}\right)^{\frac{1+\alpha}{2-\alpha}} \tag{28a}$$

with the special cases

$$D = \text{const} \cdot a^{1/2}\lambda^{3/2} \tag{28b}$$

and

$$D = \text{const} \cdot a^2\tau^3 \tag{28c}$$

In the low field case, the diffusion process becomes independent of the field and we get

$$D = \text{const} \cdot \frac{1}{N\Gamma} \left(\frac{kT}{M}\right)^{\frac{1+\alpha}{2}} \tag{29a}$$

with the special formulas

$$D = \text{const} \cdot \lambda \left(\frac{kT}{M}\right)^{1/2} \tag{29b}$$

and

$$D = \text{const} \cdot \tau \frac{kT}{M} \tag{29c}$$

The information in the formulas (29) is now new, but dependent on (27) through a universal relation first discovered by Nernst and derived independently for gases by J. J. Thomson; it is widely known as the Einstein relation. It states that

$$D = \frac{\partial \langle c_z \rangle}{\partial a} \cdot \frac{kT}{m} \tag{30}$$

Equation (30) contains of course more than is obtainable from (27) and (29), since it relates one undetermined constant to another in a known way.

The dimensional methods of this section are convenient for a rough classification of experimental material. Figs. 3 to 7 show the drift velocities

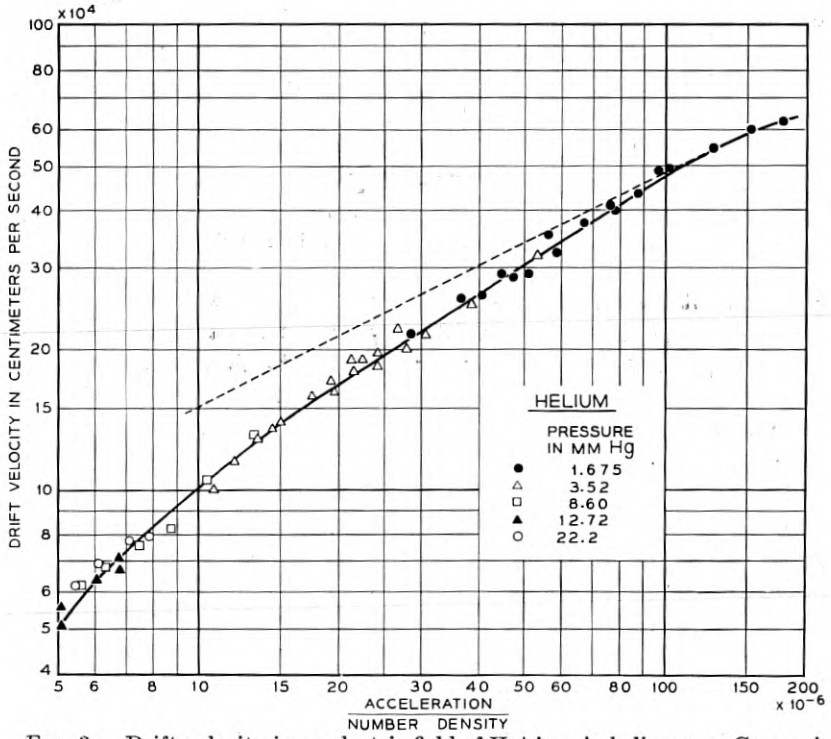

FIG. 3 — Drift velocity in an electric field of He^+ ions in helium gas. Comparison of observed results with an "asymptotic" straight line of slope ½.

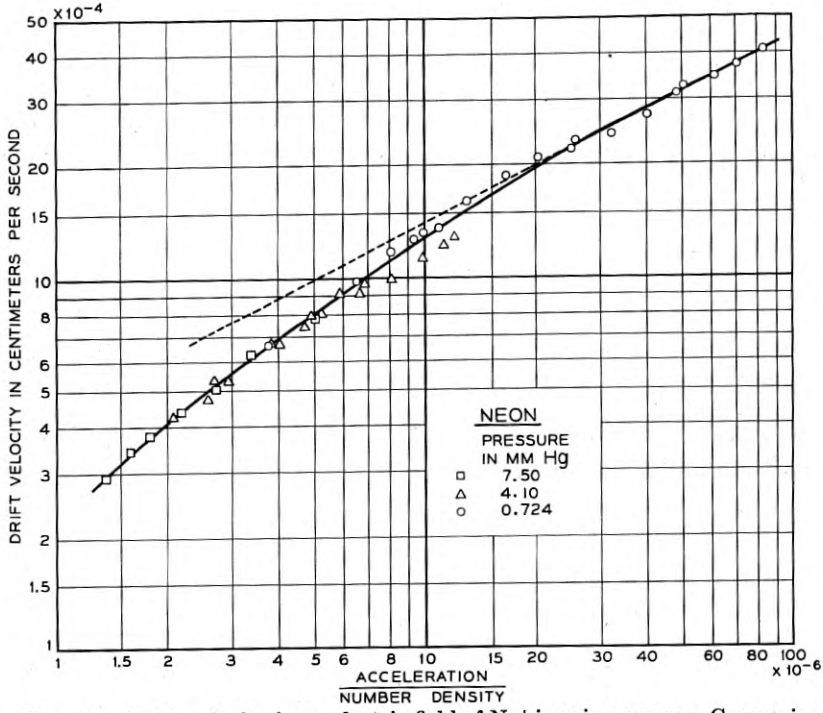

FIG. 4 — Drift velocity in an electric field of Ne^+ ions in neon gas. Comparison of observed results with an "asymptotic" straight line of slope ½

in the parent gas, observed for He^+, Ne^+, A^+, Kr^+ and Xe^+. The plot is a log-log plot of these quantities against a/N, a variety of fields having been used to determine each point. The data are taken from measurements of J. A. Hornbeck[16, 17] and R. N. Varney.[18] These data verify in the first place that the drift velocity depends on a and N only in the combination a/N. Beyond this we see that the curves consist of two straight line portions: in the lower field portion $\langle c_z \rangle$ is proportional to a/N, in the higher is proportional to $\sqrt{a/N}$. We recognize in this latter region the high field dependence predicted in equation (26b). We learn from this that the collision cross section between noble gas atoms and their ions is approximately constant in the experimentally significant velocity range. To determine these collision cross sections the computation of only a single number, namely the one entering into

[16] Hornbeck, J. A. and G. H. Wannier., Phys. Rev., **82**, p. 458, 1951.
[17] Hornbeck, J. A., Phys. Rev., **84**, p. 615, 1951.
[18] Varney, R. N., Phys. Rev., **88**, p. 362, 1952.

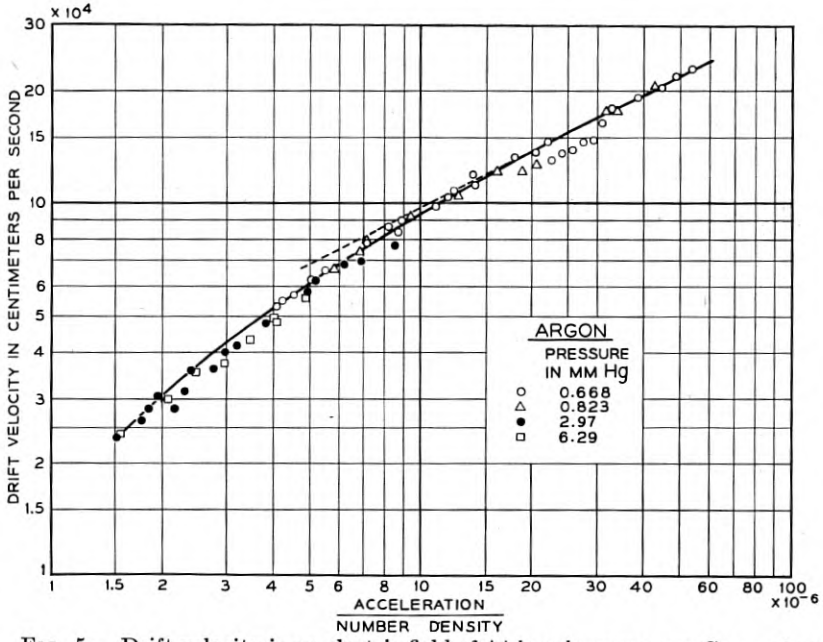

FIG. 5 — Drift velocity in an electric field of A^+ ions in argon gas. Comparison of observed results with an "asymptotic" straight line of slope ½.

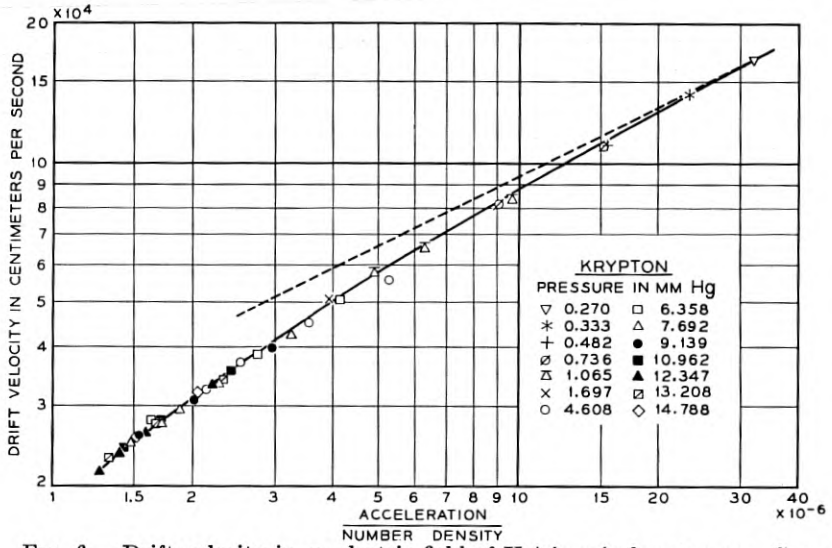

FIG. 6 — Drift velocity in an electric field of Kr^+ ions in krypton gas. Comparison of observed results with an "asymptotic" straight line of slope ½.

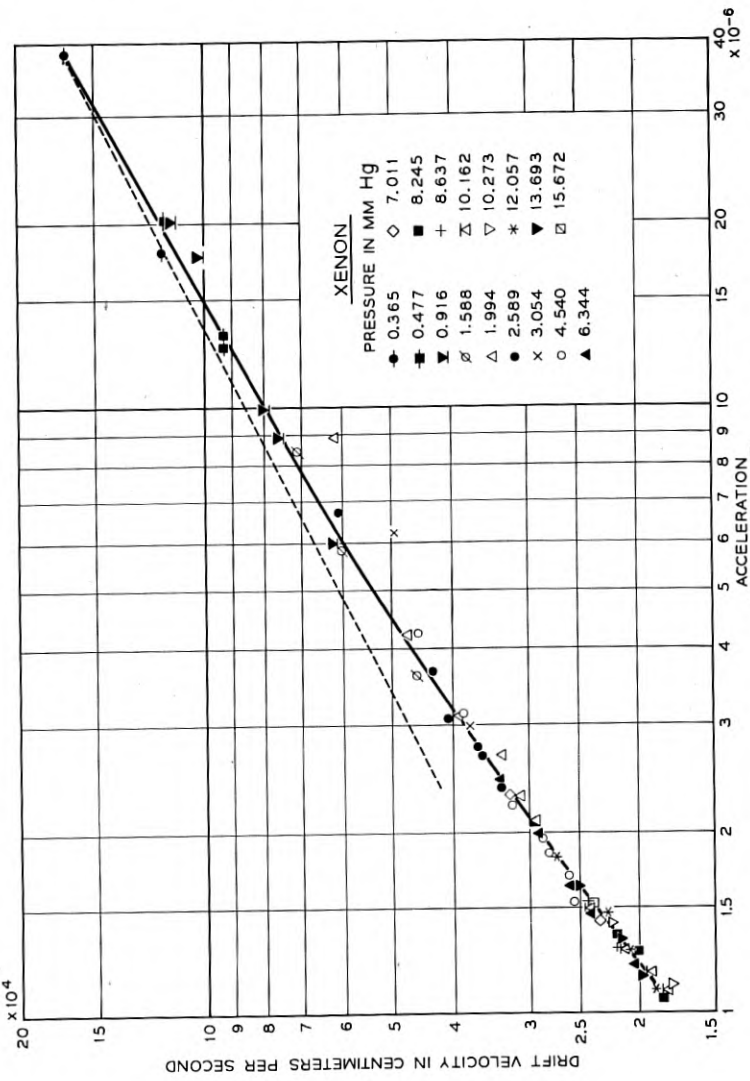

FIG. 7 — Drift velocity in an electric field of Xe^+ ions in xenon gas. Comparison of observed results with an "asymptotic" straight line of slope ½.

(26b) is required. The linear range of this plot is not as informative as the high field one. The slope unity is common to all formulas (27), and the temperature dependence of the mobility is needed to give the correct interpretation with the methods developed here. There is a certain likelihood that the parameter α of equation (25) drifts from 0 to 1 as the speed of the ions is reduced; this was pointed out for the special case of He^+ in He in section IA. A qualitatively similar situation appears to prevail for the other noble gases.

Part II — The Motion of Uniform Ion Streams in the High Field Case

IIA. Formulations of the Boltzmann Equation

The dimensional analysis of the last section shows that there is an intrinsic simplicity to the high field case which is comparable to the low field case, while the intermediate case is more difficult. With one exception,[6] however, theoretical analysis has occupied itself with the low field case only. We shall try to remedy this in the following. To begin with, a tractable but accurate formulation of the problem has to be found. Such a formulation cannot treat the field term of equation (13) as a perturbation term, but must try instead to make use of the basically simple features of the problem, notably those exhibited by the dimensional analysis of Section ID.

The equation governing the high field properties of the ions is obtained simply by substituting δ-functions for the Maxwellian velocity distributions in equation (13). This gives

$$a\frac{\partial h(\mathbf{c})}{\partial c_z} + \frac{1}{\tau(c)}h(\mathbf{c}) = \frac{1}{4\pi}\iint \delta(\mathbf{C}')h(\mathbf{c}')\frac{1}{\tau(c')}\Pi(\chi)\, d\Omega_{\gamma'}\, d\mathbf{C} \qquad (31)$$

A reduction of the number of integrations from five to two must be possible in the integral term of (31), owing to the presence of the δ-function. To achieve this we must transform the variables of integration so as to make three of the differentials equal to $d\mathbf{C}'$. We do this in the following way. First observe that

$$\boldsymbol{\gamma} = \mathbf{c} - \mathbf{C}$$

and that $\mathbf{c}$ is a constant vector. Hence we may replace $d\mathbf{C}$ by $d\boldsymbol{\gamma}$. The five-fold integration reads then

$$d\Omega_{\gamma'}\, d\mathbf{C} = \gamma^2\, d\gamma\, d\Omega_{\gamma}\, d\Omega_{\gamma'} \qquad (32)$$

that is, it goes over the magnitude γ which the two vectors have in

common, and their orientations, for which they are independent. It is known that in integrating over the two angles defining an orientation the polar axis may be chosen freely. We shall, in the following, adopt $\mathbf{c}$ as our polar axis with ψ, κ being pole distances of γ and γ' to $\mathbf{c}$ and φ, ω the corresponding azimuths. In Fig. 8 these angles are exhibited on the unit sphere. All vectors are assumed to be plotted from the center of the sphere, and show up through their piercing points. The angles between the vectors then show up as sides and the azimuths as angles. The expression (32) becomes then

$$\gamma^2 \, d\gamma \sin \psi \, d\psi \, d\varphi \sin \kappa \, d\kappa \, d\omega$$

The main transformation consists now in introducing the three components of $\mathbf{C}'$ in the place κ, ψ and φ. The transformation formulas follow from the vector identity

$$\mathbf{C}' = \mathbf{c} - \frac{M}{M+m}\gamma - \frac{m}{M+m}\gamma' \tag{33}$$

and read in full

$$C'_\xi = -\frac{M\gamma}{M+m}\sin\psi\cos\varphi - \frac{m\gamma}{M+m}\sin\kappa\cos\omega$$

$$C'_\eta = -\frac{M\gamma}{M+m}\sin\psi\sin\varphi - \frac{m\gamma}{M+m}\sin\kappa\sin\omega$$

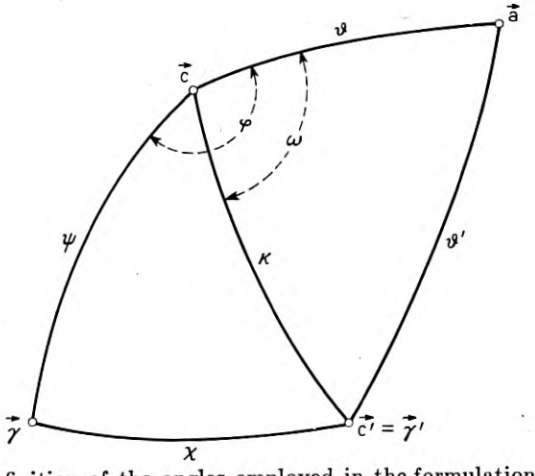

Fig. 8 — Definition of the angles employed in the formulations of the Boltzmann equation for the high field case.

$$C'_\zeta = -\frac{M\gamma}{M+m}\cos\psi - \frac{m\gamma}{M+m}\cos\kappa + c$$

From these equations the value of the Jacobian comes out to be

$$\frac{\partial(C'_\xi, C'_\eta, C'_\zeta)}{\partial(\kappa, \psi, \varphi)}$$

$$= \frac{mM^2\gamma^3}{(M+m)^3}\sin\psi\,\{\cos\psi\sin\kappa - \sin\psi\cos\kappa\cos(\varphi-\omega)\}$$

We needs its value only at the position $C'_\xi = C'_\eta = C'_\zeta = 0$. If we take the above equations for C'_ξ, C'_η, C'_ζ and multiply them respectively by $\cos\kappa\cos\omega$, $\cos\kappa\sin\omega$, $-\sin\kappa$, add and set $\mathbf{C}' = 0$ we get the identity

$$\frac{M\gamma}{M+m}\{\cos\psi\sin\kappa - \sin\psi\cos\kappa\cos(\varphi-\omega)\} = c\sin\kappa$$

The curly bracket is exactly the one occurring in the Jacobian which therefore reduces to

$$\left[\frac{\partial(C'_\xi, C'_\eta, C'_\zeta)}{\partial(\kappa, \psi, \varphi)}\right]_{\mathbf{C}'=0} = \frac{Mm}{(M+m)^2}c'^2 c\sin\psi\sin\kappa$$

and hence

$$\gamma^2\,d\gamma\,d\Omega_\gamma\,d\Omega_{\gamma'} = \frac{(M+m)^2}{Mmc}\,d\mathbf{C}'\,dc'\,d\omega$$

Substituting finally this expression into (32) and (31) we get the Boltzmann equation in the form

$$a\frac{\partial h(\mathbf{c})}{\partial c_z} + \frac{1}{\tau(c)}h(\mathbf{c}) =$$

$$= \frac{(M+m)^2}{4\pi Mmc}\int_c^{\frac{M+m}{|M-m|}c}\frac{1}{\tau(c')}\,\Pi(\chi)\,dc'\int_0^{2\pi}h(\mathbf{c}')\,d\omega \tag{34}$$

The equation is in need of additional elucidation as regards the exact meaning of $\mathbf{c}'$ as a vector and as regards the auxiliary variable χ. As to the first point we may describe the integration as occurring over a surface in velocity space. This surface is obtained from the relation

$$\mathbf{C}' = 0 \qquad \mathbf{c}' = \boldsymbol{\gamma}' \tag{35}$$

which substituted into (33) becomes

$$(M+m)\mathbf{c} - m\boldsymbol{\gamma}' = M\boldsymbol{\gamma} \tag{36}$$

Squaring this and using (9) we get

$$(M - m)c'^2 + 2m\mathbf{c}'\cdot\mathbf{c} - (M + m)c^2 = 0 \tag{37}$$

This is the equation of a sphere in velocity space which passes through the point $\mathbf{c}' = \mathbf{c}$. For all other points c' is bigger than c (collision with a stationary object always brings energy loss). The center of the sphere lies on the line joining $\mathbf{c}$ to the origin; it lies on the side of the origin from $\mathbf{c}$ when $m < M$, at infinity (making the sphere a plane) when $m = M$ and away from the origin when $m > M$. We make use of (37) to express the polar angle κ of $\mathbf{c}'$ with respect to $\mathbf{c}$ (which does not occur as an integration variable in (34)) in terms of c'. We get

$$\cos\kappa = \frac{(M + m)c^2 - (M - m)c'^2}{2mcc'} \tag{38}$$

The angle of scattering in the center of mass system also results from squaring of (36) if the term $m\gamma'$ is first taken to the right. We find

$$\cos\chi = \frac{(M + m)^2}{2Mm}\frac{c^2}{c'^2} - \frac{M^2 + m^2}{2Mm} \tag{39}$$

There is a more useful form of equation (34) which results if χ is taken as one of the integration variables rather than c'. Substitution is made from the equation (39) above; it yields

$$a\frac{\partial h(\mathbf{c})}{\partial c_z} + \frac{1}{\tau(c)}h(\mathbf{c}) = \frac{1}{4\pi}\int_0^\pi \sin\chi\, d\chi \frac{\Pi(\chi)}{\tau(c')}\left(\frac{c'}{c}\right)^3 \int_0^{2\pi} h(\mathbf{c}')\, d\omega \tag{40}$$

The magnitude of $\mathbf{c}'$ and its polar angle with respect to $\mathbf{c}$ are now auxiliary parameters; the first is obtained from (39)

$$c' = c\frac{M + m}{\sqrt{M^2 + m^2 + 2Mm\cos\chi}} \tag{41}$$

and the second from (38) and (41)

$$\cos\kappa = \frac{m + M\cos\chi}{\sqrt{M^2 + m^2 + 2Mm\cos\chi}} \tag{42}$$

As previously, the azimuth ω of $\mathbf{c}'$ about $\mathbf{c}$ is an independent variable.

The simplifications of the equation (31) exhibited in (34) and (40) still leave a double integral in the fundamental equation. The integration over $d\omega$ will now be eliminated by decomposition of $h(\mathbf{c})$ in spherical harmonics about the field direction. There is no loss of generality in this step.

$$h(\mathbf{c}) = \sum_{\nu=0}^{\infty} h_\nu(c)P_\nu(\cos\vartheta) \tag{43}$$

We have now to consider simultaneously the three vectors $\mathbf{c}$, $\mathbf{c}'$ and $\mathbf{a}$ as well as the angles between them. These angles are defined in Fig 8. We study equation (34) or (40) term by term in order to see what becomes of it upon substitution of (43). Starting with $h(\mathbf{c}')$ under the integral sign we get from Fig. 8 and the addition theorem for spherical harmonics

$$h(\mathbf{c}') = \sum_{\nu=0}^{\infty} h_\nu(c')[P_\nu(\cos\vartheta)P_\nu(\cos\kappa) + 2\sum_{\mu=1}^{\nu}\frac{(\nu-\mu)!}{(\nu+\mu)!}P_\nu^\mu(\cos\vartheta)P_\nu^\mu(\cos\kappa)\cos\mu\omega]$$

For this expression, the integration over ω is elementary and gives

$$\int_0^{2\pi} h(\mathbf{c}')\,d\omega = 2\pi\sum_{\nu=0}^{\infty} h_\nu(c')P_\nu(\cos\vartheta)P_\nu(\cos\kappa) \tag{44}$$

Further, we get for the derivative in (34) or (40)

$$\frac{\partial}{\partial c_z}\left(\sum_{\nu=0}^{\infty} h_\nu(c)P_\nu(\cos\vartheta)\right) = \sum_{\nu=0}^{\infty}\frac{dh_\nu(c)}{dc}\frac{1}{2\nu+1}\{(\nu+1)P_{\nu+1}(\cos\vartheta)+\nu P_{\nu-1}(\cos\vartheta)\} + \frac{1}{c}h_\nu(c)\frac{\nu(\nu+1)}{2\nu+1}\{P_{\nu-1}(\cos\vartheta)-P_{\nu+1}(\cos\vartheta)\} \tag{45}$$

Through the equations (43), (44) and (45), all terms in equation (34) or (40) are developed in spherical harmonics with respect to the angle ϑ between c and the field direction. We can therefore annul separately the coefficient of each Legendre polynomial in $\cos\vartheta$. This gives the following set of equations

$$\frac{(M+m)^2}{2Mmc}\int_c^{\frac{M+m}{|M-m|}c}\frac{h_\nu(c')}{\tau(c')}P_\nu(\cos\kappa)\,\Pi(\chi)\,dc' - \frac{h_\nu(c)}{\tau(c)} = \frac{\nu a}{2\nu-1}\left(\frac{dh_{\nu-1}(c)}{dc}-\frac{\nu-1}{c}h_{\nu-1}(c)\right) + \frac{(\nu+1)a}{2\nu+3}\left(\frac{dh_{\nu+1}(c)}{dc}+\frac{\nu+2}{c}h_{\nu+1}(c)\right) \tag{46}$$

or

$$\frac{1}{2}\int_0^\pi \frac{h_\nu(c')}{\tau(c')}\left(\frac{c'}{c}\right)^3 P_\nu(\cos\kappa)\,\Pi(\chi)\sin\chi\,d\chi$$

$$-\frac{h_\nu(c)}{\tau(c)} = \frac{\nu a}{2\nu-1}\left\{\frac{dh_{\nu-1}(c)}{dc} - \frac{\nu-1}{c}h_{\nu-1}(c)\right\} \tag{47}$$

$$+\frac{(\nu+1)a}{2\nu+3}\left\{\frac{dh_{\nu+1}(c)}{dc} + \frac{\nu+2}{c}h_{\nu+1}(c)\right\}$$

where

$$\nu = 0, 1, 2, 3 \cdots .$$

The auxiliary parameters entering are given by (38) and (39) for equation (46), and (41) and (42) for equation (47).

The equations (46) or (47) obtained by Legendre decomposition still are, in general, mixed integral-differential equations in one independent variable. Further simplification is possible only in special cases some of which will be discussed later. An even more simple and tractable form of the Boltzmann equation can be achieved in general, however, if one gives up the idea of determining the velocity distribution function and concentrates instead on its moments. In other words, the Boltzmann equation can be looked upon as a system of relations between velocity averages, and as such it becomes a linear algebraic system.

To carry out this reduction we multiply equation (47) by c^{s+2} and integrate from 0 to ∞. The second term on the left is then a simple velocity average. The same is true on the right hand side if two integrations by part are permissible and leave no integrated out part. $s \geqq -1$ is probably adequate for this. The integral over the integral term at first looks as follows

$$\frac{1}{2}\int_0^\infty c^{s+2}\,dc\int_0^\pi \frac{h_\nu(c')}{\tau(c')}\left(\frac{c'}{c}\right)^3 P_\nu(\cos\kappa)\,\Pi(\chi)\sin\chi\,d\chi$$

In this expression we pass from c to c' as the independent variable. From (41) we see that

$$\frac{dc}{c} = \frac{dc'}{c'}$$

Hence the expression becomes

$$\int_0^\infty c'^{s+2}\frac{h_\nu(c')}{\tau(c')}\,dc'\,\frac{1}{2}\int_0^\pi \left(\frac{c}{c'}\right)^s P_\nu(\cos\kappa)\,\Pi(\chi)\sin\chi\,d\chi$$

From (41) and (42) it is seen that this is actually the product of two independent integrals if the angular distribution $\mathrm{II}(\chi)$ is independent of the velocity of encounter c'. The first integral is then identical with the one arising from the second term in (47), and the second is a collision integral having no connection with the velocity distribution. Even if this is not the case, the second integral is still a dynamic average which can be evaluated as a function of c' previously to any knowledge of $h(\mathbf{c}')$. We express this by introducing the abbreviation

$$I_{s,\nu} = \left(\frac{c}{c'}\right)^s P_\nu (\cos \kappa)$$

Using (41) and (42) we see that $I_{s,\nu}$ is the following function of χ

$$I_{s,\nu}(\chi) = \left(\frac{\sqrt{M^2 + m^2 + 2Mm \cos \chi}}{M + m}\right)^s \cdot P_\nu\left(\frac{m + M \cos \chi}{\sqrt{M^2 + m^2 + 2Mm \cos \chi}}\right) \tag{48a}$$

which, for the particular case of equal masses, takes the simple form

$$I_{s,\nu}(\chi) = \cos^s \tfrac{1}{2}\chi \, P_\nu (\cos \tfrac{1}{2}\chi) \tag{48b}$$

With this definition the integrated equation (47) reads

$$\int_0^\infty \left\langle \frac{1 - I_{s,\nu}(\chi)}{a\tau(c)} \right\rangle h_\nu(c) c^{s+2} \, dc = \frac{\nu(\nu + s + 1)}{2\nu - 1} \int_0^\infty h_{\nu-1}(c) c^{s+1} \, dc + \frac{(\nu + 1)(s - \nu)}{2\nu + 3} \int_0^\infty h_{\nu+1}(c) c^{s+1} \, dc$$

or in terms of averages

$$(2\nu + 1)\left\langle \frac{1 - I_{s,\nu}(\chi)}{a\tau(c)} c^s P_\nu (\cos \vartheta) \right\rangle = \nu(\nu + s + 1)\langle c^{s-1} P_{\nu-1} (\cos \vartheta)\rangle + (\nu + 1)(s - \nu)\langle c^{s-1} P_{\nu-1} (\cos \vartheta)\rangle \tag{49}$$

I believe that equation (49) contains all possible derivable relations between averages as special cases. Some of the most notable ones are

listed below

$$s = 1, \qquad \nu = 1$$

$$\left\langle \frac{1 - \cos \chi}{a\tau(c)} c \cos \vartheta \right\rangle = \frac{M + m}{M} \tag{50a}$$

$$s = 2, \qquad \nu = 0$$

$$\left\langle \frac{1 - \cos \chi}{a\tau(c)} c^2 \right\rangle = \frac{(M + m)^2}{Mm} \langle c \cos \vartheta \rangle \tag{50b}$$

$$s = 2, \qquad \nu = 2$$

$$\left\langle \frac{3M \sin^2 \chi + 4m(1 - \cos \chi)}{a\tau(c)} c^2 P_2 (\cos \vartheta) \right\rangle = \\ = \frac{4(M + m)^2}{M} \langle c \cos \vartheta \rangle \tag{50c}$$

While the averages entering into (49) are not always the desired ones, it remains true nevertheless that all solution methods evolved in the following use this equation system as a starting point rather than other forms of the Boltzmann equation.

IIB. THE MEAN FREE TIME MODEL AT HIGH FIELD

If the angular distribution in the center of mass system is independent of speed and the collision cross section varies inversely as the speed then the developments of the previous section permit actually a solution of the Boltzmann equation. It is a solution in the sense that all significant velocity averages can be obtained directly without the knowledge of the velocity distribution function.

Before developing these facts from the equations of the last section, I should point out that the derivation to follow is in a sense artificial. It has been shown already by Maxwell[19] for related problems that if the mean free time between collisions is assumed constant specially simple techniques may be employed to get constants of experimental importance. These techniques can be employed here; they consist essentially in multiplying (13) by a suitable multiplier, followed by integration over **c**. However, if we were to follow this procedure we would have to duplicate for a special model in an unsystematic way the work done systematically for all laws of interactions in the preceding section. A further advantage of using systematic procedure is that we can see at a

[19] Maxwell, J. C., Collected Papers, Vol. II, p. 40.

glance what averages can or cannot be obtained, and what the relationship is between the high field and the general averages.

For this reason we limit ourselves at present to the high field averages obtainable from (49). For the special case under discussion this equation system takes the form

$$(2\nu + 1)\left\langle \frac{1 - I_{s,\nu}(\chi)}{a\tau} \right\rangle \langle c^s P_\nu (\cos\vartheta)\rangle = \nu(\nu + s + 1)\langle c^{s-1} P_{\nu-1} (\cos\vartheta)\rangle + (\nu + 1)(s - \nu)\langle c^{s-1} P_{\nu+1} (\cos\vartheta)\rangle \tag{51}$$

that is we have a system of linear relations connecting the averages $\langle c^s P_\nu (\cos\vartheta)\rangle$. The connection between these averages is made apparent in Fig. 9. Each average $\langle c^s P_\nu (\cos\vartheta)\rangle$ is marked in this figure as a dot in an s-ν-plane if s is integer. The equations (51) connecting these averages are shown as lines with different equations leading to the same dot shown in different outline. These equations generally have the shape of a V;

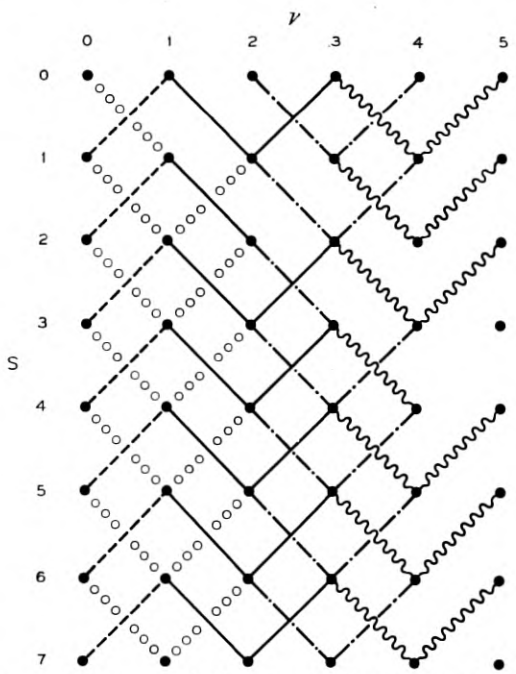

Fig. 9 — Interconnection established by the Boltzmann equation among the averages $\langle c^s P_\nu (\cos\vartheta)\rangle$; case of constant mean free time.

there are two notable exceptions to this rule, however, which make the recurrence method possible, the equations $\nu = 0$ have no left leg and the equations $s = \nu$ have no right leg. Starting out with the average $s = 0$, $\nu = 0$, which equals unity by definition one can thus proceed systematically as shown in Fig. 10, to get other averages. The averages reached are the ones for which s and ν are non-negative integers of equal parity with the restriction $s \geqq \nu$. One verifies easily that this set is equivalent to the set of all products of integer powers of the velocity components.

The first three relations one uses in the path outlined in Fig. 10, are the simplified forms of the three equations (50). We find

$$\langle c_z \rangle = \frac{M + m}{M} \Big/ \left\langle \frac{1 - \cos \chi}{a\tau} \right\rangle \tag{52}$$

$$\langle c^2 \rangle = \frac{(M + m)^3}{M^2 m} \Big/ \left\langle \frac{1 - \cos \chi}{a\tau} \right\rangle^2 \tag{53}$$

$$\langle c^2 P_2 (\cos \vartheta) \rangle = \frac{4(M + m)^3}{M^2 \left\langle \frac{3M \sin^2 \chi + 4m(1 - \cos \chi)}{a\tau} \right\rangle \left\langle \frac{1 - \cos \chi}{a\tau} \right\rangle}$$

or, more conveniently with the help of (53)

$$\langle c_z^2 \rangle = \frac{(M + m)^3 \left\langle \frac{M \sin^2 \chi + 4m(1 - \cos \chi)}{a\tau} \right\rangle}{M^2 m \left\langle \frac{3M \sin^2 \chi + 4m(1 - \cos \chi)}{a\tau} \right\rangle \left\langle \frac{1 - \cos \chi}{a\tau} \right\rangle^2} \tag{54}$$

The three equations (52), (53) and (54) give the drift velocity, the total energy, and the energy partition of the travelling ion. Equation (52) gives a constant mobility and can actually be derived from a low field theory. Formula (52) thus states that for problems involving a constant mean free time the high field and low field mobilities are numerically identical. One would suspect that the intermediate field value would have to fall in line too. This is indeed the case as will be shown in Section IIIA.

A convenient interpretation of (53) may be had by combining (52) and (53) in the following way

$$\langle mc^2 \rangle = m\langle c_z \rangle^2 + M\langle c_z \rangle^2 \tag{55}$$

The left side is essentially the total energy of the ion, the first term on the right is the energy visible in the drift motion; it follows therefore

that the second term is the "invisible" or random part of the mean energy. Formula (55) thus states that

$$\frac{\text{random energy}}{\text{visible energy}} = \frac{\text{molecular mass}}{\text{ion mass}} \tag{56}$$

that is, it exhibits in a quantitative way the capacity of storing energy in the form of random motion which light ions travelling in a heavy gas possess; for ions travelling in the parent gas the ordered and the random part of the energy are just equal; for heavy ions in a light gas the disordered fraction becomes negligible.

There are various ways of understanding the implications of equation (54). One way is to derive the mean energy in a direction at right angles to the field by the use of (53). We find

$$\langle c_x^2 \rangle = \frac{(M+m)^3 \left\langle \frac{\sin^2 \chi}{a\tau} \right\rangle}{Mm \left\langle \frac{3M \sin^2 \chi + 4m(1-\cos\chi)}{a\tau} \right\rangle \left\langle \frac{1-\cos\chi}{a\tau} \right\rangle^2} \tag{57}$$

Now from (54) and (57) the partition of the energy e may be obtained.

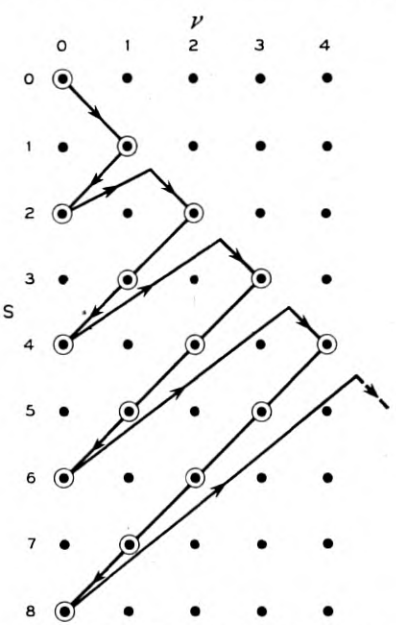

Fig. 10 — Order to be followed in computing by recursion the averages $\langle c^s P_\nu (\cos\vartheta) \rangle$; case of constant mean free time.

It comes out to be

$$e_x : e_y : e_z = \left\langle \frac{M \sin^2 \chi}{\tau} \right\rangle : \left\langle \frac{M \sin^2 \chi}{\tau} \right\rangle : \left\langle \frac{M \sin^2 \chi + 4m(1 - \cos \chi)}{\tau} \right\rangle \tag{58}$$

This shows up immediately the equipartition property for small m/M and the overwhelming preponderance of the motion in the field direction for large m/M. As an analysis of the random motion, however, equation (58) is deficient because the three directions become only comparable after the square of the drift velocity (52) is subtracted out of the z-component. We find

$$\langle c_z^2 \rangle - \langle c_z \rangle^2 = \frac{(M + m)^2 \left\langle \frac{M \sin^2 \chi + 2m(1 - \cos \chi)^2}{a\tau} \right\rangle}{Mm \left\langle \frac{3M \sin \chi + 4m(1 - \cos \chi)}{a\tau} \right\rangle \left\langle \frac{1 - \cos \chi}{a\tau} \right\rangle^2} \tag{59}$$

and from this a more refined partition formula which only counts random motion

$$e_x : e_y : e_z^* = \left\langle \frac{\sin^2 \chi}{\tau} \right\rangle : \left\langle \frac{\sin^2 \chi}{\tau} \right\rangle : \left\langle \frac{M \sin^2 \chi + 2m(1 - \cos \chi)^2}{(M + m)\tau} \right\rangle \tag{60}$$

For small m/M this result does not differ essentially from (58) but if m/M is large the z component of the random energy does not grow indefinitely the way the total energy does. Instead it stops at a value which is about four times one of the other two values.

A discussion of these expressions for special models will be delayed until the equations are extended to intermediate field conditions. This will be done in Part III.

IIC. THE CASE OF LARGE MASS RATIOS: ELECTRONS OR HEAVY IONS

The distribution of velocities for a small value of m/M is treated in the literature because it applies to electrons.[4,5] However, for the sake of completeness the derivation will be carried out here for the high field case. In this derivation all features of the law of scattering are left open, except that conservation of the kinetic energy is assumed.

The development in spherical harmonics carried out in Section IIA is the suitable starting point for small m/M, because, in this case, the distribution is almost spherically symmetrical and the expansion in spherical harmonics is also an expansion in powers of m/M. If we keep only $h_0(c)$ and $h_1(c)$ in the system (47) and treat m/M as small we get two equations, one for $\nu = 0$ and the one for $\nu = 1$. We may then use

(41) and (42) in the simplified form

$$\frac{c'}{c} \sim 1 + \frac{m}{M}(1 - \cos \chi) \tag{61}$$

$$\cos \kappa \sim \cos \chi + \frac{m}{M} \sin^2 \chi \tag{62}$$

and develop the equations in powers of m/M. Starting out with the simpler equation $\nu = 1$ we find

$$\left\langle \frac{1 - \cos \chi}{a\tau(c)} \right\rangle h_1(c) = -\frac{dh_0(c)}{dc} \tag{63}$$

This same procedure is not adequate for the equation $\nu = 0$ because the two left hand terms in (47) cancel in zero approximation. We must therefore develop the integral up to linear terms in m/M; this is a perfectly straightforward, though somewhat cumbersome, step. It leads to the following equation

$$\frac{m}{M}\left[c\frac{d}{dc}\left\{\left\langle \frac{1 - \cos \chi}{a\tau(c)} \right\rangle h_0(c)\right\} + 3\left\langle \frac{1 - \cos \chi}{a\tau(c)} \right\rangle h_0(c)\right] = \frac{1}{3}\left[\frac{dh_1(c)}{dc} + \frac{2}{c} h_1(c)\right]$$

The equation may be integrated by multiplication with c^2; this yields

$$h_1(c) = 3\frac{m}{M}\left\langle \frac{1 - \cos \chi}{a\tau(c)} \right\rangle c h_0(c) \tag{64}$$

Elimination of $h_1(c)$ from (63) and (64) gives a differential equation for $h_0(c)$ which is easily solved by quadrature; the result is

$$h_0(c) = \exp\left[-3\frac{m}{M}\int_0^c \left\langle \frac{1 - \cos \chi}{a\tau(c)} \right\rangle^2 c\, dc\right] \tag{65}$$

Except for the dependence on the angular law of scattering this formula may be found in the literature.[4] Its most important special cases are obtained for $\tau = \text{const}$ (Pseudo-Maxwellian distribution) and $\tau = \text{const}/c$ (Druyvesteyn distribution).

The derivation of (65) should be completed by a proof that indeed $h_2(c)$ is small compared to $h_1(c)$. This is not true for all values of $\mathbf{c}$; on the contrary, the argument below shows that near the origin where $c \sim a\tau(c)$, $h_2(c)$ is actually comparable to $h_1(c)$. For our purposes, however, it is sufficient if it is true in the overwhelming majority of cases. As the proof applies equally to all h_ν's we will run it in this manner. Assuming

provisionally that

$$h_{\nu+1} \ll h_\nu \ll h_{\nu-1}$$

we see that we can drop the terms in $h_{\nu+1}$ in the system (47). The integral in (47) is evaluated in zero order with the help of (61) and (62); it becomes then

$$\frac{h_\nu(c)}{\tau(c)} \cdot \frac{1}{2} \int_0^\pi P_\nu(\cos\chi)\Pi(\chi) \sin\chi \, d\chi$$

This means that we may solve explicitly for $h_\nu(c)$ in terms of $h_{\nu-1}(c)$. The formula is

$$h_\nu(c) \approx \frac{\nu}{2\nu - 1} \frac{-\frac{dh_{\nu-1}(c)}{dc} + \frac{\nu - 1}{c} h_{\nu-1}(c)}{\left\langle \frac{1 - P_\nu(\cos\chi)}{a\tau(c)} \right\rangle} \tag{66}$$

To estimate the order of magnitude of this we may neglect $P_\nu(\cos\chi)$ as compared to 1 and assume ν large. The operation in the numerator will lead to two kinds of terms: some of the form

$$\frac{1}{c} h_{\nu-1}(c)$$

and others of the type

$$\frac{m}{M} \frac{c}{\{(a\tau(c)\}^2} h_{\nu-1}(c)$$

coming from differentiation of an exponent of the type (65). Now we find from (65) that the overwhelming majority of particles have speeds c which in order of magnitude satisfy

$$c \sim (M/m)^{1/2} a\tau(c) \tag{67}$$

because this is the range within which the exponent remains comparable to 1. Applying this to the two types of terms arising from (66) we find for them in order of magnitude

$$\nu a\tau(c) \frac{1}{c} h_{\nu-1}(c) \sim \nu \left(\frac{m}{M}\right)^{1/2} h_{\nu-1}(c)$$

and

$$a\tau(c) \frac{m}{M} \frac{c}{\{(a\tau(c)\}^2} h_{\nu-1}(c) \sim \left(\frac{m}{M}\right)^{1/2} h_{\nu-1}(c)$$

If we substitute this into (66) we see that the h_ν's decrease as $(m/M)^{\frac{1}{2}\nu}$, with a possible $\nu!$ slowing up the final convergence. In any case $h_2(c)$ comes out small compared to $h_1(c)$ which is all that is needed to make equation (65) approximately correct.

While the case of small m/M is generally known it appears to be otherwise for large m/M. The intuitive basis for the solution of this case is the fact observable from (52), (57) and (59) that $\langle \mathbf{c} \rangle$ increases indefinitely with m/M, but that the relative deviation from the mean decreases so that the distribution function approaches a δ-function. The structure of this limiting function may be explored, starting directly from (40), because in the limit of large m/M the sphere of integration shrinks as may be verified from (37). This makes it possible to replace the integral in (40) by differential terms. This becomes clearer if (40) is written in the form

$$a \frac{\partial h(\mathbf{c})}{\partial c_z} = \frac{1}{2} \int_0^\pi \Pi(\chi) \sin \chi \, d\chi \frac{1}{2\pi} \int_0^{2\pi} d\omega \left\{ \left(\frac{c'}{c}\right)^3 \frac{h(\mathbf{c}')}{\tau(c')} - \frac{h(c)}{\tau(c)} \right\} \tag{68}$$

Let us call the inner average $\mathcal{J}$. It exhibits the differential properties discussed earlier: the curly bracket is the difference of two terms which are almost identical. Hence we approximate the value of $\mathcal{J}$ by expanding the slowly varying terms to first order in $\mathbf{c}' - \mathbf{c}$, while the fast varying $h(\mathbf{c}')$ will be expanded to square terms. This expansion is obviously permissible for everything except the rapidly varying function $h(\mathbf{c}')$. For $h(\mathbf{c}')$ itself no justification can be offered except success. By proceeding to square terms in this expansion we mitigate any possible error committed, but it is quite possible that structural details are lost in the procedure.

The development of $\mathcal{J}$ is straightforward. We proceed as follows

$$\mathcal{J} = \frac{1}{2\pi} \int_0^{2\pi} d\omega \left\{ h(\mathbf{c}) \left(\left(\frac{c'}{c}\right)^3 \frac{1}{\tau(c')} - \frac{1}{\tau(c)} \right) + \left(\frac{c'}{c}\right)^3 \frac{1}{\tau(c')} \left(h(\mathbf{c}') - h(\mathbf{c})\right) \right\}$$

$$\approx h(\mathbf{c}) \left\{ \left(\frac{c'}{c}\right)^3 \frac{1}{\tau(c')} - \frac{1}{\tau(c)} \right\} + \frac{1}{\tau(c)} \cdot \frac{1}{2\pi} \int_0^{2\pi} d\omega \, \{h(\mathbf{c}') - h(\mathbf{c})\}$$

The expansion of the first term involves only (41) which to this order reads

$$c' - c \approx c \frac{M}{m} (1 - \cos \chi)$$

This formula does not contain the azimuth ω which therefore disappears trivially. In the second term on the contrary we are dealing with a vectorial difference involving all three polar coordinates c', κ, ω of $\mathbf{c}'$.

If we set

$$c'_\xi = c' \sin \kappa \cos \omega$$

$$c'_\eta = c' \sin \kappa \sin \omega$$

$$c'_\zeta = c' \cos \kappa$$

we find

$$h(\mathbf{c}') - h(\mathbf{c}) = \frac{\partial h(\mathbf{c})}{\partial c_\xi} c'_\xi + \frac{\partial h(\mathbf{c})}{\partial c_\eta} c'_\eta + \frac{\partial h(\mathbf{c})}{\partial c_\zeta} (c'_\zeta - c) + \frac{1}{2} \frac{\partial^2 h(\mathbf{c})}{\partial c_\xi^2} c'^2_\xi$$
$$+ \frac{1}{2} \frac{\partial^2 h(\mathbf{c})}{\partial c_\eta^2} c'^2_\eta + \frac{1}{2} \frac{\partial^2 h(\mathbf{c})}{\partial c_\zeta^2} (c'_\zeta - c)^2 + \frac{\partial^2 h(\mathbf{c})}{\partial c_\xi \partial c_\eta} c'_\xi c'_\eta$$
$$+ \frac{\partial^2 h(\mathbf{c})}{\partial c_\xi \partial c_\zeta} c'_\xi (c'_\zeta - c) + \frac{\partial^2 h(\mathbf{c})}{\partial c_\eta \partial c_\zeta} c'_\eta (c'_\zeta - c) + \cdots$$

With the formulas given, integration over ω is elementary. We find

$$\mathcal{J} \approx h(\mathbf{c}) \left\{ \left(\frac{c'}{c} \right)^3 \frac{1}{\tau(c')} - \frac{1}{\tau(c)} \right\} + \frac{1}{\tau(c)} \left\{ \frac{\partial h(\mathbf{c})}{\partial c_\zeta} (c' \cos \kappa - c) \right.$$
$$\left. + \frac{1}{4\tau(c)} \left(\frac{\partial^2 h(\mathbf{c})}{\partial c_\xi^2} + \frac{\partial^2 h(\mathbf{c})}{\partial c_\eta^2} \right) c'^2 \sin^2 \kappa + \frac{1}{2\tau(c)} \frac{\partial^2 h(\mathbf{c})}{\partial c_\zeta^2} (c' \cos \kappa - c)^2 \right\}$$

All coefficients are to be evaluated only to the lowest non-vanishing order in M/m. From the equations (41) and (42) we get

$$c' \cos \kappa - c \approx (M/m)\, c(1 - \cos \chi)$$

$$c'^2 \sin^2 \kappa \approx (M/m)^2 c^2 \sin^2 \chi$$

The first term in $\mathcal{J}$ is simplified further by introduction of the parameter α used in the dimensional analysis of Section ID. According to that section we have that

$$\frac{d\, \ell n \tau(c)}{d\, \ell n c} = -1 + \alpha \tag{69}$$

We can generalize the original definition for any $\tau(c)$ by the above equation, where α is now a function of c. Eliminating also the tempo-

rary device of a ξ, η, ζ coordinate system we find for $\mathcal{J}$:

$$\mathcal{J} = \frac{M}{m}(4 - \alpha(c))(1 - \cos\chi)\frac{h(\mathbf{c})}{\tau(c)} + \frac{M}{m}\frac{(1 - \cos\chi)}{\tau(c)}\left\{c_x\frac{\partial h}{\partial c_x} + c_y\frac{\partial h}{\partial c_y} + c_z\frac{\partial h}{\partial c_z}\right\} + \left(\frac{M}{m}\right)^2\frac{\sin^2\chi}{4\tau(c)}\left\{\frac{\partial^2 h}{\partial c_x^2} + \frac{\partial^2 h}{\partial c_y^2} + \frac{\partial^2 h}{\partial c_z^2}\right\} + \left(\frac{M}{m}\right)^2\frac{2(1 - \cos\chi)^2 - \sin^2\chi}{4\tau(c)}\, c\left\{c_x\frac{\partial}{\partial c_x} + c_y\frac{\partial}{\partial c_y} + c_z\frac{\partial}{\partial c_z}\right\} \cdot \left\{\frac{c_x}{c}\frac{\partial h}{\partial c_x} + \frac{c_y}{c}\frac{\partial h}{\partial c_y} + \frac{c_z}{c}\frac{\partial h}{\partial c_z}\right\}$$

If the last term is evaluated exactly terms of the order $(M/m)^2$ are added to the first derivative terms in $h(\mathbf{c})$. They are obviously negligible. Finally integration over χ yields the following form for equation (68)

$$a\frac{\partial h(\mathbf{c})}{\partial c_z} = \langle\mathcal{J}\rangle = \frac{M}{m}\left\langle\frac{1 - \cos\chi}{\tau(c)}\right\rangle\{4 - \alpha(c)\}h(\mathbf{c}) + \frac{M}{m}\left\langle\frac{1 - \cos\chi}{\tau(c)}\right\rangle\sum_{i=1}^{3} c_i\frac{\partial h}{\partial c_i} + \frac{1}{4}\left(\frac{M}{m}\right)^2\left\langle\frac{\sin^2\chi}{\tau(c)}\right\rangle c^2\sum_{i=1}^{3}\frac{\partial^2 h}{\partial c_i^2} + \frac{1}{4}\left(\frac{M}{m}\right)\left\langle\frac{(1 - \cos\chi)(1 - 3\cos\chi)}{\tau(c)}\right\rangle\sum_{i,k=1}^{3} c_i c_k\frac{\partial^2 h}{\partial c_i \partial c_k} \tag{70}$$

When equation (70) is considered up to linear terms in M/m it yields a δ-function about the drift velocity $\langle c_z\rangle$ which results from the implicit equation

$$\langle c_z\rangle = \frac{m}{M}\, a \Big/ \left\langle\frac{1 - \cos\chi}{\tau(c)}\right\rangle_{c=\langle c_z\rangle} \tag{71}$$

The δ-function takes here the aspect of a non integrable function which in a special case can be seen to equal

$$h \approx \{c_x^2 + c_y^2 + (c_z - \langle c_z\rangle)^2\}^{-3/2}$$

When normalization is imposed on such a function it is made to vanish everywhere except at the point corresponding to the drift velocity.

The square terms in M/m are necessary to gain information about the functional form of $h(\mathbf{c})$. Since the region in velocity space in which h is appreciable is still small we may take α as constant in that region. A further simplication results from order of magnitude considerations on $\mathbf{c}$:

$$c^2 \sim c_z^2 \sim \langle c_z\rangle^2 \gg c_x^2 \sim c_y^2 \sim (c_z - \langle c_z\rangle)^2$$

In working out the details we see that division by

$$\frac{M}{m}\left\langle\frac{1-\cos\chi}{\tau(c)}\right\rangle$$

puts the first two terms on the right hand side of (70) in a simple form. The coefficient of the field term becomes then

$$\frac{am}{M\left\langle\frac{1-\cos\chi}{\tau(c)}\right\rangle}=\frac{am}{M\left\langle\frac{1-\cos\chi}{\tau(c)}\right\rangle_{c=\langle c_z\rangle}}\cdot\left(\frac{\langle c_z\rangle}{c}\right)^{1-\alpha}$$

The first factor is just $\langle c_z\rangle$ by the identity (71). For the second factor we have up to linear small terms

$$c=\sqrt{c_x^2+c_y^2+c_z^2}=\sqrt{c_x^2+c_y^2+(\langle c_z\rangle+c_z-\langle c_z\rangle)^2}$$

$$\approx\langle c_z\rangle\left\{1+\frac{c_z-\langle c_z\rangle}{\langle c_z\rangle}\right\}$$

and hence for the coefficient of the field term

$$\langle c_z\rangle\left(\frac{\langle c_z\rangle}{c}\right)^{1-\alpha}\approx\langle c_z\rangle-(1-\alpha)(c_z-\langle c_z\rangle)$$

After division by

$$\frac{M}{m}\left\langle\frac{1-\cos\chi}{\tau(c)}\right\rangle,$$

the square terms still contain another small factor M/m; it appears sufficient, therefore, to keep only the leading terms which are the ones containing $\langle c_z\rangle^2$ as factor. All these terms are multiplied with the ratio of two angular averages over $d\chi$; these may be taken as independent of c to a good approximation. Equation (70) thus takes the form

$$(4-\alpha)h(\mathbf{c})+c_x\frac{\partial h(\mathbf{c})}{\partial c_x}+c_y\frac{\partial h(\mathbf{c})}{\partial c_y}+(2-\alpha)(c_z-\langle c_z\rangle)\frac{\partial h(\mathbf{c})}{\partial c_z}$$

$$+\frac{1}{4}\frac{M}{m}\frac{\left\langle\frac{\sin^2\chi}{\tau}\right\rangle}{\left\langle\frac{1-\cos\chi}{\tau}\right\rangle}\langle c_z\rangle^2\left\{\frac{\partial^2 h(\mathbf{c})}{\partial c_x^2}+\frac{\partial^2 h(\mathbf{c})}{\partial c_y^2}\right\}\tag{72}$$

$$+\frac{1}{2}\frac{M}{m}\frac{\left\langle\frac{(1-\cos\chi)^2}{\tau}\right\rangle}{\left\langle\frac{(1-\cos\chi)}{\tau}\right\rangle}\langle c_z\rangle^2\frac{\partial^2 h(\mathbf{c})}{\partial c_z^2}=0$$

The equation can be solved explicitly in Cartesian coordinates by the method of separation of variables. The result is

$$h(\mathbf{c}) = \exp\left[-2\frac{m}{M}\frac{\left\langle\frac{1-\cos\chi}{\tau}\right\rangle}{\left\langle\frac{\sin^2\chi}{\tau}\right\rangle}\frac{c_x^2+c_y^2}{\langle c_z\rangle^2} - (2-\alpha)\frac{m}{M}\frac{\left\langle\frac{1-\cos\chi}{\tau}\right\rangle}{\left\langle\frac{(1-\cos\chi)^2}{\tau}\right\rangle}\frac{(c_z-\langle c_z\rangle)^2}{\langle c_z\rangle^2}\right] \tag{73}$$

This is a Maxwellian distribution with elliptic distortion and shifted origin, that is, the type shown in Fig. 1 (b).

The result (73) indicates the main features of the solution for heavy ions. Because of the neglect of derivatives higher than the second in $h(\mathbf{c}')$ it is not certain that (73) is correct in all details, even in the limit of very large m/M.

IID. THE CASE OF EQUAL MASSES; IONS TRAVELLING IN THE PARENT GAS

The developments of the previous section show that if the ion mass is either large or small in comparison to the molecular mass, analytical methods can be applied successfully to determine the velocity distribution of the ions. No such possibility was found for the mass ratio unity, which one would judge to be of particular interest because it applies to ions travelling in the parent gas. There exist isolated fragments of such an analytical theory; for instance, if we assume isotropic scattering in (46), that is $\Pi(\chi) = 1$, then the zeroth equation (46) becomes explicitly integrable and yields

$$h_0(c) = -\frac{1}{3}a\tau(c)\frac{dh_1(c)}{dc} \tag{74}$$

This is a curious reversal of the differential relationship (63) derived for electrons and implies a rather strong condition on the structure of $h_1(c)$. However, I have not been able to consolidate these fragments into something which can be used successfully in computation. The high field distribution function for mass ratio unity appears, however, sufficiently interesting to warrant the use of other methods.

A numerical determination of the velocity distribution function was undertaken in cooperation with R. W. Hamming by the so-called Monte Carlo method. The Monte Carlo method is a way of gaining

statistical information about a system by following an individual member through a large number of random processes. The result of such a procedure is knowledge about one member of the assembly for a long period of time. Time averages of various kinds can be obtained from such data; these time averages are then set equal to instantaneous averages over the assembly, in accordance with ergodic theory. In our case, an ion was followed through 10,000 collisions. On an average, the collisions were isotropic in the center of mass system ($\Pi(\chi) = 1$) and obeyed a mean free time condition $\tau = \text{const}$. Actually, both the free time and the scattering angles varied from collision to collision; the angles varied in a random fashion over a unit sphere and τ was random within an exponential distribution.

A Monte Carlo calculation of this type consists of three parts. In the first part the random numbers having the required distributions are obtained and recorded. In the present problem there were three such random numbers required for each collision, namely a time and two angles. These numbers were placed on 10,000 IBM cards, along with suitable identification. In the second part a calculating machine simulates the successive collisions and keeps a record of the initial and final velocities for each one. The third part consists in analyzing statistically the numerical material accumulated in the second. For the first part of the calculation particular values must be chosen for the acceleration a and the mean free time τ. These values were

$$a = 1$$

$$\tau = \log_{10} e = 0.43429$$

However, the dimensional analysis of Section ID shows us at this point that these two constants enter into the problem only through their product $a\tau$ which scales all velocities. It is therefore convenient at the statistical stage to remove these factors and to analyze the results in terms of a dimensionless variable which by (26c) we take in the form

$$\mathbf{w} = \frac{\mathbf{c}}{a\tau} \tag{75}$$

In view of the a priori information for mean free time problems which is gathered in Section IIB we can use the statistical data from the Monte Carlo calculation in two ways. We may (a) check the numerical computation itself or (b) gain new information not available otherwise.

(a) The check of the numerical calculation by statistical analysis proceeds as follows. From deductive reasoning we have obtained the

averages (52), (54) and (57) for c_z, c_x^2 and c_z^2. These formulas ought to be verified in the Monte Carlo calculation. This is indeed approximately true. A sampling covering 9492 out of the 10,000 collisions gives

	by Monte Carlo	by deduction
$\langle w_z \rangle$	1.912	2
$\langle w_x^2 \rangle$	0.801	$\frac{8}{9} = 0.889$
$\langle w_z^2 \rangle$	5.165	$\frac{56}{9} = 6.222$

The agreement is essentially there but deviations are noticeable. In judging these we have to realize that fluctuations are quite large in this problem. For instance if the calculation is broken down into ten runs of approximately 1,000 events each one finds the following time averages for the partial runs:

	$\langle w_z \rangle$	$\langle w_x^2 \rangle$	$\langle w_z^2 \rangle$
	1.821	0.837	4.453
	1.975	0.748	5.531
	1.915	0.785	5.132
	1.954	0.829	5.441
	1.868	0.731	4.794
	2.003	0.807	5.581
	1.766	0.793	4.811
	1.962	0.827	5.360
	2.003	0.901	5.471
	1.914	0.727	5.200
predicted	2.000	0.889	6.222

Among these runs there are some having averages higher than the predicted values, but the data clearly show that the Monte Carlo averages are generally lower. In the search for reasons it was first felt that perhaps the desired mean value for τ is not actually reached, perhaps through systematic errors introduced by the operator when rejecting certain runs. This seems indeed to be the case. The mean free time obtained from the 9492 runs mentioned above is

$$\tau = 0.4269$$

which is slightly low. Indeed it is observed that the runs with high τ were particularly troublesome in the calculation and were preferably rejected by the operator. It seems doubtful however that this error

could account for the entire discrepancy, particularly in the mean squares, although it must be emphasized that the runs with high τ make a more than proportional contribution to the total average. The angles of scattering have not been subjected to a similar analysis so that we cannot make a statement whether the aimed at isotropy in the law of scattering was realized or not. We conclude therefore by saying that while the Monte Carlo calculation gives results in general agreement with the deductive theory there are small but noticeable systematic errors in it whose origin is only partly explained. Similar errors must exist in the new results which cannot be compared with theoretical predictions.

(b) In this part we will discuss the velocity distribution function which may be constructed from the Monte Carlo results. In constructing such a function we make use of the fact that, between collisions, the velocity is accelerated at a uniform rate. Thus, in each period between two collisions, the velocity vector traces out a straight line parallel to the w_z axis covering equal distances in equal times. The Monte Carlo calculation furnishes us with a number of such straight lines as shown in Fig. 11. The density of these straight line tracks in velocity space is the velocity distribution function. The actual procedure used to obtain it was to lay a grid with a mesh of 0.23 in a half-plane with coordinates w_z and $w_\rho = \sqrt{w_x^2 + w_y^2}$ and to count the number of lines crossing each horizontal square edge. When the resultant count is converted to density

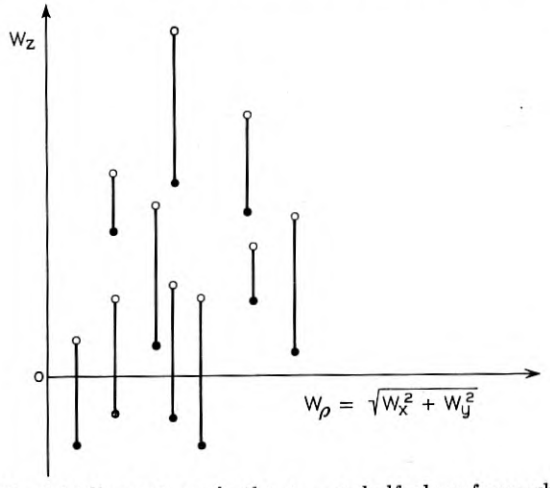

Fig. 11 — Straight line pattern in the $w_z - w_\rho$ half plane from which the velocity distribution is constructed; the Monte Carlo calculation furnishes the initial and final velocities (dots and rings).

and normalized to 1 we get a distribution in the $w_z - w_\rho$ half plane which is shown in Fig. 12. Division by $2\pi w_z$ will transform it into a conventional distribution function in velocity space; a plot of this function is shown in Fig. 13. What distinguishes this distribution function from functions previously proposed is the elongated probability contours. This feature is not unexpected in view of the unequal energy partition apparent in the equations (58) and (60).

The probability contours shown in Figs. 12 and 13 give a reliable general picture but we must not expect from them fine detail. Indeed we will now prove that the distribution function is infinite along the entire positive c_z-axis, a feature which is not obvious from inspection of Fig. 13.

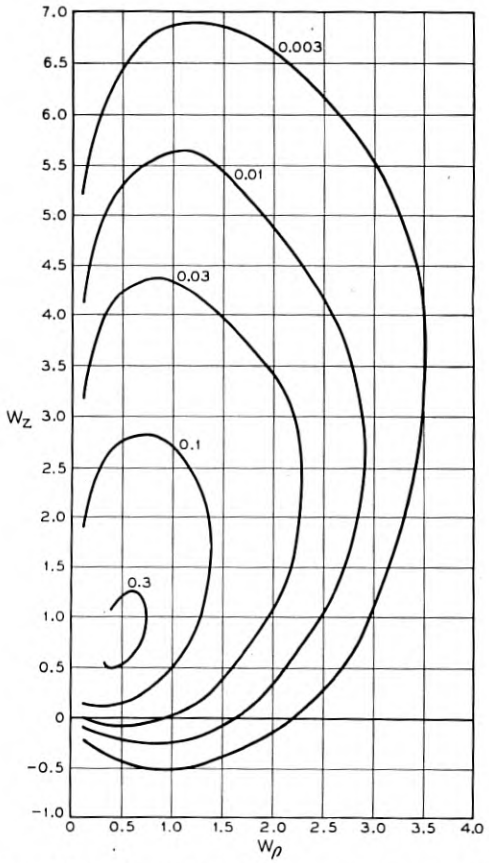

Fig. 12 — Motion of ions through the parent gas in a high field; distribution of velocities in the $w_z - w_\rho$ half plane resulting from the Monte Carlo calculation.

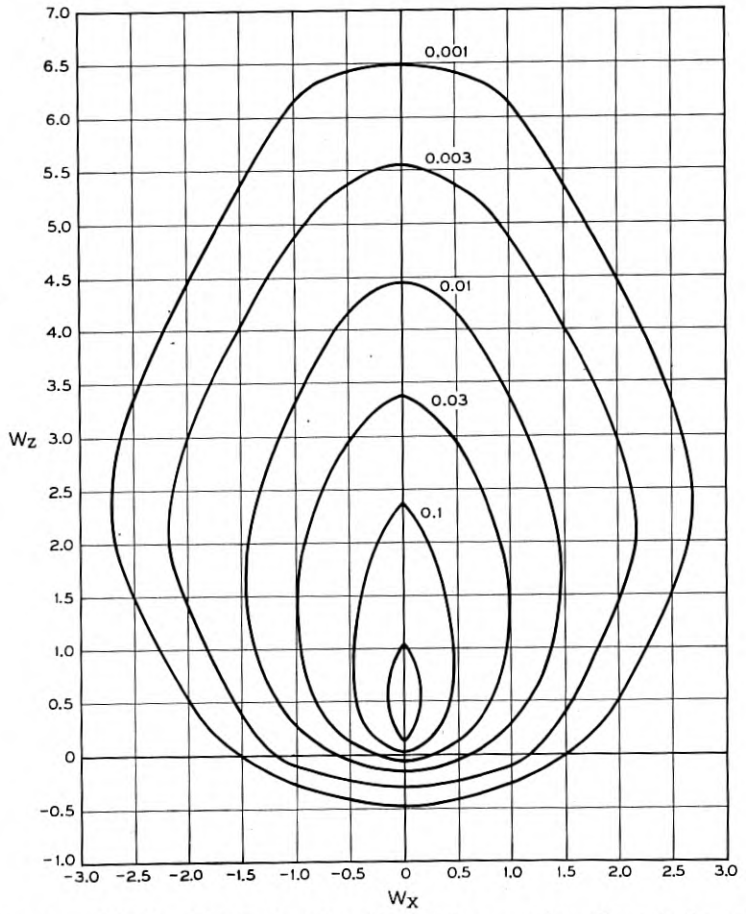

Fig. 13 — Velocity distribution function of ions moving through the parent gas in a high field; contours constructed from the Monte Carlo results of Fig. 12.

A simple physical proof of this statement goes as follows. Suppose an ion and a molecule make a collision which is almost central, but has a small impact parameter b. The collision will bring the ion almost to rest because the atom was originally at rest by hypothesis. Because the collision was not quite central, however, the ion will have a small residual velocity $\mathbf{c}_f$ at right angles to its original velocity $\mathbf{c}_i$. For any reasonable law of scattering this quantity c_f will be proportional to c_i and to b.

$$c_f \propto c_i \cdot b$$

The probability for a value of b between b and $b + db$ is proportional

to $b\ db$. Thus even if all c_i's were equally probable the probability for c_f would vary as $c_f\, dc_f$. Actually very small c_i's may be specially probable as the theorem states and this fact may or may not increase the probability for small c_f. This means that $P(c_f)dc_f$ probably varies as $c_f dc_f$, and may perhaps even contain a smaller power of c_f. When such a probability function is plotted in velocity space it will vary as $1/c_f$. Thus we know that the distribution function $\varphi(\mathbf{c})$ for ions immediately following a collision has a singularity at the origin at least as $1/c_f$. The actual distribution function $h(\mathbf{c})$ is derived from this one by spreading each point out in the forward direction as shown in Fig. 11. For the mean free time case we can write this out explicitly in the form

$$h(\mathbf{c}) = \frac{1}{\tau}\int_0^\infty \varphi(\mathbf{c} - \mathbf{a}t)\, e^{-\frac{t}{\tau}}\, dt \qquad (76)$$

For the case of a mean free path or other laws the formula is more awkward but they all differ from the above only in replacing $e^{-\frac{t}{\tau}}$ by a more complicated weight function. The singularity of h arises out of the singularity of φ which contains at least a factor $1/\sqrt{c_x^2 + c_y^2 + (c_z - at)^2}$. Along the c_z-axis, this become a factor $1/(c_z - at)$; this factor makes the integral diverge for all positive c_z ; as we approach the origin from negative c_z's the distribution function will become infinite at least as $\ell n\ 1/c_z$.

The reasoning given is intrinsicly classical because of the use of "infinitely small" impact parameters. We should not hasten to conclude, however, that the quantization of the angular momentum will necessarily remove the singularity. Indeed we know that the only mechanical information which has to be put into the Boltzmann equation (31) is the differential cross-section for scattering. If this quantity does not differ essentially in the 180° direction from a classical cross section then it will not modify the conclusion we have reached.

Conclusions which are more informative, but less "anschaulich" may be obtained from a study of Boltzmann's equation either in its closed form (34) or (40) or its "Legendre" form (46) or (47). In view of the proof given we will give only an outline of the reasoning. First we can remove the second term in (40) by the substitution

$$h(\mathbf{c}) = \exp\left[-\int_0^{c_z} \frac{dc_z}{a\tau(c)}\right] h^*(\mathbf{c})$$

The exponential is easily seen to be always positive and finite for finite c. The Boltzmann equation then takes the form

$$\exp\left[-\int_0^{c_z} \frac{dc_z}{a\tau(c)}\right] a\frac{\partial h^*}{\partial c_z} = \frac{1}{c}\cdot(\text{an integral})$$

For equal masses the integral on the right runs over a plane in velocity space. Its integrand is always positive; hence the integral can never vanish and is always positive; it is conceivable that it could be infinite for a special position provided the infinity is integrable. Such an infinity would only make matters worse. At any rate the equation shows that $h^*(\mathbf{c})$ increases monotonely with increasing c_z. When we approach the origin from negative c_z we get a logarithmic divergence (or worse). As the function $h^*(\mathbf{c})$ can nowhere decrease with increasing c_z the infinity along the positive c_z axis is confirmed and its logarithmic nature is made very likely.

Information obtainable from equation (46) confirms this conclusion. Lowest powers in the entire recursion system can be made to cancel by assuming that for small c

$$h_0 \sim -A \ell n w$$

$$h_i \sim B_i \qquad i > 0$$

with suitable relationships existing between these quantities.

A defect of all three approaches is that they give no information concerning the nature of the infinity for $c_z > 0$. One is tempted to conclude from Fig. 13 that it cannot be very strong. Something like a singularity is discernible at the origin, particularly if the contour 0.1 is drawn back to cut the w_z-axis at a negative value; this is perfectly compatible with the available information. For large positive w_z, on the other hand, the picture almost contradicts the theorem just proved. One concludes from this that the singularity, for large c_z, becomes a weak and narrow ridge rising more or less abruptly in an otherwise well behaved function.

IIE. THE CASE OF EQUAL MASSES; A NEW COMPUTATIONAL PROCEDURE

The foregoing sections have accumulated substantial evidence that there are many analytical details involved when one discusses the structure of a velocity distribution function. These details are of little interest to the experimenter who may want nothing but a formula for the drift velocity or the average energy. In view of this situation it appears very desirable to find a method whereby such quantities can be derived directly and accurately from the Boltzmann equation without a full knowledge of the entire distribution.

Maxwell's original work shows us how to achieve this for molecules obeying the mean free time condition of Section IIB. In the following, a general method is described which will permit determination of such

averages for an arbitrary law of force between the ions and the gas molecules, and an arbitrary mass ratio. The application will be limited to the case of the mass ratio 1 whose study was begun in the preceding section.

The basis of the method is an observation on the equation system (46) or (47), which is the form taken by the Boltzmann equation after inserting the Legendre decomposition (43). It would appear at first sight that these recursion relations are of such a structure that an arbitrary function $h_0(c)$ could be substituted into the "zeroth" equation and that the relations would then successively determine h_1, h_2, $h_3 \cdots$. Upon closer inpsection this is found not to be the case. Suppose we have obtained somehow functions h_0, h_1, $h_2 \cdots h_n$ and we are trying to use the nth equation to determine h_{n+1}. This equation is of the form

$$\frac{dh_{n+1}}{dc} + \frac{n+2}{c} h_{n+1} = \text{known material} \tag{77}$$

We solve for h_{n+1} by multiplying with c^{n+2} and integrating. This gives

$$c^{n+2} h_{n+1}(c) = \int^c (\text{known material})\, dc$$

The left-hand side is of such a structure that it must vanish both for $c = 0$ and $c = \infty$. It follows that the right-hand integral when taken between the limits 0 and ∞ must equal zero. This condition is indeed obeyed for any $h_0(c)$ when $n = 0$. The integral condition reads in this case

$$\frac{1}{2} \int_0^\infty c^2\, dc \int_0^\pi \Pi(\chi) \sin \chi\, d\chi \frac{h_0(c')}{\tau(c')} \left(\frac{c'}{c}\right)^3 - \int_0^\infty \frac{h_0(c)}{\tau(c)} c^2\, dc = 0$$

If we invert the order of integration in the double integral, then introduce c' as variable of integration by equation (41) and finally invert again this becomes

$$\int_0^\infty \frac{h_0(c)}{\tau(c)} c^2\, dc \left[\frac{1}{2} \int_0^\pi \Pi(\chi) \sin \chi\, d\chi - 1\right] = 0$$

This equation is trivally obeyed because the square bracket vanishes in virtue of the definition of $\Pi(\chi)$. For values of n higher than 0, the integrability condition deduced from (77) is not generally obeyed for any function $h_0(c)$. Such a statement may be proved by examples; these examples will arise in the course of the calculations to follow. Thus we find that except in the passage from $h_0(c)$ to $h_1(c)$, the recursion system is such that at each stage it imposes a condition upon the h_ν's already determined if the new $h_{n+1}(c)$ is to exist at all. With such an infinity of

conditions one can improve indefinitely an initial trial function assumed for $h_0(c)$.

The integrability conditions whose general structure is thus indicated have actually already been written down. They are the equations (49) for the special case $s = \nu$. Generally speaking, the relations (49) are also of the recursion type, permitting us to start with arbitrary averages $\langle c^s \rangle$, and computing successively $\langle c^s P_1 (\cos \vartheta) \rangle$ etc. At every stage, however, there is the exception mentioned: the equation for which $s = \nu$ has no third member, and therefore it imposes a condition upon averages already known from the previous equations. We shall refer to this type of equation as a "truncated" relation.

It is reasonable to assume that $1/\tau(c)$ can be developed into a power series in c because it equals the known constant polarization value for $c = 0$. If this can be assumed then each truncated relation $s = \nu$ is equivalent to a unique relation among velocity averages involving $h_0(c)$ only. One obtains this relation by applying to each member in the truncated relation its own recursion formula and repeating this process until ν is brought down to zero. This process will never lead into another truncated relation $s' = \nu'$ because at each step s' increases by at least two units with respect to ν'.

In order to test the method for a known case, it will be applied first to the case of constant mean free time. This case is adequately described by the theoretical treatment of Section IIB and the Monte Carlo calculation of Section IID. We have seen that the equations (49) reduce in this case to the form (51) which dovetails as shown in Fig. 9; this dovetailing leads to explicit values for certain averages as shown in Fig. 10. A "computational method" is only needed when one tries to get an average outside this selected list. In the present case the reduction of the truncated relations to a condition on $h_0(w)$ is particularly simple as is seen from Fig. 9. A singular relation which starts out as between $\langle c^{\nu-1} P_{\nu-1} (\cos \vartheta) \rangle$ and $\langle c^\nu P_\nu (\cos \vartheta) \rangle$ actually yields the numerical value of the latter because the former has been obtained numerically in a previous stage. This numerical value yields in combination with previous information $\langle c^{\nu+1} P_{\nu-1} (\cos \vartheta) \rangle$, $\langle c^{\nu+2} P_{\nu-2} (\cos \vartheta) \rangle$ etc and finally $\langle c^{2\nu} \rangle$. Thus we end up with the set of even moments of $h_0(c)$ which may be used in succession to determine $h_0(c)$ more and more closely. There is no guarantee that this procedure converges mathematically, since the general theorems usually require the knowledge of all integer moments.[20]

[20] Shohat, J. A., and J. D. Tamarkin, The Problem of Moments. Am. Math. Soc., 1943. The original three-dimensional formulation appears a little more favorable for a proof because, in this case, we know indeed all integer moments.

The justification for the method rests therefore on an empirical basis at this point.

Assuming isotropic scattering, as in the "Monte Carlo" calculation we express our results in terms of the dimensionless variable **w** defined in (75). The equation system (51) becomes then

$$\begin{aligned}(2\nu - 1)(1 - \langle I_{s,\nu}(\chi)\rangle)\langle s, \nu\rangle &= \\ = \nu(\nu + s + 1)\langle s - 1, \nu - 1\rangle + (\nu + 1)(\nu - s)\langle s - 1, \nu + 1\rangle &\end{aligned} \tag{78}$$

where the abbreviation $\langle s, \nu\rangle$ has been introduced for $\langle w^s P_\nu (\cos \vartheta)\rangle$ and the quantities $\langle I_{s,\nu}(\chi)\rangle$ are simple numbers computable from (48b) and the assumption of isotropic scattering. The first truncated relation is $s = \nu = 1$. It yields

$$\langle 1, 1\rangle = 2$$

Reducing it with the relation (78) for which $s = 2$, $\nu = 0$ we get

$$\langle 2, 0\rangle = 8 \tag{79}$$

The next truncated relation is $s = \nu = 2$, which yields

$$\langle 2, 2\rangle = \frac{16}{3}$$

and the reduction gives

$$\begin{aligned}\langle 3, 1\rangle &= \frac{92}{3} \\ \langle 4, 0\rangle &= 184\end{aligned} \tag{80}$$

Similarly in the next stage

$$\begin{aligned}\langle 3, 3\rangle &= \frac{128}{7} \\ \langle 4, 2\rangle &= \frac{18112}{133} \\ \langle 5, 1\rangle &= \frac{421600}{399} \\ \langle 6, 0\rangle &= \frac{3372800}{399}\end{aligned} \tag{81}$$

As an example of an average which cannot be had explicitly we may take the mean absolute value of the speed, that is $\langle 1, 0\rangle$. We find this

value by picking a sequence of trial functions for $h_0(w)$ with the appropriate number of parameters and imposing successively (79), (80) and (81) upon this sequence; this leads us to a sequence of values for $\langle w \rangle$ which can then be examined. In such a procedure careful consideration of the trial functions is an important element. The following information is available. It was proved in Section IID that $h_0(w)$ is logarithmically infinite at the origin. At infinity, on the other hand, $h_0(w)$ falls as e^{-w} times some power of w. One way to check this is to drop the terms containing $1/c$ as factor in (46); the solution of the recursion system becomes then

$$h_\nu(w) \sim (2\nu + 1)e^{-w}w^k$$

where k is some unknown exponent. Armed with this fore-knowledge, we shall use the following sequence of trial function for $h_0(w)$

$$h_0(w) = pEi(w) + qK_0(w) + re^{-w} + swK_1(w) \tag{82}$$

where

$$Ei(w) = \int_w^\infty \frac{e^{-u}}{u}\, du$$

and $K_0(w)$, $K_1(w)$ are the modified Hankel functions of order zero and 1.[21]

We find in zeroth approximation from normalization only

$$p^{(0)} = \frac{3}{2} \qquad q^{(0)} = r^{(0)} = s^{(0)} = 0$$

$$\langle w \rangle^{(0)} = 2.2500 \tag{83a}$$

in first approximation, using (79)

$$p^{(1)} = \frac{5}{6}$$

$$q^{(1)} = \frac{4}{9} \qquad r^{(1)} = s^{(1)} = 0$$

$$\langle w \rangle^{(1)} = 2.3818 \tag{83b}$$

---

[21] This definition, which is in accord with the tables of Jahnke-Emde, differs from the usual one by a factor $2/\pi$. This change is suggested by Watson, Bessel Functions, p. 79, and proves convenient in the following.

in second approximation, using (79) and (80)

$$p^{(2)} = \frac{7}{12}$$

$$q^{(2)} = \frac{32}{45} \qquad s^{(2)} = 0$$

$$r^{(2)} = -\frac{1}{20}$$

$$\langle w \rangle^{(2)} = 2.3858 \tag{83c}$$

and in third approximation, using (79), and (80) and (81)

$$p^{(3)} = \frac{3079}{7980}$$

$$q^{(3)} = \frac{202544}{209475}$$

$$r^{(3)} = -\frac{2507}{18620}$$

$$s^{(3)} = \frac{3152}{209475}$$

$$\langle w \rangle^{3} = 2.3864 \tag{83d}$$

Appearances indicate strongly that the sequence (83) for $\langle w \rangle$ approaches a limit which one would guess to be

$$\langle w \rangle = 2.3865 \tag{84}$$

More evidence that the conclusion drawn is correct can be obtained by using the set of trial functions

$$h_0(w) = pK_0(w) + qe^{-w} + rwe^{-w}$$

We find then the following sequence of values for $\langle w \rangle$.

$$\langle w \rangle^{(0)} = 2.546 \qquad \langle w \rangle^{(1)} = 2.395 \qquad \langle w \rangle^{(2)} = 2.388$$

This descending sequence confirms (84) by approaching this same value from above.

Further evidence for the correctness of the procedure can be obtained by deriving a function $h_0(w)$ from the Monte Carlo function $h(\mathbf{w})$ discussed in Section IID and comparing it with our trial function. The function was constructed by covering Fig. 13 with a grid of concentric

circles and horizontal lines and replacing the integration

$$h_0(w) = 2\pi \int_{\vartheta=0}^{\vartheta=\pi} h(\mathrm{w})\, d\,(\cos\vartheta)$$

by a summation over grid points. The function $h_0^+(w)$ so obtained is compared in Table I with $h_0^{(0)}(w)$, $h_0^{(1)}(w)$ and $h_0^{(2)}(w)$ as defined by (82) and the numbers following. We observe that the first approximation

TABLE I

Comparison of the Monte Carlo $h_0^+(w)$ for $h_0(w)$ with successive approximations obtained by the new method.

w	$h_0^+(w)$	$h_0^{(0)}(w)$	$h_0^{(1)}(w)$	$h_0^{(2)}(w)$
0		∞	∞	∞
0.5	0.74	0.8397	0.7281	0.7144
1	0.29	0.3291	0.3019	0.3002
1.5	0.15	0.1500	0.1438	0.1440
2	0.081	0.0734	0.0730	0.0733
2.5	0.0412	0.0374	0.0384	0.0387
3	0.0199	0.0196	0.0207	0.0208
3.5	0.0118	0.0105	0.0114	0.0114
4	0.0063	0.0057	0.0063	0.0063
4.5	0.0030	0.0031	0.0035	0.0036
5	0.0014	0.0017	0.0020	0.0020
6	0.0004	0.0005	0.0007	0.0006
7	0.0001	0.0002	0.0002	0.0002

is an improvement over the zeroth one, while the second one makes little difference, considering the accuracy to which $h_0^+(w)$ is given. In individual cases the sequence drifts away from $h_0^+(w)$; this is not surprising because the latter function is very rough; this is to be expected from its mode of derivation.

The application of this method to the hard sphere model of ion-atom collisions offers no new feature of principle. The actual working out of results is somewhat more complicated, mainly because the connection diagram for the recursion system (49) is more involved. According to equation (26b) the dimensionless variable to be used in this work is

$$\mathrm{w} = \frac{\mathrm{c}}{\sqrt{a\lambda}} \tag{85}$$

We denote its averages $\langle w^s P_\nu\,(\cos\vartheta)\rangle$ by $\langle s, \nu\rangle$ as previously. The equa-

tion system (49) then takes the form

$$(2\nu + 1)(1 - \langle I_{s,\nu}(\chi)\rangle)\langle s + 1, \nu\rangle = \\ = \nu(\nu + s + 1)\langle s - 1, \nu - 1\rangle + (\nu + 1)(s - \nu)\langle s - 1, \nu + 1\rangle \tag{86}$$

The numbers $\langle I_{s,\nu}(\chi)\rangle$ were already discussed in connection with the system (78). What distinguishes (86) from (78) is the way in which the variables are connected; the new connection diagram which replaces Fig. 9 is shown in Fig. 14. The truncated relations no longer dovetail into each other as they did before. Only the first stage proceeds in a similar way, yielding explicit expressions for $\langle 2, 1\rangle$ and $\langle 4, 0\rangle$. In the next stage we

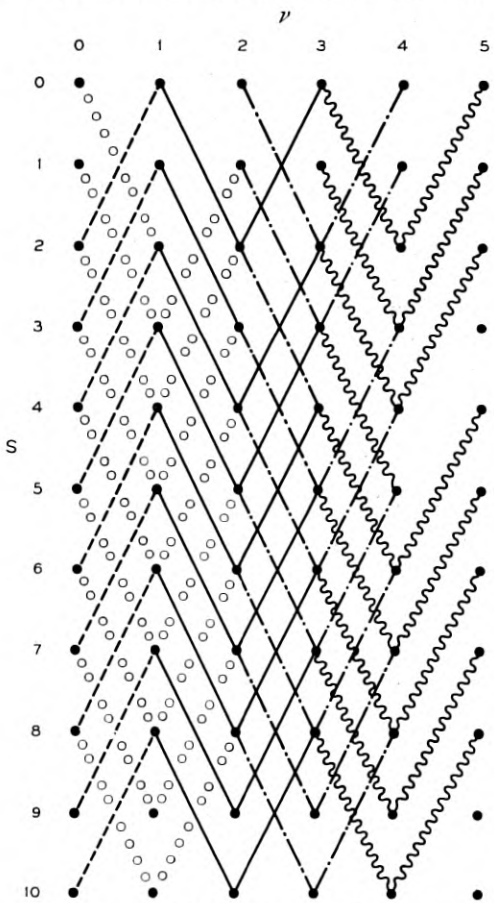

Fig. 14 — Interconnection established by the Boltzmann equation among the averages $\langle c^s P_\nu(\cos\vartheta)\rangle$; case of constant mean free path.

start out with a relation between $\langle 1, 1\rangle$ and $\langle 3, 2\rangle$. By the use of regular recursion formulas we can successively transform this into a relation between $\langle 3, 0\rangle$ and $\langle 3, 2\rangle$, then $\langle 3, 0\rangle$ and $\langle 5, 1\rangle$ and finally between $\langle 3, 0\rangle$ and $\langle 7, 0\rangle$. Here we have for the first time the normal situation in which we do not get the actual value of a moment of $h_0(c)$ but only a relation between two or more of such moments; the reason for this is that the system fails to connect up with $\langle 0, 0\rangle$ which equals unity a priori. A similar situation prevails for the next truncated relation; it is originally a relation between $\langle 2, 2\rangle$ and $\langle 4, 3\rangle$ and is finally reduced to one between $\langle 2, 0\rangle$, $\langle 6, 0\rangle$ and $\langle 10, 0\rangle$. Similarly, the next truncated relation reduces to a relation between $\langle 5, 0\rangle$, $\langle 9, 0\rangle$ and $\langle 13, 0\rangle$ and so forth. The first three of these reduced relations come out to be

$$\langle w^4\rangle = 10 \tag{87}$$

$$3\langle w^7\rangle = 112\langle w^3\rangle \tag{88}$$

$$\frac{295}{56}\langle w^6\rangle = 27\langle w^2\rangle + \frac{17}{330}\langle w^{10}\rangle \tag{89}$$

These formulas will now be imposed upon a sequence of trial functions for $h_0(w)$ suitably chosen. Again, we may make use of the information of Section IID, according to which $h_0(w)$ is logarithmically singular at the origin. For large w we proceed as previously from (46) leaving off the terms of $1/c$. We get then

$$h_\nu(w) \sim (2\nu + 1)e^{-\frac{1}{4}w^2}w^k$$

This suggests the following trial function for $h_0(w)$

$$h_0(w) = pEi(\tfrac{1}{2}w^2) + qK_0(\tfrac{1}{2}w^2) + re^{-\frac{1}{4}w^2} + sw^2K_1(\tfrac{1}{2}w^2) \tag{90}$$

The best zero order approximation is actually obtained by the function $K_0(\frac{1}{2}w^2)$. We find

$$q^{(0)} = \frac{\pi}{2\Gamma^2(\frac{3}{4})} = 1.04605 \qquad p^{(0)} = r^{(0)} = s^{(0)} = 0$$

In first order we get, using (87)

$$p^{(1)} = -0.46543 \qquad r^{(1)} = s^{(1)} = 0$$
$$q^{(1)} = 1.45285$$

In the second order, using (87) and (88)

$$p^{(2)} = -0.80856$$
$$q^{(2)} = 1.88127 \qquad s^{(2)} = 0$$
$$r^{(2)} = -0.09804$$

In third order, using (87) and (88) and (89)

$$p^{(3)} = -1.15071$$
$$q^{(3)} = 2.37034$$
$$s^{(3)} = 0.02062$$
$$r^{(3)} = -0.29016$$

These successive approximations lead to the following sequence for the drift velocity $\langle w \cos \vartheta \rangle$

$$\langle w \cos \vartheta \rangle^{(0)} = 1.04605 \tag{91a}$$
$$\langle w \cos \vartheta \rangle^{(1)} = 1.14256 \tag{91b}$$
$$\langle w \cos \vartheta \rangle^{(2)} = 1.14616 \tag{91c}$$
$$\langle w \cos \vartheta \rangle^{(3)} = 1.14661 \tag{91d}$$

We conclude from this sequence that

$$\langle w \cos \vartheta \rangle = 1.1467 \tag{92}$$

In addition to the drift velocity there is some interest in the energy and the energy partition. For the energy the following numbers are obtained

$$\langle w^2 \rangle^{(0)} = 2.1884 \tag{93a}$$
$$\langle w^2 \rangle^{(1)} = 2.3395 \tag{93b}$$
$$\langle w^2 \rangle^{(2)} = 2.3511 \tag{93c}$$
$$\langle w^2 \rangle^{(3)} = 2.3531 \tag{93d}$$

giving

$$\langle w^2 \rangle = 2.353 \tag{94}$$

A zero order value for $\langle w^2 \cos^2 \vartheta \rangle$ cannot be said to exist because the first truncated relation is the condition that a distribution function $h_2(w)$ exists at all. Thus, we can get only three numbers in a sequence approximating $\langle w^2 \cos^2 \vartheta \rangle$

$$\langle w^2 \cos^2 \vartheta \rangle^{(1)} = 1.8005 \tag{95a}$$
$$\langle w^2 \cos^2 \vartheta \rangle^{(2)} = 1.7696 \tag{95b}$$
$$\langle w^2 \cos^2 \vartheta \rangle^{(3)} = 1.7685 \tag{95c}$$

giving

$$\langle w^2 \cos^2 \vartheta \rangle = 1.768 \tag{96}$$

We can understand the results (92), (94) and (96) by giving the fraction of the total energy in ordered motion and the fraction of the energy in motion along the z-direction. We find for the first ratio

$$\frac{\langle w \cos \vartheta \rangle^2}{\langle w^2 \rangle} = 0.559 \tag{97}$$

and for the second

$$\frac{\langle w^2 \cos^2 \vartheta \rangle}{\langle w^2 \rangle} = 0.751 \tag{98}$$

The ratio (97) equals 0.5000 for all mean free time models; the ratio (98) is 0.778 for the mean free time model with isotropic scattering. Thus, the deviations from the earlier results are not drastic. However, in certain derived relations the difference is more noticeable. For instance, a good measure of the anistropy of the diffusion process is furnished by the ratio of the random energy along the field to the energy at right angles.[22] From (97) and (98) we find for this number

$$\frac{\langle w^2 \cos^2 \vartheta \rangle - \langle w \cos \vartheta \rangle^2}{\frac{1}{2}(\langle w^2 \rangle - \langle w^2 \cos^2 \vartheta \rangle)} = 1.54 \tag{99}$$

For the mean free time case this number equals 2.50. Hershey[6] in his work assumes this number to be 1.000.

A comprehensive list of velocity averages is attached in Table II. As a comment I may add that the obvious mode of constructing such a table, namely by computing the column $\nu = 0$ from (90) and then using the recursion system (86) for the others, runs into some difficulty. First of all, a series of cancellations reduces the accuracy as ν increases; finally, at the positions marked "impossible" we find the missing third members of the truncated relations. These elements cannot be computed by recursion at all, but would require an explicit solution of the equation system (47) for $h_{\nu+1}(c)$. In the table, this more arduous path is not followed. Instead, the recursion method is used for the numbers in italic type and a few numbers are added by extrapolation. The numbers so obtained will be needed in Section IVB.

The calculations on the hard sphere model are immediately applicable to the experimental data of Figs. 3 to 7, which exhibit the drift velocity of

[22] See below, equations (147) and (165).

the noble gas ions in the parent gas as functions of the parameter a/N. These data have a high field range in which the drift velocity varies as the square root of a/N. This is the variation for a model with constant mean free path, as seen from the dimensional formula (26b). It was indicated furthermore in the Section IA that we have good reason to think of the scattering between an ion and an atom as nearly isotropic.[23] These two features characterize uniquely the hard sphere model whose treatment we have just completed. To the extent that they are verified

TABLE II

Dimensionless high-field velocity averages $\langle s, \nu \rangle$ for the hard sphere model and mass ratio unity.

ν	0	1	2	3	4
s					
0	*1.0000*	0.7845	impossible		
1	*1.3923*	*1.1467*	*0.8022*	impossible	
2	*2.3534*	*2.0000*	*1.4759*	0.990	impossible
3	*4.5868*	*3.9853*	*3.0578*	2.134	
4	*10.0000*	*8.8353*	*6.992*	*5.0602*	3.474
5	*23.912*	*21.405*	17.330	12.84	
6	*61.847*	*55.97*	*46.177*	35.36	
7	*171.241*	*156.3*	130.91		
8	*503.7*	*462.81*			
9	*1563*	1445			
10	*5090.9*	4750			

the model is applicable to the experimental data. The formula to apply is (92) in combination with (85):

$$\langle c_z \rangle = 1.147 \sqrt{\frac{a}{N\sigma}} \tag{100}$$

In the logarithmic plot of $\langle c_z \rangle$ vs a/N the intercept of the straight line of slope 1/2 which fits the high field data thus equals

$$\log \frac{1.147}{\sqrt{\sigma}}.$$

The values for σ which result from this are shown in Table III. For comparison are shown the corresponding atomic cross section as determined from viscosity data.[24] It is interesting to observe that the ratio of

[23] A quantitative discussion of this point for the polarization force will follow in Section IIIB.

[24] Landolt-Börnstein, 1950 edition, Vol. I, part 1, page 325.

the two retains very nearly the constant value 3 throughout the table. The fact that the ratio is substantially larger than unity is explained by the resonance feature of the ion-atom scattering process as discussed in Section IA. The fact that it is constant is perhaps an indication of the fact that both processes are governed by overlap conditions of essentially the same wave functions.

I would like to point out in connection with the calculations of this section that the method developed is potentially of very wide application. One question that comes up, for instance, is whether a careful kinetics calculation is necessarily restricted to certain models or whether an ion-atom cross section known numerically could be used to derive therefrom kinetic properties. This is indeed possible. Suppose, for instance, that the cross section $\sigma(c)$ were available as a function of c for collision of He^+-ions and He-atoms and suppose that this cross section were to satisfy the condition of isotropy $\Pi(\chi) = 1$ to a good approxi-

TABLE III

Cross sections for ion-atom and atom-atom collisions for the noble gases.

Gas	ion-atom cross section $\times 10^{16}$ cm^2	atom-atom cross section $\times 10^{16}$ cm^2
He	54	15.0
Ne	65	21.0
A	134	42.0
Kr	157	49
Xe	192	67

mation; we may then derive for this eventuality conditions on $h_0(c)$ which are more general, respectively, than (79) or (87), (80) or (88), (81) or (89). Since we are outrunning here the experimental evidence we shall limit ourselves to the derivation of the first of these relations. The first truncated relation is exactly (50a) which, for isotropy and equal masses, reads

$$\left\langle \frac{c_z}{a\tau(c)} \right\rangle = 2 \tag{101}$$

The reduction of this formula to a condition on $h_0(c)$ requires the relationship $\nu = 0$ of the set (46). This relation is always integrable to yield $h_1(c)$ in terms of $h_0(c)$, as was pointed out early in this section. For the special circumstances assumed the integrated equation is equation (74)

$$h_1(c) = 3 \int_c^\infty \frac{h_0(\gamma)}{a\tau(\gamma)}\, d\gamma \tag{102}$$

The elimination of $h_1(c)$ is achieved by forming the average (101) on the function (102). This leaves the required condition on $h_0(c)$; it may be given the following form

$$\left\langle \frac{\sigma(c)}{c} \int_0^c \gamma^4 \sigma(\gamma)\, d\gamma \right\rangle = 2\,\frac{a^2}{N^2} \tag{103}$$

The equations (79) and (87) are manifestly special cases of this more general relation. Adaptations of this procedure to other cases are clearly possible whenever the need arises.

The calculations of this section are meant to suggest that it is possible to compute reliably average values from a Boltzmann equation without solving it completely. The method employed here for this purpose resembles a Ritz method in that it works with trial functions which must be guessed at, and like that method it is capable of indefinite improvement. The numerical results suggest strongly that we are converging toward a definite answer; however, a mathematical proof of this fact has not been presented. The method will be applied once more in the section on diffusion.

Part III — Motion of Uniform Ion Streams in Intermediate Fields

IIIA. A convolution theorem

Whenever we deal with the motion of a given type of charged particle in a gas of given composition, then there exists a wide range of densities n and N as discussed in Section IA in which the motion of these particles depends only on a/N and kT. For this range the motion is governed by equation (13). Since deriving that equation, all our efforts were dealing with the "high field" equation (34) or (40), in which the gas temperature is taken to be zero and the electric field often scales out, as in (26), (75) and (85). The accomplished solution of this restricted problem, together with the low field solutions available in the literature, brings us back to the more general equation (13) and the question what can be done with it. The topic of Part III so defined is definitely inferior in importance to the one in Part II. For we are studying here an intermediate range of variables which can be handled qualitatively, both in concept and practice, by some sort of interpolation between the high and low field regimes. For precise measurements, conditions can always be chosen so as to satisfy one or the other of the two extremes. For this reason the intermediate field case will only be pushed as far as it will go conveniently, without appeal to numerical methods.

In this Section IIIA we shall give a complete solution of the inter-

mediate field problem for the mean free time models discussed in Section IIB. This solution is achieved by the following theorem: *Given the general equation (13) for constant mean free time*

$$a\tau \frac{\partial f(\mathbf{c})}{\partial c_z} + f(\mathbf{c}) = \frac{1}{4\pi} \iint M(\mathbf{C}')f(\mathbf{c}')\Pi(\chi)\, d\Omega_{\gamma'}\, d\mathbf{C} \tag{104}$$

and the "high field" equation derived from it by setting the gas temperature equal zero

$$a\tau \frac{\partial h(\mathbf{c})}{\partial c_z} + h(\mathbf{c}) = \frac{1}{4\pi} \iint \delta(\mathbf{C}')h(\mathbf{c}')\Pi(\chi)\, d\Omega_{\gamma'}\, d\mathbf{C} \cdot \tag{105}$$

and the Maxwellian equation derived from (104) by dropping the field term

$$m(\mathbf{c}) = \frac{1}{4\pi} \iint M(\mathbf{C}')m(\mathbf{c}')\Pi(\chi)\, d\Omega_{\gamma'}\, d\mathbf{C} \tag{106}$$

then the solution $f(\mathbf{c})$ of (104) is the convolution of the solution $h(\mathbf{c})$ of (105) and the solution $m(\mathbf{c})$ of (106):

$$f(\mathbf{c}) = \int h(\mathbf{u})m(\mathbf{c} - \mathbf{u})\, d\mathbf{u} \tag{107}$$

We carry through the proof by constructing explicitly the equation satisfied by the convolution. We replace the running variables $\mathbf{c}$, $\mathbf{c}'$, $\mathbf{C}$, $\mathbf{C}'$ in (105) by $\mathbf{u}$, $\mathbf{u}'$, $\mathbf{U}$, $\mathbf{U}'$ and multiply in $m(\mathbf{c} - \mathbf{u})$. We get

$$a\tau \frac{\partial h(\mathbf{u})}{\partial u_z} m(\mathbf{c} - \mathbf{u}) + h(\mathbf{u})m(\mathbf{c} - \mathbf{u}) =$$

$$= \frac{1}{4\pi} \iint \delta(\mathbf{U}')h(\mathbf{u}')m(\mathbf{c} - \mathbf{u})\Pi(\chi_u)\, d\Omega_{\eta'}\, d\mathbf{U}$$

We now *define* $f(\mathbf{c})$ by the relation (107), and integrate the above equation over $\mathbf{u}$. The second member on the left comes out to be $f(\mathbf{c})$. For the first member, we carry out an integration by parts:

$$\begin{aligned}
\int \frac{\partial h(\mathbf{u})}{\partial u_z} m(\mathbf{c} - \mathbf{u})\, d\mathbf{u} &= -\int h(\mathbf{u}) \frac{\partial m(\mathbf{c} - \mathbf{u})}{\partial u_z}\, d\mathbf{u} \\
&= +\int h(\mathbf{u}) \frac{\partial (m(\mathbf{c} - \mathbf{u}))}{\partial c_z}\, d\mathbf{u} \\
&= \frac{\partial}{\partial c_z} \int h(\mathbf{u})m(\mathbf{c} - \mathbf{u})\, d\mathbf{u} \\
&= \frac{\partial f(\mathbf{c})}{\partial c_z}
\end{aligned}$$

For the right hand member we observe that we have the eightfold integration

$$d\Omega_{\eta'}\, d\mathbf{U}\, d\mathbf{u},$$

that is an integration over the collision angles and all final velocity components. By a general principle of kinetic theory[25] we can invert in this integration the final and the initial quantities and write

$$d\Omega_{\eta'}\, d\mathbf{U}\, d\mathbf{u} = d\Omega_{\eta}\, d\mathbf{U}'\, d\mathbf{u}' \tag{108}$$

This puts us in a position to eliminate the δ-function by integration. We find

$$a\tau \frac{\partial f(\mathbf{c})}{\partial c_z} + f(\mathbf{c}) = \frac{1}{4\pi} \iint h(\mathbf{u}')m(\mathbf{c} - \mathbf{u})\Pi(\chi_u)\, d\Omega_{\eta}\, d\mathbf{u}' \tag{109}$$

with the side condition that $\mathbf{u}$, $\mathbf{U}$, $\mathbf{u}'$, $\mathbf{U}'$ form a quadruple of vectors in the sense discussed in Section IB for which in addition

$$\mathbf{U}' = 0$$

If we substitute (107) into (104), denoting the dummy variable by $\mathbf{u}'$ instead of $\mathbf{u}$, then the two equations (104) and (109) take on a very similar look. A proof of their identity hinges upon proving the identity of the integral terms:

$$\int h(\mathbf{u}')\, d\mathbf{u}' \int m(\mathbf{c} - \mathbf{u})\Pi(\chi_u)\, d\Omega_{\eta} = \int h(\mathbf{u}')\, d\mathbf{u}' \iint M(\mathbf{C}')m(\mathbf{c}' - \mathbf{u}')\Pi(\chi_c)\, d\Omega_{\gamma'}\, d\mathbf{C} \tag{110}$$

The form of this relation suggests the assumption that the expressions are identical before integration over $\mathbf{u}'$; this assumption is proved by the events below. The complicated function $h(\mathbf{u}')$ thus disappears from the problem. The other such function, namely $\Pi(\chi)$ disappears then also; for it is by assumption arbitrary, hence could be replaced by a δ-function for a fixed, but arbitrary χ. The two sides of (110) must therefore be equal before we integrate over χ_u or χ_c, and the two χ's are to be taken equal and fixed. Defining angles as shown in the spherical diagram Fig. 15 we thus get (110) in the form

$$\int m(\mathbf{c} - \mathbf{u})\, d\epsilon = \iint M(\mathbf{C}')m(\mathbf{c}' - \mathbf{u}')\, d\phi\, d\mathbf{C} \tag{111a}$$

[25] See Reference 4, Section 3.52.

This is to be true with the side conditions

$$\mathbf{c} = \text{fixed} \tag{111b}$$

$$\mathbf{u}' = \text{fixed} \tag{111c}$$

$$\mathbf{U}' = 0 \tag{111d}$$

$$\chi_c = \chi_u = \chi = \text{fixed} \tag{111e}$$

Equation (111) is an identity involving only elementary functions. Thus the relation itself is in a sense elementary. Those who wish to believe it, may consider the theorem proved; for completeness, however, the proof of (111) will now follow.

Call the left side of (111a) X, the right side Y. To determine X, we substitute from (7) and (111d)

$$\mathbf{u} = \frac{m}{M+m}\,\mathbf{u}' + \frac{M}{M+m}\,\eta$$

with

$$\eta^2 = u'^2$$

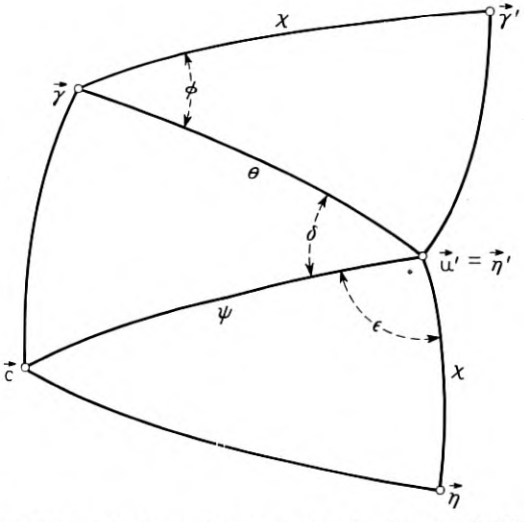

Fig. 15 — Definition of the angles occurring in the proof of the convolution theorem.

because of (9). This yields with the angles as shown on Fig. 15

$$X = \int_0^{2\pi} m(\mathbf{c} - \mathbf{u})\, d\epsilon = \left(\frac{\beta m}{\pi}\right)^{3/2} \exp\left[- \beta m c^2 \right.$$

$$\left. - \beta m u'^2 \frac{M^2 + m^2 + 2Mm\cos\chi}{(M+m)^2} + 2\beta m c u' \frac{m + M\cos\chi}{M+m}\cos\psi \right]$$

$$\cdot \int_0^{2\pi} \exp\left[2\beta \frac{mM}{M+m} c u' \sin\chi \sin\psi \cos\epsilon \right] d\epsilon$$

The integral is evaluated by a formula known from the theory of Bessel functions

$$\int_0^{2\pi} e^{z\cos\epsilon}\, d\epsilon = 2\pi I_0(z) \tag{112}$$

and yields

$$X = \frac{2}{\sqrt{\pi}} (\beta m)^{3/2} \exp\left[- \beta m c^2 - \beta m u'^2 \frac{M^2 + m^2 + 2Mm\cos\chi}{(M+m)^2} \right.$$

$$\left. + 2\beta m c u' \frac{m + M\cos\chi}{M+m}\cos\psi \right] \cdot I_0\left(\frac{2\beta Mm}{M+m} c u' \sin\chi \sin\psi \right) \tag{113}$$

Passing now to the right hand side of (111a) we may replace in the first place $d\mathbf{C}$ by $d\boldsymbol{\gamma}$, because of (111b). $\mathbf{c}'$ and $\mathbf{C}'$ are then replaced by the expressions

$$\mathbf{c}' = \mathbf{c} - \frac{M}{M+m}\boldsymbol{\gamma} + \frac{M}{M+m}\boldsymbol{\gamma}'$$

$$\mathbf{C}' = \mathbf{c} - \frac{M}{M+m}\boldsymbol{\gamma} - \frac{M}{M+m}\boldsymbol{\gamma}'.$$

With the angles defined in Fig. 15, we thus get for Y

$$Y = \left(\frac{\beta M}{\pi}\right)^{3/2} \left(\frac{\beta m}{\pi}\right)^{3/2} \exp\left[- \beta M c^2 - \beta m (\mathbf{c} - \mathbf{u}')^2 \right]$$

$$\cdot \iiint\!\!\int \gamma^2\, d\gamma \sin\theta\, d\theta\, d\delta\, d\phi$$

$$\exp\left[- \beta M \gamma^2 + \beta M c \gamma (\cos\psi\cos\theta + \sin\psi\sin\theta\cos\delta) \right.$$

$$\left. - 2\beta \frac{mM}{M+m} u'\gamma (\cos\theta - \cos\theta\cos\chi - \sin\theta\sin\chi\cos\phi) \right]$$

Integrations over δ and ϕ again go with (112). Beofore writing down the result we shall pass over to a cylindrical coordinate system defined by

$$\gamma_{||} = \gamma \cos\theta \qquad \gamma_{\perp} = \gamma \sin\theta$$

$$\gamma^2 \, d\gamma \sin\theta \, d\theta = \gamma_{\perp} \, d\gamma_{\perp} \, d\gamma_{\perp}$$

The result of the first two integrations then reads

$$Y = \frac{4}{\pi} \beta^3 (Mm)^{3/2} \exp\left[-\beta M c^2 - \beta m(\mathbf{c} - \mathbf{u}')^2\right]$$

$$\int_{-\infty}^{+\infty} d\gamma_{||} \exp\left[-\beta M \gamma_{||}^2 + 2\beta M \gamma_{||} \left(c \cos\psi - \frac{mu'}{M+m}(1 - \cos\chi)\right)\right]$$

$$\int_0^{\infty} \gamma_{\perp} \, d\gamma_{\perp} \exp\left[\beta M \gamma_{\perp}^2\right] \cdot I_0(2\beta M c \gamma_{\perp} \sin\psi) \cdot I_0\left(2\beta \frac{Mm}{M+m} u' \gamma_{\perp} \sin\chi\right)$$

The first of these two integrals is elementary, the other is Weber's second exponential integral[26] which equals

$$\int_0^{\infty} \exp(-p^2 t^2) I_0(at) I_0(bt) t \, dt = \frac{1}{2p^2} \exp\left(\frac{a^2 + b^2}{4p^2}\right) I_0\left(\frac{ab}{2p^2}\right) \qquad (114)$$

This yields for Y exactly the expression (113). The identity (111) is thus proved, and with it the convolution theorem.

The theorem just proved reduces the velocity distribution for arbitrary field and temperature to two components, one containing the field, but not the temperature, the other the temperature but not the field. In each of these components, in turn, the variable parameter scales out; thus the general distribution reduces to two basic ones one of which is the Maxwellian one: the other is worked out partially in the calculations of the Sections IIC and IID. The special case of heavy ion mass has been published independently by Kihara[27] without any apparent knowledge of this theorem which was available in the literature without complete proof.[1] Kihara's form of the theorem is that heavy ions in a light gas have an off-set Maxwellian distribution, with the gas temperature as parameter if the mean free time condition is obeyed for their collisions. Such a function is indeed the convolution of a Maxwellian distribution and the δ-function discussed in the Sections IIB and IIC.

The general distribution function resulting from (107) cannot be written down explicitly because this goal was never achieved for $h(\mathbf{c})$. However we do find a result which is almost a full substitute for this,

[26] Watson, G. N., *A Treatise on the Theory of Bessel Functions*. Cambridge University Press, Section 13.31, 1922.

[27] Kihara, Taro, Rev. Mod. Phys., **24**, p. 45, 1952.

namely that all averages of products of integer powers of the Cartesian velocity components, which were shown to be computable in the high field case, can be computed for the intermediate and low field range as well. The calculation proceeds as follows. Suppose we wish to compute the velocity average

$$\langle c_x^m c_y^n c_z^p \rangle = \int c_x^m c_y^n c_z^p f(\mathbf{c})\, d\mathbf{c} \tag{115}$$

for m, n, p integer or zero. We apply the convolution theorem (107) to $f(\mathbf{c})$, decompose the three factors into

$$\begin{aligned} c_x^m &= \{u_x + (c_x - u_x)\}^m \\ c_y^n &= \{u_y + (c_y - u_y)\}^n \\ c_z^p &= \{u_z + (c_z - u_z)\}^p \end{aligned}$$

and expand each of them by the binomial theorem. We find

$$\langle c_x^m c_y^n c_z^p \rangle = \sum_{\mu=0}^{m} \sum_{\nu=0}^{n} \sum_{\pi=0}^{p} \binom{\mu}{m} \binom{\nu}{n} \binom{\pi}{p} \tag{116}$$

$$\int h(\mathbf{u})\, u_x^{\mu} u_y^{\nu} u_z^{\pi}\, d\mathbf{u} \int m(\mathbf{v}) v_x^{m-\mu} v_y^{n-\nu} v_z^{p-\pi}\, d\mathbf{v}$$

The second integral is a thermal average, the first a high field average computable by the method of Section IIB. Thus the average (116) is a finite sum of products of computable averages and is itself computable.

When formula (116) is applied to the averages (52), (53), (54), (57) and (59) very simple results are found because of the symmetry of the function $m(\mathbf{v})$. For the drift velocity $\langle c_z \rangle$ we get from (52)

$$\langle c_z \rangle = \frac{M + m}{M} \Big/ \left\langle \frac{1 - \cos \chi}{a\tau} \right\rangle \tag{117}$$

This is the same formula as (52) which is thus proved to hold independently of the gas temperature. In the energy formulas we find simple addition of the thermal and high field values because the middle term in (116) drops out by symmetry. Inserting (53), (54), (57) and (59) we find

$$\langle mc^2 \rangle = 3kT + \frac{(M + m)^3}{M^2} \Big/ \left\langle \frac{1 - \cos \chi}{a\tau} \right\rangle^2 \tag{118}$$

$$\langle mc_z^2 \rangle = kT + \frac{(M + m)^3 \left\langle \dfrac{M \sin^2 \chi + 4m(1 - \cos \chi)}{a\tau} \right\rangle}{M^2 \left\langle \dfrac{3M \sin^2 \chi + 4m(1 - \cos \chi)}{a\tau} \right\rangle \left\langle \dfrac{1 - \cos \chi}{a\tau} \right\rangle}, \tag{119}$$

$$\langle mc_z^2 \rangle - m\langle c_z \rangle^2 =$$

$$= kT + \frac{(M+m)^2 \left\langle \frac{M \sin^2 \chi + 2m(1-\cos\chi)^2}{a\tau} \right\rangle}{M \left\langle \frac{3M \sin^2 \chi + 4m(1-\cos\chi)}{a\tau} \right\rangle \left\langle \frac{1-\cos\chi}{a\tau} \right\rangle^2} \tag{120}$$

$$\langle mc_x^2 \rangle = kT + \frac{(M+m)^3 \left\langle \frac{\sin^2 \chi}{a\tau} \right\rangle}{M \left\langle \frac{3M \sin^2 \chi + 4m(1-\cos\chi)}{a\tau} \right\rangle \left\langle \frac{1-\cos\chi}{a\tau} \right\rangle^2} \tag{121}$$

The interpretation of these formulas is implicit in the discussion of the high field formulas given earlier. In particular the combination of the equations (55) and (116) can be given the elegant form

$$m\langle c^2 \rangle = M\langle C^2 \rangle + m\langle c_z \rangle^2 + M\langle c_z \rangle^2 \tag{122}$$

It states that the energy of an ion is obtained by adding the energy of a gas molecule, the energy visible in the drift motion and a storage term which is M/m times the energy in the drift motion; this term becomes important for electrons in a gas. A low field approximation to this formula (in which the second term on the right may be neglected) has been published in the article of Kihara.[28]

IIIB. RESULTS FOR THE POLARIZATION FORCE AND THE ISOTROPIC "MAXWELLIAN" MODEL

The polarization force between ions and molecules which predominates over other forces at sufficiently low temperature satisfies the mean free time requirement of the preceding section. It follows that the complete theory given for those conditions applies to this force. The magnitude of force was given in (4). Its potential equals

$$V = \frac{1}{2} \frac{e^2 P}{\rho^4} \tag{123}$$

Classical theory is usually applicable to the scattering by the potential (123) because angular momentum quantum numbers run as high as 30 or 50 in normal situations.[29] This classical type theory, first developed by Langevin,[30] follows standard elementary methods for computing the

[28] Reference 27, formula 5.12.

[29] Holstein, Theodore, private communication, see also Reference 11.

[30] Langevin, Ann. de Chim. et de Phys., **5**, p. 245, 1905.

angle of deflection χ due to a potential of the type (123). The result is

$$\chi = \pi - 2\int_0^{u_1} \frac{du}{\left\{\frac{1}{b^2} - u^2 + \frac{e^2P(M+m)}{Mmb^2\gamma^2}\,u^4\right\}^{1/2}} \tag{124}$$

Here b is the "impact parameter", and u_1 is the lower of the two positive roots of the polynomial in the denominator; if the polynomial has no real root, the integration goes from 0 to ∞. The question whether the denominator has a real root or not is tied up with the nature of the orbit. If b is sufficiently large a root exists and the orbit looks like a hyperbola, Fig. 2(a); for small b no root exists and the two particles are "sucked" towards each other in a spiralling orbit as shown in Fig. 2(b). The two regimes are separated by a limiting orbit in which the particles spiral asymptotically into a circular orbit. This limiting orbit is found by setting the discriminant of the square root in (124) equal to 0. We find

$$b_{\text{lim}}^4 = \frac{4e^2P(M+m)}{Mm\gamma^2} \tag{125}$$

From this value of b_{lim} a cross section and a mean free time τ_s for spiralling collisions can be derived. We find

$$\tau_s = \frac{1}{2\pi eN}\left\{\frac{Mm}{(M+m)P}\right\}^{1/2} \tag{126}$$

This is indeed a constant mean free time as stated, the speed of encounter γ having dropped out.

$1/\tau$ is the dimensional quantity entering into the averages $\langle\varphi(\chi)/\tau\rangle$ which occur in the Sections IIB and IIIA. In working them out in detail as was done by Hassé[9] one has to take into account hyperbolic collisions also; for them a τ cannot be defined or comes out to be zero in the mean. This is due to small angle deflections which are infinitely probable. However, any quantity $\varphi(\chi)/\tau$ to be averaged in a physical problem contains a $\varphi(\chi)$ which vanishes for such impacts. Hence finite averages result which do not give overdue weight to these types of collisions. Following Hassé[9], we do this in the following way for the present case. We write (124) in the form

$$\chi = \pi - 2\int_0^{v_1} \frac{dv}{\left\{1 - v^2 + \frac{v^4}{4\beta^4}\right\}^{1/2}} \tag{127}$$

Here v equals bu and the parameter β equals b/b_{lim}. It is the parameter β introduced by Hassé. Now by the definition of τ we have

$$\left\langle \frac{\varphi(\chi)}{\tau} \right\rangle = N \int \gamma\varphi(\chi)\, d\sigma(\gamma,\chi)$$

$$= N \int_0^\infty \gamma\varphi(\chi) b\, db \int_0^{2\pi} d\epsilon$$

$$= \pi N \gamma b_{\text{lim}}^2 \int_0^\infty \varphi(\chi)\, d(\beta^2)$$

From (125) and (126) the factor in front of the integral just equals $1/\tau_s$; the integral on the other hand is a computable pure number independent of γ which is obtained by inserting into it the relationship (127) between χ and β. Hence we may write

$$\left\langle \frac{\varphi(\chi)}{\tau} \right\rangle = \frac{1}{\tau_s} \int_0^\infty \varphi(\chi)\, d(\beta)^2 \tag{128}$$

The three equations (126), (127) and (128) completely define the nature of the averages appearing in previous sections. The integral (128) has to be computed by numerical methods. It is seen in the course of the evaluations that it naturally decomposes into two parts. The part for which β varies from 0 to 1 deals with spiralling collisions and exists for any $\varphi(\chi)$. For β between 1 and ∞ we get the contribution of the hyperbolic collisions to the average. This part is only finite if $\varphi(\chi)$ vanishes for small angle deflections.

The averages (52), (53), (54), (57) and (59), as well as (117) to (121) contain numerous averages of the form (128) all of which satisfy the predicted condition $\varphi(0) = 0$. They are obtained by linear combination of two basic types: $\langle (1 - \cos\chi)/\tau \rangle$ and $\langle \sin^2 \chi/\tau \rangle$. The first average is given in Hassé.[9] Separating the parts due to spiralling and hyperbolic collisions we find

$$\int_0^1 (1 - \cos\chi)\, d(\beta^2) = 0.8979$$

$$\int_1^\infty (1 - \cos\chi)\, d(\beta^2) = 0.2073$$

This combines to give

$$\left\langle \frac{1 - \cos\chi}{\tau} \right\rangle = \frac{1}{\tau_s} \cdot 1.1052 \tag{129}$$

The analogous result for $\sin^2 \chi$ was obtained by the computing group of Bell Telephone Laboratories

$$\int_0^1 \sin^2 \chi \, d(\beta^2) = 0.511$$

$$\int_1^\infty \sin^2 \chi \, d(\beta^2) = 0.261$$

which gives

$$\left\langle \frac{\sin^2 \chi}{\tau} \right\rangle = \frac{1}{\tau_s} \cdot 0.772 \tag{130}$$

We may now rewrite the major results of Section IIIA for the polarization force. From equation (117) we get

$$\langle c_z \rangle = \frac{0.9048}{2\pi} \sqrt{\frac{1}{M} + \frac{1}{m}} \frac{E}{N\sqrt{P}} \tag{131}$$

This formula may be found in the literature.[31] What is new about (131) is the realization that it is exact at high as well as low electric field.

The formula for the total energy needs no discussion for a special model; it does not involve the angular distribution law when written in the form (122). Thus we would obtain, for instance, for an ion travelling in the parent gas that its total energy is obtained by doubling its apparent energy observable in the drift and adding to this the thermal energy $\frac{3}{2}kT$.

For the partition of the high field component of the energy in the three coordinate directions we have two formulas, formula (58) partitions the entire field contribution of the kinetic energy, formula (60) only its random component. The first formula gives

$$e_x : e_y : e_z = M : M : (M + 6.73m) \tag{132}$$

Formula (60) gives

$$e_x : e_y : e_z^* = (M + m) : (M + m) : (M + 3.72m) \tag{133}$$

It is convenient to apply the general formulas also to the case of constant mean free time, coupled with the assumption of isotropic scattering. This combination of assumptions represents, strictly speaking, an im-

[31] The formula is equation (3), p. 39 of Reference 2, in the limit $\lambda = 0$; or also the last unnumbered equation on p. 919 of Reference 6.

possibility; for we know of no mechanical force which realizes this arrangement. This model was already taken as the basis of the Monte Carlo calculation in Section IID. It will be seen now that it has a wider significance than one might anticipate. The necessary angular averages are

$$\langle 1 - \cos \chi \rangle = 1 \tag{134}$$

$$\langle \sin^2 \chi \rangle = \tfrac{2}{3} \tag{135}$$

This yields for (117)

$$\langle c_z \rangle = \frac{M + m}{M} a\tau \tag{136}$$

As usual, the formula for the energy does not involve the law of scattering if written in the form (122). If we choose the form (118) instead we get in agreement with (79)

$$\langle mc^2 \rangle = 3kT + \frac{(M+m)^3}{M^2} a^2\tau^2 \tag{137}$$

The partition formula (58) becomes

$$e_x : e_y : e_z = M : M : (M + 6m) \tag{138}$$

the partition formula (60) which counts random energy only becomes

$$e_x : e_y : e_z^* = (M + m) : (M + m) : (M + 4m) \tag{139}$$

Comparison of these expressions with the ones for the polarization force shows that the difference between it and the isotropic model is remarkably small from a kinetic standpoint. We may see this by comparing (132) and (138) or (133) and (139). For the other formulas, we may compare more specifically the polarization results with an isotropic case having its mean free time τ given by

$$\tau = 0.9048\ \tau_s \tag{140}$$

Equation (136) becomes then identical with (131) and because of (122) the same identity persists for the energy formula (137). In the light of this we may say that it is very nearly correct to state that scattering is isotropic for the polarization force. This qualitatively correct fact was repeatedly made use of in the preceding sections of the paper. The reason for it is chiefly the predominant effect of spiralling collisions. Indeed, equation (140) shows that a modification of τ_s by only 10 per cent takes into account the main influence of hyperbolic collisions.

From the discussion in Section IA it may be seen that the results

obtained for the polarization force have a potentially wide field of application when measurements of ion drift are extended to low temperature. In the meantime, the results apply occasionally at room temperature, whenever we deal with a small ion and are not bothered by special scattering mechanisms having large cross section. An example of this are the molecular noble gas ions in the parent gas whose drift velocities were measured by Hornbeck[16, 17] and Varney.[18] Table IV shows the measured mobility at standard gas density measured for these ions, in comparison with a value obtained from equation (131). The field range from which the observed mobility was obtained is intermediate. There is not only good numerical agreement, but the experiments follow the theory also in that there is little variation of the observed value

TABLE IV

Mobilities at standard density of the noble gas molecular ions. Comparison of the experiment with a formula based on the polarization force only.

Gas	$\mu_{obs.}$ $\frac{cm^2}{Volt\ sec}$	μ_{calc} $\frac{cm^2}{Volt\ sec}$
He	18	18.2
Ne	6.5	6.21
A	1.9	2.09
Kr	1.2	1.18
Xe	0.7	0.74

with the field. The discrepancy between the two columns can be used to determine a hard collision cross section which is to be superimposed on the polarization force, as is suggested in the so-called Langevin model.[30]

IIIC. VELOCITY DISTRIBUTION FUNCTION FOR ELECTRONS

We have almost exhausted the results achieved for intermediate field conditions. For the sake of completeness I shall mention shortly the intermediate field distribution function for electrons whose derivation we owe to the ingenuity of Davydov.[32]

The derivation does not differ in principle from the one presented in Section IIC for the electrons in the high field case. The distribution function is first expanded in spherical harmonics. For group theoretical reasons the scattering term in the Boltzmann equation is diagonal in

[32] Davydov, B., Phys. Zeits. Sowjetunion., **8**, p. 59, 1935. See also Reference 4, pp. 349–350.

such a decomposition even in the presence of molecular agitation. Thence a generalized form of (47) may be derived containing essentially the same terms. Finally, all but the first two spherical harmonics are dropped and two equations analogous to (63) and (64) are obtained. In fact, it is found that equation (63) is maintained entirely. An extremely complicated reasoning is required, on the other hand, to find the generalization of (64). The result is

$$f_1(c) = \left\{\frac{3kT}{M}\frac{df_0}{dc} + 3\frac{m}{M}cf_0(c)\right\}\left\langle\frac{1-\cos\chi}{a\tau(c)}\right\rangle \tag{141}$$

Combining (63) and (141) we find

$$\left(\frac{1}{\left\langle\frac{1-\cos\chi}{a\tau(c)}\right\rangle} + \frac{3kT}{M}\right)\frac{df_0}{dc} + 3\frac{m}{M}cf_0 = 0$$

and hence

$$f_0(c) = \exp\left[-m\int_0^c \frac{c\,dc}{\frac{\frac{1}{3}M}{\left\langle\frac{1-\cos\chi}{a\tau(c)}\right\rangle^2} + kT}\right] \tag{142}$$

This is the so-called Davydov distribution which is a generalization containing within itself the Maxwellian distribution as well as the high field distribution (65).

The mean energy and the drift velocity of electrons may be calculated from (63) and (142). They are obtainable from the literature and will not be discussed here any further. Equipartition of the energy exists at all field conditions.

Part IV — Diffusive Motion of Ions

IVA. Diffusion for Mean Free Time Models

It was proved in Section IC that if there are spatial inequalities in the distribution of the charge carriers then a smoothing out process sets in which can be described as diffusion. This derivation of principle can be supplemented for "Maxwellian" molecules by an explicit computation of the two components of the tensor (24), that is an evaluation of the integral (23). We shall do this by following the method of Maxwell[19]

rather than by generalizing the formal procedure of the Sections IIA, IIB and IIIA. Such a generalization would no doubt be possible, but would increase unduly the bulk of this paper. We shall operate therefore directly on equation (20). To get out the integral (23) we multiply the equation vectorially with $\mathbf{c}$ and integrate over $d\mathbf{c}$. This operation makes the first term vanish completely. This is obvious from symmetry for the components c_x and c_y of the multiplier $\mathbf{c}$. For c_z we have

$$a \int \frac{\partial g}{\partial c_z} c_z \, d\mathbf{c}$$

An integration by parts brings this in the form (18) and thus makes it equal to zero.

Temporarily, we may break the integral term of (20) into two parts, using some artificial procedure to eliminate small angle collisions. The first half of the integral term reads then simply

$$\frac{1}{\tau} \int g(c)\mathbf{c} \, d\mathbf{c}$$

This is already the desired average (23). On the second half we use the identity (108) to give it the form

$$-\frac{1}{4\pi\tau} \iint M(\mathbf{C}')g(\mathbf{c}')\mathbf{c}\Pi(\chi) \, d\Omega_\gamma \, d\mathbf{C}' \, d\mathbf{c}'$$

We now use (7) to replace $\mathbf{c}$ by the expression

$$\mathbf{c} = \frac{m}{M+m}\mathbf{c}' + \frac{M}{M+m}\mathbf{C}' + \frac{M}{M+m}\boldsymbol{\gamma}$$

Only $\boldsymbol{\gamma}$ is affected by the integration over $d\Omega_\gamma$ which we take up first. Using $\boldsymbol{\gamma}'$ as the axis of a polar coordinate system we may write

$$\boldsymbol{\gamma} = \boldsymbol{\gamma}_{||} + \boldsymbol{\gamma}_{\perp}$$

For every value of χ, $\boldsymbol{\gamma}_{||}$ has the fixed value $\boldsymbol{\gamma}' \cos\chi$. On the other hand the average of $\boldsymbol{\gamma}_{\perp}$ vanishes through integration over all azimuths. Thence we may write

$$\frac{1}{4\pi}\int \mathbf{c}\Pi(\chi) \, d\Omega_\gamma = \frac{m}{M+m}\mathbf{c}' + \frac{M}{M+m}\mathbf{C}' + \frac{M}{M+m}(\mathbf{c}' - \mathbf{C}')\langle\cos\chi\rangle$$

$$= \frac{m + M\langle\cos\chi\rangle}{M+m}\mathbf{c}' + \frac{M\langle 1 - \cos\chi\rangle}{M+m}\mathbf{C}'$$

We now multiply with $M(\mathbf{C}')g(\mathbf{c}')$ and integrate over $d\mathbf{c}' \, d\mathbf{C}'$. The integration of the term containing $\mathbf{C}'$ obviously vanishes for two independent reasons. The integration of the term in $\mathbf{c}'$, finally, yields again the average

(23). Combining the two pieces, we find

$$\frac{M}{M+m}\left\langle\frac{1-\cos\chi}{\tau}\right\rangle\int g(\mathbf{c})\mathbf{c}\,d\mathbf{c}$$

In this expression, the artificial exclusion of small angle scattering is no longer necessary and can be dropped. Completing the integrating of equation (20) we see that the right hand side gives averages over the unperturbed velocity distribution $f(\mathbf{c})$. Combining pieces, using (23) and indices 1, 2, 3 for the x, y and z components we get

$$\mathbf{j}_i(\mathbf{r},t) = -n(\mathbf{r},t)\sum_{\nu=1}^{3}k_\nu\left[\frac{M+m}{M\left\langle\frac{1-\cos\chi}{\tau}\right\rangle}\{\langle c_ic_\nu\rangle-\langle c_i\rangle\langle c_\nu\rangle\}\right] \tag{143}$$

According to (16) and (24), the square bracket in (143) is the diffusion tensor. It has two distinct components which equal respectively

$$D_{\parallel} = \frac{M+m}{M}\frac{\langle c_z^2\rangle-\langle c_z\rangle^2}{\left\langle\frac{1-\cos\chi}{\tau}\right\rangle} \tag{144}$$

$$D_{\perp} = \frac{M+m}{M}\frac{\langle c_x^2\rangle}{\left\langle\frac{1-\cos\chi}{\tau}\right\rangle} \tag{145}$$

The velocity averages entering are (120), and (121), that is the directional components of the random part of the energy. Substituting we get finally

$$D_{\parallel} = \frac{(M+m)kT}{Mm\left\langle\frac{1-\cos\chi}{\tau}\right\rangle} + a^2\frac{(M+m)^3\left\langle\frac{M\sin^2\chi+2m(1-\cos\chi)^2}{\tau}\right\rangle}{M^2m\left\langle\frac{3M\sin^2\chi+4m(1-\cos\chi)}{\tau}\right\rangle\left\langle\frac{1-\cos\chi}{\tau}\right\rangle^3} \tag{146}$$

$$D_{\perp} = \frac{(M+m)kT}{Mm\left\langle\frac{1-\cos\chi}{\tau}\right\rangle} + a^2\frac{(M+m)^4\left\langle\frac{\sin^2\chi}{\tau}\right\rangle}{M^2m\left\langle\frac{3M\sin^2\chi+4m(1-\cos\chi)}{\tau}\right\rangle\left\langle\frac{1-\cos\chi}{\tau}\right\rangle^3} \tag{147}$$

The diffusion coefficients have the simple property that they are obtained by adding the low field and the high field limiting expressions.

This is a consequence of the limited form of the convolution theorem proved in Section IIIA; it probably implies also that the theorem can be extended in some form to include the case of diffusion.

It has been mentioned in the Section ID that the Nernst-Townsend relation (30) applies only to ions moving in a low field. We are now in a position to examine possible extensions of it to general fields. Equations (144), (145) and (117) suggest the form

$$\frac{D_n}{\text{mobility}} = \frac{2 \times \text{mean random energy along } n}{e} \tag{148}$$

where n stands for one of the principal directions of the diffusion tensor. This formula contains equation (30) as a specialization to the low field case.

Formula (148) is one of the formulas obtained in this study of ion motion in which model parameters do not appear. It is valid (a) for all interactions at low field and (b) for the mean free time case at all fields. It also holds dimensionally at high field for models obeying (25); this may be seen from (26a) and (28a). It appears a reasonable conjecture that (148) is approximately true for any law of interaction; the question will be taken up again in the next section.

Let us, in conclusion, write down the formulas resulting from (146) and (147) for the two special mean free time models studied in detail in Section IIIB: the polarization force and the isotropic model. The necessary averages are (129), (130), (134) and (135). They yield for the polarization force

$$D_{\|} = \frac{M+m}{Mm} 0.905\tau_s \cdot kT + \frac{1}{3}\frac{(M+m)^3(M+3.72m)}{M^2m(M+1.908m)} a^2(0.905\tau_s)^3 \tag{149}$$

$$D_{\perp} = \frac{M+m}{Mm} 0.905\tau_s \cdot kT + \frac{1}{3}\frac{(M+m)^4}{M^2m(M+1.908m)} a^2(0.905\tau_s)^3 \tag{150}$$

and for the case of isotropic scattering

$$D_{\|} = \frac{M+m}{Mm}\tau kT + \frac{1}{3}\frac{(M+m)^3(M+4m)}{M^2m(M+2m)} a^2\tau^3 \tag{151}$$

$$D_{\perp} = \frac{M+m}{Mm}\tau kT + \frac{1}{3}\frac{(M+m)^4}{M^2m(M+2m)} a^2\tau^3 \tag{152}$$

Just as in the earlier study the results for the two models do not differ appreciably.

IVB. LONGITUDINAL DIFFUSION FOR THE HARD SPHERE MODEL

Whenever the mean free time condition for collisions is not fulfilled, then the computation of diffusion coefficients requires a procedure analogous to that of Section IIE. Since this entails some numerical work the calculation was only carried out for a case which was thought to be of experimental interest, namely for longitudinal diffusion of ions in the parent gas. In other words, we are extending the numerical computation at the end of Section IIE to include longitudinal diffusion. The computation to provide us with the undetermined constant of equation (28b) for the special case when m and M are equal; it also offers, incidentally, a good test case for applying the method of Section IIE outside the area for which it was designed originally.

Since the equation is only to be solved in the high field case we may apply to (20) the reduction method of Section IIA. If we introduce also the specialization warranted by the hard sphere model and unit mass ratio then, in analogy to equation (40), we get the following starting equation

$$\frac{\partial g(\mathbf{w})}{\partial w_z} + wg(\mathbf{w}) - \frac{1}{4\pi}\int_0^\pi \frac{w'^4 \sin\chi\, d\chi}{w^3}\int_0^{2\pi} g(\mathbf{w}')\, d\omega = -\lambda k\{w_z - \langle w_z\rangle\}h(\mathbf{w}) \tag{153}$$

Here the dimensionless variable $\mathbf{w}$ defined by (85) has been employed instead of $\mathbf{c}$.

Equation (153) is the fundamental equation of our problem; it is an inhomogeneous version of equation (40). We solve the equation in the same way as we did previously, namely by decomposing $g(\mathbf{w})$ into spherical harmonics and forming moments. In other words we follow step by step the procedure of Section IIA, the only difference being the presence of an inhomogeneous term. We shall not enumerate all these steps again. We shall only note in passing the inhomogeneous form of (47) which is

$$\frac{1}{2}\int_0^\pi \frac{w'^4}{w^3} g_\nu(w')P_\nu(\cos\kappa)\sin\chi\, d\chi - wg_\nu(w) - \frac{\nu}{2\nu-1}\left\{\frac{dg_{\nu-1}}{dw} - \frac{\nu-1}{w}g_{\nu-1}(w)\right\} - \frac{\nu+1}{2\nu+3}\left\{\frac{dg_{\nu+1}}{dw} + \frac{\nu+2}{w}g_{\nu+1}(w)\right\} = \lambda k\left[\frac{\nu}{2\nu-1}wh_{\nu-1}(w) + \frac{\nu+1}{2\nu+3}wh_{\nu+1}(w) - \langle w_z\rangle h_\nu(w)\right]$$

Having introduced moments in the manner described earlier we arrive at the inhomogeneous version of (49) or (86)

$$\begin{aligned}\nu(s+\nu+1)\{s-1,\nu-1\}+(\nu+1)(s-\nu)\{s-1,\nu+1\}\\ -(2\nu+1)(1-\langle I_{s,\nu}\rangle)\{s+1,\nu\}=-(2\nu+1)\langle 1,1\rangle\langle s,\nu\rangle\\ +\nu\langle s+1,\nu-1\rangle+(\nu+1)\langle s+1,\nu+1\rangle\end{aligned}\tag{154}$$

Here the curly brackets $\{s, \nu\}$ are normalized moments over $g(\mathbf{w})$ defined as follows

$$\{s,\nu\}=\frac{1}{k\lambda}\int g(\mathbf{w})w^sP_\nu(\cos\vartheta)\,d\mathbf{w}\tag{155}$$

The equations (154) show that the quantities $\{s, \nu\}$ are numbers, the variable density gradient nk having been eliminated by the definition (155). The system does permit that arbitrary amounts of the pointed averages be added to the curly ones. This indeterminacy is removed by the supplementary condition (18) which, in the present notation reads

$$\{0,0\}=0\tag{156a}$$

The connectivity of the equation system (154) is the same as that of (86). Hence it will have the same properties as that earlier system. We may, therefore, reduce it in the manner followed previously and get inhomogeneous versions of the equations (87), (88) and (89). They read

$$\{4,0\}=-\frac{5}{3}\langle 4,1\rangle+\frac{5}{3}\langle 1,1\rangle\langle 3,0\rangle.\tag{157a}$$

$$\begin{aligned}112\{3,0\}-3\{7,0\}&=4\langle 7,1\rangle-4\langle 1,1\rangle\langle 6,0\rangle\\ &+\frac{56}{5}\langle 5,0\rangle+\frac{112}{5}\langle 5,2\rangle\\ &-\frac{168}{5}\langle 1,1\rangle\langle 4,1\rangle\\ &-\frac{1344}{25}\langle 3,1\rangle+\frac{1344}{25}\langle 3,3\rangle\\ &+\frac{448}{5}\langle 1,1\rangle\langle 2,0\rangle-\frac{448}{5}\langle 1,1\rangle\langle 2,2\rangle\end{aligned}\tag{158a}$$

$$-\frac{10}{3}\langle 2,0\rangle-\frac{20}{3}\langle 2,0\rangle+10\langle 1,1\rangle^2$$

$$54\{2,0\} - \frac{295}{28}\{6,0\} + \frac{17}{165}\{10,0\} = -\frac{17}{135}\langle 10,1\rangle + \frac{17}{135}\langle 1,1\rangle\langle 9,0\rangle$$
$$- \frac{17}{36}\langle 8,0\rangle - \frac{17}{18}\langle 8,2\rangle + \frac{17}{12}\langle 1,1\rangle\langle 7,1\rangle$$
$$+ 6\langle 6,1\rangle - \frac{21}{5}\langle 6,3\rangle - \frac{44}{5}\langle 1,1\rangle\langle 5,0\rangle + 7\langle 1,1\rangle\langle 5,2\rangle \qquad (159a)$$
$$+ \frac{54}{5}\langle 4,0\rangle + \frac{108}{7}\langle 4,2\rangle - \frac{288}{35}\langle 4,4\rangle$$
$$- \frac{162}{5}\langle 1,1\rangle\langle 3,1\rangle + \frac{72}{5}\langle 1,1\rangle\langle 3,3\rangle$$

The pointed averages over the distribution $h(\mathbf{w})$ may be found in Table II. Substituting them we get

$$\{0,0\} = 0 \qquad (156b)$$

$$\{4,0\} = -10.494 \qquad (157b)$$

$$112\{3,0\} - 3\{7,0\} = 647.8 \qquad (158b)$$

$$54\{2,0\} - \frac{295}{28}\{6,0\} + \frac{17}{165}\{10,0\} = -566.4 \qquad (159b)$$

The form (90) that was assumed for $h_0(w)$ will again be taken for $g_0(w)$ with new undetermined coefficients p, q, r, s and a factor $k\lambda$ evident from (153) or (155):

$$g_0(w) = k\lambda[pEi(\tfrac{1}{2}w^2) + qK_0(\tfrac{1}{2}w^2) + re^{-\frac{1}{2}w^2} + sw^2K_1(\tfrac{1}{2}w^2)] \qquad (160)$$

This is a rather poor assumption because the form (90) was adopted for $h_0(w)$ after an extensive study of the properties of the distribution function $h(\mathbf{w})$. For $g(\mathbf{w})$ we know little beyond the fact that it is some kind of distorted p-type function. The $g_0(w)$ derived from this is not likely to resemble $h_0(w)$ very closely. Thus the choice (160) is mainly based on ignorance and convenience; this explains the slower convergence observed here than in (91), (93) and (95). To start with, the zero order is completely lost because (156) yields a zero coefficient. We find in first order, using (156) and (157)

$$p^{(1)} = 4.8842 \qquad r^{(1)} = s^{(1)} = 0$$
$$q^{(1)} = -4.2689$$

in second order, using (156), (157) and (158)

$$p^{(2)} = -10.542$$
$$q^{(2)} = +14.993 \qquad s^{(2)} = 0$$
$$r^{(2)} = -4.408$$

in third order, using (156), (157), (158) and (159)

$$p^{(3)} = -0.8710$$
$$q^{(3)} = +1.1754$$
$$r^{(3)} = +1.0140$$
$$s^{(3)} = -0.5809$$

The longitudinal diffusion coefficient results from these numbers by the use of (23), (24) and (160). With the notation (155) the formula becomes

$$D_{||} = -a^{1/2}\lambda^{3/2}\{1, 1\} \tag{161}$$

The formula (154) yielding $\{1, 1\}$ from $g_0(w)$ is $s = 2$, $\nu = 0$

$$\{1, 1\} = \tfrac{1}{4}\{3, 0\} + \tfrac{1}{2}\langle 3, 1\rangle - \tfrac{1}{2}\langle 1, 1\rangle\langle 2, 0\rangle \tag{162a}$$

or numerically from the Table II

$$\{1, 1\} = \tfrac{1}{4}\{3, 0\} + 0.6433 \tag{162b}$$

The result is

$$\{1, 1\}^{(1)} = -0.3695 \tag{163a}$$
$$\{1, 1\}^{(2)} = -0.2075 \tag{163b}$$
$$\{1, 1\}^{(3)} = -0.2198 \tag{163c}$$

The numbers do not extrapolate too reliably but one would guess that

$$\{1, 1\} = -0.22$$

is essentially correct. Hence we have

$$D_{||} = 0.22a^{1/2}\lambda^{3/2} \tag{164}$$

In order to gain an appreciation of the value obtained it is worthwhile to compare it with the value that would have been predicted from the generalized Nernst-Townsend relation (148). The mobility concept is ambiguous for all but the cases discussed then. It would seem that the appropriate concept here is the differential mobility because comparison

is made between a small density gradient and a small change in the applied field. Thus we would interpret (148) to mean

$$D_{||} \approx \frac{\partial \langle c_z \rangle}{\partial a} [\langle c_z^2 \rangle - \langle c_z \rangle^2] \tag{165a}$$

which, with (85), (92) and (96), becomes

$$D_{||} \approx 0.26 a^{1/2} \lambda^{3/2} \tag{165b}$$

The error of formula (165) is thus 18 per cent, when compared to (164).

Part V — Concluding Observations

The present article is supposed to contain the essentials of a kinetic theory of charged particles moving through a gas in the presence of an intermediate or high electric field. An effort was made to make the theory general, yet many irksome restrictions will become apparent to those who will try to apply it to their particular problem. Especially those who have in mind application to electrons will find the article unsatisfactory. It is true that many sections leave the masses variable; however, the assumption of elastic collisions, which is made throughout, is almost fatal to all but the most elementary applications. Thus most of the material is slanted for ions. Within this domain, numerous awkward restrictions are still found here. The most important ones are presumably the restriction to D.C. conditions, the assumption of "low" ion density, and the omission of all magnetic effects. It is my general impression, which I gained from the convolution theorem Section IIIA and which is confirmed by a recent publication[27] that much can be done to remove these three restrictions provided the mean free time assumption is made for collisions. To many the adoption of the mean free time condition will in itself appear an awkward restriction. In a rigorous sense this is true, and calculations are made in this article for the more appropriate hard sphere model when quantitative comparison with experiment is contemplated (equations (100) and (164)). Indications are even given for a treatment which dispenses altogether with the use of models (equation (103)). However, for rapid advance and easy handling, the mean free time assumption does appear essential. It is therefore important to point out that in a wider semiquantitative sense, the use of this model is no barrier to application. In other words, there is in the mean free time formulas information which suggests a wider validity. This is particularly true for equations which do not contain model parameters, such as (55), (56), (122) and (148). Even formulas which

do contain the mean free time yield to judicious treatment. For example, we have the hard sphere formula (100) for the drift velocity of an ion. This formula happens to be limited to the high field case and mass ratio unity. On the other hand we have formula (136) which holds for all fields and all mass ratios, but assumes constant mean free time. We now adopt this formula as a general guess for the hard sphere model, interpreting τ as previously as the mean free time between collisions; this quantity is now no longer a constant, but should be taken as

$$\tau = \frac{\lambda}{\sqrt{\langle c^2 \rangle + \langle C^2 \rangle}} \tag{166}$$

The denominator is the root mean square relative velocity which is familiar from other applications. The interpretation (166) yields a tentative formula for the drift for all mass ratios and for all fields. Specializing to the high field case, we may neglect $\langle C^2 \rangle$ in (166) and then substitute for $\langle c^2 \rangle$ from (55). This yields the high field formula

$$\langle c_z \rangle \approx \frac{(M + m)^{1/4} m^{1/4}}{M^{1/2}} (a\lambda)^{1/2} \tag{167}$$

This is indeed a very successful formula. For ions in the parent gas it differs from (100) by only 4 per cent. For electrons it checks the result of Druyvesteyn[4] to within 12 per cent. Finally, for heavy ions in a light gas, we find exact agreement with equation (71). As a second specialization we may apply (166) to the low field case. We must then set

$$\langle c^2 \rangle + \langle C^2 \rangle = 3kT \left(\frac{1}{m} + \frac{1}{M} \right)$$

and get from (136) and (166)

$$\langle c_z \rangle = \frac{1}{\sqrt{3}} \left(\frac{1}{m} + \frac{1}{M} \right)^{1/2} \frac{eE\lambda}{\sqrt{kT}} \tag{168}$$

All dimensional factors in this formula are correct. Numerically (168) is somewhat inferior to (167); for the factor differs from the correct one[33] by 20 per cent. Nevertheless, the combination of (136) and (166) gives results which are semiquantitatively correct in all relevant limiting cases. This makes it a reliable interpolation formula for intermediate field conditions; for this case $\langle c^2 \rangle$ would have to be substituted from (122) and the resultant quadratic equation solved for $\langle c_z \rangle$.

From the examples given we may conclude that the mean free time formulas contain in essence information applicable to other types of elastic scattering.

[33] See Reference 2, page 40, second equation.

Part VI — Acknowledgements

This article is the outcome of years of fruitful cooperation with the gas discharge group of Bell Telephone Laboratories, and Dr. J. A. Hornbeck in particular. At one stage of the work I enjoyed the stimulation of Dr. R. W. Hamming who is responsible for all the details of the Monte Carlo calculation. Further acknowledgements are due to Miss C. L. Froelich who carried out the computation of the number in (130), and Miss M. Murray who carried a good share of the burden in the preparation of the manuscript. Finally, I express my thanks to Dr. K. G. McKay for his critical perusal of the manuscript.

Abstracts of Bell System Technical Papers* Not Published in This Journal

Experimental Verification of the Theory of Laminated Conductors. H. S. BLACK[1], C. O. MALLINCKRODT[5] and S. P. MORGAN[1]. *I.R.E., Proc.*, **40,** pp. 902–905, August, 1952.

Clogston has discovered that if a conductor is properly laminated, there exists a particular phase velocity along the conductor for maximum penetration of the fields and minimum loss due to skin effect. An experimental coaxial line was constructed whose center conductor was laminated and whose phase velocity could be varied by changing the dielectric constant of the main dielectric. As predicted by theory, the measured attenuation was critically dependent upon phase velocity. With optimum phase velocity the attenuation, though greater than predicted by theory, was less than that of a conventional coaxial cable of the same dimensions and same main dielectric. A theoretical analysis of the experimental laminated conductor is described in an append x.

ASTM Standards—Their Effect on Plastics Technology. R. BURNS[1]. *A.S.T.M. Bull.*, No. 183, pp. 78–80, July, 1952.

Typical Block Diagrams for a Transistor Digital Computer. J. H. FELKER[1]. *A.I.E.E., Trans., Commun. & Electronics Sect.*, No. 1, pp. 175–182, July, 1952.

The first electric digital computers were built around the properties of relays. The superior speed capabilities of vacuum tubes has led in recent years to their use in new computer designs to replace relays. Because of the small size, low power consumption, and expected long life of transistors, it now appears that the transistor will replace the vacuum tube as a computer element. This paper presents a study of binary computer functions with recommended mechanizations that were selected because they appeared to be readily attainable with transistors now under development. Block diagrams are presented of switches, memory units, arithmetic units, and other basic components. Estimates are given for the number of parts required in the units. It is concluded that a high-performance all-semiconductor computer can be built with germanium diodes and transistors.

* Certain of these papers are available as Bell System Monographs and may be obtained on request to the Publication Department, Bell Telephone Laboratories, Inc., 463 West Street, New York 14, N. Y. For papers available in this form, the monograph number is given in parentheses following the date of publication, and this number should be given in all requests.

[1] Bell Telephone Laboratories.

[5] Hughes Aircraft Company, Culver City, California

A Broad-Band Interdigital Circuit for Use in Traveling-Wave-Type Amplifiers. R. C. FLETCHER[1]. *I.R.E., Proc.*, **40,** pp. 951–958, August, 1952.

Because of its high power-handling capacity, the interdigital circuit has been considered for use in traveling-wave-type amplifiers. An analysis is presented here which indicates that this type of circuit can be arranged to give constant phase velocity over a wide bandwidth (30 per cent), as required to give constant gain. The analysis is qualitatively checked experimentally. The impedance parameter (proportional to the cube of the gain in db) is approximately the same as the flattened helix (such as has been used for the magnetron amplifier) and about one-third that of the conventional circular helix.

Copper as an Acceptor Element in Germanium. C. S. FULLER[1] and J. D. STRUTHERS[1]. *Phys. Rev.*, **87,** pp. 526–527, August 1, 1952.

Properties of Thermally Produced Acceptors in Germanium. C. S. FULLER[1], H. C. THEUERER[1] and W. VAN ROOSBROECK[1]. *Phys. Rev.*, **85,** pp. 678–679, February 15, 1952.

Initial Permeability and Related Losses in Ferrites. J. K. GALT[1]. *Ceramic Age*, **60,** pp. 29–33, August, 1952.

Three-Phase Power From Single-Phase Source. A. L. HOLCOMB[6], *S.M.P.T.E., Jl.*, **59,** pp. 32–39, July, 1952.

Described is the development of a nonrotating device for the conversion of single-phase 115-volt power to a three-phase 230-volt form for the synchronous operation of cameras, sound recorders and other film pulling mechanisms associated with production of motion pictures.

Dominant Wave Transmission Characteristics of a Multimode Round Waveguide. A. P. KING[1]. *I.R.E., Proc.*, **40,** pp. 966–969, August, 1952.

This paper presents some dominant wave transmission characteristics of multimode round waveguide lines in the 4-kmc range of frequencies. The use of such waveguide lines offers the advantages of lower transmsision losses than obtainable with single-mode rectangular waveguide, and relative ease of making good joints. Possible mode conversion effects, including dominant mode elliptical polarization, have been examined and found to be innocuous. As a result, cross-polarized dominant waves can be used to provide two reasonably independent signaling channels at the same frequency in one pipe. The experimental results obtained with a straight line 2.812-inch inside diameter and length of 150 feet are given.

Superconductivity in the Cobalt-Silicon System. B. T. MATTHIAS[1]. Letter to the Editor. *Phys. Rev.*, **87,** p. 380, July 15, 1952.

[1] Bell Telephone Laboratories
[6] Westrex

Simple Phase-Angle Measurement Technique. J. A. RUDISILL, JR.[3]. *Electronics*, **25,** pp. 228, 232, 236, September, 1952.

Solid State Physics in Electronics and in Metallurgy. W. SHOCKLEY[1]. *Jl. Metals*, 4, pp. 829–842, August, 1952. (Monograph 2011).

Impurity Effects in the Thermal Conversion of Germanium. W. P. SLICHTER[1] and E. D. KOLB[1]. *Phys. Rev.*, **87,** pp. 527–528, August 1, 1952.

Traffic Engineering Design of Dial Telephone Exchanges. J. A. STEWART[4]. *Midwest Engr.*, **4,** pp. 3–5, 17–18, May, 1952.

Single-Crystal Germanium. G. K. TEAL[1], M. SPARKS[1] and E. BUEHLER.[1] *I.R.E., Proc.*, **40,** pp. 906–909, August 1952.

Significant advances have been made in the development of new types of transistors, photocells, and rectifiers and in the improvement of the reproducibility and reliability of the point-contact transistor. A key factor in this development has been the use of single-crystal germanium having a high degree of lattice perfection and compositional control. Of particular interest to the device-development engineer is the fact that the rectifying barriers between the *p*-type and *n*-type sections behave in a manner predictable from the measured properties of each section. The exceptionally long lifetime of injected carriers observed in the material and the high degree of control over its chemical composition make it ideally suitable for the production of *p-n* structures. The ranges of properties of germanium single crystals which are now realizable are given, as well as their present degree of control.

Lead-Acid Stationary Batteries. U. B. THOMAS[1]. *Electrochem. Soc. Jl.*, **99,** pp. 238C–241C, September, 1952.

Polymorphism of $ND_4D_2PO_4$. E. A. WOOD[1], W. J. MERZ[1] and B. T. MATTHIAS[1]. *Phys. Rev.*, **87,** p. 544, August 1, 1952.

Lightning Protection for Mobile Radio Fixed Stations. D. W. BODLE[1]. *I.R.E. Trans., P.G.V.C.*-1, pp. 122–133, February, 1952.

Equipment in fixed stations of a mobile radio system is susceptible to damage from lightning strokes to either the antennas or the connecting power and land communication facilities unless special protection is provided. The problem, however, is not alone one of protecting the station equipment, but consideration must also be given to the protection of these connecting facilities to insure their continuity of service. The causes of and factors affecting lightning damage are discussed, including the probable incidence of strokes to the antennas. General protection principles are outlined, and the application of specific protection methods is described.

---

[1] Bell Telephone Laboratories
[3] Western Electric Company
[4] Illinois Bell Telephone Company

Matching Coax Line to the Ground-Plane Antenna. R. T. DeCamp[2]. *QST*, **36,** pp. 18–19, 120, 122, September, 1952.

This article describes a method for predetermining antenna and matching-stub dimensions for matching any selected transmission line. Although applied particularly to the ground-plane antenna, the curves are useful for half-wave dipoles if allowance is made, when necessary, for the effect of ground on the antenna characteristics.

The Telephone System in National Defense. C. M. Mapes[2]. *Military Engr.*, **44,** pp. 375–377, September–October, 1952.

Multi-Element Directional Couplers. S. E. Miller[1] and W. W. Mumford[1]. *I.R.E., Proc.*, **40,** pp. 1071–1078, September 1952.

It is shown that the backward wave in a directional coupler is related to the shape of the function describing the coupling between transmission lines by the Fourier transform. This facilitates the design of directional couplers for arbitrary directivities over any prescribed frequency band. Tightly coupled directional couplers are analyzed in simple terms, and it is shown that any desired loss ratio, including complete power transfer between lines, may be achieved. The theories are verified using waveguide models operating at 4,000, 24,000, and 48,000 mc, and it is indicated that the work is applicable to many types of electrical and acoustic transmission lines.

Segregation of Two Solutes, with Particular Reference to Semiconductors. W. G. Pfann[1]. *Jl. Metals*, **4,** pp. 861–865, August, 1952. (Monograph 2020).

The simultaneous segregation of two solutes during the directional solidification of an ingot is treated mathematically on the basis of simplifying assumptions. Expressions are derived for the difference in concentration of two solutes, and for the location and concentration gradient of a pn barrier formed in a semiconductor by the segregation of a donor and an acceptor.

Nonsynchronous Time Division with Holding and with Random Sampling. J. R. Pierce[1] and A. L. Hopper[1]. *I.R.E., Proc.*, **40,** pp. 1079–1088, September, 1952.

There is a general type of system in which an indefinitely large number of transmitters can have access to any of an indefinitely large number of receivers over a medium of limited band-width. In these systems, signal-to-noise ratio goes down as more transmitters are used simultaneously. This paper describes a particular system which sends samples by means of coded pulse groups sent at random times. The signal-to-noise ratio is good in the absence of interference and the effect of interference is minimized by holding the previous sample if a sample is lost. An experimental system worked satisfactorily and gave close to the predicted signal-to-noise ratio might be used to provide communication and automatic switching in rural telephony, or for other applications.

---

[1] Bell Telephone Laboratories
[2] American Telephone and Telegraph Company

An Improved Electrolysis Switch. V. B. PIKE[1]. *Corrosion*, **8,** pp. 311–313; disc. pp. 322–323, September, 1952. (Monograph 2021).

An improved electrolysis switch has been placed in use in the Bell System for mitigation of electrolysis of cables by stray currents. It consists of three relays operating in sequence, namely control, intermediate and drain relays. It automatically closes a drainage bond between a lead-covered underground cable and a power return ground when stray current is picked up by the cable to such an extent as to make some of the sheath positive with respect to its environment. It also opens bond when the drainage current falls to zero and drainage is no longer required. Two sizes are used capable of draining 200 amperes and 400 amperes, respectively. Its action is very fast, only 0.015 second elapsing from the time the control circuit releases until opening of the drainage bond. Separate adjustments are available for setting the voltage at which the switch closes and opens the drainage bond. A capacitive voltage booster enables switch operation over longer battery power supply wires than is possible if the booster is not used. A power supply unit consisting of a stepdown transformer and selenium rectifier is also available to operate the switch with power drawn from an AC power source if the switch must be installed beyond reach of the battery power. A sealed steel housing enables the switch to withstand submersion and permits installation in very damp locations.

Statistics of the Recombinations of Holes and Electrons. W. SHOCKLEY[1] and W. T. READ, JR.[1]. *Phys. Rev.*, **87,** pp. 835–842, September 1, 1952. (Monograph 2022).

The statistics of the recombination of holes and electrons in semiconductors is analyzed on the basis of a model in which the recombination occurs through the mechanism of trapping. A trap is assumed to have an energy level in the energy gap so that its charge may have either of two values differing by one electronic charge. The dependence of lifetime of injected carriers upon initial conductivity and upon injected carrier density is discussed.

Motion of Gaseous Ions in a Strong Electric Field. G. H. WANNIER[1]. *Phys. Rev.*, **87,** pp. 795–798, September 1, 1952. (Monograph 2023).

This paper continues an earlier one on the same subject. Its object is to elucidate the nature of the random motion of an ion about its drift. In Section F it is shown that this motion can be described as a diffusion with a diffusion tensor axially symmetric about the field. If the mean free time between the collisions of an ion with molecules is independent of speed, then explicit expressions may be deprived for the two diffusion coefficients; these expressions are written down without proof in Section G; they are connected with the mobility by a natural extension of the Einstein relation. In Section H, the longitudinal diffusion coefficient is computed numerically for the hard sphere model, high field, and mass ratio 1; the method of computation is the same as in Section D. Finally, it is shown in Section I how approximate formulas of wider validity can be inferred from the ones obtained.

---

[1] Bell Telephone Laboratories

High-Frequency Crystal Units for Primary Frequency Standards. A. W. WARNER[1]. *I.R.E., Proc.*, **40,** pp. 1030–1033, September, 1952.

A new approach to the design of crystal units for primary frequency standard use has resulted in crystal units in the 3- to 20-mc frequency range characterized by high Q and low capacitance in the series arm of the equivalent electrical circuit. By utilizing the overtone frequency of specially shaped AT-cut quartz plates, both Q and the rate of impedance change with frequency are enhanced together, and in addition the stability with time of the crystal unit is increased because of a larger frequency-determining dimension. Additional characteristics of the crystal units include small size, stability under conditions of vibration and shock, and low-temperature coefficient. Crystal-oscillator stabilities of one part in 10^8 per month have been achieved without recourse to stabilized circuits.

---

[1] Bell Telephone Laboratories

Contributors to this Issue

WALLACE C. BABCOCK, A.B., Harvard University, 1919; S.B., Harvard University, 1922. U. S. Army, 1917–19. American Telephone and Telegraph Company, 1922–34; Bell Telephone Laboratories, 1934–. Mr. Babcock was engaged in crosstalk studies until World War II, when he studied radio countermeasure problems for the N.D.R.C. Since then he has been concerned with antenna development for mobile radio and point-to-point radio telephone systems and has been engaged in other systems studies. Member of I.R.E. and Harvard Engineering Society.

JOHN BARDEEN, B.S., in E.E., University of Wisconsin, 1928; M.S. in E.E., University of Wisconsin, 1929; Ph.D., Princeton University, 1936. Gulf Research and Development Company, 1930–33; Harvard University, 1935–38; University of Minnesota, 1938–41; Naval Ordnance Laboratory, 1941–45; Bell Telephone Laboratories, 1945–51; University of Illinois, 1951–. At Bell Telephone Laboratories, Dr. Bardeen, co-inventor with Dr. Walter Brattain of the point-contact transistor, was primarily concerned with theoretical problems in solid state physics, including the study of semi-conductors, diffusion in solids, and superconductivity. Associate editor of *The Physical Review*, 1949–51. Stuart Ballantine Medal of the Franklin Institute, 1952. Fellow, American Physical Society; member, American Association for the Advancement of Science.

WALTER H. BRATTAIN, B.S., Whitman College, 1924; M.A., University of Oregon, 1926; Ph.D., University of Minnesota, 1929. Bureau of Standards, 1928–29; Bell Telephone Laboratories, 1929–. Dr. Brattain, co-inventor with Dr. John Bardeen of the point-contract transistor, has been primarily concerned with the study of semi-conductors at Bell Laboratories. During World War II he worked for the Division of War Research of Columbia University and is currently spending the fall term of the academic year 1952–53 as a visiting lecturer at Harvard University. Stuart Ballantine Medal of the Franklin Institute, 1952. Fellow, American Physical Society and American Association for the Advancement of Science; member, Sigma Xi and Phi Beta Kappa.

KENNETH BULLINGTON, B.S. in E.E., University of New Mexico, 1936; S.M., Massachusetts Institute of Technology, 1937; Bell Telephone Laboratories, 1937–. Until World War II, Mr. Bullington was occupied with systems engineering work on wire transmission circuits. Since 1942, he has been concerned with transmission engineering on radio systems, especially with radio propagation studies. Member of I.R.E., Phi Kappa Phi, Sigma Tau, and Kappa Mu Epsilon.

R. H. COLLEY, A.B., Dartmouth College, 1909; A.M., Harvard University, 1912; Ph.D., George Washington University, 1918; Austin Teaching Fellow in Botany, Harvard University, 1910–12; Instructor in Botany, Dartmouth College, 1909–10 and 1912–16; Pathologist, Division of Forest Pathology, Bureau of Plant Industry, U. S. Department of Agriculture, 1916–28. Bell Telephone Laboratories, 1928–1952. Dr. Colley was chairman of Committee 05—Wood Poles, of the American Standards Association for nearly twenty years. He was president of the American Wood-Preservers' Association 1943–44. During his years with the Laboratories he worked particularly on development and research problems connected with material and preservative treatment specifications for poles and other timber products used in outside plant. His more recent activities were directed toward improvement of laboratory techniques for evaluating wood preservatives, and toward the development of a coordinated plan for fundamental research on oil preservatives. He was Timber Products Engineer for the Laboratories from 1940 to 1950, and Timber Products Consultant from 1950 to 1952. His article in this issue of the JOURNAL was prepared before his retirement on May 31, 1952.

KARL K. DARROW, B.S., University of Chicago, 1911. He studied at the Universities of Paris and Berlin in 1911 and 1912, specializing in physics and mathematics; Ph.D., University of Chicago, 1917. He then joined the staff of Bell Telephone Laboratories, at that time known as the Engineering Department of Western Electric Company. Here his work has included the study, correlation, and representation of scientific information for his colleagues, keeping them informed of current advances made by workers in fields related to their own activities. As a corollary to his work, Dr. Darrow appears from time to time before scientific and lay audiences to lecture on current topics in physics and the related sciences. He has taken an active interest in education, teaching physics during summer and other sessions at Stanford, Chicago, and Columbia Universities and at Smith College. From 1944 to 1946, he served

as consultant to the Metallurgical Laboratory in Chicago. Dr. Darrow is the author of *Introduction to Contemporary Physics* (1926 and 1939), *Electrical Phenomena in Gases* (1932), *Resonance of Physics* (1936), and *Atomic Energy* (1948) and of many articles in this and other journals. He is a member of the American Physical Society, which he has served as secretary since 1941, the Physical Society of London, Société Francaise de Physique, the American Philosophical Society, of which he was a counsellor for four years. From 1949 to 1951 he was vice-president of the International Union of Pure and applied Physics. In 1949 he received an honorary doctorate from the Université de Lyon and was made Chevalier de la Légion d'Honneur in 1951.

JOHN RIORDAN, B.S., Sheffield Scientific School of Yale University, 1923. United Electric Light and Power Company, now a part of the Consolidated Edison Company, 1923–1926. Department of Development and Research of the American Telephone and Telegraph Company, 1926–1934. Bell Telephone Laboratories, 1934–. With the American Telephone and Telegraph Company, his work was largely on circuit and transmission theory, particularly in relation to inductive interference from electrified railways. This work was continued in Bell Telephone Laboratories after the Development and Research Department was consolidated with it in 1934. Since 1940 Mr. Riordan has been engaged in mathematical work: Boolean algebra in switching, number theory in cable splicing, and combinatorial and probability studies of traffic. Mr. Riordan is a member of the American Mathematical Society, the Mathematical Association of America, the Institute of Mathematical Statistics, and is a Fellow of the American Association for the Advancement of Science.

GREGORY H. WANNIER, Louvain University, 1930–31; University of Cambridge, 1933–34; Ph.D., University of Basel, 1935. Assistant, University of Geneva, 1935–36; Swiss-American Exchange Fellow, Princeton University, 1936–37; instructor, University of Pittsburgh, 1937–38; assistant lecturer, Bristol University, 1938–39, instructor, University of Texas, 1939–41; University of Iowa, 1941–46. Socony-Vacuum Laboratories, 1946–49. Bell Telephone Laboratories, 1949–. Mr. Wannier, a theoretical physicist in the Physical Research Department, has worked on photoconductivity and related phenomena, and the motion of ions in gases. Also Mission to Germany, 1945. Member of American Physical Society and Schweizer Physikalishe Gesellschaft.