THE BELL SYSTEM
TECHNICAL JOURNAL

VOLUME XLIII MAY 1964 NUMBER 3

Digital Light Deflection

By T. J. NELSON

(Manuscript received July 30, 1963)

A digital method of deflecting a light beam using n optical modulators and n uniaxial crystals to provide 2^n positions of the beam is described. The input-output relations for one special configuration are derived. Optical problems and limitations are investigated and, in particular, it is found that the upper limit to the density of positions is about 10^6/sq. in. Presently available modulators are considered, and it is found that a KDP modulator has a power limitation above 1 mc operation for a total number of positions of about 70,000. Finally, the applications of the method as a semipermanent memory, a PCM decoder, and a digital delay line are briefly considered.

I. INTRODUCTION

It is desirable to substitute a light beam for the electron beam used in the class of devices of which the flying spot scanner and the Williams tube are examples. In this class of devices the electron beam is used to probe a suitable target and read or store information on it. The substitution is desirable because a light beam has negligible inertia and can, in principle, be deflected rapidly. The subject of this paper*· †· ‡ is a new

* The contents of this paper were discussed by the author at the Twenty-First Annual Conference on Electron Device Research (IEEE), Salt Lake City, Utah, June 26–28, 1963.

† A brief description of methods that can be used to deflect a light beam is given by U. J. Schmidt (Schmidt, U. J., The Problem of Light Beam Deflection at High Frequencies, Proceedings of the Symposium on Optical Processing of Information, ed. Pollack, D. K., Koester, C. J., and Tippett, J. T., Spartan Books, Inc., Baltimore, 1963, p. 98).

‡ *Note Added in Proof*. An article on some aspects of digital light deflection has recently appeared in the literature: see Kulcke, W., Harris, T. J., Kosanke, K., and Max, E., IBM Journal of Research and Development, **8**, 1964, pp. 64–67.

method of deflecting a light beam. This method, which is called digital light deflection, employs n optical modulators and n uniaxial crystals. By applying appropriate two-level electrical inputs to the n optical modulators, it is possible to deflect a light beam to 2^n positions. This deflection technique is inherently digital and the required optical modulator inputs are binary signals.

In the present paper we shall consider:

(a) the basic principles of digital light deflection

(b) the address logic for a special configuration

(c) optical problems which limit the density of positions that can be achieved

(d) presently available optical modulators and the speed limitations which they impose on a digital light deflection system

(e) experimental results that have been obtained with a four-unit (16-position) digital light deflection system.

II. PRINCIPLES OF DIGITAL LIGHT DEFLECTION

2.1 *The Binary Unit*

It is well known that uniaxial crystals have the property of displacing light of one polarization, called the extraordinary ray, while the orthogonal polarization, the ordinary ray, obeys Snell's law. If the two rays are parallel upon entering the crystal, they will be parallel upon leaving, but are not parallel inside the anisotropic medium.

Fig. 1 shows a uniaxial crystal oriented so that its optic axis lies in an xz plane. If the electric field vector of a plane-polarized beam is in the

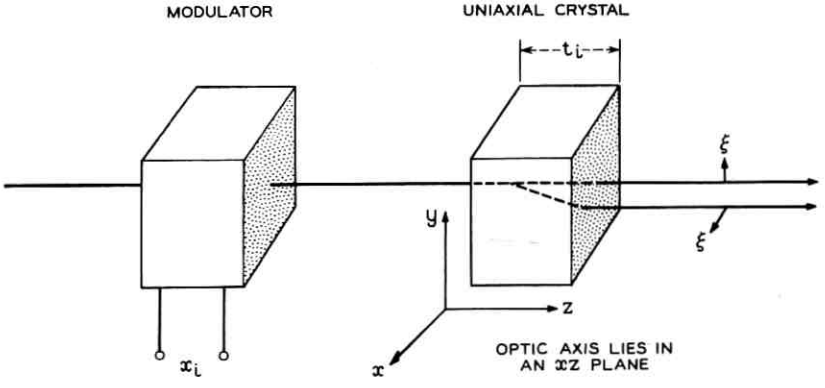


Fig. 1 — The binary unit.

x direction, it will be displaced by an amount proportional to the thickness of the crystal. The exact dependence of the deflection on the crystal thickness and orientation is given in Appendix A. If the beam is polarized in the y direction, its electric field vector is normal to the optic axis, and the beam will pass through the crystal in a straight line.

If a modulator precedes the uniaxial crystal and is capable of rotating the plane of polarization from the x direction to the y direction, and inversely, under the influence of an input signal, then it is possible to switch the beam from one position to another. Furthermore, an error in the input signal will cause the beam to be split and light transmitted simultaneously to the two positions, rather than to some other position, as would be the case with an analog deflector.

The combination of optical modulator and uniaxial crystal, in Fig. 1, is referred to as a binary unit. Because the use of a multiplicity of binary units will be considered next, subscripts identify the variables of a particular binary unit. For example, t_i represents the thickness of the uniaxial crystal in the ith binary unit and x_i describes the state of the ith modulator. If $x_i = 1$, then the plane of polarization of the incident light beam is rotated 90° by the ith modulator; whereas if $x_i = 0$ the rotation is zero.

2.2 *The Deflection Bank*

If a linearly polarized light beam is made to traverse a sequence of n binary units of the type described, then a maximum of 2^n positions of the beam can be realized. Such a combination of n binary units is designated as a deflection bank, and if the deflections are all in the x direction, as the x-deflection bank. The resulting pattern of deflections produced by a bank depends upon the thickness and orientation of the various uniaxial crystals in the bank.

A linear pattern of 2^n positions, uniformly separated, can be obtained with an n-unit bank if all uniaxial crystals have the same orientation and if their thicknesses are respectively $t_0, 2t_0, \cdots, 2^{n-1}t_0$. If the beam displacement in the thinnest crystal is d_0, then the separation between positions is also d_0. Each binary unit in such a bank is unique, as its uniaxial crystal differs in thickness from all other uniaxial crystals. The binary units in such a linear deflection bank may be arranged in any order, but we shall confine any further discussion to a configuration where the light beam successively encounters uniaxial crystals of thickness given by

$$t_i = 2^{n-i}t_0; \qquad i = 1, 2, \cdots, n. \tag{1}$$

This configuration will be shown later to have interesting and desirable properties.

A three-unit example of the configuration that has been singled out is shown in Fig. 2. The polarization of the beam before it enters the bank is orthogonal to the optic axes of the uniaxial crystals and the thickness of successive uniaxial crystals decreases by two. The two columns at the right of Fig. 2 give the positions of the beam in binary form and the modulator inputs which are required to deflect the beam to these positions.

2.3 *Address Logic*

In this section, a derivation of input-output relations and a discussion of code conversion are given for the special linear bank configuration which was mentioned in the previous section.

2.3.1 *Input-Output Relations*

For the purpose of deriving the deflection as a function of the modulator inputs, it is assumed that the polarization of the beam, as it enters the bank, is orthogonal to the optic axes of the uniaxial crystals in the bank. Then if none of the modulators are operated, the beam is undeflected by the bank. Since the displacement of the beam by unit i is determined by the plane of polarization of the beam after it has passed the modulator in unit i, whether deflection occurs is determined by the number of times the polarization of the beam has been rotated up to, and

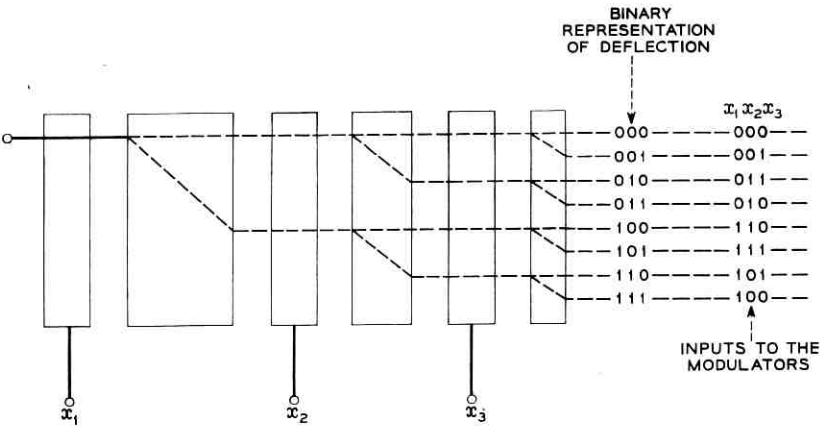


Fig. 2 — Three-unit deflection bank.

including, unit i. If we designate the deflection at unit i as a binary variable, d_i, then the displacement of the beam is $2^{n-i} d_0$ if $d_i = 1$ and 0 if $d_i = 0$. The general formula for the deflection is

$$d_i = S_{\text{odd}}(x_1, x_2, \cdots, x_i) \tag{2}$$

and

$$d_x = \sum_{i=1}^{n} 2^{n-i} d_i d_0 \tag{3}$$

where d_x is the total deflection of the beam in the x-deflection bank. S_{odd} is the notation for a symmetric switching function and its value is one if an odd number of the indicated variables have the value one, and zero otherwise.[1] Thus unit i will add its displacement to the total if the polarization of the beam has been rotated an odd number of times up to, and including, the ith unit.

The use of (2) and (3) is illustrated by considering the three-unit bank in Fig. 2. Suppose the input to the modulators is

$$x_1 x_2 x_3 = 111$$

and we wish to find the corresponding deflection. According to (2) we count, starting from the left, the number of ones appearing up to and including the position under consideration. If the result is odd we enter a one at the corresponding position of the deflection variable. Hence in this case

$$d_1 d_2 d_3 = 101.$$

Substitution of these deflection variables into (3) yields that the total deflection $d_x = 5d_0$.

In general, the addressing or input-output relations have not been completely specified until the inverses of (2) and (3) are given. If we desire to have unit i add its contribution to the total displacement but not unit $i - 1$, or if we desire the displacement of unit $i - 1$ but not that of unit i, then it is clear that the modulator in unit i must rotate the plane of polarization of the beam. This fact is the inverse of (2) and is expressed by

$$x_i = S_{\text{odd}}(d_{i-1}, d_i). \tag{4}$$

The inverse of (3) merely translates the binary number $x_1, x_2, \cdots, x_n$ from base two to base ten.

$$X = \sum_{i=1}^{n} 2^{n-i} x_i. \tag{5}$$

The use of (4) is apparent if we again consider the three-unit bank in Fig. 2. Suppose a total deflection $d_x = 5d_0$ is desired and we wish to find the necessary inputs; then (4) is appropriate. First, however, (3) is used to obtain the deflection variables d_i. Use of (3) gives in this example, as we have seen before, $d_1d_2d_3 = 101$. According to (4), if an entry in the deflection variable differs from the preceding entry, the corresponding entry in the modulator variable x_i is a one. Since the last two entries in $d_1d_2d_3 = 101$ are 01, then $x_3 = 1$. Proceeding in this fashion we obtain

$$x_1x_2x_3 = 111$$

which is the expected result.

An interesting and desirable property of the configuration that has been considered is that it is always possible to switch the beam to adjacent positions by changing the state of excitation of just one modulator. This is apparent for the three-unit bank in Fig. 2. To show this in general, consider the highest numbered unit, j, for which the number of rotations of the plane of polarization up to and including unit j is even. Evidently, all the units following j add their contributions to the total displacement. Since

$$2^j - 1 = 2^{j-1} + 2^{j-2} + \cdots + 2^{j-j+1} + 2^0 \qquad (6)$$

we have only to change the state of excitation to modulator j to increase the total deflection by the incremental distance, d_0. Also, consider the highest numbered unit, k, for which the number of rotations up to and including unit k is odd. Then unit k adds its contribution to the total deflection, but none of the units following k do. Hence, by changing the state of excitation to the modulator in unit k, we decrease the deflection by d_0. The reasoning breaks down in the first case if the excitation is $100 \cdots 0$, where the number of rotations is odd for all units, and in the second case if the excitation is $00 \cdots 0$. However, the total deflection is then either a maximum or zero, and these two conditions are reached from each other by changing the state of excitation to the modulator in unit 1. Hence the input-output relations for this configuration are cyclic; i.e., adjacent positions can always be reached from each other by changing the state of excitation to exactly one modulator.

2.3.2 *Code Conversion*

One can ask, what type of operation can be performed on the input signals to further simplify the input-output relations? If the inputs to a switching network are designated by the W_i's and the outputs by the x_i's, which are, as before, the direct inputs to the modulators, one type of

code conversion could be

$$x_i = S_{\text{odd}}(W_{i-1}, W_i). \tag{7}$$

It is known from (4) that

$$x_i = S_{\text{odd}}(d_{i-1}, d_i); \tag{8}$$

hence, from an inspection of (7) and (8) it is evident that

$$d_i = W_i.$$

Thus, the deflection and input variables are caused to be the same binary numbers. If the W_i's are the outputs of a series of bistable multivibrators which are triggered from a series of harmonically related sine waves, then $W_1, W_2, \cdots, W_n$ and hence $d_1, d_2, \cdots, d_n$ assume the binary numbers in increasing order, and a linear sweep would thereby be effected.

Fig. 3 shows the details of one switching network which can be used for

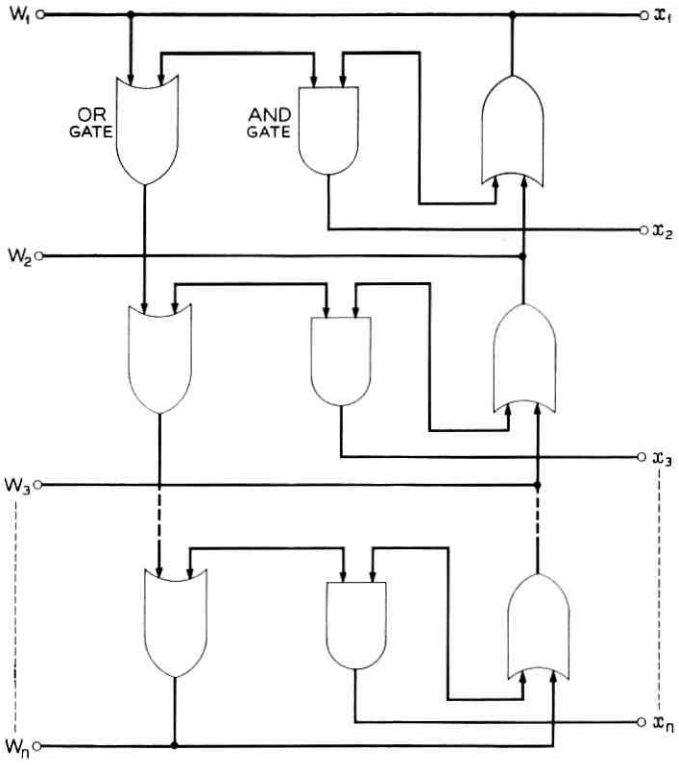


Fig. 3 — Code conversion network.

such code conversion. Two AND gates and one OR gate per binary unit are required, with the exception of the input to unit 1, which requires no alteration.

III. OPTICAL CONSIDERATIONS OF DIGITAL LIGHT DEFLECTION

3.1 *Incorporation of Lens System*

Up to this point, consideration has been limited to plane waves; however, a higher density of resolvable positions is achieved with focused light. If an x-deflection bank and a y-deflection bank are incorporated into the object and image spaces of a lens as shown in Fig. 4, it is possible to focus an aperture into the image plane in a rectangular array of positions. The extra modulator which precedes the y-deflection bank in Fig. 4 insures that the input polarization to the y bank is in the x direction. To accomplish this, the input to the modulator, T, must then be given by

$$T = S_{\text{even}}(x_1, x_2, \cdots, x_n)$$

where S_{even} is a symmetric function whose value is one if an even number of the indicated x variables have the value one, and zero otherwise. Thus, the deflection in the y direction, dx, is related to the y variables in the same manner as the deflection in the x direction, dx, is related to the x variables. Hence the input-output relations developed in the previous section apply to both the x bank and the y bank.

If there are n units in the x bank and m units in the y bank, then the resulting pattern is a matrix of $2^n \times 2^m$ positions. The incorporation of a

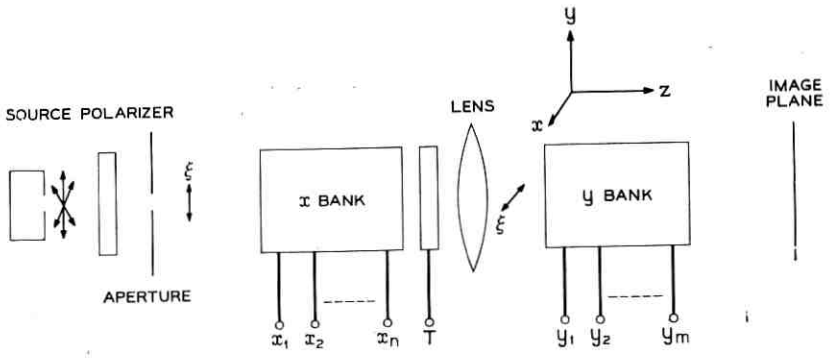


Fig. 4 — Incorporation of x bank and y bank into lens system.

lens into the system causes the rays through the system to have a range of angles with the z axis. Since the deflection in the uniaxial medium is angle-dependent, the maximum angle must be kept small. The optical modulators also limit the range of angles that are tolerable.

For simplicity in the following discussions, the distance from the aperture to the lens is assumed equal to the distance from the lens to the image plane. That is to say, the optical system has unity magnification. This simplifies some of our considerations with little loss in generality, and also unity magnification probably would be used in any practical system.

3.2 *Diffraction Effects*

Because of the anisotropy of the uniaxial crystals and possibly the modulator crystals, we require a high f number for the system. However the minimum resolvable separation between spots is roughly proportional to the f number. If the ratio of the focal length of the lens to the diameter of the lens opening is 15, that is, $f/15$, then the maximum angle any ray through the system can have with the axis of symmetry is $3/\pi$ degrees. The following calculations are based on this compromise.

We define the crosstalk ratio to be

$$C = 10 \log_{10} [P(0)/P(d)] \tag{9}$$

where $P(0)$ is the integrated intensity falling on a circle of radius a, the radius of the aperture, centered on the image and $P(d)$ is the integrated intensity falling in a circle of radius a separated from the image by a center-to-center distance of d. Fig. 5 shows curves of constant cross-

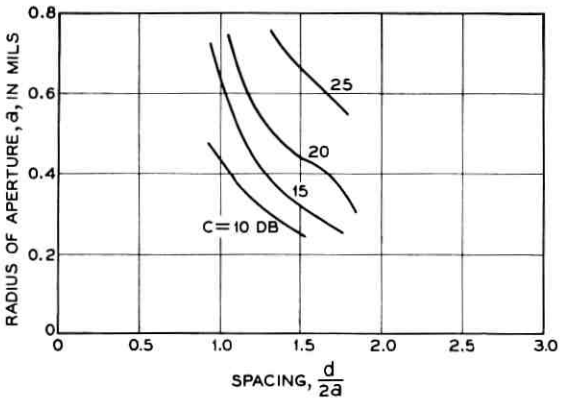


Fig. 5 — Radius of aperture vs center-to-center spacing at constant crosstalk, $f/15$ and 6943 Å.

talk ratio plotted with a as ordinate and $d/2a$ as abscissa. Fig. 6 is a plot of the loss from the integrated intensity admitted by the aperture to that falling on a circle of radius a centered on the image. These data were obtained by numerically evaluating the integrals derived in Appendix B. Figs. 5 and 6 are given for $f/15$ and $\lambda = 6943$ Å. Plane wave illumination of the aperture was assumed, and small angle approximations were made.

From the standpoint of crosstalk, no optimum seems to exist, although we find that for $a = 0.3$ mil and $d = 1.1$ mils, the value $C = 20$ db results. Therefore we find that the spot density, D, can be made as high as

$$D = \frac{1}{(1.1 \times 10^{-3})^2} = 0.826 \times 10^6/\text{in}^2$$

at $f/15$, and 6943 Å. Fig. 6 indicates that the loss would be 11 db for this case.

3.3 *Refractive Index Effects*

In general, a deflected beam passes through some of the uniaxial crystals as an extraordinary beam; hence it encounters a different refractive index from the undeflected beam which passes through all crystals as an ordinary beam. As a result the deflected image of the aperture will be formed at a different z coordinate in the optical system than the undeflected image. For paraxial imaging, which is consistent with our assumption of a fairly large f number, a calculation yields that this shift in the image is given by

$$\Delta Z = (Z_d - Z_0) = \left(\frac{n}{n_r^2} - \frac{1}{n_0}\right) \cot \psi [dx + dy] \qquad (11)$$

where Z_d is the z coordinate of the deflected image and Z_0 is the z co-

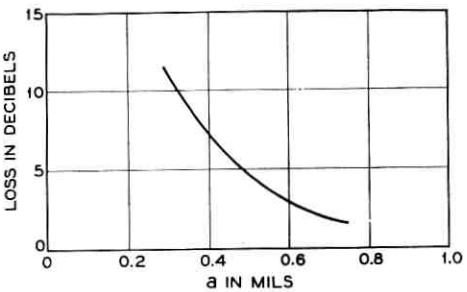


Fig. 6 — Loss vs radius of aperture at $f/15$ and 6943 Å.

ordinate of the undeflected image. Here z is positive in the image space and is measured from the center of the lens. Also, dx and dy are the x and y displacements of the deflected image. In (11)

$$n = \left(\frac{n_0^2 + n_e^2}{2}\right)^{\frac{1}{2}}, \tag{12}$$

$$n_r^2 = 2n_0^2 n_e^2/(n_0^2 + n_e^2) \tag{13}$$

and for the orientation ψ of the uniaxial crystal which gives the largest deflection (see Appendix A)

$$\cot \psi = 2n_e n_0/(n_0^2 - n_e^2). \tag{14}$$

The refractive indices of calcite at 589 mμ are[2]

$$n_0 = 1.65803, \qquad n_e = 1.4864 \tag{15}$$

hence from (11), (12), (13), (14) and (15)

$$\Delta Z = 0.36\, dx + 0.36\, dy. \tag{16}$$

We should note that since this is indeed the equation of a plane, we have merely to tilt any target to be used at the appropriate angle to get a sharp image at each of the possible positions of the spot.

IV. OPTICAL MODULATORS

Two possible ways of achieving 90° rotation of the plane of polarization have been investigated. A high specific rotation is possible in yttrium iron garnet in a magnetic field due to the Faraday effect. The necessary fields are quite high, however, for 90° rotation in a crystal of reasonable thickness; moreover, YIG has high optical attenuation at the wavelengths for which the specific rotation is greatest.[3]

Due to the linear Pockels effect, a ZO plate of KH_2PO_4 can cause 90° rotation with the half-wave voltage applied in the z direction if the induced principal axes are at 45° to the x and y directions.[4] Since KDP is a ferroelectric, the half-wave voltage is directly proportional to the temperature above the Curie temperature.

Fig. 7 defines the dimensions and orientation of a KDP crystal to be considered for a modulator.

The clamped dielectric constant, loss tangent, and half-wave voltage in the z direction are,[5,6]

$$\epsilon \approx \frac{2.27 \times 10^3 + 4.7T}{T - 119} \tag{17}$$

$$\tan \delta \approx \frac{8.42 \times 10^{-1}}{T - 119} \tag{18}$$

$$V_{\lambda/2} \approx 42(T - 119) \text{ volts.} \tag{19}$$

The stored energy and the dissipated energy per cycle are

$$W_s = \tfrac{1}{2} C V_{\lambda/2}^{\,2} \tag{20}$$

$$\approx \{1.77 + 3.67 \times 10^{-3} T\} \, (T - 119) \times 10^{-7} (ab/c) \text{ joules} \tag{21}$$

$$W_{\text{dis}} = 2\pi W_s \tan \delta \tag{22}$$

$$\approx \{0.938 + 1.94 \times 10^{-3} T\} \times 10^{-6} (ab/c) \text{ joules} \tag{23}$$

where the dimensions a, b, and c are the dimensions of the crystal in the x, y, and z directions, respectively, and are understood to be in centimeters.

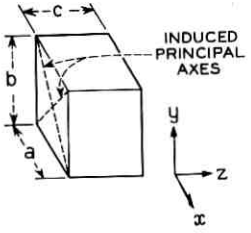


Fig. 7 — Dimensions of KH_2PO_4 modulator.

In the section on diffraction, we noticed that the center-to-center spacing of the spots would have to be greater than about one mil. If there are eight binary units in the x-deflection bank and eight in the y bank, then a square array of 65,536 positions results, 256 on a side. If the incremental spacing is one mil, then with some loss in intensity to the spots on the edge of the pattern we could choose the diameter of the lens opening, and hence a and b, to be 1 cm.

For this relatively small number of positions, and for a KDP modulator as thick as 1 cm, the dissipated energy per cycle will be on the order of one microjoule. To reduce the stored energy per cycle to this value, we would have to cool the modulator to about 6° above its Curie temperature. Thus a KDP modulator has a severe power limitation for nominal system capacities, if operation above 1 mc is desired.

V. EXPERIMENTAL RESULTS

In order to examine the optical limitations of digital light deflection, a four-unit system has been constructed. Fig. 8 is a pictorial diagram and

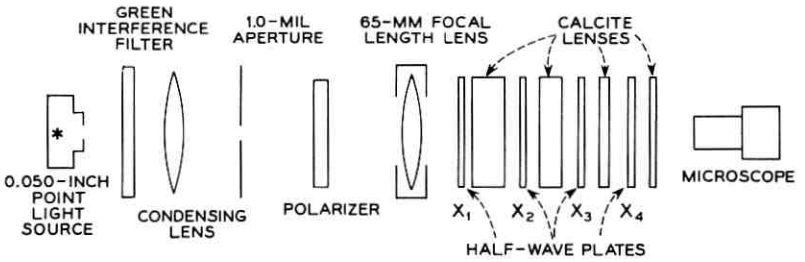


Fig. 8 — Pictorial diagram of four-unit deflection system.

Fig. 9 is a photograph of the experiment. The essential elements are: a 50-mil point light source which illuminates a 1.0-mil diameter aperture, a 65-mm lens which focuses the aperture at unity magnification into the image plane, a four-unit deflection bank, and a microscope for viewing the real image of the spot in the image plane. The four-unit deflection bank consists of four rotators and four calcite disks. The rotators were constructed by aligning two disks of Bausch and Lomb quarter-wave plastic. The rotators were mounted so that they could be rotated through 45°. In one orientation the principal directions of the so-constructed half-wave plate were parallel to the two possible polarizations of the beam and there was no rotation. In the other orientation the principal direc-

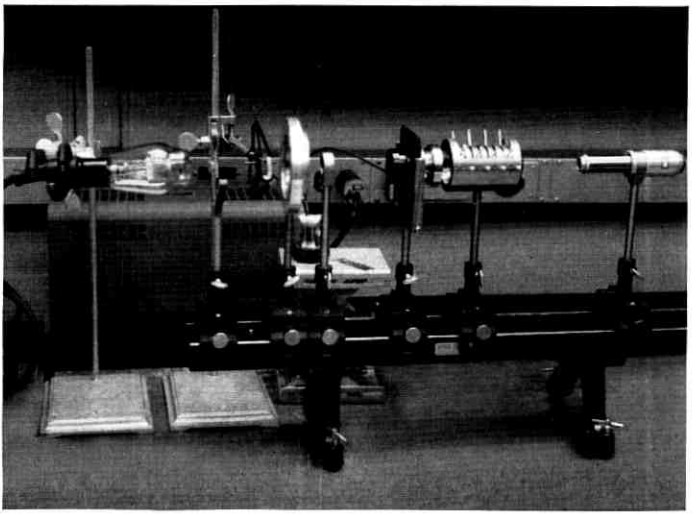


Fig. 9 — Photograph of four-unit experiment.

tions of the half-wave plate were at 45° to the possible polarizations of the beam, and the polarization of the beam was rotated by 90°.

The system was designed to give 16 positions of the 1.0-mil spot, with center-to-center distances of 2.5 mils. This design called for calcite disks of 0.0266, 0.0452, 0.0903, and 0.1806 in. thick. The disks obtained were 0.020, 0.045, 0.090, and 0.180 in. thick, the largest error occurring in the thinnest disk.

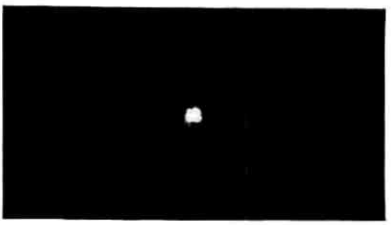
It was found that, for the maximum deflection of nominally 37.5 mils, the dispersion of the calcite in the deflection was about 1.0 mil over the visible spectrum. The half-wave plates were found to pass appreciable amounts of the unwanted polarization at the red end of the spectrum, and the dichroic polarizer used was known to be less efficient at the blue. For these reasons, the cleanest spot pattern was obtained with a green filter. It was further found that the relative intensity of the diffraction rings increased appreciably above an f number of 22. Figs. 10 and 11 are photographs of the spot pattern for various modulator settings with white and green light, respectively. The nomenclature $x_1x_2x_3x_4$ has the same meaning as that described above. These photographs were taken through the microscope at $f/22$ and $f/16$ of the lens and camera respectively, and the exposure times are indicated. The photographs show that the error in the thinnest calcite disk causes a noticeable "pairing" of the spots.

In order to provide a check on the diffraction analysis, we set the lens opening to $f/15$ and used a 666-mμ filter. The spot pattern was enlarged six times, and a 6-mil diameter aperture was centered at the position of the first spot in the new image plane. The light transmitted through this aperture was then detected by a photomultiplier. With this arrangement, x_4 was alternated, and the resulting change in photomultiplier current noted. These first two spots are "paired," as may be easily seen by inspecting Figs. 10 and 11. We measured their center-to-center spacing in the red light to be 1.04 ± 0.25 mils. For $f/15$, $\lambda = 694\ \mu$, and $d = 1.04$ mils, the predicted value of the crosstalk ratio is 12.8 db, whereas the measured value was 10.2 db.

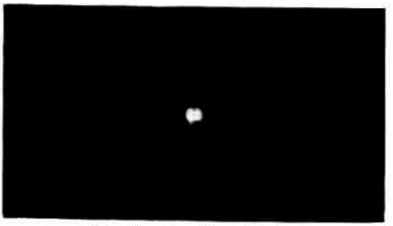
VI. APPLICATIONS

There are many functions that could be performed by digital light deflection. We shall briefly discuss three: a semipermanent memory, a pulse code modulation (PCM) decoder, and a digital delay line. In these applications, deflection banks similar to those described above can be used. The principal differences reside in the numbers of possible positions of the beam and the type of target used.

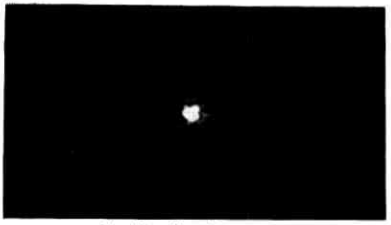
X_1, X_2, X_3, X_4 IN
INTERMEDIATE SETTINGS
20 SECONDS



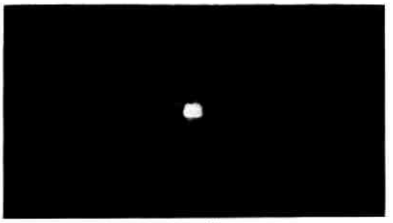
$X_1, X_2, X_3, X_4 = 0000$
5 SECONDS



$X_1, X_2, X_3, X_4 = 1010$
5 SECONDS



$X_1, X_2, X_3, X_4 = 0101$
5 SECONDS



$X_1, X_2, X_3, X_4 = 1000$
5 SECONDS

Fig. 10 — Microphotographs of spot patterns with no filter.

6.1 *The Semipermanent Memory*

Digital light deflection may be used to provide access to a target where information is stored in the form of a matrix of potential light paths. We place a photomultiplier tube, or some other photosensitive device, behind the target. When one of the possible combinations of inputs is applied to the modulators, the beam falls on the position of the target corresponding to that combination. If the target is transparent at that position, then we get an output from the photosensitive device which

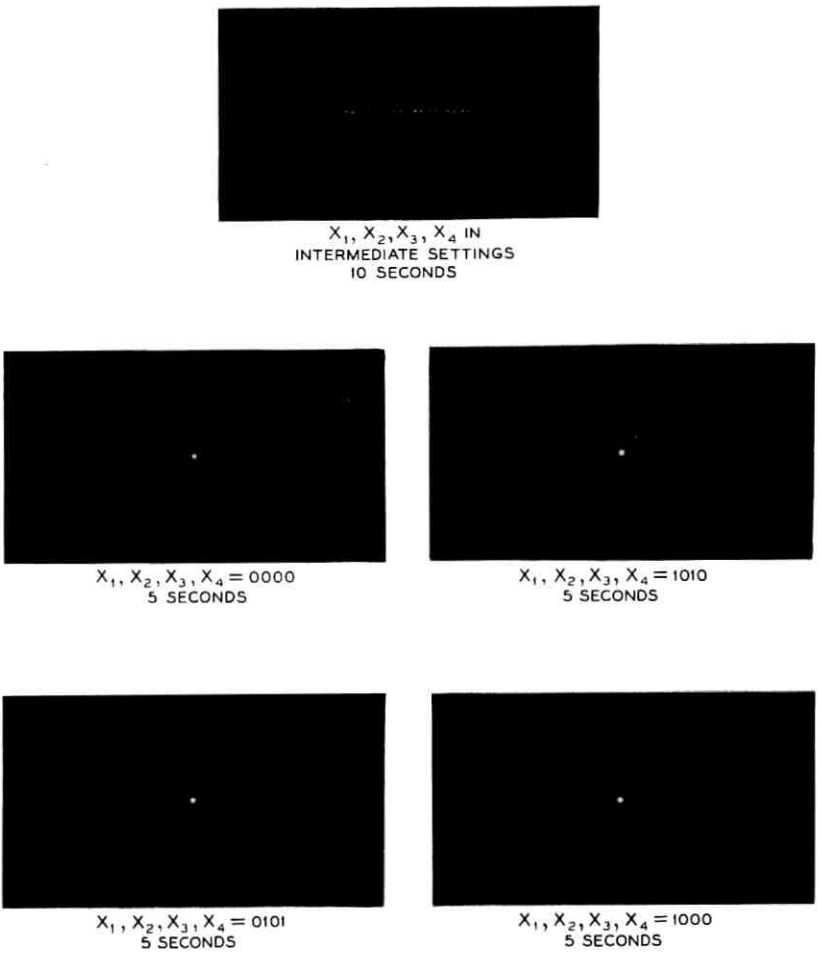


Fig. 11 — Microphotographs of spot patterns with green filter.

might represent a stored "one." If the target is opaque, the photomultiplier does not respond and a stored "zero" is inferred. A simple target could be a card with holes punched through it at some of the positions in the matrix. Since we could substitute cards with different information content, this is a semipermanent memory. In a memory application, we desire a very high storage capacity, and hence a matrix consisting of very many positions of the beam. We have shown in our section on diffraction that a high density of positions is possible; however, it then becomes difficult to position the target accurately, and this is an inherent problem

with the device. This application has the virtue that raw binary numbers serve as inputs to the storage element, and these are the type of signals most immediately available in computing machines.

6.2 *The PCM Decoder*

In the application of digital light deflection to the decoding of PCM signals, the target consists of an array of 2^n light paths, where n is the number of bits in a code group. The transparency of the light paths is quantized so that the amplitude of the output signal is different for each position. The target could be a positive photographic plate partially exposed at each of the accessible positions. The deflection banks could be conveniently used to expose the plate. The output of the photosensitive element would then be quantized PAM, and we recover the original analog signal by the usual method of passing the PAM signal through a low-pass filter. The target positioning problem is alleviated in this application because the capacity of the system is low.

6.3 *The Digital Delay Line*

Fig. 12 shows an x-deflection bank, quarter-wave plate, polarizer, delay line, analyzer, and photosensitive element arranged in line. The x-deflection bank enables us to displace the light beam in a digital manner along the x axis. Since the output light from the deflection bank is plane polarized, and the plane of polarization alternates from one position to the next, we use a quarter-wave plate and a polarizer to fix the plane of polarization that falls on the delay line. At time zero we generate

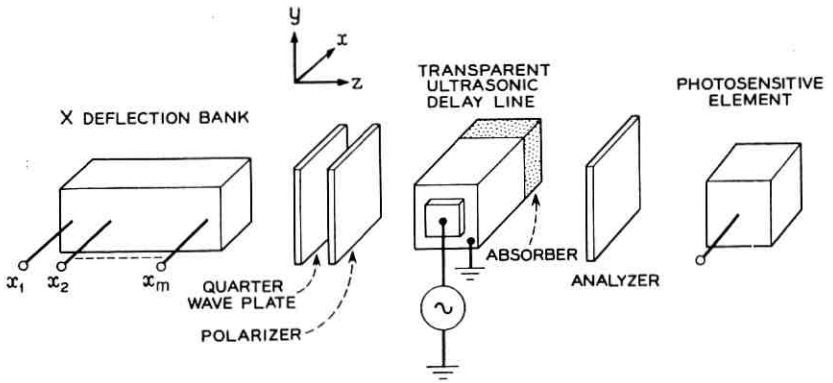


Fig. 12 — Application of digital light deflection to a digital delay line.

an ultrasonic wave in the delay line. In the absence of such a disturbance, the delay medium is transparent and isotropic, but as the disturbance passes through the position of the light beam the strain birefringence causes the light beam to be elliptically polarized. The analyzer is crossed with the polarizer and hence only transmits light to the photosensitive element when the ultrasonic wave passes through the position of the beam. Then we get an output from the photosensitive element. Thus the time at which an output results is proportional to the displacement of the beam, which in turn is determined by the state of the input variables. Thus digital light deflection can be used to make an electrically variable digital delay line. We should note that for a maximum delay line length of one inch the range of the device would be approximately 10 μsec, since sound travels in solids at roughly 0.1 in. per μsec. The time resolution would depend on the bandwidth of the delay line and the shape of the amplitude over the cross section of the light beam.

VII. CONCLUSIONS

A simple method of deflecting a light beam in a digital manner has been demonstrated. The method has the virtue of requiring only n two-level inputs for a set of 2^n possible outputs, and thus the inputs to the deflecting mechanisms are binary signals. Deflection rates above about 1 mc and with reasonable power requirements will depend on the development of optical modulator materials substantially more efficient than KDP.

VIII. ACKNOWLEDGMENTS

It is a pleasure to acknowledge the many helpful suggestions of and discussions with Messrs. J. E. Geusic, T. R. Meeker, J. H. Rowen, H. E. D. Scovil, and J. C. Skinner. I am also pleased to acknowledge the help of Miss B. T. Dale and Miss M. C. Grey, who programmed the diffraction analysis.

APPENDIX A

Fig. 13 shows the angles and directions in the uniaxial crystal with which we shall be concerned. The faces of the crystal are normal to the z axis; the optic axis of the crystal lies in an xz plane and at an angle θ with the z axis. The incident light is assumed to be a plane wave propagating in the z direction. In general the beam will be split into the ordinary ray, which passes through the crystal in a straight line, and the extraordinary ray, which is displaced.

Fig. 14 is a diagram showing the geometrical configuration of the wave normal, $\bar{s}$, and the optic axis, c. We have chosen to solve the more general

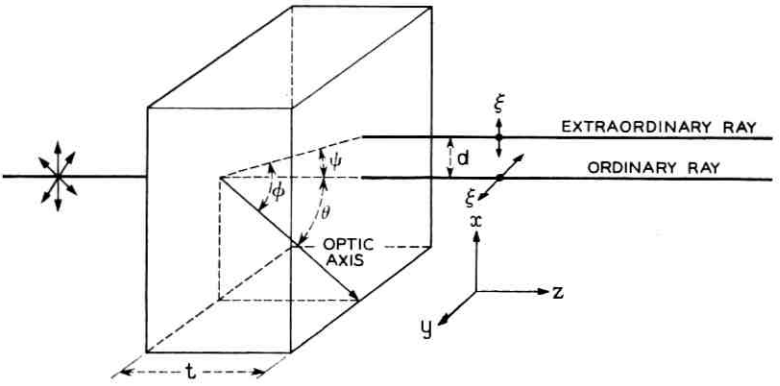


Fig. 13 — Angles and directions in uniaxial crystal.

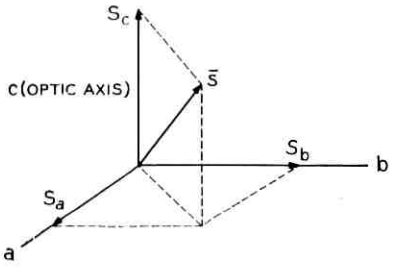


Fig. 14 — Wave normal and the crystallographic direction in the uniaxial media.

problem since the added complexity introduced is not great. In the uniaxial medium, we have[7]

$$V_a = V_b = V_0 = c/n_0 \tag{24}$$

and

$$V_c = V_e = c/n_e . \tag{25}$$

Making use of Fresnel's equation of wave normals,

$$\frac{s_a^2}{V_p - V_a^2} + \frac{s_b^2}{V_p^2 - V_b^2} + \frac{s_c^2}{V_p^2 - V_c^2} = 0, \tag{26}$$

where V_p is the normalized phase velocity, we find the two solutions

$$V_p^2 = V_0^2 \tag{27}$$

$$V_p^2 = V_0^2 \cos^2 \theta + V_e^2 \sin^2 \theta. \tag{28}$$

Born and Wolf[7] give the relations between the components of the ray vector and the components of the wave normal

$$t_k = \frac{s_k}{V_p V_r}\left(V_p^2 + \frac{g^2}{V_p^2 - V_k^2}\right) \tag{29}$$

where

$$g^2 \equiv V_p^2(V_r^2 - V_p^2) = \frac{1}{\left(\frac{s_a}{V_p^2 - V_a^2}\right)^2 + \left(\frac{s_b}{V_p^2 - V_b^2}\right)^2 + \left(\frac{s_c}{V_p^2 - V_c^2}\right)^2}. \tag{30}$$

Substituting in $V_p^2 = V_0^2$ we find that

$$g^2 = 0 \tag{31}$$

$$V_r^2 = V_p^2 = V_0^2 \tag{32}$$

$$t_k = s_k . \tag{33}$$

Thus, for the ordinary ray, the ray vector and wave normal coincide. If we substitute our second solution for the phase velocity into (29) and (30) we find

$$g^2 = (V_c^2 - V_0^2)^2 \sin^2\theta \cos^2\theta \tag{34}$$

$$V_r^2 = \frac{V_e^4 \sin^2\theta + V_0^4 \cos^2\theta}{V_e^2 \sin^2\theta + V_0^2 \cos^2\theta} \tag{35}$$

$$t_k = \frac{S_k}{V_p V_r}\left[\frac{V_e^2(V_e^2 - V_k^2)\sin^2\theta + V_0^2(V_0^2 - V_k^2)\cos^2\theta}{V_e^2 \sin^2\theta + V_0^2 \cos^2\theta - V_k^2}\right]. \tag{36}$$

This is as far as we care to carry the general case. For our more special consideration, let the b axis of Fig. 14 coincide with the y axis of Fig. 13; of course, the c axis is the optic axis, and the wave normal $\bar{s}$ points along the z axis. Thus $\bar{s}$ lies in the ac plane so that

$$s_a = \sin\theta, \qquad s_b = 0, \qquad s_c = \cos\theta. \tag{37}$$

If φ is the angle between the ray vector and the c direction, then by use of (36), we find

$$\tan\varphi = (V_e^2/V_0^2)\tan\theta = (n_0^2/n_e^2)\tan\theta. \tag{38}$$

Then the tangent of the angle between the ray vector and the z axis will be

$$\tan\psi = \tan(\varphi - \theta) = \frac{\tan\varphi - \tan\theta}{1 + \tan\varphi\tan\theta} \tag{39}$$

$$= \frac{\left(\frac{n_0^2}{n_e^2} - 1\right) \tan \theta}{1 + \frac{n_0^2}{n_e^2} \tan^2 \theta}. \tag{40}$$

Also, (35) becomes

$$V_r^2 = \frac{1}{n_r^2} = \frac{\frac{1}{n_e^4} \sin^2 \theta + \frac{1}{n_0^4} \cos^2 \theta}{\frac{1}{n_e^2} \sin^2 \theta + \frac{1}{n_0^2} \cos^2 \theta}, \tag{41}$$

hence

$$n_r^2 = \frac{1 + \frac{n_0^2}{n_e^2} \tan^2 \theta}{\frac{1}{n_0^2} + \frac{n_0^2}{n_e^4} \tan^2 \theta}, \tag{42}$$

where n_r is the refractive index for the extraordinary ray. By setting the first derivative of $\tan \psi$ with respect to $\tan \theta$ equal to zero, we may find the orientation which gives the greatest displacement of the extraordinary ray. This results in

$$\tan \theta = \frac{n_e}{n_0} \tag{43}$$

$$\tan \psi = \frac{n_0^2 - n_e^2}{2 n_e n_0} \tag{44}$$

$$n_r = \sqrt{\frac{2 n_0^2 n_e^2}{n_0^2 + n_e^2}}. \tag{45}$$

If t is the dimension of the uniaxial crystal in the z direction, then the displacement, d, of the extraordinary ray will be

$$d = \left(\frac{n_0^2 - n_e^2}{2 n_e n_0}\right) t \tag{46}$$

if the crystal has the optimum orientation.

We may note that in the negative uniaxial crystal, where $n_0 > n_e$, the extraordinary ray propagates away from the optic axis, whereas the situation is reversed in the positive crystal.

APPENDIX B

The light incident on the aperture, of radius a, is assumed to be a plane wave with wave number $k = 2\pi/\lambda$ and amplitude A. The distance S

from an element of area in the aperture to a point, P, on a sphere of radius R is expanded in a Taylor series expansion about the radial distance, ρ, from the center of the aperture (and sphere) to the element of area in the aperture.

$$S = R - \rho \sin \theta \cos \Phi \cdots . \tag{47}$$

Here θ is the angle between the radius vector to the element of area on the sphere and the cylindrical axis, Z; and Φ is the angle between the plane containing R and the cylindrical axis, and the radius vector to the element of area in the aperture. Fig. 15 shows these elements in their

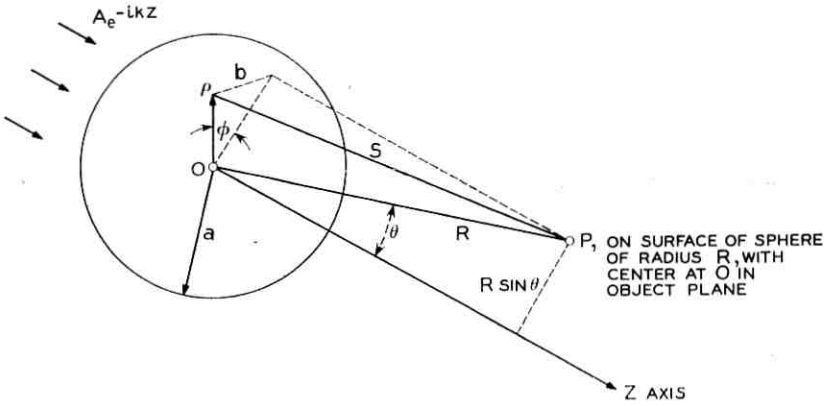


Fig. 15 — Angles and dimensions used to determine $E(\theta)$.

proper relation to one another. If the incident light has an electric field vector of the form

$$E = Ae^{-ikz} \tag{48}$$

then the electric field intensity on the sphere as a function of θ is, for small θ

$$E(\theta) = \frac{k}{2\pi} \int_0^{2\pi} \int_0^a \frac{A \exp[-ik(R - \rho \sin \theta \cos \Phi)]}{R - \rho \sin \theta \cos \Phi} \rho \, d\rho \, d\Phi. \tag{49}$$

Since $\rho \sin \theta \cos \Phi \ll R$, only the phase contribution is significant.

$$E(\theta) = \frac{kAe^{-ikR}}{2\pi R} \int_0^{2\pi} \int_0^{\rho} \rho \exp (ik\rho \sin \theta \cos \Phi) \, d\rho \, d\Phi. \tag{50}$$

Equation (50) is a standard integral with the value

$$E(\theta) = \frac{ka^2 A e^{-ikR}}{R} \left(\frac{J_1(ka \sin \theta)}{ka \sin \theta} \right). \tag{51}$$

Since the case of unity magnification is to be considered, it is assumed that the only significant function of the lens would be to cause the image and object spaces to be mirror images of each other if there were no stop at the lens. Thus, we have a section of a spherical wave to consider with a known phase and amplitude distribution. Fig. 16 shows the angles and distances under consideration. The angle θ retains its original meaning and Φ is once again a dummy variable; ρ now is the distance from the

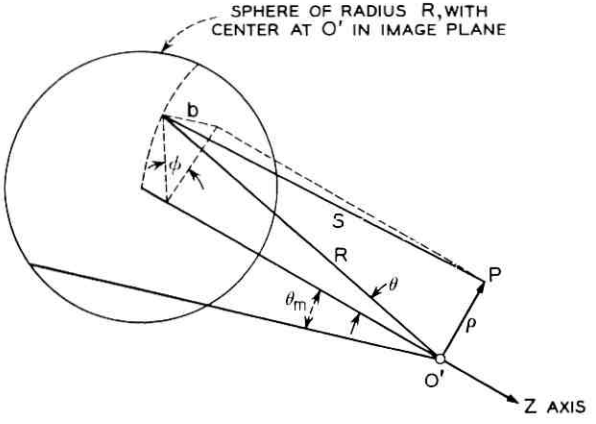


Fig. 16 — Angles and dimensions used to determine $E(\rho)$.

center of the image to the point in the image under consideration. Here S is the distance from an element of area on the sphere to the point, P, in the image plane, and is expanded about ρ.

$$S = R - \rho \sin \theta \cos \Phi + \cdots . \tag{52}$$

Hence the intensity at ρ is

$$E(\rho) = \frac{k}{2\pi} \int_0^{\theta_m} \int_0^{2\pi} \frac{E(\theta) \exp\left[-ik(R - \rho \sin \theta \cos \Phi)\right]}{R - \rho \sin \theta \cos \Phi} R \sin \theta \, d\theta \, d\Phi. \tag{53}$$

The approximation is once again made that $\rho \sin \theta \cos \Phi \ll R$:

$$E(\rho) = \frac{kRe^{-ikR}}{2\pi} \int_0^{\theta_m} \int_0^{2\pi} E(\theta) \exp\left(ik\rho \sin \theta \cos \Phi\right) \sin \theta \, d\theta \, d\Phi \tag{54}$$

and, putting in the previously determined expression for $E(\theta)$,

$$E(\rho) = (ka)^2 A e^{-i2kR} \int_0^{\theta_m} \int_0^{2\pi} \frac{J_1(ka \sin \theta)}{ka} \cdot \exp\,(ik\rho \sin \theta \cos \Phi)\, d\theta\, d\Phi \tag{55}$$

and integrating first over Φ,

$$E(\rho) = (ka) A e^{-i2kR} \int_0^{\theta_m} J_1(ka \sin \theta) J_0(k\rho \sin \theta)\, d\theta. \tag{56}$$

The integral in (56) is not available analytically; hence we resort to numerical procedures. The f number of the system and θ_m are intimately related,

$$\tan \theta_m = \frac{1}{4F} \tag{57}$$

and for small angles,

$$\theta_m \approx \frac{1}{4F}. \tag{58}$$

Since we wish to evaluate the crosstalk ratio, it is also necessary to integrate $[E(\rho)]^2$ over a circle of radius a located in the image plane. Fig. 17 defines the center-to-center separation as d and the local cylindrical coordinates ρ' and φ. We find by the usual geometrical considerations that

$$\rho = [d^2 - 2\, d\rho' \cos \varphi + \rho'^2]^{\frac{1}{2}}. \tag{59}$$

However, no approximations are in order here, as ρ' and d are of the same order of magnitude. Therefore the integral

$$P(d) = \int_0^{2\pi} \int_0^{a} \rho' E^2([d^2 - 2d\rho' \cos \varphi + \rho'^2]^{\frac{1}{2}})\, d\rho'\, d\varphi \tag{60}$$

results in the power falling on the circle of radius a and being displaced from the center of the image by distance d. The crosstalk ratio, C, will

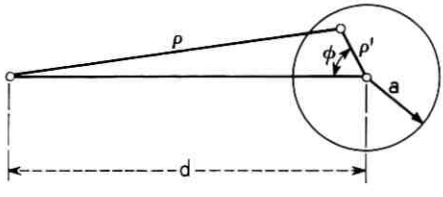


Fig. 17 — Circle of integration for determining crosstalk ratio and loss.

be defined as

$$C = 10 \log_{10} \left(\frac{P(0)}{P(d)} \right) \text{db}. \tag{61}$$

Also, we are in a position to evaluate the loss through the system

$$L = 10 \log_{10} \left(\frac{\pi a^2 A^2}{P(0)} \right) \text{db}, \tag{62}$$

that is, the db ratio of the power admitted to the system by the aperture to the power falling on a circle of radius a centered on the image. The amount of light lost from that emitted by the source depends strongly on the type of source and focusing at the aperture. If the source is focused on the lens system in the absence of the aperture, then our equations are still good approximations, although such illumination would not be plane.

REFERENCES

1. Caldwell, S. H., *Switching Circuits and Logical Design*, Wiley, New York, 1959, Chap. 7.
2. Sears, F. W., *Optics*, Addison-Wesley, Boston, 1949, p. 181.
3. Dillon, J. F., Jr., Optical Properties of Several Ferrimagnetic Garnets, J. Appl. Phys., **29**, Mar., 1958, p. 539.
4. *American Institute of Physics Handbook*, ed. Gray, D. E., McGraw-Hill, New York, 1957, pp. 6-94 to 6-97.
5. Zwicker, B. and Scherrer, P., Electro-Optical Properties of the Rochelle Salt Electrical Crystals, KH_2PO_4 and KD_2PO_4 , Helvetica Physica Acta, **17**, 1944, pp. 346–373.
6. Kaminow, I. P., and Harding, G. O., Complex Dielectric Constant of KH_2PO_4 at 9.2 Gc/sec, Phys. Rev., **129**, Feb., 1963, pp. 1562–66.
7. Born, M. and Wolf, E., *Principles of Optics*, Pergamon Press, New York, 1959, Chap. 14.

Perturbation Methods for Satellite Orbits

By F. T. GEYLING

(Manuscript received October 14, 1963)

The literature in astrodynamics abounds with perturbation techniques for satellite orbits. Various formulations have been generated in terms of orbit elements, the satellite position and velocity vectors, or combinations thereof. The computational effectiveness of any perturbation scheme depends largely on the definitions used for the dynamic state variables. Some methods are aimed at long-range predictions and orbit lifetime studies, others at short-range predictions for guidance. This paper may serve as an introduction to this field for the nonspecialist, in that it reviews the classical variation-of-parameters technique and discusses several engineering analyses that were generated in the post-Sputnik era. It also points to some connections between these relatively simple approaches and more elaborate methods of celestial mechanics. Thus it may contribute toward a comparison of several "professional" approaches whose relative merits are often debated among experts.

I. INTRODUCTION

This paper is a discussion of various perturbation techniques for satellite orbits which were investigated by the author and his colleagues during the past few years. The effort began with a tutorial "orbit seminar" several years ago and it seemed appropriate to collect some of this material here as a companion paper for R. B. Blackman's "Methods of Orbit Refinement."

It is a symptom of our times that aerospace engineers are taking a new look at the established methods of dynamical astronomy. The orbital geometries and vehicle characteristics encountered with artificial celestial bodies often require departures from the formulations of classical astronomy and, in fact, have stimulated several new (or at least independent) approaches during the post-Sputnik era. The number of publications in this time has been formidable, and in many discussions the names attached to various formulations serve as passwords for the ideas they represent. The uninitiated find themselves at a loss

concerning the methods that stand behind these names, their degree of originality, and their relations with each other.

In view of this situation the following article is addressed to two kinds of readers:

(*i*) The newcomers in the field of orbital mechanics who seek a tutorial survey and an introduction to some of the literature. A bare minimum of definitions is given for their benefit; a discussion of basic order-of-magnitude relations and certain intuitive notions which would strengthen the beginner's grasp of the physical problem had to be omitted for lack of space but can be found in the literature.[1,2]

(*ii*) The specialists in orbital mechanics who have not had occasion to correlate some of the better-known contributions in the literature and who may find this work a step in that direction. Typical issues in such comparisons are the choice of coordinates, the accuracy and elegance achieved by various transformations of the variables, and the precision obtainable from series expansions of the solution in terms of various small parameters.

The simultaneous need for conciseness of presentation and discussion of certain analytic detail presents somewhat of a dilemma. As a compromise, much of the development between the explicitly quoted results is covered in a descriptive way and the reader is referred to the literature for all standard derivations.[1-4] Most of our discussion concerns orbits of moderate eccentricity, which are representative of satellite missions. However, in many places an extension to the highly eccentric orbits of space probes follows readily.

We begin by devoting Section II to a statement of the fundamental equations of motion, the definition of so-called orbit parameters, and a description of various disturbing functions. Section III summarizes the classical treatment of satellite perturbations as gradual changes of the orbit parameters. (From a general point of view, this formulation, due to Lagrange, is derivable from the canonical systems governing the satellite problem.) It is hoped that this covers a sufficient amount of standard material to introduce the concepts and the parlance of orbital mechanics.

In Section IV we examine several perturbation methods for aerospace applications which are based on variously defined spherical and moving Cartesian coordinates. This includes the well-known contributions by Blitzer et al., Anthony et al., and Roberson. They could serve as an introduction to the discussion of more elaborate formulations by King-Hele et al. and Brenner et al. In Section V we treat one more formulation in this general category which was specifically designed for guidance studies.

A logical continuation of this paper would cover the methods of Breakwell et al. and Diliberto, Kyner's averaging technique, and the one suggested by Struble. Ultimately the hierarchy of perturbation methods leads to the Hamilton-Jacobi techniques expounded by Brouwer, Garfinkel, and Vinti. These represent a very popular approach to higher-order perturbations and the coupling between simultaneous disturbances of satellite orbits.

II. PRELIMINARIES AND DEFINITIONS

We remember that the underlying phenomenon of undisturbed satellite motion (in a central force field, i.e. around a spherically symmetric body) is Newton's law of inverse square attraction. In a Cartesian coordinate system this spells out to be

$$\ddot{x} = -(Gm_e x/r^3) \tag{1}$$

$$\ddot{y} = -(Gm_e y/r^3) \tag{2}$$

$$\ddot{z} = -(Gm_e z/r^3) \tag{3}$$

where G is the universal gravitational constant, m_e the central mass, and we shall usually take $Gm_e = k$ for brevity. m_e shall be the mass of the earth in all our discussions. [Strictly speaking, formulas (1) to (3) should show the sum of m_e and the satellite mass instead of just m_e.] r is the distance from the origin, and dots indicate time derivatives. The x-y plane is usually taken to coincide with the equator, while the positive x axis points to the vernal equinox. The above equations simply state that each acceleration component is due to the corresponding component of the gravitational attraction — the minus signs indicating a direction toward the origin. The solutions of (1) to (3) are the well-known Kepler orbits — ellipses, parabolas, and hyperbolas.

Such orbits can be conveniently described by a set of six parameters that give the plane of the motion, the shape of the orbit and its orientation in that plane, and the timing of the satellite motion along this path. These quantities may be considered the constants of integration for a solution of (1) to (3). A standard set of such orbit elements for elliptic motion is illustrated in Fig. 1. They are:

a, the semimajor axis,
e, the eccentricity,
ω, the argument of perigee, (4)
i, the inclination,
Ω, the nodal angle, and
τ, the time of perigee passage.

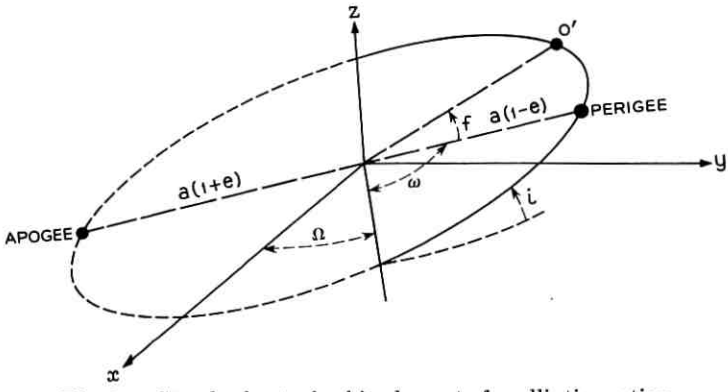


Fig. 1 — Standard set of orbit elements for elliptic motion.

The last quantity establishes a time scale for the entire motion in that it serves as "epoch" and fixes one particular passage through perigee. If the satellite has swept out the angle f since that passage, the elapsed time is given by

$$t - \tau = (a^3/k)^{\frac{1}{2}} \left\{ 2 \tan^{-1} \left[\left(\frac{1-e}{1+e} \right)^{\frac{1}{2}} \tan \frac{f}{2} \right] - e\,(1-e^2)^{\frac{1}{2}} \frac{\sin f}{1 + e \cos f} \right\}_0^f \tag{5}$$

where t is the time pertaining to the position $0'$. f is known as the true anomaly and (5) holds for all values of this angle. If we set $f = 2\pi$ this corresponds of course to a full revolution around the orbit, and the elapsed time interval is

$$T = 2\pi\,(a^3/k)^{\frac{1}{2}}, \tag{6}$$

which is known as the anomalistic period. The instantaneous position $0'$ can also be defined in terms of other angles, the so-called eccentric anomaly E or the mean anomaly M, which will be defined later. They can be related to time in similar ways.

The parameters in (4) represent a typical set of orbit elements. The position and velocity vectors at some epoch are an alternative suggested by (1)–(3). Most astrodynamical theories use variations and combinations of all these, but from the general standpoint of analytical dynamics most sets of six parameters (if they are independent of each other) may be regarded as sets of canonic variables. Before we proceed to detailed formulations we examine briefly the various physical disturbances which cause these parameters to change in time.

2.1 *The Effect of Extraterrestrial Gravitation*

If we consider the attractions from masses other than the earth we speak of "extraterrestrial" gravitation. In the presence of a disturbing body P, (1) becomes

$$\ddot{x} = -\frac{Gm_e x}{r^3} - Gm_p\left(\frac{x - x_p}{r_{ip}^3} + \frac{x_p}{r_p^3}\right) \tag{7}$$

where

m_p = the mass of P,

$r_{ip} = [(x - x_p)^2 + (y - y_p)^2 + (z - z_p)^2]^{\frac{1}{2}}$,

the distance from the satellite to P, and

$$r_p = [x_p^2 + y_p^2 + z_p^2]^{\frac{1}{2}},$$

the geocentric distance of P.

The corresponding equations for $\ddot{y}$ and $\ddot{z}$ are obvious. Now it is often convenient in analytical dynamics to express the disturbing terms in $\ddot{x}, \ddot{y}, \ddot{z}$ as partial derivatives $(\partial\tilde{R}/\partial x)$, $(\partial\tilde{R}/\partial y)$, $(\partial\tilde{R}/\partial z)$ of a disturbing function $\tilde{R}$. For the present case we would have

$$\tilde{R} = Gm_p\left[\frac{1}{r_{ip}} - \frac{xx_p + yy_p + zz_p}{r_p^3}\right], \tag{8}$$

as can be easily verified.

We see that the ratio of the second term in (7) to the first is of the order $m_p r^3/m_e r_p^3 = \kappa$. Typical values for κ in the planetary system are equal to or less than 10^{-5}. Its smallness is vital to the entire rationale of a perturbation technique.

2.2 *The Effect of the Earth's Oblateness*

The potential field for the nonspherical earth can be represented to various levels of accuracy by a series of spherical harmonics. If we restrict ourselves to terms with rotational symmetry about the polar axis, we obtain the following disturbing function

$$\tilde{R} = \frac{Gm_e R^2}{r^3}\left[\frac{J}{3}(1 - 3\sin^2\varphi) + \frac{H}{5}\frac{R}{r}(3\sin\varphi - 5\sin^3\varphi) + \cdots\right] \tag{9}$$

where

R = the earth's equatorial radius,
φ = the geocentric latitude of the satellite,
$J = 1.6239 \times 10^{-3}$, and
$H = 6.04 \times 10^{-6}$.

This two-term series is sufficiently accurate for our purposes.

2.3 *The Effect of Atmospheric Drag*

The resistance encountered by a satellite from the atmosphere is a subject of considerable uncertainty and continued research. For one thing, the density of atmospheric gases as a function of geographic location, altitude and time is not well known; moreover, the laws of interaction between a satellite and this rarefied medium are incompletely understood. Doubts exist as to the transition from a continuum behavior of the atmosphere to the gas-kinetic regime and the extent to which electric interactions play a role. Nevertheless, the classical drag law yields useful results in many cases and we shall concentrate on it. We let

$$F_D = -(C_D A/2)\rho v_a^2 \tag{10}$$

where

F_D = the total drag force on the satellite
A = the frontal area of the satellite
C_D = the drag coefficient
ρ = the atmospheric density
v_a = the satellite velocity relative to the atmosphere.

The monotonic decay of ρ with altitude covers approximately ten orders of magnitude within typical satellite altitudes and remains the subject of extensive study. The relative velocity v_a is simply the difference between the satellite's inertial velocity vector $\mathbf{v}(\dot{x},\dot{y},\dot{z})$ and the rotational velocity of the atmosphere $V = r\sigma \cos \varphi$, where σ can usually be taken as the earth's angular motion (the diurnal rate) and V always points due east. One is frequently justified in employing an approximate vector representation of (10):

$$\mathbf{F}_D \approx (C_D A/2)\ \rho\ v(\mathbf{V} - \mathbf{v}). \tag{11}$$

For typical earth satellites this force is at least two orders of magnitude smaller than the central attraction, i.e., $\kappa \leqq 10^{-2}$.

2.4 *The Effect of Radiation Pressure*

As the reader knows, solar illumination exerts some pressure on every satellite. The magnitude of this force depends on the reflectivity and geometry of the satellite and, strictly speaking, on the distance from the satellite to the sun. It frequently suffices to represent this disturbance as a constant force β per unit mass and to note that it is many orders of magnitude smaller than the central gravity force.

III. PERTURBATIONS IN THE ELEMENTS

The six orbit elements [see (4)] were constants for the case of central inverse-square attraction. However, if any additional forces act on the satellite these parameters will be subject to change. To emphasize their time dependence we might write them as $a(t)$, $e(t)$ etc. In fact, their numerical values at any time t describe the ellipse the orbiting body would follow if all perturbations vanished as of that instant. This trajectory is obviously tangent to the actual flight path at t and is known as the "osculating" orbit. The relation between the satellite position in the osculating orbit and in the x, y, z frame follows from the geometry of conic section trajectories:

$$\begin{aligned} x &= \frac{a(1-e^2)}{1+e\cos f}\,[l_1\cos f + l_2\sin f] \\ y &= \frac{a(1-e^2)}{1+e\cos f}\,[m_1\cos f + m_2\sin f] \\ z &= \frac{a(1-e^2)}{1+e\cos f}\,[n_1\cos f + n_2\sin f] \end{aligned} \tag{12}$$

where $l_1, l_2, \cdots n_2$ are functions of i, ω, and Ω. Hence x, y, z are representable as functions of a, e, i, ω, Ω, f. [As mentioned with (5), we could also work in terms of the independent variable E or M instead of f.] However, the complete definition of the osculating orbit also entails that

$$\begin{aligned} \dot{x} &= \frac{\partial x}{\partial f}\,(a,e,i,\omega,\Omega,f)\,\frac{df}{dt} \\ \dot{y} &= \frac{\partial y}{\partial f}\,(a,e,i,\omega,\Omega,f)\,\frac{df}{dt} \\ \dot{z} &= \frac{\partial z}{\partial f}\,(a,e,i,\omega,\Omega,f)\,\frac{df}{dt} \end{aligned} \tag{13}$$

where a, e, i, ω, Ω, τ are treated as constants. In other words, the velocity as well as the position in the osculating orbit are representative of the actual motion. This is the full extent of the "condition of osculation."

A large part of classical celestial mechanics has been based on the concept of osculating orbits, and during the post-Sputnik era Lagrange's classical treatment of the perturbations in these elements has been exploited *ad ultimo*. Its inclusion in this article is justified mainly by the need for completeness in an introductory survey such as this. It also serves as a point of reference for the nonclassical "perturbations in the coordinates" in the next section and for the Hamilton-Jacobi techniques frequently used by astronomers.

In essence, the Lagrange method consists of transforming the basic equations

$$\begin{aligned} \ddot{x} &= -\frac{Gm_e x}{r^3} + \frac{\partial \tilde{R}}{\partial x} \\ \ddot{y} &= -\frac{Gm_e y}{r^2} + \frac{\partial \tilde{R}}{\partial y} \\ \ddot{z} &= -\frac{Gm_e z}{r^3} + \frac{\partial \tilde{R}}{\partial z} \end{aligned} \tag{14}$$

to six first-order equations in the orbit elements and approximating their solutions by quadratures. Remembering that these parameters represented the constants of integration for the Kepler problem, (1)–(3), we note that the transition to $a(t)$, $e(t)$, $\cdots$ is nothing but Lagrange's "variation of constants" designed to accommodate the terms $[\partial \tilde{R}/\partial(x,y,z)]$ in (14). In the process of transforming (14) by means of (12) we avoid the occurrence of second derivatives of the orbit elements by demanding that

$$\frac{dx}{dt} = \frac{\partial x}{\partial t}, \text{ i.e.,}$$

$$\frac{\partial x}{\partial a}\dot{a} + \frac{\partial x}{\partial e}\dot{e} + \frac{\partial x}{\partial i}\frac{\partial i}{\partial t} + \frac{\partial x}{\partial \omega}\dot{\omega} + \frac{\partial x}{\partial \Omega}\dot{\Omega} + \frac{\partial x}{\partial \tau}\dot{\tau} = 0 \tag{15}$$

etc. This of course results in the reduction of the system of three second-order equations (14) to six first-order equations. Equations (15) are simply a restatement of (13), the condition of osculation.

In principle, (14) could be transformed by a straightforward substitution of (12). In order to simplify the algebra, however, one may

symmetrize these equations to a form where all the labor reduces to the evaluation of such quantities as

$$\left(\frac{\partial x}{\partial \alpha_i}\frac{\partial \dot{x}}{\partial \alpha_j} - \frac{\partial x}{\partial \alpha_j}\frac{\partial \dot{x}}{\partial \alpha_i}\right) + \left(\frac{\partial y}{\partial \alpha_i}\frac{\partial \dot{y}}{\partial \alpha_j} - \frac{\partial y}{\partial \alpha_j}\frac{\partial \dot{y}}{\partial \alpha_i}\right) + \left(\frac{\partial z}{\partial \alpha_i}\frac{\partial \dot{z}}{\partial \alpha_j} - \frac{\partial z}{\partial x_j}\frac{\partial \dot{z}}{\partial \alpha_i}\right) = [\alpha_i, \alpha_j], \tag{16}$$

with $i, j = 1 \cdots 6$.

The shorthand symbol that we have adopted for this expression is known as a Lagrange bracket, and α_i and α_j stand for any two of the orbit elements. These brackets have the properties

$$[\alpha_i, \alpha_i] = 0, \qquad [\alpha_i, \alpha_j] = -[\alpha_j, \alpha_i] \tag{17}$$

and

$$\frac{d}{dt}[a_i, a_j] = 0,$$

which make them useful devices in numerous manipulations of analytic dynamics. With their help the equations (14) become

$$\dot{a} = -\frac{2a^2}{k}\frac{\partial \tilde{R}}{\partial \tau} \tag{18}$$

$$\dot{e} = -\frac{a(1-e^2)}{ke}\frac{\partial \tilde{R}}{\partial \tau} - \frac{1}{e}\left(\frac{1-e^2}{ka}\right)^{\frac{1}{2}}\frac{\partial \tilde{R}}{\partial \omega} \tag{19}$$

$$\frac{di}{dt} = \frac{1}{[ka(1-e^2)]^{\frac{1}{2}}\sin i}\left[\cos i\,\frac{\partial \tilde{R}}{\partial \omega} - \frac{\partial \tilde{R}}{\partial \Omega}\right] \tag{20}$$

$$\dot{\Omega} = \frac{1}{[ka(1-e^2)]^{\frac{1}{2}}\sin i}\frac{\partial \tilde{R}}{\partial i} \tag{21}$$

$$\dot{\omega} = \left(\frac{1-e^2}{ka}\right)^{\frac{1}{2}}\frac{1}{e}\left[\frac{\partial \tilde{R}}{\partial e} - \frac{e\cot i}{1-e^2}\frac{\partial \tilde{R}}{\partial i}\right] \tag{22}$$

and a corresponding equation for $\dot{\tau}$ which will be discussed a little later. The five equations given here describe the changing geometry of the satellite orbit. All six equations together are known as Lagrange's planetary or "variational" equations.

In (18)–(22) we assumed that the perturbing forces were conservative, i.e., expressible as $(\partial \tilde{R}/\partial x)$, $(\partial \tilde{R}/\partial y)$, and $(\partial \tilde{R}/\partial z)$. In some situations, as for example in the case of drag, this is not so. Under these conditions it is convenient to represent the disturbing force by com-

ponents S, T, N which are in the radial direction, the direction of increasing true anomaly, and normal to the orbit plane, respectively. The derivation leading to (18)–(22) can be repeated with the appropriate modifications to yield a set of differential equations with S, T, N in the right-hand sides. For example

$$\dot{a} = 2\left[\frac{a^3}{k(1-e^2)}\right]^{\frac{1}{2}}[S\ e \sin f + T\ (1 + e\cos f)], \text{ etc.} \qquad (23)$$

These are known as Gauss's form of the planetary equations. We note that they contain the true anomaly as an independent variable. More will be said about this presently.

It is possible to show that the planetary equations, especially in the last form, lend themselves to an alternative derivation which appeals to intuition. It is based on the idea that any continuously acting perturbation may be interpreted as a sequence of infinitesimal impulses whose cumulative time response can be represented by a convolution integral. This approach leads to equations like (18)–(23) without the manipulations involving Lagrange brackets.[6]

Inspection of (18)–(22) shows that their nonlinear right-hand sides preclude an exact solution except for very special forms of $\tilde{R}$. Unfortunately, none of the perturbations encountered in nature fall into this category. One therefore resorts to a process of successive approximations.

Assuming that all disturbances represented by $\tilde{R}$ are small in relation to the central attraction (i.e.,

$$\left(\frac{\partial \tilde{R}}{\partial(x,y,z)}\right) \Big/ \frac{k(x,y,z)}{r^{(0)3}} = \kappa \ll 1,$$

as discussed in Section II) we consider the solution for the undisturbed motion, i.e. the Kepler problem, as a "zero-order" approximation to the actual case. Let its parameters be denoted $a^{(0)}$, $e^{(0)}$ $\cdots$. If we insert them into the right-hand sides of (18)–(22), these equations reduce to quadratures yielding a first approximation to the effects of $\tilde{R}$ on the orbit. These results are denoted $a^{(1)}$, $e^{(1)}$ $\cdots$ and known as "first-order" perturbations. We observe that $(a^{(1)}/a^{(0)})$, $(e^{(1)}/e^{(0)})$, $\cdots = 0(\kappa)$. In principle, this process can be repeated indefinitely by substituting $a^{(n-1)}$ $\cdots$ into the right-hand sides to integrate for $a^{(n)}$ $\cdots$. The limit is usually reached when the results for a, e $\cdots$ have settled or human endurance is exhausted. (The latter constraint may be eventually eliminated by computer routines for symbol manipulations.) The convergence of this

process to the exact solution has been established by Poincaré and is of fundamental interest to the mathematician. Suffice it here to say that the "smallness" of perturbations discussed in Section II should be such as to justify the iterative process.

When the right-hand sides of (18)–(22) are written out explicitly for any particular case, they tend to become awkward because a transcendental angle-time relation like (5) enters. Since the geometric description of a perturbation $\tilde{R}$ (or S,T,N) usually involves an angle like one of the anomalies very directly, it is convenient to use one of them as independent variable. The time relation which interconnects the anomalies for an osculating orbit (Kepler's equation) can be stated in terms of the eccentric anomaly E, the true anomaly, f and the mean anomaly M as follows:

$$\begin{aligned} E - e \sin E &= 2 \tan^{-1}\left[\left(\frac{1-e}{1+e}\right)^{\frac{1}{2}} \tan\frac{f}{2}\right] - \frac{e(1-e^2)^{\frac{1}{2}} \sin f}{1 + e\cos f} \\ &= (k/a^3)^{\frac{1}{2}}(t - \tau) = M. \end{aligned} \tag{24}$$

This may serve as a definition of E and M. The quantity $(k/a^3)^{\frac{1}{2}} \equiv n$ is referred to as the "mean angular rate."

If we work in terms of the true anomaly, we can write the left-hand sides of Lagrange's equations as

$$\dot{a} = \frac{da}{df}\frac{df}{dt}, \text{ etc.}$$

where an expression must now be found for $\dot{f}$. If we choose to consider f as the true anomaly in some osculating orbit valid at time t_0, then it can be related to time by (24) in terms of the unperturbed elements $a^{(0)} \equiv a_0$, $e^{(0)} \equiv e_0$ etc. We call it an "unperturbed" anomaly and designate it by $f^{(0)}$. From (24) one finds

$$\dot{f}^{(0)} = \left[\frac{k}{a_0^3(1-e_0^2)^3}\right]^{\frac{1}{2}} (1 + e_0 \cos f^{(0)})^2 \tag{25}$$

which transforms (18)–(22) to

$$\frac{da^{(1)}}{df^{(0)}} = -2\left[\frac{a_0^7(1-e_0^2)^3}{k^3}\right]^{\frac{1}{2}} (1 + e_0 \cos f^{(0)})^{-2} \frac{\partial \tilde{R}}{\partial \tau_0}, \text{ etc.} \tag{26}$$

When integrated, these expressions represent first-order perturbations in terms of the unperturbed anomaly, i.e., $a^{(1)}(f^{(0)})$, etc.

They suggest the following procedure for generating a first-order satellite ephemeris: if t_0 is the starting epoch we evaluate $a^{(1)}(f^{(0)})$ etc.

between the limits $f_0^{(0)}$ and $f_1^{(0)}$. The time t_1 at the upper limit follows from (24) in terms of a_0, $e_0 \cdots$. Using t_1 and $\tilde{a}_1 = a_0 + a^{(1)}(f_1^{(0)})$, etc., in (24), we find $\tilde{f}_1$, the true anomaly for the new osculating orbit. Changing the notation from $\tilde{a}_1$ to a_1, etc., and $\tilde{f}_1$ to $f_1^{(0)}$, we can now repeat the procedure for the next integration interval. The "updating" of orbit elements in the right-hand side of (26) amounts to a partial allowance for higher-order perturbations, while the recalculation of f at the beginning of each step represents essentially a first-order perturbation of the true anomaly.

Instead of doing the latter by discrete increments, we can work with a "perturbed" anomaly by differentiating (24) with proper allowance for the time dependence of a, e and τ. Using (25) one finds that

$$\begin{aligned}\frac{df}{df^{(0)}} = \frac{\dot{f}}{\dot{f}^{(0)}} = 1 &- \frac{2(1-e^2)^{\frac{1}{2}}a^2}{k(1+e\cos f)^2}\left(\frac{\partial \tilde{R}}{\partial a}\right) - \frac{a(1-e^2)^{\frac{1}{2}}}{ek(1+e\cos f)^2}\frac{\partial \tilde{R}}{\partial e} \\ &+ \frac{a^{\frac{3}{2}}(1-e^2)}{k^{\frac{1}{2}}(1+e\cos f)^3}\left[(1+\cos f)\tan\frac{f}{2}\right. \\ &\left. - (2e^2-1)\sin f - \frac{(1-e^2)e\sin f\cos f}{1+e\cos f}\right] \\ &\times\left[\frac{a(1-e^2)}{ke}\frac{\partial \tilde{R}}{\partial \tau} + \frac{1}{e}\left(\frac{1-e^2}{ak}\right)^{\frac{1}{2}}\frac{\partial \tilde{R}}{\partial \omega}\right]\end{aligned} \tag{27}$$

where $(\partial \tilde{R}/\partial a)$ means that $\tilde{R}$ is to be differentiated with respect to the semimajor axis wherever the latter appears explicitly but not when it is contained in n. This avoids the occurrence of a term with $(t - \tau)$. An expression analogous to (27) can be derived in terms of S,T,N; see for example Ref. 7, p. 4. Now, it is immaterial in a first-order approximation such as (27) whether we consider f or $f^{(0)}$ as the independent variable in the right-hand side. Let us assume the former and use the symbol $df/df^{(0)} = \nu(f)$. Then (26) becomes

$$\frac{da^{(1)}}{df} = \frac{-2}{\nu}\left[\frac{a_0^7(1-e_0^2)^3}{k^3}\right]^{\frac{1}{2}}(1+e_0\cos f)^{-2}\frac{\partial \tilde{R}}{\partial \tau_0}, \text{ etc.} \tag{28}$$

The integration procedure now runs between consecutive limits f_0, f_1, $\cdots$ f_j, with updating being required only in the orbit elements. The corresponding epochs t_j are of course computable by substituting a_j, $e_j \cdots f_j$ into (24). The relative advantages of integrating the perturbative equations in terms of $f^{(0)}$ or f depend on the problem at hand. As we shall see later, the choice between an unperturbed or perturbed independent variable is available in most perturbation methods.

Up to this point we have restricted our discussion to the first five

orbit parameters, which describe the geometry of the osculating orbit. Wherever τ appeared, as in (24), we assumed that it would be available from a suitable sixth equation. This parameter is needed to correlate the independent variable, such as f, with time.

An equation for τ, corresponding to (18)–(22), can be obtained by the process outlined before, which yields

$$\tau = \frac{2a^2}{k}\frac{\partial \tilde{R}}{\partial a} + \frac{a(1-e^2)}{ke}\frac{\partial \tilde{R}}{\partial e}. \tag{29}$$

Sometimes it is convenient to work with the slightly different parameter $\chi = -n\tau$. The differential equation for it reads

$$\dot{\chi} = -2(a/k)^{\frac{1}{2}}\frac{\partial \tilde{R}}{\partial a} - \frac{1-e^2}{e(ak)^{\frac{1}{2}}}\frac{\partial \tilde{R}}{\partial e}. \tag{30}$$

[A superficial comparison of these two equations gives the startling impression that (30) is obtained by multiplying (29) with $-n$, thus neglecting the $\dot{n}$ term that should appear. This term is really absorbed in the difference between $(\partial \tilde{R}/\partial a)_\tau$ and $(\partial \tilde{R}/\partial a)_\chi$, as implied by the two equations; i.e. partial derivatives of $\tilde{R}$ with respect to the semimajor axis, holding τ or χ constant as required.]

One could transform (29) and (30) to f (or E) as independent variable and pursue the quadrature as we did before. However, since $\partial \tilde{R}/\partial a$ involves $\partial f/\partial a$ (or $\partial E/\partial a$), we notice from (24) that this introduces the factors $\tan^{-1}\{[(1-e)/(1+e)]^{\frac{1}{2}}\tan(f/2)\}$ (or E) into the integrands for $\tau^{(1)}$ (or $\chi^{(1)}$). They can be quite awkward.

Several devices have been developed to circumvent this difficulty. According to one approach we transform (29) from τ to $M = n(t-\tau)$. The necessary compensating factors arise thereby which eliminate all aperiodic terms. Transforming the integrated equation back to τ, we have

$$\begin{aligned}\tilde{\tau}_1\tilde{n}_1 = \tau_0 n_0 + t_1(\tilde{n}_1 - n_0) + \left[\frac{a_0^3(1-e_0^2)^3}{k}\right]^{\frac{1}{2}} \int_{f_0^{(0)}}^{f_1^{(0)}} \\ (1+e_0\cos f^{(0)})^{-2}\left\{-\frac{3k^{\frac{1}{2}}}{2a_0^{\frac{5}{2}}}a_1^{(1)} + \frac{1-e_0^2}{e_0(ka_0)^{\frac{1}{2}}}\frac{\partial \tilde{R}}{\partial e_0}\right. \\ \left. + 2(a_0/k)^{\frac{1}{2}}\left(\frac{\partial \tilde{R}}{\partial a_0}\right)\right\} df^{(0)}\end{aligned} \tag{31}$$

where

$$\tilde{n}_1 = (k/\tilde{a}_1^3)^{\frac{1}{2}}, \qquad n_0 = (k/a_0^3)^{\frac{1}{2}}, \qquad \tilde{\tau}_1 = \tau_0 + \tau_1^{(1)}, \qquad \tilde{a}_1 = a_0 + a_1^{(1)},$$

and $(\partial \tilde{R}/\partial a_0)$ has the previously established meaning. Equation (31) can

be obtained in terms of the perturbed anomaly f if it is understood that $a_1^{(1)}$ in the quadrature and in $\tilde{a}_1$ is obtained by (28) and if the integrand of (31) is multiplied by $1/\nu$.

3.1 *Oblateness Effects*

We briefly illustrate some results from Lagrange's method. In the well-known example of oblateness perturbations, the first-order solutions for a, e, and i turn out to be entirely periodic and not very interesting.[8] The remaining elements, however, exhibit secular terms. Using only the J-term of (9) we find from (21), (22) and (31)

$$\tilde{\Omega}_1 = \Omega_0 - \frac{JR^2}{a_0^2(1-e_0^2)^2}(\cos i_0)(f_1 - f_0) + \text{periodic terms} \tag{32}$$

$$\begin{aligned}\tilde{\omega}_1 = \omega_0 &- (\cos i_0)(\tilde{\Omega}_1 - \Omega_0) + \frac{JR^2}{a_0^2(1-e_0^2)^2}\left(1 - \frac{3}{2}\sin^2 i_0\right) \\ &\cdot (f_1 - f_2) + \frac{JR^2}{a_0^2 e_0(1-e_0^2)^2} \times \text{p.t.}\end{aligned} \tag{33}$$

$$\begin{aligned}\tilde{\tau}_1 = \tau_0 &\frac{n_0}{\tilde{n}_1} - \frac{n_0}{\tilde{n}_1}\frac{JR^2}{a_0^2(1-e_0^2)^3} \\ &\times \{t(1 + e_0\cos f)^3\,[1 - 3\sin^2 i_0 \sin^2(\omega_0 + f)]\}_{f_0}^{f_1} \\ &+ \frac{1}{\tilde{n}_1}(1-e_0^2)^{\frac{1}{2}}\,[(\tilde{\omega}_1 - \omega_0) + (\cos i_0)(\tilde{\Omega}_1 - \Omega_0)] \\ &- \frac{JR^2}{\tilde{n}_1 a_0^2(1-e_0^2)^{\frac{3}{2}}}\left(1 - \frac{3}{2}\sin^2 i_0\right)(f_1 - f_0) + \text{p.t.}\end{aligned} \tag{34}$$

where f, f_1, f_0 represent the unperturbed anomaly. (We have omitted the superscript zero for convenience.) Equation (32) confirms the well-known secular behavior of the node. It turns out to shift westward for $0 < i_0 < \pi/2$ and eastward for $\pi/2 < i_0 < \pi$. At $i_0 = \pi/2$ it remains stationary, as would be expected from symmetry.

The secular component of $\omega^{(1)}$, according to (33), reduces to the well-known term

$$\frac{JR^2(5\cos^2 i_0 - 1)}{2a_0^2(1-e_0^2)^2}.$$

It represents an advance of perigee for $0 \leqq i_0 < 63°26'$ and for $116°34' < i_0 \leqq \pi$. For $63°26' < i_0 < 116°34'$ perigee regresses, and at the "critical" angles $63°26'$ and $116°34'$ it is reduced to periodic motions (as far as the first-order analysis indicates). We note that the periodic terms in

(33) contain e_0 in the denominator, and we expect $\omega^{(1)}$ to behave unstably for near-circular orbits (as one might expect for geometric reasons). Indeed, this singular behavior can be expected also in other examples, according to (19), (22) and (31). Furthermore, some difficulties will arise with small values of i_0, according to (20) and (21). These cases of near-circular and near-equatorial orbits can be accommodated by redefining the orbit elements in various ways. While such modified elements are less accessible to a geometric interpretation, they do not encumber the calculation of perturbed satellite positions as a function of time. For the sake of brevity we must forego additional details here.

3.2 *Luni-Solar Gravitation*

We omit a discussion of $\bar{a}_1$, since it shows periodic perturbations only. Substitution of (8) into (19)–(22) and (31) yields

$$\bar{e}_1 = e_0 - \frac{15 m_p a_0^{\,3} e_0 h_1 h_2}{m_e r_p^{\,5}} (1 - e_0^{\,2})^{\frac{1}{2}} \left\{ \tan^{-1}\left[\left(\frac{1 - e_0}{1 + e_0} \right)^{\frac{1}{2}} \tan \frac{f}{2} \right] \right\}_{f_0}^{f_1} + \text{p.t.} \tag{35}$$

where

$$h_1 = l_1 x_p + m_1 y_p + n_1 z_p$$

$$h_2 = l_2 x_2 + m_2 y_p + n_2 z_p$$

$$\bar{i}_i = i_0 + \frac{3 m_p a_0^{\,3}}{m_e r_p^{\,5} \sqrt{1 - e_0^{\,2}}} [(1 + 4e_0^{\,2}) h_1 h_{2_i} - (1 - e_0^{\,2}) h_2 h_{1_i}] \cdot \left\{ \tan^{-1}\left[\left(\frac{1 - e_0}{1 + e_0} \right)^{\frac{1}{2}} \tan \frac{f}{2} \right] \right\}_{f_0}^{f_1} + \text{p.t.} \tag{36}$$

$$\bar{\Omega}_1 = \Omega_0 + \frac{3 m_p a_0^{\,3} [(1 + 4e_0^{\,2}) (h_1^{\,2})_i + (1 - e_0^{\,2}) (h_2^{\,2})_i]}{2 m_e r_p^{\,5} \sin i_0 (1 - e_0^{\,2})^{\frac{1}{2}}} \cdot \left\{ \tan^{-1}\left[\left(\frac{1 - e_0}{1 + e_0} \right)^{\frac{1}{2}} \tan \frac{f}{2} \right] \right\}_{f_0}^{f_1} + \text{p.t.} \tag{37}$$

$$\bar{\omega}_1 = \omega_0 + \cos i_0 (\Omega_0 - \bar{\Omega}_1) + \frac{3 m_p a_0^{\,3}}{m_e r_p^{\,5}} (1 - e_0^{\,2})^{\frac{1}{2}} [4h_1^{\,2} - h_2^{\,2} - r_p^{\,2}] \cdot \left\{ \tan^{-1}\left[\left(\frac{1 - e_0}{1 + e_0} \right)^{\frac{1}{2}} \tan \frac{f}{2} \right] \right\}_{f_0}^{f_1} + \text{p.t.} \tag{38}$$

$$\tilde{\tau}_1 = \tau_0 \frac{n_0}{\tilde{n}_1} + \frac{(1 - e_0^2)^{\frac{1}{2}}}{\tilde{n}_1} [\tilde{\omega}_1 - \omega_0 + (\tilde{\Omega}_1 - \Omega_0) \cos i_0]$$

$$+ t_1 \left(1 - \frac{n_0}{\tilde{n}_1}\right) + \frac{3Gm_p(1 - e_0^2)^2(t_1 - t_0)}{2n_0 r_p^5 (1 + e_0 \cos f_0)^2}$$

$$\cdot [r_p^2 - 3(h_1 \cos f_0 + h_2 \sin f_0)^2] + \frac{7Gm_p}{2n_0^2 \tilde{n}_1 r_p^3} \tag{39}$$

$$\cdot \left[2 + 3e_0^2 - \frac{3h_1^2}{r_p^2}(1 + 4e_0^2) - \frac{3h_2^2}{r_p^2}(1 - e_0^2)\right]$$

$$\cdot \left\{\tan^{-1}\left[\left(\frac{1 - e_0}{1 + e_0}\right)^{\frac{1}{2}} \tan \frac{f}{2}\right]\right\}_{f_0}^{f_1} + \text{p.t.}$$

The subscripts i in (36) and (37) denote partial differentiation with respect to the inclination. The secular term of $e^{(1)}$ is a significant feature of our present results. It indicates that even near-circular satellite orbits can experience an unstable buildup of eccentricity due to luni-solar perturbations. The rate of this perturbation is proportional to the factor $m_p a_0^3 / m_e r_p^3$ and is usually very small; moreover, the coordinates x_p, y_p, z_p of the perturbing body are really time-dependent, which would modify the first-order result. Nevertheless, a long-period change of the eccentricity due to luni-solar gravitation has been observed in some satellite orbits.

The explicit form of the periodic terms in $a^{(1)}$ and $e^{(1)}$ contains the factor $1/e_0$, while $\Omega^{(1)}$, $\omega^{(1)}$ and $\tau^{(1)}$ contain $1/(\sin i_0)$. Again this necessitates the use of specially modified elements for low e_0 and i_0. One set of elements which is particularly suited to the problem of interplanetary perturbations is due to Strömgren. He utilized the fact the Ω, i, ω are nothing but a set of Euler angles orienting a system of orbital coordinates with one axis through pericenter, one at $f = \pi/2$, and one normal to the orbit. The rotation of these axes with respect to inertial space conveys the same information as the perturbations of Ω, i, ω. The idea is akin to Roberson's method for anticipating secular terms in the perturbation of coordinates (Section 4.3 and Ref. 15).

3.3 *Higher-Order Analyses*

The preceding examples are indicative of results to be found in the vast literature on perturbations in the osculating elements. We have merely covered the gist of this approach and several ideas which will be useful in the appraisal of other methods. Some of the better-known

contributions in terms of osculating elements are contained in papers by Krause, O'Keefe, Kozai, and Iszak.

In principle the quadratures (18)–(22) and (31) could be evaluated iteratively to generate higher-order results. This procedure rests directly on Poincaré's convergence proof, and a formal technique based on this approach is commonly attributed to Poisson. In several aerospace publications this has been done to obtain second- and third-order secular terms for oblateness effects. The algebraic labor is considerable, though typical secular terms such as (40) and (41) tend to be reasonably compact (see Refs. 7 and 9):

$$\Delta a^{(2)} = \frac{9\pi J^2 R^4 e_0}{2a_0^3(1 - e_0^2)^5}(1 - 3\sin^2 i_0 \sin^2\theta_0)[1 + e_0 \sin(\omega_0 + \theta_0)]^2 \times (5\sin^2 i_0 - 4)\cos(\omega_0 + \theta_0) \tag{40}$$

$$\Delta i^{(2)} = \frac{9\pi J^2 R^4 \sin 2i_0}{4a_0^4(1 - e_0^2)^4}\{\tfrac{1}{12}e_0^2 \sin 2\omega_0 - \tfrac{1}{4}(5\sin^2 i_0 - 4)[\tfrac{1}{2}e_0^2 \sin 2\omega_0 - e_0 \cos\omega_0(\cos\theta_0 + \tfrac{1}{3}\cos 3\theta_0) - e_0 \sin\omega_0(\sin\theta_0 - \tfrac{1}{3}\sin 3\theta_0)]\}. \tag{41}$$

These results represent secular increments over a 2π step in θ, the central angle measured from the node, where θ_0 is the initial value of θ. Considerable emphasis must be placed in the derivation of such expressions on checks from the conservation of energy and angular momentum and duplicate execution of the algebra. (One likes to think that more elaborate explicit expressions will be attainable with the advent of computer algebra.) A notable contribution in this area was made by Merson,[5] who presents second-order secular terms for J and first-order secular terms for the next four higher harmonics of the earth's potential. He also advocates the use of "smoothed" elements which reduce the amplitude of first-order periodic terms that might otherwise be inimical to prediction accuracy.

IV. PERTURBATIONS IN THE COORDINATES

We turn now to a description of satellite motions directly in terms of the position and velocity vectors. While these are dynamically equivalent to the instantaneous orbit elements, we note that the time dependence of the dynamic state variables in this form reflects the anomalistic motion as a primary effect. Therefore the long-time, secular changes of the orbit may not be obtainable with the same clarity or pre-

cision as in terms of the osculating parameters. On the other hand, the position-time history gives a direct account of the satellite motion in space and is useful for many aerospace applications. This prospect has stimulated several engineering analyses in recent years.

To the analyst with a general background in mechanics it would seem quite natural to approach a system of equations of the type (14) by standard perturbation techniques. Thus one could assume the perturbation series

$$\begin{aligned} x(t) &= x^{(0)}(t) + \kappa x^{(1)}(t) + \kappa^2 x^{(2)}(t) + \cdots \\ y(t) &= y^{(0)}(t) + \kappa y^{(1)}(t) + \kappa^2 y^{(2)}(t) + \cdots \\ z(t) &= z^{(0)}(t) + \kappa z^{(1)}(t) + \kappa^2 z^{(2)}(t) + \cdots \end{aligned} \tag{42}$$

and determine successively higher-order terms from the appropriate governing equations, following essentially Poisson's procedure. The traditional Encke method pursues this line of attack. However, Cartesian inertial coordinates have not enjoyed as much popularity as spherical ones, which seem more compliant with the geometry of satellite orbits. In the following, therefore, we shall concentrate on reference frames of this general type.

4.1 *Perturbations in Equatorial Spherical Coordinates*

A rather well-known perturbation analysis for oblateness effects is that due to Blitzer, Weissfeld, and Wheelon.[10] It uses the conventional equatorial spherical coordinates, r, φ, ψ (see Fig. 2) in terms of which the equations of motion read:

$$\ddot{r} - r\dot{\varphi}^2 - r\cos^2\varphi\,\dot{\psi}^2 = -\frac{k}{r^2} - \frac{3JkR^2}{r^4}\left[\tfrac{1}{3} - \sin^2\varphi\right] \tag{43}$$

$$\frac{d}{dt}(r^2\dot{\varphi}) + r^2\sin\varphi\cos\varphi\,\dot{\psi}^2 = -\frac{2JkR^2}{r^3}\sin\varphi\cos\varphi \tag{44}$$

$$\frac{d}{dt}(r^2\cos^2\varphi\,\dot{\psi}) = 0. \tag{45}$$

Here we have considered the first aspherical term in the earth's potential only. Since this is a zonal harmonic and does not contain ψ, the last equation has a vanishing right-hand side. Then a first integral of this equation

$$r^2\cos^2\varphi\,\dot{\psi} = p = \text{const.}, \tag{46}$$

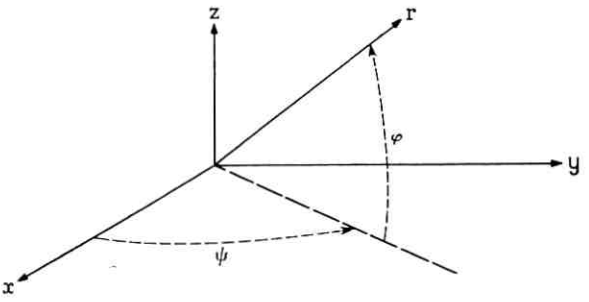


Fig. 2 — Equatorial spherical coordinate system.

representing the conservation of angular momentum about the polar axis, permits a change to ψ as independent variable. In addition, it is convenient to introduce the definitions

$$1/r = u, \qquad \tan \varphi = S, \qquad \text{and} \qquad [1/(r \cos \varphi)] = W. \tag{47}$$

Equations (43) and (44) now become

$$W'' + W = \frac{k}{p^2(1 + S^2)^{\frac{1}{2}}} + \frac{3JkR^2u^2}{p^2(1 + S^2)^{\frac{1}{2}}}\left[\frac{1}{1 + S^2} - \frac{2}{3}\right] \tag{48}$$

and

$$S'' + S = \frac{2JkR^2}{p^2}\frac{Su}{(1 + S^2)^2} \tag{49}$$

where "primes" denote differentiations with respect to ψ. In this form they are readily accessible to a perturbative procedure. We let

$$\begin{aligned} S &= S^{(0)} + \sum_{n=1}^{\infty} J^n S^{(n)} \\ u &= u^{(0)} + \sum_{n=1}^{\infty} J^n u^{(n)} \\ W &= W^{(0)} + \sum_{n=1}^{\infty} J^n W^{(n)}. \end{aligned} \tag{50}$$

Now we note that ψ in (46) was understood to represent the actual, i.e. perturbed, longitude of the satellite at all times. In order to make the connection between this perturbed independent variable and the time we find that

$$t - t_0 = \frac{1}{p}\int_{\psi_0}^{\psi} \frac{d\psi}{W^2} = \frac{1}{p}\int_{\psi_0}^{\psi} \frac{1}{W^{(0)2}}\left(1 - 2\frac{W^{(1)}}{W^{(0)}} + \cdots\right) d\psi$$
$$= t^{(0)} + \sum_{n=1}^{\infty} J^n t^{(n)} \tag{51}$$

i.e. the time itself evolves as a perturbation series. In the zero-order solution of (48) to (51) the right-hand side of (49) vanishes and we get

$$S^{(0)} = A \sin(\psi + \delta), \tag{52}$$

where A and δ are integration constants. On the right-hand side of (48) we retain the first term after substitution of $S^{(0)}$. Then $W^{(0)}$ follows in a straightforward manner. Substituting it into (51), the usual transcendental expression for time in Keplerian orbits results. It is somewhat obscured by the fact that its argument is given in terms of the longitude rather than one of the anomalies. We record the simplified form of these results for circular orbits, where $u^{(0)} = 1/r_0$:

$$W^{(0)} = \frac{1}{r_0}[1 + A^2 \sin^2(\psi + \delta)]^{\frac{1}{2}} \tag{53}$$

$$t^{(0)} = \frac{r_0^2}{2p}\left\{\frac{2}{(1+A^2)^{\frac{1}{2}}} \tan^{-1}[(1 + A^2)^{\frac{1}{2}} \tan(\psi + \delta)]\right\}_{\psi_0}^{\psi}. \tag{54}$$

It is important to note that the expressions (52) and (53) represent Keplerian (in fact circular) motion only for the unperturbed case: i.e., if (54) represents the entire time equation and no higher-order terms as in (51) exist. For any perturbed motion, where ψ is perturbed in relation to time, the zero-order terms of course retain the Keplerian forms (52) to (54), but they do not actually represent Keplerian motion.

If we now proceed to the first-order solution and retain only terms of $0(J)$ in (48) and (49), we find

$$S^{(1)\prime\prime} + S^{(1)} = -\frac{2kR^2}{p^2}\frac{S^{(0)}u^{(0)}}{[1 + S^{(0)2}]^2} \tag{55}$$

and

$$W^{(1)\prime\prime} + W^{(1)} = \frac{-3kS^{(0)}S^{(1)}}{p^2[1 + S^{(0)2}]^{\frac{5}{2}}} + \frac{3kR^2u^{(0)2}}{p^2[1 + S^{(0)2}]^{\frac{3}{2}}}\left[\frac{1}{1 + S^{(0)2}} - \frac{2}{3}\right]. \tag{56}$$

Since the right-hand side of (55) contains only zero-order quantities,

we begin our solution there and the result is

$$S^{(1)} = \frac{kR^2A}{p^2r_0}\frac{\sin(\psi+\delta)}{A^2[1+A^2\sin^2(\psi+\delta)]} + \frac{\cos(\psi+\delta)}{1+A^2} \cdot\left\{\frac{1}{\sqrt{1+A^2}}\tan^{-1}[\sqrt{1+A^2}\tan(\psi+\delta)] - \frac{\sin 2(\psi+\delta)}{1+A^2\sin^2(\psi+\delta)}\right\}. \tag{57}$$

Here we do not show a complementary solution, since it is of the same form as $S^{(0)}$ and can be absorbed with the constants A and δ. We could now substitute (57) into (56) to find $W^{(1)}$ and then use (51) to calculate the time. However, the inverse tangent in $S^{(1)}$ constitutes a secular term, which is considered an objectionable feature for some applications.

This disadvantage can be avoided by inserting an additional transformation between ψ and the argument of $S^{(0)}$. Instead of using $\psi+\delta$ for the latter let it be

$$\sigma = \lambda\psi + \delta \tag{58}$$

where

$$\lambda = 1 + \sum_{n=1}^{\infty} J^n\lambda_n \tag{59}$$

and the λ_n are constants. This device is commonly attributed to Lindstedt.[12] To obtain the zero-order solution we need only substitute σ for the angular arguments in (52) and (53). However the equation for $S^{(1)}$ changes significantly, viz.:

$$S^{(1)\prime\prime} + S^{(1)} = \frac{-2kR^2S^{(0)}u^{(0)}}{p^2[1+S^{(0)2}]^2} - 2\lambda_1S^{(0)\prime\prime} \tag{60}$$

where the primes now denote differentiations with respect to σ. Thus we find

$$S^{(1)} = \lambda_1 A\sin\sigma - \frac{kR^2(A^2\cos^2\sigma - 1 - A^2)}{p^2Ar_0(1+A^2)(1+A^2\sin^2\sigma)} - \cos\sigma\left\{\lambda_1A\sigma - \frac{kR^2A}{p^2r_0(1+A^2)^{\frac{1}{2}}}\tan^{-1}[(1+A^2)^{\frac{1}{2}}\tan\sigma]\right\}. \tag{61}$$

The appearance of the free parameter λ_1 in this result gives us the

opportunity to suppress the secular term. Thus, if we choose

$$\lambda_1 = \frac{kR^2}{p^2 r_0 (1 + A^2)^{\frac{3}{2}}} \tag{62}$$

(61) becomes

$$S^{(1)} = \frac{kR^2 A \cos \sigma}{p^2 r_0 (1 + A^2)^{\frac{3}{2}}} \{\tan^{-1} [(1 + A^2)^{\frac{1}{2}} \tan \sigma] - \sigma\}. \tag{63}$$

Here we have again omitted all terms of the same form as $S^{(0)}$. The net contribution from the terms in braces is cyclic and has the period 2π in σ.

According to (58) and (62) this amounts to a period of

$$2\pi \left[1 - \frac{JkR^2}{r_0 p^2 (1 + A^2)^{\frac{1}{2}}}\right] \tag{64}$$

in ψ. In effect, Lindstedt's transformation distorts the independent variable to absorb the secular effect. We shall see more of this later.

In principle we could transform (56) to σ and solve for W in a straightforward manner. However, to simplify the algebra, a redefinition of W will be convenient. We may backtrack to the explicit form of (48) in terms of S, u, and σ.

$$\begin{aligned}[1 + S^{(0)2}]^2 u^{(1)\prime\prime} + 2[1 + S^{(0)2}] S^{(0)} S^{(0)\prime} u^{(1)\prime} \\ + [S^{(0)\prime 2} + S^{(0)2} + 1] u^{(1)} \\ = \frac{3kR^2}{p^2} u^{(0)2} \left[\frac{1}{1 + S^{(0)2}} - \frac{2}{3}\right] \\ - 2[S^{(0)} S^{(1)} + S^{(0)\prime} S^{(1)\prime} + S^{(0)\prime 2} \lambda_1] u^{(0)}\end{aligned} \tag{65}$$

and take

$$W^{(1)} = (1 + S^{(0)2})^{\frac{1}{2}} u^{(1)}. \tag{66}$$

Then we obtain*

$$\begin{aligned}W^{(1)\prime\prime} + W^{(1)} = \frac{kR^2 u^{(0)2} [1 - 2S^{(0)2}]}{p^2 \Delta^5} \\ - \frac{2u^{(0)}}{\Delta^3} [S^{(0)} S^{(1)} + S^{(0)\prime} S^{(1)\prime} + S^{(0)\prime 2} \lambda_1]\end{aligned} \tag{67}$$

where $\Delta = (1 + A^2 \sin^2 \sigma)^{\frac{1}{2}}$.

Substituting (63) and ignoring the complementary solution for $W^{(1)}$ we

* Note that the formulas (36) and (38) in Ref. 9 contain several misprints.

have

$$W^{(1)} = \frac{-kR^2}{3p^3r_0^2(1+A^2)^3\Delta^3}\,[4A^4(A^4-1)\sin^4\sigma + A^2(9A^4+2A^2-7)\sin^2\sigma + 5A^4+2A^2-3]. \tag{68}$$

Now it only remains to find a relation between σ and time. From (46) it is clear that

$$t - t_0 = \frac{1}{p}\int_{\psi_0}^{\psi} r^2\cos^2\varphi\,d\psi = \frac{1}{p}\int_{\sigma_0}^{\sigma}\frac{d\sigma}{u^2(1+S^2)\lambda}, \tag{69}$$

which we expand up to $0(J)$. This results in

$$t^{(0)} = \frac{r_0^2}{p(1+A^2)^{\frac{1}{2}}}\tan^{-1}[(1+A^2)^{\frac{1}{2}}\tan\sigma] \tag{70}$$

and

$$\begin{aligned} t^{(1)} &= -\frac{r_0^2}{p}\int_{\sigma_0}^{\sigma}\frac{1}{\Delta^2}\left[\lambda_1 + \frac{2r_0}{\Delta}W^{(1)} + \frac{2A\sin\delta\,S^{(1)}}{\Delta^2}\right]d\sigma \\ &= -\frac{kR^2r_0}{p^3(1+A^2)^{\frac{3}{2}}}\left\{\frac{(1+A^2)^{\frac{1}{2}}}{12\Delta^2}[A^2\sin 2\sigma + 12(1+A^2)^{\frac{1}{2}}\sigma]\right. \\ &\quad + \left[2 - 3A^2 + \frac{A^2}{2\Delta^2}((1+A^2)\sin^2\sigma - \cos^2\sigma)\right] \\ &\quad \left.\cdot\tan^{-1}[(1+A^2)^{\frac{1}{2}}\tan\sigma]\right\} \end{aligned} \tag{71}$$

and completes this analysis of near-circular orbits.

Throughout the foregoing discussion we have used a perturbed coordinate, ψ or σ, as the independent variable. In principle, we could have done without Lindstedt's device, and we could have used the unperturbed longitude $\psi^{(0)}$ as the independent variable. This approach has the attractive feature that the zero-order solution (in terms of $\psi^{(0)}$) represents true Keplerian motion. The perturbed longitude could be expressed in terms of $\psi^{(0)}$ as

$$\psi = \psi^{(0)} + \sum_{n=1}^{\infty} J^n\psi^{(n)}(\psi^{(0)}). \tag{72}$$

However, if one develops the governing differential equations for $S^{(1)}$ and $W^{(1)}$, he discovers that they are completely coupled for this particular example. This approach, therefore, loses its practical value.

In conclusion we note that, in spite of various transformations, the

final results (63), (68), (71) of this analysis seem rather awkward, considering the fact that they represent the relatively trivial first-order oblateness perturbations of a near-circular orbit. This is of course due to the choice of spherical equatorial coordinates to represent the motion in a nonequatorial orbit. That disadvantage was eliminated in other formulations, to be considered next.

4.2 *Perturbations in Orbital Spherical Coordinates*

As the title of this section indicates, it is more natural to use the plane of unperturbed motion as the fundamental plane for inclined orbits. A typical analysis in this category is that by Anthony, Fosdick et al.[13,14] The coordinates r, θ, β of their reference frame (see Fig. 3) take the place of r, ψ, $(\pi/2) - \varphi$ in Fig. 2. The angle $\alpha = (\pi/2) - \varphi$ (Fig. 3) is introduced occasionally for trigonometric simplifications.

The left-hand sides of the equations of motion of course are not altered by this change of coordinates, but the right-hand sides (representing oblateness perturbations) acquire the forms shown in (73) to (75). As usual, we introduce $u = 1/r$ and $p = r^2\dot{\theta} = \dot{\theta}/u^2$ and change the independent variable from t to θ. We note that θ is the perturbed central angle in the nominal orbit plane. Thus, the equations of motion become

$$(pu') + pu(\beta'^2 + \sin^2\beta) = (k/p)[1 + JR^2u^2(1 - 3\cos^2\alpha)] \tag{73}$$

$$\begin{aligned}(p\theta')' - p\sin\beta\cos\beta = {} & (-kJR^2u/p)[(\sin^2 i\sin^2\theta \\ & - \cos^2 i)\sin 2\beta \\ & + \sin 2i\cos 2\beta\sin\theta]\end{aligned} \tag{74}$$

$$\begin{aligned}(p\sin^2\beta)' = {} & (-kJR^2u/p)[\sin^2 i\sin^2\beta\sin 2\theta \\ & + \tfrac{1}{2}\sin 2i\sin 2\beta\cos\theta]\end{aligned} \tag{75}$$

where primes denote derivatives with respect to θ. We subject this variable to the first-order Lindstedt transformation

$$\sigma = \theta(1 + J\lambda_1) \tag{76}$$

and use the "ansatz"

$$\begin{aligned}u &= u^{(0)}(\sigma) + Ju^{(1)}(\sigma) \\ p &= p^{(0)}(\sigma) + Jp^{(1)}(\sigma) \\ \beta &= (\pi/2) + J\beta^{(1)}(\sigma).\end{aligned} \tag{77}$$

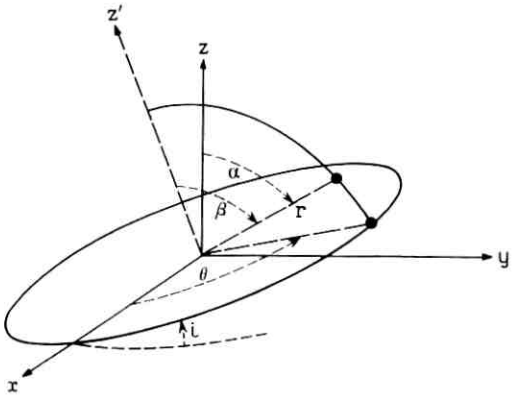


FIG. 3 — Orbital spherical coordinate system.

Without an excessive loss of generality we may assume a horizontal launch at the node and, by the definition of the θ plane as nominal orbit plane, the initial velocity vector $\mathbf{v}_0$ lies in it. Thus, at $t = 0$: $r = r_0$, $\dot{r} = 0$, $\beta = \pi/2$, $\dot{\beta} = 0$, and $\dot{\theta} = v_0/r_0$. In general, v_0 will be such as to produce an elliptic orbit. The zero-order results are

$$p^{(0)} = r_0 v_0$$

and

$$u^{(0)} = (k/r_0^2 v_0^2)[1 + e \cos \sigma] \tag{78}$$

where

$$e = (r_0 v_0^2/k) - 1$$

and has the form of a Keplerian eccentricity. As in Section 4.1, the zero-order solution (78) will represent Keplerian motion only for the unperturbed case, i.e. when $\sigma = \theta = \theta^{(0)}$. Now the first-order solutions follow in a straightforward manner:

$$p^{(1)} = (-k^2/2r_0 v_0^3)(R/r_0)^2 \sin^2 i \{1 + \tfrac{4}{3}e - e \cos \sigma - \tfrac{1}{3}e \cos 3\sigma - \cos 2\sigma\} \tag{79}$$

$$\begin{aligned} u^{(1)} = (k^3 R^2/r_0^6 v_0^6)\{&1 + \tfrac{1}{2}e^2 + \sin^2 i(-\tfrac{1}{2} + \tfrac{4}{3}e - \tfrac{5}{8}e^2) \\ &- \tfrac{1}{6}(e^2 + \sin^2 i) \cos 2\sigma - \tfrac{5}{24}e \sin^2 i \cos 3\sigma \\ &- \tfrac{1}{24}e^2 \sin^2 i \cos 4\sigma \\ &- [1 + \tfrac{1}{3}e^2 - \sin^2 i(\tfrac{2}{3} - \tfrac{9}{8}e + \tfrac{2}{3}e^2)] \cos \sigma\}, \end{aligned} \tag{80}$$

where we had to select the Lindstedt parameter as

$$\lambda_1 = (k^2R^2/r_0^4v_0^4)(\tfrac{3}{2}\sin^2 i - 1) \tag{81}$$

to avoid a secular term in (80). Finally

$$\beta^{(1)} = \frac{k^2R^2 \sin 2i}{2r_0^4v_0^4}[(1 + \tfrac{2}{3}e)\sin\sigma - \sigma\cos\sigma - \tfrac{1}{3}e\sin 2\sigma]. \tag{82}$$

It is clear that the secular term which was absorbed by the Lindstedt parameter has to do with the apsidal precession. In the absence of additional Lindstedt parameters we have no countermeasures against the secular term in (82), which reflects the nodal regression. (Note that the latter was counteracted by the Lindstedt transformation of Section 4.1, since it was the only secular effect to be considered for near-circular orbits.)

The time equation for this example can be written in a straightforward fashion. From the definition of p it follows that

$$t_1 - t_0 = \int_{\sigma_0}^{\sigma_1} \frac{(1 - J\lambda_1)}{pu^2}\, d\sigma$$

where σ_0 and σ_1 correspond to the time limits t_0 and t_1. An expansion to first-order terms yields

$$\begin{aligned} t^{(0)} + Jt^{(1)} = &\int_{\sigma_0}^{\sigma_1} \frac{d\sigma}{p^{(0)}u^{(0)2}} \\ &- J\int_{\sigma_0}^{\sigma_1} \frac{1}{p^{(0)}u^{(0)2}}\left[\lambda_1 + \frac{p^{(1)}}{p^{(0)}} + 2\frac{u^{(1)}}{u^{(0)}}\right] d\sigma. \end{aligned} \tag{83}$$

This leads to

$$\begin{aligned} t^{(0)} = \frac{r_0^3v_0^3}{k^2}\Bigg\{ &\frac{-e\sin\sigma}{(1 - e^2)(1 + e\cos\sigma)} \\ &+ \frac{2}{(1 - e^2)^{\frac{3}{2}}}\tan^{-1}\left[\left(\frac{1 - e}{1 + e}\right)^{\frac{1}{2}}\tan\frac{\sigma}{2}\right]\Bigg\}_{\sigma_0}^{\sigma_1}, \end{aligned} \tag{84}$$

which we recognize as being of strictly Keplerian form but in terms of the perturbed angle σ. Similarly

$$
\begin{aligned}
t^{(1)} = \frac{R^2}{r_0 v_0} \Bigg\{ & \frac{\sin\sigma(4 - e^2 + 3e\cos\sigma)}{(1 - e^2)^2(1 + e\cos\sigma)^2} \left[1 + \tfrac{2}{3}e + \tfrac{1}{3}e^2 + \tfrac{2}{3}e^3 \right. \\
& \left. + \left(\frac{1}{6e} - \tfrac{2}{3} + \tfrac{1}{2}e + \tfrac{2}{3}e^2 - \tfrac{2}{3}e^3 \right) \sin^2 i \right] \\
& - \frac{\sin\sigma}{(1 - e^2)(1 + e\cos\sigma)} \left[2 - \tfrac{1}{3}e + \tfrac{2}{3}e^2 \right. \\
& \left. + \left(\frac{2}{3e} - \tfrac{4}{3} + \tfrac{11}{6}e - \tfrac{2}{3}e^2 \right) \sin^2 i \right] \\
& - \frac{[2(1 + e)^3 + \tfrac{4}{3}e^3 \sin^2 i]}{(1 - e^2)^{\frac{3}{2}}} \tan^{-1}\left[\left(\frac{1 - e}{1 + e} \right)^{\frac{1}{2}} \tan\frac{\sigma}{2} \right] \Bigg\}_{\sigma_0}^{\sigma_1}.
\end{aligned}
\tag{85}
$$

The set of results (78) to (85) gives a reasonably convenient description of first-order oblateness perturbations which might be useful in the targeting and guidance of space vehicles. Extensions to near-parabolic and hyperbolic trajectories follow quite readily. As in Section 4.1, we note that the analysis might have been executed in terms of an unperturbed independent variable, viz. $\theta^{(0)}$ instead of θ, and in that case the zero-order solution would represent true Keplerian motion.

The inclusion of secular perturbations in the independent variable σ serves the same purpose as the definitions of "mean elements" introduced by Breakwell et al., by Hansen, in the von Zeipel method, and in modern averaging techniques. The Lindstedt transformation is not the most powerful device in this category but it can be extended to absorb secular effects in more than one coordinate. This will be illustrated in the next section in terms of "secular rotations" of the reference frame.

4.3 *Perturbations in Rotating Spherical Orbital Coordinates*

The idea of using suitable coordinate transformations with arbitrary parameters to neutralize secular trends was exploited in a more general way by R. E. Roberson.[15] His approach uses the orbital coordinates r, θ, δ [$= (\pi/2) - \beta$], in agreement with Section 4.2, but assumes that the entire reference frame will be subjected to three monotonic rotations, corresponding to three Euler angles, such that the satellite motion relative to this reference frame exhibits only periodic perturbations. This kinematic outlook on secular trends forms an interesting parallel to several classical procedures. Roberson himself makes some illuminating comparisons between engineering analyses such as Refs. 9, 10, and 12 to 16 and traditional formulations in terms of mean variables. He restricts his analysis to first-order perturbations, realizing that a con-

sistent higher-order theory would have to include contributions from other physical effects and various coupling terms. Some of his remarks seem quite perspicacious in comparison with the other aerospace literature of that time.

The angular velocities stipulated for the reference frame must of course depend on the secular effects that need to be absorbed. In the presence of several physical disturbances the different angular motions of the coordinate system can be superposed to first order, and the resultant motion of the reference frame will succeed in neutralizing all the secular effects simultaneously. This seems intuitively obvious and can be demonstrated in a straightforward fashion.[15]

In Fig. 4 the angles $\bar{\Omega}$ and $\bar{\imath}$ define a mean orbit plane, in that each of them manifests a secular rate. Now the satellite position is given in terms of the orthogonal system $\tilde{x}$, $\tilde{y}$, $\tilde{z}$, which displays a secular variation with respect to the node (and this corresponds to the third Eulerian rotation). Let the three secular rotations be denoted $\kappa\dot{\Omega}^{(1)}$, $\kappa(di^{(1)}/dt)$, $\kappa\dot{\omega}^{(1)}$ where κ is the perturbation parameter. They will in general be functions of $\bar{\Omega}$, $\bar{\imath}$ and the characteristics of the perturbation source.

Turning to the problem of first-order oblateness perturbations, we set $\kappa = J$ and assume the appropriate form for the perturbing potential. Now let

$$\theta = \theta_0 + \bar{f} = \theta_0 + \bar{f}^{(0)} + J\bar{f}^{(1)}, \tag{86}$$

where $\bar{f} = 0$ at $t = 0$. We adopt $\bar{f}$ as independent variable and let de-

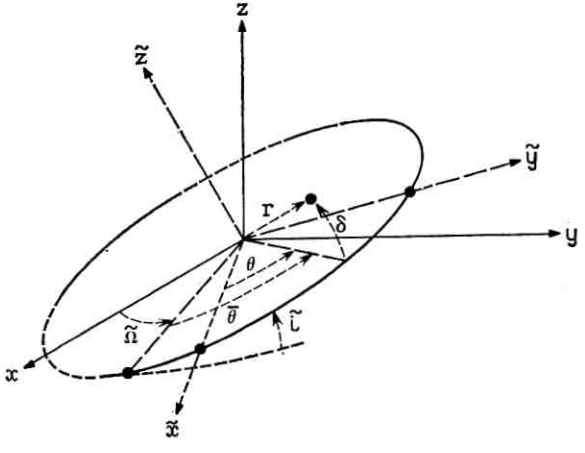


FIG. 4 — Rotating spherical orbital coordinates.

rivatives with respect to it be denoted by primes. Assuming that the secular rates are constants, we let

$$\begin{aligned} \bar{\Omega} &= \Omega^{(0)} + J\hat{\Omega}^{(1)}\bar{f} \\ \bar{\imath} &= i^{(0)} + J\hat{\imath}^{(1)}\bar{f} \\ \bar{\theta} &= \theta + J\hat{\omega}^{(1)}\bar{f}. \end{aligned} \tag{87}$$

As usual, the equations of motion are transformed by means of

$$u = 1/r \quad \text{and} \quad p = r^2\dot{\bar{f}} \tag{88}$$

and we find:

$$\begin{aligned} &(pu')' + pu[(\delta' - \hat{\Omega}^{(1)} \cos\theta \sin i)^2 \\ &\quad + (\cos\delta + \hat{\omega}^{(1)} \cos\delta + \hat{\Omega}^{(1)} \cos\varphi \cos\nu)^2] \\ &\quad - (k/p)[1 + 3JR^2u^2(1 - 3\sin^2\varphi)] = 0 \end{aligned} \tag{89}$$

$$\begin{aligned} &[p(\delta' - \hat{\Omega}^{(1)} \sin i \cos\theta)]' + p[(1 + \hat{\omega}^{(1)})^2 \sin\delta \cos\delta \\ &\quad + (1 + \hat{\omega}^{(1)})\hat{\Omega}^{(1)}(\sin\delta \cos\nu \cos\varphi + \cos\delta \sin\varphi) \\ &+ \hat{\Omega}^{(1)^2} \sin\varphi \cos\nu \cos\varphi] + (JkR^2 6u/p) \sin\varphi \cos\varphi \cos\nu = 0 \end{aligned} \tag{90}$$

$$\begin{aligned} &[p \cos\delta\ (\bar{\theta}' \cos\delta + \hat{\Omega}^{(1)} \cos\varphi \cos\nu)]' \\ &\quad + p[(1 + \hat{\omega}^{(1)})\hat{\Omega}^{(1)} \sin\delta \cos\delta \sin i \cos\theta \\ &\quad + \hat{\Omega}^{(1)^2} \cos\theta \sin i \sin\delta \cos\nu \cos\varphi \\ &\quad - \delta'\hat{\Omega}^{(1)} \sin i \sin\theta] \\ &\quad + (6JkR^2u/p) \sin\varphi \cos\delta \sin i \cos\theta = 0, \end{aligned} \tag{91}$$

where φ is the latitude and ν is defined in Fig. 5. We have made a slight digression from orderly progress in this step by setting $\hat{\imath}^{(1)} = 0$. This is prompted by previous experience with this problem — viz., that no first-order secular perturbations occur in i — and would have developed from the later calculations in any event.

Using the forms

$$\begin{aligned} u &= u^{(0)} + Ju^{(1)} \\ p &= p^{(0)} + Jp^{(1)} \\ \delta &= J\delta^{(1)} \end{aligned} \tag{92}$$

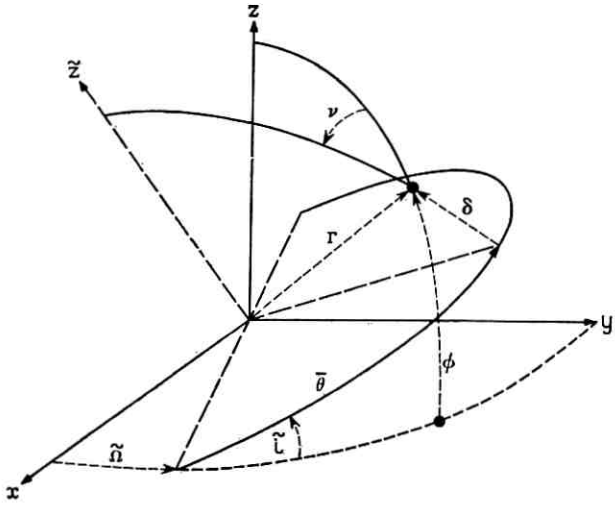


Fig. 5 — Definition of ν.

we reduce (89)–(91) to equations of $0(1)$ and $0(J)$. The zero-order solution is of course

$$\delta^{(0)} \equiv 0, \qquad p^{(0)} = \text{const.},$$
$$\text{and} \qquad u^{(0)} = (k/p^{(0)2})[1 + e \cos(\bar{f} - \alpha)]. \tag{93}$$

As in previous examples, we see that this will represent Keplerian motion only in the absence of perturbations, i.e. if $\bar{f} \equiv \bar{f}^{(0)}$ and $\hat{\Omega}^{(1)} = \hat{\omega}^{(1)} = 0$, yielding an inertial reference frame. We assume that the constants of integration ($p^{(0)}$, e, α) are chosen such that (93) with $\bar{f} = 0$ yields the satellite position and velocity at $t = 0$.

Proceeding with the solutions to $0(J)$ in the usual fashion, we require that

$$\hat{\Omega}^{(1)} = -\frac{3Jk^2R^2 \cos i^{(0)}}{p^{(0)4}} \tag{94}$$

in order to avoid a secular term in $\delta^{(1)}$ and

$$\hat{\omega}^{(1)} = (3R^2k^2/2p^{(0)4})(5 \cos^2 i^{(0)} - 1) \tag{95}$$

to avoid one in $u^{(1)}$. These of course reflect the nodal and apsidal precessions. The complementary solution for $p^{(1)}$ introduces one constant of integration, and the complementary solutions for $\delta^{(1)}$ and $u^{(1)}$ have

the form

$$A \sin \tilde{f} + B \cos \tilde{f}. \tag{96}$$

Since the zero-order solution already accounts for the dynamic state of the satellite at $t = 0$, the first-order solution encounters homogeneous initial conditions as far as they do not reflect the rotation of the reference frame. Thus at $\tilde{f} = 0$:

$$u^{(1)} = \delta^{(1)} = 0, \qquad u^{(1)\prime} = 0$$
$$\delta^{(1)\prime} - \hat{\Omega}^{(1)} \sin i^{(0)} \cos \theta_0 = 0,$$

and

$$(1/\dot{\tilde{f}}^{(0)})(2u^{(1)}u^{(0)}p^{(0)} + u^{(0)^2}p^{(1)}) + \hat{\omega}^{(1)} + \hat{\Omega}^{(1)}(\cos i^{(0)} - \sin i^{(0)} \sin \theta_0) = 0. \tag{97}$$

These govern the first-order constants of integration. [A little reflection shows that the forms (96) for $\delta^{(1)}$ and $u^{(1)}$ can be interpreted geometrically as constant changes of Ω and θ to $0(J)$. Roberson anticipates this by introducing such constants in (87) and using them in place of two of the integration constants for $\delta^{(1)}$ and $u^{(1)}$. However, nothing seems to be gained by this artifice and, if anything, it distracts from a systematic procedure.]

Finally, the time equation follows as usual in terms of $\tilde{f}$ to $0(J)$. Roberson proceeds to invert it, though the computational gains do not seem to justify this algebraic labor.

So much for our sketch of Roberson's procedure. Its extension to higher-order analyses is fairly obvious. At every level of refinement, $0(J^n)$, three coordinate rotations may be introduced — which are commensurate with the three degrees of freedom of the satellite problem whose secular trends we are trying to neutralize.

For "medium-range" prediction formulas it seems an open issue whether the rationale described in this section and traditional astronomical devices (like the "auxiliary ellipse" used by Hansen or the von Zeipel transformations based on Hamilton-Jacobi techniques) offer a computational advantage over the straightforward development of the Poisson method for successive higher-order terms. With the advent of computer algebra the latter technique may be quite satisfactory for many applications. However, for "long-range" predictions and lifetime studies it seems advisable to employ the accredited astronomical techniques of "extracting" secular effects and anticipating long-period terms in one way or another.

V. MOVING RECTANGULAR ORBITAL COORDINATES

We close this article with a formulation which calculates the position offsets for a satellite from its unperturbed orbit in an explicit form.[18,19] Instead of reckoning the perturbations in terms of the quantities r, θ, β or δ, which are defined relative to the center of the earth, we now consider a coordinate system whose origin is the nominal satellite position $0'$ on the unperturbed orbit (see Fig. 6). The latter may be defined by the initial conditions at time t_0, viz. $\mathbf{r}_0$ and $\mathbf{v}_0$. We establish an orthogonal triad about the moving point $0'$ with ξ in the radial direction, η in the direction of anomalistic motion, and ζ normal to the orbit plane. In the guidance engineer's language

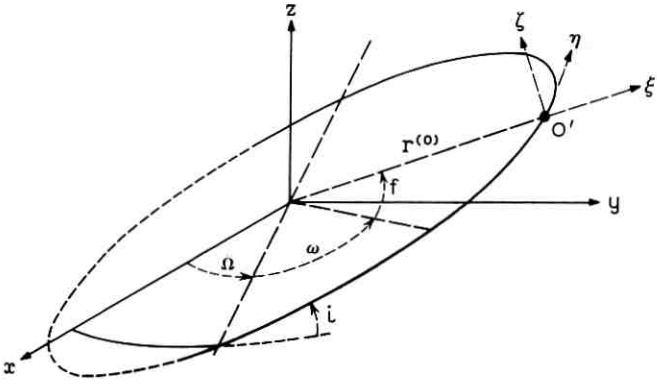


FIG. 6 — Moving rectangular coordinates centered at nominal satellite position.

these represent offsets in altitude, range, and cross-range. Any nonvanishing coordinates in this system are the effects of errors in the initial conditions or of geophysical forces. It is clear that this description of the perturbed motion can be quite useful in guidance studies, e.g. to exhibit the relative motion between a space station (given by $0'$) and a transfer vehicle (located at ξ, η, ζ) in a homing maneuver. In the subsequent discussion f will always represent the unperturbed true anomaly in the nominal orbit and $\theta = \omega + f$ the unperturbed central angle.* No Lindstedt transformations or perturbative coordinate rotations will be employed to develop this theory into a more sophisticated prediction scheme. Instead, we concentrate on the geometric interpretation of various results.

* We depart from earlier notations by omitting the superscript (0) from unperturbed quantities for simplicity.

The equations of motion can be derived by the standard Lagrangian or Hamiltonian formalism,[18] and their linearization to $0(\xi, \eta, \zeta)$ yields

$$\xi'' - 2\eta' - \xi - 2[\xi + e(\xi' - \eta)\sin f]/(1 + e\cos f) = -\frac{a^3(1-e^2)^3}{k}\tilde{V}_\xi/(1+e\cos f)^4 \tag{98}$$

$$\eta'' + 2\xi' - \eta - [-\eta + 2e(\eta' + \xi)\sin f]/(1 + e\cos f) = -\frac{a^3(1-e^2)^3}{k}\tilde{V}_\eta/(1+e\cos f)^4 \tag{99}$$

$$\zeta'' + [\zeta - 2\zeta' e\sin f]/(1 + e\cos f) = -\frac{a^3(1-e^2)^3}{k}\tilde{V}_\zeta/(1+e\cos f)^4 \tag{100}$$

where primes denote derivatives with respect to f and $\tilde{V}$ is the perturbative potential, which exists in addition to the central body term $-k/r$. The subscripts of $\tilde{V}$ denote partial derivatives with respect to ξ, η, or ζ.

The solution of the homogeneous set (98) and (99), where $\tilde{V} \equiv 0$, represents a complementary solution for the cases where $\tilde{V} \not\equiv 0$ and will be needed to satisfy the initial conditions. For an elliptic nominal orbit this solution has the form

$$\xi = \delta a\left[\frac{1-e^2}{1+e\cos f} - \frac{3e}{2}\left(\frac{k}{(1-e^2)a^3}\right)^{\frac{1}{2}}(t-\tau)\sin f\right] - \delta e\, a\cos f - \delta\tau\, e\left(\frac{k}{(1-e^2)a}\right)^{\frac{1}{2}}\sin f \tag{101}$$

$$\eta = -\delta a\frac{3}{2}\left(\frac{k}{(1-e^2)a^3}\right)^{\frac{1}{2}}(1+e\cos f)(t-\tau) + \delta e\, a\sin f\frac{(2+e\cos f)}{1+e\cos f} - \delta\tau\left(\frac{k}{(1-e^2)a}\right)^{\frac{1}{2}} \cdot(1+e\cos f) + \delta\omega\frac{a(1-e^2)}{1+e\cos f} \tag{102}$$

$$\zeta = \frac{a(1-e^2)}{(1+e\cos f)}[\delta i\sin\theta - \delta\Omega\sin i\cos\theta]. \tag{103}$$

This result can be adapted to hyperbolic, parabolic, and near-parabolic orbits without much trouble. The constants of integration $\delta a \cdots \delta\tau$ are of course just a set of numbers to be determined from the initial conditions, but the symbols we use for them indicate the parameter changes

of the nominal orbit that they represent. Alternatively, these constants could be given in terms of $\xi_0 \cdots \zeta_0'$, the perturbations of the position and velocity at t_0. That form is more descriptive for various guidance applications. Thus we find for nominally circular orbits

$$\xi = 2\eta_0' + 4\xi_0 - (2\eta_0' + 3\xi_0)\cos(f - f_0) + \xi_0' \sin(f - f_0) \tag{104}$$

$$\eta = \eta_0 - 2\xi_0' - 3(\eta_0' + 2\xi_0)(f - f_0) + 2(2\eta_0' + 3\xi_0)\sin(f - f_0) + 2\xi_0'\cos(f - f_0) \tag{105}$$

$$\zeta = \zeta_0' \sin(f - f_0) + \zeta_0 \cos(f - f_0). \tag{106}$$

(Since a nominal perigee does not exist for this case, we assume that the angle ω, whose existence is still implied by the notation f, has some arbitrary value $\bar{\omega}$. Without loss of generality we can set $f_0 = 0$ so that $\theta_0 = \bar{\omega}$.)

In a guidance application $\xi_0 \cdots \zeta_0'$ might represent the errors resulting from a position and velocity determination or a corrective thrust maneuver and (104) to (106) would then describe the "run-out" as a function of time. In an "orbit sensitivity" study these expressions can be used to demonstrate the effect of $\xi_0 \cdots \zeta_0'$ on the orbit parameters. In a homing maneuver the same expressions would represent the relative motion between the two vehicles attempting a rendezvous. In principle, two relative position measurements $\mathbf{X}$: ξ, η, ζ at separate times suffice to determine all the constants in (104) to (106), and a corrective maneuver could be planned to drive the residuals to zero at a specified instant or by successive approximations.

Particular solutions of (98) to (100) can be found in a straightforward manner if $e = 0$. For $e \neq 0$ we construct these solutions as power series in e, for lack of a better expedient. We consider the series to $0(e)$ and let them be denoted by

$$\begin{aligned} \xi &= \bar{\xi}_1 + \bar{\xi}_2 + e \sum_j f_j \\ \eta &= \bar{\eta}_1 + \bar{\eta}_2 + e \sum_j g_j \qquad j = 1, 2, 3 \\ \zeta &= \bar{\zeta}_1 + \bar{\zeta}_2 + e \sum_j h_j \end{aligned} \tag{107}$$

where

$\bar{\xi}_1\bar{\eta}_1\bar{\zeta}_1$ = the complementary solution (104) to (106)
$\bar{\xi}_2\bar{\eta}_2\bar{\zeta}_2$ = a particular solution representing $\bar{V}$ to $0(\kappa)$
$f_1g_1h_1$ = solution reflecting $e \cdot (\bar{\xi}_1, \bar{\eta}_1, \bar{\zeta}_1)$ on the right-hand sides of (98) to (100)

$f_2g_2h_2$ = solution reflecting $e\cdot(\bar{\xi}_2, \bar{\eta}_2, \bar{\zeta}_2)$ on the right-hand sides
$f_3g_3h_3$ = solution reflecting $e\cdot(\tilde{V}_\xi, \tilde{V}_\eta, \tilde{V}_\zeta)$ on the right-hand sides.
The following explicit general forms can be given for these solutions:

$$\begin{aligned}
\bar{\xi}_2 &= (a^3/k)\Big[-2\int \tilde{V}_\eta\, df + 2\cos f\int \tilde{V}_\eta \cos f\, df \\
&\quad + \cos f\int \tilde{V}_\xi \sin f\, df + 2\sin f\int \tilde{V}_\eta \sin f\, df \\
&\quad - \sin f\int \tilde{V}_\xi \cos f\, df\Big] \\
\bar{\eta}_2 &= (a^3/k)\Big[3\iint \tilde{V}_\eta\, df\, df + 2\int \tilde{V}_\xi\, df - 4\sin f\int \tilde{V}_\eta \cos f\, df \\
&\quad - 2\sin f\int \tilde{V}_\xi \sin f\, df + 4\cos f\int \tilde{V}_\eta \sin f\, df \\
&\quad - 2\cos f\int \tilde{V}_\xi \cos f\, df\Big] \\
\bar{\zeta}_2 &= (a^3/k)\Big[\cos f\int \tilde{V}_\zeta \sin f\, df - \sin f\int \tilde{V}_\zeta \cos f\, df\Big]
\end{aligned} \tag{108}$$

where we take $f_0 = 0$ for the lower limit of all quadratures.

$$\begin{aligned}
f_1 &= (\eta_0 - 2\xi_0')\sin f - \tfrac{5}{2}(\eta_0' + 2\xi_0)\cos f - 3(\eta_0' + 2\xi_0) \\
&\quad \cdot f\sin f \\
g_1 &= 7(\eta_0' + 2\xi_0)\sin f + (\eta_0 - 2\xi_0')\cos f - 3(\eta_0' + 2\xi_0) \\
&\quad \cdot f\cos f - (\xi_0'/2)\cos 2f - (\eta_0' + \tfrac{3}{2}\xi_0)\sin 2f \\
h_1 &= -(\zeta_0/2) - (\zeta_0'/2)\sin 2f - (\zeta_0/2)\cos 2f.
\end{aligned} \tag{109}$$

The terms f_2, g_2, and h_2 are obtainable from expressions analogous to (108) but with the following substitutions:

$$\begin{aligned}
&2(\bar{\xi}_2' - \bar{\eta}_2)\sin f - 2\bar{\xi}_2\cos f && \text{for } (-a^2/k)\tilde{V}_\xi \\
&2(\bar{\eta}_2' + \bar{\xi}_2)\sin f + \bar{\eta}_2\cos f && \text{for } (-a^3/k)\tilde{V}_\eta \\
&2\bar{\zeta}_2'\sin f + \bar{\zeta}_2\cos f && \text{for } (-a^3/k)\tilde{V}_\zeta
\end{aligned} \tag{110}$$

and $f_3g_3h_3$ follow from (108) if we substitute

$$(4a^3/k)\tilde{V}_{\xi,\eta,\zeta} \qquad \text{for } (-a^3/k)\tilde{V}_{\xi,\eta,\zeta}\,. \tag{111}$$

Since the differential operators for all of these partial solutions are of the form

$$\begin{aligned} &\xi'' - 2\eta' - 3\xi \\ &\eta'' + 2\xi' \\ &\zeta'' + \zeta, \end{aligned} \tag{112}$$

no explicit complementary solution of $0(e)$ is provided: i.e., the constants $\xi_0 \cdots \zeta_0'$ will be used to satisfy the i.c.'s to all levels of accuracy. These will differ from zero only if the nominal orbit is taken to differ from $\mathbf{r}_0$ and $\mathbf{v}_0$ at t_0.

Specific results may now be obtained by the above formulas, which lend themselves to a geometric interpretation of perturbed satellite motions. For the oblateness effect one finds

$$\begin{aligned} \bar{\xi}_2 &= (JR^2/a)[-1 + \sin^2 i(\tfrac{3}{2} + \tfrac{1}{6}\cos 2\theta)] \\ \bar{\eta}_2 &= (JR^2/a)[(2 - 3\sin^2 i)f + \tfrac{1}{12}\sin^2 i \sin 2\theta] \\ \bar{\zeta}_2 &= (JR^2/2a)\sin 2i[f\cos\theta - \tfrac{1}{2}\sin\theta]. \end{aligned} \tag{113}$$

The terms in 2θ reflect the doubly symmetric distortion of the orbit due to the oblateness of the gravitational field. The constant term in $\bar{\xi}_2$ and the secular term in $\bar{\eta}_2$ reflect the additional mass of the equatorial bulge. Combining (113) with (104) to (106) into a complete solution, we observe that the constant term in ξ is

$$\Delta a = (JR^2/a)[-1 + \tfrac{3}{2}\sin^2 i] + 2\eta_0' + 4\xi_0 \tag{114}$$

and the secular term in η

$$a\Delta\theta = f[(JR^2/a)(2 - 3\sin^2 i) - 3(\eta_0' + 2\xi_0)],$$

which represent the differences between the nominal circular orbit and the mean circular orbit resulting from the perturbations. Since $\Delta a = 0$ and $\Delta\theta = 0$ do not yield linearly independent conditions for ξ_0 and η_0', we cannot effect a launch so that the radius and the mean angular rate coincide with the nominal ones (determined for a spherical earth) unless $\sin i = \sqrt{2/3}$. On the other hand, it turns out that we can preserve the nominal inclination of the orbit by choosing $\zeta_0 = 0$ and

$$\zeta_0' = (JR^2/2a)\sin 2i\cos\theta_0. \tag{115}$$

Now, if we designate $\Delta\bar{\xi}_2 = [\bar{\xi}_2]_{\theta=0}^{\theta=2\pi}$, we find for the nodal regression

$$\dot{\Omega} = \frac{-\Delta\bar{\xi}_2 k^{\frac{1}{2}}}{\sin i\, 2\pi a^{\frac{3}{2}}} = -nJ\left(\frac{R}{a}\right)^2 \cos i \tag{116}$$

and this agrees with the well-known result.

In the case of drag perturbations one replaces $\tilde{V}_\xi$, $\tilde{V}_\eta$, $\tilde{V}_\zeta$ of (110) by the appropriate components of (11):

$$\begin{aligned} F_\xi &= 0 \\ F_\eta &= -(C_D A\rho_0/2m)[(k/a)^{\frac{1}{2}} - \sigma a \cos i](k/a)^{\frac{1}{2}} \\ F_\zeta &= -(C_D A\rho_0/2m)\,\sigma \sin i(ka)^{\frac{1}{2}} \cos\theta \end{aligned} \tag{117}$$

and finds

$$\begin{aligned} \tilde{\xi}_2 &= (2a^3/k)F_\eta f \\ \tilde{\eta}_2 &= (a^3F_\eta/k)[4 - \tfrac{3}{2}f^2] \\ \tilde{\zeta}_2 &= (a^3F_\zeta/4k)[2f\tan\theta - \cos\theta]. \end{aligned} \tag{118}$$

Noting that

$$\frac{1}{2\pi a}[\tilde{\zeta}_2']_{\theta=0}^{\theta=2\pi} = \frac{di}{df} = -\frac{C_D A\rho_0 a^3\sigma \sin i}{4mk^{\frac{1}{2}}}, \tag{119}$$

we have agreement with standard results for the orbital precession due to diurnal winds.

If we extend this drag analysis to $0(e)$ we find

$$\begin{aligned} f_2 + f_3 &= (a^3F_\eta/2k)[\tfrac{5}{2}\sin f - f\cos f - 3f^2\sin f] \\ g_2 + g_3 &= (a^3F_\eta/2k)[6f\sin f + 9\cos f - 3f^2\cos f] \\ h_2 + h_3 &= [a^3F_\zeta/(4k\cos\theta)][\tfrac{5}{2}\cos(\bar{\omega}+2f) - \tfrac{11}{2}\cos\bar{\omega} - f\sin\bar{\omega} \\ &\quad - f\sin(\bar{\omega}+2f)], \end{aligned} \tag{120}$$

which are simple enough to permit a further extension to cases where $\rho_0 = \rho(f)$ is variable around the orbit. The details are straightforward.[19]

It is of course understood that any of these results should be accompanied by $\bar{\xi}_1\bar{\eta}_1\bar{\zeta}_1$ if a general solution is desired. This, however, adds nothing to the characteristics of a particular perturbation. The formulas (104) to (106) and (108) to (111) can also be applied to a variety of other effects such as luni-solar perturbations and radiation pressure.

The motivation behind the results of this section was to give a geometrically tangible account of perturbed satellite motions over a fractional period or just a few periods. This may be useful in various targeting, intercept, and rendezvous operations. On the other hand, the formulations of Sections III, 4.2, and 4.3 form the beginnings of ephemeris computing techniques and orbit lifetime studies. These subjects

have been pursued further in several higher-level methods (see Section I), some of which deal partly with the elements and partly with coordinates and make occasional use of contact transformations. They may be considered a stepping stone to full-fledged astronomical perturbation analyses, about which there exists an extensive literature.

VI. ACKNOWLEDGMENTS

The author wishes to acknowledge many helpful discussions with his colleagues, Messrs. A. J. Claus, A. G. Lubowe, and H. R. Westerman, especially in connection with Section III.

REFERENCES

1. Brouwer, D., and Clemence, G. M., *Methods of Celestial Mechanics*, Academic Press, New York, 1961.
2. Geyling, F. T., and Westerman, H. R., *Dynamics of Space Vehicles*, to be published.
3. Moulton, F. R., *An Introduction to Celestial Mechanics*, MacMillan, New York, 1914.
4. Smart, W. M., *Celestial Mechanics*, Longmans, Green and Co., New York, 1953.
5. Merson, R. H., The Motion of a Satellite in an Axi-Symmetric Gravitational Field, Geophysical Journal of the Roy. Astr. Soc., **4**, 1961, p. 17.
6. Danby, J. M. A., *Fundamentals of Celestial Mechanics*, MacMillan, New York, 1962, p. 238; and Lure, A. I., Equations of Disturbed Motion in the Kepler Problem, p. 288, of *Artificial Earth Satellites*, Vol. 4, ed. Kurnosova, L. V., Plenum Press, 1961.
7. Lidov, M. L., Evolution of the Orbits of Artificial Satellites of Planets as Affected by the Gravitational Perturbations from External Bodies, J. AIAA, Aug. 1963, p. 1985.
8. *Dynamics of Space Vehicles*, Ch. VI.
9. Claus, A. J., and Lubowe, A. G., A High-Accuracy Perturbation Method with Direct Application to Communication Satellite Orbit Prediction, Astronautica Acta, in preparation.
10. Blitzer, L., Weisfeld, M., and Wheelon, A. D., Perturbations of a Satellite Orbit Due to the Earth's Oblateness, J. Appl. Phys., **27**, Oct. 1956, p. 1141.
11. Moe, M. M., Solar-Lunar Perturbations of the Orbit of an Earth Satellite, J. ARS, **30**, No. 5, 1960, p. 485.
12. Lindstedt, A., Beitrag zur Integration der Differentialgleichungen der Störungstheorie, Abh. K. Akad. Wiss., St. Petersburg, **31**, No. 4, 1882.
13. Anthony, M. L., and Fosdick, G. E., An Analytical Study of the Effects of Oblateness on Satellite Orbits, Research Report, the Martin Company, Denver, Colo., April, 1960.
14. Fosdick, G. E., and Hewitt, M., Effects of the Earth's Oblateness and Atmosphere on a Satellite Orbit, Martin-Baltimore Engineering Report 8344, June, 1956.
15. Roberson, R. E., Orbital Behavior of Earth Satellites, (Parts I and II), J. Franklin Inst., **264**, Sept. and Oct., 1957.
16. Roberson, R. E., Oblateness Correction to Impact Points of Ballistic Rockets, J. Franklin Inst., Dec., 1958.
17. Roberson, R. E., Effect of Air Drag on Elliptic Satellite Orbits, ARS Paper 466-57, June, 1957.
18. Geyling, F. T., Satellite Perturbations from Extra-Terrestrial Gravitation and Radiation Pressure, J. Franklin Inst., **269**, No. 5, May, 1960, p. 375.
19. Geyling, F. T., Drag Displacements and Decay of Near-Circular Satellite Orbits, J. AIAA, in preparation.

Methods of Orbit Refinement

By R. B. BLACKMAN

(Manuscript received January 20, 1964)

During the past six or seven years, several methods of orbit refinement were developed specifically for use with artificial satellites and spacecraft. This article describes these methods, and the classical method, in a uniform mathematical formalism in order to facilitate comparisons of their relative advantages and disadvantages for practical systems applications. However, such comparisons are made in this article only to the extent that motivated the development of the new methods.

I. INTRODUCTION

The accuracy to which the position of a satellite or spacecraft can be predicted depends upon the accuracy to which the "initial" position and velocity vector, or related orbital parameters, are known. Since these parameters can be determined only from observational data which inevitably contain observational errors, the accuracy to which they can be known depends upon the nature, the quantity, the accuracy, and the distribution (in space and time) of the observational data, and the way in which these data are processed. The accuracy of the orbital parameters, and of prediction, depend also upon the accuracy to which all of the forces acting on the satellite or spacecraft are known and taken into account. Clearly, the term "accuracy" must be taken largely in a statistical sense.

Orbit refinement is essentially data smoothing for the purpose of accurate prediction. Given the nature of the observational data, and the statistical properties of the observational errors, it is possible to formulate a method of data smoothing and prediction which is optimum in the sense of giving predictions with the greatest possible accuracy. However, such an optimum method will, in general, be useful only as a standard of comparative performance for more practical methods. The reason for this is that it has not been difficult to find simpler and therefore more practical methods which are nearly as accurate as the optimum method. (For example, see Ref. 1.)

It should be noted also that in the practical applications of data-

smoothing and prediction methods each application is usually characterized by a unique set of practical constraints. Thus, it has been a frequent experience that a method which was judged to be the most practical for one application was usually not judged to be the most practical for another application. In fact, it has frequently happened that a new method, perhaps new only in the sense that it is a composite of parts of older methods, was developed for a particular system.

The classical method of orbit refinement, the so-called "differential corrections" method (which is essentially the method of least squares, developed by K. F. Gauss in 1795) has served astronomers very well for over 150 years. However, this method becomes quite unwieldy for large quantities of observational data. Hence, the quantity of data to be processed was frequently reduced by the supplementary use of "normal places" and/or "smoothing," described briefly in Section 6.6.1, pp. 141–142 of Ref. 2, and later in this article. Thus, up to about 1955 no need was felt for another method of orbit refinement, although the basis for the development of alternative methods was implicit in a number of articles published in the field of general statistical analysis, such as Refs. 3–6.

With the development of artificial satellites and space probes, the need for alternative methods of orbit refinement began to be felt in some quarters. The first definite proposal of an alternative method, as far as the author is aware, was made by P. Swerling (Refs. 7–9). A somewhat different method, independently developed by the author (Ref. 10) was used in the Telstar I experimental satellite communications system (Refs. 11, 12). Some difficulties experienced with this method after about four weeks of successful operation led A. J. Claus (Ref. 13) to develop another method which is slightly different from Swerling's method. In addition to these methods, it is worthwhile to include a method of space navigation described by R. H. Battin (Ref. 14) because it involves a practical detail which, under favorable circumstances, may be profitably introduced into the other methods.

The essential details of these methods will be described here in a uniform mathematical formalism in order to reveal their basic similarities and differences, and in order to facilitate comparisons of their relative advantages and disadvantages for practical systems applications.

II. CLASSICAL DIFFERENTIAL CORRECTIONS METHOD. LEAST SQUARES

Let $\tilde{\varphi}$ be an n-rowed vector representation of the observational (angular) data. Every component of this vector is assumed to be labeled to identify it as either a declination angle or a right ascension angle, and to

specify the time at which it was observed. Let ϵ be the 6-rowed vector representation of a set of values of the orbital elements, and let $\varphi(\epsilon)$ be an n-rowed vector representation of the angles which would have been observed, assuming that the actual orbital elements are exactly represented by ϵ, and assuming that observations are made with ideal accuracy. If the observational errors are independently and normally distributed, with zero means and equal variances, the best estimate of the orbital elements is that value of ϵ which minimizes the quadratic form

$$Q = [\tilde{\varphi} - \varphi(\epsilon)]' \cdot [\tilde{\varphi} - \varphi(\epsilon)],$$

where the prime stands for transposition. This quadratic form is simply the sum of the squares of the components of the vector difference $\tilde{\varphi} - \varphi(\epsilon)$.

Let ϵ_0 be an initially assumed value for ϵ, close to the true value. Then, to the first-order term in $(\epsilon - \epsilon_0)$,

$$\varphi(\epsilon) = \varphi(\epsilon_0) + J \cdot (\epsilon - \epsilon_0), \tag{1}$$

where J is the n by 6 matrix symbolized by

$$J = \partial\varphi(\epsilon_0)/\partial\epsilon_0 . \tag{2}$$

Hence, to second-order terms,

$$Q = [r - J \cdot (\epsilon - \epsilon_0)]' \cdot [r - J \cdot (\epsilon - \epsilon_0)],$$

where r is the n-rowed vector residual

$$r = \tilde{\varphi} - \varphi(\epsilon_0). \tag{3}$$

Now, Q is a minimum with respect to ϵ if

$$J' \cdot [r - J \cdot (\epsilon - \epsilon_0)] = 0.$$

Written in the form

$$J' \cdot J \cdot (\epsilon - \epsilon_0) = J' \cdot r,$$

this corresponds to the set of 6 equations commonly called "normal equations." If $J' \cdot J$ is nonsingular, and if the value of ϵ which satisfies this equation is denoted by $\bar{\epsilon}$, then,

$$\bar{\epsilon} = \epsilon_0 + (J' \cdot J)^{-1} \cdot J' \cdot r. \tag{4}$$

This $\bar{\epsilon}$ is then substituted for ϵ_0 in (2), in (3), and in the right-hand member of (4), in order to obtain another $\bar{\epsilon}$. This substitution procedure is iterated until $\bar{\epsilon}$ has essentially converged. The final $\bar{\epsilon}$ is the least squares estimate of the orbital elements.

In a statistical sense, this estimate is unbiased to the first order in the observational errors, assuming that the errors are not biased. It is only asymptotically unbiased to the second order, but it is a "consistent" estimate in the sense that the probability is unity that it will be correct to the second order as $n \to \infty$. (See Section 4.3 for clarification.)

2.1 *Classical Method. Weighted Least Squares*

The method described in the preceding section is quite unwieldy for large values of n on account of the size of the J matrix. Some relief was obtained by resorting to various artifices. For example, if a number of observations were made at sufficiently short intervals of time, a straight average of these observations would be taken and treated as a single observation called a "normal place." (More elaborate methods of deriving normal places directly from the observational data are called "smoothing" by Baker and Makemson in Ref. 2.) Since normal places would be more accurate than single actual observations, in proportion to the number of actual observations which went into each of them, it was necessary to generalize the differential corrections method to some extent.

The quadratic form to be minimized is now

$$Q = [\tilde{\varphi} - \varphi(\epsilon)]' \cdot W \cdot [\tilde{\varphi} - \varphi(\epsilon)],$$

where W is an n by n diagonal matrix. In expanded form, it is

$$Q = \sum w_{ii} \cdot [\tilde{\varphi}_i - \varphi_i(\epsilon)]^2,$$

where the w_{ii} are the components of W. Thus, the quadratic form is simply a weighted sum of the squares of the components of the vector difference $\tilde{\varphi} - \varphi(\epsilon)$.

The analysis in this case will not be pursued beyond this point, since it is a special case of the analysis given in the next section. Suffice it to say that if the analysis were carried out for this special case, the results would be equivalent to the method used by astronomers when they deal with normal places, or with uncorrelated observations of different degrees of accuracy.

2.2 *General Form of the Classical Method.*

If the observational data are not all of the same nature (angles, ranges, and range-rates), and especially if some of the errors in the data are correlated, let Φ be the n by n covariance matrix of the n-rowed vector $\tilde{\varphi}$. For further generality, let ϵ be an m-rowed vector representation of a set

of values of parameters which, in addition to the 6 orbital elements, may include such quantities as the frequency of a satellite-borne Doppler source, and instrumental biases. Then the quadratic form to be minimized is

$$Q = [\tilde{\varphi} - \varphi(\epsilon)]' \cdot \Phi^{-1} \cdot [\tilde{\varphi} - \varphi(\epsilon)]. \tag{5}$$

Under the assumption that the components of $\tilde{\varphi}$ obey a joint n-dimensional normal (i.e., Gaussian) distribution with covariance matrix Φ, the value of ϵ which minimizes the quadratic form (5) is the maximum likelihood estimate of the true value of the parameters. Under any other symmetrical probability distribution of the errors in the observational data, this value of ϵ is simply the weighted least squares estimate of the parameters.

Substituting (1) into (5) we get

$$Q = [r - J \cdot (\epsilon - \epsilon_0)]' \cdot \Phi^{-1} \cdot [r - J \cdot (\epsilon - \epsilon_0)], \tag{6}$$

where J and r are defined by (2) and (3) except that J is now an n by m matrix, and ϵ_0 is an m-rowed vector. Now, Q is minimum with respect to ϵ if

$$J' \cdot \Phi^{-1} \cdot [r - J \cdot (\epsilon - \epsilon_0)] = 0.$$

Denoting the value of ϵ which satisfies this equation by $\bar{\epsilon}$, we have

$$\bar{\epsilon} = \epsilon_0 + C \cdot \rho, \tag{7}$$

where

$$C = (J' \cdot \Phi^{-1} \cdot J)^{-1}, \qquad \text{an } m \text{ by } m \text{ matrix}, \tag{8}$$

$$\rho = J' \cdot \Phi^{-1} \cdot r, \qquad \text{an } m\text{-rowed vector}. \tag{9}$$

Equation (7) is the generalization of (4) and, as in the case of (4), it is to be used iteratively until $\bar{\epsilon}$ has essentially converged.

After $\bar{\epsilon}$ has converged, C is its covariance matrix. This follows from the fact that (6) may be expressed in the form

$$Q = (\epsilon - \bar{\epsilon})' \cdot C^{-1} \cdot (\epsilon - \bar{\epsilon}) + \text{terms independent of } \epsilon. \tag{10}$$

In case the data are all of the same nature (all angles, or all ranges, or all range-rates), are all of the same accuracy, and the errors are not correlated, then, $C = \sigma^2 \cdot (J' \cdot J)^{-1}$.

The availability of the covariance matrix C of the estimate $\bar{\epsilon}$ offers the possibility of using the classical method in the intrapass stage of the pass-by-pass method described in Section IV, in order to reduce the amount

of observational data to be processed at any one time. However, it is not the only way to reduce the amount of observational data to be processed at any one time. There are other ways which, in particular, avoid the computation of the n by m matrix J, where n may be of the order of 400 for each pass. One such way is Swerling's method described in the next section. Another such way is cited as an example in Section 4.3. A modified form of the second way is described briefly in Section V.

III. SWERLING'S METHOD

Let ϵ_1 be a 6-rowed vector estimate of the osculating orbital elements at epoch t_1, and let C_1 be its covariance matrix. Let $\bar{\varphi}$ be the n-rowed vector of new observational data more or less centered at epoch t_2, where $t_2 > t_1$, and let Φ be its covariance matrix. To obtain the least squares estimate of the osculating orbital elements at epoch t_2, we must first update (i.e., extrapolate) ϵ_1 and C_1. If $\hat{\epsilon}_1$ is the result of updating ϵ_1, the updated C_1 is

$$\hat{C}_1 = M \cdot C_1 \cdot M',$$

where M is the 6 by 6 matrix symbolized by

$$M = \partial \hat{\epsilon}_1 / \partial \epsilon_1 .$$

Then, assuming that the errors in the new data are not correlated with the errors in the old data, the quadratic form to be minimized is

$$Q = (\epsilon - \hat{\epsilon}_1)' \cdot \hat{C}_1^{-1} \cdot (\epsilon - \hat{\epsilon}_1) + [\bar{\varphi} - \varphi(\epsilon)]' \cdot \Phi^{-1} \cdot [\bar{\varphi} - \varphi(\epsilon)]. \quad (11)$$

This is essentially the sum of the right-hand members of (5) and (10). Now, to the first-order term in $(\epsilon - \hat{\epsilon}_1)$,

$$\varphi(\epsilon) = \varphi(\hat{\epsilon}_1) + J \cdot (\epsilon - \hat{\epsilon}_1), \quad (12)$$

where J is the n by 6 matrix symbolized by

$$J = \partial \varphi(\hat{\epsilon}_1) / \partial \hat{\epsilon}_1 . \quad (13)$$

Then, to second-order terms,

$$Q = (\epsilon - \hat{\epsilon}_1)' \cdot \hat{C}_1^{-1} \cdot (\epsilon - \hat{\epsilon}_1) + [r - J \cdot (\epsilon - \hat{\epsilon}_1)]' \cdot \Phi^{-1} \cdot [r - J \cdot (\epsilon - \hat{\epsilon}_1)],$$

where

$$r = \bar{\varphi} - \varphi(\hat{\epsilon}_1). \quad (14)$$

Now, Q is minimum with respect to ϵ if

$$\hat{C}_1^{-1} \cdot (\epsilon - \hat{\epsilon}_1) - J' \cdot \Phi^{-1} \cdot [r - J \cdot (\epsilon - \hat{\epsilon}_1)] = 0.$$

Denoting the value of ϵ which satisfies this equation by $\bar{\epsilon}$, we have

$$\bar{\epsilon} = \hat{\epsilon}_1 + C \cdot \rho, \tag{15}$$

where

$$C = [\hat{C}_1^{-1} + J' \cdot \Phi^{-1} \cdot J]^{-1}, \tag{16}$$

$$\rho = J' \cdot \Phi^{-1} \cdot r. \tag{17}$$

Finally, since the quadratic form, to second-order terms, may be expressed in the form

$$Q = (\epsilon - \bar{\epsilon})' \cdot C^{-1} \cdot (\epsilon - \bar{\epsilon}) + \text{terms independent of } \epsilon, \tag{18}$$

it follows that C is the covariance matrix of $\bar{\epsilon}$. Equations (11), (15), (16), and (17) correspond respectively to equations (16), (25), (19), and (23) in Swerling's JAS paper (Ref. 8). Further, since ϵ_1 and C_1 may have been computed at the preceding stage in exactly the same way as $\bar{\epsilon}$ and C at the last stage, (11) corresponds also to equation (30) in Swerling's JAS paper.

With regard to the updating of ϵ_1 and C_1 it should be noted that estimates of "time of perigee (or nodal) passage" should be labeled with the serial number of the passage. Thus, even in the hypothetical case of a purely Keplerian orbit, unless the serial number of the passage is intended to be the same in $\bar{\epsilon}$ as it is in ϵ_1, the vector $\hat{\epsilon}_1$ will differ from the vector ϵ_1. The component T_1 (time of perigee or nodal passage) of ϵ_1 will be increased to $\hat{T}_1$ in $\hat{\epsilon}_1$, where $\hat{T}_1$ is T_1 plus an integral multiple of the period estimate $2\pi a_1^{3/2}/\sqrt{\mu}$, where a_1 is the semimajor axis component of ϵ_1. The matrix M will therefore be a unity matrix except for an off-diagonal component $\partial \hat{T}_1/\partial a_1$ which is an integral multiple of $3\pi\sqrt{a_1/\mu}$. In this connection, A. J. Claus has found advantages in using the period instead of the semimajor axis as a component of ϵ_1, especially when perturbing forces are taken into account.

In the classical method n must be at least equal to the number of orbital elements, and it must include all of the available observational data — old observational data (which has been processed at least once before) as well as new. In Swerling's method n may be less than the number of orbital elements (possibly $n = 1$), and old observational data are represented by the 6 components of ϵ_1, the 21 distinct components of the symmetrical matrix C_1, and the epoch t_1. The chief objection to Swerling's method, as far at least as some applications are concerned, is its inability to omit any part of the old observational data without reprocessing the remainder of it. This objection is less serious if Swerling's

method is used only in the intrapass stage of the pass-by-pass method described in the next section.

IV. PASS-BY-PASS METHOD

This method consists essentially of two stages per pass. (These stages are not quite the same as Swerling's stages.) In the first (or *intrapass*) stage, a set of estimates of the orbital elements, and an associated covariance matrix, are computed from the observational data *for each pass.* In this stage, any method of computation which yields a covariance matrix for the estimates of the orbital elements may be used. In the second (or *interpass*) stage, the sets of single-pass (i.e., *intrapass*) estimates of the orbital elements are combined cumulatively (and possibly selectively), by a method in which the single-pass (i.e., *intrapass*) covariance matrices play important roles.

Let $\hat{\epsilon}_1$ and $\hat{C}_1$ have the same significance as in the description of Swerling's method, but let the new data be processed separately to obtain an independent vector estimate ϵ_2 of the osculating orbital elements at epoch t_2, with covariance matrix C_2. Then, the quadratic form to be minimized is

$$Q = (\epsilon - \hat{\epsilon}_1)' \cdot \hat{C}_1^{-1} \cdot (\epsilon - \hat{\epsilon}_1) + (\epsilon - \epsilon_2)' \cdot C_2^{-1} \cdot (\epsilon - \epsilon_2). \quad (19)$$

This is the sum of two terms similar to the right-hand member of (10). Now, Q is minimum with respect to ϵ if

$$\hat{C}_1^{-1} \cdot (\epsilon - \hat{\epsilon}_1) + C_2^{-1} \cdot (\epsilon - \epsilon_2) = 0.$$

Denoting the value of ϵ which satisfies this equation by $\bar{\epsilon}$, we have

$$\bar{\epsilon} = C \cdot (\hat{C}_1^{-1} \cdot \hat{\epsilon}_1 + C_2^{-1} \cdot \epsilon_2) \quad (20)$$

where

$$C = (\hat{C}_1^{-1} + C_2^{-1})^{-1}. \quad (21)$$

Finally, since (19) may be expressed in the form

$$Q = (\epsilon - \bar{\epsilon})' \cdot C^{-1} \cdot (\epsilon - \bar{\epsilon}) + \text{terms independent of } \epsilon, \quad (22)$$

it follows that C is the covariance matrix of $\bar{\epsilon}$. (See Appendix for a more illuminating derivation).

Note that a "fading memory" can be introduced into the pass-by-pass method (or any other method in which the old data are represented by ϵ_1, C_1, and t_1) simply by substituting kC_1 for C_1, where k is a scalar constant ($k > 1$) or

$$k = \exp [\gamma \cdot (t_2 - t_1)]$$

where γ is a constant ($\gamma > 0$). However, at the cost of storing a number of single-pass estimates of the orbital elements and the associated covariance matrices, old single-pass estimates of the orbital elements may be completely omitted at any time without having to reprocess any of the observational data on which the more recent single-pass estimates are based.

4.1 *An Example*

It is illuminating to see how the method of combination described in the preceding section works out in a simple problem. Consider an object traveling along the x axis at a known acceleration a. Let

$$\epsilon_1 = \begin{bmatrix} x_1 \\ v_1 \end{bmatrix}, \qquad \epsilon_2 = \begin{bmatrix} x_2 \\ v_2 \end{bmatrix},$$

be the estimates of position and velocity at epochs t_1 and t_2, and let

$$C_1 = C_2 = \begin{bmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_v^2 \end{bmatrix}.$$

If $t_2 - t_1 = \tau$, then

$$\hat{\epsilon}_1 = \begin{bmatrix} x_1 + v_1\tau + \frac{1}{2}a\tau^2 \\ v_1 + a\tau \end{bmatrix},$$

whence

$$M = \frac{\partial \hat{\epsilon}_1}{\partial \epsilon_1} = \begin{bmatrix} 1 & \tau \\ 0 & 1 \end{bmatrix}.$$

Then,

$$\hat{C}_1 = M \cdot C_1 \cdot M' = \begin{bmatrix} \sigma_x^2 + \tau^2\sigma_v^2 & \tau\sigma_v^2 \\ \tau\sigma_v^2 & \sigma_v^2 \end{bmatrix}.$$

Note that the correlation coefficient, which is

$$\frac{1}{\sqrt{1 + (\sigma_x/\tau\sigma_v)^2}},$$

is very close to unity for large values of τ. For this reason, or for other reasons, we may expect serious difficulties in numerical computations based on (20) and (21) as they stand. These difficulties are considerably reduced by the transformations described in the next section.

In the example under consideration there is of course no difficulty about inverting matrices analytically. It is found finally that

$$\bar{\epsilon} = \begin{bmatrix} \bar{x} \\ \bar{v} \end{bmatrix} \text{ at epoch } t_2 ,$$

where

$$\bar{x} = \frac{1}{4\sigma_x^2 + \tau^2\sigma_v^2} \{\sigma_x^2[2(x_1 + x_2) + \tau(v_1 + v_2)] + \sigma_v^2\tau^2 x_2\} ,$$

$$\bar{v} = \frac{1}{4\sigma_x^2 + \tau^2\sigma_v^2} \{2\sigma_x^2(v_1 + v_2 + a\tau) + \sigma_v^2\tau(x_2 - x_1 + \tfrac{1}{2}a\tau^2)\} ,$$

and

$$C = \frac{\sigma_x^2}{4\sigma_x^2 + \tau^2\sigma_v^2} \begin{bmatrix} 2\sigma_x^2 + \tau^2\sigma_v^2 & \tau\sigma_v^2 \\ \tau\sigma_v^2 & 2\sigma_v^2 \end{bmatrix} .$$

Under the assumptions implicit in C_1 and C_2, that x_1, v_1, x_2, and v_2 are not correlated, the equations for $\bar{x}$ and $\bar{v}$ may be verified by a more familiar method. We have in fact two independent estimates of the position at epoch t_2, viz.,

$$x_1 + \frac{v_1 + v_2}{2}\tau \quad \text{with variance } \sigma_x^2 + \frac{\tau^2\sigma_v^2}{2} ,$$

and

$$x_2 \quad \text{with variance } \sigma_x^2 .$$

Taking the weighted average of these two estimates, each estimate being weighted in inverse proportion to its variance, we get $\bar{x}$, as expressed above. The variance of $\bar{x}$ is the harmonic mean of the two variances, so that

$$\text{var } \{\bar{x}\} = \frac{\sigma_x^2(2\sigma_x^2 + \tau^2\sigma_v^2)}{4\sigma_x^2 + \tau^2\sigma_v^2} ,$$

in agreement with the equation for C.

Similarly, we have three independent estimates of the velocity at epoch t_2, viz.,

$$v_1 + a\tau \quad \text{with variance } \sigma_v^2 ,$$

$$v_2 \quad \text{with variance } \sigma_v^2 ,$$

and

$$\frac{x_2 - x_1}{\tau} + \frac{1}{2}a\tau \quad \text{with variance } \frac{2\sigma_x^2}{\tau^2} .$$

Taking the weighted average of these three estimates, each estimate being weighted in inverse proportion to its variance, we get $\bar{v}$, as expressed above. The variance of $\bar{v}$ is the harmonic mean of the three variances, so that

$$\text{var}\,\{\bar{v}\} = 2\sigma_x^2\sigma_v^2/(4\sigma_x^2 + \tau^2\sigma_v^2),$$

in agreement with the equation for C.

The covariance of $\bar{x}$ and $\bar{v}$ may also be verified, but the algebra is more involved.

4.2 *Transformation of Equations*

Omitting the explicit notational reference to updating (i.e., extrapolation) in (20) and (21), we have

$$\bar{\epsilon} = C\cdot(C_1^{-1}\cdot\epsilon_1 + C_2^{-1}\cdot\epsilon_2), \tag{23}$$

where

$$C = (C_1^{-1} + C_2^{-1})^{-1}. \tag{24}$$

These equations require the inversion of three matrices which, at least in the case of angles-only data, are usually extremely ill-conditioned. Hence, they have been transformed in order to reduce this difficulty. The transformations described here were developed by A. J. Claus and the author.

Introducing the identity

$$C_1^{-1}\cdot\epsilon_1 = (C_1^{-1} + C_2^{-1})\cdot\epsilon_1 - C_2^{-1}\cdot\epsilon_1$$

into (23), we get

$$\bar{\epsilon} = \epsilon_1 - C\cdot C_2^{-1}\cdot(\epsilon_1 - \epsilon_2).$$

Now,

$$\begin{aligned} C\cdot C_2^{-1} &= (C_1^{-1} + C_2^{-1})^{-1}\cdot C_2^{-1} = [C_2\cdot(C_1^{-1} + C_2^{-1})]^{-1} \\ &= (1 + C_2\cdot C_1^{-1})^{-1} = [(C_1 + C_2)\cdot C_1^{-1}]^{-1} \\ &= C_1\cdot(C_1 + C_2)^{-1}. \end{aligned}$$

Next, let

$$P_1 = S\cdot C_1\cdot S, \qquad P_2 = S\cdot C_2\cdot S, \tag{25}$$

where S is a diagonal matrix in which each diagonal term is the reciprocal of the square root of the sum of the corresponding diagonal terms of

C_1 and C_2. (Thus, every diagonal term of $P_1 + P_2$ is unity.) Then,

$$C_1 \cdot (C_1 + C_2)^{-1} = S^{-1} \cdot P_1 \cdot (P_1 + P_2)^{-1} \cdot S.$$

Hence,

$$\bar{\epsilon} = \epsilon_1 - S^{-1} \cdot P_1 \cdot (P_1 + P_2)^{-1} \cdot S \cdot (\epsilon_1 - \epsilon_2).$$

Similarly,

$$\bar{\epsilon} = \epsilon_2 + S^{-1} \cdot P_2 \cdot (P_1 + P_2)^{-1} \cdot S \cdot (\epsilon_1 - \epsilon_2).$$

Next, let W_1 and W_2 be arbitrary square matrices whose sum is a unity matrix. Then,

$$\bar{\epsilon} = W_1 \cdot \epsilon_1 + W_2 \cdot \epsilon_2 - R \cdot (\epsilon_1 - \epsilon_2),$$

where

$$R = (W_1 \cdot S^{-1} \cdot P_1 - W_2 \cdot S^{-1} \cdot P_2) \cdot (P_1 + P_2)^{-1} \cdot S.$$

Since $P_1 + P_2$ may yet be ill conditioned for inversion, we now use a well-known artifice, and write

$$R = [W_1 \cdot S^{-1} \cdot (P_1 \cdot G) - W_2 \cdot S^{-1} \cdot (P_2 \cdot G)] \cdot [(P_1 + P_2) \cdot G]^{-1} \cdot S,$$

where G may be regarded as another arbitrary square matrix although it merely represents a set of rules for combining rows and/or columns of $P_1 + P_2$, as well as of P_1 and P_2 individually. [The normalization of the matrix sum $C_1 + C_2$ to $P_1 + P_2$ and the preservation of its symmetry by the introduction of the matrix S, as in (25), simplifies the implementation of the matrix G as a set of operational rules.] Finally, W_1 and W_2 are restricted to diagonal matrices, so that

$$R = S^{-1} \cdot [W_1 \cdot (P_1 \cdot G) - W_2 \cdot (P_2 \cdot G)] \cdot [(P_1 + P_2) \cdot G]^{-1} \cdot S. \quad (27)$$

Noting that the right-hand member of (23) reduces to $2C$ if ϵ_1 is replaced by C_1 and ϵ_2 is replaced by C_2, it follows from (26) that

$$C = \tfrac{1}{2}[W_1 \cdot C_1 + W_2 \cdot C_2 - R \cdot (C_1 - C_2)]. \quad (28)$$

For further details, see Ref. 12.

4.3 *Debiasing Single-Pass Estimates*

Depending upon the method used to obtain single-pass estimates of the orbital elements in the first (or intrapass) stage of the pass-by-pass method of orbit refinement described in Section IV, the single-pass estimates may be biased on the average even if the errors in the observa-

tional data are not biased. Unless these biases are tolerable, they must of course be removed. In this section we will describe a method of removing these biases in the single-pass estimates of the orbital elements. However, biases in the single-pass estimates of the orbital elements due to biased errors in the observational data will not be removed by this method.

In the interest of simplicity, the description will be by analogy with a much simpler problem which involves only one observable coordinate, one parameter to be estimated, and does not involve time at all. We will consider two methods of estimation, of which the first is analogous to the classical method of orbit refinement, and the second is analogous to any method of obtaining single-pass estimates which are biased on the average even if the errors in the observational data are not biased.

Let y be a function of the observable coordinate x, and let

$$x_i = x_0 + \epsilon_i \qquad (i = 1, 2, \cdots, n)$$

be the observed values of x, where the ϵ's are uncorrelated random errors with ave $\{\epsilon_i\} = 0$ and var $\{\epsilon_i\} = \sigma^2$ for every i. The most direct way of estimating $y_0 \equiv y(x_0)$ is obviously to compute first

$$\bar{x} = (1/n)\sum x_i ,$$

and then $y(\bar{x})$. The nature of the estimate obtained in this way is determined as follows. Since

$$\bar{x} = x_0 + \alpha,$$

where

$$\alpha = (1/n)\sum \epsilon_i ,$$

then, to second-order terms in the ϵ_i's,

$$y(\bar{x}) = y_0 + a_0\alpha + \tfrac{1}{2}b_0\alpha^2,$$

where $a_0 = dy_0/dx_0$, and $b_0 = d^2y_0/dx_0^2$. Now,

$$\text{ave } \{\alpha\} = 0, \qquad \text{ave } \{\alpha^2\} = \sigma^2/n.$$

Hence, to second-order terms in σ,

$$\text{ave } \{y(\bar{x})\} = y_0 + (b_0\sigma^2/2n),$$

and

$$\text{var } \{y(\bar{x})\} = \text{ave } \{[y(\bar{x})]^2\} - [\text{ave } \{y(\bar{x})\}]^2 = a_0^2\sigma^2/n.$$

If σ^2 is not known, an estimate of var $\{y(\bar{x})\}$ is given by

$$\text{var } \{y(\bar{x})\} \approx \frac{a_0^2}{n(n-1)} \sum (x_i - \bar{x})^2 .$$

Since

$$\lim_{n \to \infty} \text{ave } \{y(\bar{x})\} = y_0 ,$$

the estimate is asymptotically unbiased; and since

$$\lim_{n \to \infty} \text{var } \{y(\bar{x})\} = 0,$$

the estimate is "consistent" in the sense that the probability is unity that it will be correct to at least the second order as $n \to \infty$. These results are indicative of the nature of the estimates of orbital elements obtained by the classical method of orbit refinement, in which the estimates are such as to "predict" values of the observable coordinates which agree with the actual observations in the least squares sense.

Now, consider another way of estimating y_0. We compute $y(x)$ for each observed value of x, and define the estimate of y_0 as

$$\tilde{y}_0 = (1/n) \sum y_i \quad \text{where} \quad y_i = y(x_i).$$

The purpose of estimating y_0 in this way is to permit the estimation of var $\{\tilde{y}_0\}$ without using a_0. Thus,

$$\text{var } \{\tilde{y}_0\} \approx \frac{1}{n(n-1)} \sum R_i^2$$

where

$$R_i = y_i - \tilde{y}_0 .$$

The importance of this is that the analog of a_0 in the computation of orbital elements from, say, two complete radar fixes (each fix consisting of a range and two angles) is the inverse of a 6 by 6 matrix whose components are functions of the epochs of the two fixes. The computation of this matrix may be avoided by computing a set of orbital elements from each pair of radar fixes, averaging over the sets, and *estimating* the covariance matrix from the residuals.

The nature of the estimate $\tilde{y}_0$ is determined as follows. Since, to second-order terms,

$$y_i = y_0 + a_0 \epsilon_i + \tfrac{1}{2} b_0 \epsilon_i^2 ,$$

then

$$\tilde{y}_0 = y_0 + a_0\alpha + \tfrac{1}{2}b_0\beta,$$

where α is as previously defined, and

$$\beta = (1/n)\sum \epsilon_i^2.$$

Now,

$$\text{ave } \{\beta\} = \sigma^2.$$

Hence,

$$\text{ave } \{\tilde{y}_0\} = y_0 + (b_0\sigma^2/2),$$

and

$$\text{var } \{\tilde{y}_0\} = \text{ave } \{[\tilde{y}_0]^2\} - [\text{ave } \{\tilde{y}_0\}]^2 = a_0^2\sigma^2/n.$$

Thus, $\tilde{y}_0$ is a biased estimate of y_0. In particular, it should be noted that, while the variance of this estimate decreases with increasing n, the bias in the estimate is independent of n. Hence, the variance bears no relation to the accuracy of the estimate.

The bias may be removed by the following supplementary procedure.

1. Compute $\hat{x}_0$ such that

$$y(\hat{x}_0) = \tilde{y}_0.$$

This is analogous to computing artificial tracking data (at the same epochs as the actual tracking data) from the biased single-pass estimates of the orbital elements analogous to $\tilde{y}_0$. The method of computing tracking data from orbital elements must of course be numerically compatible with the method of computing orbital elements from tracking data in the absence of observational errors.

2. Compute

$$\hat{x}_i = 2\hat{x}_0 - x_i, \qquad i = 1, 2, \cdots, n.$$

This is analogous to combining the actual tracking data with the artificial tracking data computed in the preceding step. (The choice of a combination such that the random error in each $\hat{x}_i$ is equal in magnitude but opposite in sign to the random error in the corresponding x_i was suggested by D. R. Brillinger.)

3. Compute

$$\hat{y}_i = y(\hat{x}_i), \qquad i = 1, 2, \cdots, n.$$

This is exactly the same procedure used in computing $y_i = y(x_i)$.

4. Compute

$$\bar{y}_i = \tfrac{1}{2}(3y_i - \hat{y}_i), \qquad i = 1, 2, \cdots, n.$$

5. Compute

$$\bar{y}_0 = (1/n)\sum \bar{y}_i \quad \text{as the estimate of } y_0 .$$

The nature of the estimate $\bar{y}_0$ is determined as follows. To second-order terms,

$$\hat{x}_0 = x_0 + \alpha - \frac{b_0}{2a_0}(\alpha^2 - \beta) .$$

Hence,

$$\hat{x}_i = x_0 - (\epsilon_i - 2\alpha) - \frac{b_0}{a_0}(\alpha^2 - \beta), \qquad i = 1, 2, \cdots, n,$$

$$\hat{y}_i = y_0 - a_0(\epsilon_i - 2\alpha) + \frac{b_0}{2}(\epsilon_i^2 - 4\alpha\epsilon_i + 2\alpha^2 + 2\beta),$$

$$\bar{y}_i = y_0 + a_0(2\epsilon_i - \alpha) + \frac{b_0}{2}(\epsilon_i^2 + 2\alpha\epsilon_i - \alpha^2 - \beta),$$

$$\bar{y}_0 = y_0 + a_0\alpha + \tfrac{1}{2}b_0\alpha^2 .$$

Thus, to second-order terms in the ϵ_i's, $\bar{y}_0$ is the same as $y(\bar{x})$. It is asymptotically unbiased, and it is a consistent estimate of y_0. It may be absolutely debiased, to second-order terms in the ϵ_i's, by changing step 4 of the supplementary procedure to compute

$$\bar{y}_i = \frac{1}{2}\left[\left(3 + \frac{1}{n-1}\right)y_i - \left(1 + \frac{1}{n-1}\right)\hat{y}_i\right].$$

Then, in determining the nature of the estimate $\bar{y}_0$, we now have

$$\bar{y}_i = y_0 + \frac{a_0}{n-1}[(2n-1)\,\epsilon_i - n\alpha]$$

$$+ \frac{b_0}{2(n-1)}[(n-1)\,\epsilon_i^2 + 2n\alpha\epsilon_i - n\alpha^2 - n\beta],$$

$$\bar{y}_0 = y_0 + a_0\alpha + \frac{b_0}{2(n-1)}(n\alpha^2 - \beta).$$

Hence,

$$\text{ave}\,\{\bar{y}_0\} = y_0 .$$

Generally, if step 2 of the supplementary procedure is changed to compute

$$\hat{x}_i = \tfrac{1}{2}[(1 + w)\hat{x}_0 + (1 - w)x_i],$$

and step 4 is changed to compute

$$\bar{y}_i = \tfrac{1}{2}[(1 + W)y_i + (1 - W)\hat{y}_i],$$

then,

$$\bar{y}_0 = y_0 + a_0\alpha + \frac{(1 - W)b_0}{4}\left\{\frac{1 - w^2}{4}\alpha^2 + \left[\frac{3 + w^2}{4} + \frac{1 + W}{1 - W}\right]\beta\right\},$$

whence,

$$\text{ave } \{\bar{y}_0\} = y_0 + \frac{(1 - W)b_0\sigma^2}{4}\left[\frac{1 - w^2}{4n} + \frac{3 + w^2}{4} + \frac{1 + W}{1 - W}\right].$$

Hence, the estimate $\bar{y}_0$ is asymptotically unbiased if

$$W = \frac{w^2 + 7}{w^2 - 1},$$

and is absolutely unbiased if

$$W = \frac{w^2 + \dfrac{7n + 1}{n - 1}}{w^2 - 1}.$$

The choice of w should be made with some regard to the fact that

$$\text{var } \{\hat{x}_i\} = \sigma^2\left[1 - \frac{(n - 1)\ (1 + w)\ (3 - w)}{4n}\right].$$

The choice is $w = 3$ in step 2 of the supplementary procedure, and $W = 2$ in step 4, so that var $\{\hat{x}_i\} = \sigma^2 =$ var $\{x_i\}$.

V. CLAUS'S METHOD

The pass-by-pass method described in Section IV was used at the Andover, Maine, station of the Telstar I experimental satellite communications system. The intrapass estimates of the orbital elements were computed from angles-only data by a method described in some detail in Refs. 11 and 12. Suffice it here to say that a set of orbital elements is computed from each set of four sightlines (of which there may be as many as 200 in a single pass), the sets of orbital elements are averaged, and an *estimate* of the covariance matrix is computed from the residuals.

(Compare this outline with that of the example cited in Section 4.3, which uses complete radar fixes, including range. It should be noted that unless the number of sets of orbital elements is at least equal to the number of orbital elements the covariance matrix will be singular.) This method gave excellent results for about four weeks. After that period it began to give sporadically bad results — typical errors of 10^5 feet in single-pass estimates of the semimajor axis.

Extensive simulation studies by W. C. Ridgway III showed that most of the trouble was probably due to the sensitivity of single-pass estimates of the orbital elements to a bias error in the elevation angles, where these estimates are derived from single-tracker angles-only data. This result was subsequently confirmed by some formal analysis by the author. Ridgway's studies showed, in particular, that the sensitivity increases rapidly with decreasing length of pass, and with increasing maximum elevation angle, although sightlines at elevation angles over 82.5 degrees (or under 7.5 degrees) were not used. This is consistent with the fact that the sporadically bad results began to occur when the perigee of Telstar I had precessed sufficiently to make the passes at Andover substantially shorter than they were during the first week, and a substantially larger proportion of the passes had high maximum elevation angles. Ridgway's studies showed that, under these conditions, an error of 10^5 feet in the single-pass estimate of the semimajor axis could easily be due to a bias of 0.01 degree in the elevation angles.

In order to overcome the sensitivity of a single-pass single-tracker angles-only method to a bias error in the elevation angles, on short passes with high maximum elevation angles, A. J. Claus has developed a method of orbit refinement which permits the use of a few, perhaps only one or two, measurements of range in each pass. This method was intended to be used in the intrapass stage of the pass-by-pass method described in Section IV, but it may be used as a self-sufficient method, just as Swerling's method may be used either in the intrapass stage of the pass-by-pass method or as a self-sufficient method.

Although Claus developed his method with no foreknowledge of Swerling's method, it turns out that his method differs from Swerling's method only in the explicit introduction of an iterative routine which, as in the classical method, improves the estimates of the orbital elements.

Instead of (12) we now substitute (1) into (11) so that

$$Q = (\epsilon - \hat{\epsilon}_1)' \cdot \hat{C}_1^{-1} \cdot (\epsilon - \hat{\epsilon}_1) + [r - J \cdot (\epsilon - \epsilon_0)]' \cdot \Phi^{-1} \cdot [r - J \cdot (\epsilon - \epsilon_0)], \tag{29}$$

where J and r are defined by (2) and (3), and ϵ_0 has the same signifi-

cance as in the classical method. Now, Q is minimum with respect to ϵ if

$$\hat{C}_1^{-1} \cdot (\epsilon - \hat{\epsilon}_1) - J' \cdot \Phi^{-1} \cdot [r - J \cdot (\epsilon - \epsilon_0)] = 0.$$

Denoting the value of ϵ which satisfies this equation by $\bar{\epsilon}$, we have

$$\bar{\epsilon} = C \cdot [\hat{C}_1^{-1} \cdot \hat{\epsilon}_1 + J' \cdot \Phi^{-1} \cdot J \cdot \epsilon_0 + \rho], \tag{30}$$

where

$$C = [\hat{C}_1^{-1} + J' \cdot \Phi^{-1} \cdot J]^{-1}, \tag{31}$$

$$\rho = J' \cdot \Phi^{-1} \cdot r. \tag{32}$$

Equation (30) may be expressed in the form

$$\bar{\epsilon} = \hat{\epsilon}_1 + C \cdot [J' \cdot \Phi^{-1} \cdot J \cdot (\epsilon_0 - \hat{\epsilon}_1) + \rho], \tag{33}$$

where C, J, and ρ (the latter through r as well as J) depend implicitly on ϵ_0. Comparing this with (15), it will be seen that the iterative use of (30) or (33), substituting $\bar{\epsilon}$ for ϵ_0 at each iteration, until the difference between ϵ_0 and $\bar{\epsilon}$ is negligible, is precluded in Swerling's method by the constraint $\epsilon_0 \equiv \hat{\epsilon}_1$.

Finally, since (29) may be expressed in the form

$$Q = (\epsilon - \bar{\epsilon})' \cdot C^{-1} \cdot (\epsilon - \bar{\epsilon}) + \text{terms independent of } \epsilon, \tag{34}$$

it follows that C is the covariance matrix of $\bar{\epsilon}$.

VI. BATTIN'S METHOD

This method, as described in Ref. 14, is essentially a special case of Swerling's method. However, by introducing new data one at a time, Battin's method avoids the inversion of matrices. (Swerling's JAS article contains equations whereby matrix inversions are avoided if new data are introduced one at a time, but these equations appear at the end of the section entitled "Statistics of Propagated Errors," and their use for avoiding matrix inversions is not explicitly stated.)

In the notation of Section III, (15) may be written in the form

$$\bar{\epsilon} = \hat{\epsilon}_1 + [J' \cdot J + \sigma^2 \hat{C}_1^{-1}]^{-1} \cdot J' \cdot r, \tag{35}$$

where σ^2 is the variance of the scalar $\bar{\varphi}$, J is a 1 by 6 matrix (i.e., J' is a 6-rowed vector), and r is a scalar. Now, if

$$a = J \cdot \hat{C}_1 \cdot J' + \sigma^2 \quad \text{(a scalar)}, \tag{36}$$

then,

$$J' = J' \cdot \frac{J \cdot \hat{C}_1 \cdot J' + \sigma^2}{a},$$

$$= \frac{1}{a}[J' \cdot J \cdot \hat{C}_1 + \sigma^2] \cdot J',$$

$$= \frac{1}{a}\left[J' \cdot J + \sigma^2 \hat{C}_1^{-1}\right] \cdot \hat{C}_1 \cdot J'. \tag{37}$$

Substituting (37) for the last J' in (35), we get

$$\bar{\epsilon} = \hat{\epsilon}_1 + \frac{r}{a} \hat{C}_1 \cdot J'. \tag{38}$$

Further, since (16) may be written in the form

$$C = \sigma^2[J' \cdot J + \sigma^2 \hat{C}_1^{-1}]^{-1},$$

then,

$$C - \hat{C}_1 = \sigma^2[J' \cdot J + \sigma^2 \hat{C}_1^{-1}]^{-1} - \hat{C}_1,$$

and, therefore,

$$[J' \cdot J + \sigma^2 \hat{C}_1^{-1}] \cdot (C - \hat{C}_1) = -J' \cdot J \cdot \hat{C}_1.$$

Substituting (37) for J' in the right-hand member, we get

$$C = \hat{C}_1 - \frac{1}{a} \hat{C}_1 \cdot J' \cdot J \cdot \hat{C}_1. \tag{39}$$

Equations (36), (38), and (39) are equivalent to equations (30), (29), and (33) in Battin's paper (Ref. 14).

From this description of Battin's method, it is evident that the inversion of matrices may be avoided also in Claus's method if new data (perhaps only the range data) are introduced one at a time. Comparing (33), (31), and (32) with (15), (16), and (17), it is clear that (38) and (39), with (36), are valid in Claus's method if

$$r = \tilde{\varphi} - \varphi(\epsilon_0) + J \cdot (\epsilon_0 - \hat{\epsilon}_1), \tag{40}$$

and

$$J = \partial\varphi(\epsilon_0)/\partial\epsilon_0. \tag{41}$$

The initial value of ϵ_0 may be taken equal to $\hat{\epsilon}_1$, but thereafter $\bar{\epsilon}$ is repeatedly substituted for ϵ_0 until the difference between $\bar{\epsilon}$ and ϵ_0 is negligible.

VII. MOD-S AND MOD-C METHODS

Swerling's method and Claus's method can be modified to permit the introduction of new data n at a time, where $n < 6$, at the cost of having to invert a matrix of order n. This might be worthwhile for $n = 2$ or 3.

With regard to Swerling's method, let

$$A = J \cdot \hat{C}_1 \cdot J' + \Phi \qquad (\text{an } n \text{ by } n \text{ matrix}). \tag{42}$$

Then,

$$\begin{aligned} J' \cdot \Phi^{-1} &= J' \cdot \Phi^{-1} \cdot [J \cdot \hat{C}_1 \cdot J' + \Phi] \cdot A^{-1} \\ &= [J' \cdot \Phi^{-1} \cdot J + \hat{C}_1^{-1}] \cdot \hat{C}_1 \cdot J' \cdot A^{-1}. \end{aligned} \tag{43}$$

Substituting this into (17), substituting the resultant expression for ρ into (15), and taking account of (16), we get

$$\bar{\epsilon} = \dot{\epsilon}_1 + \hat{C}_1 \cdot J' \cdot A^{-1} \cdot r. \tag{44}$$

Equation (36) is a special case of (42), (38) is a special case of (44), and (39) is a special case of

$$C = \hat{C}_1 - \hat{C}_1 \cdot J' \cdot A^{-1} \cdot J \cdot \hat{C}_1 . \tag{45}$$

With regard to Claus's method, (42), (44), and (45) are valid if r and J are defined by (40) and (41), and the repeated substitution of $\bar{\epsilon}$ for ϵ_0 is carried out until the difference between ϵ_0 and $\bar{\epsilon}$ is negligible.

VIII. SUMMARY AND CLOSING REMARKS

8.1 *Summary*

The classical "differential corrections" method of orbit refinement, occasionally supplemented by the use of "normal places," is appropriate for astronomical bodies whose relative positions change very slowly, whose relative angular positions, viewed from the earth, can be measured with extreme optical precision, and which therefore require comparatively small quantities of observational data to establish their orbits with great accuracy. However, that method is very unwieldy for artificial earth satellites or short-range space probes, where the relative inaccuracy of the observational data must be offset by greater quantities of observational data.

For artificial earth satellites or short-range space probes, Swerling's method is more practical, chiefly because it does not require all of the observational data to be processed together. However, to omit any part

of the old observational data which has been processed by Swerling's method it is necessary to start all over again in the sense that all of the old observational data which are to be retained must be reprocessed in the same way, along with any new observational data which might be available.

The pass-by-pass method, on the other hand, operates on the principle of computing an independent set of estimates of the orbital elements for each pass, and combining the sets of single-pass estimates in an optimum way. Thus, entire blocks of observational data may be omitted without actually reprocessing any of the old observational data from which single-pass estimates of the orbital elements have already been computed. The method of obtaining the single-pass (intrapass) estimates is optional. In the Telstar I experiments, it consisted in dividing up the data (no range data) into mutually interlaced independent sets of four sightlines, computing a set of orbital elements from each set of four sightlines, averaging over the sets of orbital elements, and computing a covariance matrix for the average set, from the residuals. This particular intrapass method introduces biases (apart from the biases in the data) and a method of eliminating these computational biases was developed but was not used in the Telstar I experiments. After four weeks of successful operation, a more serious source of trouble arose, which was traced to the increasing sensitivity to bias (residual boresight error and sample bias) in the elevation angle data. This increasing sensitivity was associated with the precession of perigee to latitudes close to that of the tracker.

To overcome the sensitivity of the angles-only intrapass method to bias in the elevation angle data, Claus developed a method, intended to be used chiefly in the intrapass stage of the pass-by-pass method for Telstar II, which could accept occasional range data. This method is essentially Swerling's method supplemented by an iterative routine which improves its accuracy.

In Swerling's or Claus's method, the inversion of six-by-six or higher-order matrices can be avoided by borrowing a detail from Battin's method of spacecraft navigation, provided that the observational data are processed only one at a time, as in Battin's method. However, at the cost of inverting n-by-n matrices, where $n < 6$ (say, 2 or 3), the observational data may be processed n at a time, by modified forms of Swerling's or Claus's method.

The comparisons of the methods described in this paper have been made only to the extent that motivated the development of the newer methods. The practical details of these methods are interchangeable to

a large extent, so that different and more appropriate combinations of these details may be made for other specific practical applications.

8.2 *Closing Remarks*

Claus has pointed out that, in principle, none of the alternative methods described in this paper, including the use of normal places in the classical method, can be as efficient, in a statistical average sense, as the classical method without normal places. The reason for this is simply that the classical method without normal places allows the maximum possible freedom in fitting the estimates of the orbital elements to the observational data. Hence, the choice of a method, or combination of methods, for a particular application, usually involves a small sacrifice in accuracy.

IX. ACKNOWLEDGMENTS

I wish to thank T. M. Burford, A. J. Claus, S. Darlington, F. T. Geyling, and F. W. Sinden for many comments on, and enlightening discussions of, the subject of this paper.

APPENDIX

Equations (20) and (21) may be derived in another way (essentially that used by Aitken in Ref. 3) which does not depend upon the minimization of a quadratic form and provides further insight into the significance of these equations.

Let x be the true value of an n-rowed vector (one-column matrix). Let $\tilde{x}$ be an unbiased estimate of x, with

$$\text{ave}\,\{(\tilde{x} - x)\cdot(\tilde{x} - x)'\} = \tilde{C},$$

where the prime stands for transposition and "ave" stands for ensemble average. The nth-order square matrix $\tilde{C}$ is the covariance matrix of $\tilde{x}$.

Let $\hat{x}$ be another unbiased estimate of x, with

$$\text{ave}\,\{(\hat{x} - x)\cdot(\hat{x} - x)'\} = \hat{C},$$

and let it be assumed that $\tilde{x}$ and $\hat{x}$ are independent, so that

$$\text{ave}\,\{(\tilde{x} - x)\cdot(\hat{x} - x)'\} = 0.$$

Now, consider the weighted linear average

$$\bar{x} = \tilde{W}\cdot\tilde{x} + \hat{W}\cdot\hat{x} \tag{46}$$

where $\tilde{W}$ and $\hat{W}$ are nth-order square matrices, with

$$\tilde{W} + \hat{W} = I \tag{47}$$

where I is the nth-order unit matrix. Since

$$\bar{x} - x = \tilde{W} \cdot (\tilde{x} - x) + \hat{W} \cdot (\hat{x} - x),$$

it readily follows that if

$$\bar{C} = \text{ave}\,\{(\bar{x} - x) \cdot (\bar{x} - x)'\}$$

then

$$\bar{C} = \tilde{W} \cdot \tilde{C} \cdot \tilde{W}' + \hat{W} \cdot \hat{C} \cdot \hat{W}'. \tag{48}$$

Each of the diagonal elements of $\bar{C}$, viz.,

$$\bar{C}_{ii} = \sum_{j,k} (\tilde{W}_{ij} \cdot \tilde{C}_{jk} \cdot \tilde{W}_{ik} + \hat{W}_{ij} \cdot \hat{C}_{jk} \cdot \hat{W}_{ik})$$

will be minimized under constraints on $\tilde{W}_{ik} + \hat{W}_{ik}$ by minimizing

$$\bar{C}_{ii} - 2 \sum_{k} \lambda_{ik} \cdot (\tilde{W}_{ik} + \hat{W}_{ik})$$

where the λ_{ik}'s are Lagrange multipliers. This requires that

$$\sum_{j} \tilde{W}_{ij} \cdot \tilde{C}_{jk} = \lambda_{ik}, \qquad \sum_{j} \hat{W}_{ij} \cdot \hat{C}_{jk} = \lambda_{ik},$$

for every i, k. In matrix notation,

$$\tilde{W} \cdot \tilde{C} = \lambda, \qquad \hat{W} \cdot \hat{C} = \lambda$$

where λ is the nth-order square matrix of the λ_{ik}'s. Hence,

$$\tilde{W} = \lambda \cdot \tilde{C}^{-1}, \qquad \hat{W} = \lambda \cdot \hat{C}^{-1}, \tag{49}$$

where, to satisfy (47),

$$\lambda = (\tilde{C}^{-1} + \hat{C}^{-1})^{-1}. \tag{50}$$

By (48) and (49)

$$\bar{C} = \lambda \cdot (\tilde{C}^{-1} + \hat{C}^{-1}) \cdot \lambda',$$

$$= \lambda' \text{ by (50)},$$

but, since λ is a symmetrical matrix,

$$\bar{C} = \lambda. \tag{51}$$

Finally, by (46), (49), and (51),

$$\bar{x} = \bar{C} \cdot (\tilde{C}^{-1} \cdot \tilde{x} + \hat{C}^{-1} \cdot \hat{x}), \tag{52}$$

where, by (50) and (51),

$$\bar{C} = (\tilde{C}^{-1} + \hat{C}^{-1})^{-1}. \tag{53}$$

This derivation shows that the diagonal elements of $\bar{C}$, which are the variances of the components of $\bar{x}$, have been minimized *independently of one another.*

REFERENCES

1. Blackman, R. B., Smoothing and Prediction of Time Series by Cascaded Simple Averages, 1960 I.R.E., Internatl. Conv. Record, **8**, Part 2, pp. 47–54, and I.R.E. Trans. Circuit Theory, **CT-7**, Special Supp., Aug., 1960, pp. 136–143.
2. Baker, R. M. L., Jr., and Makemson, M. W., *An Introduction to Astrodynamics*, Academic Press, New York, 1960.
3. Aitken, A. C., On Least Squares and Linear Combination of Observations, Proc. Roy. Soc. Edinb. A, **55**, 1934, pp. 42–47.
4. David, F. N., and Neyman, J., Extension of the Markoff Theorem on Least Squares, Statistical Research Memoir, London, **2**, 1938, pp. 105–116.
5. Plackett, R. L., A Historical Note on the Method of Least Squares, Biometrika, **36**, 1949, pp. 458–460.
6. Cohen, E. R., The Basis for the Criterion of Least Squares, Rev. Mod. Phys., **25**, 1953, pp. 709–713.
7. Swerling, P., A Proposed Stagewise Differential Correction Procedure for Satellite Tracking and Prediction, Rand Corporation Report P-1292, Jan. 8, 1958.
8. Swerling, P., First Order Error Propagation in a Stagewise Smoothing Procedure for Satellite Observations. J. Astronautical Sci., **6**, Autumn, 1959, pp. 46–52.
9. Swerling, P., Comment on A Statistical Optimizing Navigation Procedure for Space Flight, J. Amer. Inst. of Aeronautics and Astronautics, **1**, Aug. 1963, p. 1968.
10. Blackman, R. B., unpublished work.
11. Claus, A. J., Orbit Determination for Communication Satellites from Angular Data Only, paper delivered at the American Rocket Society 17th Annual Meeting, Nov. 13–18, 1962, Los Angeles.
12. Claus, A. J., Blackman, R. B., Halline, E. G., and Ridgway, W. C., III, Orbit Determination and Prediction, and Computer Programs, B.S.T.J., **42**, July, 1963, pp. 1357–1382.
13. Claus, A. J., private communication.
14. Battin, R. H., A Statistical Optimizing Navigation Procedure for Space Flight, Jour. ARS, **32**, Nov., 1962, pp. 1681–1696.

On the Response of Nonlinear Control Systems to Periodic Input Signals

By I. W. SANDBERG

(Manuscript received June 20, 1963)

In this paper we study a broad class of nonlinear control systems containing a single memoryless nonlinear element. We present conditions under which there exists a unique periodic response, with a given period, to an arbitrary periodic input with the same period, and we derive an upper bound on the mean-square error incurred by applying the describing-function technique. The expression for the error reflects the intuitive engineering arguments that are often employed to justify the use of the describing-function method. Conditions are also presented under which subharmonic response components and self-sustained oscillations cannot occur.

I. INTRODUCTION

The describing-function technique is often used to determine the response of nonlinear control systems to sinusoidal input signals. In this approach,* which is applicable to systems of any order but which is ordinarily restricted to systems containing only one nonlinear element, it is assumed that the response is periodic, with only the component at the input frequency significant.

Although the describing-function technique is of considerable practical value and indeed is one of the most powerful analytical tools available to the control system synthesist, it appears that, except with regard to predicting the existence of self-sustained oscillations,[6] there has been no rigorous discussion of its validity.†

In this paper we study a broad class of nonlinear control systems containing a single memoryless nonlinear element. We present conditions under which there exists a unique periodic response, with a given period, to an arbitrary periodic input with the same period, and we

* The describing-function technique was discovered independently by engineers in at least five different countries.[1-5]

† However, some interesting relevant ideas have been presented by Johnson.[7]

derive an upper bound on the mean-square error incurred by applying the describing-function technique. The expression for the error reflects the intuitive engineering arguments that are often employed to justify the use of the describing-function method. Conditions are also presented under which subharmonic response components and self-sustained oscillations cannot occur.

Some mathematical preliminaries are considered in Section II. In Section III we describe the physical system to be studied, introduce some assumptions and notation, and discuss the describing-function technique. The remaining sections are concerned with mathematical results relating to the functional equation that governs the behavior of the physical system.

II. MATHEMATICAL PRELIMINARIES

Let $\mathfrak{R} = [\Theta,\rho]$ be an arbitrary metric space.[8] A mapping $\mathbf{A}$ of the space $\mathfrak{R}$ into itself is said to be a contraction if there exists a number $k < 1$ such that

$$\rho(\mathbf{A}x,\mathbf{A}y) \leqq k\rho(x,y)$$

for any two elements $x,y \varepsilon \Theta$. The contraction-mapping fixed-point theorem[8] is basic to much of the subsequent discussion. It states that every contraction-mapping defined in a complete metric space $\mathfrak{R}$ has one and only one fixed point (i.e., there exists a unique element $z \varepsilon \Theta$ such that $\mathbf{A}z = z$). Furthermore $z = \lim_{n\to\infty} \mathbf{A}^n x_0$, where x_0 is an arbitrary element of Θ.

Let T be a real positive constant. The space of real-valued periodic functions of t with period T which are square-integrable over a period is denoted by $\mathfrak{K}$. The norm of $g \varepsilon \mathfrak{K}$ is denoted by $\| g \|$ and is defined by

$$\| g \|^2 = \frac{1}{T}\int_0^T g^2 dt$$

(i.e., $\| g \|$ is the rms value of g). With this norm $\mathfrak{K}$ is a Banach space.

If $g \varepsilon \mathfrak{K}$,

$$g = \underset{N\to\infty}{\text{l.i.m.}} \sum_{n=-N}^{N} g_n e^{in\omega_0 t}$$

where $\omega_0 = 2\pi/T$ and the Fourier coefficients g_n are given by

$$g_n = \frac{1}{T}\int_0^T g(t)e^{-in\omega_0 t}dt.$$

Parseval's identity reads:

$$\sum_{-\infty}^{\infty} |\, g_n \,|^2 = \| g \|^2.$$

Two elements of $\mathcal{K}$, g and h, are equivalent if $\| g - h \| = 0$.

In accordance with the usual notation, the norm of a linear operator $\mathbf{Q}$ defined on $\mathcal{K}$ is denoted by $\| \mathbf{Q} \|$.

The symbol $\mathcal{L}_{1R}$ denotes the space of real-valued absolutely integrable functions defined on the real interval $(-\infty, \infty)$. We take as the definition of the Fourier transform of $f(t) \in \mathcal{L}_{1R}$:

$$F(i\omega) = \int_{-\infty}^{\infty} f(t) e^{-i\omega t} dt.$$

The symbol $\mathbf{I}$ is used throughout to denote the identity operator.

III. DESCRIPTION OF THE PHYSICAL SYSTEM, THE DESCRIBING-FUNCTION TECHNIQUE, AND THE PROJECTION OPERATOR $\mathbf{P}$

We shall be concerned with the familiar nonlinear control system shown in Fig. 1.

Assumption I: It is assumed throughout that $\mathbf{F}$ *(in Fig. 1) is a linear operator. Let* $\mathfrak{F} = \{ \cdots, F_{-2}, F_{-1}, F_0, F_1, F_2, \cdots \}$ *denote a countable set of complex constants such that* $\sup_n |\, F_n \,| < \infty$ *and* F_n *is equal to the complex conjugate of* F_{-n}. *Unless stated otherwise, it is assumed that the restriction of* $\mathbf{F}$ *to* $\mathcal{K}$ *is a bounded linear mapping of* $\mathcal{K}$ *into itself with the property that if* $g \in \mathcal{K}$ *and* $h = \mathbf{F}g$, *then* $h_n = F_n g_n$ *in which* g_n *and* h_n, *respectively, are the nth Fourier coefficients of* g *and* h. *(According to the Riesz-Fischer theorem,* $\mathbf{F}$ *is completely defined on* $\mathcal{K}$ *by* $\mathfrak{F}$.)

The class of operators consistent with Assumption I includes the important special case in which

$$\mathbf{F}g = \int_{-\infty}^{\infty} f(t - \tau) g(\tau) d\tau, \qquad g \in \mathcal{K}$$

where $f(t) \in \mathcal{L}_{1R}$ (see Appendix A). Here $F_n = F(in\omega_0)$ where $F(i\omega)$ is the Fourier transform of $f(t)$.

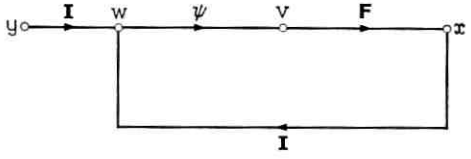


Fig. 1 — Nonlinear control system.

Assumption II: The nonlinear function ψ in Fig. 1, which introduces the constraint $v(t) = \psi[w(t)]$, is assumed throughout to be real-valued, independent of t, and such that there exist two real constants α and β ($\beta > 0$) with the properties that $\frac{1}{2}(\alpha + \beta) = 1$ and

$$\alpha(\mu_1 - \mu_2) \leqq \psi(\mu_1) - \psi(\mu_2) \leqq \beta(\mu_1 - \mu_2)$$

for any real $\mu_1 \geqq \mu_2$.

The normalization $\frac{1}{2}(\alpha + \beta) = 1$ permits some simplification in the subsequent statements of results.

In Fig. 1 the input signal y is related to the output signal x by the functional equation

$$x = \mathbf{F}\psi[x + y].$$

3.1 *The Sinusoidal Response of the System in Fig. 1*

The response of the system in Fig. 1 to a sinusoidal input is frequently of engineering interest. Typically the system is of high order* so that the well-known special techniques applicable to second-order systems cannot be used. The describing-function approach simplifies the problem by assuming that the output is periodic and that the only significant frequency component of the output is that component at the input frequency. Hence it is assumed that the input to the nonlinear device is a sinusoid and that ψ is characterized by the ratio of the fundamental component of its output to the amplitude of the sinusoidal input (this ratio is called the describing function for ψ). Thus the nonlinear element is treated as an element with a gain that varies with input signal level, and to the extent that the describing-function approximation (sometimes called the "first harmonic approximation") is valid, the usual frequency response methods can be employed.

The first harmonic approximation is often "justified" on three grounds: first, no significant subharmonic components of $x(t)$ are ordinarily present; second, the harmonics of the output of ψ are ordinarily of smaller amplitude than the fundamental and, third, in most feedback systems $\mathbf{F}$ behaves as a low-pass filter with the result that the higher harmonics are significantly attentuated.

Aside from at least two computational difficulties[9] associated with the describing-function method, which can be remedied to a considerable extent with machine aids, "The third and most basic difficulty is related to the inaccuracy of the method and, in particular, to the

* That is, the nonlinear differential equation governing the system is typically of high order.

uncertainty throughout the analysis about the accuracy. There is [in the literature] no simple method for evaluating the accuracy of the describing-function analysis of a nonlinear system and no definite assurance that the results derived with the describing function are even approximately correct."[9] However, it should not be inferred that the accuracy is necessarily poor.[9,10] "Indeed the correlation between experimental and theoretical results is in many cases better than the accuracy of the design data."[9]

3.2 *The Role of the Projection Operator* **P**

A moment's reflection will show that the describing-function technique as applied to the system in Fig. 1 amounts to analyzing the approximating system that results by replacing the operator **F** with the operator $\tilde{\mathbf{F}}$ defined by

$$\frac{1}{T}\int_0^T [\tilde{\mathbf{F}}g]e^{-in\omega_0 t}dt = F_n g_n, \qquad n = \pm 1$$

$$= 0, \qquad n \neq \pm 1$$

where $g \in \mathfrak{K}$, g_n is the nth Fourier coefficient of g and $T = 2\pi/\omega_0$ is the period of the input sinusoid.

At this point it is convenient to introduce

Definition I: Let $\mathfrak{N}$ *denote a set of integers such that* $-m \in \mathfrak{N}$ *if* $m \in \mathfrak{N}$. *Let* g *be an arbitrary element of* $\mathfrak{K}$ *with nth Fourier coefficient* g_n. *The projection operator* **P** *is a linear mapping of* $\mathfrak{K}$ *into itself defined by*

$$\frac{1}{T}\int_0^T [\mathbf{P}g]e^{-in\omega_0 t}dt = g_n, \qquad n \in \mathfrak{N}$$

$$= 0, \qquad n \notin \mathfrak{N}.$$

An obvious generalization of the describing-function technique is to take as the approximating system the system that is obtained by replacing **F** in Fig. 1 with **PF**. The results to be presented relate to this more general situation. Of course in the case of principal interest, $\mathfrak{N} = \{-1,1\}$ and $\mathbf{PF} = \tilde{\mathbf{F}}$.

IV. RESULTS RELATING TO THE FUNCTIONAL EQUATION $x = \mathbf{F}\psi[x + y]$

The proof of the following simple preliminary result is given in Appendix B.

Theorem I:

$$\| \mathbf{F} \| = \sup_n | F_n |.$$

If $\inf_n | 1 - F_n | > 0$, *the operator* $(\mathbf{I} - \mathbf{F})$ *possesses a bounded inverse on* $\mathcal{K}$ *and*

$$\| (\mathbf{I} - \mathbf{F})^{-1}\mathbf{F} \| = \sup_n \left| \frac{F_n}{1 - F_n} \right|.$$

The principal result of this section is

Theorem II: Let $\mathbf{F}$, ψ, *and* β *be as defined in Section III. Let* $y \varepsilon \mathcal{K}$. *Suppose that*

$$r = \sup_n \left| \frac{F_n}{1 - F_n} \right| (\beta - 1) < 1.$$

Then there exists a unique $x \varepsilon \mathcal{K}$ *such that* $x = \mathbf{F}\psi[x + y]$. *In fact,* $x = \lim_{m \to \infty} x_m$ *where*

$$x_{m+1} = (\mathbf{I} - \mathbf{F})^{-1}\mathbf{F}\{\psi[x_m + y] - x_m\}$$

and x_0 *is an arbitrary element of* $\mathcal{K}$. *The mth approximation* x_m *satisfies*

$$\| x_m - x \| \leqq \frac{r^m}{1 - r} \| x_1 - x_0 \|.$$

Proof:

Let $\psi[w] = \psi_0 w + \tilde{\psi}[w]$, where ψ_0 is a real constant such that $\inf_n | 1 - \psi_0 F_n | > 0$ (since $r < 1$, there exists such a ψ_0). According to Theorem I, $(\mathbf{I} - \psi_0\mathbf{F})$ possesses a bounded inverse on $\mathcal{K}$. Hence the functional equation $x = \mathbf{F}\psi[x + y]$ can be written as $x = \mathbf{M}x$ where

$$\mathbf{M}x = (\mathbf{I} - \psi_0\mathbf{F})^{-1}\mathbf{F}\tilde{\psi}[x + y] + \psi_0(\mathbf{I} - \mathbf{F})^{-1}\mathbf{F}y.$$

In order to prove Theorem II it is sufficient to consider the case in which $\psi_0 = 1$. However, we prefer to bring out the fact that this choice of ψ_0, the median of α and β, is optimal in a significant sense.

It is evident that $\mathbf{M}$ is a mapping of $\mathcal{K}$ into itself. Let us consider the determination of a condition under which $\mathbf{M}$ is in fact a contraction-mapping of $\mathcal{K}$ into itself. Let $g,h \varepsilon \mathcal{K}$ and observe that

$$\begin{aligned} \| \mathbf{M}g - \mathbf{M}h \| &= \| (\mathbf{I} - \psi_0\mathbf{F})^{-1}\mathbf{F}\{\tilde{\psi}[g + y] - \tilde{\psi}[h + y]\} \| \\ &\leqq \| (\mathbf{I} - \psi_0\mathbf{F})^{-1}\mathbf{F} \| \| \tilde{\psi}[g + y] - \tilde{\psi}[h + y] \|. \end{aligned}$$

Since

$$\| \tilde{\psi}[g + y] - \tilde{\psi}[h + y] \| = \left\| \left(\frac{\psi[g + y] - \psi[h + y]}{g - h} - \psi_0 \right) (g - h) \right\|,$$

$$\| \tilde{\psi}[g + y] - \tilde{\psi}[h + y] \| \leqq \eta(\psi_0) \| g - h \|$$

where

$$\eta(\psi_0) = \beta - \psi_0, \qquad \psi_0 \leqq 1$$
$$= \psi_0 - \alpha, \qquad \psi_0 \geqq 1.$$

Thus

$$\| \mathbf{M}g - \mathbf{M}h \| \leqq \| (\mathbf{I} - \psi_0\mathbf{F})^{-1}\mathbf{F} \| \eta(\psi_0) \| g - h \|,$$

and $\mathbf{M}$ is a contraction if

$$q(\psi_0) = \| (\mathbf{I} - \psi_0\mathbf{F})^{-1}\mathbf{F} \| \eta(\psi_0) < 1.$$

Using Theorem I,

$$q(\psi_0) = \sup_n \left| \frac{F_n}{1 - \psi_0 F_n} \right| \eta(\psi_0).$$

Assuming that there exists a ψ_0 such that $q(\psi_0) < 1$, the following result, which is proved in Ref. 11, implies that $\inf_{\psi_0} q(\psi_0) = q(1)$.

Lemma I: Let ξ be a complex number and suppose that $| \xi - \psi_0 |^{-1}\eta(\psi_0) < 1$. *Then*

$$| \xi - \psi_0 |^{-1}\eta(\psi_0) \geqq | \xi - 1 |^{-1}\eta(1).$$

From this point on we assume that $\psi_0 = 1$ and we set $q(1) = r$. Thus the assumptions stated in Theorem II imply that $\mathbf{M}$ is a contraction. In view of the contraction-mapping fixed-point theorem, this establishes the existence and uniqueness of the function $x(t)$ and the fact that it can be determined in accordance with the stated iteration procedure.

The upper bound on $\| x_m - x \|$ follows directly from the fact that x can be written as

$$x = x_0 + \sum_{j=0}^{\infty} [x_{(j+1)} - x_j],$$

in which, for all $j \geqq 1$

$$\| x_{(j+1)} - x_j \| = \| \mathbf{M}x_j - \mathbf{M}x_{(j-1)} \| \leqq r \| x_j - x_{(j-1)} \|.$$

Remarks:

Observe that a nontrivial self-sustained periodic oscillation with period T cannot exist in the system of Fig. 1 if the hypotheses of Theorem II are satisfied and $\psi(0) = 0$ (since then $y = 0$ implies that $x = 0$).

When $\mathbf{F}$ is defined by

$$\mathbf{F}g = \int_{-\infty}^{\infty} f(t - \tau)g(\tau)d\tau$$

where $f(t) \varepsilon \mathfrak{L}_{1R}$, Theorem II implies the following simple necessary condition for the occurrence of jump-resonance phenomena[9] in the system of Fig. 1:

$$\sup_{\omega} \left| \frac{F(i\omega)}{1 - F(i\omega)} \right| (\beta - 1) \geqq 1.$$

For example, let $F(i\omega) = -k[(i\omega + a)(i\omega + 1)]^{-1}$ where k and a are positive constants. Let

$$\begin{aligned} \psi[w] &= 2w, && |w| \leqq c \\ &= 2c \operatorname{sign}(w), && |w| > c \end{aligned}$$

where c is a positive constant. Then it is a routine matter to show that jump-resonance phenomena can occur only if $k > \frac{1}{2}(1 + a^2)$.

4.1 *Two Further Consequences of Theorem II*

Corollary I: Suppose that the hypotheses of Theorem II are satisfied. Then

$$\| x \| \leqq \frac{1}{1 - r} \| (\mathbf{I} - \mathbf{F})^{-1} \mathbf{F} \psi[y] \|.$$

Proof:

Set $m = 0$ and $x_0 = 0$ in the upper bound for $\| x_m - x \|$.

Corollary II: Suppose that the hypotheses of Theorem II are satisfied and that $\hat{x} \varepsilon \mathfrak{K}$ satisfies $\hat{x} = \mathbf{PF}\psi[\hat{x} + y]$. Then

$$\| x - \hat{x} \| \leqq \frac{1}{1 - r} \| (\mathbf{I} - \mathbf{F})^{-1} \mathbf{F} (\mathbf{I} - \mathbf{P}) \psi[\hat{x} + y] \|.$$

Proof:

With $m = 0$ and $x_0 = \hat{x}$, the upper bound for $\| x_m - x \|$ yields

$$\| x - \hat{x} \| \leqq \frac{1}{1 - r} \| (\mathbf{I} - \mathbf{F})^{-1} \mathbf{F} \{ \psi[\hat{x} + y] - \hat{x} \} - (\mathbf{I} - \mathbf{PF})^{-1} \mathbf{PF} \{ \psi[\hat{x} + y] - \hat{x} \} \|.$$

Since $(\mathbf{I} - \mathbf{PF})^{-1} = (\mathbf{I} - \mathbf{P}) + (\mathbf{I} - \mathbf{F})^{-1}\mathbf{P}$ and $(\mathbf{I} - \mathbf{P})\hat{x} = 0$,

$$\| x - \hat{x} \| \leqq \frac{1}{1 - r} \| (\mathbf{I} - \mathbf{F})^{-1} \mathbf{F} (\mathbf{I} - \mathbf{P}) \psi[\hat{x} + y] \|.$$

Remarks:

Note that the hypotheses of Theorem II imply that there exists a unique $\hat{x} \varepsilon \mathfrak{K}$ such that $\hat{x} = \mathbf{PF}\psi[\hat{x} + y]$.

The bound on $\| x - \hat{x} \|$ can be expressed with the aid of Parseval's identity as

$$\| x - \hat{x} \| \leqq \frac{1}{1 - r} \left(\sum_{n \neq \mathfrak{N}} \left| \frac{F_n}{1 - F_n} p_n \right|^2 \right)^{\frac{1}{2}} \tag{1}$$

where p_n is the nth Fourier coefficient of $\psi[\hat{x} + y]$. Consider the usual describing-function case in which $\mathfrak{N} = \{-1,1\}$, and y is a sinusoid with period T. Assuming that ψ is an odd function so that $p_0 = 0$, (1) clearly shows that $\| x - \hat{x} \|$ is small when the amplitudes of the harmonics* of $\psi[\hat{x} + y]$ are sufficiently small or when the attenuation of the harmonics by $\mathbf{F}$ is sufficiently large. Thus subject to the key inequality

$$\sup_n \left| \frac{F_n}{1 - F_n} \right| (\beta - 1) < 1,$$

(1) makes precise the usual intuitive engineering arguments regarding the applicability of the describing-function method for determining the sinusoidal response of the system in Fig. 1.

It can be shown[12] that if $\alpha \geqq 0$, $\psi(0) = 0$, and $y = \mathbf{P}y$:

$$\| \mathbf{P}\psi[\hat{x} + y] \| \geqq \alpha \| \hat{x} + y \|.$$

Under these conditions

$$\begin{aligned} \| (\mathbf{I} - \mathbf{P})\psi[\hat{x} + y] \| &= (\| \psi[\hat{x} + y] \|^2 - \| \mathbf{P}\psi[\hat{x} + y] \|^2)^{\frac{1}{2}} \\ &\leqq (\beta^2 - \alpha^2)^{\frac{1}{2}} \| \hat{x} + y \|, \end{aligned}$$

and

$$\begin{aligned} \| x - \hat{x} \| &\leqq \frac{1}{1 - r} \| (\mathbf{I} - \mathbf{F})^{-1}\mathbf{F}(\mathbf{I} - \mathbf{P}) \| (\beta^2 - \alpha^2)^{\frac{1}{2}} \| \hat{x} + y \| \\ &\leqq \frac{1}{1 - r} \sup_{n \neq \mathfrak{N}} \left| \frac{F_n}{1 - F_n} \right| (\beta^2 - \alpha^2)^{\frac{1}{2}} \| \hat{x} + y \|. \end{aligned}$$

Under the conditions stated in Theorem II, the response $x(t)$ can be determined in accordance with an iteration procedure for which the successive approximations converge in the mean-square sense at least a geometric rate. In particular, if we take $x_0 = \hat{x}$, the solution given by the describing-function method, the second approximation is

$$x_1 = (\mathbf{I} - \mathbf{F})^{-1}\mathbf{F}\{\psi[\hat{x} + y] - \hat{x}\},$$

* Of course here all even harmonic components vanish.

which can be evaluated in a relatively simple manner once $\hat{x}$ has been determined. Using Theorem II and the expression for $\| x_1 - x_0 \|$ obtained above,

$$\| x - x_1 \| \leqq \frac{r}{1 - r} \| (\mathbf{I} - \mathbf{F})^{-1} \mathbf{F} (\mathbf{I} - \mathbf{P}) \psi[\hat{x} + y] \|.$$

4.2 *An Observation Relating to the Necessity of the Assumptions in Theorem II*

The basic assumption in Theorem II is:

$$\sup_n \left| \frac{F_n}{1 - F_n} \right| (\beta - 1) < 1.$$

This inequality is satisfied if and only if the numbers $(F_n)^{-1}$ are bounded away from the disk centered in the complex plane at (1,0) and having radius $(\beta - 1)$. It is of interest to note that there is a function ψ, in fact a linear ψ, that satisfies Assumption II and possesses the property that there is no function $x(t) \varepsilon \mathcal{K}$ such that $x = \mathbf{F}\psi[x + y]$ if, for some integer k:

(*i*) $F_k \neq 0$

(*ii*) F_k is a point on the real-axis diameter of the disk mentioned above, and

(*iii*) the kth Fourier coefficient of y does not vanish.

To prove this assertion observe that if the three conditions are satisfied, $\alpha \leqq (F_k)^{-1} \leqq \beta$, and $x = \mathbf{F}\psi[x + y]$ with $\psi[w] = (F_k)^{-1} w$, possesses no solution belonging to $\mathcal{K}$.

This observation suggests that the assumptions made in Theorem II are not too far from being necessary.

4.3 *On the Boundedness of the Solution*

Theorem III: Let $\mathbf{F}$ *be defined by*

$$\mathbf{F}g = \int_{-\infty}^{\infty} f(t - \tau) g(\tau) d\tau, \qquad g \varepsilon \mathcal{K}$$

where $f \varepsilon \mathcal{L}_{1R}$, *and let* ψ *satisfy Assumption II. Let* $x \varepsilon \mathcal{K}$ *satisfy* $x = \mathbf{F}\psi[x + y]$, $y \varepsilon \mathcal{K}$. *Suppose that* $(1 + |t|) f(t)$ *is square-integrable on* $(-\infty, \infty)$. *Then* $|x(t)|$ *is uniformly bounded on* $0 \leqq t \leqq T$.

Proof:

Let $h = \psi[x + y]$ and note that $h \varepsilon \mathcal{K}$. Using the Schwarz inequality,

$$|x(t)|^2 = \left| \int_{-\infty}^{\infty} (1 + |\tau|) f(\tau) \frac{h(t-\tau)}{(1+|\tau|)} \, d\tau \right|^2$$

$$\leqq \int_{-\infty}^{\infty} [(1 + |\tau|) f(\tau)]^2 d\tau \int_{-\infty}^{\infty} \left| \frac{h(t-\tau)}{1+|\tau|} \right|^2 d\tau.$$

The last integral can be bounded as follows:

$$\int_{-\infty}^{\infty} \left| \frac{h(t-\tau)}{1+|\tau|} \right|^2 d\tau = \sum_{n=-\infty}^{\infty} \int_{nT}^{(n+1)T} \left| \frac{h(t-\tau)}{1+|\tau|} \right|^2 d\tau$$

$$\leqq 2 \left(1 + (T)^{-2} \sum_{n=1}^{\infty} n^{-2} \right) \int_0^T |h(t)|^2 d\tau.$$

Thus $|x(t)|$ is uniformly bounded on $0 \leqq t \leqq T$.

Remarks:

The hypothesis regarding $f(t)$ is almost always satisfied in cases of engineering interest.

If the hypotheses of Theorem III are satisfied and $|y(t)|$ is uniformly bounded on $[0,T]$, it follows that $x(t)$ is continuous on $[0,T]$, since $|h(t)|$ is uniformly bounded on $[0,T]$ and

$$\int_{-\infty}^{\infty} |f(t+\delta) - f(t)| \, dt \to 0 \text{ as } \delta \to 0.$$

V. RESULTS RELATING TO AN IMPORTANT SPECIAL CLASS OF CONTROL SYSTEMS

One frequently encounters[9,10] discussions of systems of the type shown in Fig. 1 in which ψ is an odd function and the operator corresponding to $\mathbf{F}$ is characterized in the frequency domain by a transfer function (typically a real rational function in $i\omega$) with at least one pole at $\omega = 0$. The techniques presented earlier can be applied to situations of this type if $\mathbf{F}$ is replaced with $\mathbf{F}'$, the restriction of $\mathbf{F}$ to the subspace*

$$\mathcal{K}' = \left\{ g \mid g \in \mathcal{K}; \int_0^T g(t) e^{-in\omega_0 t} dt = 0, n \text{ even} \right\}.$$

The operator $\mathbf{F}'$ is completely defined on $\mathcal{K}'$ by the set of complex numbers $\mathfrak{F}' = \{\cdots, F_{-3}, F_{-1}, F_1, F_3, \cdots\}$. The result analogous to Theorem II is

Theorem IV: Let $\mathbf{F}'$ and $\mathcal{K}'$ be as defined above. Let ψ and β be as defined

* It is a simple matter to show that the linear manifold $\mathcal{K}'$ is in fact a subspace of $\mathcal{K}$. Consider any Cauchy sequence of elements of $\mathcal{K}'$. Since $\mathcal{K}$ is complete, the sequence converges to a function g belonging to $\mathcal{K}$. However, a direct application of Parseval's identity shows that this is impossible unless $g \in \mathcal{K}'$.

in Section III, with the further qualification that $\psi(-\mu) = -\psi(\mu)$ *for any real* μ. *Let* $y \in \mathcal{K}'$. *Suppose that*

$$q = \sup_{n \text{ odd}} \left| \frac{F_n}{1 - F_n} \right| (\beta - 1) < 1.$$

Then the conclusion of Theorem II remains valid if $\mathcal{K}$ *is replaced with* $\mathcal{K}'$, r *is replaced with* q, *and* $\mathbf{F}$ *is replaced with* $\mathbf{F}'$.

Proof:

The proof is essentially the same as for Theorem II. The assumption that ψ is odd is needed to verify that the operator corresponding to $\mathbf{M}$ (with $\psi_0 = 1$) is a mapping of the Banach space $\mathcal{K}'$ into itself.*

Of course Theorem IV implies results entirely analogous to Corollaries I and II.

VI. ON THE LACK OF SUBHARMONIC COMPONENTS IN THE RESPONSE

One of the key assumptions relating to the describing-function analysis of the system in Fig. 1 is that the response $x(t)$ (assuming it exists) is a periodic function with period T when $y(t)$ is a sinusoid with period T. In particular $x(t)$ is assumed not to contain subharmonic components. The techniques described earlier can be used to obtain explicit conditions under which this assumption is valid. The following theorem contains one such result. A preliminary fact[13] that is needed is: if $f(t) \in \mathcal{L}_{1R}$ and $F(i\omega) \neq 1$ for all real ω, then there exists a function $h(t) \in \mathcal{L}_{1R}$ with Fourier transform $F(i\omega)\,[1 - F(i\omega)]^{-1}$.

Theorem V: Let η *denote the space of bounded real-valued measurable functions defined on* $(-\infty, \infty)$. *Let* $\mathbf{F}$ *be defined by*

$$\mathbf{F}g = \int_{-\infty}^{\infty} f(t - \tau) g(\tau) d\tau, \qquad g \in \eta$$

where $f \in \mathcal{L}_{1R}$ *and let* ψ *and* β *be as defined in Section III. Suppose that* $F(i\omega) \neq 1$ *for all real* ω *and that*

$$(\beta - 1) \int_{-\infty}^{\infty} | h(t) | \, dt < 1,$$

where $h(t)$ *has Fourier transform* $F(i\omega)\,[1 - F(i\omega)]^{-1}$. *Let* $y \in \eta$. *Then there exists a unique* $x \in \eta$ *such that* $x = \mathbf{F}\psi[x + y]$. *Further,* $x(t)$ *is continuous and if* $y(t + T) = y(t)$, *then* $x(t + T) = x(t)$.

* If $g \in \mathcal{K}'$, $g(t) = -g(t + \frac{1}{2}T)$ for almost every t. Since ψ is odd, $\mathbf{M}$ preserves this property and hence $\mathbf{M}$ maps $\mathcal{K}'$ into itself.

Proof:

Arguments very similar to those presented in Ref. 13 can be used to show that under the conditions stated in the theorem, $\mathbf{F}$ is a bounded mapping of η into itself, that the operator $(\mathbf{I} - \mathbf{F})$ possesses a bounded inverse on η and that

$$(\mathbf{I} - \mathbf{F})^{-1}\mathbf{F}g = \int_{-\infty}^{\infty} h(t - \tau)g(\tau)d\tau, \qquad g \varepsilon \eta.$$

Thus, paralleling the proof of Theorem II, the functional equation $x = \mathbf{F}[x + y]$ can be written as $x = \mathbf{L}x$ where

$$\mathbf{L}x = (\mathbf{I} - \mathbf{F})^{-1}\mathbf{F}\{\psi[x + y] - x\}.$$

Let the norm of an element of η be defined by

$$\| g \|_{\infty} = \sup_{t} | g(t) |, \qquad g \varepsilon \eta.$$

With this norm η is a Banach space. It is a routine matter to verify that under the conditions stated in the theorem $\mathbf{L}$ is a contraction mapping of η into itself. This implies the existence of a unique solution $x(t) \varepsilon \eta$.

The continuity of $x(t)$ follows directly from the fact that $x = \mathbf{F}\psi[x + y]$ in which $\psi[x + y]$ is bounded and $f \varepsilon \mathfrak{L}_{1R}$.

If $y(t + T) = y(t)$, $\mathbf{L}$ is a contraction mapping of the following subspace of η into itself: $\{g \mid g \varepsilon \eta, g(t) = g(t + T)\}$ and hence there exists a unique solution belonging to this subspace.* This completes the proof of Theorem V.

VII. FINAL REMARKS

It seems likely to this writer that the contraction-mapping fixed-point theorem, and more generally the techniques of functional analysis, can be exploited with considerable profit by the control system synthesist. Indeed one objective of this paper is to stimulate engineering interest in these techniques.

The results in this paper can be extended to cover the analogous multiloop multi-nonlinear-element case (i.e., the case in which y, w, v, and x in Fig. 1 are N-vector valued functions of t, and ψ represents N nonlinear elements of the type considered earlier). In particular, the

* Alternatively, observe that $x(t + T)$ is a solution of the functional equation when $y(t + T) = y(t)$. Since the solution is unique, $x(t + T) = x(t)$.

corresponding extension of Theorem II is given in Theorem IV of Ref. 14.*

The writer is indebted to V. E. Beneš and H. O. Pollak for reading the draft.

APPENDIX A

Let **F** be defined by

$$\mathbf{F}g = \int_{-\infty}^{\infty} f(t-\tau)g(\tau)d\tau, \qquad g \in \mathfrak{K}$$

where $f(t) \in \mathfrak{L}_{1R}$.

We show first that **F** is a bounded mapping of $\mathfrak{K}$ into itself. Consider

$$h(t) = \int_{-\infty}^{\infty} f(t-\tau)g(\tau)d\tau.$$

Using Schwarz's inequality,

$$\begin{aligned} |h(t)| &\leqq \int_{-\infty}^{\infty} |g(t-\tau)|\cdot|f(\tau)|^{\frac{1}{2}}\cdot|f(\tau)|^{\frac{1}{2}}d\tau \\ &\leqq \left(\int_{-\infty}^{\infty} |g(t-\tau)|^2\cdot|f(\tau)|\,d\tau\right)^{\frac{1}{2}}\left(\int_{-\infty}^{\infty}|f(\tau)|\,d\tau\right)^{\frac{1}{2}} \end{aligned}$$

from which

$$\int_0^T |h(t)|^2 dt \leqq \int_0^T \left[\int_{-\infty}^{\infty} |g(t-\tau)|^2\cdot|f(\tau)|\,d\tau\right] dt \int_{-\infty}^{\infty}|f(\tau)|\,d\tau. \quad (2)$$

Since g has period T and $f \in \mathfrak{L}_{1R}$,

$$\int_{-\infty}^{\infty}\left[\int_0^T |g(t-\tau)|^2 dt\right]|f(\tau)|\,d\tau = \int_0^T |g(t)|^2\,dt \int_{-\infty}^{\infty}|f(\tau)|\,d\tau < \infty.$$

Hence Fubini's theorem implies that the order of integration in (2) can be interchanged. Thus

$$\int_0^T |h(t)|^2 dt \leqq \int_0^T |g(t)|^2 dt \left(\int_{-\infty}^{\infty}|f(\tau)|\,d\tau\right)^2$$

and since $h(t)$ is clearly real-valued and periodic in t with period T, **F** is a bounded mapping of $\mathfrak{K}$ into itself.

Consider now the relation between the Fourier coefficients of h and g

* In Theorem IV of Ref. 14, the basic functional equation is written in terms of what corresponds here to w (since $x = w - y$, this is, of course, an unimportant difference), and the nonlinear functions are permitted to depend periodically on t with period T.

stated in Section III. We have

$$\int_0^T h(t)e^{-in\omega_0 t}\,dt = \int_0^T \left[\int_{-\infty}^{\infty} f(\tau)g(t-\tau)d\tau\right] e^{-in\omega_0 t}dt.$$

By the same argument as before, the order of integration can be interchanged. Thus

$$\begin{aligned}\int_0^T h(t)e^{-in\omega_0 t}dt &= \int_{-\infty}^{\infty}\left[\int_0^T g(t-\tau)e^{-in\omega_0 t}dt\right]f(\tau)d\tau \\ &= \int_{-\infty}^{\infty}\left[\int_0^T g(t-\tau)e^{-in\omega_0(t-\tau)}dt\right]e^{-in\omega_0\tau}f(\tau)d\tau \\ &= F(in\omega_0)\int_0^T g(t)e^{-in\omega_0 t}dt\end{aligned}$$

in which, clearly, $\sup_n |F(in\omega_0)| < \infty$.

APPENDIX B

Proof of Theorem I

The norm of $\mathbf{F}$ is: $\sup\{\|\mathbf{F}g\| ; g \in \mathcal{K}, \|g\| = 1\}$. Using Parseval's identity, $\|\mathbf{F}g\|^2 = \sum_{-\infty}^{\infty} |F_n|^2 \cdot |g_n|^2$. Thus $\|\mathbf{F}\| \leqq \sup_n |F_n|$, and since it is clear that there exists a $g \in \mathcal{K}$ such that $\|g\| = 1$ and $\|\mathbf{F}g\| \geqq \sup_n |F_n| - \delta$, where δ is an arbitrary positive number, $\|\mathbf{F}\| = \sup_n |F_n|$.

Next, consider the invertibility of the operator $(\mathbf{I} - \mathbf{F})$ when $\inf_n |1 - F_n| > 0$. Let $w \in \mathcal{K}$. The hypothesis implies that

$$\sum_{-\infty}^{\infty} |1 - F_n|^{-2} |w_n|^2 < \infty,$$

where w_n is the nth Fourier coefficient of w. Thus, according to the Riesz-Fischer theorem, there exists a $z \in \mathcal{K}$ with Fourier coefficients $z_n = w_n(1 - F_n)^{-1}$. Parseval's identity implies that z satisfies the equation $(\mathbf{I} - \mathbf{F})z = w$ (in the sense that $\|(\mathbf{I} - \mathbf{F})z - w\| = 0$) and that

$$\|z\| \leqq \sup_n |1 - F_n|^{-1} \|w\|.$$

Thus $(\mathbf{I} - \mathbf{F})$ possesses a bounded inverse on $\mathcal{K}$.

The expression for $\|(\mathbf{I} - \mathbf{F})^{-1}\mathbf{F}\|$ given in Theorem I follows from arguments similar to those used to obtain the expression for $\|\mathbf{F}\|$.

REFERENCES

1. Tustin, A., A Method of Analyzing the Effects of Certain Kinds of Nonlinearity in Closed-Cycle Control Systems, Jour. IEE, **94**, Pt. IIA, 1947, p. 152.
2. Goldfarb, L. C., On Some Nonlinear Phenomena in Regulatory Systems, *Frequency Response*, ed. Oldenburger, T., New York, Macmillan, 1956, pp. 239–257, (translation from the Russian original in Automat. Telemech., Moscow, 8, 1947, p. 349).
3. Oppelt, W., Locus Curve Method for Regulators with Friction, Rep. Nat. Bur. Stand., No. 1691, (translation from the German original in Z. Ver. Dtsch. Ing., 90, 1948, p. 179).
4. Kochenburger, R. J., A Frequency Response Method for Analyzing Contractor Servomechanisms, Trans. A.I.E.E., **69**, Pt. I, 1950, p. 270.
5. Dutilh, J. R., Theorie des Servomecanismes a Relais, Ondes Elect., **30**, 1950, p. 438.
6. Bass, R. W., Mathematical Legitimacy of Equivalent Linearization by Describing Functions, *Automatic and Remote Control*, ed. J. F. Coales, London, Butterworths, 1961, p. 895.
7. Johnson, E. C., Sinusoidal Analysis of Feedback Control Systems Containing Nonlinear Elements, Trans. A.I.E.E., **71**, Pt. II, July, 1952, p. 169.
8. Kolmogorov, A. N., and Fomin, S. V., *Elements of the Theory of Functions and Functional Analysis*, New York, Graylock Press, 1957.
9. Truxal, J. G., *Automatic Feedback Control System Synthesis*, New York, McGraw-Hill, 1955, pp. 562–563 and 600–601.
10. Gille, J. C., Pelegrin, M. J., and Decaulne, P., *Feedback Control Systems*, New York, McGraw-Hill, 1959.
11. Sandberg, I. W., On the Properties of Some Systems That Distort Signals — II B.S.T.J., **43**, January, 1964, p. 91.
12. Sandberg, I. W., On the Properties of Some Systems That Distort Signals — I, B.S.T.J., **42**, September, 1963, p. 2033.
13. Sandberg, I. W., Signal Distortion in Nonlinear Feedback Systems, B.S.T.J. **42**, November, 1963, p. 2533.
14. Sandberg, I. W., On Truncation Techniques in the Approximate Analysis of Periodically Time-Varying Nonlinear Networks, Proc. Allerton Conf. Circuit and System Theory, 1963.

Digital Computer Simulation of a Four-Phase Data Transmission System

By M. A. RAPPEPORT

(Manuscript received June 21, 1963)

This paper discusses the performance of a four-phase data transmission system in the presence of delay distortion and impulse noise. The tool used in this investigation is digital computer simulation, including a new technique for introducing impulse noise.

The noise studies indicate the usefulness of the eye aperture for measuring the degrading effects of delay distortion on such a four-phase system. Using the eye aperture as a criterion, the effects on performance of sinusoidal, parabolic, quartic, and parabolically bounded sinusoidal delays are studied. Curves showing the resulting degradation in performance are obtained. These are used to find bounds on allowable delay for a certain allowable degradation for lines typical of equalized voice or group bandwidths, usual nonequalized voice-bands, and loaded cable type voice-bands.

I. INTRODUCTION

1.1 *Nature of Simulation and of the Telephone Plant*

Digital computer simulation can be a powerful tool in the study of data transmission systems. This paper begins with a discussion of the techniques that have proved useful in applying this tool to various data transmission systems, including a new technique for studying the effects of impulse noise. These methods are then used for a detailed simulation study of a four-phase data transmission system.

Digital computer simulation, in general, aims at studying some physical system by a mathematical model of the system on a digital computer. In particular we are concerned with a real data transmission system in a real telephone plant. The scope of this paper is thus the performance of data systems over a time-stationary transmission path chosen at random from an ensemble of such lines.

The major forms of interference which concern us are this time-

invariant transmission distortion and random additive impulse noise. Transmission distortion is characterized as nonuniform attenuation and/or nonlinear phase as a function of frequency across the transmission band of the system. Nonlinear phase is commonly characterized in terms of nonuniform envelope delay, i.e., envelope delay distortion. Impulse noise is defined here as any randomly occurring voltage (or current) disturbances characterized by the occurrence of larger numbers of high peaks or pulses of noise than would be present in Gaussian noise of the same power, interspersed with long, low-power ("quiet") periods. There are, of course, many other forms of disturbance in the telephone plant, ranging from line dropout and frequency offset to the special case of undersea cables where Gaussian background noise is the significant disturbance. While some of these factors are amenable to simulation, this paper is concerned only with the effects of transmission distortion and impulse noise.

1.2 *Reasons for Using Simulation*

The underlying reason for studying data transmission by digital computer simulation is the complexity of the data system itself. This includes the range and nature of the ensemble of possible telephone transmission facilities, difficulties both in specifying and working with impulsive type noises in closed form, and the analytic difficulties in investigating real modulators and demodulators. This complexity presents problems both in analytic and experimental approaches.

In the laboratory there are two major difficulties: first, obtaining insight into the basic workings of the system in an environment which is very difficult to control in the laboratory; second, obtaining sufficient flexibility to provide a controlled investigation of the wide range of conditions encountered in actual practice.

Whereas in the experimental approach one of the difficulties is to see the forest for the trees, the analytical approach has the problem of having to clear away too many of the trees to make the forest visible at all. That is, the analytic approach often has to make a large number of assumptions about the performance of a real system in order to make the analysis tractable in closed form. For example, analytical approaches generally substitute the more easily handled Gaussian noise for the actually present impulse noise.

It is an attempt to get the best of both worlds that leads one to simulation. One hopes that the environment and all the factors in it can be controlled, without forcing the investigator into too many simplifying

assumptions that may lead to misleading results. Because of this sort of hybrid approach, simulation has two major goals. First, insight is sought into how a system really works. Among other particular aims, one seeks the allowable assumptions that can be made in analyzing a system, the basic and secondary factors affecting system performance, possible new directions in design and so on. Second, simulation hopes to produce a catalog of the expected performance of a real system for a realistic range of transmission and noise environments.

II. SCOPE OF FOUR-PHASE RESULTS

Sections III and IV describe the simulation techniques used and the particular four-phase system considered. In this section the results obtained are briefly listed for purpose of reference. The results can be grouped in four classes: modem (modulator-demodulator) design, choice of criterion, numerical specification of performance for particular transmission conditions, and application of these numerical results to give general transmission design specifications.

The modem design results obtained (Section 5.1) are on the modulation envelope shaping. In particular, the desirable amount of overlap between successive pulses is obtained. The same results were obtained simultaneously and independently in a laboratory test of the physical system by P. A. Baker,[1] and thus are useful also as a check on the accuracy of the simulation.

Section 5.2 presents results on the choice of a criterion for measuring performance of the system. The factors underlying choice of a criterion have been considered elsewhere.[2] The basic aim is to find a simple measure which will have the property of correlating with the performance of the system over a range of transmission facilities in the presence of impulse noise. It is shown that the aperture or opening of the "eye" pattern (see Fig. 7 of Section 5.2) is a reasonably satisfactory criterion. Therefore, the eye pattern is used for the presentation of results in the remainder of the paper.

Whenever applicable, general transmission design specifications are intertwined with the specific numerical results. Section 6.2 presents numerical results on the distortion produced when delay can be defined as a sinusoidal function of frequency. These results are interpreted to show general transmission design requirements for group (i.e., 40-kc or greater) bandwidths. Section 6.3 gives numerical results for the distortion effect of a variety of other delay shapes, for example, parabolic delay, as a function of frequency.

Sections 6.4 and 6.5 consider two general design questions. The first is: given some means of specifying allowable delay, that is, some class of transmission lines, find the line producing maximum degradation in performance. With this information one can then give bounds on performance, if bounds on the delay are specified. For a wide class of classical delay specifications, the result is essentially a two-cycle sinusoidal delay across the band. The next section considers the width of the band over which delay must be specified to predict system performance. It is shown that, for a system transmitting N bits per second, it is necessary to specify the delay over about $0.7N$ cycles of bandwidth, i.e., from the carrier $\pm 0.35N$ cycles.

Section 6.6 sums up the results on delay distortion to produce general transmission design specifications. The results are given in the form of curves of allowable delay vs maximum degradation. Three sets of curves are given to correspond to various shapes of delay in real facilities.

The final section gives some results for attenuation distortion. These latter results are not intended to be a systematic presentation, but show the magnitude of the effects.

III. SIMULATION TECHNIQUES

3.1 *Introduction*

This section considers those general simulation techniques appropriate to the study of a data transmission system. Many of these techniques were introduced by R. A. Gibby.[3] The main new approach is to introduce impulse noise to obtain the conditional probability of error given a noise of a specified nature present. The system is considered in blocks. For each block we attempt to duplicate mathematically the action the real system performs in shaping, or more generally in operating on, an electrical signal. This is not an attempt to duplicate the action of a particular capacitor or resistor but rather to present mathematically the performance of an entire system block. Further, advantage is taken of characteristics of certain system blocks to materially simplify the real system. For example, blocks in sequence which are commutative can be reversed without affecting the simulation results.

Fig. 1 shows two block diagrams of a data communications system. The upper figure is a common representation of a physical system. The lower figure is an equivalent model useful for simulation analysis. This equivalence is presented as an aid in understanding the nature of communication system simulation. Hereafter, we will call the upper figure the P (or physical) model and the lower figure the S model.

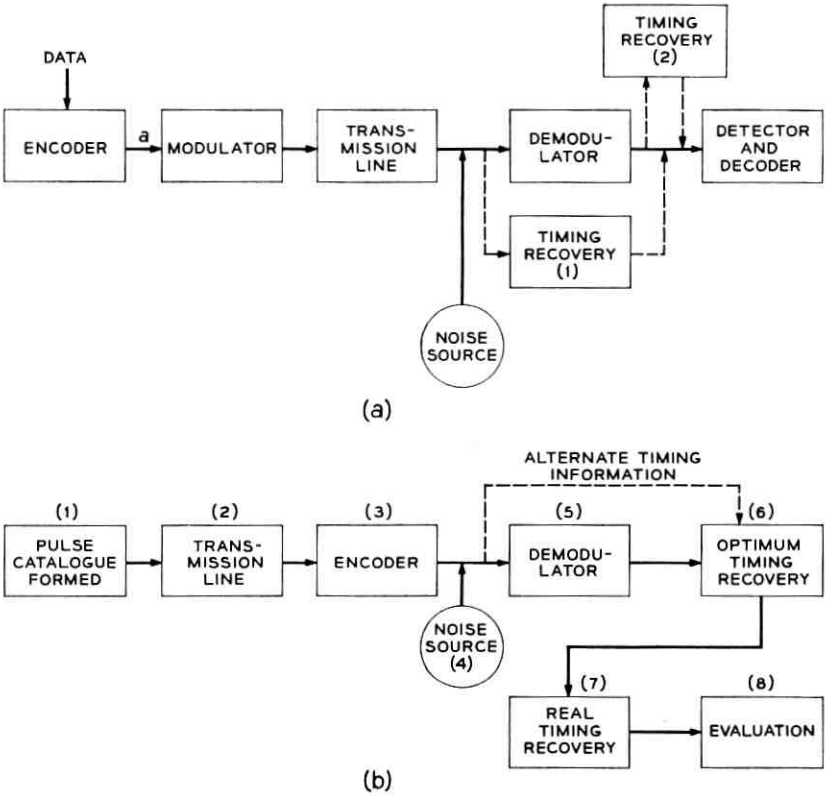


Fig. 1 — Block diagram representations of communications systems: (a) physical (P) model, (b) simulation (S) model.

3.2 *The Modulator and the Transmission Line*

The first box in the S model provides a catalog of all possible individual pulses used in the system. Since a digital computer is a discrete system, each pulse is defined by a sequence of sample values at a succession of evenly spaced time points. For a binary FSK system, for example, such a catalog would be one pulse each at the basic (mark and space) frequencies of the system.

The next step in the S model is to pass each pulse through the transmission medium. We emphasize that the pulses are acted on by the medium before they are encoded into a pulse train. This is justified by the commutative properties of filtering (i.e., action by the transmission medium) and adding together individual pulses displaced in time

(i.e., encoding of information). The third box in the S model then forms a pattern of pulses corresponding to the given pattern of encoded information.

The operation of the transmission facility on each pulse in the catalog is simulated in the frequency domain for reasons to be discussed shortly. Therefore, first a Fourier series of a particular pulse is formed in the computer. Next this Fourier series is modified by the attenuation and phase characteristics of the transmission facility. This operation will be discussed in greater detail in Section 5.3. The distorted pulses are then reformed in the time domain. These distorted pulses are then used to form a signal wave train corresponding to encoded information.

While the legitimacy of reversing blocks 2 and 3 follows from the commutativity of the two blocks, it is not yet clear why we should want to do this in a simulation. The basic motivation is the tremendous increase in speed when a simulation is performed in this way. Instead of having to process a long train of pulses through the transmission medium, it is now necessary to process only the small catalog of pulses inherent in the modulation technique. The superposition of the string of pulses is then a simple operation (which would in any case have to be done, no matter what the sequence of the blocks).

In addition, this approach gives a bonus of defining clearly individual pulses and their spectrums. We will see, in the section on sinusoidal delay distortion, the insight that is possible into performance of a system by the examination of individual pulses.

It is worthwhile to consider also why the transmission medium characteristics are simulated in the frequency and not the time domain. The underlying reason is that most practical knowledge of the plant is presently known in frequency domain parameters. By substituting for a Fourier approach a convolution of the impulse response of the line with each of the possible input signal pulses, a simulation can be easily modified to operate in the time domain. The advantages of reversing the sequence of boxes 2 and 3 of the S model hold just as strongly, and for the same reasons, in the time domain as in the frequency domain.

To make the greatest use of the long bit patterns available by simulating in this way, we desire a bit pattern that is representative of a random pulse train. What we actually use is what we will hereafter call a pseudo-random pulse train. For some integer h we desire to obtain every possible sequence of h bits the same number of times. For example, if h equals 3 there are eight possible combinations:

$$000, 001, 010, 011, 100, 101, 110, 111.$$

Any sequence in which each of these possible combinations of N

bits occurs exactly the same number of times will have the same distribution as a random sequence for all bit sequences of length up to and including N. For example, a sequence where each of the above occurs exactly once is given by

$$1\ 1\ 1\ 0\ 0\ 0\ 1\ 0\ 1\ 1.$$

A discussion of generation of such sequences can be found in Peterson.[4]

3.3 *The Demodulator*

Discussion of box 4, the introduction of noise into the simulation, will be postponed until later in this section. Box 5 of the S model represents the demodulator of the data transmission system. The signal processing of the demodulator is simulated on the computer. The basic approach is to duplicate the action of each block of the demodulator on the sequence of amplitude samples representing the data line signal. The output of the demodulator in the S model is then a sequence of amplitude samples representing the restored but distorted data pulse train.

Timing recovery in the S model is done in two separate steps. The first step, represented by box 6, is basically synchronization of the data train. The aim is to find the optimum sampling point relative to some criterion. Several criteria might be considered, but regardless of what criterion is used a synchronized or optimum timing point is obtained.

The second step in timing, represented by box 7, is to introduce the timing jitter which would occur in a physical system. Thus this box represents the physical timing recovery error, both that due to the jitter inherent in the circuits themselves and the jitter due to distortion of the data signal. The effect is to duplicate the degradation in performance due to imperfect timing. In keeping with methods used physically, the timing recovery signal can be generated either from the received line signal or the signal out of the demodulator.

3.4 *Noise Impairment — Performance Evaluation*

The last step in a simulation is to evaluate or measure the performance of the system. Some criterion of performance is chosen and is implemented in the simulation. For example, an eye pattern to measure system performance by the eye aperture might be formed. The eye pattern is formed by superimposing all possible three-bit intervals. In Fig. 7(a), a reference eye is shown. Fig. 7(b) shows some of the traces of a distorted waveform, and the resulting eye. The complete distorted eye would include all possible three-bit traces. The eye aperture is de-

fined as the minimax opening—that is, the maximum opening of the "worst" bit pattern. Various coding schemes can be introduced at this point, and the improvement they produce evaluated according to the chosen criterion.

We return now to discussion of box 4, the noise input, postponed in the previous presentation. Consider a noise burst of some particular size and shape introduced into a data system. We are interested in determining the conditional probability of error given this noise. In simulation the random occurrence of the noise pulse is replaced by the systematic introduction of a noise pulse at each of a large number of points in the data pulse train. As discussed above, the pulse train itself has each data sequence of a certain length, say 10 bits, occurring the same number of times; i.e., the possible sequences of N bits have uniform probability of occurrence. This pseudo-random data train is then mixed with noise by assuming the noise to be additive. Since a particular noise pulse is introduced at a very large (e.g., 4500) number of points, and uniformly along the data train, the effect is to introduce the noise randomly with uniform probability at all points in the data train.

The major difficulty in using an approach of this type is to choose noise pulse shapes which are representative of the telephone plant. In the mathematical sense this problem is at present unsolved. However, despite this present restricted knowledge, useful results are obtained for the following reasons. First, there is a range of system transmission conditions of interest for which relative performance is reasonably invariant under a range of different noise shapes. That is, although the absolute performance (i.e., the conditional probability of error) of each system changes from noise shape to noise shape, the relative performance, or ranking, of the systems stays the same over this range of noise shapes. Thus the degradation introduced by a particular transmission medium can be measured. Second, it is quite possible to handle a fairly large number of noise shapes. For example, in the four-phase case nine noise shapes were handled in quite reasonable computer times. Thus performance may be catalogued for various kinds of noise pending further knowledge. Finally, this implies a third basic use, which is that even if the system does perform differently under different wave shapes, this very information is useful in indicating both the basic nature of and possible improvements in a system.

The introduction of noise into the simulation can be used in various ways in measuring performance of the system. First, if the noise allows some ranking of transmission facilities, such as discussed in the paragraph above, it may be possible to directly correlate these rankings

with some criterion of performance of the system. This would then justify using this criterion in practice. For example, a justification for use of the eye aperture as a criterion for expected relative performance in four-phase transmission will be presented later. Second, it may be possible to use such a simulation in the design of better error detecting and correcting codes. That is, in the process of analyzing a long sequence of bits of a pseudo-random nature, it may appear that certain patterns of errors are more likely to occur in a realistic noise environment. Using such information, simpler error detecting and correcting codes on real lines might be obtainable. This is one approach to the construction of error codes on lines in which the memory is of a very complicated nature.

IV. FOUR-PHASE SYSTEM

4.1 *Physical*

We come now to the four-phase system which will be our prime concern for the remainder of the paper. The physical system we consider has been described by P. A. Baker.[1] The modulator of this system is shown in Fig. 2. Eight sine waves with the relative phases shown in the diagram are generated. These occur in two groups of four, as shown, and the data are encoded by choosing alternately from these two groups. The information is actually contained in the change in phase

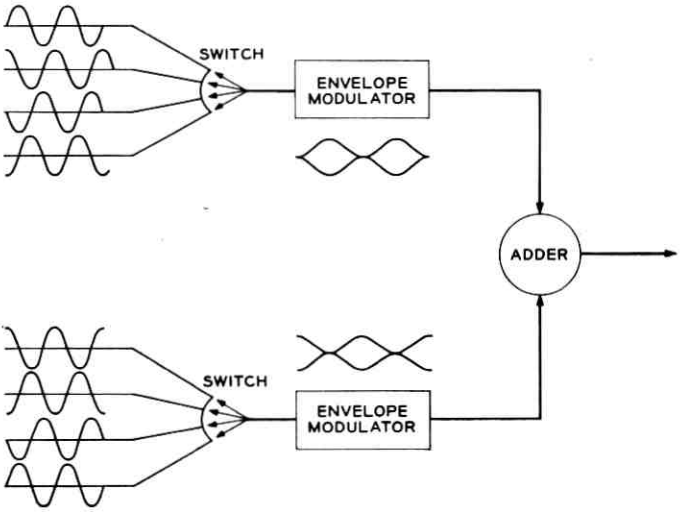


Fig. 2 — Modulator.

between successive pulses. Since each pulse encodes two bits of information, it has been given the name "dibit." The two alternating channels are used to insure there is always a change in phase between successive dibits. From the diagram it can be seen that this change is always one of four odd multiples of $\pi/4$. The actual shaping of the envelope of the sine waves is shown in the diagram. Just what this shaping should be to optimize performance was the first major check on the accuracy of the simulation. It will be considered in detail in Section 5.1.

The demodulator of the system is shown in Fig. 3. The information is recovered by comparing successive dibits to determine the change in phase between them. For this purpose the modulator uses two quite similar channels. We consider in detail only the upper one of the figure. The incoming signal is delayed by an interval of 1 dibit and then multiplied with the new incoming signal. This therefore results in a multiplication of successive dibit intervals. The output of the multiplier is then integrated. The result will be either positive or negative, depending on the relative phase of the two successive dibits. A truth table to recreate the encoded information is shown on the figure. This truth table

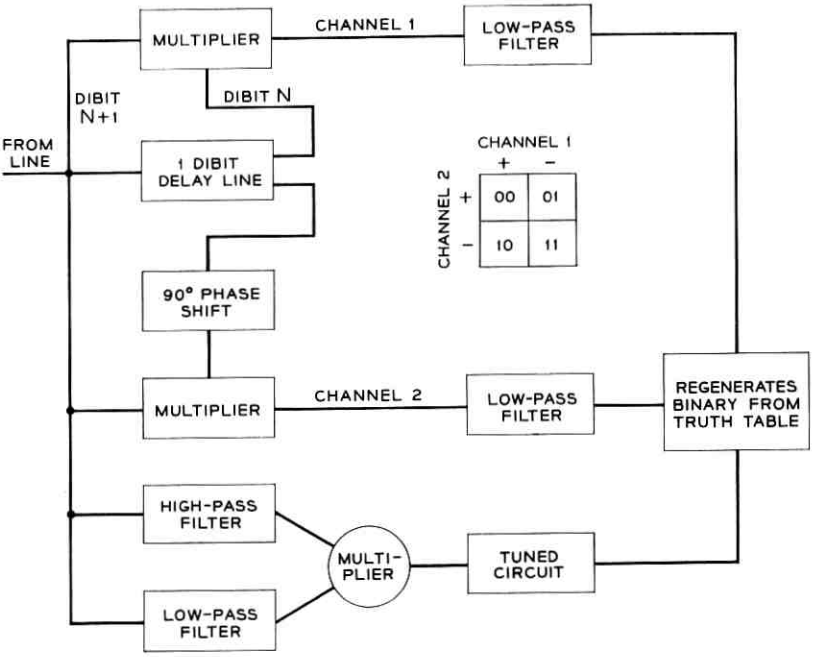


Fig. 3 — Receiver.

also demonstrates the result of the lower channel, which differs from the upper channel only in that it uses an extra $\frac{1}{4}$ cycle of delay. For greater detail the reader is referred to the paper by Baker.[1]

4.2 *Simulation*

This section is concerned with the application of the general techniques of simulation to the specific four-phase system. The simulation processes a data input train of 512 dibits. This corresponds to 1024 bits of input information, or to all possible combinations of four successive dibits. An extension of the exhaustive pattern technique mentioned above had to be developed for this four-phase case. It is no longer sufficient to make a choice between 0's and 1's for each information time slot. In the first time slot it is possible to pick any one of the eight possible signal waves, as shown in Fig. 2. Due to the alternation between the two channels of the modulator, it is possible to choose any one of four signals for each time slot after the first one. Thus, there are a total of $(8 \times 4 \times 4 \times 4 = 512)$ possible four-dibit combinations in the pseudo-random data train. For each sequence of four dibits there is another sequence different only in polarity; i.e., sequence A is simply the negative of sequence B. Since the multiplier eliminates this polarity difference, there is a seeming redundancy in using a pattern of this length. However, the noise introduced is of one polarity only. Thus this redundancy is necessary to obtain representative results with noise present.

There are two channels shown in the demodulator of Fig. 3. Initially both channels were simulated. It seems intuitively reasonable that the over-all results in these two channels should be essentially the same. For example, the signal sequence of initial phases -90, 45 should, with the extra 90° delay, give approximately the same results in the lower channel as the signal sequence 0, 45 gives in the upper channel. This is true within the limitations of slightly differing end effects between the two possible data sequences. There is such an image in the lower channel for every data sequence in the upper channel. One of the first tests performed with the working simulation was to check this hypothesis of approximately the same over-all results from the two channels. Over a fairly wide range of cases considered, no substantial difference in the performance of the two channels on an over-all basis was found. Therefore, the results given in this paper were obtained using only the upper channel of Fig. 3. The results apply to either channel operating alone, or to the system operating with both channels.

The results presented were obtained using nine basic noise wave-

shapes. These are shown in Fig. 4. In the first row are shown three sinusoidal pulses. These pulses differ in the ratio of their frequency to that of the dibit speed (or equivalently the carrier frequency) of the system. Similarly, the second row shows sinusoidal pulses subject to a one-sided exponential decay, and the third row shows two-sided exponentially decayed sinusoids. Each noise was considered over a range of noise to signal amplitudes. This range will be described more fully in Section 5.2. The noises shown were picked for two reasons. First, they represent a fairly wide range of parameters and shapes. Second, impulses idealized in this way are suggested by experimental studies of the telephone plant such as that of J. H. Fennick.[5]

V. BASIC SIMULATION RESULTS

5.1 *Simulation Check — Envelope Shaping*

The first step in using a simulation must be to check its performance against the physical model it represents. In the case of four-phase simulation this first check was provided by an investigation of the optimum envelope shaping (in some particular class of functions) in the modulator.

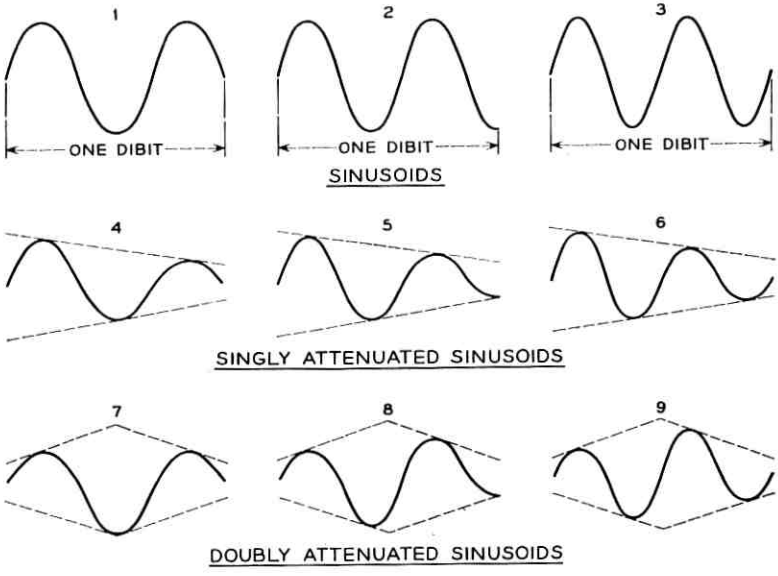


Fig. 4 — Noise waveshapes; all shapes are shown prior to detection but subsequent to transmission.

Prior to the envelope shaping the basic four-phase pulse is given by

$$f(t) = \sin [\omega_{car} t + n(\pi/4)] \qquad n = 0,1,2,3,4,5,6,7 \tag{1}$$

where n represents the information content of the wave. Consider first a raised-cosine envelope shaping. Then a four-phase pulse is denoted by (upper indicates upper channel of modulator)

$$\begin{aligned} f_{\text{upper}}(t) &= \sin\left(\omega_{car} t + n\frac{\pi}{4}\right)\left[\frac{1}{2} + \frac{1}{2}\cos\frac{\omega}{2}\,\text{dibit}^t\right] \qquad n = 0,2,4,6 \\ f_{\text{lower}}(t) &= \sin\left(\omega_{car} t + n\frac{\pi}{4}\right)\left[\frac{1}{2} - \frac{1}{2}\cos\frac{\omega}{2}\,\text{dibit}^t\right] \qquad n = 1,3,5,7 \end{aligned} \tag{2}$$

A typical sequence of such pulses is shown in Fig. 5(a). However, as can be seen in Figs. 2 or 5(a), this full raised-cosine envelope builds in an overlap between successive dibits. One alternative to such shaping is phase shift keying—that is, a jump from one phase to another at dibit intervals. However, since the transmission lines have finite bandwidth, such phase jumps would be distorted in transmission. To provide a controlled smooth transition, rather than the uncontrolled distortion resulting from the action of finite bandwidth on phase jumps, while minimizing the effect of the built-in overlap of the raised cosine, the basic functional form of the envelope given by (3) was considered

$$\begin{aligned} \text{env}_{\text{upper}}(t) &= \frac{\frac{1}{2}\cos\frac{\omega_d t}{2} - \frac{1}{2}\cos\frac{\omega_d T}{2}}{\frac{1}{2} - \frac{1}{2}\cos\frac{\omega_d T}{2}} \qquad 0 \leqq \left|\frac{\omega_d t}{2}\right| \leqq T\pi \\ &\qquad\qquad \tfrac{1}{2} \leqq T \leqq 1 \\ &\qquad\qquad T\pi \leqq \left|\frac{\omega_d t}{2}\right| \leqq \pi \\ &= 0 \\ &\qquad\qquad \omega_d = \text{dibit radian frequency} \end{aligned}$$

and similarly for the lower channel envelope. The parameter T controls the amount of built-in overlap of the modulation envelope. The optimum value for T agreed when found independently in a laboratory test of the working system by Baker and on the computer. This provided a check on the simulation.

The optimum value of T is $\frac{3}{4}$. This corresponds to a modified raised cosine with about $\frac{1}{4}$ dibit overlap, as shown in Fig. 5(b) (that is, about 50 per cent of the maximum overlap of Fig. 5a). The shaping of this modulation envelope in effect determines the amplitude spectrum of the line signal. In the course of subsequent investigation on the computer, another class of amplitude shaping was also considered. A typical mem-

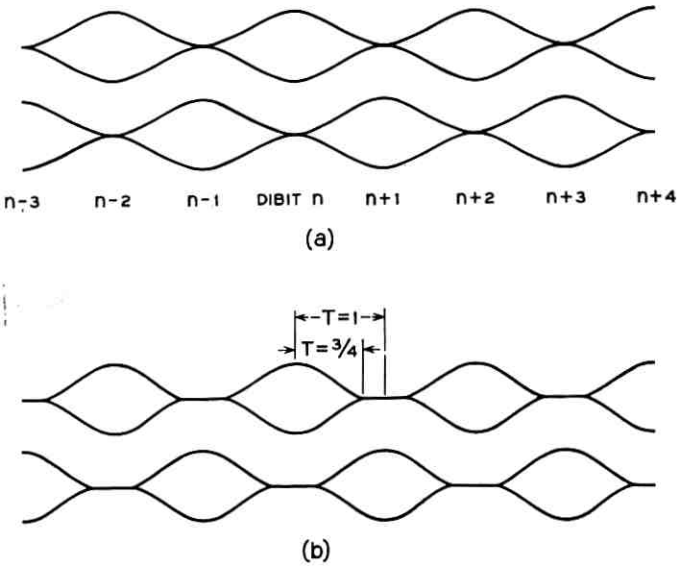


Fig. 5 — Envelope shaping: (a) full raised-cosine shaping, (b) modified shaping used in present data terminals.

ber of this class of squared-off raised-cosine amplitude spectra is shown in Fig. 6 along with the spectra resulting from a complete raised-cosine shaping (that is $T = 1$) and for the optimized $T = \frac{3}{4}$. A sequence of values of the parameter b (see Fig. 6) of the squared-off raised cosine was considered. No value of b gave performance as good as the modified raised cosine with $T = \frac{3}{4}$.

5.2 *Criterion — Eye Pattern*

The measure of performance used in obtaining the results described in the above paragraph was the eye pattern (Fig. 7). However, while the accuracy of the simulation can be checked by comparing the eye obtained in a laboratory test to that obtained by the simulation, this does not in itself justify the eye as a suitable measure of performance for the four-phase data system. This justification of the eye depends in the final analysis on demonstrating a correlation between the expected performance of the system in a real noise environment and the opening or aperture of the eye.

Impulse noises of nine different shapes were introduced into the simulation. Each of these was considered over a signal-to-noise ratio of about 14 db, from noise of +5 to −9 db relative to the signal. The

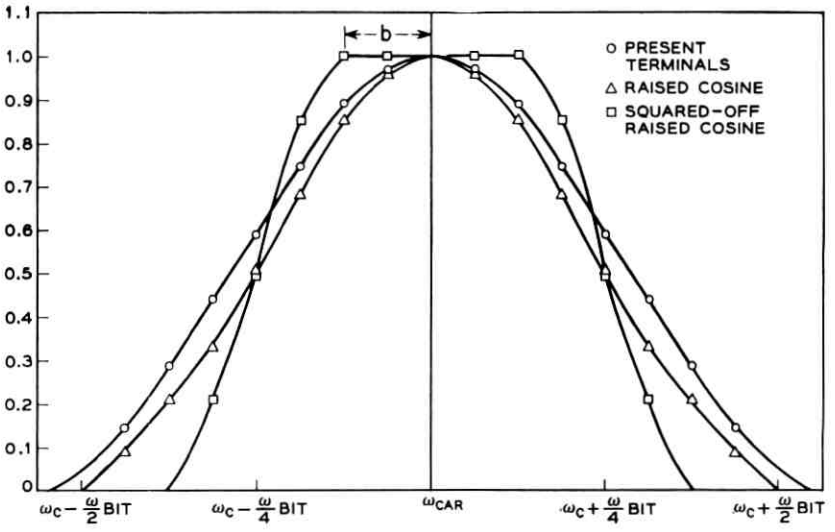


Fig. 6 — Envelopes of amplitude spectra.

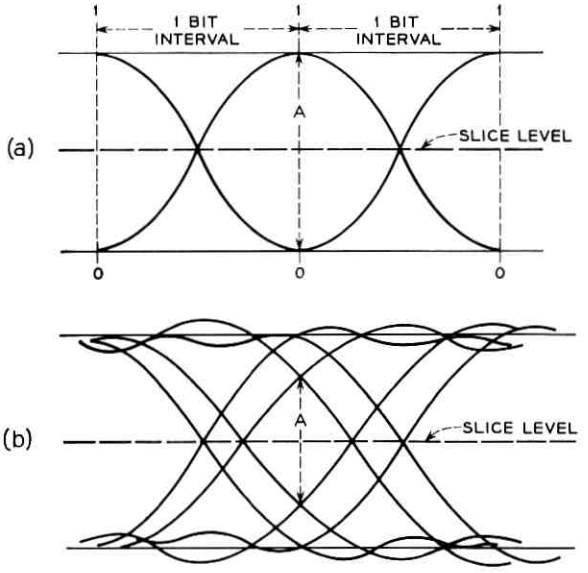


Fig. 7 — Eye patterns. (a) Undistorted eye: tracings result from patterns 000, 001, 010, 011, 110, and 111; A is eye aperture normalized for undistorted eye to 1.0. (b) Distorted eye: figure shows some tracings; total eye formed from *all* possible three-bit intervals.

signal-to-noise ratio was defined as the peak undistorted signal amplitude relative to the peak noise amplitude.

The arbitrary nature of this definition is not significant, since it does not affect the ranking or relative performance of various lines. A sample of the many cases considered is shown in Fig. 8 for two noises and two transmission lines (one of which is back-to-back or reference transmission). As might be expected, all of the lines perform approximately the same at very high noise levels. At such levels the noise completely swamps out the signal, and the resulting performance is then the totally random effect of the noise alone. We therefore choose the range of conditional probability of error for which the results are meaningful in terms of degradation in performance. To avoid misunderstanding, we note that while performance is certainly sensitive to noise waveshape,

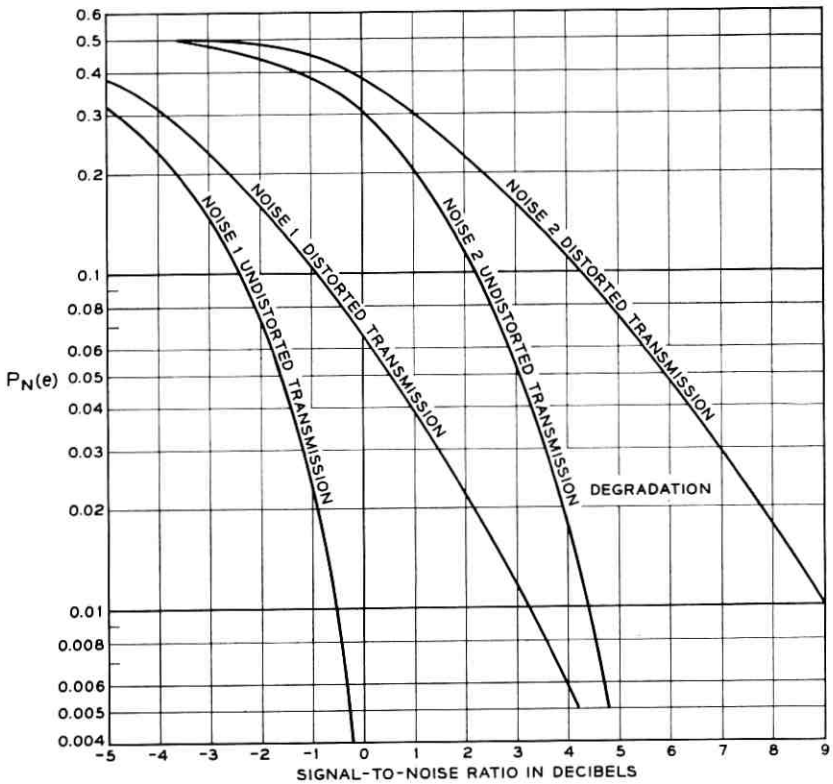


Fig. 8 — Conditional probability of error $|P_N(e)|$ vs S/N [db (peak undistorted signal/peak noise)]. Shows results of two sample noises on a distorted and an undistorted transmission line.

much of the difference seen in Fig. 8 comes from defining S/N using the peak value, and not some average total power measure for the noise.

The range of conditional probability of error chosen for these tests was between 0.005 and 0.125. For a particular transmission line and conditional probability of error, there is a range of degradation of performance as measured by the change in signal-to-noise ratio for various noise shapes. A sample of the noise results obtained is given in Table I for a range of transmission lines in the following manner. First the transmission line is described with a sufficient number of parameters to characterize it. For example, sinusoidal delay is described in terms of its amplitude (Bm), its frequency (m) and its phase θ relative to the transmission band. A detailed discussion of each parametric representation will be described later. Numerical results are then given for each of three levels of conditional probability of error, namely 0.005, 0.025, and 0.125. These results are the range in degradation and signal-to-noise ratio due to the nine noises considered for the given transmission line relative to back-to-back performance. The correlation in performance between results obtained using conditional probability of error and the eye aperture can be seen by comparing the degradation in performance measured by successively smaller $P(e)$ (i.e., for successively smaller noises) with the degradation measured by the aperture shown in the last column. It is seen that the eye seems to give a limiting value.

Figs. 11–13 of Section 6.2 are presented in terms of eye aperture. Presentation in this form is justified by the fact that, for the range of transmission lines and noise considered, the value of eye aperture correlates consistently with the degradation in performance obtained by keeping constant conditional probability of error. This justification is one of the principal results of this paper. While this does not demonstrate that the eye will be a good measure of performance for all possible distortions, the wide range of noises and transmission facilities considered indicates the eye to be an appropriate criterion for four-phase transmission, at least over lines whose principal distortion is nonuniform delay.

5.3 *Transmission Line Simulation*

A typical four-phase data pulse is given by the product of $f(t)$ of (1) and envelope (t) of (3). As mentioned previously, passing these pulses over a transmission line was done in the simulation in the frequency domain. Thus for each of these pulses a Fourier series was formed. A typical example of such a series is given in (5), where ω_0, the fundamental frequency of the Fourier series, determines the spacing of the spectral lines which represent the wave:

TABLE I—RANGES OF CONDITIONAL PROBABILITY OF ERROR

All numbers are degradation in db relative to no distortion—range represents variation over nine noises.

Delay	Degradation (in db) Measured Using Conditional Probability of Error			Degradation (in db) Measured Using Aperture
	$P(e) = 0.125$	$P(e) = 0.025$	$P(e) = 0.005$	
Sinusoidal				
$Bm = 0.5$, $m = 0.5$, $\theta = \pi/2$	0–0.3	0.5–0.8	0.7–1.2	1.1
1.5 0.5 0	1.4–1.8	3.1–3.7	3.8–4.5	4.9
1.5 0.5 $\pi/2$	2.9–3.5	6.2–6.6	8.2–8.9	9.6
1.0 1 0	1.3–1.7	3.1–3.8	4.2–5.0	4.8
1.0 1 $\pi/2$	2.8–3.7	6.4–7.0	8.6–9.3	10.2
1.0 1.25 $\pi/2$	3.0–3.8	7.1–7.6	9.7–10.5	11.7
0.5 1.5 0	0.6–1.0	1.8–2.2	2.5–3.1	3.1
1.0 1.5 0	3.9–4.4	8.5–8.8	11.4–11.7	12.8
1.0 1.75 0	4.5–5.1	9.8–10.5	13–14.5	15.8
0.5 2.0 $\pi/2$	0.3–0.7	1.5–1.8	2.1–2.8	3.1
1.0 2.0 $\pi/2$	2.0–2.7	5.6–6.3	8.1–9.0	10.1
1.5 2.5 0	5.6–6.1	12.1–12.7	17.5–19.0	24.5
Parabolic				
1.0 bit delay at $\omega_{car} \pm 0.35\ \omega_{bit}$	0.3–0.4	1.1–1.2	1.6–1.9	1.8
1.5 bits delay at $\omega_{car} \pm 0.35\ \omega_{bit}$	0.8–1.0	2.1–2.5	3.3–3.4	3.7
1.5 bits delay at $\omega_{car} + 0.275\ \omega_{bit}$ $\omega_{car} - 0.425\ \omega_{bit}$	1.1–1.7	3.2–3.7	5.0–5.1	5.8
1.5 bits delay at $\omega_{car} + 0.2\ \omega_{bit}$ $\omega_{car} - 0.5\ \omega_{bit}$	2.5–3.1	6.6–6.8	10.3–11.3	12.4
Quartic				
1.25 bits delay at $\omega_{car} \pm 0.35\ \omega_{bit}$	0.2–0.8	0.8–1.3	1.7–2.2	1.9
1.25 bits delay at $\omega_{car} + 0.275\ \omega_{bit}$ $\omega_{car} - 0.425\ \omega_{bit}$	0.7–1.3	2.5–2.7	3.5–4.0	4.8
1.0 bit delay at $\omega_{car} + 0.2\ \omega_{bit}$ $\omega_{car} - 0.5\ \omega_{bit}$	2.1–2.6	5.3–5.8	8.5–9.8	11.3
Band cutoff cases				
Quadratic delay to $\omega_{car} \pm 0.35\ \omega_{bit}$ — then delay uniform	0.6–0.9	1.4–1.7	2.3–2.5	2.5
Delay of Fig. 22, curve a	2.0–2.8	4.6–5.0	7.8–8.4	8.8
Delay of Fig. 22, curve c	1.1–1.6	3.6–3.8	5.2–5.9	6.3
Delay of Fig. 22, curve f	1.0–1.2	2.6–2.8	3.5–4.0	4.4
Attenuation cases				
Sinusoidal delay				
$Bm = 1.0$, $m = 2.5$, $\theta = 0$ 6-db slope atten. of Fig. 9(a)	2.4–3.2	5.2–6.0	7.8–8.5	8.9
Sinusoidal delay				
$Bm = 0.5$, $m = 2.0$, $\theta = 0$ 6-db slope atten. of Fig. 9(a)	2.6–3.2	4.6–6.2	7.5–8.1	9.9

$$f(t) \text{ env } (t) = g(t) \tag{4}$$

$$g(t) = \sum_{n=0}^{N} A(n\omega_0) \cos [n\omega_0 t + \psi(n\omega_0)]. \tag{5}$$

$A(n\omega_0)$ and $\psi(n\omega_0)$ are the amplitude and phase of the spectrum of the pulse across the transmission band of interest. The effect of passing such a pulse through a transmission medium is to yield an output pulse which can be represented by:

$$h(t) = \sum_{n=0}^{N} A(n\omega_0)\ R(n\omega_0) \cos [n\omega_0 t + \psi(n\omega_0) + \varphi(n\omega_0)]. \tag{6}$$

Here, $R(n\omega_0)$ and $\varphi(n\omega_0)$ represent the attenuation and phase of the transmission medium. The envelope delay $D(\omega)$ is the derivative of $\varphi(\omega)$ with respect to ω.

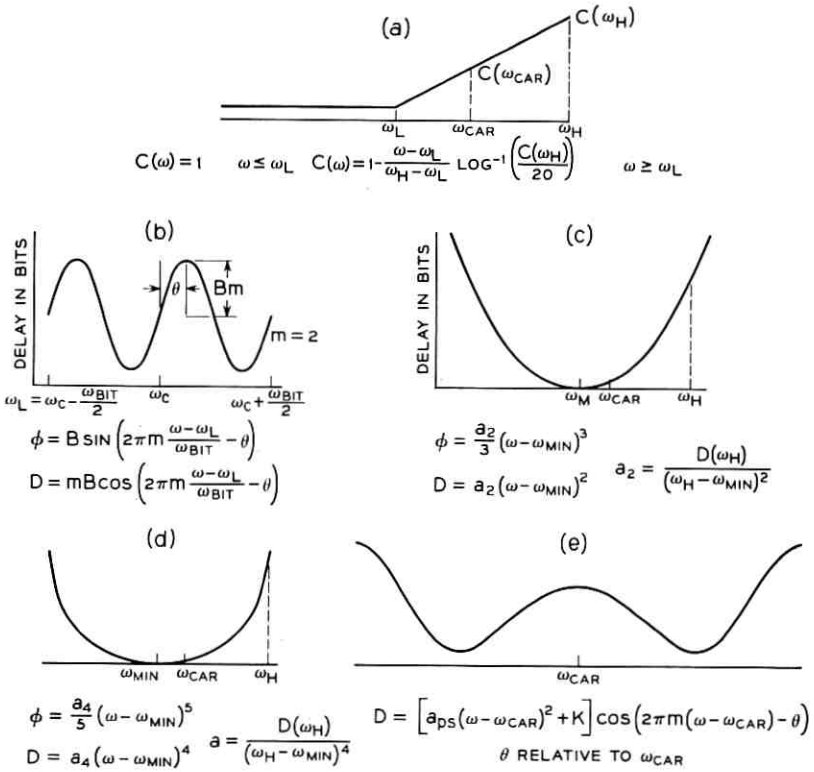


Fig. 9 — Attenuation and delay shapes.

The parametric forms of $R(\omega_0)$ and $D(\omega)$ are given in chart form in Fig. 9. Aside from a flat passband, the primary form of attenuation considered was one which seems to be typical of a large part of the voice-band plant. It consists of a flat passband out to some cutoff frequency, followed by a slope attenuation of so many db per cycle. We note that this is not db per octave; that is, the curve is linear with respect to frequency and not with respect to the log of frequency. In addition, some attenuations previously introduced, whose spectra are shown in Fig. 6, were also considered. Among these is the effect of varying the built-in overlap or intersymbol interference, which is, of course, also a variation in the amplitude spectrum of the pulses.

The general forms of delay considered were sinusoidal, parabolic, quartic and a parabolic bounded sinusoid. In addition, many of these forms were also used with variations in the band edges of the delay. For example, sinusoidal delay might be used across 70 per cent of the band and then a relatively sharp cutoff delay used at the band edge. Finally the simulation is set up to handle frequency by frequency read-in of both phase and attenuation across the transmission band.

VI. DELAY RESULTS

6.1 *Introduction*

The remainder of this article is a presentation of the results achieved using the simulation to investigate performance of a four-phase system over a telephone channel. While some raw data results on impulse noise are presented in Table I, for the reasons described above the remainder of the discussion will refer to the eye aperture results shown in curve form in a group of figures.

6.2 *Multicycle Sinusoidal Delay — Group Band Transmission Design*

We begin our discussion with the results obtained in an application which is at once the simplest to explain, has perhaps the clearest intuitive explanation of the effects of delay, and yet is of definite practical importance. This is the case illustrated in Fig. 9(b), in which the delay is a sinusoidal function of frequency across the transmission band of interest.

Such a sinusoidal delay is defined in terms of three parameters. First is the amplitude of the delay (*Bm* of the figure), which characterizes the peak delay in bit times. Second is the phase (θ), which represents the position of the sinusoidal delay relative to the carrier frequency of the system. Finally, the number of cycles of the sinusoid

across the transmission band of interest is given by m. This characterization is of particular interest in group bands. Typical delay shapes which arise due to group band separation filters might be those shown in Fig. 10(a). In order to permit transmission of data over group band channels these delay curves are equalized. The results of equalization typically yield a number of cycles of sinusoidal delay, such as shown in Fig. 10(b). The designer often has some control both over the magnitude and the number of ripples of delay, with a compromise to be made between technical factors and economics. That is, the greater the number of ripples required, in general, the more stages of equalization are needed, and thus there is an economic constraint involved.

The basic results on multicycle sinusoidal delay are shown in Figs. 11 through 13, where Fig. 13 is an overlay of the preceding two figures. A number of conclusions can be drawn from these results. For design

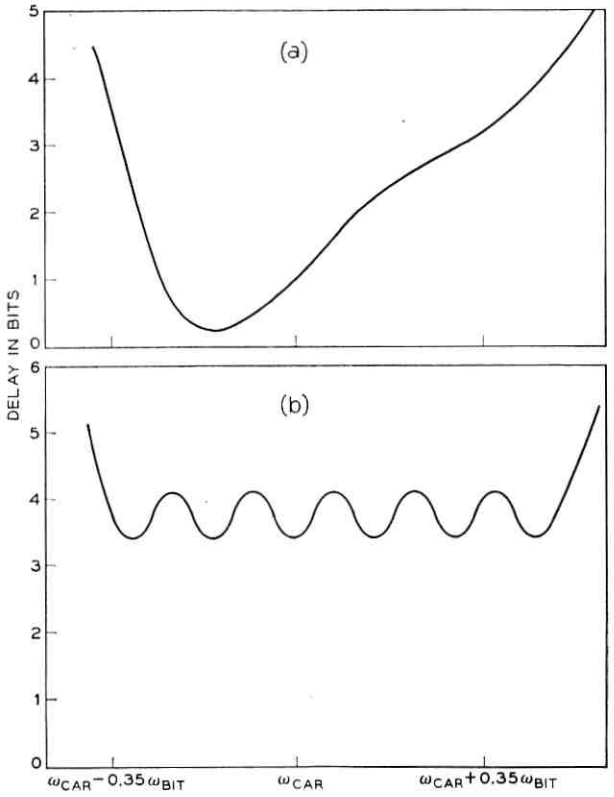


Fig. 10 — Group bands: (a) unequalized, (b) equalized.

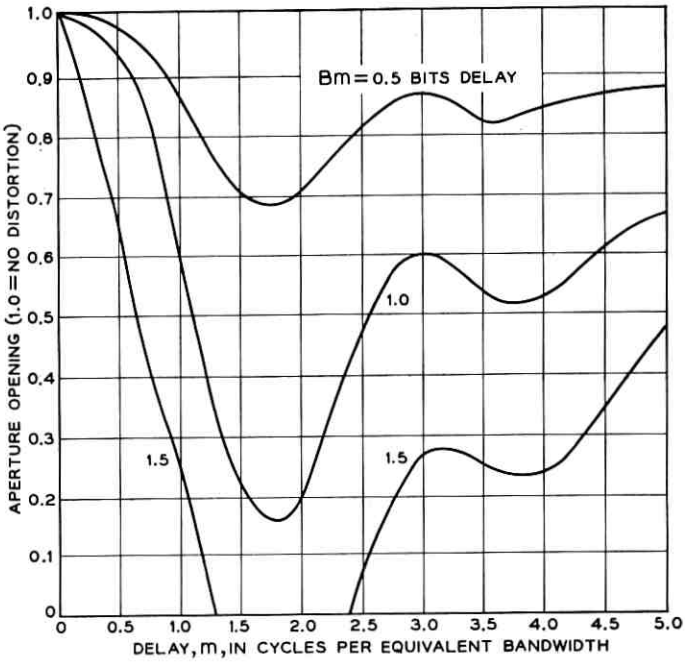


Fig. 11 — Aperture vs delay for sinusoidal delay; all curves use $\theta = 0$,

$$D = \frac{Bm}{\omega_{bit}} \cos\left[\left(\frac{\omega - \omega_c}{\omega_{bit}}\right)m + \theta\right].$$

Aperture = 0 indicates system makes errors even with no noise. Peak-to-peak delay = $2Bm$. Curves for Bm = 0.5, 1.0, and 1.5 bits peak delay.

purposes, two of these are most significant. First is the general improvement in transmission as the number of cycles of delay is increased while holding the peak delay in dibits constant. Second is the relative preference for an odd number of cycles of delay across the passband. This is clearest if $\theta = 0$ (i.e., the aperture for $m = 3$, $\theta = 0$ is larger than for $m = 4$, $\theta = 0$). If $\theta = \pi/2$ there is not much difference between $m = 3$ and $m = 4$. However, since the larger m will generally be more expensive to achieve by equalizing, $m = 3$ would again be preferred.

We consider first the improvement in transmission for constant peak delay. To a first-order approximation this can be explained on the basis of echo theory.[6] Since the delay is sinusoidal, the interference introduced is due primarily to echoes. In particular, simulation shows the first echo on each side of the pulse transmitted to be the primary cause of interference. The amplitude of any echo is proportional only to the

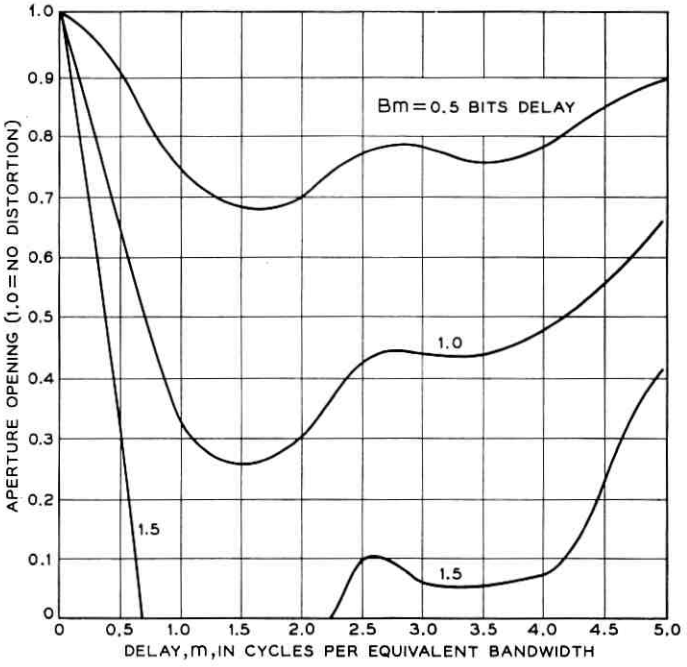


Fig. 12 — Aperture vs delay for sinusoidal delay; all curves use $\theta = \pi/2$,

$$D = \frac{Bm}{\omega_{bit}} \cos\left[\left(\frac{\omega - \omega_c}{\omega_{bit}}\right) m + \theta\right].$$

Aperture = 0 indicates system makes errors even with no noise. Peak-to-peak delay = $2Bm$. Curves for Bm = 0.5, 1.0, and 1.5 bits peak delay.

amplitude of the sinusoidal phase (B), and is independent of the other parameters of the delay. Thus as the number of cycles of sinusoidal delay increases, holding the product Bm constant, B decreases and so does the interference. Fig. 14 shows a typical four-phase system pulse and the echoes produced by the sinusoidal delay for various numbers of cycles (in frequency) with constant peak delay. The advantages of the simulation being able to show individually distorted pulses in such clear detail is most evident at this point.

On the other hand, the position of the echo is determined solely by the number of cycles of sinusoidal delay across the band; that is, the distance of the echo from the main pulse is directly proportional to m. The preference for odd numbers of cycles of delay can be explained, to a first-order approximation, as follows. The primary interference is the first echo in each direction. As the number of cycles of delay increases,

the effect on this echo is to move it farther and farther from the main pulse. We note that the amplitude of the echo is decreasing for a constant value of peak delay. In practice, this means that as the number of cycles of delay is increased, the interference progressively phases in and out with respect to the time at which the demodulated pulses are sampled. In other words, while the echo is always present, its major point of interference oscillates between the sampling instant and midway between the sampling instants of the demodulated pulse train. This oscillating phasing in and out with respect to the sampling time of the interference from the echo produces alternately more and less interference at the sampling instants.

An example is shown in Fig. 14. The peak of the first echo comes midway in the adjacent time slots of the two-cycle interference. This produces a maximum distortion in phase in the adjacent time slot —

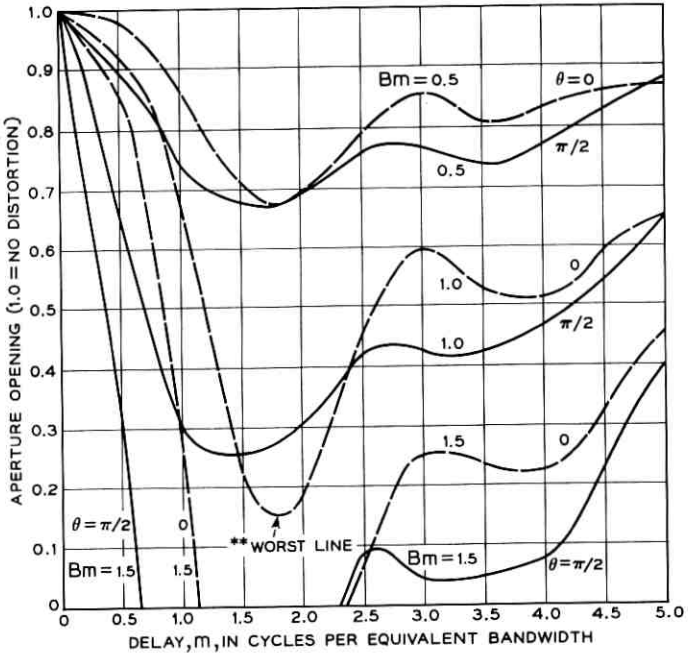


Fig. 13 — Combination of Figs. 11 and 12:

$$D = \frac{Bm}{\omega_{bit}} \cos\left[\left(\frac{\omega - \omega_c}{\omega_{bit}}\right) m + \theta\right];$$

aperture = 0 indicates system makes errors even with no noise; peak-to-peak delay = $2Bm$.

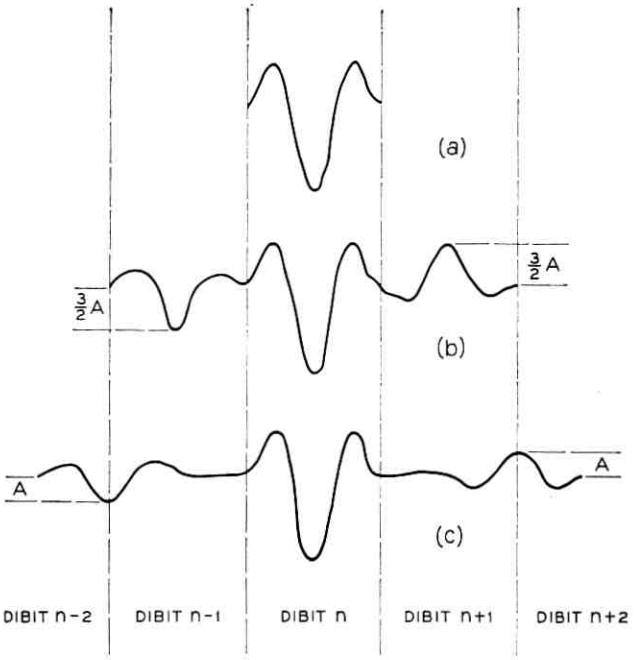


Fig. 14 — Effects of sinusoidal delay in terms of echoes: (a) basic four-phase pulse, (b) echoes from two-cycle/band sinusoid, and (c) echoes from three-cycle/band sinusoid.

that is, at the sampling instants. For the three-cycle interference, the echo reaches a peak midway between time slots. It therefore produces less distortion in either of the two time slots it interferes with, and a correspondingly lower maximum distortion.

6.3 *Delay Results Useful for Voice-Band Design*

This section introduces and discusses results over a variety of delay characteristics. The curves representing these characteristics share the property that they have no more than 3 minima. Their general shapes are such that they can be made to represent a wide variety of voice-band channels.

The first set of curves, shown in Fig. 15, is simply an enlargement of the effects of sinusoidal delay for less than two cycles of sinusoid. Fig. 16 shows some typical voice-band delay curves. For each an approximating sinusoid (with its associated parameters) is given. The results in Fig. 15 are for θ from zero to $\pi/2$ radians. However, since the

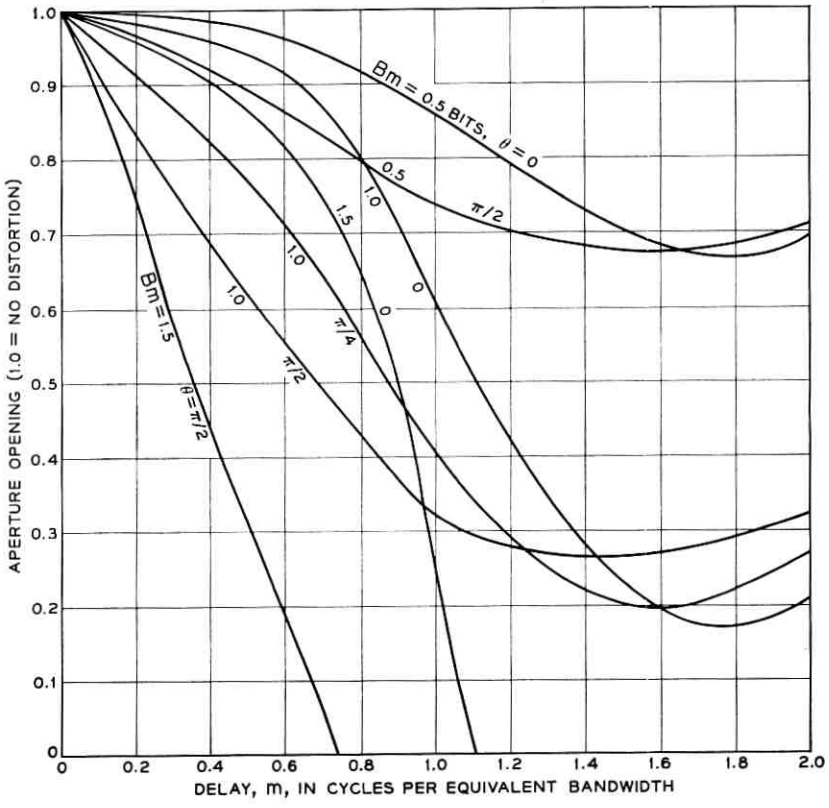


Fig. 15 — Aperture vs delay for sinusoidal delay — expansion for low m of Fig. 13.

$$\text{Delay} = \frac{Bm}{\omega_{bit}} \cos\left[\left(\frac{\omega - \omega_c}{\omega_{bit}}\right) m + \theta\right].$$

Aperture = 0 indicates system makes errors even with no noise present; peak-to-peak delay = $2Bm$.

results repeat in successive quadrants (i.e., $\theta = \pi$ gives essentially the same results as $\theta = 0$) the curves represent the full range of θ.

Figs. 17 and 18 give results for quadratic and fourth-power delay distortion. The curves represent delay symmetric with respect to the carrier, and displaced from the carrier by varying percentages of the bandwidth of interest. It is worthwhile to note again that these curves are, as are all the curves given in the results, normalized with respect to the carrier frequency and the bit speed. Thus one can apply the results to any frequency range and corresponding bit speed. The curves are

normalized for a $1\frac{1}{2}$ carrier cycle per dibit system. Thus, for example, a typical system at this speed would have an 1800 cycle carrier and transmit 2400 bits per second, i.e., 1200 dibits per second. Again, the same curves apply to a 48,000-bps system using a 36-kc carrier.

The results presented to this point represent relatively simple delay shapes—that is, sinusoidal, quadratic, or quartic delays. It will be shown in Section 6.5 that the same results apply within very reasonable limits to any delay curve which is identical with one of the delays considered over a minimum of 70 per cent of the band. In addition one class of relatively complicated delay shapes was also investigated. These delays are parabolically bounded sinusoids, and are described in Fig. 9(e). As the product of a parabola plus a constant with a sinusoid they can be used to represent a wide class of various delay shapes. The results are given in Figs. 19 and 20 for $m = 2$ and $m = 3$ and for a range of values of the ratio α_{ps} to K. It will be seen that these ratios are particularly

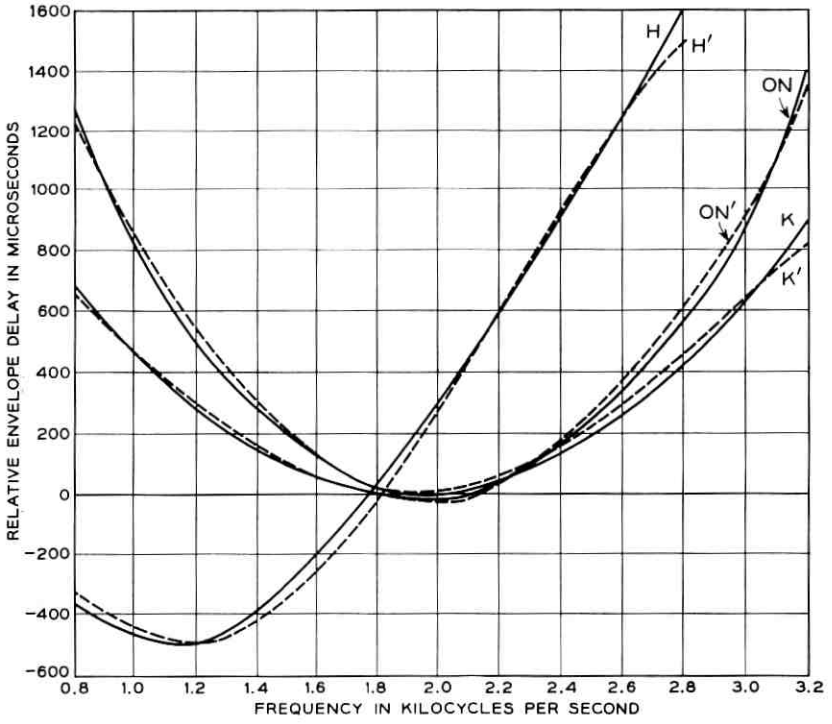


Fig. 16 — Representations of some typical lines by sinusoidal approximations (primes are approximations). K = 2 links K carrier, ON = 3 links ON carrier, and H = 300 miles of H44 loaded cable.

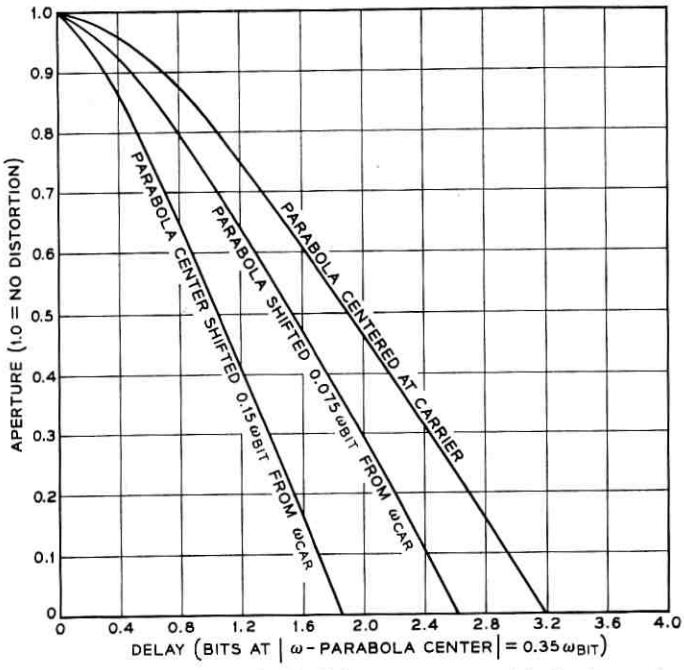


Fig. 17 — Apertures for parabolic delay; aperture = 0 indicates system makes errors with no noise.

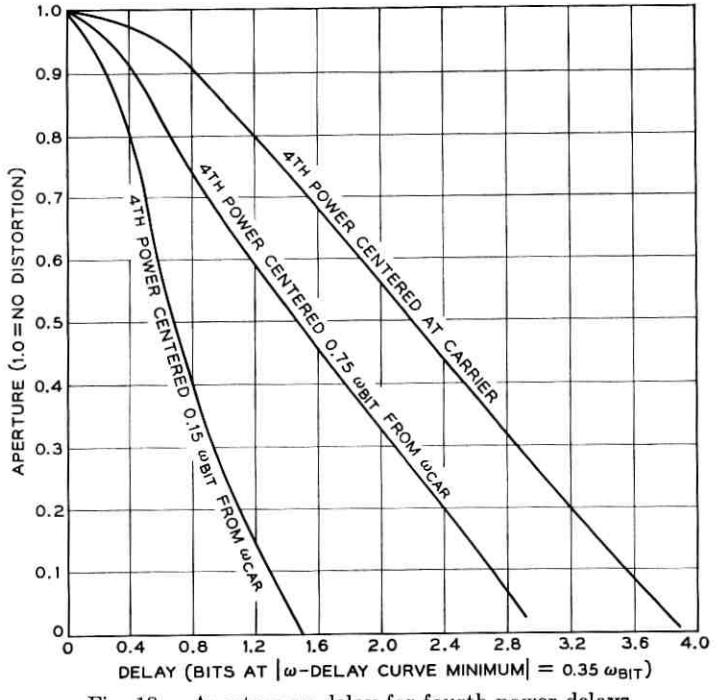


Fig. 18 — Aperture vs delay for fourth-power delays.

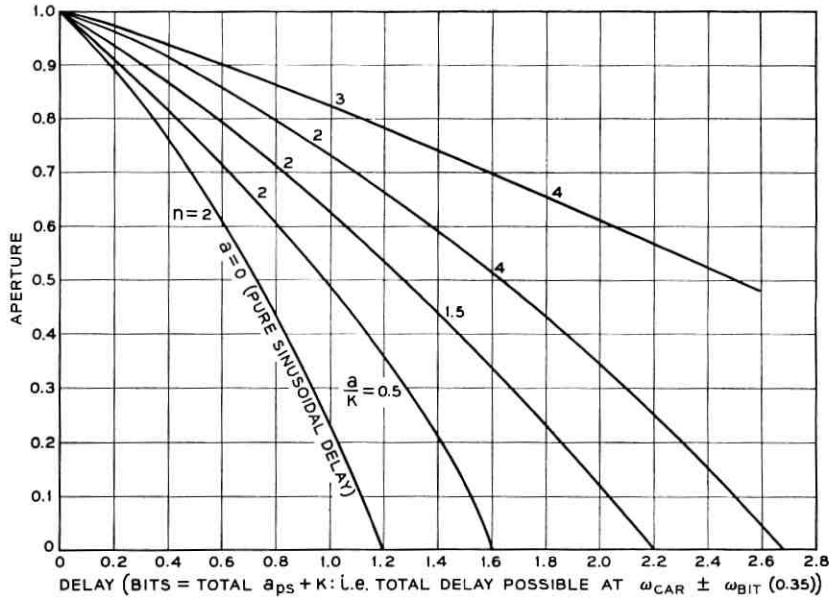


Fig. 19 — Aperture vs delay for parabolic bounded sinusoidal delays, $\theta = 0$.

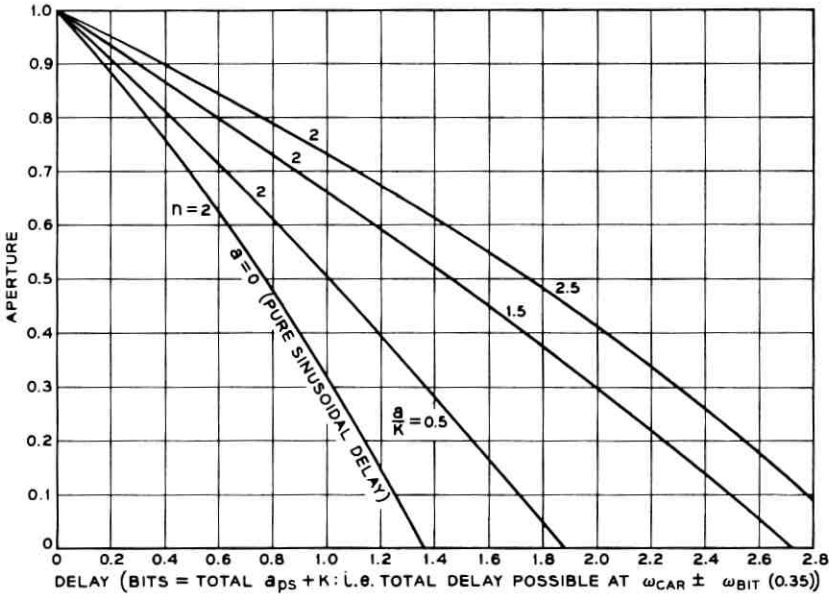


Fig. 20 — Aperture vs delay for parabolic bounded sinusoidal delays, $\theta = 90°$.

useful in the setting of curves of maximum degradation for performance in the voice band.

6.4 *Line in a Given Class Producing Most Degradation*

The simplest generally used way of characterizing delay is to specify the maximum allowable delay across some percentage of the transmission band. For example, one might say that the maximum delay across 70 per cent of the transmission band is to be no more than $\frac{1}{2}$ dibit (making a tacit assumption that the minimum delay is 0). It is of interest to specify the worst possible delay shape meeting such a requirement — that is, to find the delay shape falling within this allowable maximum which produces the worst degradation in performance. For a line of reasonably good performance, say for eye openings greater than 0.6, an answer to this problem has been provided by R. W. Lucky.[7] However, since the lines he allows include such pathologies as discontinuities in the delay, the limiting cases must often for our purposes be considered nonrealizable. It is therefore of interest to find the worst line out of the classes we have considered, using the above criterion of "worst." For a check this worst line then can be compared to that obtained by Lucky.

If the limiting value of delay is defined across 60 per cent of the band or greater, the worst performance in the classes discussed above is that obtained with a slightly less than two-cycle sinusoid across the transmission band. Fig. 21 shows in graphic form the difference in performance between such a two-cycle sinusoid and various other delay characteristics having the same maximum delay across a given percentage of the band. For the range of reasonably good transmission lines which Lucky has considered, Lucky's "worst" line produces only about one db greater degradation in performance as measured by the respective aperture values (or equivalently by the impulse noise performance; see Section 5.2) than this two-cycle sinusoid. When one remembers that Lucky's lines are in general quite drastic in their shapes, it seems reasonable to use this two-cycle sinusoid as an upper bound through the remainder of this paper.

6.5 *Effect of Delay at the Edges of the Band*

Examination of the spectrum of an individual pulse shows that the energy falls off very rapidly toward the edge of the band. Now when timing is performed using the line signal, the energy near the sides of

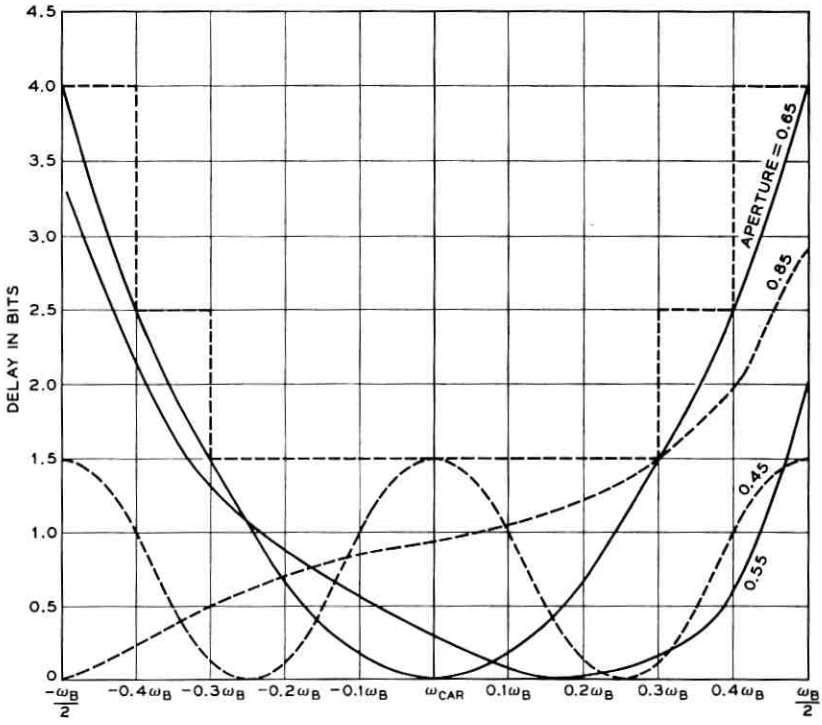


Fig. 21 — Examples of performance range meeting classical delay requirements; frequency in multiples of bit speed from carrier.

the band can (and for certain nonrandom patterns does)* contain a great deal of the necessary information. On the other hand, for timing performed from the data signal, in general for timing from random data, and for recovery of the data itself, the edges of the band contain only a relatively small amount of the information. Thus it would not appear necessary to specify the delay as accurately across the entire transmission band of interest.

A number of delay curves were used to check this supposition. Fig. 22 shows some of the cases which gave the widest variation in performance. These lines have respectively quadratic (curves a, b, c) and fourth-power (curves d, e, f) delay as a function of frequency for the frequency range $\omega_{car} - 0.35\ \omega_{bit}$ to $\omega_{car} + 0.35\ \omega_{bit}$. However, the eye apertures

* Timing for random patterns even when using the line signal gets most of its energy from the center of the band where the spectrum has much greater amplitude. This fact was pointed out to the author by M. A. Logan.[8]

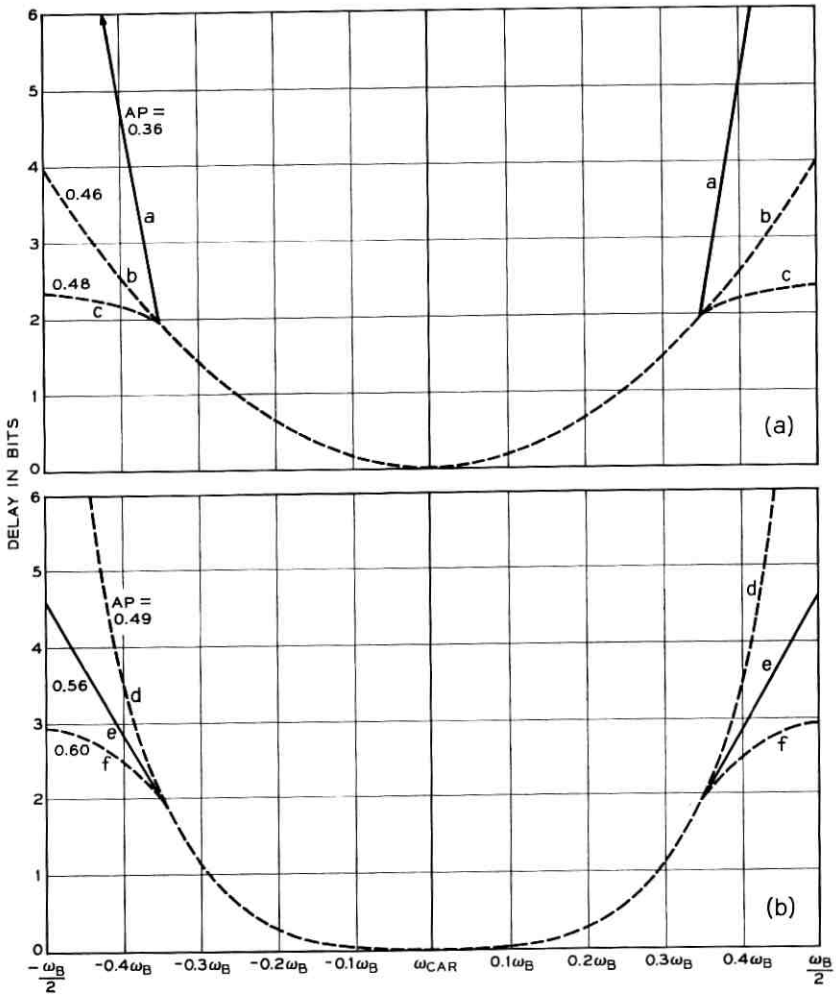


Fig. 22 — Bandwidth requirements, all curves symmetric about ω_{car} : (a) quadratic delay, (b) fourth-power delay.

for even the extreme cases shown vary by less than 2 db, and in most cases that seem to be of practical interest a variation of less than one db was found. The same results held for the impulse noise performance discussed in Section 5.2.

Similar results were found for delays in which the center band delay was a sinusoidal or parabolically bounded sinusoidal function of frequency. Therefore, one concludes that if delay is specified for the frequency range from $\omega_{car} - 0.35\ \omega_{bit}$ to $\omega_{car} + 0.35\ \omega_{bit}$, then for the

recovery of the data signal it is not worth the cost or effort to try to equalize delay beyond about 70 to 75 per cent of the transmission band.

6.6 *Voice-Band Transmission Design*

Historically, the specification of delay has been given in a staircase arrangement. This is equivalent to setting a sequence of pairs of check frequencies and the limitation on delay between them. Thus, for example, a typical delay specification might be

$$\begin{aligned} &\omega_{car} - 0.3\ \omega_{bit}\ \omega_{car} + 0.3\ \omega_{bit} && 1.5 \text{ dibits delay} \\ &\omega_{car} - 0.4\ \omega_{bit}\ \omega_{car} + 0.4\ \omega_{bit} && 2.5 \text{ dibits delay} \\ &\omega_{car} - 0.5\ \omega_{bit}\ \omega_{car} + 0.5\ \omega_{bit} && 4.0 \text{ dibits delay.} \end{aligned} \tag{7}$$

When delay is specified in this manner, no account is taken of the wide variation in performance of delays of various shapes all meeting the basic requirements. Thus, for example, the delays shown in Fig. 21 all meet the requirements listed above, yet have eye apertures ranging from 0.45 to 0.85. In effect, one is faced with the choice of either placing too strict a requirement for many delay shapes, or not in truth being able to guarantee that delays meeting a particular requirement will have no more than a certain allowable degradation.

These results suggest a somewhat more complicated specification of delay requirements for good transmission design. One way of doing this is to use a single equation and vary the parameters of this equation to allow for various delay shapes — for example, to bound the delay by $\alpha_{ps}\ \omega^2 + K$. By choosing various ratios of α_{ps} to K one can allow for a wide variety of delay shapes.

In the voice band there are three major shapes of delay which must be considered: namely, the usual single-minimum parabolic-type delay shape which arises from carrier transmission, the slope-type delay which arises from loaded cable, and the more rectangular delay with a ripple across the transmission band which results from equalizing. The three curves of Fig. 21 are examples of these types of curves. To represent these classes, consider three values of the ratios of α_{ps} to K. For the equalizer ripple-type line a ratio of α_{ps} to K of 1 to 1 was chosen. For a carrier transmission delay a ratio of α_{ps} to K of 4 to 1 was taken. For the loaded cable-type delays a ratio of 8 to 1 was chosen. For each of these delay shapes a set of design curves for various allowed degradations was derived. In each case the maximum delay was found such that a specified degradation in performance would not be exceeded. As usual, the allowable degradation is in aperture or equivalently in impulse noise performance. These curves are shown in Figs. 23, 24 and 25.

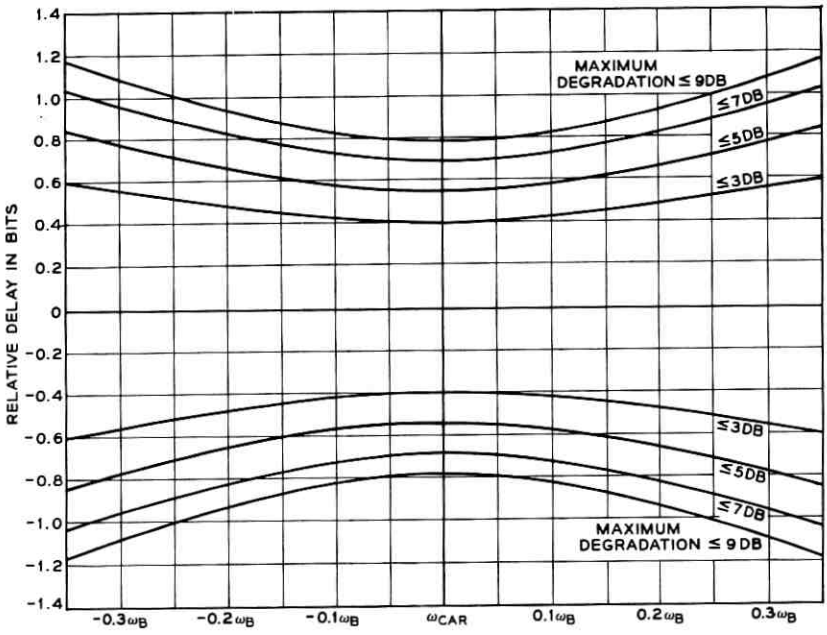


Fig. 23 — Design curve no. I. Maximum allowable delay for $\alpha_{ps}/K = 1$ to give indicated degradation:

$$D = [\alpha_{ps}/(\omega - \omega_{car})^2 + K] \cos (nw + \theta).$$

Curves give maximum degradation; in general, degradation should be less for delays falling within bounds shown. Note that flat delay introduces no distortion.

It is not necessary for the delay of a particular line to meet all the requirements of all three sets of curves in order to have a particular allowable degradation. On the contrary, if a given delay falls within any one curve for a particular allowable degradation in performance then it will meet this degradation requirement. Thus, in the field or in design, it is necessary to consider three curves of allowable delay to determine if a particular delay shape will meet a particular degradation requirement.

In the curves, delay has been specified only up to 70 per cent of the band. This is in keeping with the previous results on the non-necessity of specifying delay beyond this point. However, it is true that there is some range in the actual performance due to the effect of delay beyond the 70 per cent limit. In addition, as discussed in the section on worst lines, these delays are not the actual worst lines as found by Lucky. To allow for these two factors, an additional 1-db margin is built into each of the curves. It is felt that only in the very rarest of circumstances

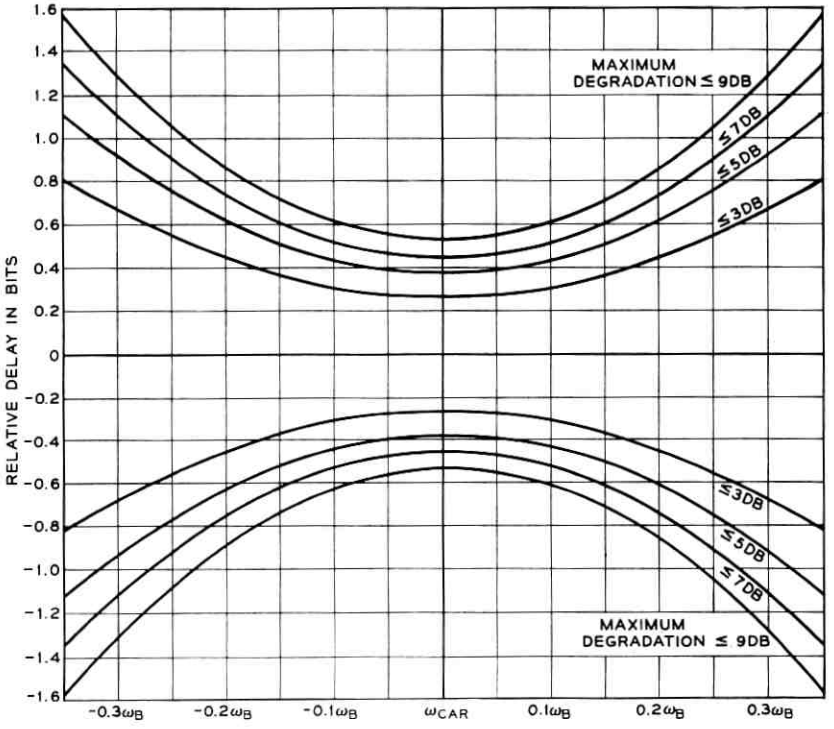


Fig. 24 — Design curve no. II. Maximum allowable delay for $\alpha_{ps}/K = 4$ to give indicated degradation:

$$D = [\alpha_{ps}/(\omega - \omega_{car})^2 + \mathrm{K}] \cos (nw + \theta).$$

Curves give maximum degradation; in general, degradation should be less for delays falling within bounds shown. Note that flat delay introduces no distortion.

will this margin be insufficient, and then only by a very small additional amount of degradation.

VII. ATTENUATION RESULTS

The scope of this study did not include a systematic investigation of the distortion due to attenuation alone or to a combination of attenuation and delay. However, certain representative results are discussed here to indicate some of the effects of attenuation. The simulation can, of course, handle any desired attenuation.

The spectrum has symmetric components with respect to the carrier. Thus for symmetric delay (in particular for no delay distortion), slope attenuation across the transmission band [i.e., attenuation = k_1 (frequency) + k_2] should produce an effect equivalent to a flat loss of value

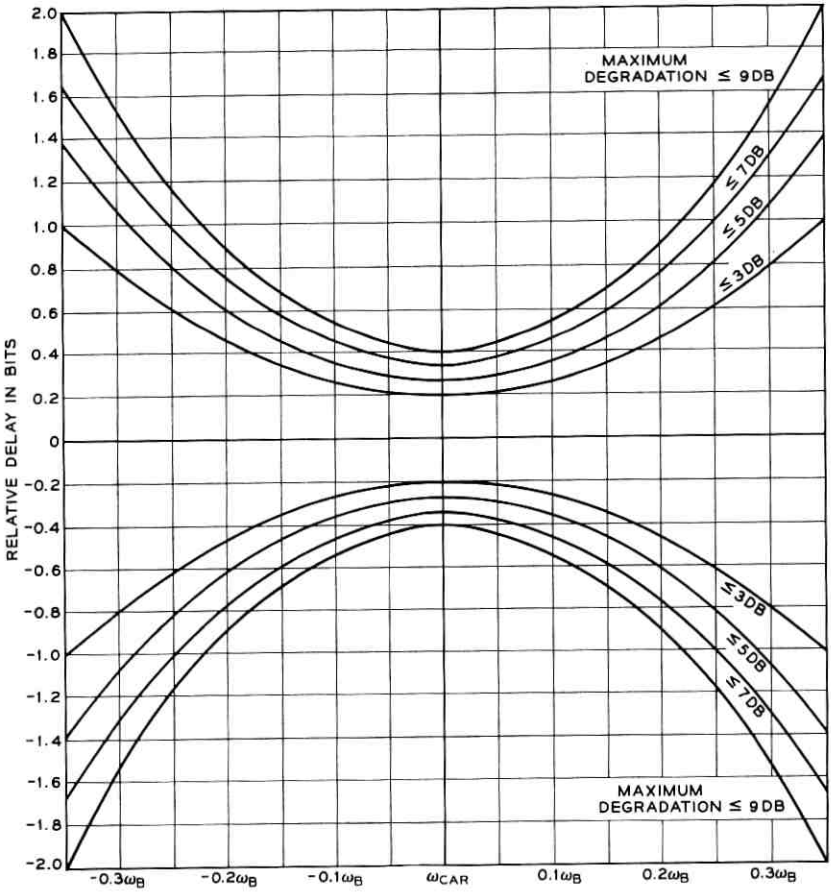


Fig. 25 — Design curve no. III. Maximum allowable delay for $\alpha_{ps}/K = 8$ to give indicated degradation:

$$D = [\alpha_{ps}/(\omega - \omega_{car})^2 + K] \cos (nw + \theta).$$

Curves give maximum degradation; in general, degradation should be less for delays falling within bounds shown. Note that flat delay introduces no distortion.

equal to the attenuation at the carrier frequency. This line of thought was essentially verified using the simulation. In the presence of delay unsymmetrical with respect to the carrier frequency, however, this is no longer true. This also indicates the nonadditive nature of the effects of attenuation and delay. In any study of the combined effects of attenuation and delay distortion, it is this nonadditive nature of the interaction which is the most important single fact.

The general form of attenuation considered was linear on a plot of db versus frequency as shown in Fig. 9(a). This attenuation shape was found to be typical of voice-bands by Alexander, Gryb, and Nast.[9] The cutoff frequency was taken as the carrier minus 20 per cent of the bit speed. For example, with an 1800-cps carrier operating at 2400 bps, this would be at 1800 − (0.2) (2400) = 1320 cps. The slope was measured from this cutoff frequency ($\omega_{car} - 0.2\ \omega_{bit}$) to $\omega_{car} + 0.35\ \omega_{bit}$. Slopes ranging from 4 to 12 db over this range were investigated.

Table II shows clearly the nonadditive interaction of the attenuation plus delay. Both delays shown in Table II are sinusoidal and respectively of odd and even symmetry with respect to the carrier frequency. The results are normalized to zero attenuation at the carrier.

TABLE II — INTERACTION OF ATTENUATION AND DELAY

Amplitude	Delay		Attenuation	Aperture Delay Only	Aperture Delay + Attenuation
Bm	θ	*m*	6-db slope	0.45	0.36
1.0	0	2.5			
Bm	θ	*m*	6-db slope	0.70	0.32
0.5	$\pi/2$	2.0			

VIII. ACKNOWLEDGMENTS

As in almost every paper it is really impossible to give credit to all those who have contributed to the author's understanding of the problem. However, special thanks are due to R. A. Gibby for guidance and motivation and for introducing the author to the simulation approach to problems of this type. In addition, R. R. Anderson, R. W. Lucky and S. Habib in analysis and simulation, M. A. Logan and P. A. Baker in the area of the physical system, and H. C. Fleming, D. W. Nast, and W. R. Young in the area of transmission design were particularly helpful in the development of this paper.

REFERENCES

1. Baker, P. A., PM Data Sets for Serial Transmission at 2000 and 2400 Bits per Second, Conference Paper CP-62-143, A.I.E.E., Fall, 1961.
2. Rappeport, M. A., Criterion Problem in Data Transmission, Conference Paper 63-534, A.I.E.E., Winter, 1963.
3. Gibby, R. A., An Evaluation of AM Data System Performance by Computer Simulation, B.S.T.J., **39**, May, 1960, p. 675.
4. Peterson, W. W., *Error Correcting Codes*, John Wiley & Sons, New York, 1961.
5. Fennick, J. H., A Report on Some Characteristics of Impulse Noise in Tele-

phone Communication Systems, Conference Paper 63-986, IEEE, June, 1963.
6. Wheeler, H. A., The Interpretation of Amplitude and Phase Distortion in Terms of Echoes, Proc. I.R.E., **27,** June, 1939, p. 359.
7. Lucky, R. W., A Functional Analysis Relating Delay Variation and Intersymbol Interference in Data Transmission, B.S.T.J., **42,** Sept., 1963, p. 2427.
8. Logan, M. A., private communication.
9. Alexander, A. A., Nast, D. W., and Gryb, R. M., Capabilities of the Telephone Network for Data Transmission, B.S.T.J., **39,** May, 1960, p. 431.

NEASIM: A General-Purpose Computer Simulation Program for Load-Loss Analysis of Multistage Central Office Switching Networks

By R. F. GRANTGES and N. R. SINOWITZ

(Manuscript received November 7, 1963)

Blocking probability is the most frequently required performance characteristic in traffic studies of complex central office switching networks. Determining this quantity without actual measurement is a difficult task. To aid the communications system designer, a simulation program has been prepared which produces useful estimates of blocking probability for a large class of networks. The program is based on a simplified mathematical model for the analysis of switching networks developed by C. Y. Lee, and thus differs from conventional simulators in that it simulates a mathematical model rather than a traffic-handling system.

Although Lee's model is widely used, its utility has been limited by computational difficulties encountered in networks of realistic size and complexity. This limitation is in most practical cases removed by the program, which features rapid input preparation, short computer runs, and specification of the desired precision of the results as an input parameter. Moreover, the program allows for the incorporation of more a priori *information about the actual behaviour of switching networks than is included in Lee's model, thereby leading to a more accurate estimate of blocking probability.*

The simulator has been programmed for the IBM 7090 computer, but the concepts are machine independent.

CONTENTS

I. INTRODUCTION

Determining the traffic performance of complex multistage central office switching systems without actual measurement can be a major problem for the communications engineer. While probability theory has been successfully applied to a wide variety of telephone traffic problems,[1,2,3] a precise formulation of a mathematical model completely describing the multistage switching system has thus far not been found.

No systematic approach exists which completely accounts for the gross complexity encountered in large-scale congestion systems, but several authors have contributed significantly to the theory, notably C. Jacobaeus,[4,6] K. Lundkvist,[5] A. Jensen,[7] C. Y. Lee,[8] A. Elldin,[9] R. Fortet,[10] and P. LeGall.[11] More recently, V. E. Beneš[12-15] has initiated "an attempt to describe a comprehensive point of view towards the subject of connecting systems."[12] Although the engineer does not yet have a comprehensive theory, he does have a valuable tool in computer simulation.

Simulation of telephone traffic flow has a long history in the Bell System. As early as 1907, a rudimentary simulation was undertaken to improve switchboard performance. Artificial traffic was generated by a card-drawing technique, and the simulation was used to verify a semi-mathematical analysis of the loads which could be handled by a team of operators meeting an average delay criterion. In the ensuing years, simulation techniques have been aids in the study of complex traffic problems, such as the effect of limited sources on graded multiple capacities, the efficiency of random slipped multiples, the capacities of various alternate routing plans, and the distribution of delays under various trunking plans. The traffic load capacity of the No. 1 crossbar network was largely determined by the load-loss relationships in the link and junctor patterns obtained from elaborate simulations begun in 1936. This was the first time that the capacity of a largely complete system had come under study by simulation methods. A 10,000-line No. 5 crossbar office was simulated in 1948 by a specially designed machine

which coordinated the efforts of four operators, providing significant data for the traffic engineering of this system.[16]

In recent years the high-speed electronic digital computer has proved an effective tool in large-scale traffic simulations.[17-22] For this class of simulations, special computer programs are written which usually contain:

(*i*) a logical description of the system under consideration,

(*ii*) a procedure for generating and offering traffic to the system, and

(*iii*) a method for extracting and recording the desired system characteristics.

These programs may be called "special-purpose" simulators — in the sense that they are written for the purpose of studying a specific traffic-handling system. A variety of performance data may be obtained, including:

(*i*) probability of blocking at various loads (load-loss data);

(*ii*) delay distribution, including average delay on calls delayed; and

(*iii*) mean queue lengths.

Of these, the most frequently required datum is the probability of loss (or delay).

These simulation programs have produced a large amount of useful information, but their application has not been widespread because of the considerable programming effort required. To reduce programming effort, various "general-purpose" traffic simulation programs have been written.[24,25,26] However, it must be understood that each program is only "general" with respect to a particular class of traffic systems.

The multistage central office switching system is an example of a class of traffic-handling systems for which no general-purpose simulator has heretofore been written, although much has been accomplished by special-purpose simulations written for specific switching network arrangements.[27] Because the use of these network simulation programs has been greatly restricted by the cumbersome programming and input preparation required, a strong need has developed for a quick, easy-to-use, general-purpose simulation technique.

To meet this need, the authors have developed a computer program which, with a minimum of user effort, will produce useful estimates of blocking probability for a very large class of multistage switching networks. The program is based on a simplified mathematical model of switching networks developed by C. Y. Lee,[8] and differs in approach from programs referred to earlier in that a complete description of the *traffic-handling system itself* is not given to the computer; rather, the program simulates the behavior of Lee's *analytical model*. Deriving its

name from this view of its operation, the program has been acronymously named NEASIM — NEtwork Analytical SIMulator.

While Lee's model is probably the most widely used analytical model for multistage matching networks, its utility has been severely limited by the computational difficulties associated with networks of realistic size and complexity. This limitation is in most practical cases removed by the program, which features rapid input preparation and short (hence economical) computer runs. Furthermore, unlike many simulations in which the reliability of results must be assessed on an *a posteriori* basis, the analytical simulator admits of an *a priori* appraisal so that desired precision becomes an input parameter.

The probability linear-graph model, basic to the NEASIM approach, is described in Section II, where its use in switching network analysis is explained. The philosophy of the program together with a general description is presented in Section III. Two key program routines are described in Section IV. Reliability considerations are given in Section V. Finally, Section VI discusses program modifications which can increase the validity of the NEASIM estimate.

II. THE PROBABILITY LINEAR-GRAPH MODEL

In 1955 C. Y. Lee,[8] extending the earlier work of Kittredge and Molina, presented a simplified mathematical model for the analysis of switching networks. Since the NEASIM program simulates the behavior of this model, a description of the model is given (Section 2.1). An example to illustrate how the model is applied is given in Section 2.2; the computational difficulties which may be encountered in the analysis of practical networks are then discussed — thus pointing up the need for the simulation program. Section 2.3 explores further the notion of blocking probability for a network and the use of the linear-graph model in determining this quantity.

2.1 *The Model*

Consider a crosspoint network in which each input can be connected to any output by the operation of appropriate crosspoints. Let $P_t(j,k)$ be the probability that all paths through the network between input j and output k are busy at time t.

Associate with each link l_i of the network a binary-valued random variable $X_t^{(i)}$ whose value represents the state of the link at time t. The NEASIM convention is: 0 represents idle and 1 represents busy.

Since the concern of telephone traffic engineers is with "busy-hour" traffic, it is usually assumed that

(a) *the busy-idle distributions of the link random variables are stationary (or homogeneous) in time.*

That is, for any N,

$$\Pr\{X_{t_n}^{(i)} = \delta_n \; ; \; n = 1, \cdots, N\} = \Pr\{X_{t_n+h}^{(i)} = \delta_n \; ; \; n = 1, \cdots, N\}$$

for all h, where $\delta_n = 0$ or 1 and the index i runs over all the links in the network. A consequence of this assumption is that $P_t(j,k)$ is also stationary in time, for all j and k, and the subscript t will henceforth be omitted.

Let us now fix our attention on two terminals, one on each side of a crosspoint network in which

(b) *all of the switches are nonblocking,**

and in which

(c) *there is no connection path between any switch and itself.*

Then the configuration of possible paths through the network between the two terminals can be represented by a two-terminal cycle-free linear graph with directed branches, in which the nodes of the graph represent network switches and the directed branches of the graph represent network links. Consider, for example, the network depicted in Fig. 1. The possible path configuration seen between terminals A and B is indicated by heavy lines, and the corresponding graph is as shown in Fig. 2(a).

Next, assume that

(d) $P(j,k)$ *is independent of* j *and* k,

and we can speak of $P(j,k) = B$ as the probability of blocking of the network. The notion of blocking probability will be further explored in Section 2.3.

Finally, Lee makes the simplifying assumption:

(e) *the link random variables,* $X^{(i)}$, *are independent.*

This assumption, which is frequently made to render analysis manageable, is the principal weakness of the model and will cause the results to depart from reality in varying degrees — depending on the particular network. In general the model will tend to overestimate blocking. The problem of obtaining realistic results is discussed further in Section VI.

Each of a large class of switching networks can therefore be repre-

---

* Lee's requirement that the switches be nonblocking is actually not restrictive, and can be relaxed by an appropriate adjustment of the link occupancies.

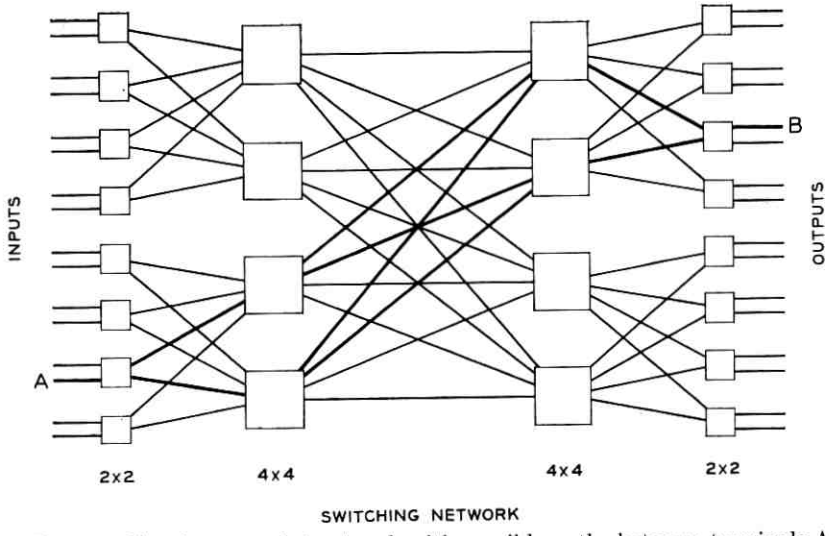


Fig. 1 — Simple crosspoint network with possible paths between terminals A and B indicated by heavy lines.

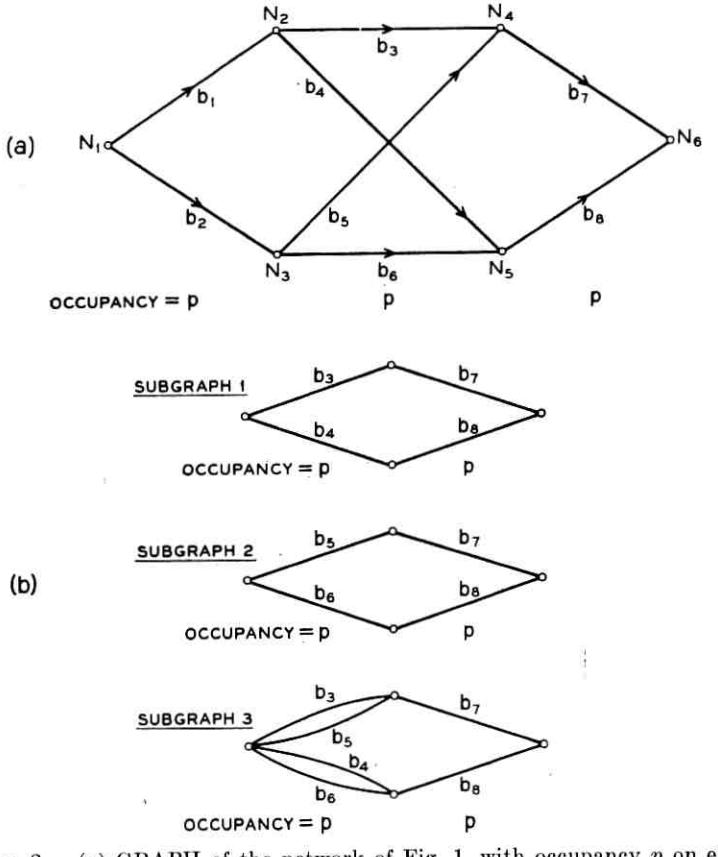


Fig. 2 — (a) GRAPH of the network of Fig. 1, with occupancy p on each of the branches. (b) The subgraphs of the GRAPH.

sented by a simplified model called, by Lee, a *two-terminal probability linear-graph* in which

(*i*) switches are represented by nodes, and links by directed branches, and

(*ii*) assumptions (a)–(e) hold.

The mathematical object, the two-terminal probability linear-graph, will be referred to as GRAPH in the sequel.

Once the GRAPH of a network is obtained, the calculation of the blocking probability can proceed in a straightforward manner.

2.2 *An Example*

As a first illustration, consider the GRAPH of Fig. 2(a). Let E be the event that there is a path through the GRAPH,* E_i be the event that branch b_i is idle, and $p_i = \Pr\{b_i \text{ is busy}\}$. Then

$$E = A_1 \cup A_2 \cup A_3 \cup A_4$$

where the paths, A_i, are

$$\begin{aligned} A_1 &= E_1 \cap E_3 \cap E_7 \\ A_2 &= E_1 \cap E_4 \cap E_8 \\ A_3 &= E_2 \cap E_5 \cap E_7 \\ A_4 &= E_2 \cap E_6 \cap E_8 . \end{aligned}$$

The blocking probability is then

$$\begin{aligned} B &= 1 - \Pr\{E\} \\ &= 1 - \Pr\{A_1 \cup A_2 \cup A_3 \cup A_4\} \\ &= 1 - \sum_{i=1}^{4} \Pr\{A_i\} + \sum_{\substack{i,j=1 \\ i<j}}^{4} \Pr\{A_i \cap A_j\} \\ &\quad - \sum_{\substack{i,j,k=1 \\ i<j<k}}^{4} \Pr\{A_i \cap A_j \cap A_k\} + \Pr\{A_1 \cap A_2 \cap A_3 \cap A_4\}. \end{aligned}$$

Now the assumption of independence gives

$$\begin{aligned} \Pr\{E_i \cdots E_j\} &= \Pr\{E_i\} \cdots \Pr\{E_j\} \\ &= q_i \cdots q_j \qquad q_i = 1 - p_i \end{aligned}$$

* Having adopted the GRAPH model, we speak of a "path through the GRAPH" rather than "a path through the network," and say that "a branch is busy or idle" rather than "a link is busy or idle."

whence, for the case in which $p_i = p$ for all i,

$$B = q^8 - 4q^7 + 2q^6 + 4q^5 - 4q^3 + 1. \tag{1}$$

In general, given a GRAPH G with m different link occupancies $p_1, \cdots, p_m$, the procedure just illustrated will yield the *blocking polynomial* of the GRAPH:

$$B_G = B_G(p_1, \cdots, p_m).$$

As a computational tool the utility of the GRAPH model decreases with increasing complexity of the GRAPH's geometrical structure. When the GRAPH geometry grows more complex, the blocking polynomial B_G becomes cumbersome — admitting a greater possibility for error in its determination. Moreover, once B_G is found, one still has to substitute numerical values for $p_1, \cdots, p_m$ into the polynomial to obtain a result. As an example, for the GRAPH of moderate complexity shown in Fig. 3, the blocking polynomial is

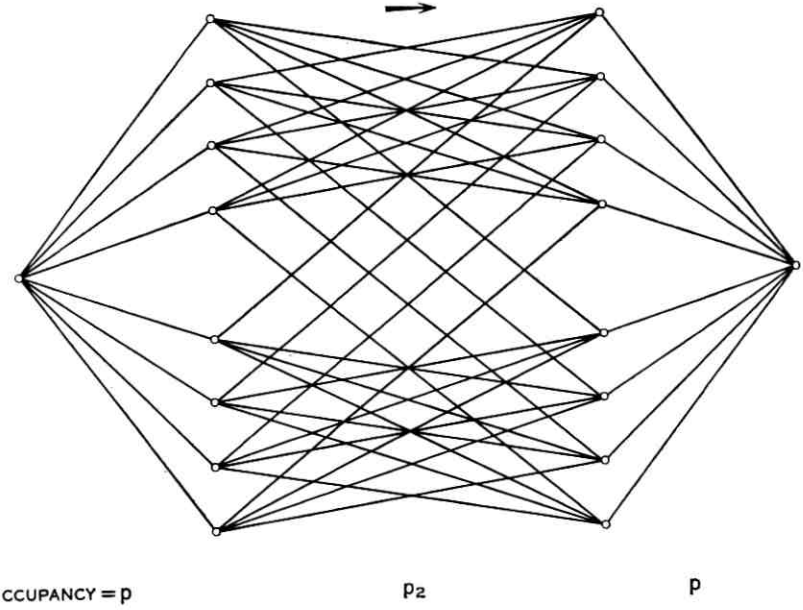


Fig. 3 — A GRAPH of moderate complexity.

$$
\begin{aligned}
B = {} & p^8 \\
& + p^7 q(8A^4) \\
& + p^6 q^2(4A^8 + 24A^4B^2) \\
& + p^5 q^3(24A^4B^4 + 32A^3B^3C) \\
& + p^4 q^4(8A^4C^4 + 48A^2B^4C^2 + 8B^6D + 6B^8) \qquad (2) \\
& + p^3 q^5(32AB^3C^3D + 24B^4C^4) \\
& + p^2 q^6(4C^8 + 24B^2C^4D^2) \\
& + pq^7(8C^4D^4) \\
& + q^8(D^8)
\end{aligned}
$$

where:

$$
\begin{aligned}
A &= p + qp_2 \\
B &= p + qp_2^2 \\
C &= p + qp_2^3 \\
D &= p + qp_2^4 \\
q &= 1 - p.
\end{aligned}
$$

Faced with computing B in (2) over a range of occupancies, an engineer would surely resort to a computer. The essential computational difficulty is that one is confronted with expressions of the form

$$\Pr\{A_i \cup \cdots \cup A_j\}$$

where the paths $A_i, \cdots, A_j$ are not disjoint.

It is interesting to note that the method of approach just described — which may be called the "path enumeration approach" — is not the only way to proceed and is indeed not the most efficient. A second procedure for finding the blocking polynomial may be called the "combinatorial approach" and is best illustrated by example.

Since all the branches in the GRAPH of Fig. 2(a) are busy with probability p, a moment's reflection shows that the blocking probability can be written as

$$B = p^2 + qp \cdot B_{\text{subgraph 1}} + pq \cdot B_{\text{subgraph 2}} + q^2 \cdot B_{\text{subgraph 3}}$$

where $B_{\text{subgraph 1}}$, $B_{\text{subgraph 2}}$, $B_{\text{subgraph 3}}$ are, respectively, the blocking probabilities of the GRAPHs indicated in Fig. 2(b). But by inspection we have

$$B_{\text{subgraph 1}} = (1 - q^2)^2$$
$$B_{\text{subgraph 2}} = (1 - q^2)^2$$
$$B_{\text{subgraph 3}} = [1 - (1 - p^2)q]^2,$$

so that

$$B = p^2 + 2pq(1 - q^2)^2 + q^2[1 - (1 - p^2)q]^2$$

is the blocking polynomial (1) expressed in another form.

The GRAPH of Fig. 4 is more representative of the type of GRAPH

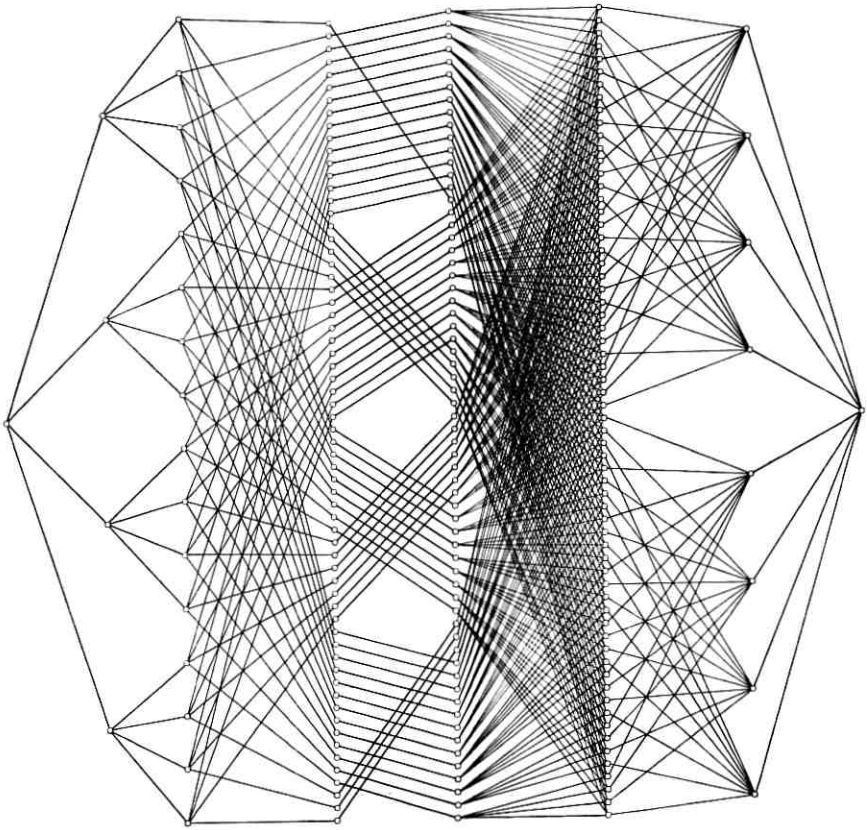


Fig. 4 — Typical GRAPH geometry of realistic central office network.

encountered in modern central office networks. Determination of the blocking polynomial in this case — by either the path enumeration or combinatorial approaches — is a formidable task indeed. If the GRAPH model is to be useful to the modern engineer, some means of handling such complex GRAPHs must be made available.

When seeking performance measures of a network, the engineer is not really concerned with the explicit polynomial representation $B_G(p_1, \cdots, p_m)$; what he would like is a curve (or set of curves) displaying this functional relationship. The simulation program described in the following sections, when given the link occupancies $p_1, \cdots, p_m$, produces, with predictable precision, a numerical estimate of B_G.

2.3 *Blocking Probability and the Linear-Graph Model*

When several kinds of traffic are handled with different disciplines in a network which may have several characteristic graphs, blocking probability for the network can only be meaningfully defined relative to the persons or terminals encountering blocking. For example, the blocking encountered by a call originating and terminating in the same central office may not be the same as the blocking encountered by an incoming call. In general, a network is required to handle several "classes" of connections and the engineer is concerned with the blocking probability for each of these classes. We shall limit our discussion of blocking probability to one class, that is, a subset of all input-output pairs in which each input-output pair has the same graph and for which it is reasonable to suppose that the traffic between every input-output pair is identical* with that of every other pair. Without loss of generality we can therefore assume that the network has one class, so that when speaking of the blocking probability of the network, we shall mean the blocking probability of the class. (These remarks form the basis of assumption (d) of Section 2.1.)

Having thus limited ourselves to one class of connections, we have still to define the blocking probability of a network in a manner which is in agreement with the generally familiar definitions. Here we again encounter difficulty. The authors agree wholeheartedly with Beneš (Ref. 15, p. 2805), that "In fact, not even the definition (let alone the calculation) of the probability of blocking has received adequate treatment . . . ."

Syski (Ref. 3, p. 198), after a long series of prefatory remarks, defines two quantities, time congestion $S(t)$ and call congestion $\pi(t)$ for the case

* That is, every input calls every output at the same rate with the same holding time distribution and with the same lost calls disposition.

of a simple full-access trunk group. The time congestion is the probability that all trunks in the group are busy at time t; the call congestion is the conditional probability that the group is blocked when a call arrives at the instant t. Under the input assumptions and for equilibrium conditions, time and call congestions are independent of time and are denoted by S and π, respectively. S is equivalent to the fraction of time during which congestion is encountered, while π is equivalent to the fraction of calls encountering congestion, and the two quantities will, in general, differ.

Of these, of course, the measure of most concern to the network designer is call congestion. Now if it is assumed that calls originate completely independent of the state of the network, time congestion will equal call congestion. Such an assumption is unjustified if calls cannot originate from busy lines, since time congestion conventionally includes busy line periods while call congestion excludes them. It is, however, reasonable to assume that idle pairs of terminals originate calls at a constant rate independent of the state of the network. In particular, there must be no change in the calling rate after a blocked call. If, under this assumption, time congestion is modified to include only periods in which both lines are idle, it will be equal to call congestion. Actually, even if the foregoing assumption is not met, the time between calls is likely to be much longer than the time taken by the network to return to equilibrium, so that, again, the modified time congestion will be close to the call congestion.

With a suitable choice of branch occupancies, Lee's model allows the computation of call congestion. Alternatively, the branch occupancies may be chosen so as not to reflect the requirement that only idle terminals are to be considered, thus allowing the computation of time congestion. In either case the computed results will be subject to the error introduced by inaccurate assumptions in the model.

Before viewing the probability linear-graph model in the light of the above remarks, it is well to make an observation on the underlying philosophy of the model. Let us perform the following conceptual experiment in a real network under a particular set of (equilibrium) traffic conditions. Suppose that we fix our attention on a particular representative input-output pair (j,k) and examine closely that portion of the network seen between them, i.e., their graph. It is reasonable to believe that, were we to examine the detailed traffic pattern within the graph for a sufficiently long time, we would ultimately come to have complete knowledge of the busy-idle state behavior of the graph links under the particular traffic conditions. The experiment could be repeated under other equilibrium traffic conditions, so that we would eventually be able

to describe completely the behavior of the graph links under all equilibrium conditions. Assuming that the particular input-output pair and connection graph studied are truly representative of the entire network (or at least of an entire connection class), we could then "discard" the rest of the network and determine the blocking probability of the network under various equilibrium conditions by computations or simulations based only on the connection graph and our complete knowledge of its behavior.

That such complete knowledge could be obtained is a practical impossibility. In the absence of complete knowledge, assumptions can be and, indeed, must be made about the detailed behavior of the graph based on such *a priori* knowledge as we have. It is reasonable to suppose that, as the assumptions made approach the real behavior of the graph, the blocking probability determined from the graph will approach the real blocking of the network. A belief in the fundamental soundness of this reasoning constitutes the basic philosophy of the graph model approach to the determination of blocking probability, beginning with Molina and continuing with C. Y. Lee and the NEASIM program.

The assumption made by Lee of link independence [(e) of Section 2.1], while obviously omitting much *a priori* knowledge of graph behavior, possesses the practical advantage of allowing computation of the blocking polynomial where graph geometry does not prohibit. The basic NEASIM program allows evaluation of the polynomial for all graphs meeting Lee's restrictions.

The utility of the results obtainable from Lee's model is well known. When the specific branch occupancies are chosen rationally,* the calculated blocking agrees well enough with real blocking figures (obtained from full-scale simulation or measurement) for many engineering and design purposes. Accuracy can be improved by "calibrating" Lee's results against real values where they are available. Furthermore, computed values lacking in absolute accuracy will reveal relative differences between networks and between various traffic conditions in the same network.

The NEASIM program, to which the remainder of this paper is devoted, is basically designed to estimate the value of the blocking polynomial. This portion of its design and use is described in Sections III–V. The design also allows additional assumptions regarding the detailed traffic behavior of connection graph link states to be incorporated in the graph model. When more *a priori* information is included,

* That is, chosen to reflect the requirement that the input-output terminals j,k are idle by (usually) subtracting the load contributed by the terminals j,k from the assumed carried link loads.

the validity of the simulation results improves as anticipated. This aspect of the NEASIM program is described in Section VI.

III. THE NEASIM PROGRAM

This part of the paper is devoted to a presentation of the program. The basic viewpoint or philosophy of the NEASIM approach is given in Section 3.1. A general description of the program is contained in Section 3.2.

3.1 *Philosophy of the Program*

We saw in the preceding section how the GRAPH model provides — in theory at least — a method for the analysis of a large class of switching networks and how the model is impractical as a tool for networks with complex geometrical structure. To evolve a practical tool, we shall change our method of approach.

Suppose we are given a GRAPH G with branch occupancies

$$p_1 , \cdots , p_m .$$

Until now we were concerned with the explicit blocking polynomial $B_G = B_G(p_1 , \cdots , p_m)$. However, we can think of $B_G(p_1 , \cdots , p_m)$ as a curve in Euclidean $(m + 1)$-space, and set as our goal a precise approximation to this curve. This point of view immediately suggests the use of simulation techniques.

For example, if all the branches in the GRAPH of Fig. 2(a) are busy with probability p, then $B_G = B_G(p)$ is a curve in 2-space. If a computer simulation were to be performed to give an approximation to $B_G(p)$, then we would require a computer program containing an algorithm which, when repeated n times, will produce an estimate $B_G^{(n)}(p)$ such that

$$\lim_{n \to \infty} B_G^{(n)}(p) = B_G(p)$$

and for which we have confidence limits on the absolute error,

$$| B_G^{(n)}(p) - B_G(p) | ,$$

for all n.

Consider the eight branches of the GRAPH of Fig. 2(a). Although all of them are busy with probability p, at any one instant each of the branches is either busy or idle — and for this particular configuration of busy and idle branches there is or is not a path through the graph. Thus, in each repetition of the above mentioned algorithm,

(*i*) we assign busy-idle states to each of the eight branches in such a way that probability of busy in each case is p;

(*ii*) after the assignment has been made we determine whether or not there is a path through the graph for the given assignment.

Let $B_G^{(n)}(p)$ represent the proportion of n repetitions of the algorithm when no path through the graph was found. We would then expect that $\lim_{n\to\infty} B_G^{(n)}(p) = B_G(p)$; that is, we would expect our estimator to converge to the true blocking curve. In general, if the GRAPH G has m different occupancies $p_1, \cdots, p_m$, the program should produce an estimator $B_G^{(n)}(p_1, \cdots, p_m)$ converging to $B_G(p_1, \cdots, p_m)$.

The preceding heuristic remarks were intended to outline the essential approach taken by the NEASIM program. The remainder of Section III is devoted to a description of the program itself. Sections IV and V contain the arguments which show that the estimator indeed converges to the true blocking curve, and how confidence statements about the precision of the results are obtained.

3.2 *Program Description*

The following four sections describe the NEASIM program. Section 3.2.1 takes a "macroscopic" point of view, beginning with an account of the input and then proceeding to give a broad outline of the program. Section 3.2.2 sketches the layout of data in the computer memory. Section 3.2.3 discusses the organization and operation of the NEASIM algorithm. The salient features of the program flow are shown in Fig. 5. Storage requirements and execution speed are given in Section 3.2.4.

3.2.1 *Macroscopic Description*

NEASIM was written for the IBM 7090 computer. The input consists of punched cards which we categorize as Graph Definition Cards and Simulation Definition Cards. Since the notion of a probability-linear graph implies a geometrical configuration together with an occupancy assignment on the branches, the computer must be supplied with both these types of information. The geometrical configuration is read into the computer via the Graph Definition Cards and the various occupancies are read in via the Simulation Definition Cards.

Graph Definition Cards

The information punched on these cards includes

(1) the total number of nodes,

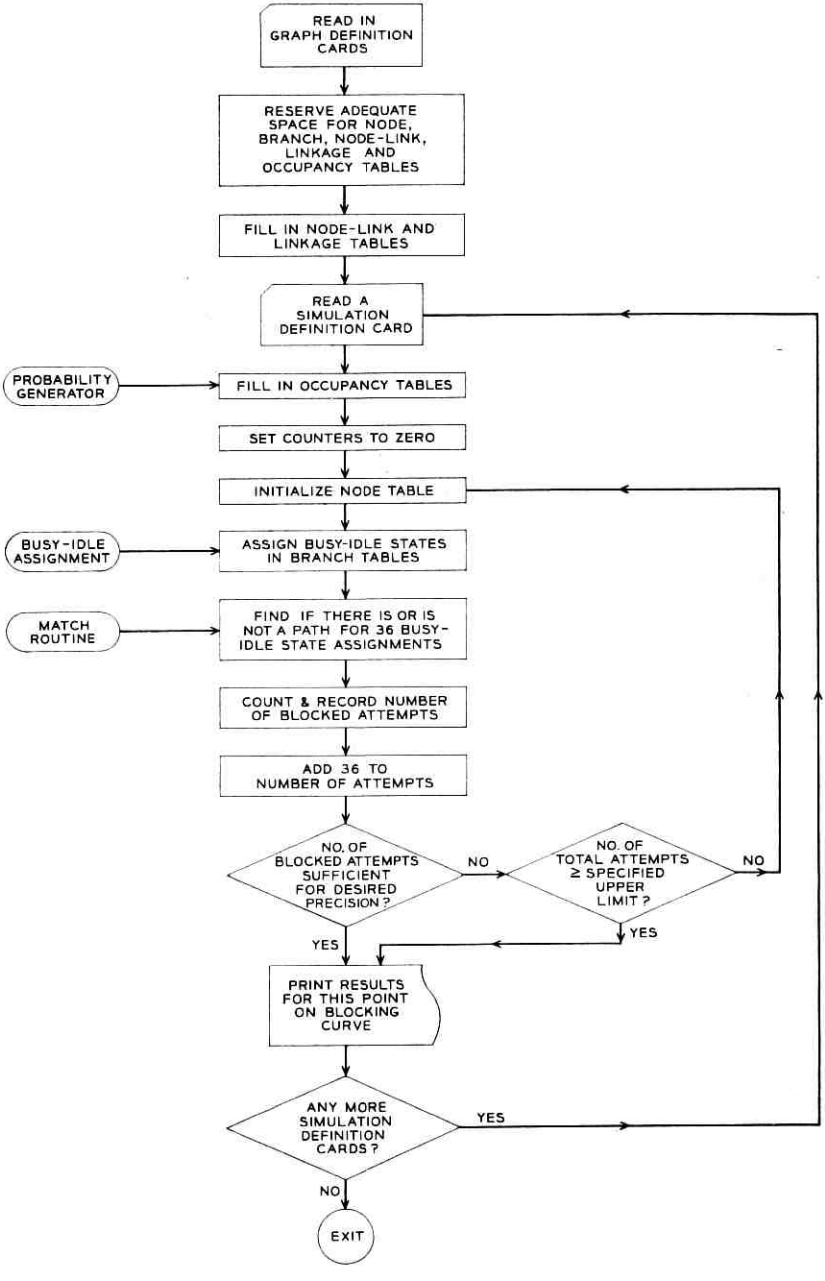


Fig. 5 — Flow diagram for the NEASIM program.

(2) the total number of branches,
(3) the total number of branch occupancies,
(4) the occupancy associated with each of the branches, and
(5) the interconnection scheme between the various nodes.

Simulation Definition Cards

Suppose there are m different branch occupancies for the GRAPH in question. These m values are punched on a Simulation Definition Card. For every set of values of the occupancies $p_1, \cdots, p_m$ — that is, for every point on the blocking curve $B_G(p_1, \cdots, p_m)$ — there is one Simulation Definition Card.

The estimator $B_G^{(n)}(p_1, \cdots, p_m)$ will converge to $B_G(p_1, \cdots, p_m)$ as $n \to \infty$. But the computer must be instructed as to when to terminate the run. Now, the NEASIM process is such that it is possible to request that the error $| B_G^{(n)} - B_G |$ lie within a given percentage of the true value B_G. This "desired precision" information is also punched on the Simulation Definition Card. It may happen, however (as will be the case whenever B_G is very small) that, in order to obtain the requested percentage error, the total number of repetitions of the NEASIM algorithm will perforce be exceedingly large. Since a lengthy computer run is economically undesirable, and since a high degree of precision is usually not needed when the blocking probability is so very low, an upper limit on the number of repetitions, n, is also supplied to the computer by being punched on the Simulation Definition Card. If, after a run has been made, a greater degree of precision is still needed, it is possible to "pick up where we left off" and continue the simulation in another computer run.

The Program

The Graph Definition Cards are read into the computer first. With this information the program constructs, in effect, a map in the computer memory of the geometrical configuration of the GRAPH. Moreover, the program associates with each branch an occupancy p_i — *whose value is as yet unspecified.* The first Simulation Definition Card is then read in, and the program *now* assigns the appropriate values to $p_1, \cdots, p_m$. Once this information is obtained, the program is ready to execute the NEASIM algorithm, which consists essentially of two parts:

(i) a busy-idle assignment is made on all of the branches in accordance with the specified occupancies $p_1, \cdots, p_m$;

(*ii*) the presence or absence of a path is determined for the particular assignment.

These two steps are repeated again and again until such time as the estimate $B_G^{(n)}(p_1, \cdots, p_m)$ has been found.

The results for this point on the blocking curve are printed out, and the second Simulation Definition Card (if there is one) is read. The program then goes through the preceding steps for this second set of values of $p_1, \cdots, p_m$ until the estimate for this point on the blocking curve has been obtained. When all the Simulation Definition Cards have been processed, the program run ends.

3.2.2 *Memory Organization*

After the Graph Definition Cards have been read in, the program prepares five main tables as follows:

(1) Node table
(2) Node-Link table
(3) Branch tables
(4) Occupancy tables
(5) Linkage table.

The Node Table

The size (i.e., the number of words) of the Node table is equal to the number of nodes in the GRAPH, there being a one-to-one correspondence between the words in this table and the nodes in the GRAPH. These words are used by the program to indicate, after a particular iteration of the NEASIM algorithm, whether or not there is a path from each particular node to the first node.

The Node-Link Table

The size of the Node-Link table is also equal to the total number of nodes in the GRAPH, with each word corresponding to a particular GRAPH node. For each node the table indicates

(*i*) the number of branches leaving the node — connecting to nodes more distant from the origin, and

(*ii*) a reference to a section of the Linkage table where further information on each branch is stored.

The Node-Link and Linkage tables together constitute the program's map of the GRAPH geometry. The other tables provide storage for busy-idle indications and path information.

The Branch Tables

There are as many Branch tables as there are different occupancies in the GRAPH. For occupancies $p_1, \cdots, p_m$ there will be m Branch tables, which we denote by $\mathrm{BRT}_{(1)}, \cdots, \mathrm{BRT}_{(m)}$. The size of $\mathrm{BRT}_{(i)}$ equals the number of branches in the GRAPH which are busy with probability p_i. There is a one-to-one correspondence between a particular word in $\mathrm{BRT}_{(i)}$ and a particular branch in the GRAPH. The association between the words in the Branch tables and the particular GRAPH branches is part of the information stored in the Linkage table. The Branch table words are used by the program to store the busy-idle state of every GRAPH branch on each iteration of the NEASIM algorithm. The total storage required for the Branch tables equals the number of branches in the GRAPH.

The Occupancy Tables

There are m Occupancy tables, $\mathrm{OCT}_{(1)}, \cdots, \mathrm{OCT}_{(m)}$, corresponding respectively to $\mathrm{BRT}_{(1)}, \cdots, \mathrm{BRT}_{(m)}$. The size of $\mathrm{OCT}_{(i)}$ is an input parameter of the program chosen to be large compared with $\mathrm{BRT}_{(i)}$. Every bit in the table $\mathrm{OCT}_{(i)}$ contains a binary one with probability p_i. The Occupancy tables are used by the program to supply random busy-idle states for assignment to the branches of the GRAPH on each iteration of the NEASIM algorithm.

The Linkage Table

As previously mentioned, this table contains the detailed interconnection information between the various nodes in the GRAPH. The size of the table is equal to the number of branches in the GRAPH. Each word in the table represents a branch, say b_j, in the GRAPH and contains

(i) the address of a word in the Node table which corresponds to the node to which b_j leads, and

(ii) the address of a word in a Branch table which stores the current busy-idle state of the branch b_j.

Consider, for example, the GRAPH of Fig. 2(a) and suppose that the occupancy of branches b_1 and b_2 is p_1; the occupancy of branches b_3, b_4, b_5 and b_6 is p_2; and the occupancy of branches b_7 and b_8 is p_3. The tables which the program would prepare are indicated in Fig. 6. The symbolic addresses, such as NW1, BRW3, etc., have been chosen for illustrative purposes only.

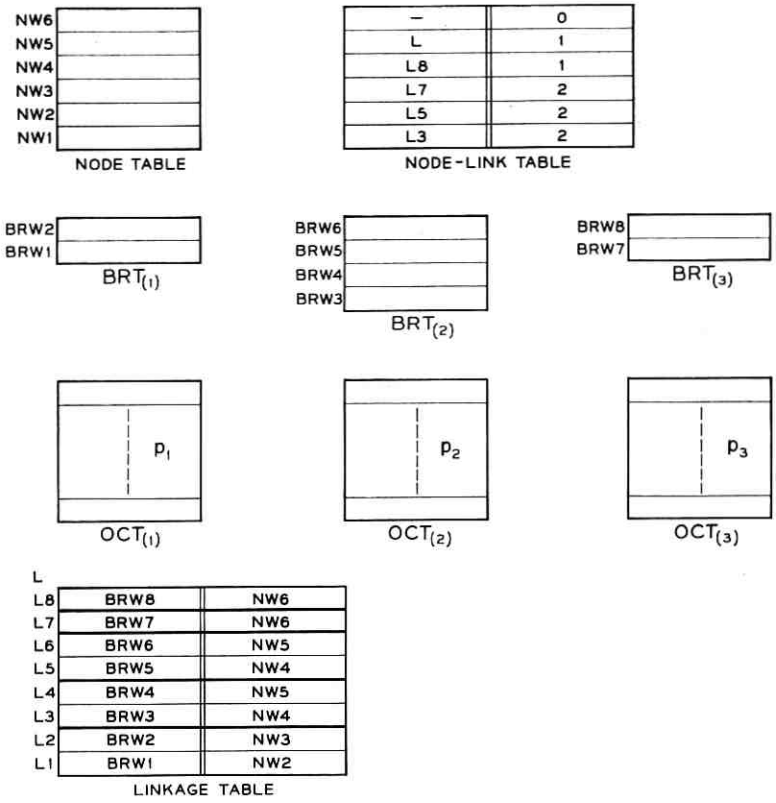


Fig. 6 — Computer memory layout for the GRAPH of Fig. 2 with occupancies p_1, p_2 and p_3.

3.2.3 *Organization of the NEASIM Algorithm*

In a previous section it was mentioned that the NEASIM algorithm requires

(*i*) an assignment of busy-idle states on the branches, and

(*ii*) a method of searching for the existence of a path.

These tasks are performed by three program segments which we shall call

(*i*) the PROBABILITY GENERATOR,

(*ii*) the BUSY-IDLE ASSIGNMENT, and

(*iii*) the MATCH routine.

For the sake of program efficiency (speed) the task of assigning busy-idle states to the branches is divided into two parts. The function of the PROBABILITY GENERATOR is to generate tables of busy-idle

bits — the Occupancy tables. The BUSY-IDLE ASSIGNMENT routine assigns busy-idle states to branches on each iteration of the algorithm. The function of the MATCH routine is to find whether or not there is a path, given a particular assignment of busy-idle states on the branches. Discussion of the internal logic of these programs will be deferred to Section IV. For the present we will be concerned only with their functions.

After the Graph Definition Cards have been read, enough information is available for the program to reserve appropriate space for the Node, Node-Link, Branch, Occupancy and Linkage tables. Once space allotment has been made and the necessary entries filled into the Node-Link and Linkage tables, the program reads the first Simulation Definition Card. Among other parameters, this card specifies the m different branch occupancies desired. At this point the PROBABILITY GENERATOR fills in the Occupancy tables. When this routine has completed its task, each bit in $\mathrm{OCT}_{(i)}$, $i = 1, \cdots, m$, will contain a binary one with probability p_i and a zero with probability $1 - p_i$. (Recall that the NEASIM convention is that one represents busy and zero represents idle.) The contents of the Occupancy tables remain unaltered for the entire time that the program is seeking an estimate for one point on the blocking curve.

The storage of many computers is organized into sequentially numbered words, each of which consists of a fixed number of contiguous bits, and each of which is addressable by the stored program for logical and algebraic manipulation. The number of bits in an IBM 7090 word is 36, so that each word in the Occupancy tables represents 36 independent busy-idle states — thus allowing for 36 independent repetitions of the NEASIM algorithm.

After the Occupancy tables have been prepared, various counters are set to zero and the Node table is initialized. As the contents of the Node table indicate for each node the presence or absence of a path from the particular node to the origin, the program takes the attitude of the man from Missouri and assumes there is no path until one is proven to exist. Thus the words of the Node table are initially set to all 1's except for the first node, which always contains zeros.

At this point the program enters the BUSY-IDLE ASSIGNMENT routine. To assign busy-idle states to the branches, this routine steps through each of the Branch tables and for each word in $\mathrm{BRT}_{(i)}$, a word from $\mathrm{OCT}_{(i)}$ is selected at random and its contents duplicated in the Branch table word. When BUSY-IDLE ASSIGNMENT has been completed, the program enters the MATCH routine.

Using the linkage information stored in the Node-Link and Linkage tables, the MATCH routine performs its logic on the Branch and Node

tables to find whether or not there is a path through the graph for each of the 36 independent busy-idle configurations. The operation of this routine is analyzed in Section 4.2.

When MATCH has completed its work, the number of 1's in the word of the Node table corresponding to the terminal node in the GRAPH (NW6 in Fig. 6) will be the number of times there was no path through the graph in the 36 trials. This number of blocked attempts is noted and 36 is entered into a "run-length" counter. The Node table is reinitialized, busy-idle states reassigned and the algorithm repeated again and again. The repetitions will terminate when one of two situations occurs at the end of the MATCH routine. Either

(*i*) enough attempts have been scored to guarantee the precision called for on the Simulation Definition Card, or

(*ii*) the maximum number of attempts specified on the card has been exceeded.

When the repetitions terminate, the proportion of blocked attempts is calculated and the results are printed out. The next Simulation Definition Card is read and the process starts over for the next point on the blocking curve.

3.2.4 *Storage Requirements and Execution Speed*

The largest GRAPH that can be handled by the present version of NEASIM is determined by the core storage available in a particular computer. Total storage required is given by the expression

$$P + 2(N + B) + O$$

where:

P is the number of storage locations required by the NEASIM program itself (about 2000 words),

N is the number of nodes in the GRAPH,

B is the number of branches in the GRAPH, and

O is the storage required for Occupancy tables — typically $m \times 1024$ where $1 \leqq m \leqq 8$ is the number of occupancy tables.

The NEASIM algorithm is designed for rapid execution on the IBM 7090. Average speed depends principally on the size of the GRAPH. Typical speeds range from about 600 trials per second (550-branch GRAPH) to about 5000 trials per second (48-branch GRAPH).

IV. PROBABILITY GENERATOR AND MATCH ROUTINES

The functions of the two routines called PROBABILITY GENERATOR and MATCH were mentioned in the previous section. The present section is concerned with the internal logic of these programs.

The algorithm used by PROBABILITY GENERATOR is due to W. C. Jones.[28] It is felt that this heretofore unpublished algorithm is of sufficiently widespread interest to be included in this paper.

A word on notation: throughout Section IV we shall use the notation $C[A]$ to mean "the contents of A," where A represents some bit in the computer memory. (The reader is therefore cautioned to distinguish between "bit A" and its contents $C[A]$.)

Two computer instructions which will be referred to repeatedly are the logical (inclusive) "OR" and logical "AND" instructions. Instructing the computer to perform an OR on two bits will guarantee that the result will be 1 if, and only if, either or both of the bits contain 1; while an AND yields 1 if, and only if, both bits contain 1. When we OR bit A to bit B, then we shall say that we "perform $[A]$ OR $[B]$;" when we AND bit A to bit B, then we say that we "perform $[A]$ AND $[B]$."

4.1 *The PROBABILITY GENERATOR*

The purpose of PROBABILITY GENERATOR is to generate the occupancies $p_1, \cdots, p_m$. We mentioned in Section 3.2.2 that this subroutine will fill up $\mathrm{OCT}_{(i)}$ $(i = 1, \cdots, m)$ with 36-bit words each of whose bits contains 1 with probability $p_i (i = 1, \cdots, m)$. In the sequel we will focus our attention on (a representative) one of these 36 bits — keeping in mind that the program is actually working on 36 bits independently and in parallel.

Suppose we wish to generate a random binary variable which takes the value 1 with probability $p (0 < p < 1)$ and to place our result in bit X. The algorithm, when terminated, will give $\Pr\{C[X] = 1\} = p$.

The first action taken by PROBABILITY GENERATOR is to express p as a binary fraction to 10 places.* This fraction is then scanned from right to left until the first bit is found which contains a 1. The algorithm uses this binary fraction of $n \leqq 10$ places determined by the scan. We can therefore, without loss of generality, express p as

$$p = 0.b_n b_{n-1} \cdots b_2 b_1 \qquad b_1 = 1; b_2, \cdots, b_n = 0 \text{ or } 1.$$

A digit selected from a random binary number will be referred to as a "random bit" in the following algorithm. In a random binary number, the value of each digit is 1 with probability $\frac{1}{2}$.

Algorithm:

(i) Set $j = 1$. Generate a random bit, say r_1, and store it in X. Thus $C[X] = r_1$.

* The number of places is arbitrary. Ten was chosen to make the round-off error smaller than 0.001, since occupancies are specified on the Simulation Definition Cards to three decimal places.

(*ii*) If $j = n$, go to step (*vii*). Otherwise, increase j by 1 and continue.

(*iii*) Generate another random bit r_j and store it temporarily in some bit, say R. Thus $C[R] = r_j$.

(*iv*) If $b_j = 0$, go to step (*v*). If $b_j = 1$, go to step (*vi*).

(*v*) Perform [R] AND [X]; store in X. Go to step (*ii*).

(*vi*) Perform [R] OR [X]; store in X. Go to step (*ii*).

(*vii*) Stop.

We observe that step (*ii*) is performed (iterated) exactly n times. Let P_j be the probability that $C[X] = 1$ after the jth iteration. The assertion is that $P_n = p$ — i.e., after the algorithm is terminated, X will contain 1 with probability p.

Proof of Algorithm:

Consider two bits A and B, and let $p_A = \Pr\{C[A] = 1\}$, and $p_B = \Pr\{C[B] = 1\}$. When the contents of A and B are independent, if we perform [A] AND [B], the probability of the result being 1 is

$$p_A p_B ,$$

while if we perform [A] OR [B], the probability of the result being 1 is

$$p_A + p_B - p_A p_B .$$

Now, $\Pr\{r_j = 1\} = \Pr\{r_j = 0\} = \frac{1}{2} \qquad j = 1, \cdots, n.$

Therefore if step (*v*) is executed,

$$P_{j+1} = \tfrac{1}{2} P_j \qquad j = 1, \cdots, n - 1 \tag{3}$$

while if step (*vi*) is executed,

$$\begin{aligned} P_{j+1} &= \tfrac{1}{2} + P_j - \tfrac{1}{2} P_j \qquad j = 1, \cdots, n - 1 \\ &= \tfrac{1}{2} P_j + \tfrac{1}{2}. \end{aligned} \tag{4}$$

But step (*v*) is executed only if $b_{j+1} = 0$, and step (*vi*) is executed only if $b_{j+1} = 1$. Hence (3) and (4) can be combined into

$$P_{j+1} = \tfrac{1}{2} P_j + \tfrac{1}{2} b_{j+1} \qquad j = 1, \cdots, n - 1.$$

Since $P_1 = \frac{1}{2}$, it follows by induction that

$$P_n = \frac{b_n}{2} + \frac{b_{n-1}}{2^2} + \cdots + \frac{b_2}{2^{n-1}} + \frac{b_1}{2^n} = p$$

and our assertion is proved.

4.2 *The MATCH Routine*

The MATCH routine is entered by the program after the busy-idle states have been assigned to the branches. Its purpose is to find whether or not there is a path through the graph for any particular assignment. In the present section we first derive a recursive set-theoretic formula which may be used to determine whether there is a path. The remainder of the section shows how the recursion is carried out by the computer to produce an estimate which converges to the blocking probability of the GRAPH.

In Section 2.2 we saw how the "path enumeration approach" could, in theory at least, be employed in determining the blocking probability. We again take the path enumeration approach, but from a slightly altered point of view:

Instead of trying to find whether there is a path through the *entire* graph at one fell swoop (which amounts to finding whether there is a path from the first node to the last node), we shall try to find whether there is a path from the first node to *each of the nodes* in the GRAPH. While this approach may appear to inject an unnecessary complication, we will see how, by stepping through the GRAPH in an orderly fashion, this approach lends itself naturally to computer programming. We begin by investigating, somewhat further, the geometrical structure of a GRAPH.

Consider the general GRAPH (whose structure is shown schematically in Fig. 7) and suppose that there are a total of ν nodes. Each GRAPH has a first node, N_1, a last node, N_ν, and several intermediate "stages" of nodes. The notion of "stage" is made more precise by the following *definition:* a node N is said to be in stage $s (s = 1, 2, \cdots)$ if, and only if, all paths from N_1 to N contain no more than $s - 1$ branches, and there exists at least one path from N_1 to N which contains exactly $s - 1$ branches.

Any GRAPH will thus contain some number $S \geqq 2$ of stages, where the first and last (Sth stage) consist, respectively, of the single nodes N_1 and N_ν. Let n_s be the number of nodes in stage s, $s = 1, \cdots, S$ (thus $n_1 = n_S = 1$); and define a quantity m_s by

$$m_s = \sum_{i=1}^{s} n_i .$$

We choose to order the nodes of the GRAPH in the following manner: to each of the n_s nodes in stage s ($s = 2, \cdots, S$) assign arbitrarily, but uniquely, one of the integers

$$m_{s-1} + 1, m_{s-1} + 2, \cdots, m_{s-1} + n_s \qquad s = 2, \cdots, S.$$

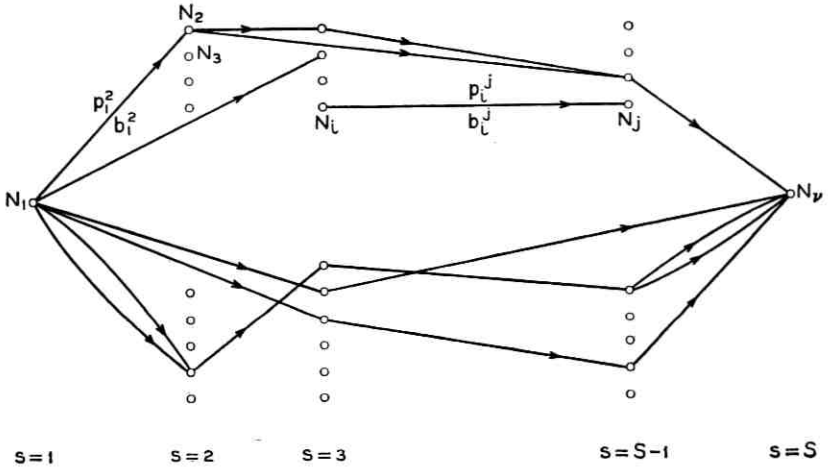


Fig. 7 — A general GRAPH.

The nodes of the GRAPH are hence totally ordered and may therefore be denoted by N_i, where $i = 1, \cdots, \nu$.

When N_i and N_j are connected by one or more parallel branches, where N_i, N_j are arbitrary nodes, we say that, for $i < j$, N_i is *directly connected* to N_j through each of these branches and write $N_i \rightarrow N_j$. We shall suppose, without loss of generality, that if $N_i \rightarrow N_j$, then there is only one connecting branch, b_i^j. (This restriction is assumed in order to avoid introducing yet another index, but will not affect the subsequent results. When a computer instruction involving b_i^j is described, we shall understand that the computer is to execute this instruction for all the b_i^j for which $N_i \rightarrow N_j$.)

Let us now examine the graph, asking the question for every node N_j, "Is there an available path from N_1 to N_j?" Our objective is to answer the question for $j = \nu$.

For each node N_j $(j > 1)$ consider the branches b_i^j for which $N_i \rightarrow N_j$. Clearly, for each of these branches, if there was no available path from N_1 to N_i or there is no available path through the branch b_i^j, or both, then, and only then, will there be no available path from N_1 to N_j which passes through N_i. More precisely, let

X_j be the event: there is no available path from N_1 to N_j,

Y_i^j be the event: there is no available path from N_1 to N_j which passes through N_i,

B_i^j be the event: branch b_i^j is not available (busy).

Then

$$Y_i^j = X_i \cup B_i^j$$

$$X_j = \bigcap_{\substack{i<j \\ N_i \to N_j}} Y_i^j = \bigcap_{\substack{i<j \\ N_i \to N_j}} (X_i \cup B_i^j).$$

Consider all the nodes N_i such that $N_i \to N_j$. Suppose there are k_j such nodes. Call them $N_{i_1}, N_{i_2}, \cdots, N_{i_{k_j}}$, where $i_1 < i_2 < \cdots < i_{k_j}$. Next, define an event $Z_j^{(m)}$ recursively by

$$Z_j^{(m)} = (X_{i_m} \cup B_{i_m}{}^j) \cap Z_j^{(m-1)} \qquad m = 2, \cdots, k_j \qquad (5)$$
$$N_{i_m} \to N_j$$
$$j = 2, \cdots, \nu$$

and

$$Z_j^{(1)} = X_{i_1} \cup B_{i_1}{}^j \qquad N_{i_1} \to N_j \qquad (6)$$
$$j = 2, \cdots, \nu.$$

It then follows that

$$Z_j^{(k_j)} = X_j \qquad j = 2, \cdots, \nu \qquad (7)$$

and in particular

$$Z_\nu^{(k_\nu)} = X_\nu$$

where X_ν is the event "there is no path through the graph." The MATCH routine is based upon these formulas.

We now focus our attention on (a representative) one bit in each word of the Node and Branch tables — again keeping in mind that the program is actually working on a full word of bits independently and in parallel. The program will thus execute 36 simultaneous iterations of the MATCH algorithm, which we now proceed to evolve.

When MATCH is entered, the following state of affairs prevails:

(*i*) To each node N_i and to each branch $b_i{}^j$ there has been uniquely assigned one bit of computer memory, so that we can henceforth refer to these *bits* as bit N_i and bit $b_i{}^j$.

(*ii*) $C[b_i{}^j] = 1$ with the appropriate occupancy $p_i{}^j = \Pr\{\text{branch } b_i{}^j \text{ is busy}\}$. (These $p_i{}^j$ were called $p_1, \cdots, p_m$ in Sections II and III.)

(*iii*) The linkage information between the node bits N_i and the branch bits $b_i{}^j$ is stored in the Node-Link and Linkage tables.

It follows that representation of the event $Z_j^{(1)} = X_{i_1} \cup B_{i_1}{}^j$ of formula (6) can be achieved in the computer by performing

$$[N_{i_1}] \text{ OR } [b_{i_1}{}^j]$$

provided that we always set $C[N_1] = 0$; and the event $Z_j^{(k_j)} = X_j$ of (7) can be represented by executing the following steps.

(*i*) Perform $[N_{i_1}]$ OR $[b_{i_1}{}^j]$; store in N_j.
(*ii*) Set $m = 1$.
(*iii*) If $m = k_j$ go to step (*vi*). Otherwise, continue.
(*iv*) Increase m by 1.
(*v*) Perform $([N_{i_m}]$ OR $[b_{i_m}{}^j])$ AND $[N_j]$; store in N_j. Go to step (*iii*).
(*vi*) Stop.

We observe that the algorithm may be condensed if we initially set $C[N_j] = 1 (j \geqq 2)$:

(*i*) Set $m = 1$.
(*ii*) Perform $([N_{i_m}]$ OR $[b_{i_m}{}^j])$ AND $[N_j]$; store in N_j.
(*iii*) If $m = k_j$, go to step (*v*). Otherwise continue.
(*iv*) Increase m by 1. Go to step (*ii*).
(*v*) Stop.

This initialization of the node bits — $C[N_1] = 0$, $C[N_j] = 1$ for $j = 2, 3, \cdots, \nu$ — was mentioned in Section 3.2.3 and can be interpreted as meaning "there is always a path from N_1 to N_1; there is no path from N_1 to N_j $(j > 1)$ until proven otherwise."

Summarizing, the steps that the computer may take to represent the event $Z_\nu^{(k_\nu)} = X_\nu$ are

(M1) Set $C[N_1] = 0$. Set $C[N_j] = 1, j > 1$.
(M2) For each bit N_j, $j = 2, \cdots, \nu$, and in the order

$$N_2, N_3, \cdots, N_\nu$$

find all bits N_i for which $N_i \rightarrow N_j$ and perform

$$([N_i] \text{ OR } [b_i{}^j]) \text{ AND } [N_j]; \text{ store in } N_j.$$

Now, the statement "find all bits N_i for which $N_i \rightarrow N_j$" implies that the input to the program is such that it specifies to the computer which nodes N_i are directly connected to N_j. It was felt that a more straightforward input format would be to specify to which nodes, N_j, each node N_i connects. With the linkage information stored in this fashion,* step (M2) may be replaced by the equivalent step

(M2′) For each bit N_i, $i = 1, \cdots, \nu - 1$, and in the order

$$N_1, N_2, \cdots, N_{\nu-1}$$

* A careful perusal of Section 3.2.2 will show that this is indeed the way the linkage information is stored in the Node-Link and Linkage tables.

find all bits N_j for which $N_i \rightarrow N_j$ and perform

$$([N_i] \text{ OR } [b_i^j]) \text{ AND } [N_j]; \text{ store in } N_j .$$

Steps (M1) and (M2′) constitute the MATCH algorithm for determining whether there is a path through the graph. When the MATCH algorithm is terminated, $C[N_\nu] = 1$ if, and only if, there is no path, so that $\Pr\{C[N_\nu] = 1\} = \Pr\{\text{no path}\}$.

Suppose the algorithm is iterated n times. In any one iteration the event $C[b_i^j] = 1$ is generated with probability p_i^j and independently of the event $C[b_i^j] = 1$ in any other iteration. But $\Pr\{C[N_\nu] = 1\}$ after a given iteration is clearly only a function of the probabilities $\Pr\{C[b_i^j] = 1\}$. Hence, N_ν will contain 1 with the same probability, B, after each iteration, and we conclude that n iterations of the algorithm constitute a sequence of n Bernoulli trials with probability B of success on each trial. Let ξ_i be a random variable such that

$$\begin{aligned} \xi_i &= 1 \text{ if } C[N_\nu] = 1 \text{ after the } i\text{th trial} \\ &= 0 \text{ if } C[N_\nu] = 0 \text{ after the } i\text{th trial} \end{aligned}$$

and let

$$B_n = \xi_1 + \xi_2 + \cdots + \xi_n .$$

An application of the Law of Large Numbers (Ref. 29, p. 189) gives

$$\lim_{n\to\infty} (B_n/n) = B.$$

Identifying B_n/n with $B_G^{(n)}(p_1, \cdots, p_m)$ of Section 3.1 and B with $B_G(p_1, \cdots, p_m)$ of the same section, this last result is equivalent to, and proves the assertion that,

$$\lim_{n\to\infty} B_G^{(n)}(p_1, \cdots, p_m) = B_G(p_1, \cdots, p_m).$$

Repeating the experiment often enough and dividing the number of times $C[N_\nu] = 1$ (i.e., the number of blocked attempts) by the number of iterations n (i.e., the number of attempts) will therefore yield a precise estimate of the blocking probability of the GRAPH. How large n has to be in order to attain any given degree of precision is discussed in the next section.

V. RELIABILITY CONSIDERATIONS

Since any simulation is nothing but an experimental measurement, and hence subject to statistical fluctuations, it is necessary to assess the

reliability of the results. Unfortunately, most of the testing for a given simulation must be done on an *a posteriori* basis and there is, in general, rarely sufficient *a priori* information on which to base the decision of how long the run is to be. The need for such information becomes even more important when the cost factor in computer simulations is considered. It will be shown below that the NEASIM procedure is one which allows for *a priori* determination of run length. The decision of how long the run is to be is based on the desired precision and is made by the computer.

In Section 4.2 we introduced the random variable B_n, the number of times $C[N_\nu] = 1$ in n repetitions of the experiment. B_n will now be called "the number of calls blocked in a simulation run of n calls."* The estimator B_n/n was seen to converge to the blocking probability B, and is indeed a maximum likelihood, mean-unbiased, minimum variance estimator.

To generate the random binary numbers of Section 4.1, NEASIM uses the well-known multiplicative congruential method. This method for pseudo-random number generation has been discussed, tested, and used by numerous investigators[31-36] since it was first proposed by Lehmer.[37] Although the method has been demonstrated to generate 35-bit random binary numbers, there is some measure of cyclic behavior in the low-order bits. NEASIM forms a word of 36 random bits by combining the most random halves (18 high-order bits) of two 35-bit words generated by this method. A further check on the randomness is provided by NEASIM itself, which, as part of its output, prints the generated branch occupancies. We therefore assume that the events $C[N_\nu] = 1$ are independent for individual trials of the experiment. The results of any given computer run should thus be binomially distributed (see also Section 4.2), and hence asymptotically normal.

As an illustration, we again turn to the GRAPH of Fig. 3. For a branch occupancy of $p = p_2 = 0.5$, formula (2) yields $B = 0.0525$. The NEASIM program, for a run of $n = 201{,}600$ calls, gave $B = 0.0529$ — an error within 1 per cent. (This run took some 40 seconds of

* The new nomenclature is chosen to indicate a different interpretation of what NEASIM is doing: NEASIM looks at the configuration of possible paths through the network between two subscribers, takes "snapshots" of the current busy-idle states of the links in these paths, and then finds if there is a path for each snapshot. Thus the program is essentially running calls through a portion of the network — the portion of the network being a representative one, and hence one from which significant statistical data can be extracted to describe the performance of the entire network. A similar approach was taken by A. Feiner, W. C. Jones, and others,[30] who wrote "abbreviated" simulations in which the busy-idle state data were taken from previously run full-scale simulations, but for which a new program had to be written for each new network to be simulated.

computer time.) To study the normality, the number of blocked calls was noted for every 1008 calls run. Since B_n is the number of blocked calls in a run of n calls, the binomial distribution function in this case is

$$\Pr\{B_n \leqq \beta\} = \sum_{k=0}^{\beta} \binom{1008}{k} B^k(1 - B)^{1008-k}$$

and the approximating normal distribution function (Ref. 29, p. 172) is $\Phi(x_{\beta+\frac{1}{2}})$, where

$$x_t = (t - 1008B)h$$

$$h = [1008B(1 - B)]^{-\frac{1}{2}}$$

and

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}y^2} dy$$

is the standard normal distribution function. Fig. 8 is a plot on probability paper of the cumulative frequency distribution determined by the computer, and the theoretical normal distribution line for $B = 0.0525$.

In order to obtain meaningful results, it was felt that the run-length should be determined by the following *criterion:* the number of trials shall be large enough so as to give 95 per cent confidence that the estimator lies within a fixed percentage of the true value of B. Since the blocking probability is unknown at the outset, this criterion is more useful than requiring the estimator to lie in a fixed interval about B. Thus, we wish to choose n large enough so that

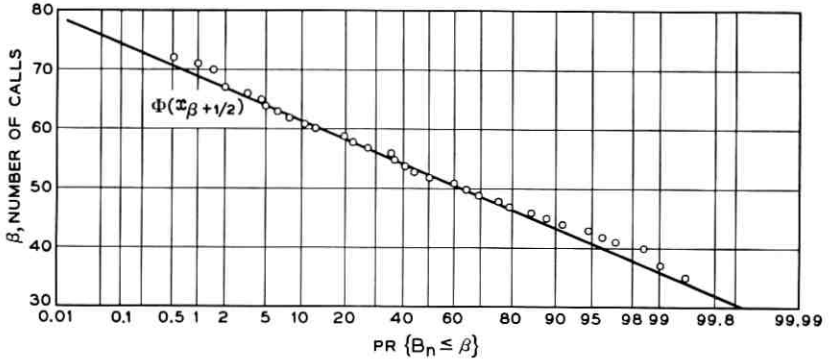


Fig. 8 — Cumulative distribution for GRAPH of Fig. 3 ($p = p_2 = 0.5$) in a run of 201,600 calls in which the number of blocked calls was noted for every 1,008 calls run. The points are experimental data; the line is the theoretical distribution, $\Phi(x_{\beta+\frac{1}{2}})$

$$\Pr\left\{\left|\frac{B_n}{n} - B\right| \leqq aB\right\} \geqq 0.95$$

where a is an input parameter to the program representing the desired percentage. Rearranging terms, we get

$$\Pr\left\{\left|\frac{B_n - nB}{\sqrt{nB(1-B)}}\right| \leqq a\sqrt{\frac{nB}{1-B}}\right\} \geqq 0.95$$

where, for n sufficiently large, the distribution of the random variable

$$\frac{B_n - nB}{\sqrt{nB(1-B)}}$$

approaches the standard normal. Our requirement will therefore be satisfied if

$$a\sqrt{\frac{nB}{1-B}} \geqq 1.96$$

or

$$n \geqq \frac{3.84}{a^2}\left(\frac{1}{B} - 1\right). \tag{8}$$

It follows that the number of trials should be increased as the blocking probability decreases in order to maintain 95 per cent confidence that the estimator lies within a fixed percentage of the true value — as intuition dictates. On the other hand, for a fixed B, as n increases, the 95 per cent confidence limits will narrow. This is displayed in Fig. 9, where the program results for the GRAPH of Fig. 3, for $B = 0.0525$, are seen to lie within the confidence limits.

Furthermore, since $(1/B) - 1 \leqq (1/B)$ for all B in the unit interval, the requirement will be met if

$$n \geqq \frac{3.84}{a^2}\cdot\frac{1}{B}.$$

But for large n, $B \approx (B_n/n)$. Hence, making the substitution we obtain

$$B_n \geqq \frac{3.84}{a^2}.$$

As long as the number of simulated calls is large enough so that the number of blocked calls is at least $3.84/a^2$, there is 95 per cent confidence that B_n/n is within aB of B.

NEASIM was written to accept several values of a ranging from 0.05

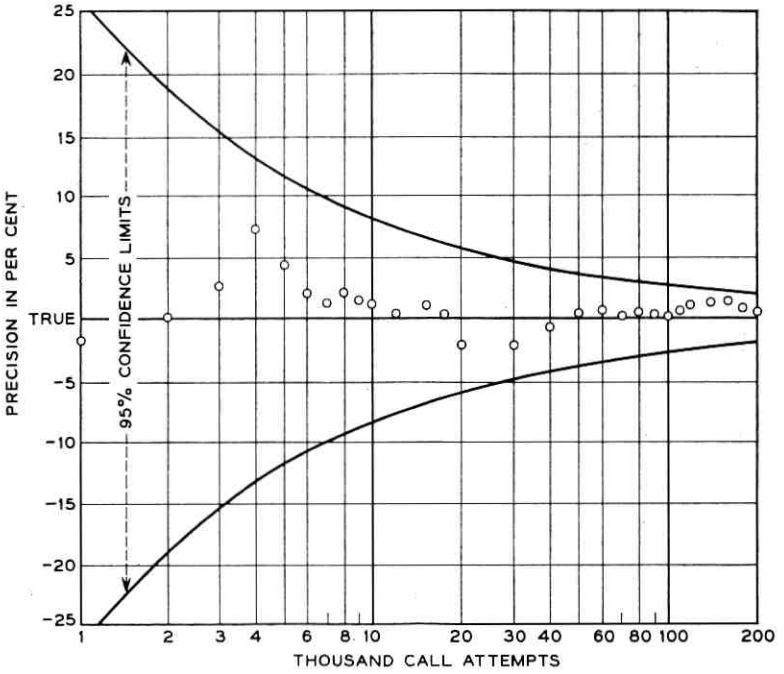


Fig. 9 — NEASIM blocking estimates for GRAPH of Fig. 3 vs run length ($p = p_2 = 0.5$, $B = 0.0525$).

to 1.00. From the specified value, it determines the minimum number of blocked calls necessary to guarantee this precision. Fig. 10 shows the results of the simulation of the GRAPH of Fig. 3 for $a = 0.10$ and 0.50. In GRAPHs where B is very small, it is usually unnecessary to estimate B with a very high degree of precision — especially at the expense of costly computer runs. An upper limit for n is therefore also specified. The computer then proceeds with the simulation until it either exceeds the lower limit on B_n or the upper limit on n.

In the latter case, the reliability can be assessed as follows. Since

$$\Pr\left\{\frac{|B_n - nB|}{\sqrt{nB(1-B)}} \leqq 1.96\right\} \geqq 0.95$$

we can rearrange terms and obtain

$$\Pr\{dB^2 + eB + f \leqq 0\} \geqq 0.95$$

where

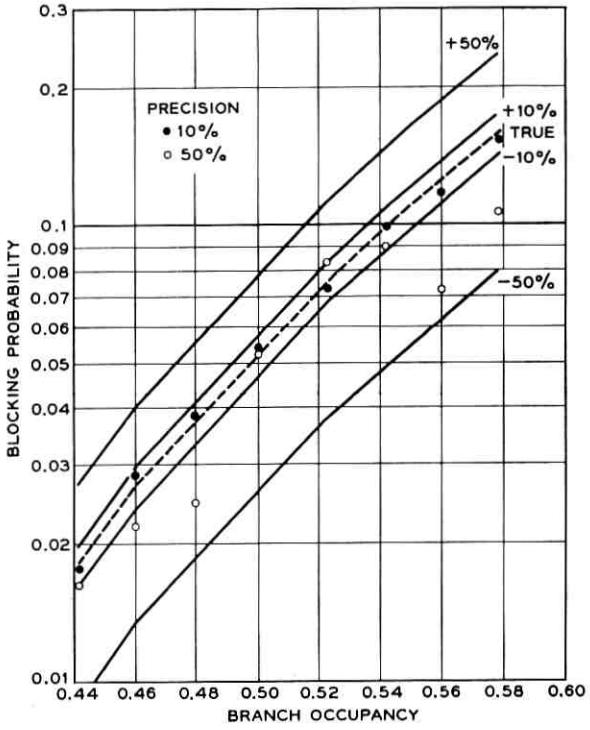


Fig. 10 — NEASIM blocking estimates for GRAPH of Fig. 3 vs branch occupancy $p = p_2$.

$$d = n^2 + 3.8416\, n$$

$$e = -(2nB_n + 3.8416\, n)$$

$$f = B_n^{\,2}.$$

It is easily verified that for all n, $B_n > 0$, the parabola $dB^2 + eB + f$ is concave upward, has real roots, and that B_n/n lies between the roots whenever $B_n < n$. Thus for any simulation run, the 95 per cent confidence interval can be obtained by solving for the roots.

Now in all cases of interest, the product of the roots,

$$f/d = B_n^{\,2}/(n^2 + 3.8416\, n),$$

can be closely approximated by $(B_n/n)^2$, so that B_n/n can be taken as the geometric mean of the roots. Suppose the higher root is B_1 and the lower root B_2 ; then

$$B_1 = kB_n/n$$

$$B_2 = (1/k)B_n/n.$$

The roots were calculated over a range of values of B_n and n, and the results are displayed in Fig. 11, where k is plotted as a function of B_n/n for various values of n. As an example, suppose that in a run of 80,000 calls, $B_n = 80$ were found blocked. Then $B_n/n = 0.001$ and Fig. 11 gives $k = 1.244$. Hence $B_1 = 1.244\ (0.001) = 0.00124$, $B_2 = 0.001/1.244 = 0.000803$, and there is 95 per cent confidence that $0.000803 \leqq B \leqq 0.001244$. The importance of the case under consideration would then determine whether a longer simulation is necessary.

Since the GRAPH geometry may be as complex as indicated in Fig. 4,

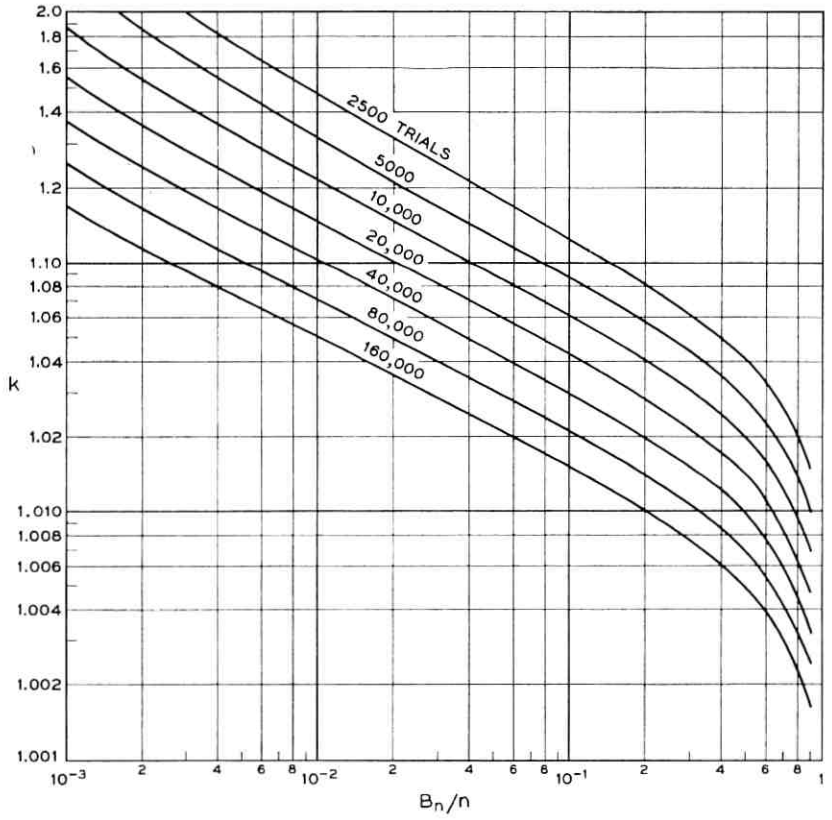


Fig. 11 — The 95 per cent confidence limits for $B : (1/k)(B_n/n) \leqq B \leqq k(B_n/n)$.

it is important to be able to verify that no error was made in mapping the geometry into the computer memory. To this end, a program was prepared by Miss D. Logan which, using the Graph Definition Cards as input and an SC-4020 Microfilm Recorder, draws a picture showing the GRAPH geometry.

VI. POSSIBILITIES FOR INCREASED REALISM

In the preceding section it was shown that the NEASIM program produces precise, reliable results in comparison with Lee's probability linear-graph model of switching networks. Unfortunately, in complex switching networks, substantial differences may exist between the estimates obtained with Lee's analytical technique and actual performance determined by field measurement or full-scale (complete) simulation. These differences appear to be largely attributable to the unreality of the assumption [(e) of Section 2.1] that the link random variables are independent.

The dependence that exists between the different links in a network stage and between the links in adjacent stages is incompletely understood, but nonetheless real. Attempts to take account of interlink dependence by judicious modification of link occupancy values have been moderately successful in calculations and may, of course, be employed in NEASIM runs. However, the nature of the NEASIM process suggests alternate approaches which may ultimately prove to be more fruitful. While considerable success has been obtained with some of the techniques discussed in this section, much remains to be accomplished.

6.1 *Dependence Effects within a Switching Stage*

Two possibilities for increased realism within the confines of a single link stage appear worthy of mention. The NEASIM program typically assigns link stage busy-idle states from a binomial distribution. An obvious suggestion (but one upon which little work has been done) would be to modify the PROBABILITY GENERATOR and/or BUSY-IDLE ASSIGNMENT routines in such a way as to produce busy-idle state assignments taken from various distributions — Jacobaeus' E distribution,[4,6] for example.

A second approach, which has been extensively used, is to incorporate program routines *between* BUSY-IDLE ASSIGNMENT and execution of the MATCH routine. These routines examine the random busy-idle assignments and make appropriate assignment changes where GRAPH

geometry or other considerations indicate. An example of such a routine is one which is designed to insure that there exists at least one "sure-idle" branch out of an $n \times n$ input switch. If calls are only placed between idle network terminals, at least one branch out of an $n \times n$ input switch must be idle. But if the branches leaving the initial GRAPH node are specified at occupancy p, then the program would normally make all these branches busy with probability p^n. The SURE-IDLE routine, upon finding such an assignment, will select an initial branch at random and make it idle. This action, of course, disturbs the otherwise binomial distribution of branch states and should be taken into account when branch occupancy values are specified.

6.2 *Interstage Dependence Effects*

An interesting technique for introducing interstage dependence exists within the NEASIM framework and is based upon an obvious extension of the SURE-IDLE routine just described. Briefly stated, a routine could be designed to examine the busy-idle assignments made on the input and output branches of each node of the GRAPH. Acting on knowledge of the switch geometry and other factors, the "dependencing" routine could change the initial busy-idle assignments where necessary to make them more realistic. Only a very simple and admittedly inaccurate routine has been used to date with, however, a remarkable increase in the "realism" of the results obtained.

The simple-minded DEPENDENCING routine currently employed assumes that every GRAPH node is in reality an $n \times n$ switch. It observes that in an $n \times n$ switch each input branch has probability $1/n$ of being connected to any particular output branch. It attempts to implement its idea of reality by, for each input branch, examining each output branch and forcing the output branch state to agree with the input branch state with probability $1/n$. While this routine can produce rather quaint effects, such as duplicating the state of one input branch on all output branches, it does possess the virtues of simplicity (rapid execution) and a basically correct notion of interstage dependence.

That the use of the SURE-IDLE and DEPENDENCING routines can be effective is demonstrated in Fig. 12. The results shown in the figure are for a moderately complex eight-stage switching network whose GRAPH has 232 branches. NEASIM results with and without dependencing and sure-idles are compared with the results of a full-scale simulation. The improvement in realism possible with the rudimentary routines just described appears quite dramatic.

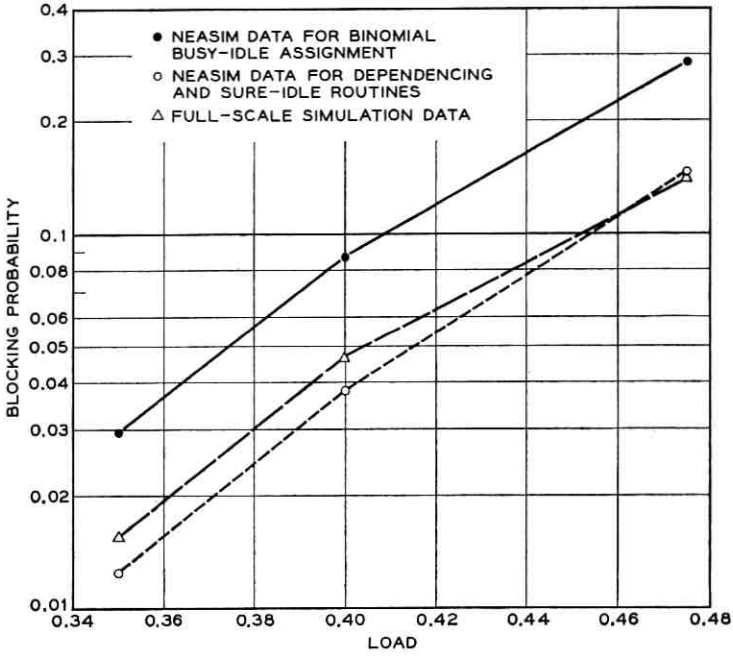


Fig. 12 — Comparison of NEASIM results (with and without dependencing and sure-idles) with full-scale simulation data for a realistically complex network.

VII. CONCLUSION

The NEASIM approach to switching network simulation has achieved its major goal, the mechanization of a widely employed — if somewhat unrealistic — technique for load-loss analysis. The applications of the GRAPH model need no longer be restricted by the overpowering computational difficulties brought on by GRAPH complexity.

Furthermore, the simulation of an analytical model concept basic to NEASIM seems to open new areas for profitable exploration in the analysis of switching systems — and perhaps other stochastic systems as well. The success of the elementary realism-injecting routines suggests that further research along this line may be rewarding.

Finally, the degree of realism in results attained so far, coupled with the ease of application, has produced what amounts to a new tool for use in both the design and engineering of new switching networks. It is now feasible to achieve relatively complete and accurate load-loss engineering data on complex switching systems well in advance of actual field experience.

VIII. ACKNOWLEDGMENTS

The authors wish to express their appreciation to W. S. Hayward, A. Descloux and J. G. Kappel for many stimulating discussions. In particular, Mr. Hayward suggested a significant improvement in the MATCH routine. The work of Miss D. Logan in writing several parts of the program is greatly appreciated.

REFERENCES

1. Kosten, L., The Historical Development of the Theory of Probability in Telephone Traffic Engineering in Europe, Teleteknik, **1,** 1957, pp. 32–40.
2. Wilkinson, R. I., The Beginning of Switching Theory in the United States, Teleteknik (English Edition), **1,** 1957, pp. 14–31.
3. Syski, R., *Introduction to Congestion Theory in Telephone Systems*, Oliver and Boyd, London, 1960.
4. Jacobaeus, C., Blocking Computations in Link Systems, Ericsson Review, No. 3, 1947, pp. 86–100.
5. Lundkvist, K., Method of Computing the Grade of Service in a Selection Stage Composed of Primary and Secondary Switches, Ericsson Review, No. 1, 1948, pp. 11–17.
6. Jacobaeus, C., A Study on Congestion in Link Systems, Ericsson Technics, **48,** 1950, pp. 1–68.
7. Jensen, A., A Basis for the Calculation of Congestion in Crossbar Systems, Teleteknik, No. 3, 1952.
8. Lee, C. Y., Analysis of Switching Networks, B.S.T.J., **34,** 1955, pp. 1287–1315.
9. Elldin, A., Applications of Equations of State in the Theory of Telephone Traffic, Thesis, Stockholm, 1957.
10. Fortet, R., and Canceill, B., Probabilité de Perte en Selection Conjuguee, Teleteknik, **1,** 1957, pp. 41–55.
11. LeGall, P., Methode de Calcul de L'encombrement dans les Systèmes Téléphoniques Automatiques a Marquage, Ann. des Telecom., **12,** 1957, pp. 374–386.
12. Beneš, V. E., Heuristic Remarks and Mathematical Problems Regarding the Theory of Connecting Systems, B.S.T.J., **41,** 1962, pp. 1201–1247.
13. Beneš, V. E., Algebraic and Topological Properties of Connecting Networks, B.S.T.J., **41,** 1962, pp. 1249–1274.
14. Beneš, V. E., A "Thermodynamic" Theory of Traffic in Connecting Networks, B.S.T.J., **42,** 1963, pp. 567–607.
15. Beneš, V. E., Markov Processes Representing Traffic in Connecting Networks, B.S.T.J., **42,** 1963, pp. 2795–2837.
16. Frost, G. R., Keister, W., and Ritchie, A. E., A Throwdown Machine for Telephone Traffic Studies, B.S.T.J., **32,** 1953, pp. 292–359.
17. Gerlough, D. L., Simulation of Freeway Traffic by an Electronic Computer, Proc. Highway Research Board, 1956, pp. 543–547.
18. Goode, H. H., Pollmar, C. H., and Wright, J. B., The Use of a Digital Computer to Model a Signalized Intersection, Proc. Highway Research Board, 1956, pp. 548–557.
19. Clapham, J. C. R., A Monte Carlo Problem in Underground Communications, Operations Research Quarterly, **9,** No. 1, 1958, pp. 36–54.
20. Jennings, N. H., and Dickens, J. H., Computer Simulation of Peak Hour Operation in a Bus Terminal, Management Science, **5,** No. 1, 1958, pp. 106–120.
21. Stark, M. C., Computer Simulation of Street Traffic, NBS Tech. Note 119, U.S. Dept. of Commerce, Off. Tech. Services, Washington, D.C., 1958.
22. Gross, F. J., Simulation of Data Switching Systems on a Digital Computer, A.I.E.E. Transactions, Part I — Communications and Electronics, **46,** 1960, pp. 796–800.

23. Katz, J. H., Simulation of a Traffic Network, Communications of the ACM, **6,** 1963, pp. 480–486.
24. Dietmeyer, D. L., Gordon, G., Runyon, J. P., and Tague, B. A., An Interpretive Simulation Program for Estimating Occupancy and Delay in Traffic-Handling Systems Which Are Incompletely Detailed, A.I.E.E. Conference Paper CP 60-1090, San Diego, 1960.
25. Gordon, G., General Purpose Systems Simulator, Proc. Eastern Joint Computer Conference, 1961, pp. 87–104.
26. Weber, J. H., Some Traffic Characteristics of Communications Networks with Automatic Alternate Routing, B.S.T.J., **41,** 1962, pp. 769–796.
27. Bader, J. A., and Hayward, W. S., Computer Simulation as a Machine Aid to Switching System Design, A.I.E.E. Conference Paper CP 61-258, New York, 1961.
28. Jones, W. C., personal communication.
29. Feller, W., *An Introduction to Probability Theory and Its Applications*, John Wiley and Sons, New York, 1957.
30. Feiner, A., Jones, W. C., *et al.*, personal communication.
31. Juncosa, M. L., Random Number Generation on the BRL High-Speed Computing Machines, Report No. 855, Ballistics Research Laboratories, Aberdeen Proving Ground, Maryland, 1953.
32. Todd, J., and Taussky, O., Generation of Pseudo-Random Numbers, *Symposium on Monte Carlo Methods*, ed. Meyer, H. A., John Wiley and Sons, New York, 1956, pp. 15–28.
33. Bofinger, E., and Bofinger, V. I., A Periodic Property of Pseudo-Random Sequences, Jour. ACM, **5,** 1958, pp. 261–265.
34. Certaine, J. E., On Sequences of Pseudo-Random Numbers of Maximal Length, Jour. ACM, **5,** 1958, p. 353.
35. Greenberger, M., Appendix to Part II of Decision-Unit Models and Simulation of the United States Economy, M.I.T., January, 1958.
36. Greenberger, M., Random Number Generators, Preprints of the 14th Natl. Conf. ACM, September, 1959.
37. Lehmer, D. H., Mathematical Methods in Large-Scale Computing Units, *Proc. of a Second Symposium on Large-Scale Digital Calculating Machinery*, Ann. of the Computation Laboratory of Harvard University, **26,** 1951, p. 141.

Filling Factor and Isolator Performance of the Traveling-Wave Maser*

By F. S. CHEN and W. J. TABOR

(Manuscript received December 16, 1963)

In designing a large gain and simultaneously large instantaneous bandwidth traveling-wave maser (TWM), the filling factor and the isolator performance should be optimized.

The filling factor is a measure of the efficiency of interaction between the spin system of the maser material and the RF magnetic fields of the slow-wave structure. For the TWM using the 90° operation of ruby and the comb as the slow-wave structure, the c axis of the ruby should be parallel to the z axis of the structure (the direction of the signal wave propagation) for the largest filling factor. The improvement of the filling factor by the proper orientation of the c axis of the ruby is larger at the lower signal frequencies because the transition probability of ruby is more nearly linear at those frequencies.

The isolator should provide sufficient reverse absorption to make the TWM short-circuit stable and yet add the minimum forward absorption to the TWM. Both the reverse and the forward absorption of the isolator depend critically on the size of the ferrite disks and the position in which they are imbedded in the comb structure.

An analysis of the filling factor and isolator performance and its comparison with measurements was made. Together with Refs. 1 and 2, this paper is intended to reduce the amount of experimental work involved in developing traveling-wave masers. Although the discussion is centered on the comb-type ruby TWM, the data provided also apply to other tape slow-wave structures using different active crystals.

I. INTRODUCTION

In order to make a large gain and simultaneously large instantaneous bandwidth traveling-wave maser (TWM), it is necessary to orient the

* This work was supported in part by the U. S. Army Signal Corps under Contract No. DA 36-039-SC-89169.

active material so that the susceptibility tensor has a maximum interaction with the RF magnetic field of the slow-wave structure. Quantitatively this is expressed as a filling factor, F, where:[1]

$$F = F_v F_p . \tag{1}$$

F_v is the volume filling factor and represents the fraction of the RF magnetic energy that is inside the active material. If all of the RF magnetic energy is within the boundaries of the active material, then this factor can have its maximum value of one. F_p is the phase filling factor and represents how well the RF field can couple to the susceptibility tensor of the active material. For example, if the active material is represented by a susceptibility with a transverse component of magnetization which is circularly polarized, and a structure with the RF magnetic field that is circular and in the same sense of rotation, then F_p is equal to one. It is clear therefore that the maximum value of the total filling factor F is one.

The present design for a TWM uses a comb structure that is loaded with active material on both sides and nearly fills all of the available space. There are, however, some regions that are not filled with active material; for example, the space occupied by the isolator, the space between the fingers, and the region near the wall of the waveguide housing that is important for the shaping of the ω-β response of the structure. However, these spaces are considered small, and therefore the volume filling factor is taken to be nearly one. Since the volume filling factor has nearly its maximum value, no further consideration will be made on F_v ; instead, our attention will be directed to F_p .

Since the comb structure has almost no RF fields in the direction of the fingers[2] and the ruby has no susceptibility component along the dc magnetic field when used in the 90° operation,[3] it is clear that the dc magnetic field must be placed parallel to the fingers of the comb for maximum efficiency. Therefore, the only degree of freedom for the c axis of the ruby is a rotation about an axis parallel to the fingers. In the comb structure the RF magnetic field changes with position and with θ, where θ is the phase difference between adjacent fingers of the structure and can have values from 0 to π over the passband of the structure. In addition, the susceptibility of the active material is a function of frequency. Hence F_p varies in a complicated fashion over the passband of the structure. It is our intention to find an expression of F_p as a function of the various dimensions of the structure, the angle between the c axis of the ruby and the z direction of the comb structure, and the phase difference θ.

The isolator incorporated in the comb structure consists of a linear array of ferrite disks arranged to have the same periodicity as the comb. The disks are so shaped that the ferrimagnetic resonance occurs at the same dc field and frequency as that required by the ruby. Two requirements of the isolator are: first, the isolator should provide reverse absorption at least equal to the round-trip gain provided by the ruby and second, the ratio of the reverse to forward attenuation should be high. Our objective in this part of the analysis is to find the effect of the various structural parameters of the comb on the performance of the isolator and to provide a guide toward the optimum design of the isolator. It is found that the structure that provides maximum gain performance is not the same as that for optimum isolator performance. Therefore, the attitude has been to design the structure for maximum gain performance and then to optimize the isolator within the constraints imposed by this structure.

In this paper, the attention will be centered on the comb as the slow-wave structure, a ferrimagnetic material such as YIG as the isolator, and ruby as the active material, where the latter is only considered in the 90° operation; i.e., the dc magnetic field is perpendicular to the crystalline c axis of the ruby. However, the analysis applies equally well to other types of taped structures and active materials, since the results are expressed as a function of θ and not of the frequency. The connection between the frequency and θ is made through the ω-β relations of the particular structure.

The assumption is made that the RF field configuration of the structure is unperturbed by the presence of the spin system of the ruby and the ferrite disks. This is justified, since the spins in ruby are very dilute and the ferrite disks are thin.

Under the following two conditions:[2] (i) no RF fields in the direction of the fingers and (ii) the field on that part of the z axis between the adjacent fingers independent of z (uniform field assumption), the RF magnetic fields in region 2 (see Fig. 1) can be expressed as:

$$H_z = j \sum_{n=-\infty}^{\infty} A_n \cosh \beta_n (D - x) e^{-j\beta_n z} \cos \frac{\pi y}{2h} \tag{2}$$

$$H_x = \sum_{n=-\infty}^{\infty} A_n \sinh \beta_n (D - x) e^{-j\beta_n z} \cos \frac{\pi y}{2h} \tag{3}$$

$$A_n = (-1)^n \sqrt{\frac{\epsilon_1}{\mu}} \frac{lE_0}{L} \frac{\sin \dfrac{\beta_n l}{2}}{\dfrac{\beta_n l}{2}} \frac{1}{\sinh \beta_n D} \tag{4}$$

$$\beta_n = \frac{\theta + 2n\pi}{L} \tag{5}$$

$$E_0 = -2j \frac{V_0}{l} \sin \frac{\theta}{2} \tag{6}$$

where the voltage on the mth finger is taken as $V_m = V_0 e^{-jm\theta}$, and h is the sum of the finger length and the correction due to the fringe capacitance at the finger tips. The rest of the notation is shown in Fig. 1. Equations (2) and (3) also assume $W = D$, which is a good approximation to the TWM with a large-gain, instantaneous-bandwidth product now in use.

Neglecting the copper and dielectric losses of the structure, the microwave signal power in the TWM can be expressed as:

$$P_o = P_i \exp\left[-\frac{L_T}{v_g} \frac{\omega \left\{ \int_m \mathbf{H} \cdot \chi_1'' \cdot \mathbf{H}^* \, dv + \int_i \mathbf{H} \cdot \chi_2'' \cdot \mathbf{H}^* \, dv \right\}}{\int_v \mathbf{H} \cdot \mathbf{H}^* \, dv} \right] \tag{7}$$

where P_i and P_o are the input and output power, respectively, L_T is the total length of the structure, v_g is the signal group velocity and χ_1'' and χ_2'' are the imaginary parts of the susceptibility tensor of the ruby and ferrite disks, respectively. The volume integral in the denominator with the subscript v encloses the whole volume of the structure per period, and the volume integrals in the numerator with the

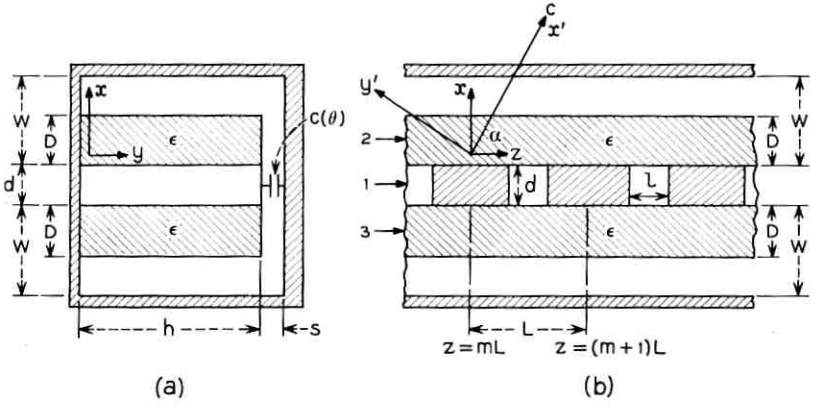


Fig. 1 — Comb structure loaded with ruby. The c axis of ruby makes an angle α with the z direction.

subscripts m and i enclose the volume of ruby or ferrite per period, respectively.

By performing the integration in (7) using the RF fields expressed in (2) and (3), the filling factor and the isolator performance can be found. The filling factor will be derived in Section II and the isolator performance will be discussed in Section III. In both sections the results of theory are compared with the measurements. In Section IV, concluding remarks will be made.

II. FILLING FACTORS

2.1 *Analysis and Discussions*

The susceptibility tensor in (7) is a classical expression, whereas the properties of ruby are given in terms of the spin Hamiltonian; it is therefore necessary to establish a connection between χ_1'' and the quantum mechanical properties of ruby. The subscript one will now be dropped from χ_1''. Throughout this paper MKS units will be used.

Classically, the absorbed power is given by

$$P_{\text{abs}} = \tfrac{1}{2}\mu_0\omega \int_m \mathbf{H}\cdot\chi''\cdot\mathbf{H}^* \, dv \tag{8}$$

where μ_0 is the permeability of free space, ω is the frequency in radians per second, and $\mathbf{H}$ is the RF magnetic field. The quantum mechanical expression for absorbed power is given by:

$$P_{\text{abs}} = \int_m W_{1\text{-}2}(\rho_1 - \rho_2)\hbar\omega \, dv \tag{9}$$

where ρ_1 and ρ_2 are the densities of spins in levels 1 and 2 respectively, $\hbar$ is Planck's constant divided by 2π, and $W_{1\text{-}2}$ is the transition probability between the levels in question. $W_{1\text{-}2}$ is in general a function of the position.

Equating (8) and (9) and dividing through by the total RF energy, $\mu_0\int_v \mathbf{H}\cdot\mathbf{H}^* dv$, one obtains:

$$\frac{\int_m \mathbf{H}\cdot\chi''\cdot\mathbf{H}^* \, dv}{\int_v \mathbf{H}\cdot\mathbf{H}^* \, dv} = \frac{2\int_m W_{1\text{-}2}(\rho_1 - \rho_2)\hbar \, dv}{\mu_0 \int_v \mathbf{H}\cdot\mathbf{H}^* \, dv}. \tag{10}$$

The left-hand side of (10) is exactly the expression occurring in (7).

One can further define:

$$F_v \equiv \frac{\int_m \mathbf{H}\cdot\mathbf{H}^*\,dv}{\int_v \mathbf{H}\cdot\mathbf{H}^*\,dv} \tag{11}$$

$$F_p \equiv \frac{\int_m \mathbf{H}\cdot\chi''\cdot\mathbf{H}^*\,dv}{(\chi_{x'x'}'' + \chi_{y'y'}'')\int_m \mathbf{H}\cdot\mathbf{H}^*\,dv} \tag{12}$$

where F_v and F_p are the volume and the phase filling factors, respectively, and $\chi_{x'x'}''$, $\chi_{y'y'}''$ are the diagonal elements of the susceptibility tensor. $\chi_{z'z'}''$ is not present in (12), since this term is zero for the 90° operation. The prime in the subscripts of χ'' refers to the crystal axes of ruby. The x' axis is taken parallel to the crystalline c axis. (see Fig. 1). In general, both the RF magnetic field and the spin magnetic moment have elliptic polarizations, and when these ellipses have the same sense of rotation, the same principal axes and the same shape (i.e., $H_x^2\chi_{yy}'' = H_y^2\chi_{xx}''$), then F_p as defined above has its maximum value of one. The actual value of F_p for the comb structure will then be a measure of how the fields depart from this ideal case.

Comparing (10), (11) and (12) one obtains:

$$F_p = \frac{2\int_m W_{1\text{-}2}(\rho_1 - \rho_2)\hbar\,dv}{\mu_0(\chi_{x'x'}'' + \chi_{y'y'}'')\int_m \mathbf{H}\cdot\mathbf{H}^*\,dv}. \tag{13}$$

It is now necessary to calculate the detailed expression for $W_{1\text{-}2}$. One starts with an unperturbed Hamiltonian $\mathcal{H}_0$ and the time-dependent Schrodinger equation

$$\mathcal{H}_0\psi = i\hbar(\partial\psi/\partial t) \tag{14}$$

whose solutions are $\psi_\lambda \exp -[(iE_\lambda/\hbar)t]$. For this calculation the ψ_λ eigenfunctions are those obtained from the spin Hamiltonian for ruby.[3] A perturbation term is now added which will represent the coupling of the spins in ruby to the RF magnetic field. If this perturbation is called $\mathcal{H}_1$, then the equation that must be solved is:

$$(\mathcal{H}_0 + \mathcal{H}_1)U = i\hbar(\partial U/\partial t) \tag{15}$$

where

$$\mathcal{H}_1 = \boldsymbol{\mu} \cdot \mathbf{H} \tag{16}$$

$\boldsymbol{\mu}$ is the magnetic moment of the spin, and $\mathbf{H}$ is the RF magnetic field. $\boldsymbol{\mu}$ is given by:

$$\boldsymbol{\mu} = g\beta\mathbf{S} \tag{17}$$

where g is the spectroscopic splitting factor, β is the Bohr electronic magneton, and $\mathbf{S}$ is the spin angular momentum operator.

The RF magnetic fields are given in (2) and (3). For convenience they can be expressed as:

$$H_z = \cos\frac{\pi y}{2h} \sum_n L_n(\theta,x) \left(-\frac{i}{2}\right) [\exp i(\omega t + \beta_n z) - \exp -i(\omega t + \beta_n z)] \tag{18}$$

$$H_x = \cos\frac{\pi y}{2h} \sum_n M_n(\theta,x) \left(\tfrac{1}{2}\right) [\exp i(\omega t + \beta_n z) + \exp -i(\omega t + \beta_n z)] \tag{19}$$

where

$$L_n = A_n \cosh \beta_n(D - x)$$

$$M_n = A_n \sinh \beta_n(D - x).$$

The coordinate system for the ruby and comb structure is shown in Fig. 1. For an angular separation of α between the z and x' axes, (16) becomes:

$$\mathcal{H}_1 = g\beta[(S_{x'} \sin\alpha + S_{y'} \cos\alpha)H_x + (S_{x'} \cos\alpha - S_{y'} \sin\alpha)H_z]. \tag{20}$$

Equation (15) will be solved in terms of an infinite series of unperturbed solutions ψ_λ, i.e.,

$$U = \sum_\lambda C_\lambda(t) \exp - [i(E_\lambda/\hbar)t] \tag{21}$$

where the expansion coefficients depend on the time. $|C_\lambda(t)|^2$ is the probability of finding the spin in the state ψ_λ at the time t. Initially the spin is assumed to be in the state ψ_1, i.e., $|C_1(0)|^2 = 1$, $|C_\lambda(0)|^2 = 0$ for $\lambda > 1$. The perturbation is then turned on and a solution for $|C_2(t)|^2$ is sought. $|C_2(t)|^2$ is the probability of finding the spin in state 2. For weak RF magnetic fields this term is equal to:

$$|C_2(t)|^2 = W_{1\text{-}2}t. \tag{22}$$

The process of obtaining $|C_2(t)|^2$ is a standard one in quantum

mechanics[4] and will not be given in detail here. The result is

$$\mathbf{W}_{1\text{-}2} = \frac{g^2\beta^2\pi}{2\hbar^2}\, g(\omega - \omega_0) \sum_{n,m} (O_n O_m + P_n P_m) \tag{23}$$

where $g(\omega - \omega_0)$ is a normalized line shape function such that:

$$\int_{-\infty}^{\infty} g(\omega - \omega_0)\, d\omega = 1$$

and

$$\begin{aligned} O_n &= (\langle S_{y'}\rangle M_n + \langle S_{x'}\rangle L_n) \cos\alpha \cos\beta_n z \\ &\quad - (\langle S_{x'}\rangle M_n + \langle S_{y'}\rangle L_n) \sin\alpha \sin\beta_n z \\ P_n &= (\langle S_{x'}\rangle M_n + \langle S_{y'}\rangle L_n) \sin\alpha \cos\beta_n z \\ &\quad + (\langle S_{y'}\rangle M_n + \langle S_{x'}\rangle L_n) \cos\alpha \sin\beta_n z \end{aligned}$$

$$\langle S_{x'}\rangle \equiv \int \psi_1 S_{x'} \psi_2^*\, dv, \qquad \langle S_{y'}\rangle \equiv \int \psi_1 S_{y'} \psi_2^*\, dv.$$

Substituting (23) into (12), one obtains:

$$F_p = \frac{g^2\beta^2\pi g(\omega - \omega_0)(\rho_1 - \rho_2) \int \sum_{n,m} (O_n O_m + P_n P_m)\, dv}{\hbar\mu_0(\chi_{x'x'}'' + \chi_{y'y'}'') \int \mathbf{H}\cdot\mathbf{H}^*\, dv}. \tag{24}$$

If we were to examine just one element of the χ'' tensor, say $\chi_{x'x'}''$, and subject it to a simple field, $H_{x'} \sin \omega t$, then by going through the same procedure [using (13)–(22)] one could establish that:

$$\chi_{x'x'}'' = \frac{g^2\beta^2\pi}{\mu_0\hbar}\, g(\omega - \omega_0)(\rho_1 - \rho_2)\, |\langle 1 \mid S_{x'} \mid 2\rangle|^2 \tag{25}$$

and similarly for the other elements in the tensor.

Using (25) to simplify (24), one obtains:

$$F_p = \frac{\int_m \sum (O_n O_m + P_n P_m)\, dv}{(\langle S_{x'}\rangle^2 + \langle S_{y'}\rangle^2) \int_m \mathbf{H}\cdot\mathbf{H}^*\, dv}. \tag{26}$$

The phase filling factor will now be calculated for two cases: (1) a ruby TWM loaded on one side and (2) a ruby TWM loaded on both sides. In both cases it is assumed that the ruby entirely fills the gap between the fingers and the outer waveguide wall and extends to the

open tip of the fingers. In addition, it is assumed that, in the case of a TWM loaded with ruby on one side only, the opposite side is still filled with a material whose dielectric constant equals that of ruby. In practice this assumption is fulfilled very closely with recently developed masers.[5] By performing the integration in (26), one obtains

$$F_p(\tfrac{1}{2}) = \tfrac{1}{2} \pm \frac{\langle S_{x'}\rangle\langle S_{y'}\rangle}{\langle S_{x'}\rangle^2 + \langle S_{y'}\rangle^2} \sum (\tfrac{1}{2}) + \frac{[\langle S_{x'}\rangle^2 - \langle S_{y'}\rangle^2] \cos 2\alpha}{2[\langle S_{x'}\rangle^2 + \langle S_{y'}\rangle^2]} \sum (1) \tag{27}$$

for the structure filled on one half side, and

$$F_p(1) = \tfrac{1}{2} + \frac{[\langle S_{x'}\rangle^2 - \langle S_{y'}\rangle^2] \cos 2\alpha}{2[\langle S_{x'}\rangle^2 + \langle S_{y'}\rangle^2]} \sum (1) \tag{28}$$

for the completely filled structure. $\sum (\frac{1}{2})$ and $\sum (1)$ are defined as follows:

$$\sum (\tfrac{1}{2}) = \frac{\sum \frac{E_0^2}{\beta_n D} \frac{\sin^2\left(\frac{\beta_n l}{2}\right)}{\left(\frac{\beta_n l}{2}\right)^2}}{\sum \frac{E_0^2}{\beta_n D} \frac{\sin^2\left(\frac{\beta_n l}{2}\right)}{\left(\frac{\beta_n l}{2}\right)^2} \frac{\cosh \beta_n D}{\sinh \beta_n D}} \tag{29}$$

$$\sum (1) = \frac{\sum \frac{E_0^2}{\beta_n D} \frac{\sin^2\left(\frac{\beta_n l}{2}\right)}{\left(\frac{\beta_n l}{2}\right)^2} \frac{1}{\sinh^2 \beta_n D}}{\sum \frac{E_0^2}{\beta_n D} \frac{\sin^2\left(\frac{\beta_n l}{2}\right)}{\left(\frac{\beta_n l}{2}\right)^2} \frac{\cosh \beta_n D}{\sinh \beta_n D}}. \tag{30}$$

The $\pm$ sign of the second term in (27) determines the nonreciprocal behavior of the ruby TWM. In the direction of propagation where the elliptically polarized fields of the comb structure are in the same sense as the near circular transition of ruby, the plus sign applies. In the opposite direction, where the fields are not as well matched, the minus sign applies.

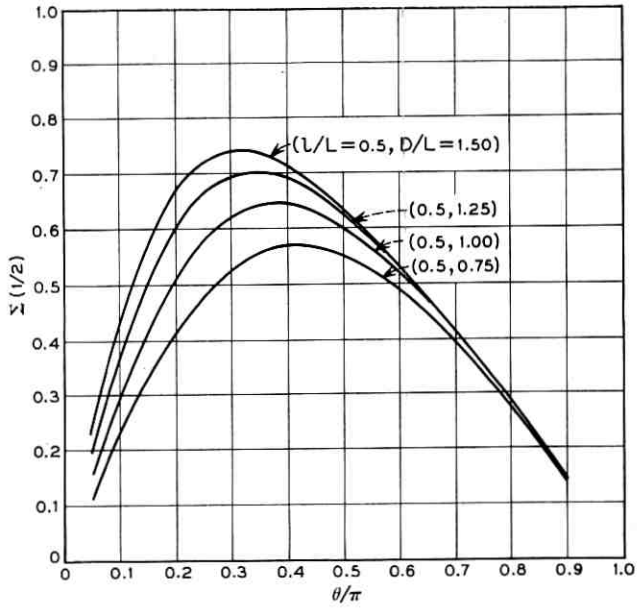


Fig. 2 — $\Sigma(\frac{1}{2})$ vs θ/π for $l/L = 0.5$ and four values of the parameter D/L.

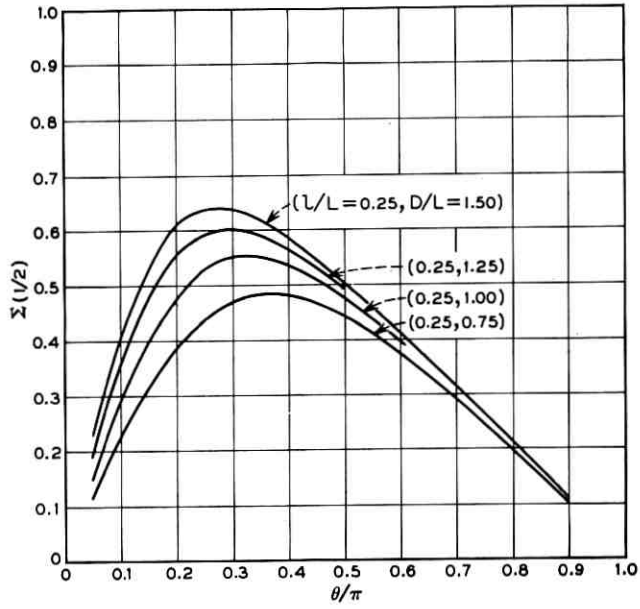


Fig. 3 — $\Sigma(\frac{1}{2})$ vs θ/π for $l/L = 0.25$ and four values of the parameter D/L.

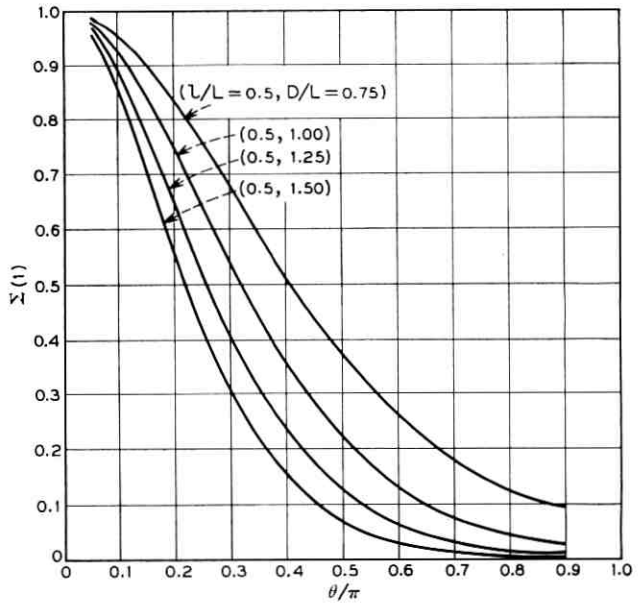


Fig. 4 — $\Sigma(1)$ vs θ/π for $l/L = 0.5$ and four values of the parameter D/L.

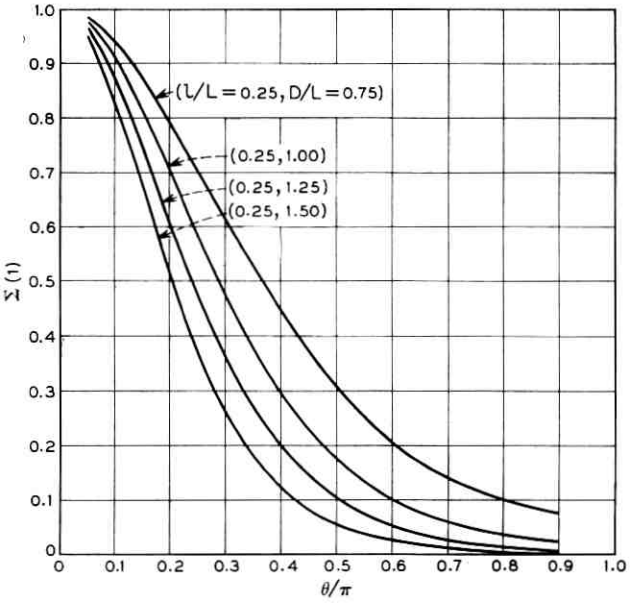


Fig. 5 — $\Sigma(1)$ vs θ/π for $l/L = 0.25$ and four values of the parameter D/L.

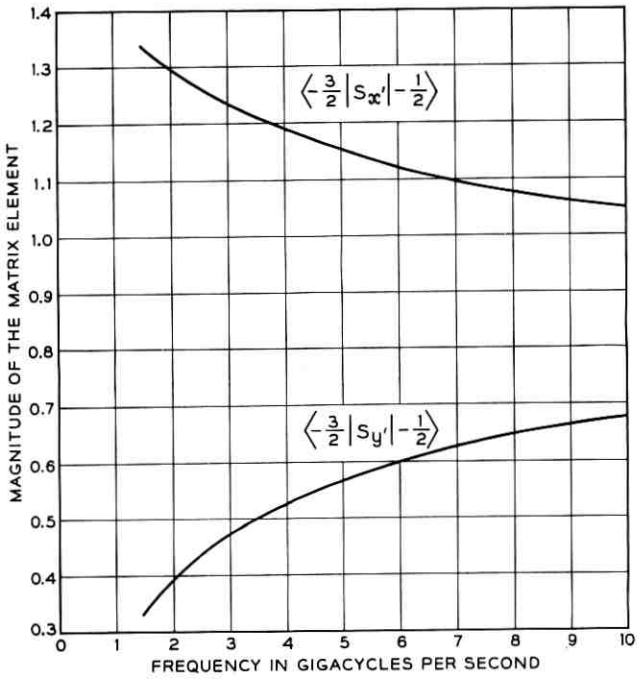


Fig. 6 — A plot of $\langle S_{x'} \rangle$ and $\langle S_{y'} \rangle$ as functions of frequency.

The $\sum$'s are in general functions of the parameters l/L, and D/L and θ, where the terms l, L and D are defined in Fig. 1. $\sum(\frac{1}{2})$ and $\sum(1)$ are plotted as a function of θ in Figs. 2, 3, 4 and 5. Fig. 6 is a plot of the absolute values of the matrix elements, $\langle S_{x'} \rangle$ and $\langle S_{y'} \rangle$, as a function of frequency. The matrix elements are calculated between the levels $-\frac{3}{2} \leftrightarrow -\frac{1}{2}$ for the 90° operation of ruby. With the data presented in Figs. 2 to 6 and (27) and (28), the phase filling factor for a ruby loaded comb structure can be closely estimated.

Equation (28) shows that the phase filling factor for the completely filled structure is $\frac{1}{2}$ if the transition probability of the ruby is circular (i.e., $\langle S_{x'} \rangle = \langle S_{y'} \rangle$) regardless of the angle α, or, if $\alpha = \pi/4$ radians, regardless of the ratio $\langle S_{x'} \rangle / \langle S_{y'} \rangle$. This can be shown to be the consequence of the symmetry of the RF magnetic fields about the plane of the fingers. Equation (28) also shows that the filling factor does not depend on the sign of α, i.e., cos 2α is an even function of α. This result is a consequence of the symmetry of the RF fields about a plane parallel to the x-y plane and centered between the fingers.

The filling factor increases as D/L becomes smaller (see Fig. 3) if $\langle S_x' \rangle$ is greater than $\langle S_y' \rangle$. This is due to the fact that the RF magnetic fields become more linear as D/L gets smaller.

The ratio of the electronic gain of the TWM with both sides loaded with ruby to that with only one side loaded with ruby is shown as a function of θ for different frequencies in Fig. 7. The curves are obtained assuming $D/L = 0.75$, $l/L = 0.5$ and $F_v = 0.5$ for a half loaded and $F_v = 1.0$ for a fully loaded comb. Even at 10 gc, where the transition probability of the ruby is becoming fairly circular, a 30 per cent increase in gain can be obtained by loading the second side with ruby. As an example, let us consider a structure with 30 db electronic gain when the ruby is loaded on one side only. The round-trip gain would be $30 + 0.3(30) = 39$ db, (assuming $\theta = \pi/2$). An isolator of at least 39 db would be required to assure short-circuit stability, and since the reverse-to-forward loss ratio of an isolator is about 20, the forward loss of the isolator would be approximately 2 db. If the structure were now loaded on both sides, the forward gain would increase to 39 db and the round-trip gain would be 78 db. The forward loss of the isolator

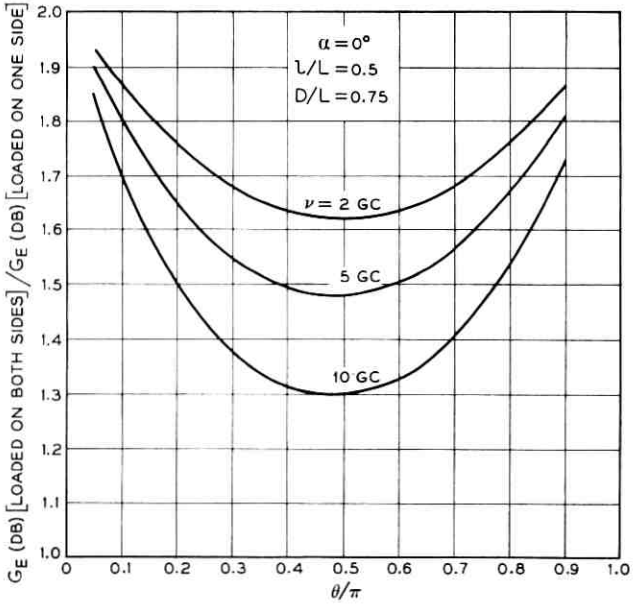


Fig. 7 — The ratio of the electronic gain (in db) of a maser when loaded on both sides to that when loaded on one side. The structures are otherwise identical, with $\alpha = 0°$, $l/L = 0.5$, $D/L = 0.75$.

for this structure would be approximately 4 db. The net gain (excluding the common copper loss) is therefore 28 db for the single-sided loading and 35 db for the double-sided case. It is therefore clear that two-sided loading is more efficient at least up to X-band, although not by a very large margin.

2.2 *Experimental Verification*

Equation (27) was checked experimentally by measuring the paramagnetic absorption in a structure loaded on one side with ruby. The other side of the comb structure was loaded with alumina which has about the same dielectric constant as ruby. The measurements were performed in absorption so that complicating factors such as a microwave pump and isolator could be avoided. Equation (7), of course, defines the absorbed power as well as the gain of the TWM. The passband of the structure was centered at approximately 5.4 gc. Since (27) is a function of the angle between the c axis of the ruby and the z direction of the structure α, two different pieces of ruby were used: one with $\alpha = 67°$ and another with $\alpha = 20°$.

In the evaluation of F_p the volume filling factor, F_v, was taken to be exactly $\frac{1}{2}$. All the measurements that enter into (7) were performed at 4.2°K. v_g was calculated from the ω-β response of the structure. The ω-β plot was obtained by a bridge technique in which the difference in path length of the two arms of the bridge was carefully accounted for and subtracted from the measurement. In this way, the correct ω-β response of the comb structure was assured.

The term $(\rho_1 - \rho_2)$ that enters into the susceptibility formula (25) is difficult to determine exactly, since it depends on the chemical analysis of the ruby. Since the major concern of this experiment is the filling factor and not the accuracy of the chemical analysis, the concentration was left as a parameter to be used as a best fit to the calculated value of F_p. The assumed concentration was then compared with that obtained by chemical analysis.[6]

Figs. 8 and 9 are plots of the experimentally measured and calculated values of F_p. Table I compares the assumed concentration to that obtained by chemical analysis.

Figs. 8 and 9 both show the same type of behavior. The ratio of experimentally measured values to the calculated values decreases as θ increases. Part of this may be explained by the fact that the value of the volume filling factor, F_v, was taken to be exactly $\frac{1}{2}$ for all values of θ. This is a good approximation when $\theta = 0°$, since no RF fields exist between the fingers, but at large values of θ, RF fields do exist between

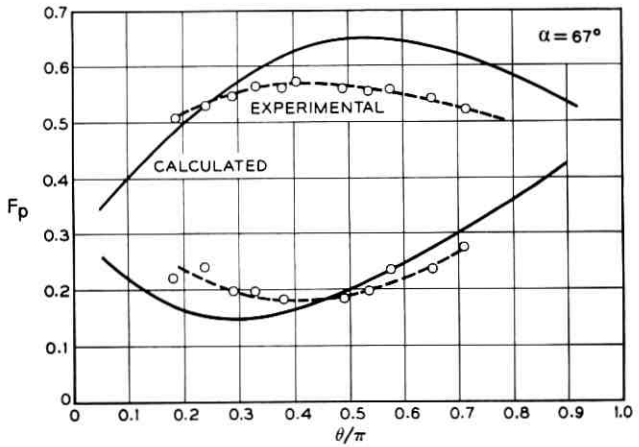


Fig. 8 — A comparison of the theoretical and experimental results for a ruby with $\alpha = 67°$. The solid line is the theoretical curve and the points are the experimental measurements.

fingers where there is no ruby, and therefore $F_v < \frac{1}{2}$. If F_v becomes less than $\frac{1}{2}$, the experimentally determined value of F_p would correspondingly increase and approach the calculated one.

The assumed chromium concentrations compare very well with those determined by chemical analysis. In fact, in the case where $\alpha = 20°$

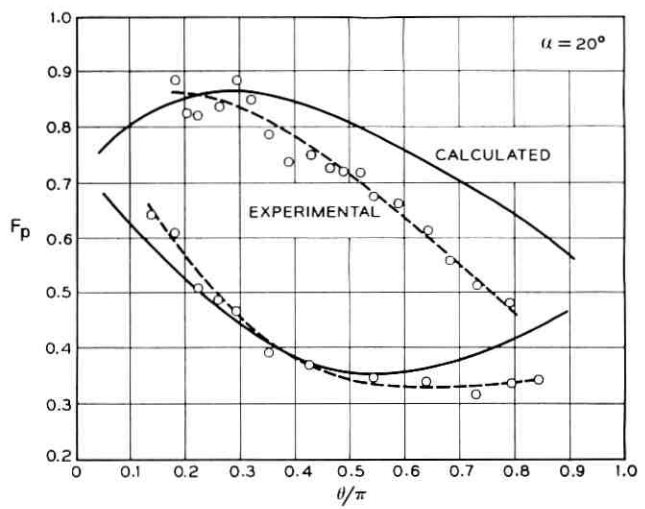


Fig. 9 — A comparison of the theoretical and experimental results for a ruby with $\alpha = 20°$. The solid line is the theoretical curve and the points are the experimental measurements.

TABLE I — COMPARISON OF ASSUMED CONCENTRATION TO CONCENTRATION DETERMINED BY CHEMICAL ANALYSIS

α	Assumed Conc. (Cr/Al Atomic %)	Chemical Analysis (Cr/Al Atomic %)
67°	0.032	0.030
20°	0.028	0.028

the concentrations are the same and therefore Fig. 9 can be considered to be free of all adjustable parameters. However, the concentrations determined by chemical analysis cannot be considered highly accurate on an absolute basis, and therefore this agreement could be fortuitous.

III. ISOLATOR PERFORMANCES

The isolator consists of thin ferrite disks imbedded periodically inside the ruby slab as shown in Fig. 10. The dc magnetic field is applied in the y direction, and it is normal to the plane of the ferrite disks. The ferrite disks are usually thin and have square cross sections. The ratio of the side dimension to the thickness of the square (aspect ratio) is so adjusted that the ferrite resonates at the same dc magnetic field as ruby for a given frequency.

In Section 3.1, the analysis and the result of the machine computation will be presented. The calculation is compared with the measurements in Section 3.2.

3.1 *Analysis and Discussion*

Since the dielectric constant of the ferrite is about the same as the surrounding ruby, the ferrite disks are thin, and the plane of the precessing spins coincides with the plane of the RF magnetic field before the introduction of the ferrite disks, then the disturbance of the RF fields due to the presence of the resonating ferrite disks should be small.

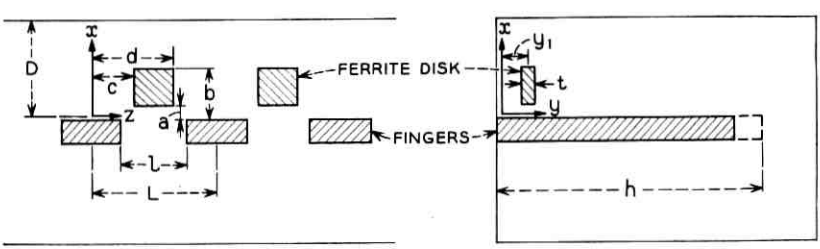


Fig. 10 — Ferrite disk isolators imbedded in the comb structure.

Perturbation theory, which assumes the RF fields inside the ferrite to be the same as those present if the ferrite were replaced by the ruby, should be appropriate for the analysis here.

We shall regard the ferrite square block as an approximation to an ellipsoid. Then the demagnetization factors N_x, N_y and N_z in the x, y and z directions, respectively, can be defined in the usual manner. In this approximation, the demagnetization factor of a square disk will be approximated by the demagnetization factor of a circular disk, i.e., $N_x = N_z$. Even if the disk is not square, $N_x \approx N_z \approx 0$ if the disk is sufficiently thin. In such cases, the imaginary part of the external susceptibility tensor χ_2'' becomes diagonal when the RF fields are expressed in circular components.

From (2) and (3), the positive and negative circular components of the RF magnetic field can be defined as

$$H_+ \equiv \frac{1}{\sqrt{2}}(H_x - jH_z) = \frac{j}{\sqrt{2}} \sum_{n=-\infty}^{\infty} A_n e^{\beta_n(D-x)} e^{-j\beta_n z} \cos \frac{\pi y}{2h} \tag{31}$$

$$H_- \equiv \frac{1}{\sqrt{2}}(H_x + jH_z) = \frac{j}{\sqrt{2}} \sum_{n=-\infty}^{\infty} A_n e^{-\beta_n(D-x)} e^{-j\beta_n z} \cos \frac{\pi y}{2h}. \tag{32}$$

Let χ_+'' and χ_-'' be the diagonal elements of the diagonalized χ_2''. They can be expressed as[7]

$$\chi_+'' = \frac{\omega_m T^{-1}}{(\omega_e - \omega)^2 + T^{-2}} \tag{33}$$

$$\chi_-'' = \frac{\omega_m T^{-1}}{(\omega_e + \omega)^2 + T^{-2}} \tag{34}$$

$$\omega_e = \gamma[H_0 + (N_x - N_y)4\pi M] \tag{35}$$

$$\omega_m = \gamma 4\pi M \tag{36}$$

$$T = \frac{2}{\gamma \Delta H}, \qquad \gamma = 2.8 \text{ mc/oe} \tag{37}$$

where $4\pi M$ is the saturation magnetization, H_0 is the externally applied dc magnetic field and ΔH is the linewidth of the ferrite. In terms of the circular components the absorption due to the ferrite is given by [see (7)]

$$\alpha_+ = \frac{\omega L_T}{v_0} \cdot \frac{\chi_+'' \int_i H_+ H_+^* \, dv + \chi_-'' \int_i H_- H_-^* \, dv}{\int_v (H_+ H_+^* + H_- H_-^*) \, dv} \text{ neper} \tag{38}$$

$$\alpha_- = \frac{\omega L_T}{v_g} \frac{\chi_+'' \int_i H_- H_-^* \, dv + \chi_-'' \int_i H_+ H_+^* \, dv}{\int_v (H_+ H_+^* + H_- H_-^*) \, dv} \text{ neper} \quad (39)$$

where α_+ and α_- are the reverse and the forward absorption of the isolator, respectively. For most of the ferrites used in TWM's, $\chi_-'' \ll \chi_+''$. Therefore, the terms containing χ_-'' can be neglected. Substituting (31) and (32) into (38) and (39) and performing the integrations, one obtains

$$\alpha_\pm = 27.3 \frac{L_T}{v_g} \left(\frac{f}{Q_\pm}\right) \text{db} \quad (40)$$

where f is the frequency and $Q_\pm$ is defined as

$$1/Q_+ = t/h \, [\cos (\pi y_1/2h)]^2 \, \chi_+'' \, A \quad (41)$$

$$1/Q_- = t/h \, [\cos (\pi y_1/2h)]^2 \, \chi_+'' \, B. \quad (42)$$

A and B are ratios of the energy of the RF magnetic field due to the positive and the negative circular components over the area occupied by the ferrite disk to the total RF magnetic field energy contained over the area in the x-z plane per period, respectively. They can be expressed as

$$\left.\begin{matrix} A \\ B \end{matrix}\right\} = \frac{1}{\sum_{n=-\infty}^{\infty} |A_n|^2 \frac{\sinh 2\beta_n D}{2\beta_n D}} \times \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} A_n A_m^* \frac{\sin \frac{(\beta_m - \beta_n)(d-c)}{2} \cos \frac{(\beta_m - \beta_n)(d+c)}{2}}{(\beta_m^2 - \beta_n^2) \, LD} \times [\exp \pm(\beta_m + \beta_n)(D-a)][\pm 1 \mp \exp \mp(\beta_m + \beta_n)(b-a)] \quad (43, 44)$$

where the upper and the lower signs are for A and B respectively. The notations a, b, c and d are defined in Fig. 10. In (43) and (44), the RF field energy contained in the region between the fingers (region 1 in Fig. 1) is neglected, since this is usually small compared to the RF field energy contained in the ruby.

Equations (43) and (44) were computed and the results are shown in Figs. 11–14. Note that A is a quantity proportional to the reverse absorption and A/B is the ratio of the reverse to the forward absorption of the isolators. Large A and A/B over a large range of θ are desirable

for the isolators. Since Figs. 11–14 are shown in terms of θ/π rather than the frequency, the frequency-phase relation (ω-β diagram) of the structure has to be known in order to convert θ into the frequency. All the dimensions involved were normalized by the length of a period of the structure L. In order to provide an easier understanding of the numerical data, we shall assume $L = 0.08$ inch (the commonly used size in our laboratory) and discuss the data with the parameters expressed in inches.

Since the isolation ratio becomes worse when the width of the waveguide housing gets narrower, we shall study the structure with small D first ($D/L = 0.75$). In Fig. 11, A and A/B vs θ/π are shown when

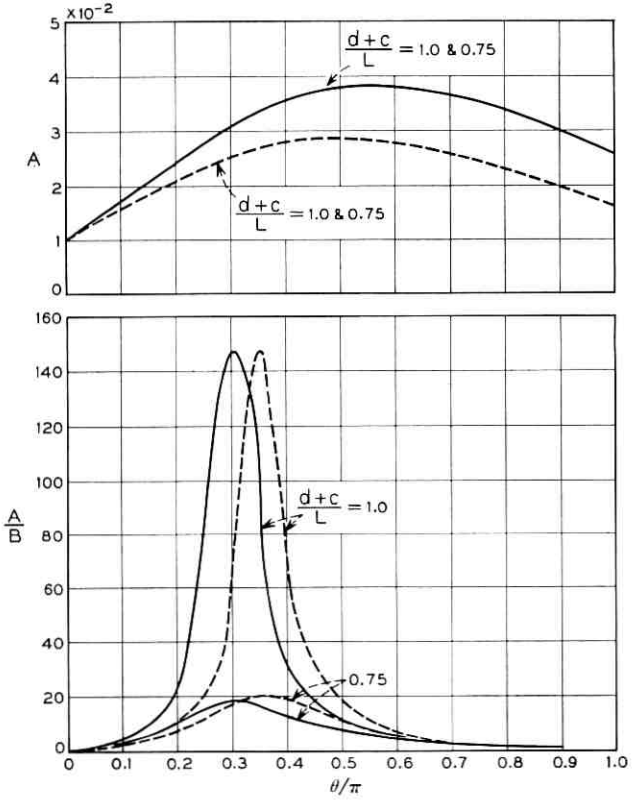


Fig. 11 — A and A/B vs θ/π for the following parameters: $D/L = 0.75$, $l/L = 0.5$; ferrite size, $(b - a)/L = (d - c)/L = 0.125$ (0.01 × 0.01 inch); solid curves for $(D - a)/L = 0.6875$ ($a = 0.005$ inch); broken curves for $(D - a)/L = 0.625$ ($a = 0.01$ inch).

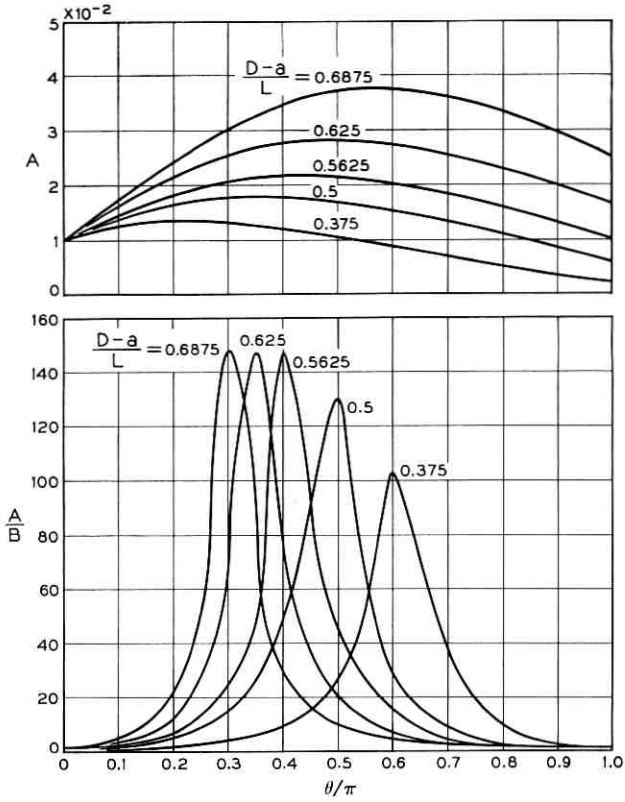


Fig. 12 — A and A/B vs θ/π for the following parameters: $D/L = 0.75$, $l/L = 0.5$; ferrite size, $(b - a)/L = (d - c)/L = 0.125$ (0.01 × 0.01 inch); $(d + c)/L = 1.0$ (ferrite on an axis midway between the adjacent fingers).

ferrite disks of 0.01 × 0.01 inch are placed in different positions of the structure. Solid lines are for the case where one edge of the disk is 0.005 inch away from the surface of the fingers, i.e., $a = 0.005$ inch (see Fig. 10). If the center of the disk is exactly at the position midway between the adjacent fingers $[(d + c)/L = 1.0]$, the isolation ratio A/B can be as large as 140 over a small range of θ. If the disk is shifted in the z direction by 0.01 inch $[(d + c)/L = 0.75]$, the isolation ratio decreases to values less than 20 while A remains almost the same. Next, let us take $a = 0.01$ inch (see the broken lines). At the position $(d + c)/L = 1.0$, the isolation ratio again reaches over 140. Again, shifting the position in the z direction by 0.01 inch decreases the isolation ratio to about 20. One sees readily that the isolator disks have to be positioned carefully

at the position midway between the adjacent fingers. A small displacement from this position deteriorates the isolation ratio rapidly. This has been known experimentally for a long time, and indeed the position of the largest A/B for a given a dimension is found in practice by moving the bar imbedded periodically with the isolator disks until the minimum forward attenuation results. Hence we shall assume the center of the ferrite disk to be on the line midway between the adjacent finger ($z = L/2$) in all of the following discussions.

Next, let us consider how the isolator performs as a is increased further. Fig. 12 shows A and A/B vs θ/π for 0.01 × 0.01 inch isolator disk as a is increased to 0.005, 0.01, 0.015, 0.02 and 0.03 inch [corre-

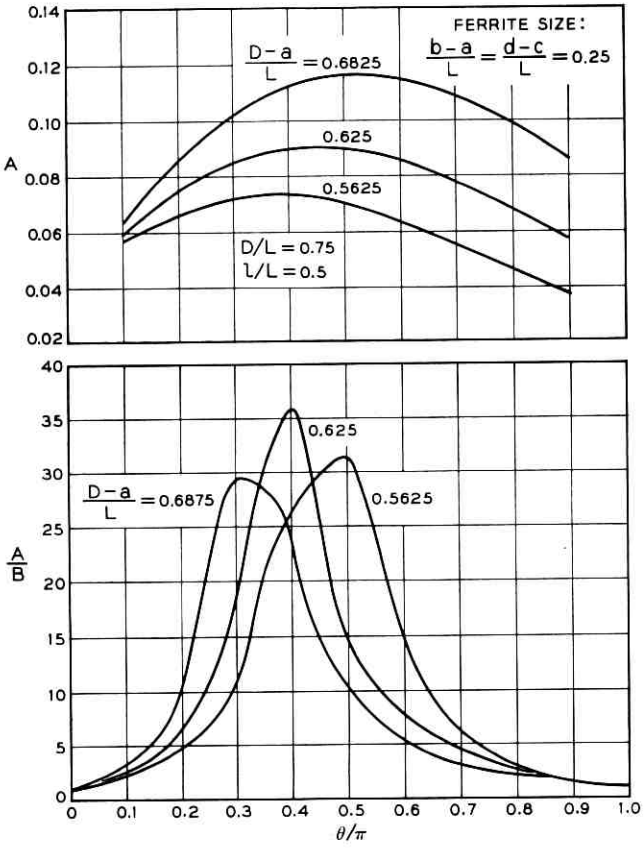


Fig. 13 — A and A/B vs θ/π: $D/L = 0.75$, $l/L = 0.5$; ferrite size, $(b - a)/L = (d - c)/L = 0.25$ (0.02 × 0.02 inch).

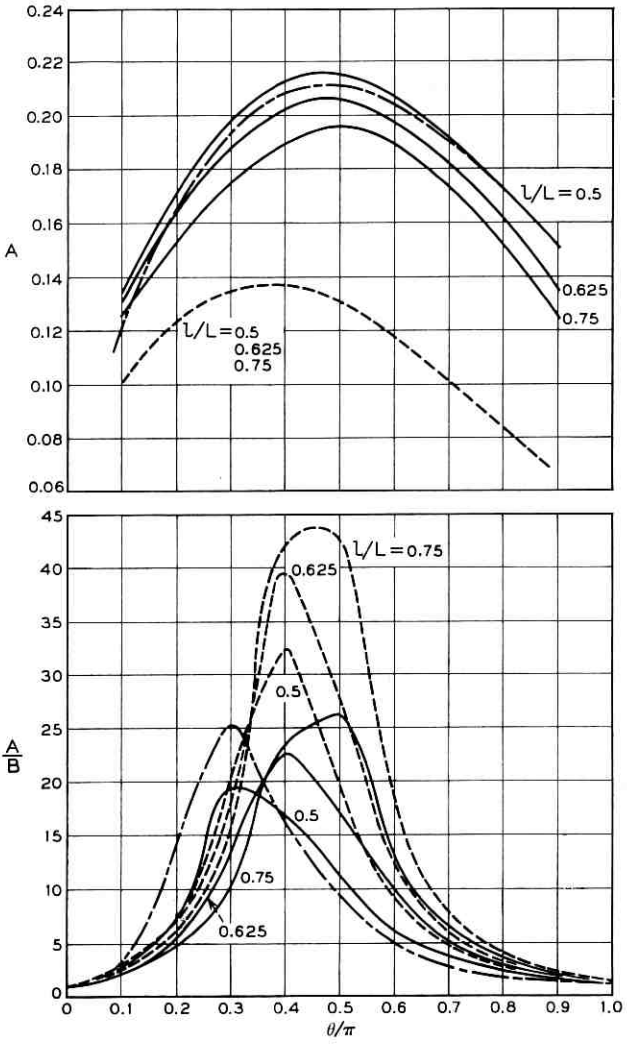


Fig. 14 — A and A/B vs θ/π: ferrite size, $(b - a)/L = (d - c)/L = 0.375$ (0.03 × 0.03 inch); solid line, $D/L = 0.75$, $(D - a)/L = 0.6875$ ($a = 0.005$ inch); dashed line, $D/L = 1.0$, $(D - a)/L = 0.8125$ ($a = 0.015$ inch); irregular dashed line, $D/L = 1.0$, $(D - a)/L = 0.9375$ (a = 0.005 inch); $l/L = 0.5$.

sponds to $(D - a)/L = 0.6875, 0.625, 0.5625, 0.5, 0.375$]. The maximum isolation ratio A/B is about the same from a = 0.005 to 0.015 inch and then it decreases for a larger a. This can be understood by noting that the RF magnetic field becomes linearly polarized at the wall of the waveguide housing, and hence A/B approaches one if the ferrite disk is placed near the wall. The value of θ at which the maximum A/B occurs gradually increases with the increase in a. If this were not so, a large value of A/B would still be maintained for ferrite disks of a larger cross section. Due to the spread of the peaks of A/B the ferrite disks of a larger cross section give small values of A/B. Larger values of a also reduce the reverse absorption (smaller A) since the RF energy is concentrated near the fingers.

In practice, the isolator disks of 0.01 × 0.01 inch cross section are too small to fabricate and also too small to provide enough reverse absorption. In order to see to which direction the size of the ferrite disks should be enlarged, let us consider two rectangular disks of 0.01 × 0.03 inch. One of them has a longer side parallel to the z axis and the other has a shorter side parallel to the z axis. Both disks have $a = 0.005$ inch. Then one can show from Figs. 11 and 12 that for the former case $A = 9 \times 10^{-2}$, $B = 3.2 \times 10^{-3}$ at $\theta = 0.3\pi$ where A/B is maximum, and for the latter case $A = 8.1 \times 10^{-2}$, $B = 1.08 \times 10^{-3}$ at $\theta = 0.35\pi$ where A/B is maximum. Apparently the latter is the better way to place rectangular ferrite disks.

Fig. 13 shows A and A/B vs θ/π for the ferrite disks of 0.02 × 0.02 inch. The edge of the ferrite disks is placed $a = 0.005$, 0.01 and 0.015 inch away from the surface of the fingers. As the ferrite disks are moved away from the surface of the fingers, the isolation ratio improves and then gets worse again. The reverse absorption steadily decreases at the same time. The sharp reduction of the isolation ratio by enlarging the size of the ferrite disks can be understood from the following considerations. At a given θ (or frequency), part of the ferrite disk at the position where B is small contributes little to the forward absorption, while the rest, located at the position where B is large, absorbs much of the forward-wave energy. Hence the forward absorption of the whole ferrite disk is larger. In another words, the peaks of A/B vs θ/π curve occurs at a different θ/π for different portions of the ferrite disk (see Fig. 12) The ferrite disks of this size are still not large enough to provide sufficient isolation for the high electronic gain attainable in present TWM's.

Fig. 14 shows A and the isolation ratio A/B vs θ/π when the size of the ferrite is increased to 0.03 × 0.03 inch (see the solid lines). The constant A increases almost twice from the value for 0.02 inch square

ferrite. Assuming the thickness of the ferrite $t = 0.1\ (d - c)$ (aspect ratio of 10), the reverse absorption of the isolator increases about 2×1.5 by increasing the size of the disk from 0.02 inch square to 0.03 inch square. However, the isolation ratio reduces from the maximum of 29.5 to 19.5 (the forward absorption increases 4.5 times).

It is clear by now that the isolation ratio of the isolator improves rapidly as the size of the ferrite disks gets smaller. On the other hand, the size of the ferrite has to be large to provide enough reverse absorption. Thus in order to improve the performance of the isolator, one faces contradicting requirements. One approach to circumvent this difficulty is to look for a ferrite of much larger susceptibility. Another solution is to look for a proper dimension of the structure to provide a larger area of circular polarization of the RF magnetic field. In Fig. 14 A and A/B vs θ/π are shown with a larger spacing between the fingers (ferrite size is 0.03 inch square). It is increased from $l = 0.04$ to 0.05 and to 0.06 inch while L is kept at 0.08 inch. The isolation ratio improves gradually.

Next consider the effect on the isolator performance when D/L is increased from 0.75 to 1.0. This will in general increase the group velocity.[2] From (7) and (40), the net db gain of TWM G_n is

$$G_n \propto \frac{1}{v_g}\left(F \mid \chi'' \mid - \frac{1}{Q_-}\right) \tag{45}$$

where F is the filling factor and $\mid \chi'' \mid$ is the susceptibility of the inverted spin system of the ruby. For the TWM to be short-circuit stable,

$$1/Q_+ > 2F \mid \chi'' \mid . \tag{46}$$

The isolator incorporated in the TWM should satisfy the condition (46) first, and then comes the consideration of how to reduce $1/Q_-$. The change in v_g will change the net gain and the reverse and the forward isolator absorption but not the stability condition (46). Hence, we shall confine our discussion here to the effect of $Q_\pm$ as D/L is increased. Suppose the condition (46) is satisfied with the isolator of $0.03 \times 0.03 \times 0.003$ inch, $a = 0.005$ inch, $l/L = 0.5$ and $D/L = 0.75$. If D/L is increased to 1.0, one sees from the irregular dashed curves in Fig. 14 that A remains almost the same, while the ruby filling factor reduces (see Fig. 4). Thus the TWM is still stable. However, the isolation ratio A/B increases from 19.5 to 25 (i.e., $1/Q_-$ decreases). Therefore the isolator performs better with larger D/L. Unfortunately, D/L is usually made small to provide a small group velocity near the center of the passband instead of being sized for the consideration of the

optimum isolator performance. Even if the forward absorption of the isolator increases with a smaller D/L, the net gain of the TWM may increase.

3.2 *Comparison of the Calculation with Measurements*

Isolators of 0.02 × 0.02 × 0.002 and 0.03 × 0.03 × 0.003 inch polycrystalline YIG disks were made and imbedded inside the comb structure of the dimensions shown in Fig. 15. The forward and the

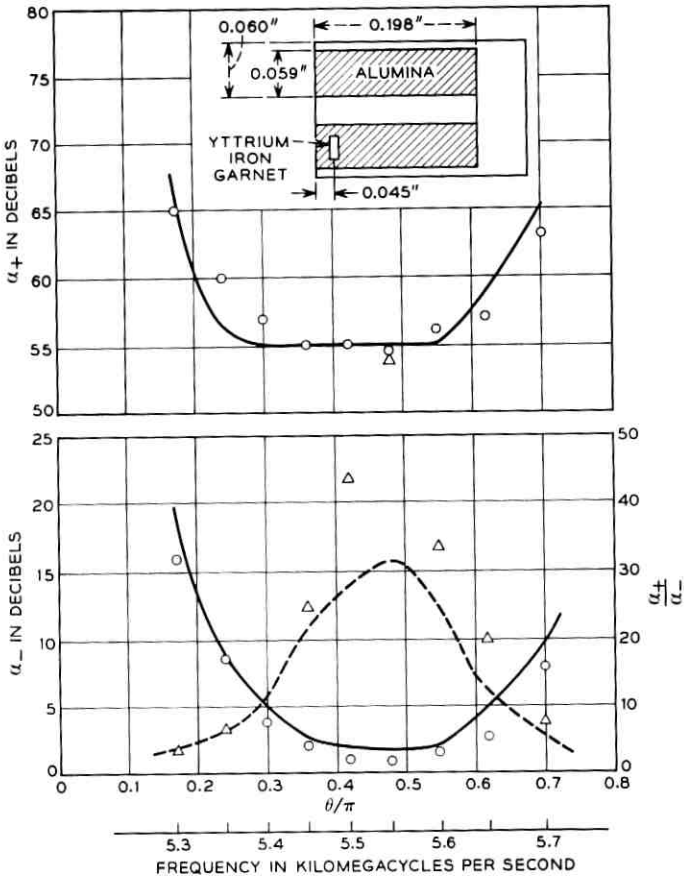


Fig. 15 — Measured and calculated forward and reverse losses. Ferrite disks 0.02 × 0.02 × 0.002 inch (71 pieces). L = 0.08 inch, D = 0.06 inch, l = 0.04 inch, a = 0.016 inch, ΔH = 223 oe, $4\pi M$ = 1750 oe. Solid lines are the calculated $\alpha_\pm$ and the circled points are measured. Broken line is the calculated α_+/α_- and the triangles are measured points.

reverse absorption were measured and shown as discrete points in Figs. 15, 16 and 17. The polycrystalline YIG has $4\pi M = 1750$ oe at room temperature. Its linewidth ΔH was determined by measuring the dc magnetic fields at which the reverse isolator loss in db is one-half of the maximum. It is more than twice the usual linewidth. This is presumably due to the variation of the aspect ratio among the YIG disks and also partly due to the shape of the ferrite disks being square instead of ellipsoidal. The finger length of the structure was 0.198 inch but in the calculation of $\alpha_{\pm}$ it was assumed that $h = 0.22$ inch to correct for the fringe capacitance at the finger tips.

The ω-β relation of the comb structure was measured by a phase bridge technique described in the previous section. This makes it possible to relate ω to θ and also gives the group velocity.

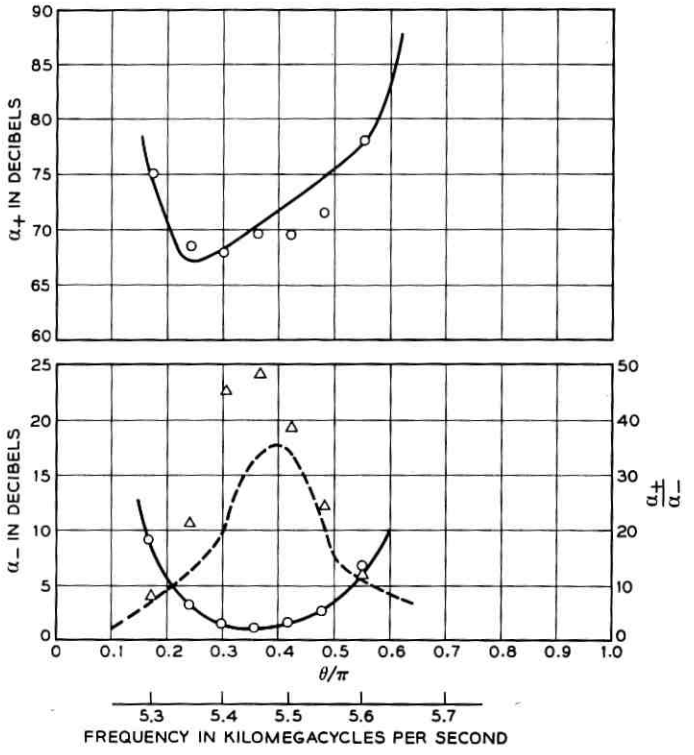


Fig. 16 — Measured and calculated forward and reverse absorption. Ferrite size: 0.02 × 0.02 × 0.002 inch (71 pieces). $L = 0.08$ inch, $D = 0.06$ inch, $l = 0.04$ inch, $a = 0.01$ inch, $\Delta H = 223$ oe, $4\pi M = 1750$ oe. Solid lines are the calculated $\alpha_{\pm}$ and the circled points are measured. Broken line is the calculated α_+/α_- and the triangles are measured points.

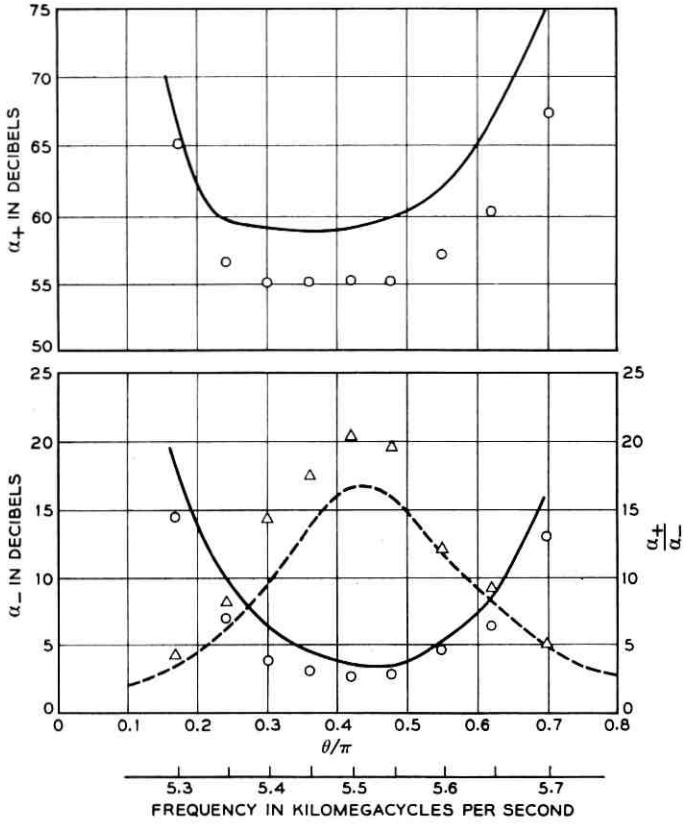


Fig. 17 — Measured and calculated forward and reverse absorption. Ferrite disks = 0.03 × 0.03 × 0.002 inch (35 pieces). L = 0.08 inch, D = 0.06 inch, l = 0.04 inch, a = 0.009 inch, ΔH = 251 oe, $4\pi M$ = 1750 oe. Solid lines are the calculated $\alpha_\pm$ and the circled points are measured. Broken line is the calculated α_+/α_- and the triangles are measured points.

Figs. 15 and 16 show the calculated and the measured data with 0.02 × 0.02 × 0.002 inch ferrite disks. Fig. 15 is for the case a = 0.016 inch and Fig. 16 is for a = 0.01 inch. The curves of Fig. 13 show that, by decreasing a from 0.016 to 0.01 inch, the reverse absorption increases more strongly at large θ than at small θ, and that the maximum value of α_+/α_- occurs at a smaller θ. These theoretical predictions are confirmed by the measurements, as can be observed from the data in Figs. 15 and 16.

The measured and the calculated $\alpha_\pm$ and α_+/α_- for the ferrite size of 0.03 × 0.03 × 0.002 inch are shown in Fig. 17. α_+ for the 71 pieces of ferrite disks was too large to be measured in our equipment. There-

fore every other ferrite disk was removed to reduce α_+ by half. Comparing the measured data shown in Figs. 16 and 17, one notices that, by increasing the size of the ferrite disk from 0.02 to 0.03 inch square, the reverse absorption α_+ increases about 1.6 times while the maximum isolation ratio α_+/α_- decreases by a factor of 2.4. These are approximately the values expected from the calculation.

The measurements agree well with the calculation in general. It should be mentioned that in calculating α_+, use was made of parameters which can be measured only within an accuracy of 10 per cent.

IV. CONCLUSIONS

It has been shown that the filling factor increases as the c axis of the ruby approaches the z axis of the structure for the usual 90° operation in ruby. The increase is more significant near the $\theta = 0$ end of the passband of the comb structure than near the $\theta = \pi$ end. For the forward-wave structure ($df/d\theta > 0$), this corresponds to larger increase of the filling factor near the lower cutoff frequency than near the upper cutoff frequency. The increase in F_p as the c axis of the ruby becomes parallel to the z axis of the structure is more pronounced at lower frequencies, since the transition probability of the spin is more linear at those frequencies.

The ratio of the reverse-to-forward absorption of the isolator incorporated in the TWM can be over a hundred if the size of the ferrite disks is very small compared to the width of the waveguide housing, $2D + d$, and the size of the fingers. At present, the isolation ratio is far from the best due to the necessity of using large-size ferrite disks to provide enough reverse absorption. The size of the ferrite disks can be kept small and yet provide a large reverse absorption if the imaginary part of the susceptibility of the ferrite, χ'', can be increased. Since at resonance $\chi'' = 4\pi M/\Delta H$, one should look for a material with a large $4\pi M$ and a smaller ΔH. The saturation magnetization, however, cannot be made arbitrarily large, since the dc magnetic field, which is fixed by the ruby resonance, must be greater than $4\pi M$ in order to saturate the isolator disks. ΔH of the ferrite should be about 20 oe, which is approximately the linewidth of ruby. ΔH of polycrystalline YIG at the temperature of 4.2°K is over 200 oe. This can be reduced to a certain extent by minimizing the scattering in the aspect ratios of the ferrite disks and the use of round disks instead of square disks, since the internal field of a round disk is more uniform than that of a square one. A single-crystal YIG with ΔH enlarged to about 20 oe

has good possibility as an improved isolator material for TWM application.

The analysis also shows that the isolation ratio gradually improves and then gets worse, while the reverse absorption steadily decreases, as the ferrite disks are moved away from the surface of the fingers toward the waveguide wall. Also, the optimum isolation ratio gradually shifts to a higher value of θ.

In order to provide enough reverse absorption, more than one isolator deck is usually stacked on top of another. It may seem as if the forward insertion loss of the isolator can be kept small over a wider range of θ if a composite isolator with different a's is used instead of one with the same a. However, by adding the curves of different a's in Fig. 12 one will find that this is not so.

The analysis and the curves provided in this paper enable one to estimate the increase in electronic gain obtained by a proper orientation of the c axis of the ruby and by the use of ruby slabs on both sides of the comb. The size of the ferrite disks necessary to provide sufficient isolation and the dependence of the isolator performance on the position of the ferrite disks are also discussed. Together with the Refs. 1 and 2, this paper constitutes part of an effort to reduce the experimental work involved in developing traveling-wave masers.

V. ACKNOWLEDGMENTS

We wish to thank W. J. C. Grant and M. Berry for the programming of the computation, and R. C. Petersen and R. P. Morris for their expert help in the measurements.

REFERENCES

1. Harris, S., DeGrasse, R. W., and Schulz-DuBois, E. O., Cutoff Frequencies of the Dielectrically Loaded Comb Structure as Used in Traveling-Wave Masers, B.S.T.J., **43**, Jan., 1964, p. 437.
2. Chen, F. S., B.S.T.J., this issue, p. 1035.
3. Schulz-DuBois, E. O., Paramagnetic Spectra of Substituted Sapphires — Part I: Ruby, B.S.T.J., **38**, Jan., 1959, p. 271.
4. Schiff, L. I., *Quantum Mechanics*, McGraw-Hill, New York, 1955.
5. Tabor, W. J., A 100-Mc Broadband Ruby Traveling-Wave Maser at 5 Gc, Proc. IEEE, **51**, August, 1963, p. 1143.
6. Dodd, D. M., Wood, D. L., and Barns, R. L., Spectrophotometric Determination of Chromium Concentration in Ruby, to be published.
7. Suhl, H., and Walker, L. R., Topics in Guided-Wave Propagation Through Gyromagnetic Media, Part I, B.S.T.J., **33**, May, 1954, pp. 579–659.

The Comb-Type Slow-Wave Structure For TWM Applications*

By F. S. CHEN

(Manuscript received December 5, 1963)

The space harmonic analysis of the dielectrically loaded comb structure as used in traveling-wave masers (TWM) is presented. The frequency-phase characteristics (the ω-β diagrams) are computed by regarding each finger of the comb structure as a capacitive loaded transmission line. The impedance of the line is based on the space harmonic analysis. Computed data are found to be in agreement with experimental results and, in particular, it is confirmed that the ω-β relation depends very critically on certain dimensions of the dielectric loading. The results of the analysis are used to derive prescriptions for the design of dielectrically loaded TWM comb structures, especially of structures with low group velocity which are suitable to provide simultaneously large gain and large instantaneous bandwidth.

I. INTRODUCTION

The comb-type structure has been used successfully as a slow-wave structure for traveling-wave masers (TWM).[1] In a TWM the small-signal gain in db is inversely proportional to the group velocity. Generally the gain obtainable from present maser materials is small. Therefore, a great deal of effort in developing a TWM is concerned with deriving a comb structure design with small group velocity at the frequency of interest. Both the group velocity and the passband of the comb structure can be found from the ω-β diagram. The shape of the ω-β diagram of the comb structure loaded with "masing" crystal depends very critically on the various dimensions of the structure as well as of the crystal. In this paper, an analysis of the comb structure will be presented. It should serve as a guide for a reasonably accurate determination of the dimensions of the comb structure, of the active maser material and of other dielectrics which give rise to a required ω-β diagram. Some additional

* This work was supported in part by the U. S. Army Signal Corps under Contract No. DA 36-039-SC-89169.

experimentation may be needed in practice for small corrections of the resulting ω-β characteristic.

Various tape structures with the tapes perpendicular to the direction of signal propagation have been proposed and analyzed[2,3,4,5] in the past. The space harmonic analysis originally introduced by Fletcher[2] was used to obtain the ω-β diagrams of these tape structures. It assumes no RF field components in the direction of the tape (TEM wave approximation); that is, the tape can be regarded as a transmission line supporting the TEM wave in the direction transverse to the direction of signal propagation. An impedance matching condition for the TEM lines in the transverse plane can be derived (transverse resonance). The resulting equation implicitly contains the ω-β relation. The authors mentioned treated only the case where the structure is immersed in a uniform dielectric, that is, mostly vacuum. For the TWM application, the structure is alway partially loaded with dielectric and the TEM approximation no longer holds. However, when the structure is nearly filled with dielectric, as in TWM's for a large gain and simultaneously large instantaneous bandwidths,[6] one finds that the TEM approximation can also be successfully used to calculate the ω-β diagram of the dielectric-loaded tape structures. An "effective dielectric constant" is then defined to take into account the fact that the structure is only partially loaded with dielectric. This approach was used in earlier calculations of the upper and lower cutoff frequencies of the comb structure by Harris, DeGrasse and Schulz-DuBois.[7] The analysis to be presented here extends their work to cover the entire ω-β diagram.

The ω-β diagram of the comb structure[1] and the "Karp structure"[8] for TWM applications had also been discussed previously, using the equivalent circuit method. However, the field analysis to be presented here gives a more detailed understanding of the structure. For instance, the filling factor of the active crystal and the performance of the isolator embedded in the structure are readily obtained by this analysis.

In the next three sections, the ω-β diagram of the comb structure, the impedance of the tapes (or fingers) and the effective dielectric constant will be derived. This is followed by a more detailed discussion of various properties of the comb structure, including the techniques available for reducing the bandwidth of the passband, higher-order transmission bands of the structure and practical design considerations. The calculations are usually compared with experiments, and they are found to be in good agreement.

The impedance of the finger and the effective dielectric constant are defined as a function of θ, the phase angle between adjacent fingers. The phase angle θ, or equivalently the phase propagation constant β, are

related to the frequency ω by the ω-β relation which is obtained from the transverse resonance matching condition. Only the ruby-loaded comb structure is treated here up to and including numerical details. It is easily possible, however, to extend the same type of analysis to other types of tape structures. This is particularly easy since both the impedance and the effective dielectric constant are evaluated here as functions of θ and not of ω.

II. THE ω-β DIAGRAM

Let us consider first a comb structure without dielectric loading (empty comb). The open end of the fingers has fringing electric fields terminating at the surrounding conductors. The effect of these fields can be expressed by a capacitance C (Fig. 1) which in general is a function of θ, $C = C(\theta)$, where θ is the phase angle between the adjacent fingers and takes values from zero to π radians over the passband. It will be assumed that the RF electric and the magnetic fields vanish in the direction of fingers (y direction) except at the finger tip. Then the finger can be regarded as a TEM transmission line with a characteristic impedance K and supporting a wave propagation in the y direction with velocity c, the velocity of light in free space. K is a function of θ, $K = K(\theta)$. At the finger tip ($y = h$), the impedance looking in $+y$ and $-y$ directions must have the same magnitude and the opposite sign; therefore there exists a matching condition

$$\frac{1}{\omega C(\theta)} = K(\theta) \tan \frac{\omega h}{c}. \tag{1}$$

According to this equation, the grounded finger presents an inductive reactance at the finger tip. The length of the finger h is therefore limited by

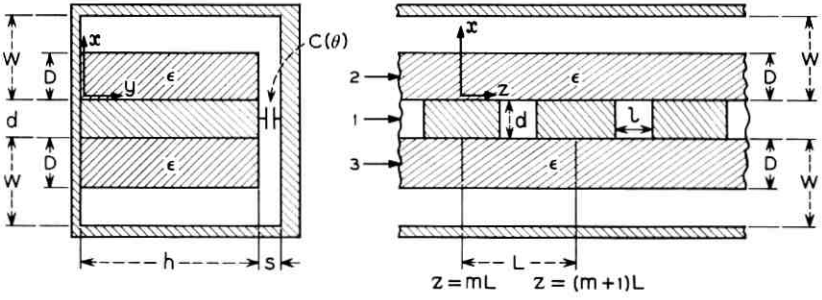


Fig. 1 — Cross sections of the comb structure.

$$(m - 1)\pi < \omega h/c < (m - \tfrac{1}{2})\pi, \qquad m = 1, 2, 3, \cdots . \tag{2}$$

For a given θ, one can find the frequency from (1), provided the functions $K(\theta)$ and $C(\theta)$ are known, and in this way the ω-β diagram (or ω-θ diagram) may be derived. For $m = 1$, the finger acts essentially as a quarter-wave resonator and (1) gives the lowest passband for a given geometry of the comb structure. For $m = 2$, the finger behaves as a three-quarter wave resonator and the next higher passband appears. Thus the comb structure provides a series of passbands separated by stop bands.

When the comb structure is partially loaded with dielectric as in Fig. 1, there appear components of RF fields in the y direction, and the finger no longer behaves as a TEM line. However, when the structure is almost completely loaded with the dielectric, the fields are again approximately TEM waves in the y direction. We will adopt this TEM approximation for the loaded structure and modify the impedance K and the propagation constant ω/c of (1) by a factor $\sqrt{\epsilon(\theta)}$. $\epsilon(\theta)$ is called the effective dielectric constant and it will be defined more rigorously in Section IV. Then the dispersion equation for the loaded comb can be expressed as

$$\frac{1}{\omega C(\theta)} = \frac{K(\theta)}{\sqrt{\epsilon(\theta)}} \tan \frac{\omega h \sqrt{\epsilon(\theta)}}{c} \tag{3}$$

for the case where the structure is loaded with dielectric of uniform thickness from $y = 0$ to $y = h$ as in Fig. 1. For simplicity $K(\theta)$, $C(\theta)$ and $\epsilon(\theta)$ will be abbreviated as K, C and ϵ from here on. The TEM approximation offers a further advantage, since it enables one to analyze the various "finger tip loadings" in a simple way.

It turns out that, in practical realizations of the comb structure, the fringe capacity C is always small, so that $\omega h\sqrt{\epsilon}/c$ is close to $(m - \frac{1}{2})\pi$ with $m = 1, 2, 3, \cdots$. Hence (3) can be simplified by defining a quantity Δh through

$$\omega \sqrt{\epsilon}(h + \Delta h)/c = (m - \tfrac{1}{2})\pi. \tag{4}$$

Substituting (4) into (3) and using the fact that $\Delta h \ll h$, one obtains

$$\Delta h \doteq (KCc/\epsilon) \tag{5}$$

where it should be noted that Δh is also a function of θ. Equation (4) becomes

$$(2m - 1)\frac{\lambda}{4} \doteq \sqrt{\epsilon}\left(h + \frac{KCc}{\epsilon}\right) \tag{6}$$

where $\lambda = c/f$ is the free-space wavelength. For the empty comb, the ω-β diagram can be obtained from (6) by letting $\epsilon = 1$. Equation (6) states that, in the passbands, the finger length h plus the correction due to the fringe capacitance at the finger tip should be $(2m - 1)\lambda/4$ when measured with the scale $\sqrt{\epsilon}$.

III. IMPEDANCE OF THE FINGER, $K(\theta)$

The admittance of a single finger as a TEM line is a somewhat abstractly defined quantity. One considers transmission lines having the cross section shown in Fig. 1(b), but without dielectric and infinitely long in the $\pm y$ directions, or alternatively suitably terminated. On this set of identical transmission lines one considers waves of equal amplitude traveling, for example, in the $+y$ direction and phased by θ between adjacent lines in the $+z$ direction. Under these conditions, the admittance of a single finger is defined as the ratio of current to voltage on a finger and is a function of θ.

Consider a comb structure as in Fig. 1, but without the dielectric and with the structure divided into three regions as shown. Then the current on a finger is the sum of the current on the surface of the finger in region 1 (between the fingers) and the current on the surface of the finger in regions 2 and 3. Thus the admittance of the finger is the sum of two admittances:

$$Y(\theta) = Y_1(\theta) + Y_2(\theta) \tag{7}$$

where $Y_1(\theta)$ is the admittance due to the current on the surface of the finger in the region 1 and $Y_2(\theta)$ is the admittance due to the current on the surface of the finger in the regions 2 and 3.

The current on the finger can be found by a line integral of the RF magnetic fields, which in turn can be found by matching boundary conditions in the x-z plane.

The potential on the mth finger may be written as

$$V_m = Ve^{-jm\theta} \tag{8}$$

where θ is again the phase angle between the adjacent fingers and may vary from 0 to π. Then, as will be shown in Appendix A, the potential along the z axis between the mth and the $(m + 1)$th finger can be expressed as

$$V(\theta,z) = Ve^{-j(m+\frac{1}{2})\theta}\left[g(z)\cos\frac{\theta}{2} - jf(z)\sin\frac{\theta}{2}\right] \tag{9}$$

where $g(z)$ and $f(z)$ specify the symmetric and the antisymmetric part of the potential distribution respectively. $g(z)$ is also the potential distribution when $\theta = 0$, and $f(z)$ is also the potential distribution when $\theta = \pi$. It should be added that the representation of the θ-dependent potential distribution by a combination of θ-independent symmetric and antisymmetric functions, $g(z)$ and $f(z)$, is at best a good approximation or, to use a more appropriate term, a guess. The representation is strictly correct only for $\theta = 0$ and $\theta = \pi$. However, it is bound to be a good approximation near $\theta = 0$ and near $\theta = \pi$ where one of the functions dominates. The calculations show that the detailed shape of the potential distribution assumed is of little influence in the midband region near $\theta = \pi/2$ and in fact up to $\theta = \pi$. It is in this sense that the expression into a symmetric and an antisymmetric part, which is a powerful tool in other electromagnetic problems, is justified here.

The RF electric field on the z axis becomes

$$\begin{aligned} E_z &= 0, \qquad mL - \frac{L-l}{2} < z < mL + \frac{L-l}{2} \\ &\qquad\qquad \text{(on the fingers)} \\ E_z &= -\frac{\partial V}{\partial z}, \qquad (m+\tfrac{1}{2})L - \frac{l}{2} < z < (m+\tfrac{1}{2})L + \frac{l}{2} \\ &\qquad\qquad \text{(between fingers).} \end{aligned} \tag{10}$$

The E_z field in region 2, E_{z2}, may be expressed by a space harmonic or generalized Fourier sum

$$E_{z2} = \sum_{n=-\infty}^{\infty} F_n \sinh \beta_n(W - x) e^{-j\beta_n z} e^{jky} \tag{11}$$

where $\beta_n L = \theta + 2n\pi$, W is the height of the waveguide as shown in Fig. 1, and F_n is the amplitude of the nth space harmonic component. Each term of the sum is the electric field of a TEM solution to Maxwell's equations, and the z dependence, $\exp(-j\beta_n z)$, assumes periodicity of the resulting field pattern from finger to finger as required by (8). The as yet unknown amplitude coefficients F_n are determined by letting $x = 0$ in (11) and equating the resulting expression with the field on the z axis (10). Thus,

$$F_n = -\frac{V e^{-j(m+\frac{1}{2})\theta}}{L \sinh \beta_n W} \int_{(m+\frac{1}{2})L-(l/2)}^{(m+\frac{1}{2})L+(l/2)} \left[g' \cos\frac{\theta}{2} - jf' \sin\frac{\theta}{2} \right] e^{j\beta_n z} dz \tag{12}$$

where $g' = dg/dz$ and $f' = df/dz$. The current on the surface of the mth

finger facing regions 2 and 3 can be found from the line integral of the z component of the RF magnetic field there, and finally Y_2 becomes

$$Y_2(\theta) = 2Y_0 \frac{L-l}{L} \sum_{n=-\infty}^{\infty} (-1)^n \frac{\sin \frac{\beta_n(L-l)}{2}}{\frac{\beta_n(L-l)}{2}} \coth \beta_n W \times \left[\cos \frac{\theta}{2} \int_{(L-l)2}^{(L+l)/2} g' \sin \beta_n z \, dz + \sin \frac{\theta}{2} \int_{(L-l)/2}^{(L+l)/2} f' \cos \beta_n z \, dz \right] \tag{13}$$

where $Y_0 = \frac{1}{377}$ mho.

In the case of constant-field approximation, as has been assumed by Fletcher,[2]

$$\begin{aligned} g &= 1 \\ f &= (2z/l) - (L/l). \end{aligned} \tag{14}$$

The integration in (13) can be readily performed. For the case where $L - l = L/2$ and $d = l$ (square finger cross section),

$$\frac{Y_2(\theta)}{Y_0} = 2 \sin \frac{\theta}{2} \sum_{n=-\infty}^{\infty} (-1)^n \left(\frac{\sin \frac{\beta_n L}{4}}{\frac{\beta_n L}{4}} \right)^2 \coth \beta_n W. \tag{15}$$

$Y_1(\theta)$ becomes, in the same constant-field approximation

$$\frac{Y_1(\theta)}{Y_0} = 4 \frac{d}{L} \sin^2 \frac{\theta}{2}. \tag{16}$$

The impedance of a finger $K(\theta) = (Y_1(\theta) + Y_2(\theta))^{-1}$ in the constant-field approximation is plotted in Fig. 2 in dashed lines as a function of θ/π for $W/L = 1.25$ and 0.75.

The constant-field assumption does not take into account the singularity in the field at the corner of the finger. It will be shown later that the calculated ω-β diagram using the impedance thus obtained disagrees severely with the measured one for $\theta < 0.5\pi$. Harris et al.[7] used the constant-field approximation for region 1 only and assumed the field produced by an infinitely thin tape for regions 2 and 3. This will be referred to as the thin-tape approximation. The calculation used conformal mapping and it is applicable directly only for $\theta = 0$, $\pi/2$ and π. The thin-tape approximation assumes a 180° singularity at the corner of the finger and thus exaggerates the actual 90° singularity there. A further difficulty in

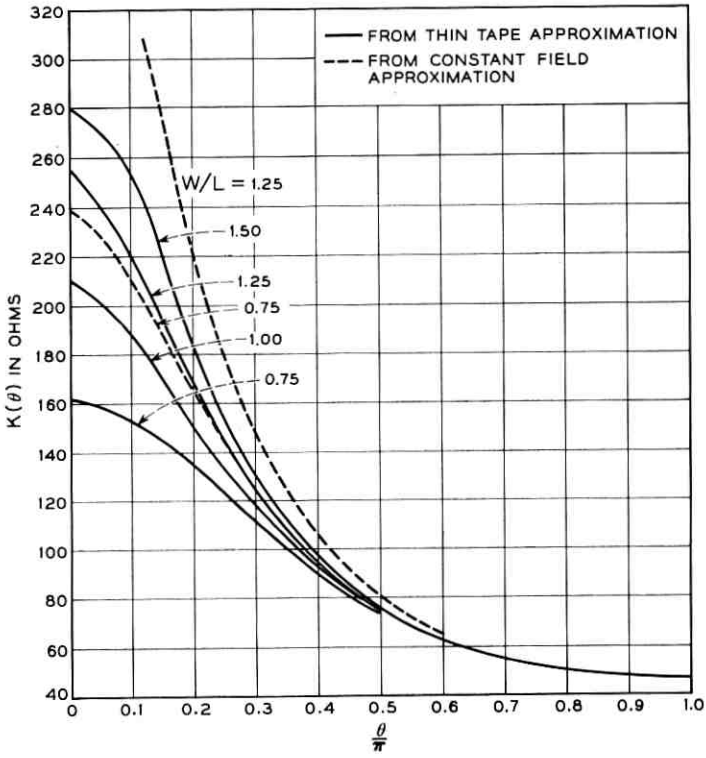


Fig. 2 — The impedance of the finger vs θ. It is assumed that $L - l = l = d$.

this approach is that the fields on the boundaries between the regions 2, 3 and the region 1 are not matched.

A third alternative would be to base the calculations on experimental data. The potential distribution functions $g(z)$ and $f(z)$ in (9) may be measured directly with a large scale two-finger model in an electrolytic tank. This is probably the most reliable method, although data are obtained with only limited accuracy and only numerically. In addition, the data are applicable directly with good accuracy only for $\theta = 0, \pi$.

The second (thin-tape) approximation will be used here, and it will be extended to cover the whole range of θ. With the thin-tape approximation, the functions g and f, which specify the potential distribution along the z-axis, for $\theta = 0$ and π respectively, can be found by Schwarz-Christoffel transformation. They have a quite complicated form involving elliptic functions, and therefore the integrals in (13) are difficult to

evaluate. Fortunately, at $\theta = 0$, $\pi/2$ and π, $K(\theta)$ can be found directly by Schwarz-Christoffel transformations without resort to (13). One can arrive at a good approximation of $Y_2(\theta)$ for the whole passband of the structure from (13) and the knowledge of Y_2 at $\theta = 0$, $\pi/2$ and π without evaluating the integral. The procedure is described in Appendix B, and the result is shown in Fig. 2 with solid lines for various values of W/L. Both the constant-field and the thin-tape approximations give about the same $K(\theta)$ for $\theta \geqq \pi/2$ but there is a large difference near $\theta = 0$. It will be shown later that $K(\theta)$ resulting from the thin-tape approximation yields a reasonably good agreement with experimental data.

IV. THE EFFECTIVE DIELECTRIC CONSTANT $\epsilon(\theta)$

If the comb is not entirely immersed in an isotropic medium, some RF field components appear in the direction of the fingers. The TEM assumption which has been used in Section II no longer holds. However, if the medium is only slightly nonuniform, the TEM assumption is still a good approximation. Fortunately, TWM's designed for high gain and large instantaneous bandwidth are so heavily loaded with an active crystal that the TEM approximation is reasonably valid and may be used in the structure analysis. In this section, an effective dielectric constant is defined. It is a function of θ — that is, of the details of the RF electric field configuration. At a particular value of θ, the effective dielectric constant of $\epsilon(\theta)$ is defined as the dielectric constant of a *uniform* medium filling the same comb structure, which results in the same total charge per comb finger as that produced by the true, incomplete dielectric loading (this definition is meaningful only for heavy dielectric loading such that the TEM approximation holds). It is obvious that this quantity $\epsilon(\theta)$ is a very helpful one for the analysis.

Referring to Fig. 1, a slab of dielectric of thickness D is placed in the regions 2 and 3. Its dielectric constant is assumed to be isotropic and equal to ϵ. TEM-type solutions of Maxwell's equations are assumed in regions 2 and 3, both inside and outside the dielectric slab. The components of the field vary as

$$\exp\left[\pm\beta_n x - jky - j\beta_n z\right]$$

where k is the plane wave propagation constant for the respective media. Boundary conditions have to be matched at $x = 0$, D and W in the y-z plane. The constant-field approximation expressed in (14) is explicitly used for the boundary condition at $x = 0$.

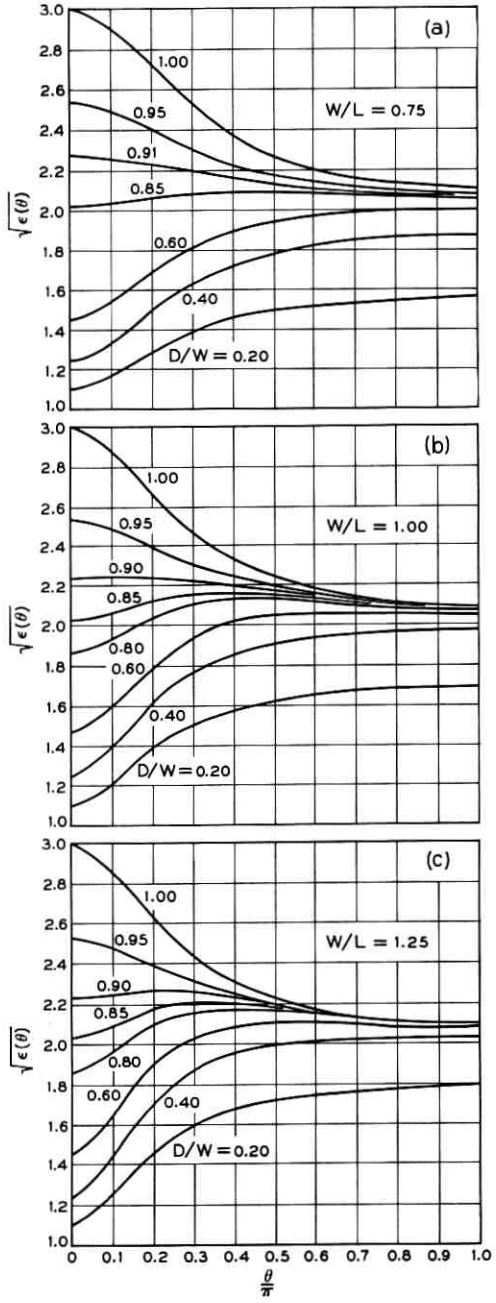


Fig. 3 — Square root of the effective dielectric constant vs θ. It is assumed that $L - l = l = d$, and $\epsilon = 9$.

Then the charge per unit length on a finger is given by

$$Q = 2\epsilon \int_{-(L-l)/2}^{(L-l)/2} E_{x2}dz + 2\int_{0}^{-d} E_{z1}dx \tag{17}$$

where E_{x2} is the x component of the electric field in region 2 at $x = 0$ and E_{z1} is the z component of the electric field in region 1. An effective dielectric constant $\epsilon(\theta)$ is now defined by requiring the same amount of charge Q to exist on the finger as if the whole structure were immersed in a medium of dielectric constant $\epsilon(\theta)$. The resulting formula can be given for combs with rectangular fingers; however, for simplicity, only the result for equally spaced square fingers ($L - l = d = L/2$) is given here

$$\epsilon(\theta) = \frac{\sin\frac{\theta}{2} + \frac{\epsilon}{2}\sum_{n=-\infty}^{\infty}(-1)^n\left(\frac{\sin\frac{\beta_n L}{4}}{\frac{\beta_n L}{4}}\right)^2 \times \left(\frac{1 + \epsilon\coth\beta_n(W-D)\coth\beta_n D}{1 + \epsilon\coth\beta_n(W-D)\tanh\beta_n D}\right)\tanh\beta_n D}{\sin\frac{\theta}{2} + \frac{1}{2}\sum_{n=-\infty}^{\infty}(-1)^n\left(\frac{\sin\frac{\beta_n L}{4}}{\frac{\beta_n L}{4}}\right)^2\coth\beta_n W} \tag{18}$$

The result of machine computations of $\epsilon(\theta)$ using (18) and assuming a dielectric constant of the loading dielectric of $\epsilon = 9$ (approximately the value of ruby) is shown in Fig. 3(a), (b), and (c). To facilitate other computations which will be discussed later, the square root $\sqrt{\epsilon(\theta)}$ is shown in these graphs rather than $\epsilon(\theta)$. For a fairly complete loading ($D/W > 0.8$), $\epsilon(\theta)$ approaches 5 near $\theta = \pi$. Near $\theta = 0$, $\epsilon(\theta)$ varies widely, depending on D/W. This can be understood by noting that near $\theta = \pi$ the RF fields concentrate near the fingers and hence $\epsilon(\theta)$ is little changed by a change in the width of the air gap near the waveguide wall. On the other hand, near $\theta = 0$ more of the fields reach the waveguide wall. Thus the width of the air gap there affects the magnitude of $\epsilon(0)$ more drastically than that of $\epsilon(\pi)$.

V. PROPERTIES OF THE COMB STRUCTURE

With the knowledge of $K(\theta)$, $C(\theta)$ and $\epsilon(\theta)$, one can calculate the ω-β diagram of both the empty and the loaded comb structure. The calculation reveals a number of interesting properties of the comb struc-

ture and also suggests various techniques for narrowing the passband of the structure.

5.1 *Empty Comb Structure*

As a maser structure, the comb is always loaded with one or more dielectrics. However, the study of the empty comb is of some interest to us, since it offers the possibility of checking the accuracy of our impedance calculations by measurements.

The fringe capacitance $C(\theta)$ for $\theta = 0$, and $\theta = \pi$ has been measured in the electrolytic tank. Values for $C(0)$ and $C(\pi)$ are shown in Fig. 4. They were obtained by resistance measurements on a large scale model of a comb finger in a tank. The values are plotted versus the distance between the finger tips and the opposite waveguide wall, s. The data are valid for fingers of square cross section, $L/2 \times L/2 = 0.040 \times 0.040$ inch, spaced center-to-center by $L = 0.08$ inch and contained in a housing of width $2W + L/2 = 0.240$ inch (aspect ratio $W/L = 1.25$). It should be mentioned here that these data can be applied to dimensions other than those indicated if one observes two facts. First, if all linear dimensions are scaled simultaneously by some factor, the capacity is scaled by the

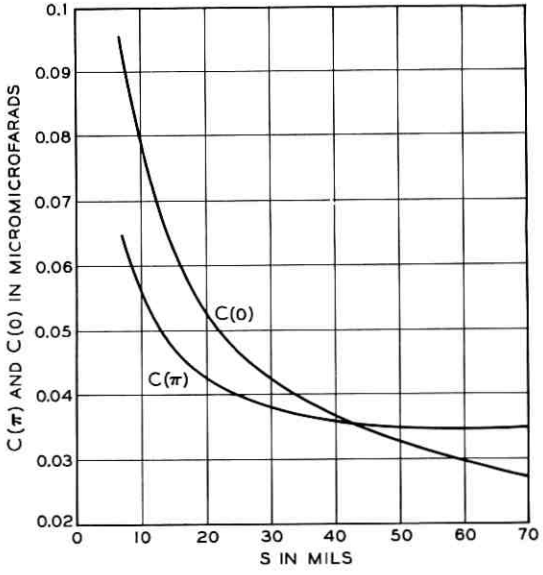


Fig. 4 — The fringe capacitance at $\theta = 0$ and $\theta = \pi$ vs the spacing between the finger tip and the waveguide wall for the case where $L - l = l = d = 0.040$ inch.

same factor. Second, experience has shown that the fringe capacity is a very slow function of the ratio W/L; no noticeable errors were found when these capacity values were used for W/L values ranging from 0.75 to 1.5. Unfortunately, $C(\theta)$ for θ other than 0, $\pi/2$ and π cannot be measured in a simple tank model. However, some indication of how $C(\theta)$ changes with θ may be obtained experimentally. One would start with a measured dispersion curve of an empty comb; one would assume that the values $K(\theta)$ in Fig. 2 for the thin-tape approximation are sufficiently accurate; then (1) offers a possibility of evaluating experimental values of $C(\theta)$. In practice it turns out, however, that this approach does not yield values $C(\theta)$ of sufficient accuracy to determine the exact shape of the $C(\theta)$ function.

In Fig. 5, calculated dispersion curves of several empty comb structures with dimensions as shown are given in solid lines. The calculation is based on $K(\theta)$ of Fig. 2 using the thin-tape approximation. $C(\theta)$ is assumed to change linearly with θ from $\theta = 0$ to $\theta = \pi$. Measured points of the dispersion curves are also shown in Fig. 5. The measurements and calculation agree well.

One case of a dispersion curve calculated by using $K(\theta)$ from the constant-field approximation is also shown in Fig. 5 by the dashed line. It deviates considerably from the measured points for $\theta < 0.5\ \pi$.

One may conclude that both the impedance of a finger calculated by using the thin-tape approximation and the assumption that $C(\theta)$ changes linearly with θ are sufficiently accurate for the present analysis.

The passband of the empty structure can be narrowed by reducing the width of the waveguide housing $2W + d$. Then the impedance values $K(\theta)$ at $\theta = \pi$ approach each other more closely and so do the Δh values at $\theta = 0$ and $\theta = \pi$. From (6), one readily sees that a narrower passband results.

5.2 *Loaded Comb Structure*

In this section, the discussion is restricted to the case of dielectric loading on both sides of the comb. Both loading slabs are parallelepipeds of equal thickness D. Both slabs cover the full finger height from the root to the tip, i.e., the height h of the fingers is also that of the loading slab. In addition, a comb of equally spaced square cross section fingers ($L - l = l = d$) is assumed.

Since (3) is only an approximation and also since the dielectric constant of ruby is neither a scalar nor exactly 9, one cannot expect to obtain a close quantitative agreement between the measured and the

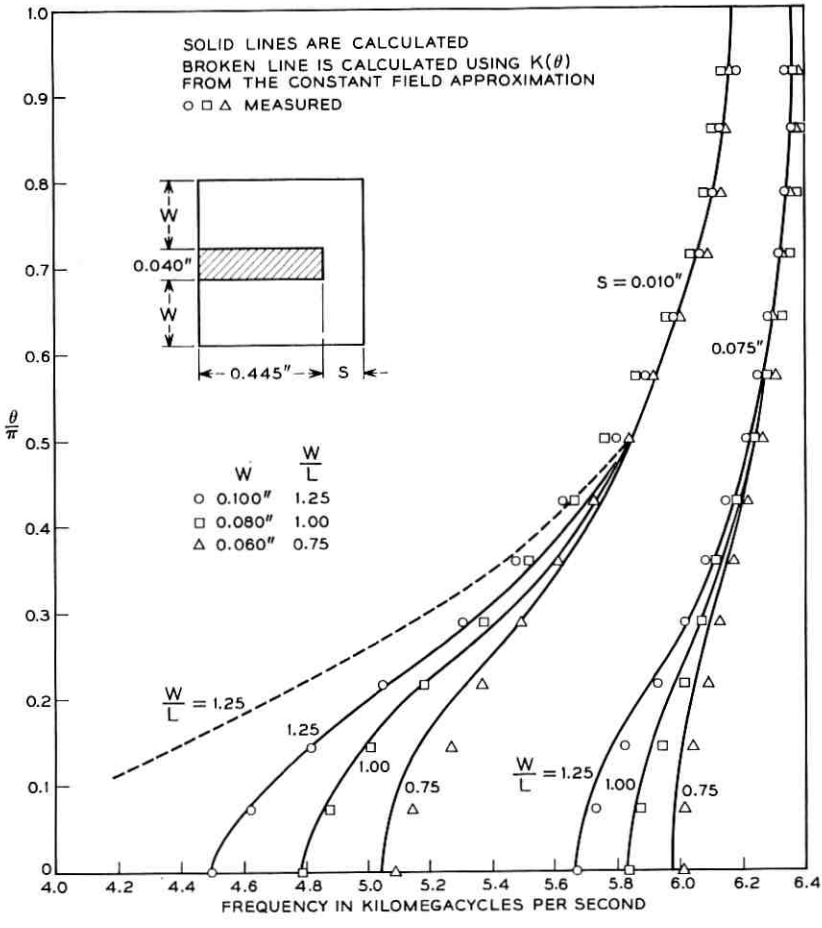


Fig. 5 — Calculated and the measured ω-β diagram of the empty comb. $L - l = l = d = 0.040$ inch.

calculated ω-β diagrams. However, the effects of various loading dimensions on the ω-β diagram are correctly predicted by the theory, and the present analysis provides a reliable basis for an initial choice of design parameters.

It will be convenient to define $f(\pi)$ and $f(0)$ as the frequency at which $\theta = \pi$ and 0, respectively. For a forward-wave structure ($df/d\theta > 0$), $f(\pi) > f(0)$, and for the backward-wave structure ($df/d\theta < 0$) $f(\pi) < f(0)$.

In Fig. 6, calculated dispersion curves are shown for $W/L = 1.25$, $K(\pi)C(\pi)c/h = 0.05$ and various values of D/W. The frequency scale is normalized to $f_0(\pi)$, where $f_0(\pi)$ is the frequency of the empty comb ($D/W = 0$) at $\theta = \pi$. Without loading ($D/W = 0$), the structure is a forward-wave structure. Relatively thin slabs of ruby loading (see the curve for $D/W = 0.2$) make it a backward-wave structure of a comparatively wide bandwidth. This can be understood by observing that the RF fields near $\theta = \pi$ are more concentrated near the fingers than the RF fields near $\theta = 0$. Hence, the RF fields at $\theta = \pi$ see more of the presence of the thin ruby slab than the fields at $\theta = 0$. In this way $f(\pi)$ is reduced while $f(0)$ remains essentially unchanged. Further increases in the width of the loading (see the curves for $D/W = 0.4$ and 0.6) reduce the bandwidth of the backward-wave structure. This happens because the RF fields near $\theta = 0$ begin to interact with the dielectric slab, while the fields near $\theta = \pi$ are almost completely contained in the initial thin slabs. This again changes the frequencies at $\theta = \pi$ and 0 to a different extent and thus reduces $f(0)$ more than $f(\pi)$.

At a still greater dielectric slab thickness (see the curve for $D/W = 0.9$) the structure is forward with a fairly narrow band and finally, with complete loading (see the curve for $D/W = 1.0$), it is forward with a

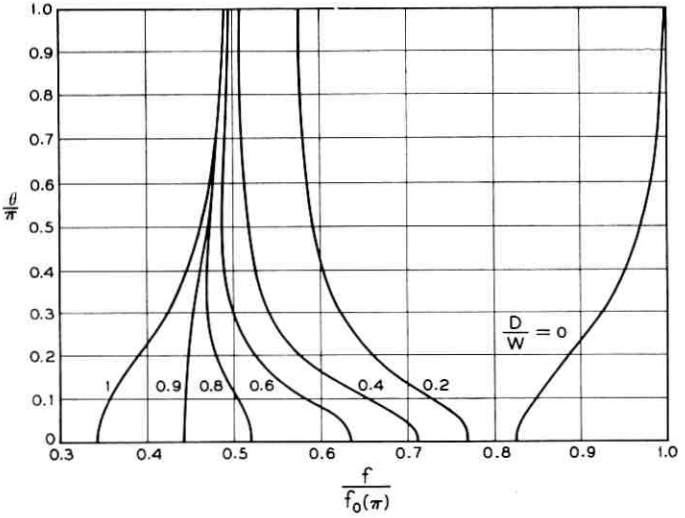


Fig. 6 — Calculated ω-β diagram as the thickness of the dielectric is changed. It is assumed that $L - l = l = d$, $W/L = 1.25$, $K(\pi)C(\pi)c/h = 0.05$, and $\epsilon = 9$. The frequency scale is normalized to $f_0(\pi)$, the frequency at $\theta = \pi$ for the empty comb.

somewhat wider band. The explanation is again based on the fact that increasing the dielectric width reduces the lower cutoff frequency while leaving the upper one the same.

The most important aspect of this behavior is seen by comparing the curves for $D/W = 0.6$, 0.8, and 0.9 in Fig. 6. It is apparent there that the transition between backward and forward structure does not involve backward-wave structures of gradually decreasing bandwidth, and then forward-wave structures of initially very narrow and eventually wider bandwidths. If this were so, it would be very easy to design comb structures with extremely high slowing of the group velocity. Instead, it is seen that the transition from backward to forward-wave structures takes place via intermediate structures showing "fold-over." By this, we mean a structure which for some range of θ is forward, whereas it is backward in the remainder of the θ range. It has been pointed out[9] that such a situation leads to instability and oscillations in a traveling-wave maser amplifier in spite of the incorporated isolator. Thus, for all practical applications, the occurrence of fold-over has to be avoided.

It may be added here that dispersion curves of combs with other dimensions vary in a similar fashion as the thickness of the dielectric slab is varied. It is of interest, however, to find out how the onset of fold-over is related to the comb geometry, i.e., to the ratio W/L, and to the finger end capacity. Here it is particularly desirable to have analytical data which indicate what choice of the W/L and D/W ratios and of the finger end capacity will result in the greatest slowing of the group velocity near the center portion of the passband, but still avoid fold-over. For this purpose, a number of dispersion curves normalized to $f(\pi)$ were calculated and are shown in Fig. 7. One notices that fold-over takes place rather abruptly for $D/W \lesssim 0.9$ for all cases. In addition, one sees that the minimum group velocity $[\alpha(df/d\theta)]$ attainable near the center of the passband without fold-over decreases by reducing W/L and $\Delta h(\pi)/h$. The group velocity near the center of the passband is more or less arbitrarily defined as

$$v_g = L\,\frac{2\pi(f\,|_{\theta=0.7\pi} - f\,|_{\theta=0.3\pi})}{0.4\pi} \tag{19}$$

It is shown as a slowing factor $S \equiv c/v_g$ in Fig. 8 vs W/L for $D/W = 0.9$ and for different values of $\Delta h(\pi)/h$. The largest slowing is obtained for the smallest W/L. It should be pointed out, however, that our present calculations were not carried out for W/L ratios of 0.5 or smaller, although such values would increase the slowing still further. Values of $W/L \lesssim 0.5$ appear unsuitable for practical maser designs for various

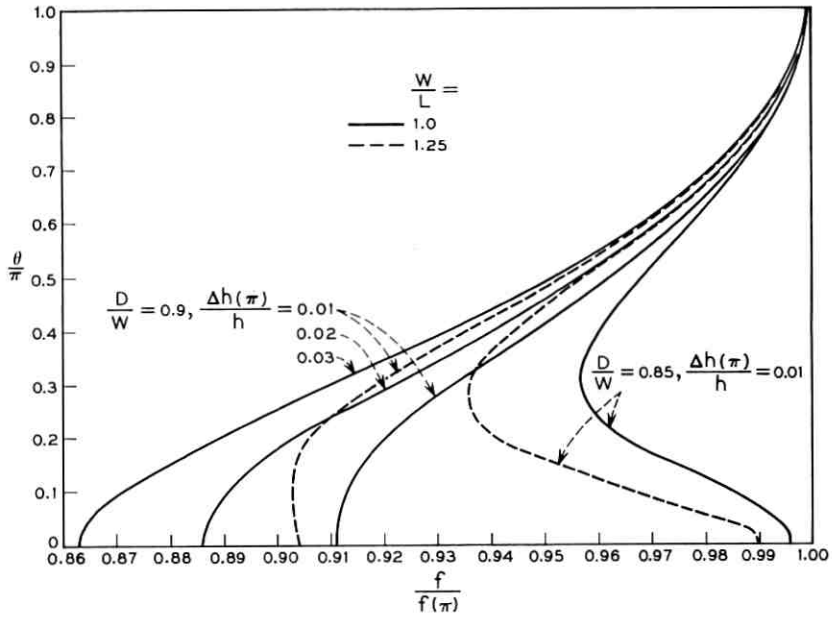


Fig. 7 — ω-β diagram for different W/L and $\Delta h(\pi)/h$. $D/W = 0.9$, $L - l = l = d$.

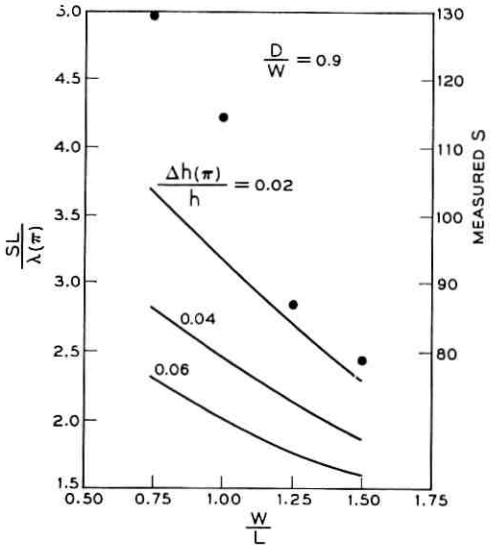


Fig. 8 — Slowing factor vs W/L. Measured points are also shown. The measurements are made with structures having $\Delta h/h = 0.02$, $\lambda(\pi) = 9.9$ cm, and $L = 0.08$ inch.

reasons, among them the difficulty of incorporating an isolator into a comb structure with such dimensions. Larger slowing results also from reducing $\Delta h(\pi)/h$. $\Delta h(\pi)$ can be varied over a narrow range by varying C, which is a function of s, the distance between the finger tip and the opposite waveguide wall. For the comb geometry used for measuring the fringe capacity, $C(\pi)$ does not appreciably decrease for s beyond $s \gtrsim 0.040$ inch. More generally, this would be done if s exceeds a value comparable to the finger "diameter." This therefore sets a minimum for $\Delta h(\pi)/h$ at a given operating frequency.

To illustrate the significance of Fig. 8, consider two TWM's A and B with exactly the same dimensions except that the finger length h of A is twice that of B. This difference in h makes the operating frequency of A about one-half of that of B. For the time being, assume also that the magnitude of the magnetic Q of the active crystal is independent of frequency. With the values $W/L = 1.0$ and $\Delta h(\pi)/h = 0.02$ for A, Fig. 8 shows that $SL/\lambda(\pi) = 3.2$. For B, $\Delta h(\pi)/h = 0.04$, so that $SL/\lambda(\pi) = 2.5$. The db gain of a TWM is proportional to fS/Q_m, which in turn, for constant Q_m, is proportional to $SL/\lambda(\pi)$. Thus one should expect that the db gain of maser A is larger than that of maser B by a factor 1.28 $(= 3.2/2.5)$. In practice, however, $|Q_m|$ usually increases toward lower frequencies, so that the gain of the lower-frequency maser A tends to be lower.

It should be added that the slowing itself, S, is inversely proportional to a scale factor $L/\lambda(\pi)$ where $\lambda(\pi) = c/f(\pi)$ is the free-space wavelength of $f(\pi)$. Thus for combs with a given period length L, the slowing is greater for lower frequencies. It follows that the slowing factor S is a meaningful parameter in comparing different comb structures only if they have essentially the same period length, L, and operating frequency range, f.

Consider another hypothetical case. Suppose the fringe capacity vanishes, $C(\pi) = C(0) = 0$, so that $\Delta h = 0$. Experimentally, this situation could be realized by a $\lambda/2$ ladder structure where fingers of twice the comb structure finger length are anchored at both ends in a waveguide enclosure. Without dielectric loading, this case would be that of the Easitron structure (see Ref. 7) which is characterized by a zero passband. With dielectric loading, however, a finite passband results nevertheless. This is due to the variation of $\epsilon(\theta)$ with θ, and in particular the difference between $\epsilon(\pi)$ and $\epsilon(0)$. Under these circumstances one might expect that the bandwidth is smaller and hence the slowing greater than in the case of a structure with finite fringe capacity and finite Δh. This is indeed the case. There is no curve shown in Fig. 8 for the parameter

$\Delta h = 0$, but it is obvious that this curve would lie above that for

$$\Delta h(\pi)/h = 0.02.$$

Measurements of dispersion curves of several comb structures loaded with alumina ($\epsilon = 9.3$) were made and they are shown in Fig. 9. Those parts of the dispersion curves where $\theta \geqq \pi/2$ depend very little on W/L and D/W. The fold-over takes place at D/W between 0.85 and 0.90. The maximum slowing factor near the center of the passband which is attainable without fold-over increases with smaller W/L. These general

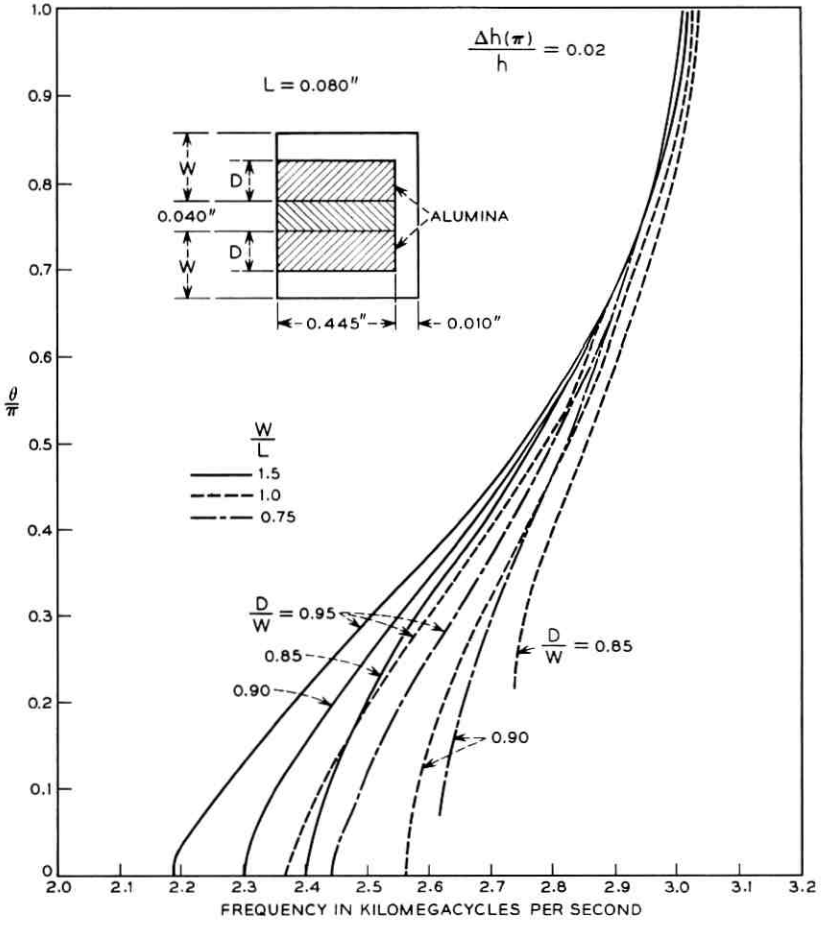


Fig. 9 — Measured ω-β diagrams of loaded comb structure. $L - l = l = d = 0.040$ inch.

features agree with the results of the calculation. There is a deviation of typically about 4 per cent near $\theta = \pi$, and of about 10 per cent near $\theta = 0$ between the absolute values of measured and calculated frequencies. These discrepancies are not too disturbing, however, since (3) is only an approximation; the fringe capacitance data derived for an empty comb change when the comb is loaded up to the tip of the fingers, and the dielectric constant of alumina is not exactly 9 as used in the calculations. The slowing factors calculated from these measurements are also shown in Fig. 8 as circled points. The measured slowing factors are about 30 per cent smaller than those calculated, but the dependence of S on W/L is correctly predicted by the theory.

VI. TECHNIQUES FOR REDUCING THE STRUCTURE BANDWIDTH

Since a structure with smaller bandwidth gives a larger slowing factor and thus a larger maser gain, the bandwidth of the structure should be kept as small as possible. It is necessary, of course, that the structure bandwidth exceed the required instantaneous or tunable design bandwidth of the maser amplifier by some reasonable, safe margin. This restriction was not important until recently. In earlier phases of traveling-wave maser development, it was difficult to design comb structures for sufficiently high slowing without running into the fold-over condition. More recently several techniques were developed which make it possible to design loaded comb structures with almost arbitrarily narrow bandwidths, down to structure bandwidths of only twice the instantaneous amplifier response.[6] These techniques were derived both by experimentation[10] and by the theoretical considerations reported in this paper. They include the following: (*i*) Near the tip of the fingers, the ruby may be shaped as in Fig. 10(a) or (b) by a step or bevel undercut. (*ii*) A strip of dielectric material of a high dielectric constant may be added next to the finger tips as in Fig. 10(c). (*iii*) The thickness of the comb fingers, d, may be reduced (see Fig. 1). These three techniques may be applied either independently or together to reduce the bandwidth of the structure.

By shaping the dielectric near the finger tip as in Fig. 10(a) and (b), the effective dielectric constant $\epsilon(\theta)$ is reduced. The reduction is not uniform across the band, but $\epsilon(\theta)$ is more drastically decreased for θ near 0. Thus, for a forward-wave structure where the lower cutoff frequency occurs at $\theta = 0$, the lower cutoff frequency increases without much change to the upper cutoff frequency. The bevel shape of Fig. 10(b) can be considered as a series of small steps, as indicated by the dashed line.

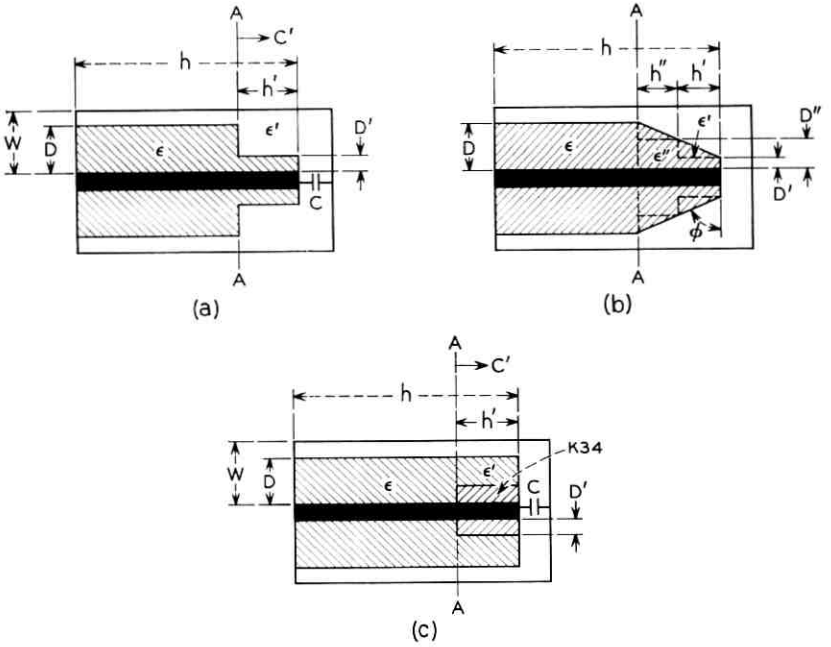


Fig. 10 — Techniques of "shaping" and "K34 loading" near the finger tip to reduce the bandwidth.

The dielectric shapes of Fig. 10(a) and (b) are approximately equivalent in their role of narrowing the bandwidth.

The finger tip loading may take the form shown in Fig. 10(c). A thin slab of dielectric material with high dielectric constant is imbedded in alumina or ruby near the finger tip. In our experiments, a ceramic with $\epsilon = 34$ manufactured by American Lava Corporation was used. This material will be referred to as K34. This type of finger tip loading increases the effective dielectric constant $\epsilon(\theta)$ near $\theta = \pi$ faster than near $\theta = 0$. For the forward-wave structure, this means that the upper cutoff frequency can be decreased faster than the lower cutoff frequency, and thus a narrowing of the bandwidth results.

Let h' be the length of the stepped dielectric (Fig. 10a) or K34 (Fig. 10c), C' be the capacitance looking toward the finger tip at the plane A-A, ϵ' be the effective dielectric constant of the section h', and D' be the thickness dimension as shown. The capacitive impedance looking toward the right side at the plane A-A, $1/\omega C'$, can be found by regarding the section h' as a TEM transmission line with the characteristic im-

pedance $K(\theta)/\sqrt{\epsilon'(\theta)}$ and a propagation constant $\omega\sqrt{\epsilon'(\theta)}/c$ and terminated by C at the end. If the electrical length of the modified section of finger line (which has the physical length h') is small compared to a quarter wavelength,

$$\tan \frac{\omega\sqrt{\epsilon'(\theta)}}{c} h' \approx \frac{\omega\sqrt{\epsilon'(\theta)}}{c} h',$$

and if the fringe capacity loading at the finger tip is small,

$$\frac{1}{\omega C} \gg \frac{K(\theta)}{\sqrt{\epsilon'(\theta)}}$$

then the effect of the stepped dielectric can be expressed by an effective capacity C'

$$C' \approx C[1 + (\epsilon' h'/KCc)] \tag{20}$$

which effectively terminates the regular TEM finger transmission lines of length h-h'. From (6) and (20) one obtains the following formula, which contains implicitly the ω-β relation

$$\frac{(2m-1)}{4}\lambda \approx \sqrt{\epsilon(\theta)}\left(h + \frac{KCc}{\epsilon} + \frac{\epsilon' - \epsilon}{\epsilon} h'\right). \tag{21}$$

For the stepped ruby as in Fig. 10(a) where $D > D'$, $(\epsilon' - \epsilon)/\sqrt{\epsilon}$ is negative. This quantity can be calculated from Fig. 3 and it is shown in Fig. 11 in solid curves. Measured values are also shown as circles. The measurements and calculations of $(\epsilon' - \epsilon)/\sqrt{\epsilon}$ agree well except near $\theta = 0$. In practical design work, it is often convenient to keep one of the cutoff frequencies unchanged while shifting the other cutoff frequency by shaping the dielectric. This requires a large ratio of

$$|(\epsilon' - \epsilon)/\sqrt{\epsilon}|$$

at the two cutoff frequencies. In Fig. 11, the curves for $D/W = 0.95$ and $0.4 \leqq D'/W \leqq 0.6$ satisfy this requirement. As a special case of stepped ruby, the value $D' = 0$ may also be considered. This corresponds to a ruby loading of rectangular cross section which does not cover the full finger height h. Then ϵ' becomes unity and the value $(1 - \epsilon)/\sqrt{\epsilon}$ is also plotted in Fig. 11 in the top curve marked $D'/W = 0.0$. Since the ratio $(1 - \epsilon)/\sqrt{\epsilon}$ at both cutoff frequencies, that is at $\theta = 0$ and at $\theta = \pi$, is not large, the reduction of the ruby height does not appear a promising way to narrow the bandwidth of the comb structure.

For K34 loading, $(\epsilon' - \epsilon)/\sqrt{\epsilon}$ becomes positive and can be evaluated

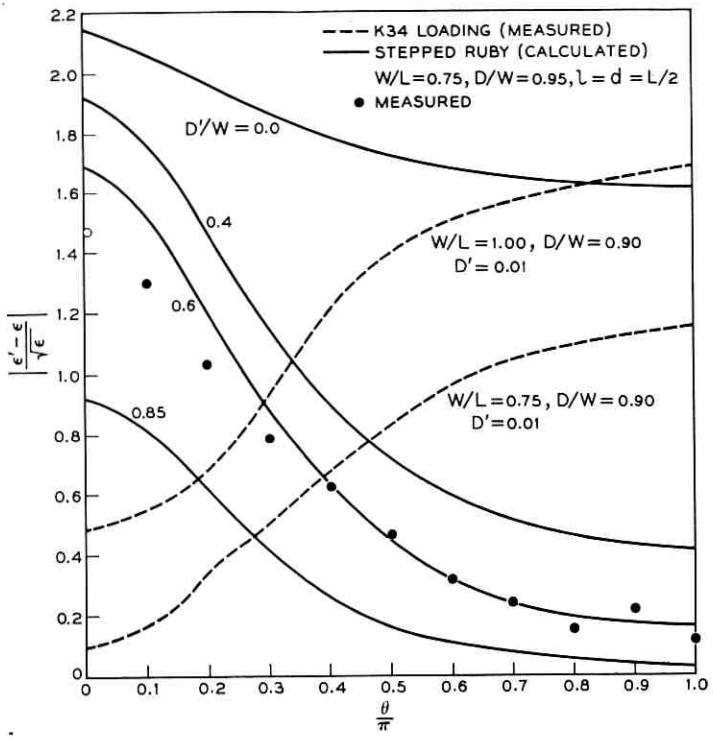


Fig. 11 — $\left|\frac{\epsilon' - \epsilon}{\sqrt{\epsilon}}\right|$ vs θ/π for ruby shaping and K34 loading.

from measured data with the help of (21). Two of the measured curves of $(\epsilon' - \epsilon)/\sqrt{\epsilon}$ are shown in Fig. 11 in broken lines. $(\epsilon' - \epsilon)/\sqrt{\epsilon}$ is larger near $\theta = \pi$ than near $\theta = 0$. For the forward-wave structure this causes the upper cutoff frequency to decrease while leaving the lower frequency almost unchanged.

The upper cutoff frequency of the forward-wave structure can also be reduced by using thin rectangular fingers (i.e., $d < L - l$). $\epsilon(\theta)$ of the comb structure with fingers of square cross section was shown in Fig. 3. It was also shown that $D/W = 0.9 \cdots 0.95$ is usually the best choice to reduce the bandwidth of the passband and yet avoid fold-over near $\theta = 0$. For this value of D/W, one notices in Fig. 3 that $\epsilon(0) > \epsilon(\pi)$. In order to reduce the bandwidth further, one may increase $\epsilon(\pi)$ so that it approaches $\epsilon(0)$. This can be done by reducing the dimension d of the fingers. Since the fields between the fingers are almost negligible at $\theta = 0$, the thickness dimension of the fingers does not affect $\epsilon(0)$. On

the other hand, near $\theta = \pi$ the fields see more of the ruby with thin fingers than with square fingers; thereby $\epsilon(\pi)$ increases.

In Fig. 12, $\sqrt{\epsilon(\pi)}$ vs $d/(L - l)$ using the expression given by Harris et al.[7] is shown. Three measured points are also indicated. The expression of $\epsilon(\pi)$ in Ref. 7 assumes a uniform field between the fingers. This approximation is less justified as the fingers become thin, although it should be qualitatively correct even for $d/(L - l) < 0.5$.

Both the K34 loading and the choice of finger thickness affect the frequencies near $\theta = \pi$. Band narrowing by thin fingers has the additional advantage that the filling factor improves somewhat compared to the use of K34 loading and of square fingers.

All the three techniques described here can be combined to narrow the passband very effectively in such a way that fold-over is still avoided.

VII. HIGHER-ORDER MODES

It has been shown in Section II that there exists a series of higher-order passbands for a comb structure.

Let us compare the first and the second modes of operation from two different approaches. Equation (6) shows that for a given θ the free-space wavelength of the first mode is three times longer than that of the second mode when all of the dimensions of the structure are kept the same. Thus the percentage bandwidth of the passbands is approximately the same for all modes. By using a frequency scale for the first mode which is

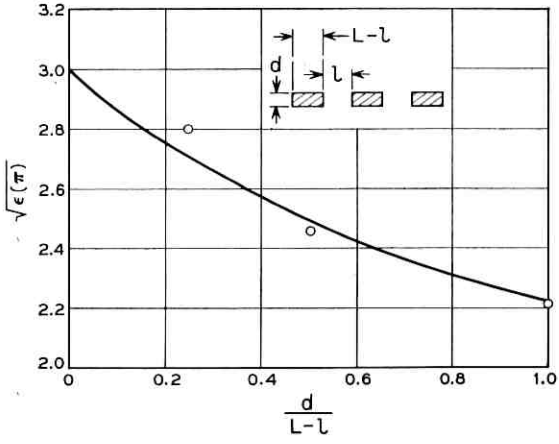


FIG. 12 — The square root of the effective dielectric constant at $\theta = \pi$ for "thin fingers." It is assumed that $\epsilon = 9$.

one-third that for the second mode, the ω-β diagram of the first and the second modes should coincide. This is indeed verified from measured ω-β diagrams of the first and the second modes of the structure designated as A in Fig. 13.

On the other hand, when W/L, D/W and s are kept unchanged while the length of the finger h is made three times longer, the passband of the smaller structure operating in its first mode will be in about the same frequency range as that of the larger structure operating in its second mode. However, the bandwidth of the smaller structure in the first mode is larger than that of the larger structure in the second mode. This is due to a larger $\Delta h/h$ for the structure with the smaller h. A structure B with a finger length of one-third that of structure A was built, and the ω-β diagram of its first mode is also shown in Fig. 13. One notices that the

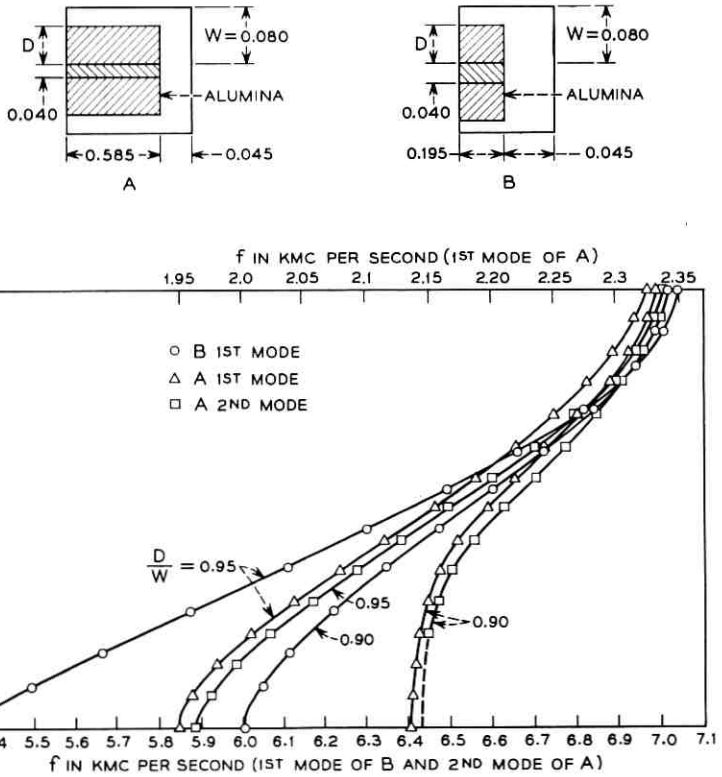


Fig. 13 — Comparisons of the first and the second passbands of the loaded comb structures. The fingers are 0.040 × 0.040 inch in cross section and l = 0.040 inch.

first mode of B has a larger bandwidth than the second mode of A. The larger slowing factor associated with the smaller bandwidth is one of the advantages of higher-order modes, but it is not an essential factor, since there are more efficient techniques of reducing the available bandwidth. However, the higher-order modes may be advantageous when the required microwave frequency of the passband is so high that it becomes difficult to fabricate comb fingers of a quarter-wave electrical length, provided the waveguide modes which may propagate in such a structure can be suppressed. Then the larger structure would permit reduced mechanical tolerances.

VIII. DESIGN CONSIDERATIONS

Since the gain of the maser increases as the bandwidth of the passband of the structure is reduced, it is preferable to design a structure having the narrowest bandwidth compatible with the requirements on the useful bandwidth of the maser.

The choice of various dimensions of the structure to obtain a given center frequency and to approach the narrowest bandwidth possible without fold-over will be discussed here.

From the curves in Fig. 8, a small value of W/L is preferred. Other design considerations may dictate the smallest permissible W/L value. In all our experiments, for example, a period length $L = 0.080$ inch was chosen. For W/L smaller than 0.7 it would seem rather difficult to incorporate a high-performance isolator into the structure. Thus $W/L \approx 0.7$ is a compromise optimum value.

With given dimensions of the empty comb and without ruby shaping near the finger tip, the bandwidth is reduced by gradually reducing D/W until fold-over sets in near $\theta = 0$. This happens at about $D/W \approx 0.9$. Fold-over appears more readily if additional "ruby shaping" is applied to the structure. Thus D/W should be larger than 0.9 in order to allow for some latitude in ruby shaping. From our experience $D/W = 0.95$ is a suitable value.

For the step in the ruby near the finger tip (Fig. 10a), D'/W may be taken between 0.4 and 0.6. Then $|(\epsilon' - \epsilon)/\sqrt{\epsilon}|$ at $\theta = 0$ is large, and the ratio of its magnitude at $\theta = 0$ and $\theta = \pi$ is also large.

The frequency near the $\theta = \pi$ end can be easily controlled by using thin fingers. The curve in Fig. 12 may serve as a guide in the choice of a suitable finger aspect ratio $d/(L - l)$.

The final parameter yet to be determined is the length of the fingers, h. For an initial design, one may use the midband value of $\sqrt{\epsilon}(\pi/2)$

from Fig. 3. Then

$$h \approx \frac{\lambda}{4\sqrt{\epsilon(\pi/2)}}$$

where λ is the free-space wavelength of the given center frequency. The three techniques discussed above for narrowing the bandwidth affect primarily the frequencies near $\theta = 0$ and $\theta = \pi$, i.e., close to both cutoffs; but the frequencies near $\theta = \pi/2$, i.e., close to midband, remain nearly unchanged.

Following the suggestions given here, one should arrive at a first-order design for a TWM which will perform fairly close to the theoretical expectation. Experience has shown that a small additional amount of fine adjustment is needed in order to have the traveling-wave maser perform according to the specifications. This may involve control of the center frequency, adjustment for more or less slowing in order to obtain the design gain, or adjustment of the curvature in the ω-β diagram so as to realize a flat gain-versus-frequency characteristic. The theoretical data provided in this paper make it rather easy to determine the appropriate design modifications.

IX. ACKNOWLEDGMENT

The author wishes to thank E. O. Schulz-DuBois for his helpful suggestions and critical reading of the manuscript. He also wishes to thank Mrs. E. Sonnenblick for the numerical computations and R. C. Peterson for the capacitance measurements in the electrolytic tank.

APPENDIX A

The Potential Distribution on the z Axis

Consider the coordinate axes shown in Fig. 14. The potential on the z axis for $-(l/2) < z < (l/2)$ can be expressed as

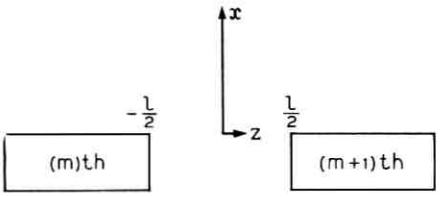


Fig. 14 The coordinate system used in Appendix A.

$$V(\theta,z) = V_s(\theta)g(z) + V_a(\theta)f(z) \tag{22}$$

where $V_s(\theta)$ and $V_a(\theta)$ are those parts of the potential which are symmetrical and antisymmetrical with respect to the x axis, respectively, and $g(z)$ and $f(z)$ describe their variation along the z axis. They are defined so that

$$g\left(-\frac{l}{2}\right) = g\left(\frac{l}{2}\right) = 1 \tag{23}$$

$$-f\left(-\frac{l}{2}\right) = f\left(\frac{l}{2}\right) = 1. \tag{24}$$

Substituting (23) and (24) into (22), the potentials of the mth and $(m+1)$th finger become, respectively,

$$Ae^{-jm\theta} = V_s - V_a \tag{25}$$

$$Ae^{-j(m+1)\theta} = V_s + V_a. \tag{26}$$

From (25) and (26),

$$V_s = Ae^{-j(m+\frac{1}{2})\theta}\cos(\theta/2) \tag{27}$$

$$V_a = -jAe^{-j(m+\frac{1}{2})\theta}\sin(\theta/2). \tag{28}$$

Then (22) becomes

$$V(\theta,z) = Ae^{-j(m+\frac{1}{2})\theta}[g(z)\cos(\theta/2) - jf(z)\sin(\theta/2)]. \tag{29}$$

Assuming the fingers to be infinitely thin, $g(z)$ and $f(z)$ can be found from a Schwarz-Christoffel transformation. This transformation is described in Ref. 7. The result can be given in closed form for the derivatives $g' = dg/dz$ and $f' = df/dz$:

$$g' = \frac{\pi}{L\sqrt{\sin^2\frac{\pi z}{2l} - 0.5}} \frac{\cos\frac{\pi z}{2l}}{\frac{\pi W}{L} + \ln\sqrt{2}} \tag{30}$$

$$f' = \frac{\pi}{L\sqrt{\sin^2\frac{\pi z}{2l} - 0.5}} \frac{1}{1.854} \tag{31}$$

where it is assumed that $l/L = 0.5$. In deriving (30), an elliptic function has been approximated by a sine function. This approximation becomes better with larger W/L. The origin of the z coordinate for (30) and (31) is the same as originally indicated in Fig. 1, which differs from that shown

in Fig. 14. It is seen that an infinity occurs both for g' and f' at the finger corners, $z = \pm l/2$.

APPENDIX B

Impedance of a Finger Based on the "Thin-Tape" Approximation

It was shown in Section III that the admittance of a finger can be found by adding the admittance between the fingers Y_1, and the admittance between the finger and the waveguide Y_2. Y_2 can be expressed in terms of the potentials f and g along the z axis of Fig. 14. f and g are the potentials at $\theta = \pi$ and $\theta = 0$, respectively. We shall assume f and g to be those obtained by conformal mapping of the thin-tape geometry.

By separating the term with $n = 0$ from the remaining summation, (13) may be rewritten in the following form

$$\frac{Y_2(\theta)}{Y_0} \doteq \frac{\sin\frac{\theta}{4}}{\frac{\theta}{4}} \coth\left(\theta\frac{W}{L}\right) \times \frac{2}{L}\left[\sin\frac{\theta}{2}\int_{-l/2}^{l/2} f' \cos\frac{\theta}{L} z\, dz + \cos\frac{\theta}{2}\int_{-l/2}^{l/2} g' \sin\frac{\theta z}{L}\, dz\right] + \sum_{\substack{n=-\infty \\ n\neq 0}}^{\infty} (-1)^{n+r} \cdot \frac{\sin\left(\frac{\theta}{4}+\frac{n\pi}{2}\right)}{\frac{\theta}{4}+\frac{n\pi}{2}} \left(\sin\frac{\theta}{2}\right)\frac{2}{L}\int_{-l/2}^{l/2} f' \cos(\theta + 2n\pi)\frac{z}{L}\, dz \tag{32}$$

for

$$L - l = l = d$$

where

$$r = 0 \quad \text{for} \quad n > 0$$

$$r = 1 \quad \text{for} \quad n < 0.$$

The terms involving g are omitted except for $n = 0$, since usually $g' \ll f'$. The coordinate origin used in this equation is at the position shown in Fig. 14.

Let Y_L, Y_M and Y_U be $Y_2(\theta)/Y_0$ at $\theta = 0$, $\pi/2$ and π, respectively. These quantities can be obtained directly by conformal mapping. Y_L and Y_U are shown in Ref. 7. Y_M is obtained by a similar procedure and

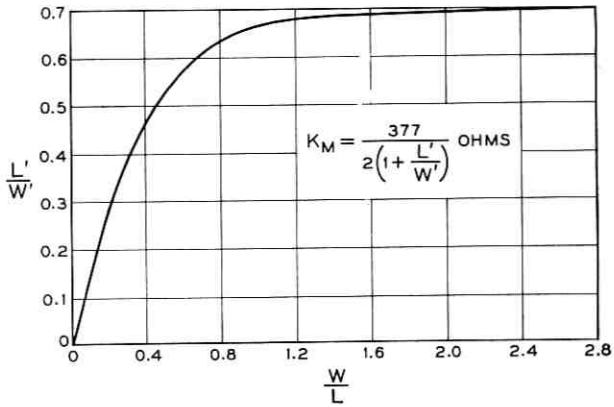


Fig. 15 — The impedance $K(\theta)$ of a finger at $\theta = \pi/2$.

the result is shown in Fig. 15. For θ other than these three values, (32) should in principle be used to find $Y_2(\theta)$. Fortunately, $Y_2(\theta)$ can be expressed approximately in terms of Y_L and Y_M or Y_U in an interpolation formula, and thus the tedious evaluation of (32) can be avoided.

Letting $\theta = 0$ and then $\theta = \pi/2$ in (32), one obtains,

$$Y_L = \frac{1}{W}\left[\int_{-l/2}^{l/2} f'\,dz + \frac{2}{L}\int_{-l/2}^{l/2} zg'\,dz\right] \tag{33}$$

$$Y_M = \frac{\sin\frac{\pi}{8}}{\frac{\pi}{8}} \coth\frac{\pi W}{2L}\left(\sin\frac{\pi}{4}\right)\frac{2}{L}\left[\int_{-l/2}^{l/2} f' \cos\frac{\pi z}{2L}\,dz + \int_{-l/2}^{l/2} g' \sin\frac{\pi z}{2L}\,dz\right] + \frac{2}{L}\sum_{\substack{n=-\infty \\ n\neq 0}}^{\infty}(-1)^{n+r} \cdot \frac{\sin\left(\frac{\pi}{8}+\frac{n\pi}{2}\right)}{\frac{\pi}{8}+\frac{n\pi}{2}} \sin\frac{\pi}{4}\int_{-l/2}^{l/2} f' \cos\left(\frac{\pi}{2}+2n\pi\right)\frac{z}{L}\,dz. \tag{34}$$

We shall assume that approximately

$$\int_{-l/2}^{l/2} f'\,dz \doteq \int_{-l/2}^{l/2} f' \cos\frac{\theta z}{L}\,dz \tag{35}$$

$$\frac{\sin\left(\frac{\theta}{4}+\frac{n\pi}{2}\right)}{\frac{\theta}{4}+\frac{n\pi}{2}} \doteq \frac{\sin\left(\frac{\pi}{8}+\frac{n\pi}{2}\right)}{\frac{\pi}{8}+\frac{n\pi}{2}}, \qquad \text{for } n \neq 0 \quad (36)$$

$$\int_{-l/2}^{l/2} f' \cos(\theta + 2n\pi)\frac{z}{L}\,dz \doteq \int_{-l/2}^{l/2} f' \cos\left(\frac{\pi}{2}+2n\pi\right)\frac{z}{L}\,dz, \qquad \text{for } n \neq 0 \quad (37)$$

$$\int_{-l/2}^{l/2} g' \sin\frac{\theta z}{L}\,dz \doteq \frac{2}{L}\int_{-l/2}^{l/2} g'\,z\,dz. \quad (38)$$

Then (32) becomes

$$Y_2(\theta) \doteq \frac{\sin\frac{\pi}{2}}{\sin\frac{\pi}{4}} Y_M + \left(\frac{\sin\frac{\theta}{4}}{\frac{\theta}{4}}\coth\theta\frac{W}{L} - \frac{\sin\frac{\pi}{8}}{\frac{\pi}{8}}\coth\frac{\pi W}{2L}\right)\frac{2W}{L} \cdot\sin\frac{\theta}{2}\,Y_L + \frac{\sin\frac{\theta}{4}}{\frac{\theta}{4}}\left(\coth\theta\frac{W}{L}\right)\left(\cos\frac{\theta}{2}-\sin\frac{\theta}{2}\right)\frac{4}{L^2}\int_{-l/2}^{l/2} g'\,z\,dz. \quad (39)$$

The last term of (39) is small compared to the other terms, and it can be found by numerical integration of g' in (30).

One notices that $Y_2(\theta)$ of (39) becomes Y_L and Y_M when $\theta = 0$ and $\theta = \pi/2$, respectively. One may estimate the error involved in (39) by letting $\theta = \pi$ and compare it with Y_U obtained by conformal mapping. For $W/L = 1$, $Y_2(\pi)$ obtained from (39) gives a value 10 per cent smaller than that obtained directly by conformal mapping. The admittance of a finger which is the sum of Y_1 and Y_2 thus has an error of about 4 per cent at $\theta = \pi$. For other values of θ, the error would be even smaller.

REFERENCES

1. DeGrasse, R. W., Schulz-DuBois, E. O., and Scovil, H. E. D., The Three-Level Solid-State Traveling-Wave Maser, B.S.T.J., **38**, Mar., 1959, pp. 305–334.
2. Fletcher, R. C., A Broadband Interdigital Circuit for Use in Traveling-Wave Tube Amplifiers, Proc. I.R.E., **40**, Aug., 1952, pp. 951–958.
3. Butcher, P. N., The Coupling Impedance of Tape Structures, J. IEE, **104B**, Mar., 1957, pp. 177–187.
4. Walling, J., Interdigital and Other Slow Wave Structures, J. Electronics, **3**, Mar., 1957, pp. 239–258.
5. Ash, E. A., and Studd, A. C., A Ladder Structure for Millimeter Waves, I.R.E. Trans. Electron Devices, **ED-8**, July, 1961, pp. 294–302.

6. Tabor, W. J., A 100-Mc Broadband Ruby Traveling-Wave Maser at 5 Gc, Proc. IEEE, **51,** August, 1963, p. 1143.
7. Harris, S., DeGrasse, R. W., and Schulz-DuBois, E. O., Cutoff Frequencies of the Dielectrically Loaded Comb Structure as Used in Traveling-Wave Masers, B.S.T.J., **43,** Jan., 1964, p. 437.
8. Haddad, G. I., and Rowe, J. E., X-Band Ladder-Line Traveling-Wave Maser, I.R.E. Trans. on Microwave Theory and Techniques, **MTT-10,** January 1962, pp. 3–8.
9. Kostelnik, J. J., DeGrasse, R. W., and Scovil, H. E. D., The Dual Channel 2390-mc Traveling-Wave Maser, B.S.T.J., **40,** July, 1961, pp. 1117–1127.
10. DeGrasse, R. W., and Hensel, M. L., private communication.

A Comparison of Permanent Electrical Connections

By G. W. MILLS

(Manuscript received February 10, 1964)

A study has been completed which compares four types of permanent electrical connections (soldered, solderless wrapped, percussive welded, and resistance welded) under environmental conditions of vibration, shock, temperature extremes, corrosion, humidity, and bending. Only good-quality connections were included in this study, and they represented the current state-of-the-art for each type. Under these conditions the connections showed no significant degradation in their electrical characteristics as long as they remained mechanically secure. Differences in the four types of connections were therefore assessed in relation to their mechanical characteristics. Consequently, one of the more important results of the study was the recognition of fatigue life as the most important mechanical connection characteristic when comparing connections which meet the high standards of the Bell System for electrical stability. Using fatigue life as a basis for comparison and soldered connections as a reference standard, the major conclusions with regard to general wiring (the connection of wires to terminals, such as surface and local cable wiring) are as follows for the conditions that existed in this study:

(a) monitored percussive welded connections are superior to soldered connections;

(b) over-all, solderless wrapped connections are essentially equivalent to soldered connections;

(c) resistance welded connections are significantly inferior to soldered connections.

Although differences were found among the types of connections, no evidence was obtained that any of the connection types are not satisfactory as presently used in normal Bell System applications.

I. INTRODUCTION

The Bell System uses many types of electrical connections. The best connection for a specific application is chosen on the basis of the relative

merits of the connections in the following general areas:

(1) adaptability of each connection for the application under consideration,

(2) reliability or life required from the connections under the environmental conditions in which the equipment must operate, and

(3) relative cost of each connection.

This study was concerned with obtaining information on the comparative reliability or life of the four main types of permanent connections (solderless wrapped, soldered, percussive welded, and resistance welded) under various environmental conditions. The probability of the occurrence of substandard connections was recognized as a factor in determining the reliability of a given type of connection. However, only good-quality connections of each type were compared in this study, and all present manufacturing standards were followed in making them.

All four types of connections are considered adequate for the applications in which they are presently used, and the results of this study should not be construed as a recommendation to change to a different type of connection in these applications.

The scope of the program is illustrated by Fig. 1. It shows general wiring application (the connection of wires to terminals such as surface and local cable wiring) and the environments considered.

Some of the environments chosen are more severe than those normally encountered in a central office. This was necessary to produce a measurable effect in a reasonable time and should not alter the results, because this was a comparison study.

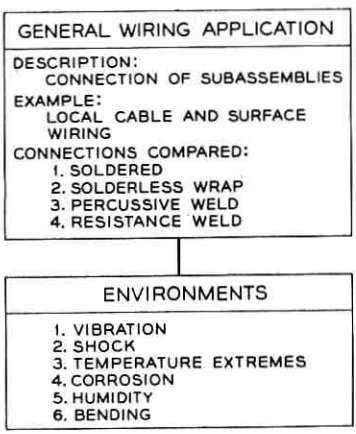


Fig. 1 — Test program for permanent connection comparison.

The end result from each of the environmental tests was the establishment of an "order of merit" for the four types of connections in that particular environment. Thus the connection which suffered the least degradation in an environment would be placed at the top of the "order of merit" list and the connection which showed the greatest degradation would be on the bottom.

Early in the test program, it was observed that none of the test environments caused any electrical failures and that very few of the test environments caused mechanical connection failures. Thus, it was concluded that the establishment of an "order of merit" would have to be obtained from some observed mechanical degradation in destructive tests. Two types of tests always led to connection failure, i.e., breakage, and these were the vibration and bending fatigue tests. The establishment of an "order of merit" for these tests was very simple: the connections which lasted the longest were the best and those which failed first were the poorest.

Since both of these tests are essentially fatigue tests, it was realized that the fatigue life of a connection could be a better basis of comparison than an ultimate-strength test, since all the test connections had met the rigid Bell System standards for electrical stability. Further analysis, as described in Section 3.1, led to the conclusion that fatigue life should be used as the comparison basis throughout this study for the following reasons:

(1) A comparison based on fatigue life offers an absolute comparison scale for the types of connections studied.

(2) Electrically stable permanent connections can be characterized by their fatigue life.

II. TEST PROCEDURES AND RESULTS

2.1 *General*

The quality and uniformity of each group of connections was determined by destructively testing approximately half the group and thus establishing the strength distribution of the remaining half, or test connections. The destructively tested or control connections were generally selected alternately by order of manufacture, from the whole group of connections. The destructive strength control data are presented in the Appendix along with a description of the destructive strength method used for each type of connection.

Because of the statistical nature of the data collected in a program of

this type it is necessary to present the data in the form of probability distributions, as shown in Fig. 2. The ends of the bar represent the 10 per cent and 90 per cent points on the distribution and the projecting line between represents the 50 per cent point.

2.2 *Environmental Tests*

Typical examples of the four types of connections compared in this study are shown in Fig. 3. Since this study is basically a comparison of the four types of connections, every effort was made to hold all of the unknown parameters to an absolute minimum, and this was accomplished to a great extent by subjecting all of the connections to the same environment at the same time. The connections were mounted on fixtures to facilitate the handling and the mounting during exposure to the various environments. These mountings were of two types, as shown in Fig. 4: (a) connections with insulation, consisting of 3-inch loops of wire connecting two terminals, the group thus containing forty connections, i.e., 20 loops of wire; (b) connections without insulation, the wire having a 90° bend near the terminal and being fastened directly to a standoff insulator. This last type was designed for the resistance change measurements (ΔR) which were required for the temperature, corrosion and humidity test environments.

This study consisted of the sixteen general wiring tests listed in Table I. Most of these were of the comparison type, but a few were studies concerned with only one type of connection. All of the vibration fatigue

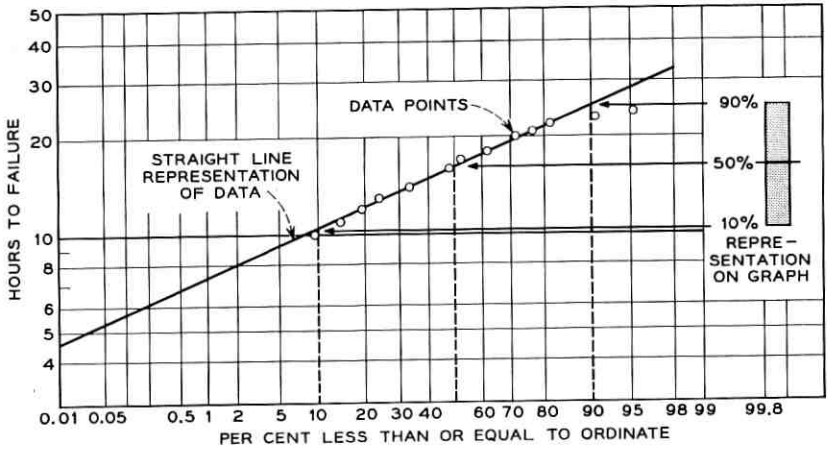


Fig. 2 — Typical test data.

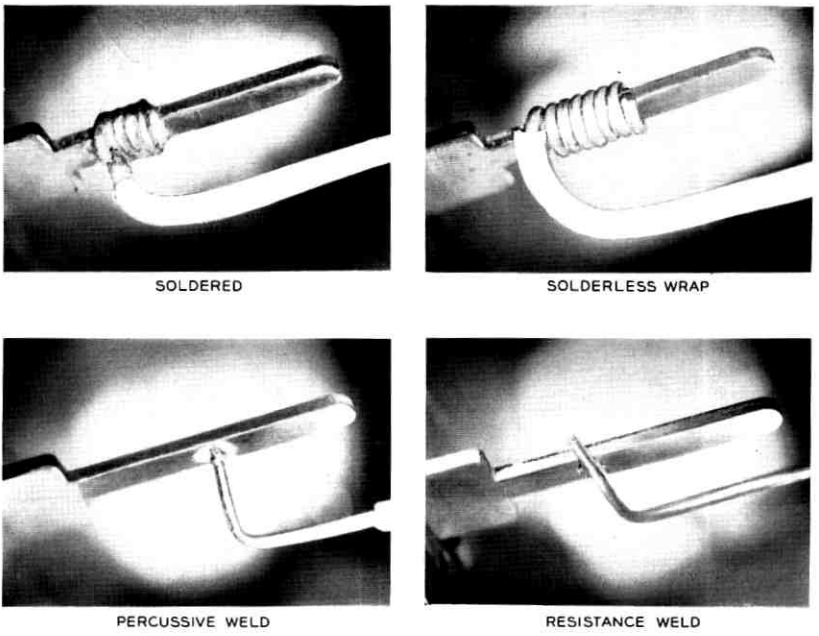


Fig. 3 — Typical examples of general wiring connections.

tests, with the exception of test 14, used the vibration configuration shown in Fig. 5.

The first footnote of Table I requires some explanation. The vibration fatigue life determinations in this study were spread over a considerable number of months because of the long exposure times for some of the

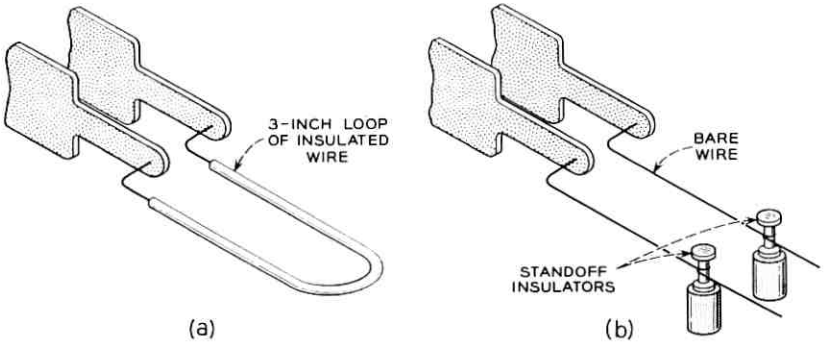


Fig. 4 — Typical general wiring connections: (a) with insulation, (b) without insulation.

TABLE I — LIST OF GENERAL WIRING TESTS

Test Description	Test Number	Insulation		Fatigue Method			Number of Connections of Each Type Tested	Figure Number of Results
		Without	With	Vibration	Lightly Loaded Bending	Heavily Loaded Bending		
Vibration test without insulation	1	X		X			80	6
Vibration test with insulation	2		X	X			40	6
Shock test, laboratory (fatigue life)	3		X	X			40	7
Shock test, laboratory (destructive test)	4		X				40	8
Shock test, railroad non-cushioned	5*		X	X			20	9
Shock test, railroad cushioned	6*		X	X			20	9
Temperature test, central office conditions	7	X		X			40	10
Temperature test, outside plant conditions	8	X		X			40	10
Three months corrosion test	9	X		X			20	11
Six months corrosion test	10*	X		X			20	11
Humidity test	11	X		X			40	12
Lightly loaded bending test (30° angular displacement)	12		X		X		40	14
Heavily loaded bending test (45° angular displacement)	13		X			X	40	16
Configuration test	14		X	X			40	17
Ultimate strength as a function of fatigue life	15	X			X		120†	18
Inferior weld test	16	X			X		20†	19

* Test conducted after recalibration of the vibration machine.
† Only percussive welded connections tested.

tests. Toward the end of the testing, the mounting springs on the table of the vibration machine had to be replaced and the machine recalibrated. This recalibration seems to have affected the fatigue life determination of some or all of the subsequent tests; however, it has not been possible to determine the extent of this difference or even to prove that

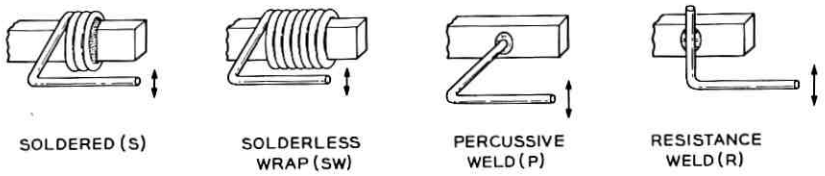


Fig. 5 — Standard vibration configuration.

it does or does not exist. If a difference does exist, it would introduce an error in the comparisons between the vibration fatigue life data of some of the later tests and the control fatigue life data (test 1 or 2) but would not affect the comparison among the four types of connections on a particular test, since they all experienced the same vibration environment.

2.2.1 *Vibration Tests*

The vibration tests have been divided into two general classifications: (a) those for wires without insulation and (b) those for wires with insulation.

The results of these two tests are shown in Fig. 6, and they cover the fatigue life of the four types of connections, with and without insulation. The connections which were connected with a loop were grouped in pairs. They were vibrated according to the schedule listed in Table II and used the vibration configuration illustrated in Fig. 5. The actual motion of the wires was a function of the mass of the loop as well as the acceleration of the connection; therefore all loops were made as close to the same size as possible.

In test 1 (without insulation) it was necessary to solder loops on the connections, since these were set up for resistance change (ΔR) measurements as shown in Fig. 4(b) and therefore had no loops. Care was taken to control the size and weight of these loops so that they would closely approximate the loops with insulation in test 2, shown in Fig. 4(a).

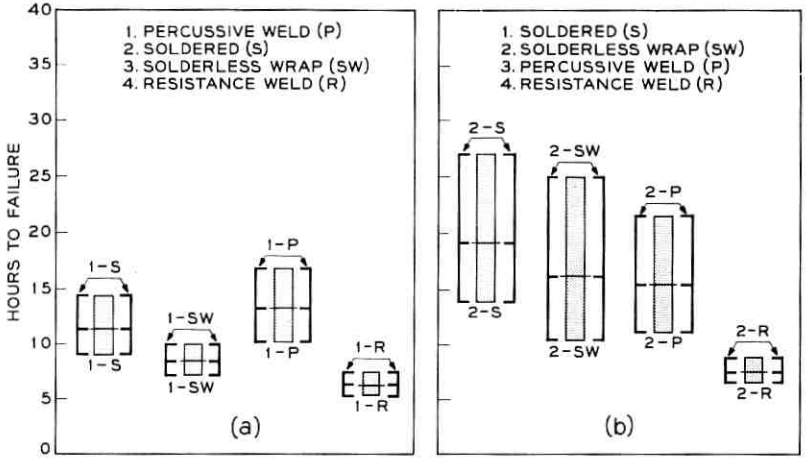


Fig. 6 — Vibration fatigue life results. (a) Test number 1 (without insulation); order of merit, based on vibration fatigue life. (b) Test number 2 (with insulation); order of merit, based on vibration fatigue life.

TABLE II — VIBRATION SCHEDULE

The vibration was sinusoidal in wave shape and its frequency varied from 5 cps to 500 cps and back to 5 cps in 2 minutes. The displacement, acceleration, and running time schedules were as follows:

Displacement (Inches)	Acceleration (G's)	Running Time (Hours)
0.1	5	2
0.2	10	2
0.3	15	2
0.4	20	2
0.5	25	to destruction

Note: The cross-over frequency was approximately 30 cps: that is, 5 to 30 cps controlled displacement and 30 to 500 cps controlled acceleration.

Since the connections were grouped in pairs and the weakest connection of the pair failed first, no usable fatigue data could be obtained from the remaining connection and it was clipped off, leaving a stub which could be destructively tested.

2.2.2 *Shock Tests*

The shock tests can be divided into two general classifications: (a) laboratory shock and (b) railroad shock. Two laboratory shock tests were conducted, one using the vibration fatigue life after the shock test as a measure of the connection degradation, and the other using the destructive strength as an indication of any degradation suffered by the connections. These tests consisted of subjecting all four types of connections to 90 high-level shocks. All of the shock test connections were mounted on a fixture and experienced the same shocks at the same time. Half sine wave shocks were used with a peak amplitude of 500 to 600 G's and a duration of 2 to 3 milliseconds. There were no connection failures due to this shock environment. In the case of the fatigue life part, test 3, the connections were then subjected to the vibration schedule (Table II) and their fatigue life determined. The results of these tests are presented in Fig. 7.

In the case of the destructive strength part, test 4, the connections were destructively tested after the 90 shocks, each in accordance with the method prescribed for it, as described in the Appendix. The results of these tests are compared to a destructive strength control and are shown in Fig. 8.

It should be noted that the order of merit for test 4 could not be based on any absolute strength scale because of the differences in the destruc-

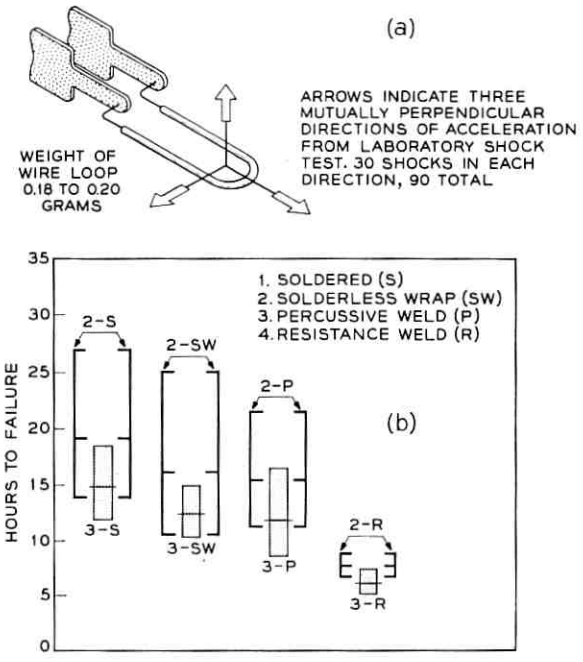


Fig. 7 — Vibration fatigue life after laboratory shock tests. (a) Typical connection orientation during shock tests. (b) Test number 3 (with insulation); order of merit, based on vibration fatigue life.

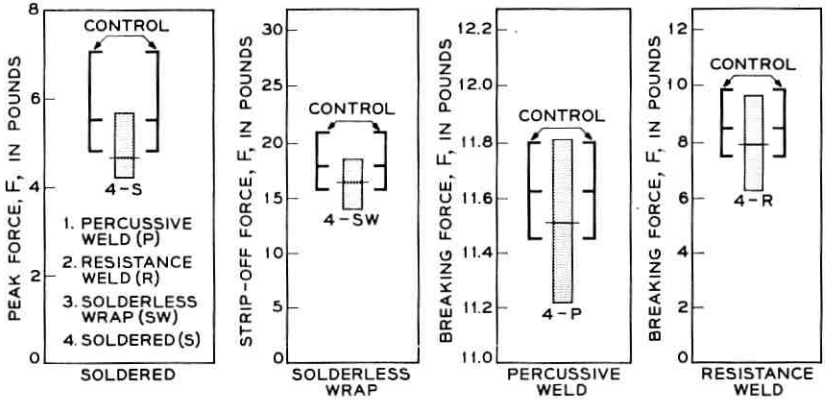


Fig. 8 — Destructive strength after laboratory shock tests (test number 4); order of merit, based on percentage of destructive test degradation.

tive strength tests. It was therefore based on the percentage degradation indicated by the difference between the mean of the control sample and the mean of the sample which had been exposed to the shock.

The railroad shock tests were of two types: (a) noncushioned, test 5, and (b) cushioned, test 6. These tests consisted of sending the four types of connections in a rigid plywood box from Columbus, Ohio, to New York City by railway express for a total of ten round trips. The connections from both tests, numbers 5 and 6, were shipped in the same container; however, the noncushioned connections were fastened directly to the plywood box, whereas the cushioned connections were supported by rubberized hair in the center of the container. The fatigue life of the connections was determined after the ten round trips by vibrating them according to the vibration schedule (Table II). The results of both tests are presented in Fig. 9. There was very little difference in degradation observed for the two tests, and they resulted in identical orders of merit based on the vibration fatigue life.

The vibration fatigue life determination for these two tests was conducted after recalibration of the vibration machine as discussed in Section 2.2, and this could affect the comparison between these tests and the control; however, it will not affect the comparison of each type of connection within tests 5 and 6.

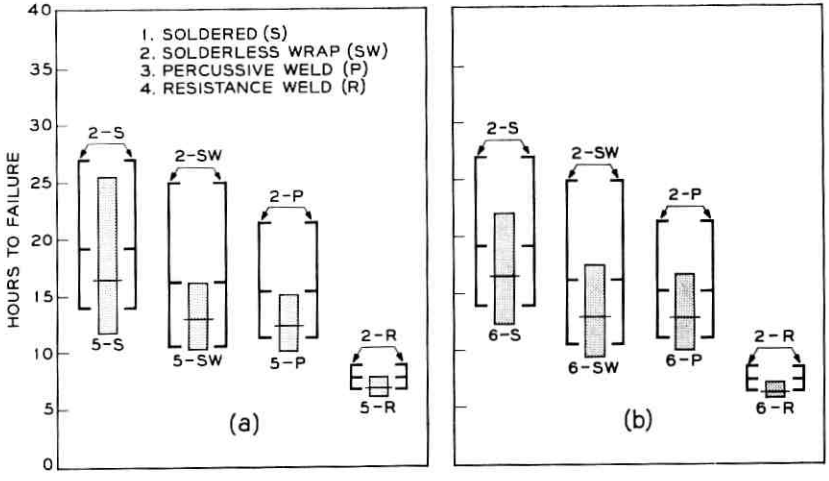


Fig. 9 — Vibration fatigue life after railroad shock and vibration. (a) Test number 5 (noncushioned); (b) test number 6 (cushioned). The order of merit, based on vibration fatigue life, was the same for both tests.

2.2.3 *Temperature Tests*

The connections were subjected to two types of temperature tests: (a) central office conditions, test 7, and (b) outside plant conditions, test 8.

The central office temperature test consisted of subjecting forty connections of each type to a temperature of 105°C for a total of 154 days. Once a week the connections were removed from the oven and allowed to come to room temperature (20°C). Every two weeks the connections were mechanically disturbed (plucked) and the change in resistance, ΔR, was measured. After the 154 days, loops were carefully soldered to the connections and they were subjected to vibration according to the vibration schedule (Table II) and their fatigue life determined. The results are presented in Fig. 10. All forty connections of each type survived the 154-day test without developing a change in resistance, ΔR, of 0.001 ohm, which would have constituted an electrical failure. The solderless wrap connections, however, did develop a considerably higher number of ΔR's, as shown in Table III.

The soldered, solderless wrap, and resistance welded connections showed a small but significant loss in fatigue life due to test 7. The percussive welded connections, however, showed a larger, more significant loss in fatigue life due to the temperature test.

The outside plant temperature test, test 8, was very similar to the central office condition temperature test. The differences were that test 8 ran for 168 days and that when the connections were removed from the

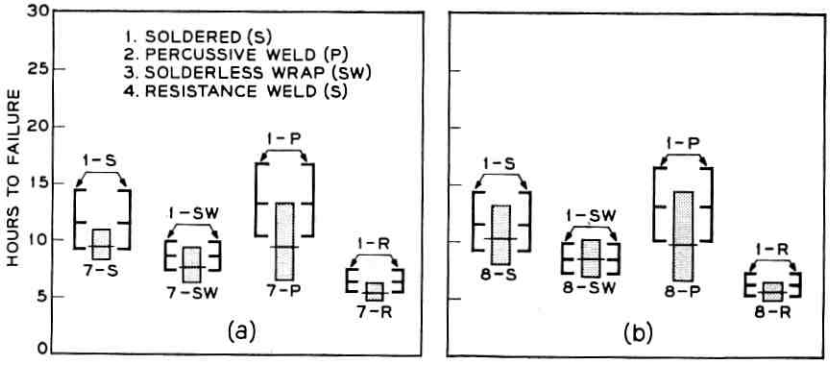


Fig. 10 — Vibration fatigue life after temperature tests. (a) Test number 7 (central office conditions); (b) test number 8 (outside plant conditions). The order of merit, based on vibration fatigue life, was the same for both tests.

TABLE III — NUMBER OF ΔR VALUES OBSERVED FOR EACH TYPE OF CONNECTION

ΔR Values in Milliohms	Soldered		Solderless Wrap		Percussive Weld		Resistance Weld	
	CO	OP	CO	OP	CO	OP	CO	OP
0.2	6	none	42	44	1	1	3	1
0.3	none	none	8	9	none	none	none	none
0.4	none	none	1	2	none	none	none	none
0.5	none	none	4	1	none	none	none	none
0.6	none	none	none	1	none	none	none	none

Note: CO = central office conditions, OP = outside plant conditions.

oven at 105°C they were immediately placed in a cold box at −40°C and allowed to stabilize at this temperature. After approximately an hour at −40°C, they were removed and allowed to come to room temperature (20°C). Resistance change data and vibration fatigue life information were obtained in the same manner as the central office temperature test; the results are presented in Table III and Fig. 10. The changes in resistance and fatigue life data for the outside plant conditions were similar to the results obtained from the central office conditions, except that the soldered and solderless wrapped connections sustained less degradation. Both temperature tests yielded the same order of merit.

2.2.4 *Corrosion Tests*

The corrosion tests were of two types: (a) three months, test 9, and (b) six months, test 10. These tests were identical in all respects with the exception of the exposure time, as indicated in Fig. 11. They consisted of exposing all four types of connections to the corrosive atmosphere of New York City on the roof of the Bell Laboratories building at West Street. Resistance change measurements were made on all of the connections before and after exposure. All corrosion test connections survived the environment without developing the 0.001-ohm resistance change which would have constituted an electrical failure.

The apparent increase in fatigue life of all the connections, except for percussive welded connections, in the six months corrosion test is unexplainable except for the recalibration of the vibration machine discussed in Section 2.2. This is the most probable cause of this inconsistency, since there is no reason to expect that exposure to a corrosive atmosphere can improve the fatigue life of any type of connection.

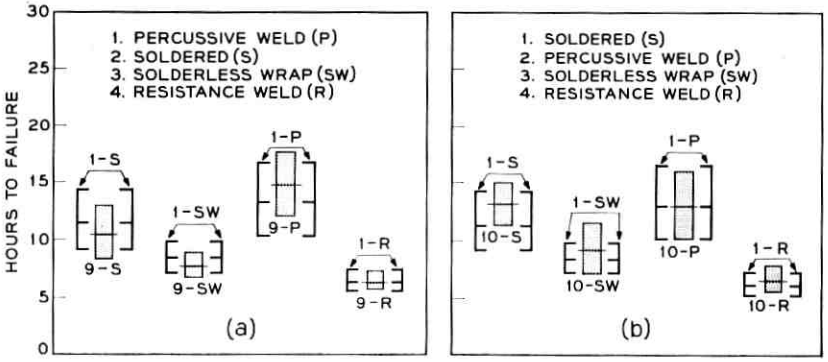


Fig. 11 — Vibration fatigue life and orders of merit after corrosion tests. (a) Test number 9 (three months); (b) test number 10 (six months).

2.2.5 *Humidity Test*

The humidity test, test 11, consisted of subjecting all four types of connections to controlled temperature and humidity conditions according to the following schedule: they were exposed to 90 per cent relative humidity and 85°F dry bulb temperature for six consecutive days, then dried at 140°F for two days. This cycle was repeated eight times for a total test time of 64 days. Resistance change measurements were made before, during, and after the test with no electrical failures being observed. The fatigue life of the connections was determined by carefully soldering loops of wire to the connections and vibrating them according to the vibration schedule (Table II). The results are presented in Fig. 12.

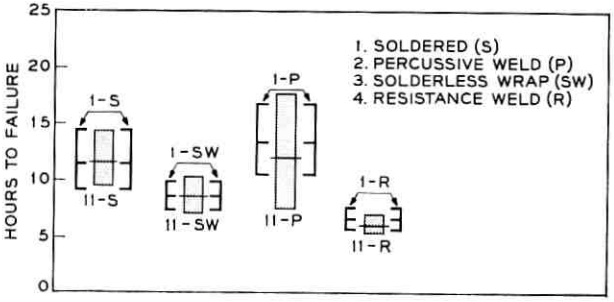


Fig. 12 — Vibration fatigue life and order of merit after humidity test, test number 11.

2.2.6 *Bending Tests*

All four types of connections were subjected to two bending tests: (a) the lightly loaded bending test, test 12, and (b) the heavily loaded bending test, test 13.

Bending, in the lightly loaded bending test, took place in a horizontal plane as shown in Fig. 13. The load on the connections came from the tension in the wire and varied from approximately 0 to 4 grams. The wire was moved 30° in one direction from its equilibrium position and then returned, and then moved 30° in the other direction and returned to its equilibrium position. This constituted one cycle. The number of such cycles to failure for each connection was the fatigue data. Forty connections of each type were tested and the results are presented in Fig. 14.

The heavily loaded bending test consisted of determining the bending fatigue life of all four types of connections by bending in a vertical plane with 300 grams hanging on the connection, as shown in Fig. 15. At a relatively slow rate the terminal was rotated 45° from its originally horizontal position and then returned. This constituted one cycle. The number of such cycles required to cause failure of each connection was the fatigue data. Forty connections of each type were tested for all connections except solderless wrap, of which 32 were tested; results are presented in Fig. 16.

2.2.7 *Additional Tests*

There were three additional tests conducted in this study; (a) configuration test, test 14, (b) ultimate strength as a function of bending fatigue life for percussive welded connections, test 15, and (c) inferior weld test, test 16.

The configuration test consisted of determining the vibration fatigue life of the four types of connections, using a different configuration. The 90° bend used in the configuration shown in Fig. 5 was eliminated and the wire was brought straight from the terminal as shown in Fig. 17(b). Vi-

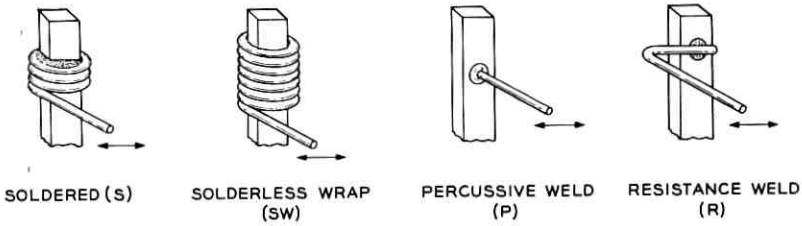


Fig. 13 — Lightly loaded bending configurations.

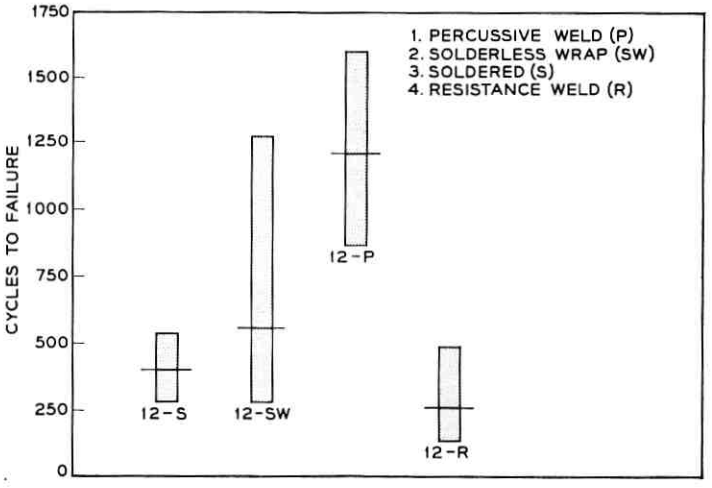


Fig. 14 — Lightly loaded bending fatigue life, test number 12 (with insulation); order of merit, based on bending fatigue life.

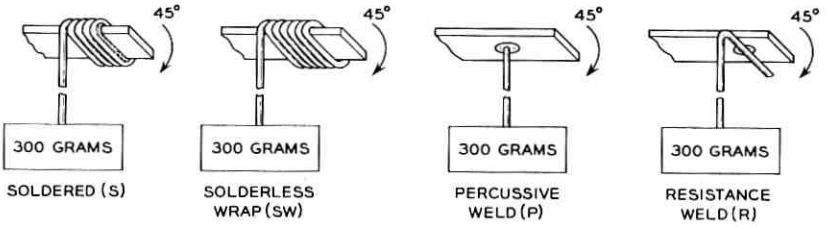


Fig. 15 — Heavily loaded bending configurations.

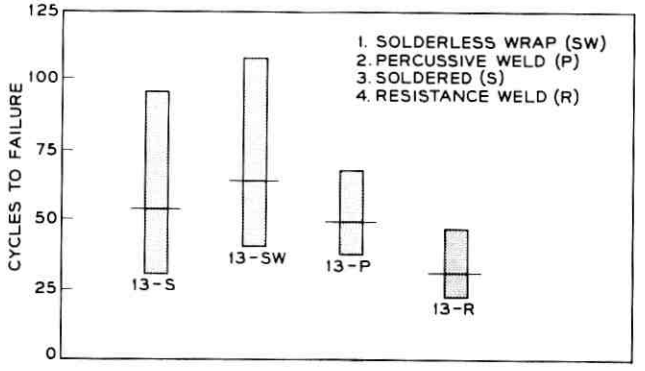


Fig. 16 — Heavily loaded bending fatigue life, test number 13 (with insulation); order of merit, based on bending fatigue life.

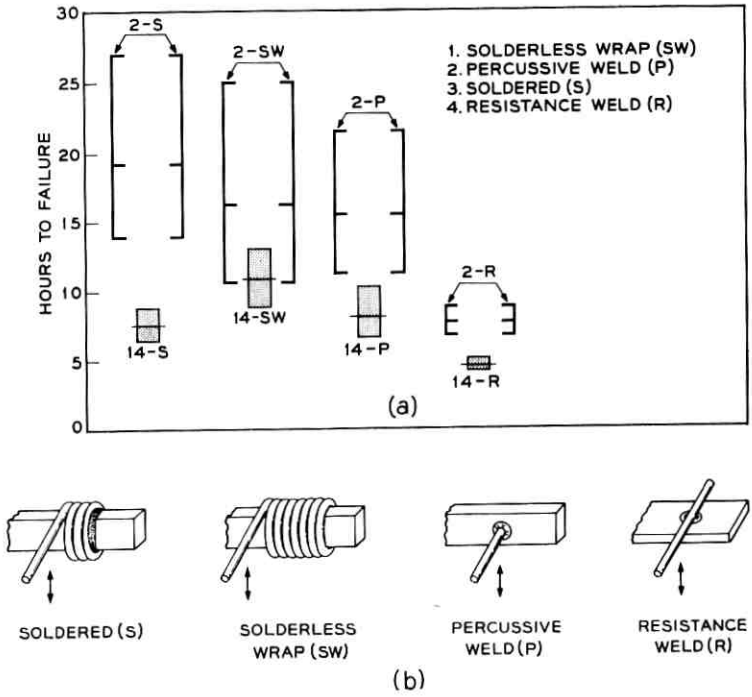


Fig. 17 — (*a*) Vibration fatigue life with configuration omitting 90° bend in the lead-off wire, test number 14 (with insulation); order of merit, based on vibration fatigue life. (*b*) Vibration configuration omitting 90° bend.

bration still took place in a vertical plane and the vibration schedule (Table II) was followed. The results of this test are also shown in Fig. 17(a). All four types of connections showed a deterioration in their fatigue life when the 90° bend was omitted. This bend apparently partially isolates the connections from external wire movements and therefore improves their fatigue life.

The objective of test 15 was to determine the ultimate strength of percussive welded connections as a function of the bending fatigue life. The fatigue method selected was the lightly loaded bending type. A total of 240 percussive welded connections were used in this test; they were divided into two groups of 120 each. The first group was destructively tested immediately after manufacture to determine their strength, and these data are included in the destructive strength control presented in the Appendix. The remaining 120 connections were divided into eleven groups: one group of 20 connections and 10 groups of 10 connections

TABLE IV — SCHEDULE FOR TEST 15

Number of lightly loaded bending cycles	500	600	700	800	900	1000	1100	1200	1300	1400
Surviving connections	10	10	9	10	9	9	7	3	6	2
Average strength of surviving connections in pounds	11.1	11.1	11.1	11.2	10.8	10.9	11.1	10.7	10.5	10.7

each. The 20-connection group was fatigued to failure by the lightly loaded bending method to establish the fatigue life of the connections. The other ten groups were each fatigued to a predetermined number of bending cycles according to the schedule shown in Table IV. After the connections in each group had been fatigued for the number of cycles assigned to that group, the connections remaining were destructively tested using the combined test.

The results of this test, presented in Fig. 18, indicate that at 100 per cent of the average fatigue life the surviving connections show only an approximate 5 per cent reduction in their ultimate strength. This indicates that the mechanical quality of a percussive welded connection is characterized better by its fatigue life than by its ultimate strength.

The object of the inferior weld test, test 16, was to determine the ultimate strength and fatigue life of inferior percussive welds and to compare these characteristics with those of good welds. The welds were made inferior by using insufficient capacitance in the welding power supply during their manufacture. The results are presented in Fig. 19. The

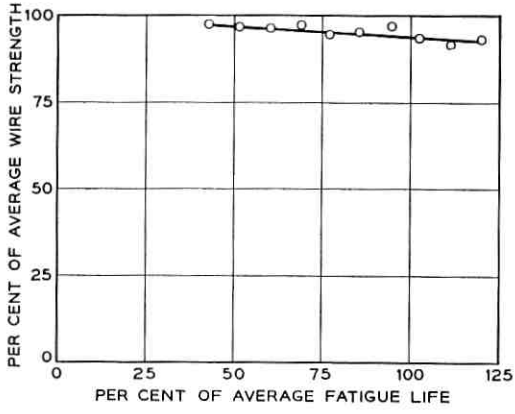


Fig. 18 — Ultimate strength as a function of bending fatigue life for percussive welded connections, test number 15 (without insulation).

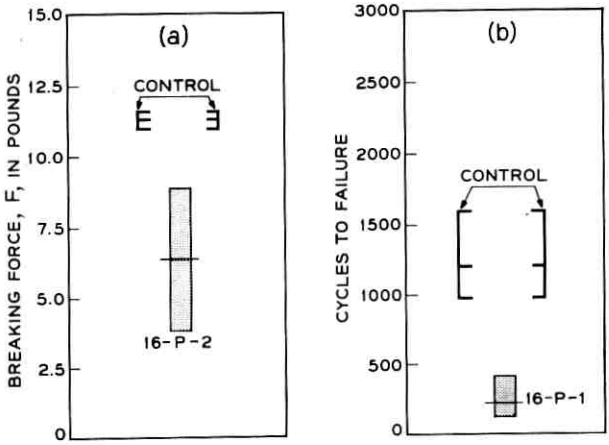


Fig. 19 — Inferior weld test, test number 16: (a) destructive strength — inferior welds show 55 per cent of the strength of good welds, based on the means of the distributions; (b) bending fatigue life (without insulation, lightly loaded bending) — inferior welds show 17 per cent of the bending fatigue life of good connections, based on the means of the distributions.

combined destructive test was used to determine the ultimate strength of the connections, and the fatigue method selected was of the lightly loaded bending type.

The welds used in this test showed only 55 per cent of the strength of good connections, and the fatigue life of the inferior type welds showed 17 per cent of the fatigue life of good connections. These results indicate that, for percussive welding, bending fatigue life is a much more sensitive indicator of weld quality than the strength of the connection as measured by the destructive combined test.

III. DISCUSSION

3.1 *Importance of Fatigue*

The measure of life expectancy chosen for this study was the fatigue life of the connections. The effect of the various nondestructive environments was measured by determining the loss of fatigue life caused by them. During testing, no significant electrical degradation was observed on any of the test connections as long as they remained mechanically secure. Comparison of the four types of connections was therefore made on the basis of their mechanical characteristics. Fatigue life was found to be the most important mechanical characteristic of permanent elec-

trical connections, as well as being one of the most sensitive indicators of connection degradation.

Fatigue life was measured in both vibration and bending. The vibration fatigue life was measured by the number of hours the connections survived a specified vibration schedule, and the bending fatigue life was measured by the number of cycles to failure of the connections in the lightly loaded and heavily loaded bending configurations. A comparison based on fatigue life offers an absolute comparison scale for all types of connections and facilitates the determination of an order of merit.

A given connection is better than another if, under identical environmental conditions, its life is longer. The life of a connection has ended when for any reason it fails to provide an adequate path for electrical current. In general, a connection may fail in two ways, electrically or mechanically. An electrical failure occurs when the connection develops an electrical characteristic such as a large constant or variable resistance which is incompatible with the equipment in which it is used. This type of failure is generally important only for pressure type connections; with properly designed and applied connections of this type it poses no serious threat to long life.

A mechanical connection failure occurs when the conducting path is physically broken. Most of the failures that are found in normal use fall into this group, and these are very small in number. Mechanical failures can be further divided into two groups, excessive force failures and fatigue failures. An excessive force failure is defined as a failure due to the application of a destructive force greater than that which a new, average, good connection can withstand. The probability of such a failure for good connections is directly related to the care with which the connections are handled and is generally very low.

A fatigue failure is defined as any mechanical failure in which the reduction in fatigue life was the primary cause of failure. If a connection is repeatedly stressed to a value below its breaking point, by definition it is being fatigued, and its time of ultimate failure will have been significantly influenced by its fatigue history. Thus most mechanical connection failures are due to fatigue, and since most connection failures are mechanical, it is apparent that most of the connections which fail in service do so because of fatigue.

The basic measure of quality of a permanent connection is its life. Any adverse environment to which the connection is subjected usually results in a reduction in either its electrical or mechanical life. There is no environment present in a typical central office which will cause a significant reduction in the electrical life of connections of the types and

qualities considered in this study so long as the connection remains mechanically secure. Two of the most severe environments in a central office are the small-amplitude vibrations due to equipment operation and the occasional bending of the connection during testing and wiring changes. These are both of a fatigue nature, and if the connection prematurely fails, it will probably be due to fatigue. Other environments such as temperature, corrosion, and humidity will in general not cause connection failure but will reduce the fatigue life, as shown in this study, and thus hasten the failure of connections. It follows, therefore, that most connection failures in service will be fatigue failures and that the basic measure of mechanical quality of a permanent connection is its fatigue life.

The results of this test program offer a number of arguments supporting the hypothesis that fatigue life is the most important mechanical characteristic of a good connection. These will be presented in the form of statements and then the supporting data and reasoning will be discussed.

(1) The ultimate mechanical strength of a connection is a poor measure of the fatigue life remaining in the connection.

Test 15 (Fig. 18) illustrates the above statement for percussive welds. In this test, connections were fatigued a predetermined number of cycles by the lightly loaded bending method and then the surviving connections were destructively tested. The results show that, at 100 per cent of the average fatigue life, the average strength of the surviving connections was approximately 95 per cent. In other words, when almost all of the fatigue life of the connection was expended, it showed only a 5 per cent loss in strength. The surviving connections were actually in bad shape, but a destructive strength test would have indicated hardly any degradation.

A second statement, closely associated with the first, is as follows:

(2) Loss in the ultimate mechanical strength of a connection is generally coincident with an even greater loss in its fatigue life.

Test 15 supports this statement. Further support is provided by the inferior weld test, Fig. 19. Inferior percussive welds were manufactured on purpose by using insufficient capacitance in the welding power supply, and the destructive strength and bending fatigue life distributions determined for these inferior welds. The results show that the destructive strength average dropped to 55 per cent of the good weld value and the bending fatigue average dropped to 17 per cent of the good weld value. Thus the fatigue life is approximately three times as sensitive an indication of inferior welds as destructive strength tests.

3.2 *General*

The most important factor affecting the fatigue life of a permanent connection is the manner in which the wire is brought from the terminal or its configuration. The configuration test, test 14, showed that when the 90° bend in the standard configuration was omitted, the fatigue life of all four types of connections was drastically reduced. The configuration used in test 14 may or may not have been the worst possible fatigue life configuration for the connections, but there can be no doubt that it is possible for the configuration to alter the fatigue life by approximately a factor of two. The results of the configuration test lead to the conclusion that the 90° bend in the standard configuration apparently partially isolates the connection from external wire movements and therefore improves the fatigue life of connections.

In general, the results of the shock tests indicate that all of the connections suffered some mechanical degradation from this environment.

The laboratory shock test (fatigue life) and both railroad shock tests show that in general it is the high fatigue life connections, i.e., those on the high end of the distribution, which are most affected by the shock environment.

Laboratory shock tests 3 and 4 afford one of the few opportunities for a direct comparison of the fatigue life and destructive strength test methods of measuring the degradation of the connections. The connections of both tests were exposed to the same shock environment and the degradation due to this environment was measured in two ways, (1) vibration fatigue life (Fig. 7) and (2) a destructive test (Fig. 8). Table V summarizes the results of these two tests and presents the degradation as a percentage of the control value.

Inspection of Table V indicates that for all four types of connections fatigue life is the more sensitive indicator for the measurement of shock degradation. This table thus further substantiates the statement that the

TABLE V — SHOCK DEGRADATION AS MEASURED BY FATIGUE LIFE AND DESTRUCTIVE TESTS (PERCENTAGE DEGRADATION DUE TO SHOCK)

Connection Type	Fatigue Life	Destructive Test
Soldered	23%	15%
Solderless wrap	24%	10%
Percussive weld	23%	1%
Resistance weld	22%	8%

mechanical strength of a connection is a poor measure of the fatigue life remaining in the connection.

The results of the central office and outside plant temperature tests, 7 and 8 respectively, showed, in general, a small but significant loss in fatigue life for all types of connections except for percussive welded connections. The degradation experienced by percussive welded connections from both temperature tests was greater than that of the other three types and amounted to approximately 30 per cent of the control fatigue life. Even this significant loss, however, did not prevent percussive welded connections from placing second in the order of merit.

3.2.1 *Composite Vibration Fatigue Life*

The vibration fatigue life data from all of the tests without insulation have been added together to form a composite or summary for each type connection. This summary is presented in Fig. 20 and Table VI; the table contains a list of the tests and the number of connections of each type which were included. The order of merit based on the vibration fatigue life for the connections without insulation is as follows:

1. percussive weld
2. soldered
3. solderless wrap
4. resistance weld.

A similar summary for the connections with insulation is presented in Fig. 21 and Table VII. It should be noted that the data scatter, as indicated by the length of the bar in the soldered and solderless wrap distributions, has increased, compared to the distribution without insulation, whereas the scatter of percussive and resistance welded data

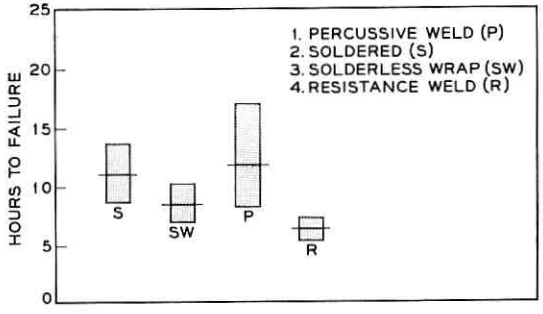


Fig. 20 — General wiring vibration fatigue life summary (without insulation); order of merit, based on vibration fatigue life.

TABLE VI — VIBRATION TEST DATA SUMMARY (WITHOUT INSULATION)

Test Description	Test Number	Number of Data Points From Each Test			
		Soldered	Solderless Wrap	Percussive Weld	Resistance Weld
Vibration	1	40	40	37	40
Temperature					
central office	7	20	20	20	20
outside plant	8	20	20	20	20
Corrosion					
3 months	9	10	10	10	10
6 months	10	10	10	10	10
Humidity	11	20	20	20	20
Totals		120	120	117	120

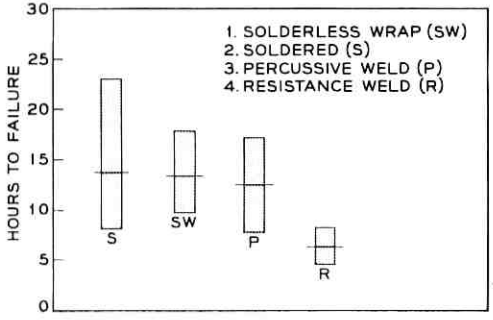


Fig. 21 — General wiring fatigue life summary (with insulation); order of merit, based on vibration fatigue life.

TABLE VII — VIBRATION TEST DATA SUMMARY (WITH INSULATION)

Test Description	Test Number	Number of Data Points From Each Test			
		Soldered	Solderless Wrap	Percussive Weld	Resistance Weld
Vibration	2	20	20	20	20
Configuration	14	20	20	20	20
Shock					
Laboratory	3	14	20	18	20
Railroad					
Noncushioned	5	10	10	10	10
Cushioned	6	10	10	10	10
Totals		74	80	78	80

has remained about the same. This is probably due to the insulation effect on the soldered and solderless wrapped connections. These connections were manufactured with the insulation in close proximity to the terminal, which is the normal wiring procedure; however, the exact location of the insulation was not accurately controlled and this resulted in a somewhat random insulation effect.

The order of merit based on the vibration fatigue life of the connections with the insulation is as follows:

1. solderless wrap
2. soldered
3. percussive weld
4. resistance weld.

It is the opinion of the author that the "without insulation" data is more important than the "with insulation" data, because the presence of insulation close to the connection cannot be depended upon in the case of soldered and solderless wrap connections. The worst case, without insulation, is therefore the most important. Inspection of the bar graphs in Fig. 20 and 21 shows that there is very little difference in the low end of the soldered and solderless wrap distributions, which indicates that the low fatigue life connections in the "with insulation" distribution were essentially without insulation. Since it is the low fatigue life connections which are most vulnerable to early failure, it follows that the most meaningful comparison should be based on the "without insulation" data.

If the presence of insulation could be assured, as in the case of the modified solderless wrap which has one turn of insulated wire as part of the connection, the fatigue life would undoubtedly be consistently higher and the "with insulation" data would be more meaningful.

The above two summaries were obtained by adding the fatigue life data of the tests involved to form a composite distribution. This procedure weights the final distribution according to the number of connections in each type of test, and this is certainly not the only way to treat the data. However, all four types of connections were treated equally, and so a comparison between them is meaningful.

3.2.2 *Orders of Merit*

An order of merit based on vibrational fatigue life statistical distributions was obtained from all environmental tests which compared the four types of connections and which used vibrational fatigue as a measure of degradation. It is also possible to obtain an order of merit for these tests from the number of stub connections of each type which survived the complete vibrational fatigue life determination. When one connec-

tion of a pair failed during the vibration test, the loop was clipped off, leaving the remaining good connection with a wire stub long enough to be destructively tested. Then the vibration test was continued until all pairs of connections had at least one failure. This additional vibration caused some of the stubs from the surviving connections to fail. Since all four types of connections initially had the same number of test connections, an order of merit can be obtained from the number of surviving stub connections by assigning the first position to the connection type which had the highest number of surviving stubs, second place to the type having the next highest number, and so forth. This procedure has meaning because all stub connections remained on the vibration machine until the end of the vibration period. Table VIII compares the orders of merit obtained from the number of stubs remaining with those obtained by consideration of the fatigue life distribution of the four types of connections.

In combining the orders of merit in Table VIII, it seems reasonable to assume that the soldered and percussive welded connections are of approximately equal merit and should share the number 1 position in an over-all vibrational order of merit as follows:

1. { soldered
   { percussive weld
3. solderless wrap
4. resistance weld.

It seems desirable to establish an over-all order of merit, shown in Table IX, for the four types of connection considered in the general wiring portion of this study. It is possible to do this because approximately equal numbers of all four types of connections were subjected to the various environments at the same time and under identical conditions. It should be remembered, however, that the best connection in the

TABLE VIII — ORDER OF MERIT COMPARISON

	Order of Merit							
	Based on Number of Surviving Stubs				Based on Fatigue Life			
	First	Second	Third	Fourth	First	Second	Third	Fourth
Percussive weld (P)	6	2	3	0	2	6	3	0
Soldered (S)	4	6	1	0	8	2	1	0
Solderless wrap (SW)	1	3	7	0	1	3	7	0
Resistance weld (R)	0	0	0	11	0	0	0	11

Numbers represent the number of firsts, seconds, thirds, and fourths for each rating method.

TABLE IX — OVER-ALL ORDER OF MERIT

Description	Number of Connections	Order of Merit			
		First	Second	Third	Fourth
Vibrational fatigue	1600	S&P	—	SW	R
Lightly loaded bending fatigue	160	P	SW	S	R
Heavily loaded bending fatigue	152	SW	P	S	R
Over-all order of merit based on fatigue	1912*	P	S	SW	R

Legend: S, soldered; SW, solderless wrap; P, percussive welded; R, resistance welded.
* Sum of above three groups of connections.

over-all order of merit was not necessarily the best under every environmental test, but was best only in an average sense.

Percussive welded connections were assigned first place because they shared first place with the soldered in the most important group of tests, vibrational fatigue (1600 test connections), and were also first in the lightly loaded bending tests. Soldered connections were assigned second place on the basis of their sharing first place in the vibrational fatigue tests. Resistance welded connections were assigned last place for obvious reasons, leaving third place for solderless wrapped connections.

It should be remembered that the soldered and solderless wrapped connections used in this study were of the same quality as those in wide-scale use in the telephone plant today. These connections have given and continue to give satisfactory service in the central office environment.

On the other hand, the percussive welded connections used in this study were of higher quality than those which have been used in special applications. This higher quality results from the use of the monitoring technique described in Ref. 1. These monitoring criteria are strongly recommended for all applications of percussive welding to assure the high quality of which the process is capable.

3.2.3 *Summary*

Table X, connection suitability as a function of environment (based on fatigue life), attempts to present the results of the general wiring portion of this study in a manner which would aid in the selection of a connection for a given environment. A rating number, based on the fatigue life of the connections, is assigned to each type of connection for

TABLE X — CONNECTION SUITABILITY AS A FUNCTION OF ENVIRONMENT (BASED ON FATIGUE LIFE)

Test Description	Test No.	Fatigue Method*	Without Insulation				With Insulation			
			Soldered	Solderless Wrap	Percussive Weld	Resistance Weld	Soldered	Solderless Wrap	Percussive Weld	Resistance Weld
Vibration test	1&2	VIB	7	5	8	4	10	8	8	4
Shock test laboratory	3	VIB					8	7	6	3
Shock test railroad noncushioned	5	VIB					9	7	6	3
Shock test railroad cushioned	6	VIB					9	7	7	3
Temperature test central office conditions	7	VIB	6	4	6	3				
Temperature test outside plant conditions	8	VIB	7	5	6	3				
Three months corrosion	11	VIB	7	5	8	4				
Six months corrosion	12	VIB	7	5	8	4				
Humidity tests	13	VIB	7	5	7	3				
Configuration test	14	VIB					4	6	4	2
Composite	—	VIB	7	5	8	4	7	7	6	3
Lightly loaded bending test	12	LLB					3	5	10	2
Heavily loaded bending test	13	HLB					8	10	8	5

* Fatigue method: VIB — vibration; LLB — lightly loaded bending; HLB — heavily loaded bending.

each test conducted. The table is divided into three general areas defined by the double lines. The top area is concerned with the vibration fatigue life both with and without insulation, the middle area is concerned with the lightly loaded bending fatigue method, and the last area at the bottom is concerned with the heavily loaded bending fatigue method. The rating numbers within any of the three areas are consistent among themselves. However, comparison of rating numbers from different areas has no meaning.

The rating system used is as follows: the number 10 was assigned to the highest value of fatigue life in a given area. The other rating numbers in the same area were generally assigned according to the percentage of fatigue life they had, compared to the highest value. Some adjustment of the rating numbers has been made to take into consideration the orders

of merit. In general, the rating numbers are closely associated with the fatigue life of the connections involved and offer a reasonably good index as to how well a connection will perform in a given environment.

Table X should be particularly valuable in selecting a connection for a given environment. The value of Table X is derived from the fact that, within a given fatigue area, valid cross comparisons can be made. Thus it is possible to get an idea of the degradation caused by shock on a percussive welded connection as compared to that caused by temperature or humidity or some other environment totally different from shock on, say, a soldered or solderless wrapped connection. These cross comparisons are possible because the common parameter chosen to measure degradation was fatigue life.

The "composite" listed under test description in the vibration fatigue area of Table X was obtained from the results of the vibration fatigue life summaries of Section 3.2.1.

The effect of insulation on the wire in the vicinity of a connection is apparent from the vibration portion of Table X. The percussive and resistance welded connections are relatively unaffected by the presence or absence of insulation. This result was expected, since the manufacture of these connections requires the insulation to be removed from the area near the welds. The soldered and solderless wrapped connections show a significant increase in their fatigue life when insulation is close to the connection.

Since improved vibration fatigue life is obtained when a soldered or solderless wrapped connection has insulation close to the terminal, it seems reasonable to inquire about the possibility of manufacturing these connections with the insulation always close to the terminal in order to take advantage of the increased fatigue life. It seems unlikely that this could be done for the soldered connections, because of the adverse effect of the heat on the wire insulation. In the case of the solderless wrapped connection, however, modified wrapping bits are available which place a turn of insulated wire around the terminal. The average vibration fatigue life of these modified wraps would undoubtedly be greater than either the "with" or "without" insulation connections tested in this program. It is the opinion of the author that the modified solderless wrapped connections could show enough improvement in their fatigue life to take over first place in the order of merit as opposed to third place without the modified wrap. In summary, it can be stated that a significant improvement in vibration fatigue life could be obtained on solderless wrapped connections by using a modified wrap.

It should be pointed out, however, that the need for this increased

fatigue life in normal Bell System field applications appears unnecessary in view of the fact that billions of solderless wrapped connections are in use in the telephone plant today and have given excellent service since their introduction ten years ago.

IV. CONCLUSIONS

4.1 *General*

The following three conclusions apply to all four types of electrical connections covered in this study. These connections were made with one size and type of terminal (0.025 by 0.062 inch nickel-silver) and with one type of wire (24-gauge solid copper); consequently, the conclusions may not hold for all types and sizes of terminals and wire.

(1) Assuming adequate electrical stability, a permanent electrical connection can be characterized by its fatigue life. Thus, if the fatigue life of a connection is well defined for various fatigue methods and the effect of adverse environments is determined on these fatigue lives, the mechanical quality of the connection has been established.

(2) A bend in the wire in the vicinity of the connection partially isolates it from external wire movements and significantly improves its fatigue life.

(3) The ultimate mechanical strength of a connection is a poor measure of the fatigue life of a connection.

4.2 *Specific*

(1) Using fatigue life as a basis of comparison and soldered connections as a reference standard, it is concluded that:

(a) monitored percussive welded connections are, in general, superior to soldered connections for the conditions which existed during this study,

(b) over-all, solderless wrapped connections are essentially equivalent to soldered connections, for the conditions which existed during this study, and

(c) resistance welded connections are significantly inferior to soldered connections for the conditions which existed during this study.

These conclusions represent over-all averages for all of the conditions tested. In some specific environments they may be interchanged or reversed. See Table X for details.

(2) Table X gives a good estimate of the comparative fatigue life which can be expected from the four types of connections under the various environmental test conditions.

(3) All four types of connections show loss in fatigue life and destructive strength due to repeated high level shocks.

(4) The shock and vibration experienced by all four types of connections when shipped by railroad express can result in loss of fatigue life.

(5) In general, all four types of connections show some loss in fatigue life due to the temperature tests environment. Percussive welded connections show the most significant loss.

(6) A significant improvement in the fatigue life of solderless wrapped connections can be expected through use of a modified wrap which places one turn of insulation around the terminal.

V. ACKNOWLEDGMENTS

The author gratefully acknowledges the general guidance and encouragement given by C. B. Brown and H. M. Knapp. Thanks are also due to J. C. Coyne for many valuable discussions and to J. J. Dunbar for assistance in the laboratory.

APPENDIX

Connection Manufacture and Quality Control

This appendix describes the destructive strength testing methods used for each type of connection. It also presents the data resulting from these tests in the form of statistical distributions.

A.1 *General Wiring*

All of the connections were made using one type of terminal and wire from two spools.

The terminal was of the solderless wrapped type shown in Fig. 22. It

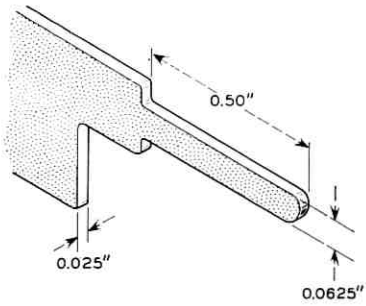


Fig. 22 — General wiring terminal; material is nickel-silver.

was made of nickel silver 0.025 inch thick and $\frac{1}{16}$ inch wide in the area of the connection. The soldered and solderless wrapped connections were made to this terminal in the conventional manner and the percussive and resistance welded connections were made on the $\frac{1}{16}$ inch wide side.

The wire used in all general wiring connections was standard switchboard wire, 24-gauge, solid, tinned copper wire with polyvinyl chloride insulation. Toward the end of the test sequence, the first spool of wire was exhausted and it was necessary to use another spool. The elongations of the two wires, measured over a 10-inch length, were as follows:

1. first spool — 16 to 19 per cent elongation
2. second spool — 15 to 18 per cent elongation.

A.1.1 *Soldered*

The soldered connections were manufactured in accordance with the present standards. They were actually of the wrapped and soldered type, since more uniformity could be obtained by wrapping three to three and one-half turns on the terminal with a wrapping tool and then soldering.

In the case of the soldered connections, there was no generally accepted method for determining the strength of a connection, so a testing procedure was devised. The testing procedure consisted of mounting the test terminal in a universal joint type of arrangement as shown in Fig. 23 and measuring the peak force required to pull the wire completely off the terminal.

This procedure for determining the strength of soldered connections was evaluated by using it to measure the strength of the extreme conditions of soldered connections: (1) a good connection and (2) an extremely poor soldered connection made by wrapping 3 to $3\frac{1}{2}$ turns around the terminal and applying no solder to the connection. Twenty-five connections of each type were tested, and the results are shown in Table XI.

A total of 1040 soldered connections was manufactured for this test program under conditions which assured high quality and uniformity. Half of these were used in determining their vulnerability to various environmental conditions, as described in the main body of the report,

TABLE XI — STRENGTH OF SOLDERED CONNECTIONS

Type Connection	Force Reading in Pounds		
	Maximum	Minimum	Mean
Good connection	7.5	4.8	5.76
Poor connection	3.5	2.6	3.04

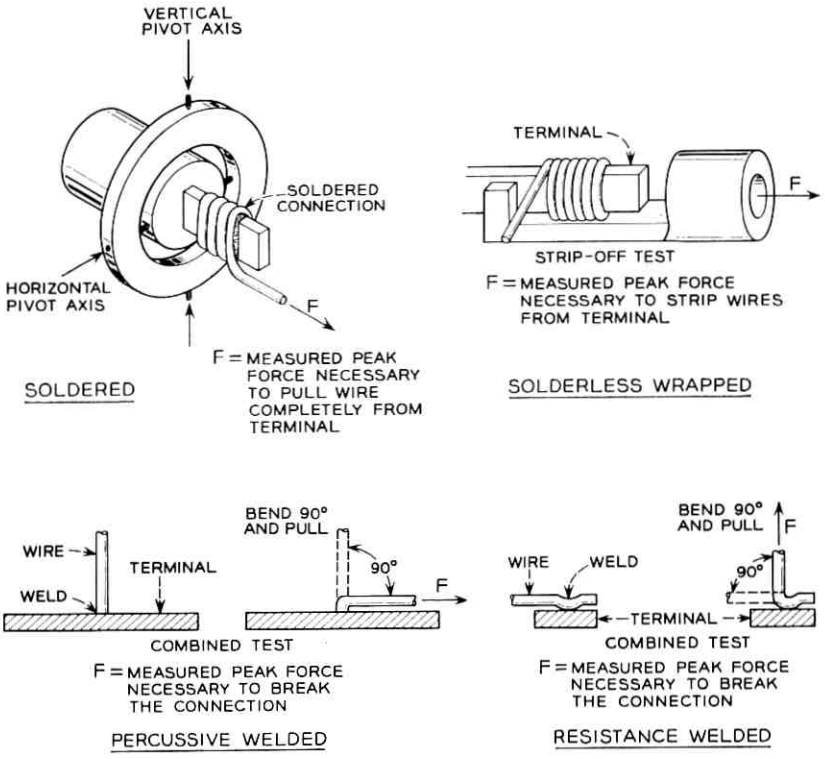


Fig. 23 — Destructive tests.

and the other half were destructively tested according to the procedure outlined above. The results of these destructive tests are presented in Fig. 24.

A.1.2 *Solderless Wrap*

The solderless wrapped connections were manufactured according to present standards, using qualified wrapping bits and wire of the proper elongation. The operator and wrapping bit effect on connection quality was minimized by having half the connections made by each of two operators, who in turn made half of their connections with one qualified bit and half with another. A total of 1040 connections was manufactured, half of which were destructively tested immediately after manufacture in order to determine the quality of the remaining connections, which were used in the comparison tests. The destructive tests were of two

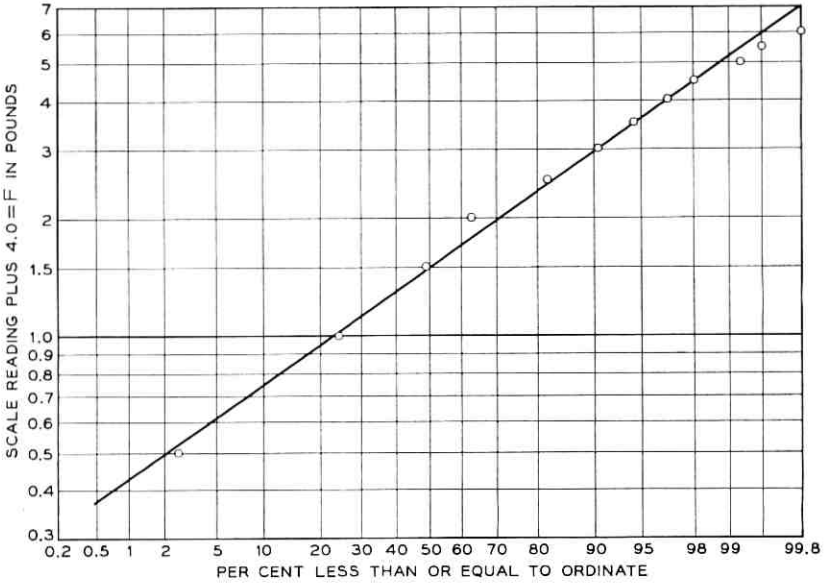


Fig. 24 — Soldered connection destructive strength distribution, general wiring; 590 data points.

types: (1) a standard strip-off test, shown in Fig. 23, which measures the maximum force required to strip the connection from its terminal and (2) a standard unwrap test, which requires that the wire shall be capable of being unwrapped completely from the terminal without breaking.

The distribution of strip-off values for the 260 connections destructively tested in this manner is shown in Fig. 25.

A.1.3 *Percussive Welded*

All of the general wiring percussive welds, with the exception of the inferior weld test, were made using the monitoring criteria developed by J. C. Coyne and reported in Ref. 1. These criteria resulted in welds of high quality, and a quantitative measure of this quality was obtained by destructively testing alternate welds (by order of manufacture). The destructive test used was the combined test, illustrated in Fig. 23, which consists of bending the vertical weld 90° to a horizontal position and determining the peak force necessary to break the connection. A total of 1540 welds (not counting the inferior weld test connections) was manufactured for this test program, and the destructive test data are presented in Fig. 26.

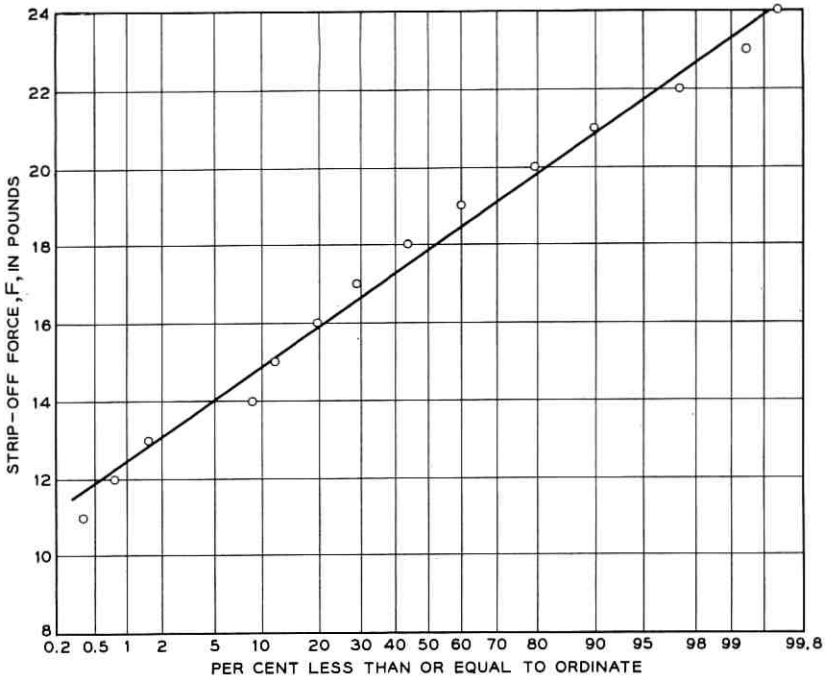


Fig. 25 — Solderless wrapped connection destructive strength distribution, general wiring; 260 data points.

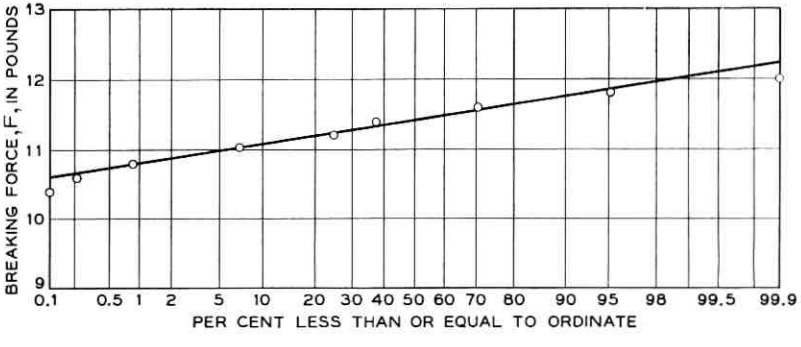


Fig. 26 — Percussive welded connection destructive strength distribution, general wiring; 940 data points.

A.1.4 *Resistance Welded*

Although resistance welding has been used for electrical connections for a number of years, no generally accepted quality requirements or standard method of measuring the weld strength could be found. The resistance welds used in this test program were the best which could be produced after a reasonable amount of experimentation with the process. The destructive strength test chosen was similar to that used for percussive welds — that is, the horizontal wire was bent 90° to a vertical position and the peak force necessary to break the connection was determined. This procedure is illustrated in Fig. 23. Some consideration was given to a test which consisted of pulling the wire horizontally along its original manufactured direction. This type of test yielded very little information about the weld, however, because the connections almost always broke in the wire, well away from the weld area. The destructive test chosen always broke at the weld and thus gave a much better indication of the weld strength.

The weld quality was evaluated by destructively testing alternate connections, by order of manufacture, and thus assuring the quality of the remaining connections for the environmental testing program. A total

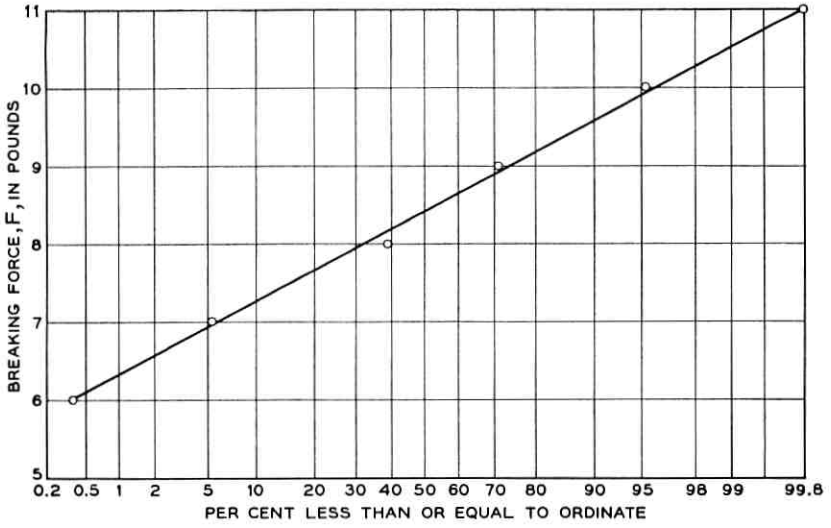


Fig. 27 — Resistance welded connection destructive strength distribution, general wiring; 520 data points.

of 1040 resistance welds was manufactured for this test program, and the destructive test data are presented in Fig. 27.

REFERENCE

Coyne, J. C., Monitoring the Percussive Welding Process for Attaching Wires to Terminals, B.S.T.J., **42**, Jan., 1963, p. 55.

Evaluation of Solar Cells by Means of Spectral Analysis

By H. K. GUMMEL and F. M. SMITS*

(Manuscript received February 20, 1964)

An approach toward testing of solar cells is outlined, and a test set and evaluation procedure of test results are described. Outer space short-circuit current is calculated from spectral response measurements performed on the cells. From this and additional measurements that determine the forward diode characteristic, the maximum obtainable power and the voltage at which maximum power is delivered are computed. The accuracy of outer space short-circuit current predictions is ±2 to 3 per cent when suitable standards are employed.

I. INTRODUCTION

The solar cell is a device which converts light energy into electrical energy. The electrical output of the cell, or response, depends on the spectral composition of the incident light. For space applications, determination of the response of solar cells to sunlight not filtered by the atmosphere is of great importance.

In characterization of the performance of a solar cell under outer space illumination, the most important parameter is the short-circuit current under such illumination. Once this current is known, the current-voltage output characteristic can be measured under a light source of arbitrary spectral distribution but of an intensity adjusted to produce the outer space short-circuit current.

This paper shows that it is practical to obtain the outer space short-circuit current from measurements of the spectral response of solar cells. Of the alternative approaches, one method relies on measurements under terrestrial sunlight while another method attempts to simulate the spectrum of the sun. Each of these two methods has practical difficulties that offset its inherent simplicity.

The direct sunlight measurements must be made outdoors unless

* Sandia Corporation, Albuquerque, New Mexico.

special buildings or sun tracking facilities are available and should be made at high mountain altitudes. Only a few hours around noon on clear days can be used for precision work. These requirements constitute a considerable difficulty, especially for laboratories located on the East Coast. Even if measurements are made at high altitudes, corrections must still be applied for the alteration of the spectrum and the reduction of over-all intensity by atmospheric absorption and scattering.

Direct solar simulation requires a light source which is constant in time and which has a spectral composition equivalent to that of the sun. Absolute spectral measurements of high accuracy are required for calibration and maintenance. This problem is eased, however, if calibrated solar cells are available against which the calibration of the simulator can be checked.

The only information available from a solar simulator, as well as from direct sunlight measurements, is the total current of a solar cell; detailed spectral response information is not obtained.

In the method which is described here, spectral response measurements are performed on the solar cells themselves. The spectral response at a given wavelength, multiplied by the sun's intensity at the same wavelength, gives the contribution to the short-circuit current at this particular wavelength. The total short-circuit current is obtained by integration of these contributions over all wavelengths. The spectrum of the sun is thus introduced only in calculations; the problem of building and maintaining a sun simulator is avoided. The accuracies with which the spectral response of the cells can be measured are, in principle, comparable to the accuracies with which the spectral distribution of the output of a solar simulator can be measured. However, if appropriately calibrated standard solar cells[1] are used, the accuracy of the outer space currents as determined by the test equipment can be considerably higher than that of the spectral measurements themselves.

An automatic testing facility based on such principles has been developed which, in addition to the spectral measurements, evaluates the current-voltage characteristic of the solar cell under test. The measurements are recorded on IBM cards to facilitate data handling. Evaluation of the measurements and any statistical analysis of the results is then easily done on an electronic computer.

This paper gives a description of the test equipment and the testing procedure. First, a description of the theory is given that relates spectral information and short-circuit current. There follows a detailed account of the design of the test set. The procedure for evaluation of the test results is described in the last section.

II. THEORY

In the derivation of the relation between the spectral response of solar cells and short-circuit current it is convenient to represent spectral response in terms of quantum efficiency, defined here as the number of electrons delivered into a short circuit per photon incident on the solar cell. This represents an over-all efficiency and includes the effects of light reflection at the surface, internal carrier-pair generation efficiency, and loss of carriers due to recombination.

If the quantum efficiency is known, the outer space short-circuit current, I_{scos}, can be calculated by integration over all wavelengths of the product of the incident outer space photon flux density[2] $\varphi(\lambda)$ and the quantum efficiency $Q(\lambda)$ of the cell

$$I_{scos} = Aq \int Q(\lambda)\varphi(\lambda)\, d\lambda \tag{1}$$

where A is the cell area and q the electronic charge. The quantum efficiency of a solar cell is a smoothly varying function of wavelength. For this reason it is adequate to sample the quantum efficiency only at a small number of discrete wavelengths. The integral can be approximated by a sum

$$I_{scos} = Aq \sum_i Q(\lambda_i)\varphi(\lambda_i)\Delta_i \tag{2}$$

where the photon flux entering the sum must be smoothed according to the intervals at which the quantum efficiency is sampled.

The λ_i are the wavelengths at which the quantum efficiency is measured, and the Δ_i are determined by the wavelength intervals between these points and the integration scheme used, e.g., trapezoidal approximation, Simpson's rule, Gaussian quadrature, etc.

The quantum efficiencies are measured as shown schematically in Fig. 1. Light from a tungsten light source is passed through interference filters which are mounted on a turntable. To eliminate the influence of fluctuations and drift in the light level, the ratio of the response of the cell to be measured to that of a monitor cell is taken. In terms of such ratios, R_i, the quantum efficiency of the sample cell can be expressed as

$$Q(\lambda) = R_i Q_{\text{ref}}(\lambda_i) \tag{3}$$

where $Q_{\text{ref}}(\lambda_i)$ is the quantum efficiency of the monitor cell. The sum (2) can thus be written:

$$I_{scos} = \sum_i R_i[AqQ_{\text{ref}}(\lambda_i)\varphi(\lambda_i)\Delta_i]. \tag{4}$$

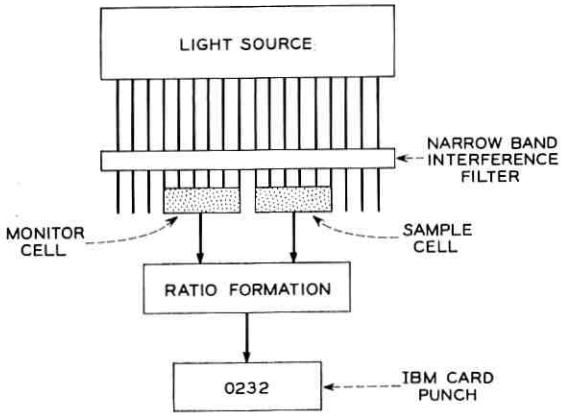


Fig. 1 — Method of measuring ratio of quantum efficiencies.

Defining the quantities in the brackets as "weighting factors"

$$W_i = A q Q_{\text{ref}}(\lambda_i)\varphi(\lambda_i)\Delta_i\,, \tag{5}$$

one obtains for the outer space short-circuit current

$$I_{scos} = \sum_i R_i W_i\,. \tag{6}$$

The weighting factors include the spectral composition of outer space sunlight and the quantum efficiency of the monitor cell. In the derivation, outer space sunlight served only as an example. Clearly, the method is not restricted to outer space illumination but can be applied, with appropriate weighting factors, for any illumination of fixed spectral composition and intensity. Practical schemes for obtaining the weighting factors in connection with the calibration of a set of solar cells will be discussed below.

As a preliminary step in obtaining weighting factors for outer space current prediction, a set of weighting factors referred to as terrestrial weights was determined for the short-circuit current of solar cells measured on a clear day near noon at a high altitude. The set of solar cells which were to be calibrated as standards had a wide range in spectral response, as shown in Fig. 2. It included cells bombarded with various fluxes — up to $2 \times 10^{16}/\text{cm}^2$ — of 1-Mev electrons and virgin cells.

To find the terrestrial weighting factors, first the quantum efficiency of the monitor cell at the various wavelengths is measured on a relative scale by comparison of the monitor cell output with that of a spectrally

flat detector, such as a thermocouple. Combining this information with relative spectral intensities of the sun under which the observation was taken, one obtains the short-circuit current to within a common constant factor for all cells. This factor is readily determined from the actual measurements of the short-circuit currents. It is significant that only

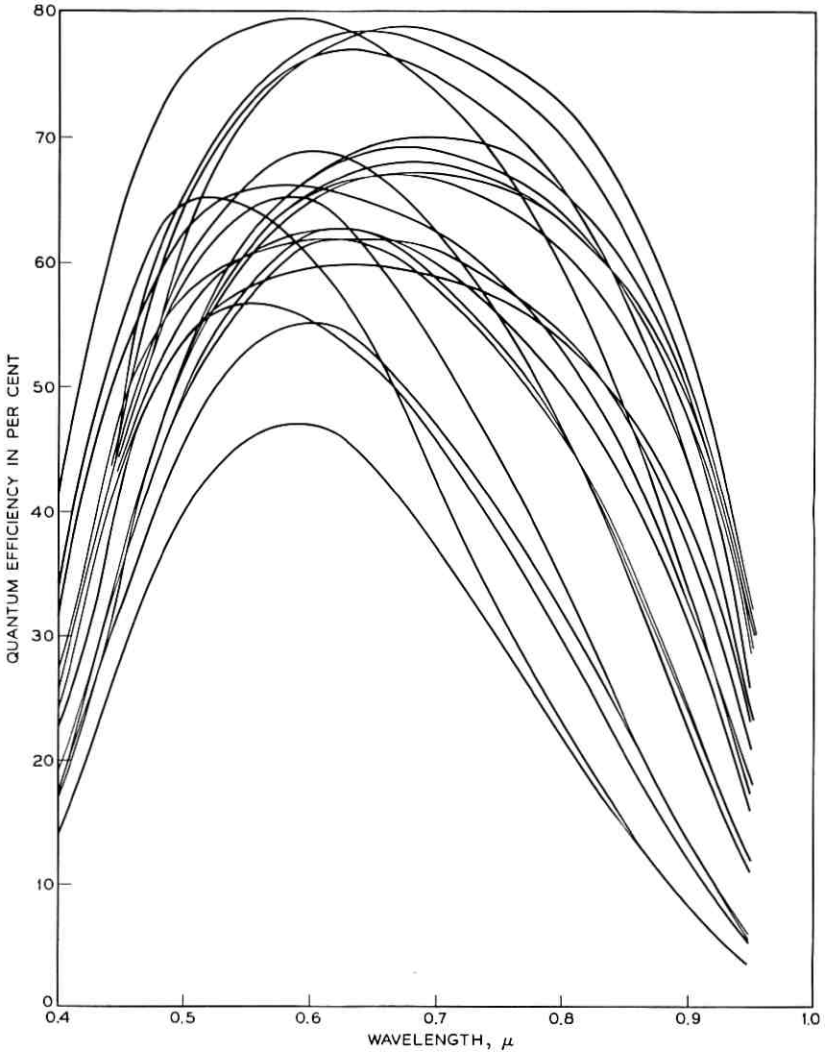


Fig. 2 — Quantum efficiencies of set of cells used as standards.

the current measurements determine the absolute scale, while all other measurements need to be performed on a relative scale.

However, if the set of solar cells used in such measurements has a sufficiently wide spread in spectral response characteristics, it is possible to find the weighting factors even without any prior information on the quantum efficiency calibration and on the spectral composition of the sun. The appropriate weights are simply those which reproduce the current of all cells with the least error.

In the present work, the terrestrial weighting factors were determined by a judicious combination of the two methods. Fig. 3 shows the resulting comparison of the measured short-circuit current and the calculated current. In this figure the most heavily prebombarded cells appear at the lower left-hand corner.

To convert the terrestrial weighting factors into outer space weighting factors, it is necessary to increase the weighting factors at each wavelength by the ratio of outer space solar intensity to terrestrial solar intensity at this particular wavelength. Such information is readily available from solar spectral recordings performed by the Smithsonian Institution.[1] For the set of solar cells shown in Fig. 2 an outer space

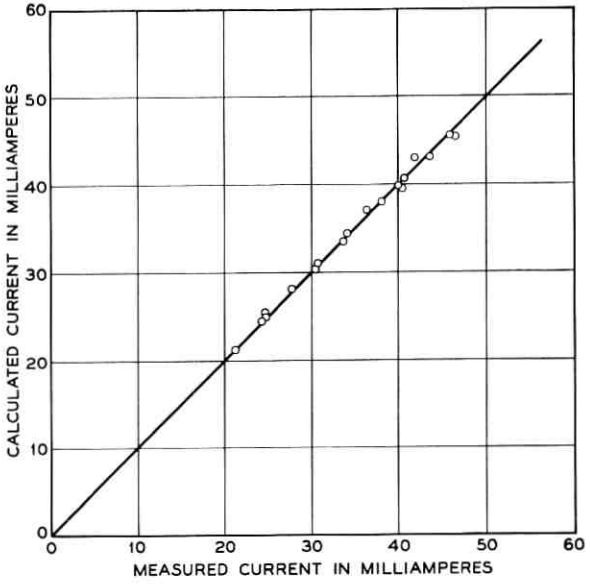


Fig. 3 — Comparison of directly measured current with current calculated from spectral information.

extrapolation was performed by this procedure. Thus, weighting factors were determined for the prediction of outer space short-circuit currents from the spectral measurements performed on the test set.

The calibrated cells serve as standards for the weighting factors. Thus, the monitor cell in the test equipment need not provide the long-time standard. It serves only as a short-time standard that can be compared readily with the calibrated cells.

For further characterization of solar cells, additional measurements under tungsten light filtered through heat-absorbing glasses, later referred to as "white light," are performed. These are the short-circuit current, the open-circuit voltage, the current delivered into a 10-ohm load, and the current delivered into 0.45 volts. The reverse leakage and the forward voltage with 50 ma passing through the solar cell are measured in the dark.

In a special mode of operation the test equipment can form the weighted sum, thus giving the short-circuit current directly. In general, however, it is preferred to evaluate all the test results on a digital computer.

The obtained outer space short-circuit current is then used in conjunction with the measurements under the white light to calculate the over-all output characteristic for outer space illumination, in particular the maximum-power point and the voltage at maximum power.

In addition, the short-circuit current under the white light source is also calculated from the spectral response with an appropriate set of weighting factors. The percentage difference between the calculated and the measured short-circuit current is determined and recorded. Generally, this difference is small, indicating that the outer space short-circuit current calculation can be trusted. Occasionally, solar cells are observed in which the calculated current deviates substantially from the measured current. This indicates that the particular solar cell has a nonlinear response, e.g., its quantum efficiency is light level dependent. The results for the outer space short-circuit current obtained on such solar cells are correspondingly in error. Alternative methods of calculation that are applicable to such nonlinear cells are described in Section IV.

III. TEST EQUIPMENT

As mentioned, two groups of measurements are performed in the test set. One group of eight measurements evaluates the spectral response of the cells, while the measurements in the other group evaluate the current-voltage characteristic of the solar cell.

For those tests in which the output is porportional to the incident light intensity, the ratio of the response of the sample cell to that of a reference cell is measured. This eliminates the effect of drift and fluctuations in the light source and the interference filters.

The set consists of the following major components:

1. optical system comprising light source and interference filters,
2. mechanical system providing for optical filter transport and actuation of control switches,
3. electronics and ratio-formation,
4. control system, giving control commands for operation of counters and punch-out, and providing for the routing of the signals through appropriate amplifiers and attenuators, etc., and
5. output channels: printer or translator and card punch.

3.1 *Optical and Mechanical System*

The optical system is shown schematically in Fig. 4. A 1-kw incandescent projection lamp (General Electric type PH/1M/T20MP) serves as the light source. The light is focused by a spherical mirror (diameter 12 inches, focal length $5\frac{3}{4}$ inches) and directed by a plane mirror towards the sample and reference cells, which are placed adjacent to each other in a light-tight box. To insure uniform illumination of both cells, diffusing glass is interposed into the light beam. Long-wavelength light is attenuated by heat-absorbing glass and nearly monochromatic light at various wavelengths is selected by narrow-band interference filters. These filters are mounted on a disk that is driven intermittently by a Geneva motion and interposes the filters sequentially into the optical path. The filters have transmission bandwidths of the order 0.01 μ. For the evaluation of silicon cells, filters at the following wavelengths are used: 0.4, 0.45, 0.5, 0.6, 0.7, 0.8, 0.9, and 0.95 μ. For measurements on the solar cells in the dark, four positions are blocked; the remaining four positions admit white light for measurements of additional solar cell characteristics. The white light intensity is adjusted to generate in a typical solar cell a short-circuit current that is comparable to the short-circuit current expected in outer space.

The important factors of the mechanical system are also shown in Fig. 4. A 1-rps motor drives the Geneva motion, which advances the filter disk. Thus the time available for any individual test is 1 sec. The motor axle contains a set of adjustable cams that actuate microswitches for the timing of various functions within a single test cycle. The wipers of a set of 16-position rotary switch decks are driven by the filter disk axle to permit the selection of the different types of tests to be performed.

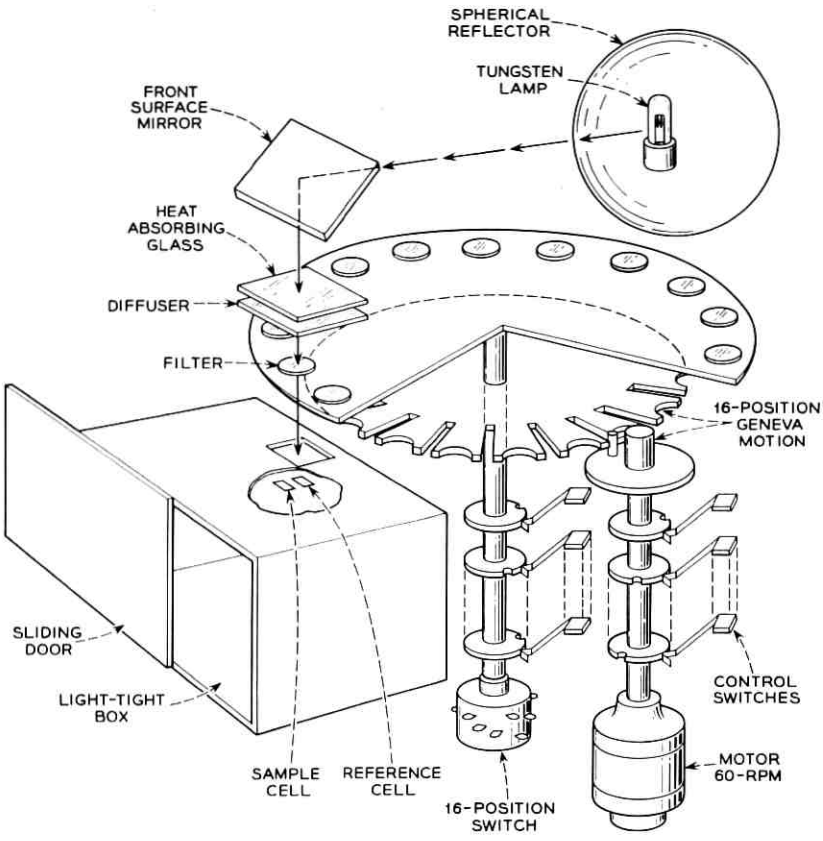


Fig. 4 — Optical and mechanical system.

3.2 *Electronics and Ratio Formation*

The various signals encountered are measured by a combination of voltage-to-frequency converter (Dymec 1122B) and counter. This combination, when used in both the sample and reference channel, makes fast and accurate ratio formation possible.

If the input voltage at the sample channel is V_s volts (see Fig. 5), and if the voltage-to-frequency converter (vfc) gives A cycles per second per volt, the total count N_s obtained in t seconds is

$$N_s = V_s A t. \tag{7}$$

Similarly one obtains in the reference channel

$$N_r = V_r A t. \tag{8}$$

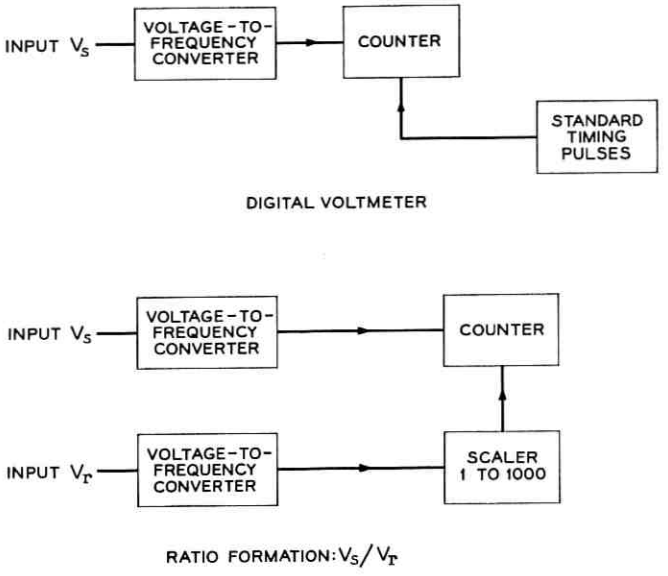


Fig. 5 — Method of ratio formation.

To form the ratio V_s/V_r the reference counter is used as a scaler, in such a way that it gives an output pulse for every 1000 input pulses. The time between output pulses is then

$$t = \frac{1000}{AV_r} . \tag{9}$$

The output pulses from the scaler gate the counter in the sample channel in such a way that one pulse turns it on and the next one turns it off. Thus t in (7) is identical to that in (9). This leads to

$$N_s = 1000\ (V_s/V_r) \tag{10}$$

i.e., the count in the sample channel counter gives the desired ratio of the voltages with adequate resolution.

The voltage-to-frequency converter and counter combination, when used as a voltmeter, gives an output that represents the input voltage averaged over the time of measurement. Thus fluctuations are smoothed out. The effective noise bandwidth is therefore the reciprocal of the time of measurements. In spite of the small noise bandwidth, however, this combination has a fast transient response. Thus, high-resolution, low-noise measurements during short time intervals are possible. With a

wideband amplifier (Kintel type 112) preceding the voltage-to-frequency converter, the equivalent noise at the input that actually has been observed is of the order of 1 μv for 0.1-second measurement intervals even after an estimated 750 hours of operation of the system. The chopper at the input of the preamplifier deteriorates with time; for a new chopper the noise may be considerably below 1 μv.

The signals to be measured are in the range 0.5 mv to 10 mv except for the spectral response measurement at the shortest wavelength, 0.4 μ, where the output of the light source is low and the signal is about 100 μv. For all other wavelengths the amplifier noise thus contributes less than 0.2 per cent to the measurement error. Since the contribution of the 0.4 μ spectral region to the total short-circuit current is small, the error contributed by the amplifier to the total short-circuit current is also below 0.2 per cent.

The short-circuit current is obtained from the measured voltage drop across a shunting resistor. The resistor has to be of sufficiently low value that the voltage developed across it would result in a forward current in the dark that is negligible compared to the short-circuit current. For the spectral measurements, a resistor of 10 ohms is adequate, while for the white light measurements a 0.1-ohm resistor is used. The switching of resistors at the low signal levels has to be done carefully to prevent introduction of spurious thermal emf's. Mercury relays are employed in an arrangement shown in Fig. 6. For the measurement of the open-circuit voltage ($\approx\frac{1}{2}$ volt) and other voltages of comparable magnitude, the preamplifier is bypassed and the signal is fed directly into the voltage-to-frequency converter.

3.3 *The Control System*

The programming of the various tests is accomplished through the 16-position switch coupled to the axis of the drive motor. Microswitches control reset commands at the beginning of each test; approximately in the middle of each test, (after the amplifier transients have died down) a new counting cycle is initiated. The last operation within each test is the print command to the printer or IBM punch.

The switching of the signal paths for the various tests is accomplished by relays which are activated from the 16-position switch through a diode matrix as shown in Fig. 7.

The following quantities are measured in the tests: in the eight spectral readings, the ratio of the short-circuit currents of sample cell and reference cell at eight wavelengths,

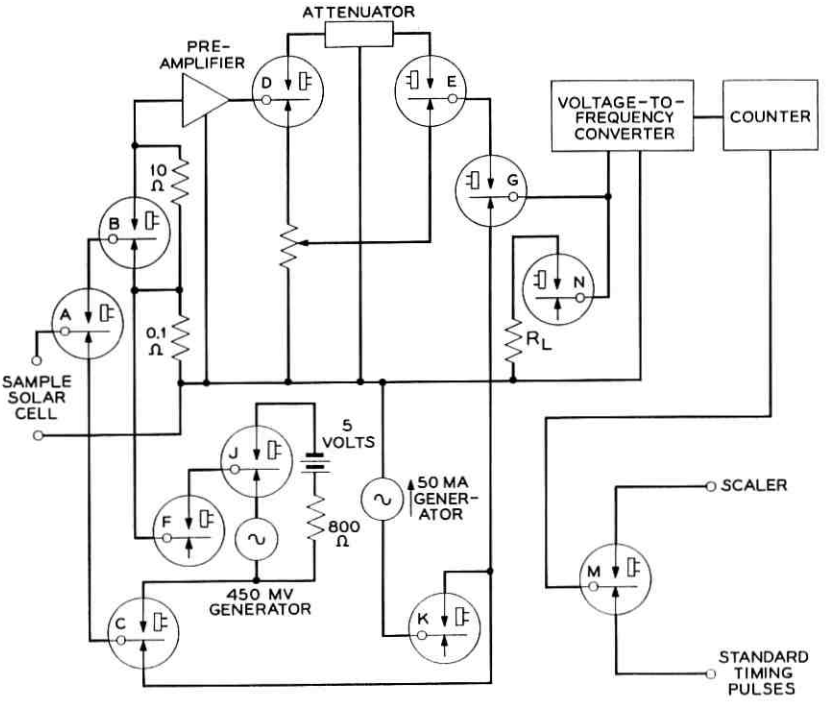


Fig. 6 — Switching of signal paths.

I_{rev}, the current flowing through the cell when a reverse voltage of 5 v is applied through an 800-ohm resistor,

V_L, the voltage developed across a load resistor R_L ($\approx$10 ohms) when the cell is illuminated by white light (see optical system),

V_f, the voltage across the cell when a forward current I_f, normally 50 ma, is passed through the cell in the dark,

I_{sc}, the short-circuit current measured directly under white light.

I_{scR}, the ratio of the short-circuit current of the sample cell under white light to that of the reference cell (In this measurement the amplifier gains are set such that for proper adjustment of the light source the numerical value of I_{scR} is the same as that of I_{sc}.),

I_{45}, the current delivered by the cell under white light illumination into a 0.45-volt voltage source, and

V_{oc}, the open-circuit voltage of the cell under white light illumination.

The short-circuit current is measured both directly (I_{sc}) and normalized (I_{scR}). In the linearity test, the normalized current is to be used, as it is compared with the normalized spectral readings. In the evalua-

tion of the output characteristics, the unnormalized short circuit current is used in conjunction with the open-circuit voltage and the voltage across a load resistor.

IV. EVALUATION OF THE SHORT-CIRCUIT CURRENT

The short-circuit current is evaluated from the raw data according to (6). This evaluation can be performed either directly by the test equipment concurrently with the measurement of the R_i or by a separate calculation, preferably on an electronic computer.

For the evaluation in the test set the counter is not reset between the spectral response measurements R_i. The multiplication by W_i is carried out by interposing attenuators between the preamplifier and voltage-to-frequency converter of the sample channel. Such attenuators are switched by the 16-position switch, which is coupled with the axle of the filter disk. The number added into the counter at each filter position equals the current contributed by the wavelength band represented by the filter, and the final number on the counter gives the total short-circuit current.

The direct test set evaluation of the short-circuit current is of importance if no computing facilities are available or if results are needed

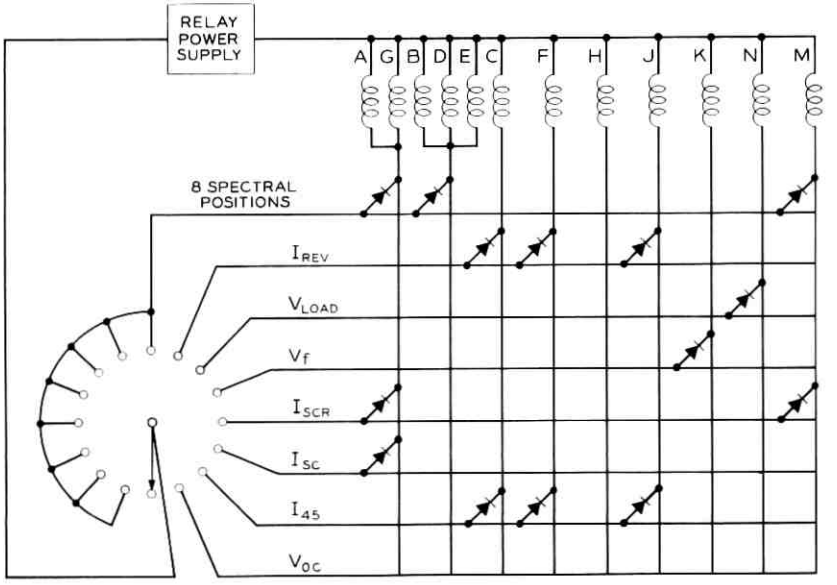


Fig. 7 — Diode matrix for activation of relays.

immediately. On most other occasions, it is preferable to have the computer process the raw R_i data, since one set of data can then be analyzed for cell response under various light sources.

In addition to providing short-circuit current under outer space illumination, the computer program, as used for routine evaluation of solar cells at Bell Telephone Laboratories, evaluates the short-circuit current for the white light that is used in some of the tests of the output characteristic, as discussed in Section II. For a check of consistency, the program then computes the percentage deviation of this current from the short-circuit current measured directly under white light. This check is important for the detection of cells whose short-circuit current changes nonlinearly with light intensity.

As an additional feature, the program computes the ratio of the computed outer space current to the computed white light current. This quantity, CC, is of interest for a number of reasons. If the maximum intensity of the white light falls at a different wavelength than the maximum intensity of outer space sunlight, then CC provides a "color index" for the cells. In the test set described here the white light is deficient in infrared and red light due to filtering with heat-absorbing glass. Therefore, red-sensitive cells have large CC numbers while blue-sensitive ones have smaller CC numbers.

In the standard procedure for calculating the outer space short-circuit current, the measured R_i values are multiplied by proper weighting factors W_i and summed as in (6). Fluctuations in intensity and spectral composition of the light source cause no first-order error, since the R_i are ratios. However, if the cell current varies nonlinearly with light intensity, the current at full solar illumination, as calculated from spectral measurements at low light intensity, is in error. From the comparison of measured white light response and calculated white light response one knows whether or not a cell is linear. Nonlinear cells are encountered rarely enough so that they present no serious problem. Nevertheless, the program can be used — if required — in such a way that errors due to cell nonlinearity are minimized; however, errors will be introduced if the spectral content of the white light changes. The procedure is to multiply the measured white light current I_{scR} by the factor CC. Since nonlinearities cause errors in both numerator and denominator, the influence is reduced to second-order effects.

V. THE OUTPUT CHARACTERISTIC

As discussed before in connection with the short-circuit current for outer space illumination, the output characteristic of the solar cell could

be established under a light source of arbitrary spectral composition but adjusted in intensity to produce the known outer space short-circuit current. To do this, however, would be a very tedious and time-consuming procedure. As an alternative, one measures the output characteristic under a light source that generates a short-circuit current close to, but not necessarily equal to, the outer space short-circuit current. It then is possible to calculate the outer space response from these measurements and the known outer space short-circuit current.

For such calculations of the output characteristic the following model of a solar cell is used:

$$I = I_{sc} - I_o \left\{ \exp \left[\frac{q(V + IR_s)}{nkT} \right] - 1 \right\}. \tag{11}$$

Here I is the output current at voltage V, I_{sc} is the light induced current (under the given illumination) and corresponds to a current generator in parallel with the solar cell diode. R_s is a lumped resistance approximating the contact resistance and sheet resistance in the front layer of the cell. I_o, the diode forward-current constant, is orders of magnitude smaller than I_{sc}; therefore the term (-1) in (11) can be neglected. With this simplification one can express I_o in terms of the open-circuit voltage, V_{oc}, i.e. that voltage for which the left side of (11) is zero, which leads to

$$I = I_{sc} \left[1 - \exp \frac{q(V + R_s I - V_{oc})}{nkT} \right]. \tag{12}$$

An ideal diode would have a value of 1 for the coefficient n in the exponent. A number of effects* cause deviations from this simple behavior, and frequently the output curve cannot be fitted by a constant value of n over the entire voltage range and for different light intensities. However, one may fit the output characteristic over a limited range near the maximum power point with a constant n for a particular light intensity. Only small errors are introduced if one uses the same value of n near the maximum power point under slightly different illumination.

Similarly the value of the resistance R_s in (12) is not strictly a constant, since it involves a spreading resistance in the thin diffused layer. Nevertheless, in general R_s will be a very weak function of current and light level so that, for this calculation, it can be treated as a constant.

With these assumptions the output characteristic under outer space illumination is determined by the four quantities I_{sc}, V_{oc}, n, and R_s, which must be established by measurement. The first two of these de-

* Among these effects are space-charge recombination and sheet resistance. See also Ref. 3.

pend on the particular light level. From the measurements of I_{sc}, V_{oc}, and V_L under the white light, and from the measurement of V_f in the dark, all four quantities can be extracted. By replacing I_{sc} in (12) by the known outer space short-circuit current I_{scos} and setting $I = 0$, one obtains the open-circuit voltage V_{ocos} for outer space illumination. The parameters I_{scos}, V_{ocos}, n, and R_s are now used to characterize the output characteristic of the cell under outer space illumination. This permits a calculation of all quantities of interest according to (12), in particular the maximum power point. An approach for performing these calculations on an electronic computer is given in the Appendix.

Most of the measurements on the test are reproducible over periods of hours with an rms deviation from the mean of about 0.5 per cent. If calibrations on the test set are to be made to an accuracy approaching 0.5 per cent or less, the fine adjustments become very tedious, as the criterion has to be the result of a statistical analysis of several measurements.

A preferable procedure is to make adjustments on the test set to a moderate accuracy and to have the computer apply fine corrections. Two modes of operation are used. In one mode these corrections are applied manually as parameters that are entered with the data. In the other mode, a set of standard cells is measured along with the cells to be evaluated. The data cards of the standard cells also contain their calibrated outer space short-circuit current value. The computer can then determine the percentage deviation of the calculated value of the outer space short-circuit current from the calibration value for each of the standard cells. The calculated values of I_{scos} for the cells under test are then corrected on the basis of the average of these deviations. A similar correction is applied to the white light current.

With the latter mode of operation, the long time standard is provided by a group of standard cells, and the built-in monitor cell serves only as a short-time reference. If there are no uncertainties introduced by long-time drifts of the standards, absolute accuracies of outer space short-circuit current predictions of 2 to 3 per cent should be realizable. Comparison of preflight predictions with flight data on the Telstar satellites and the Anna 1B satellite[4] confirm that an accuracy of 3 per cent or better has been achieved.

VI. ACKNOWLEDGMENTS

The authors acknowledge the contributions of P. J. Kamps, who did the mechanical design, and of W. G. Ansley, who provided the digital recording system.

APPENDIX

A.1 *Method of Calculation of n and R_s*

The quantity n is determined by fitting the output characteristic between V_L and V_{oc}. As one has to make allowances for R_s, the series resistance, an iterative scheme is convenient. Assume for a moment that the series resistance R_s is zero. One can then compute the value of n or nkT/q from (12)

$$\frac{nkT}{q} = \frac{V_{oc} - V_L}{\ln\left(1 - \frac{V_L}{R_L I_{sc}}\right)}. \tag{13}$$

Using this value of nkT/q one can compute the value of the voltage across the cell for a current $I_{sc} - I_f$ through the illuminated cell

$$V_c = V_{oc} + \frac{nkT}{q} \ln\left(\frac{I_f}{I_{sc}}\right). \tag{14}$$

If there were no series resistance, then the measured voltage V_f should be equal to V_c. Because of the series resistance, the voltage V_f is higher by $R_s \cdot I_f$. One thus obtains an initial estimate of the resistance

$$R_s = \frac{V_f - V_c}{I_f}. \tag{15}$$

Now this value of R_s is used for an improvement of (13). During the measurement of V_L, current is flowing out of the cell, and the junction voltage is the measured voltage V_L increased by the voltage drop $(V_L/R_L)/R_s$. Thus one obtains

$$\frac{nkT}{q} = \frac{V_{oc} - V_L\left(1 + \frac{R_s}{R_L}\right)}{\ln\left(1 - \frac{V_L}{R_L I_S}\right)}. \tag{16}$$

The new value of nkT/q thus obtained can be used to compute an improved value of V_c, and in turn, of R_s, and so on. The convergence of this procedure is quite rapid, since nkT/q in (14) is multiplied by $\ln (I_f/I_{sc})$ and I_f was chosen to be near I_{sc}.

A.2 *Method of Calculation of Outer Space Quantities*

Using the quantities I_{sc}, V_{oc}, (nkT/q) and R_s, as determined in the previous section, the outer space quantities are evaluated as follows.

Outer space open-circuit voltage:

$$V_{ocos} = V_{oc} + \frac{nkT}{q} \ln \frac{I_{scos}}{I_{sc}}. \tag{17}$$

For further evaluation of the characteristics it is convenient to introduce the following normalizations:

$$y = \frac{qV}{nkT} \tag{18}$$

$$u = \frac{I}{I_{scos}}$$

$$x = \frac{I_{scos}}{I_{scos} - I} = \frac{1}{1 - u} \tag{19}$$

$$z = \frac{qV_{ocos}}{nkT} \tag{20}$$

$$\beta = \frac{qR_sI_{scos}}{nkT}. \tag{21}$$

The normalized current u at the normalized voltage y is obtained from a solution of the equation

$$u = 1 - e^{y+\beta u-z}. \tag{22}$$

In the computer program, this equation is again solved by iteration. The power at the voltage corresponding to y is now given by

$$P = (nkT/q)I_{scos}yu. \tag{23}$$

To obtain the maximum power, (22) is rewritten so that the normalized voltage appears as a function of the normalized current

$$y = z - \beta u + \ln (1 - u). \tag{24}$$

The normalized power π is obtained by multiplying (24) by u

$$\pi = u[z - \beta u + \ln (1 - u)]. \tag{25}$$

The maximum is obtained by equating the derivative to zero. This yields the condition

$$z = \frac{u}{1 - u} + 2\beta u - \ln (1 - u). \tag{26}$$

In (25), u is required as a function of z and β. Again, on the computer a solution is conveniently found by iteration. To get (25) into a form

suitable for this, a switch is made to the variable x, defined in equation (19). Then

$$x = z + 1 - 2\beta\left(1 - \frac{1}{x}\right) - \ln x \tag{27}$$

and in this form the equation can be solved by iteration, starting with $x = 1$ on the right-hand side. One now obtains for the maximum power

$$P_{\max} = I_{scos}V_{ocos}\frac{(x-1)^2}{x}\left(1 + \frac{\beta}{x}\right), \tag{28}$$

and for the voltage at which maximum power is delivered

$$V_{mp} = \frac{nkT}{q}(x-1)\left(1 + \frac{\beta}{x}\right). \tag{29}$$

REFERENCES

1. Gummel, H. K., Smits, F. M., and Froiland, A. R., A Method for Terrestrial Determination of Solar Cell Short Circuit Current under Outer Space Solar Illumination, Wescon 1961, Paper 7/3.
2. In this work the spectral solar irradiance data of Johnson, F. S., J. Meteorology, **11**, 1954, p. 431, were used.
3. Queisser, H. J., Forward Characteristics and Efficiencies of Silicon Solar Cells, Solid State Electronics, **5**, No. 1, 1962, and Gettering Effects on the Forward Characteristics of Silicon Solar Cells, Proc. I.R.E., **50**, 1962, p. 486.
4. The authors are indebted to Dr. R. E. Fischell of the Johns Hopkins University, Applied Physics Laboratory, for flight performance data on experimental cells on which preflight measurements were made on the test set described here.

A Theory of a Unilateral Parametric Amplifier Using Two Diodes

By J. HAMASAKI

(Manuscript received December 18, 1963)

This paper describes the theory of a unilateral parametric amplifier which contains two variable-capacitance diodes separated by a quarter wavelength at the signal frequency and a half wavelength at the idler frequency. It is shown that a broadband signal circuit is essential in order to obtain unilateral gain, and that matching conditions are obtainable even with a high gain. The optimum noise figure is slightly worse than that of a single-diode reflection-type amplifier; however, this amplifier has advantages if it is refrigerated at liquid helium temperature, because it does not require a circulator in front of its input port. The amplifier usually requires an isolator at its output port, since it does not have substantial loss in the reverse direction.

I. INTRODUCTION

The performance of a single-diode reflection-type parametric amplifier is often limited by the availability of a good circulator, which is essential to a practical amplifier. This becomes a more serious problem when the amplifier is to be refrigerated down to liquid helium temperature, since satisfactory circulator performance is then difficult to obtain.

A unilateral parametric amplifier with two diodes, originally proposed by Baldwin,[1] does not require any circulator or isolator in front of the amplifier if the signal source impedance is reasonably well matched. Therefore, this amplifier might avoid the difficult circulator problems.

This paper is prepared to show the theoretical characteristics of a unilateral parametric amplifier with two diodes separated by a quarter wavelength at the signal frequency. In Section II, an exact expression for the scattering matrix of a simplified model of the amplifier is obtained. In Sections III and IV, expressions for power gain, noise figure and bandwidth are calculated. In Sections V and VI, the optimum noise figure and some design considerations of the amplifier are described.

II. PRESENTATION OF THE SCATTERING MATRIX OF THE AMPLIFIER

Fig. 1 shows a basic configuration of the unilateral parametric amplifier. The black box shown in Fig. 1 is a symmetrical lossless reciprocal two-terminal-pair network whose image impedance Z and phase constant θ are assumed to be real quantities. Stationary susceptances of the diodes are considered as a part of the black box, and losses in the diodes are included in the external loads. C and C' in Fig. 1 represent the sinusoidally varying shunt capacitances which play an essential role in the mechanism of amplification. Parameters concerned with the right-side arm of the black box are indicated by primed letters, and those of signal and idler frequencies are indicated by suffixes 1 and 2 respectively.

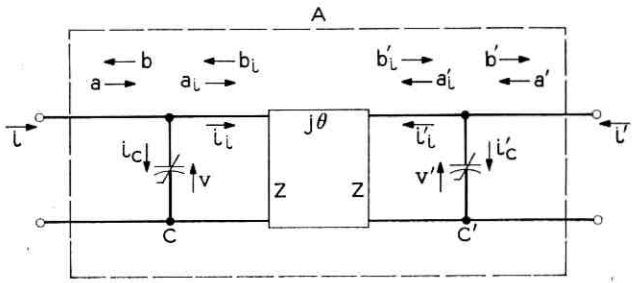


Fig. 1 — Basic configuration of the unilateral parametric amplifier.

Currents and voltages of the varying capacitance C are related by the following equations:

$$\begin{aligned} i_{c1} &= j\omega_1(c/2)v_2^* \\ i_{c2}^* &= -j\omega_2(c^*/2)v_1 \end{aligned} \tag{1}$$

where ω is angular frequency and the asterisk indicates a complex conjugate. The terminal currents and voltages of the capacitance C are expressed in terms of incident and reflected waves a, b, a_i, and b_i normalized by Z as follows:

$$\begin{aligned} v &= \sqrt{Z}(a + b) = \sqrt{Z}(a_i + b_i) \\ i &= \frac{1}{\sqrt{Z}}(a - b) \\ i_i &= \frac{1}{\sqrt{Z}}(a_i - b_i). \end{aligned} \tag{2}$$

The directions of flow of these waves are indicated by arrows in Fig. 1. The continuity equation of current at the terminal C is given by

$$i = i_c + i_i . \tag{3}$$

Combining (1), (2), and (3), we obtain the following equations:

$$\begin{aligned} a_{i1} &= a_1 - j\omega_1\dot{\kappa}a_2^* - j\omega_1\dot{\kappa}b_2^* \\ b_{i1} &= b_1 + j\omega_1\dot{\kappa}a_2^* + j\omega_1\dot{\kappa}b_2^* \\ a_{i2}^* &= a_2^* + j\omega_2\dot{\kappa}^*a_1 + j\omega_2\dot{\kappa}^*b_1 \\ b_{i2}^* &= b_2^* - j\omega_2\dot{\kappa}^*a_1 - j\omega_2\dot{\kappa}^*b_1 \end{aligned} \tag{4}$$

where $\dot{\kappa}$ is a complex number defined as:

$$\dot{\kappa} = \frac{c}{2}\cdot\frac{\sqrt{Z_1Z_2}}{2}. \tag{5}$$

Equations similar to (4) are obtainable for the varying capacitance C' by replacing unprimed parameters by primed ones in (4).

Waves of the lossless reciprocal two-terminal-pair network are related by the following equations.

$$\begin{aligned} b_i &= e^{-j\theta}a_i' \\ b_i' &= e^{-j\theta}a_i . \end{aligned} \tag{6}$$

Substituting (4) and the similar equations for C' into (6), we obtain the following set of equations:

$$\begin{aligned} b_1 + j\omega_1\dot{\kappa}b_2^* + j\omega_1\dot{\kappa}'e^{-j\theta_1}b_2'^* &= e^{-j\theta_1}a_1' - j\omega_1\dot{\kappa}a_2^* - j\omega_1\dot{\kappa}'e^{-j\theta_1}a_2'^* \\ b_1' + j\omega_1\dot{\kappa}e^{-j\theta_1}b_2^* + j\omega_1\dot{\kappa}'b_2'^* &= e^{-j\theta_1}a_1 - j\omega_1\dot{\kappa}e^{-j\theta_1}a_2^* - j\omega_1\dot{\kappa}'a_2'^* \\ -j\omega_2\dot{\kappa}^*b_1 - j\omega_2\dot{\kappa}'^*e^{j\theta_2}b_1' + b_2^* &= j\omega_2\dot{\kappa}^*a_1 + j\omega_2\dot{\kappa}'^*e^{j\theta_2}a_1' + e^{j\theta_2}a_2'^* \\ -j\omega_2\dot{\kappa}^*e^{j\theta_2}b_1 - j\omega_2\dot{\kappa}'^*b_1' + b_2'^* &= j\omega_2\dot{\kappa}^*e^{j\theta_2}a_1 + j\omega_2\dot{\kappa}'^*a_1 + e^{j\theta_2}a_2^* \end{aligned} \tag{7}$$

Assuming the same magnitude for the complex quantities $\dot{\kappa}$ and $\dot{\kappa}'$, and solving (7) with respect to b_1, b_1', b_2^*, and $b_2'^*$, we obtain an exact form of the scattering matrix for the circuit shown in Fig. 1. The matrix is as follows:

$$\begin{bmatrix} b_1 \\ b_1' \\ b_2^* \\ b_2'^* \end{bmatrix} = \begin{bmatrix} S_{11} & S_{12} & S_{13} & S_{14} \\ S_{21} & S_{22} & S_{23} & S_{24} \\ S_{31} & S_{32} & S_{33} & S_{34} \\ S_{41} & S_{42} & S_{43} & S_{44} \end{bmatrix} \cdot \begin{bmatrix} a_1 \\ a_1' \\ a_2^* \\ a_2'^* \end{bmatrix} \tag{8}$$

$$
\begin{aligned}
S_{11}\cdot\Delta = S_{22}\cdot\Delta &= 2\omega_1\omega_2\kappa^2 \exp(-j\theta_1)\,[\cos\theta_1 + \exp(j\theta_2)\cos\theta_p] \\
&\quad - 4\omega_1^2\omega_2^2\kappa^4 \exp[j(\theta_2-\theta_1)]\sin\theta_1\sin\theta_2 \\
S_{33}\cdot\Delta = S_{44}\cdot\Delta &= 2\omega_1\omega_2\kappa^2 \exp(j\theta_2)\,[\cos\theta_2 + \exp(-j\theta_1)\cos\theta_p] \\
&\quad - 4\omega_1^2\omega_2^2\kappa^4 \exp[j(\theta_2-\theta_1)]\sin\theta_1\sin\theta_2 \\
S_{12}\cdot\Delta &= \exp(-j\theta_1)\,\{1 + j2\omega_1\omega_2\kappa^2 \\
&\quad \cdot\exp[j(\theta_2+\theta_p)]\sin\theta_1\} \\
S_{21}\cdot\Delta &= \exp(-j\theta_1)\,\{1 + j2\omega_1\omega_2\kappa^2 \\
&\quad \cdot\exp[j(\theta_2-\theta_p)]\sin\theta_1\} \\
S_{34}\cdot\Delta &= \exp(j\theta_2)\,\{1 - j2\omega_1\omega_2\kappa^2 \\
&\quad \cdot\exp[-j(\theta_1+\theta_p)]\sin\theta_2\} \\
S_{43}\cdot\Delta &= \exp(j\theta_2)\,\{1 - j2\omega_1\omega_2\kappa^2 \\
&\quad \cdot\exp[-j(\theta_1-\theta_p)]\sin\theta_2\}
\end{aligned}
\tag{9}
$$

$$
\begin{aligned}
S_{13}\cdot\Delta = -(\omega_1/\omega_2)S_{42}\cdot\Delta &= -j\omega_1\kappa \exp(j\theta_p/2) \\
&\quad \cdot\{1 + \exp[j(\theta_2-\theta_1-\theta_p)] - 4\omega_1\omega_2\kappa^2 \\
&\quad \cdot\exp[j(\theta_2-\theta_1)]\sin\theta_1\sin\theta_2\} \\
S_{24}\cdot\Delta = -(\omega_1/\omega_2)S_{31}\cdot\Delta &= -j\omega_1\kappa \exp(-j\theta_p/2) \\
&\quad \cdot\{1 + \exp[j(\theta_2-\theta_1+\theta_p)] - 4\omega_1\omega_2\kappa^2 \\
&\quad \cdot\exp[j(\theta_2-\theta_1)]\sin\theta_1\sin\theta_2\} \\
S_{14}\cdot\Delta = -(\omega_1/\omega_2)S_{32}\cdot\Delta &= -j\omega_1\kappa \exp(j\theta_p/2)\,\{\exp(j\theta_2) \\
&\quad + \exp[-j(\theta_1+\theta_p)]\} \\
S_{23}\cdot\Delta = -(\omega_1/\omega_2)S_{41}\cdot\Delta &= -j\omega_1\kappa \exp(-j\theta_p/2)\,\{\exp(j\theta_2) \\
&\quad + \exp[-j(\theta_1-\theta_p)]\}
\end{aligned}
$$

and

$$
\begin{aligned}
\Delta = 1 &- 2\omega_1\omega_2\kappa^2(1 + \exp[j(\theta_2-\theta_1)]\cos\theta_p) \\
&+ 4\omega_1^2\omega_2^2\kappa^4 \exp[j(\theta_2-\theta_1)]\sin\theta_1\sin\theta_2
\end{aligned}
\tag{10}
$$

where κ and θ_p are real numbers, and represent the magnitudes of $\dot{\kappa}$ and $\dot{\kappa}'$, and a difference of pump phases at two diodes respectively as shown in the following equations:

$$\begin{aligned} \dot{\kappa} &= \kappa e^{j\theta_p/2} \\ \dot{\kappa}' &= \kappa e^{-j\theta_p/2}. \end{aligned} \tag{11}$$

Assuming the following phase relationships

$$\begin{aligned} \theta_1 &= (2m + 1)(\pi/2) \\ \theta_2 &= n\pi \\ \theta_p &= (2m + 2n + 1)(\pi/2), \end{aligned} \tag{12}$$

we obtain the following special relations for the waves:

$$\begin{aligned} b_1 &= (-)^{m+1} j a_1' \\ b_1' &= (-)^{m+1} j \frac{1 + 2\omega_1\omega_2\kappa^2}{1 - 2\omega_1\omega_2\kappa^2} a_1 + (-)^{n+1} j \frac{2\omega_1\kappa e^{-j\theta_p/2}}{1 - 2\omega_1\omega_2\kappa^2} a_2^* \\ &\quad - j \frac{2\omega_1\kappa e^{-j\theta_p/2}}{1 - 2\omega_1\omega_2\kappa^2} a_2'^* \\ b_2^* &= j \frac{2\omega_2\kappa e^{-j\theta_p/2}}{1 - 2\omega_1\omega_2\kappa^2} a_1 + \frac{2\omega_1\omega_2\kappa^2}{1 - 2\omega_1\omega_2\kappa^2} a_2^* + (-)^n \frac{1}{1 - 2\omega_1\omega_2\kappa^2} a_2'^* \\ b_2'^* &= (-)^n j \frac{2\omega_2\kappa e^{-j\theta_p/2}}{1 - 2\omega_1\omega_2\kappa^2} a_1 + (-)^n \frac{1}{1 - 2\omega_1\omega_2\kappa^2} a_2^* \\ &\quad + \frac{2\omega_1\omega_2\kappa^2}{1 - 2\omega_1\omega_2\kappa^2} a_2'^*. \end{aligned} \tag{13}$$

In (13) it is found that signal waves are completely matched in each direction, and that the signal incident wave a_1' propagates from the right to the left with no gain nor loss and has no interaction with the other three waves. When both idler ports are properly terminated, the transducer gain of this amplifier for the incident wave a_1 — that is, $|S_{21}|^2$ — equals the square of the ratio $(1 + 2\omega_1\omega_2\kappa^2)/(1 - 2\omega_1\omega_2\kappa^2)$. If the surge impedance of the circulator of the reflection-type amplifier at the signal frequency equals the image impedance Z_1 and the idler load impedance at the idler frequency equals the image impedance Z_2, the gain formula is identical to that of a single-diode reflection-type amplifier, $|(1 + 4\omega_1\omega_2\kappa^2)/(1 - 4\omega_1\omega_2\kappa^2)|^2$, except for a factor of $\frac{1}{2}$ in the pumping term $2\omega_1\omega_2\kappa^2$. Thus, a perfectly matched and unilateral parametric amplifier can be achieved by means of the phase synchronization shown in (12).

If all frequency characteristics of θ_1, θ_2, Z_1 and Z_2 are known, it is possible to calculate the frequency characteristic of the amplifier from

(9) and (10). Since this method is rather tedious, a simplified method will be developed in the following sections.

III. EFFECTS OF REFLECTIONS, BANDWIDTH

For the sake of analytical simplicity we first assume that phase constants are frequency independent and satisfy the condition given by (12), but we introduce the frequency-dependent reflection coefficients of the external loads. These reflection coefficients can be modified to include the effects of frequency-dependent phase constants.

The incident and reflected idler waves are related by the idler reflection coefficients Γ_2 and Γ_2' as follows:

$$\begin{aligned} a_2^* &= \Gamma_2^* b_2^* \\ a_2'^* &= \Gamma_2'^* b_2'^*. \end{aligned} \tag{14}$$

Substituting these relations into (13) and eliminating a_2^*, $a_2'^*$, b_2^* and $b_2'^*$ from the equation, we obtain the following equations:

$$\begin{aligned} b_1 &= S_{12}' a_1' \\ b_1' &= S_{21}' a_1 \end{aligned} \tag{15}$$

where

$$\begin{aligned} S_{12}' &= (-)^{m+1} j \\ S_{21}' &= (-)^{m+1} j \left[\frac{1 + 2\omega_1\omega_2\kappa^2}{1 - 2\omega_1\omega_2\kappa^2} \right. \\ &\left. + \frac{8\omega_1\omega_2\kappa^2 \left\{ 1 + \frac{1}{2}\left(\frac{1}{\Gamma_2^*} + \frac{1}{\Gamma_2'^*} \right) \right\}}{\left(2\omega_1\omega_2\kappa^2 - \frac{1 - 2\omega_1\omega_2\kappa^2}{\Gamma_2^*} \right)\left(2\omega_1\omega_2\kappa^2 - \frac{1 - 2\omega_1\omega_2\kappa^2}{\Gamma_2'^*} \right) - 1} \right]. \end{aligned} \tag{16}$$

Equation (15) shows that any reflection in the idler circuits does not deteriorate the unilateral characteristic of the amplifier. (A change in idler phase, θ_2, however, causes deterioration in the unilateral characteristics of the amplifier: see Appendix.)

Next let us consider a special case in which the amplifier has symmetrical idler terminations, $\Gamma_2' = \Gamma_2$. Substituting this condition into the second equation of (16), we simplify the equation as follows:

$$S_{21}' = (-)^{m+1} j \frac{\dfrac{1 + 2\omega_1\omega_2\kappa^2}{1 - 2\omega_1\omega_2\kappa^2} - \Gamma_2^*}{1 - \Gamma_2^* \dfrac{1 + 2\omega_1\omega_2\kappa^2}{1 - 2\omega_1\omega_2\kappa^2}}. \tag{17}$$

The reflection coefficient Γ_2 is expressed in terms of the image impedance Z_2 and the idler load admittance Y_{T2} as follows:

$$\Gamma_2 = \frac{1 - Y_{T2}Z_2}{1 + Y_{T2}Z_2}, \tag{18}$$

where

$$Y_{T2} = G_{T2} + jB_{T2}. \tag{19}$$

This admittance may include parasitic elements of the diode. Substituting (5), (11) and (18) into (17), we obtain the following relation:

$$S_{21}' = (-)^{m+1} j \frac{Y_{T2}^* + \frac{1}{8}\omega_1\omega_2 \mid c \mid^2 Z_1}{Y_{T2}^* - \frac{1}{8}\omega_1\omega_2 \mid c \mid^2 Z_1}. \tag{17'}$$

This equation shows that the idler image impedance Z_2 does not affect the gain. Only the idler load admittance, Y_{T2}, is of primary importance for transmission gain of the amplifier. Therefore, we can assume hereafter that the idler image impedance equals $1/G_{T2}$ in order to avoid the reflection at the center frequency of amplification. The square root of the power gain at the center frequency, $\sqrt{PG}$, is given by the following relation:

$$\sqrt{PG} = g \equiv \frac{G_{T2} + \frac{1}{8}\omega_1\omega_2 \mid c \mid^2 Z_1}{G_{T2} - \frac{1}{8}\omega_1\omega_2 \mid c \mid^2 Z_1} \tag{20}$$

and a half-gain bandwidth, $(\Delta f)_{3db}$, is approximately determined from the frequencies where the susceptance component of the idler load equals the denominator of (20), i.e.,

$$\mid B_{T2} \mid = G_{T2} - \tfrac{1}{8}\omega_1\omega_2 \mid c \mid^2 Z_1 \tag{21}$$

where the right-hand side of (21) is a slowly varying function of frequency in comparison with B_{T2}.

Now let us consider effects of small reflections in the signal circuit. We assume that both the signal source and load admittances are represented by $(1/R_{01}) + jB_1$ as shown in Fig. 2, and waves a_{01}, b_{01}, a_{01}' and b_{01}' are normalized to R_{01} instead of Z_1. Expressing voltages and currents in terms of waves as was done in (2), we obtain the following equations from the equations of continuity of currents and voltages:

$$\begin{aligned} a_1 &= (k_1 - j\mu_1)a_{01} + (l_1 - j\mu_j)b_{01} \\ b_1 &= (l_1 + j\mu_1)a_{01} + (k_1 + j\mu_1)b_{01} \end{aligned} \tag{22}$$

where

$$\begin{aligned} k_1 &= \frac{1}{2}\left(\sqrt{\frac{R_{01}}{Z_1}} + \sqrt{\frac{Z_1}{R_{01}}}\right) \\ l_1 &= \frac{1}{2}\left(\sqrt{\frac{R_{01}}{Z_1}} - \sqrt{\frac{Z_1}{R_{01}}}\right) \\ \mu &= \frac{B_1}{2}\sqrt{R_{01}Z_1}\,. \end{aligned} \tag{23}$$

Substituting (23) into (15), we obtain the following relations:

$$\begin{aligned} b_{01} &= S_{11}{}^{(0)}a_{01} + S_{12}{}^{(0)}a_{01}' \\ b_{01}' &= S_{21}{}^{(0)}a_{01} + S_{22}{}^{(0)}a_{01}' \end{aligned} \tag{24}$$

$$S_{11}{}^{(0)} = S_{22}{}^{(0)} = \frac{k_1 - j\mu_1}{k_1 + j\mu_1}\cdot\frac{-\Gamma_1{}^* + S_{12}'S_{21}'\Gamma_1}{1 - S_{12}'S_{21}'\Gamma_1{}^2}$$

$$S_{12}{}^{(0)} = \frac{k_1 - j\mu_1}{k_1 + j\mu_1}\cdot\frac{S_{12}'(1 - |\,\Gamma_1\,|^2)}{1 - S_{12}'S_{21}'\Gamma_1{}^2} \tag{25}$$

$$S_{21}{}^{(0)} = \frac{k_1 - j\mu_1}{k_1 + j\mu_1}\cdot\frac{S_{21}'(1 - |\,\Gamma_1\,|^2)}{1 - S_{12}'S_{21}'\Gamma_1{}^2}$$

where Γ_1 is a reflection coefficient of the signal source and load normalized to Z_1, and is defined as follows

$$\Gamma_1 = \frac{l_1 - j\mu_1}{k_1 + j\mu_1} = \frac{\dfrac{1}{Z_1} - \left(\dfrac{1}{R_{01}} + jB_1\right)}{\dfrac{1}{Z_1} + \left(\dfrac{1}{R_{01}} + jB_1\right)}. \tag{26}$$

Equation (25) reveals that any reflections in signal circuits cause an

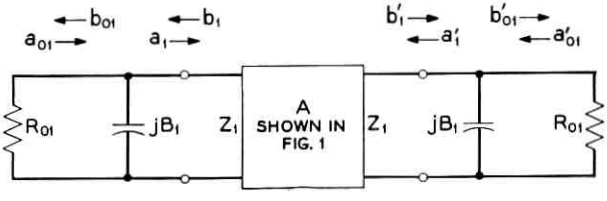


Fig. 2 — Forward and reverse waves in the signal circuits.

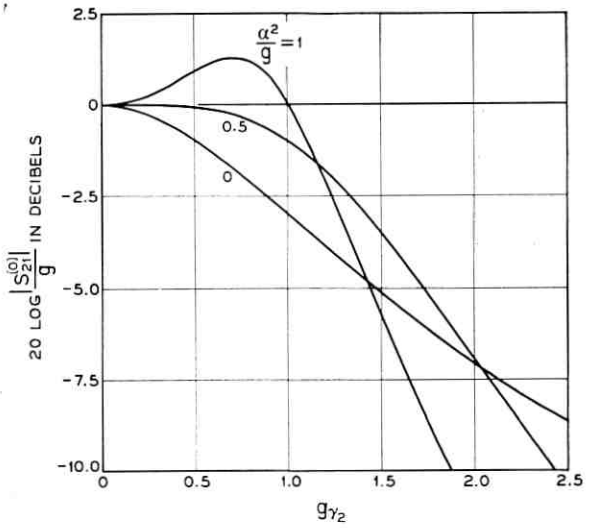


Fig. 3 — Normalized forward gain for various values of α^2/g vs normalized idler reflection coefficient $g\gamma_2$.

internal feedback and deteriorate the unilateral characteristic. To simplify the analysis, we assume that Γ_1 is proportional to Γ_2 and that both of them are small imaginary numbers, i.e., $\Gamma_2 = j\gamma_2$, $\Gamma_1 = j\alpha\gamma_2$, $|\Gamma_1|^2 \ll 1$, and $|\Gamma_2| \ll 1$. Substituting the first equation of (16) and (17) into (25), we obtain the following scattering matrix elements, which characterize the frequency dependence of the amplifier if $g \gg 1$:

reflection:

$$|S_{11}^{(0)}| = |S_{22}^{(0)}| \approx \frac{|\alpha||g\gamma_2|}{\sqrt{\left\{1 - \frac{\alpha^2}{g}(g\gamma_2)^2\right\}^2 + (g\gamma_2)^2}}$$

reverse gain:

$$|S_{12}^{(0)}| \approx \frac{\sqrt{1 + (g\gamma_2)^2}}{\sqrt{\left\{1 - \frac{\alpha^2}{g}(g\gamma_2)^2\right\}^2 + (g\gamma_2)^2}} \tag{27}$$

forward gain:

$$|S_{21}^{(0)}| \approx \frac{g}{\sqrt{\left\{1 - \frac{\alpha^2}{g}(g\gamma_2)^2\right\}^2 + (g\gamma_2)^2}}.$$

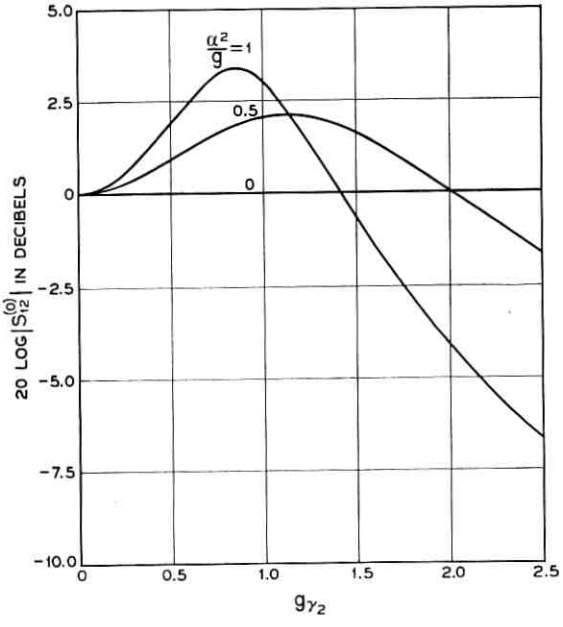


Fig. 4 — Reverse gain for various values of α^2/g vs normalized idler reflection coefficient $g\gamma_2$.

The characteristics of these elements with various values of the parameter α^2/g, as a function of $g\gamma_2$, are shown in Figs. 3, 4, and 5.

It is shown that the amplifier has a maximally flat gain characteristic when $\alpha^2/g = \frac{1}{2}$, as follows:

$$| S_{21}{}^{(0)} | \approx \frac{g}{\sqrt{1 + \frac{(g\gamma_2)^4}{4}}}. \tag{28}$$

A half-gain bandwidth is determined from the frequencies where the following relation is held:

$$| g\gamma_2 | \approx \sqrt{2}.$$

Similarly, for the amplifier with a matched signal source and load, the bandwidth is derived from the following equation:

$$| g\gamma_2 | \approx 1.$$

Therefore, by introducing a proper mismatch in the signal source and load and assuming that γ_2 is proportional to a frequency change, we can

obtain about $\sqrt{2}$ times larger bandwidth than that of the matched amplifier. But this enlarged bandwidth is obtainable only at the expense of deterioration of unilateral characteristics, as shown in Fig. 4.

IV. EFFECTS OF LOSSES; NOISE FIGURE

In the previous discussions, losses in the diode have been considered a part of the external circuits. However, in order to study the noise performance of the amplifier the diode losses have to be separated from the external circuits.

In Fig. 6, G_d and G_d' represent the equivalent loss conductance of the diodes, and they are accompanied by the noise current generators i_{nd} and i_{nd}'. G_L and G_s represent the load and signal source conductance, respectively.

In order to eliminate feedback effect due to mismatches at the signal ports, we assume that the signal output port is perfectly matched to its image impedance Z_1. Thus, we can eliminate any feedback effect even if there is mismatch at the signal input port. The matching condition at the signal output port is expressed as follows:

$$(G_L + G_d')Z_1 = 1. \tag{29}$$

In order to feed the maximum power into the black box A shown in Fig. 6, the signal source impedance has to be matched to the input impedance of the amplifier, which includes the diode loss conductance G_d. This matching condition is expressed as follows:

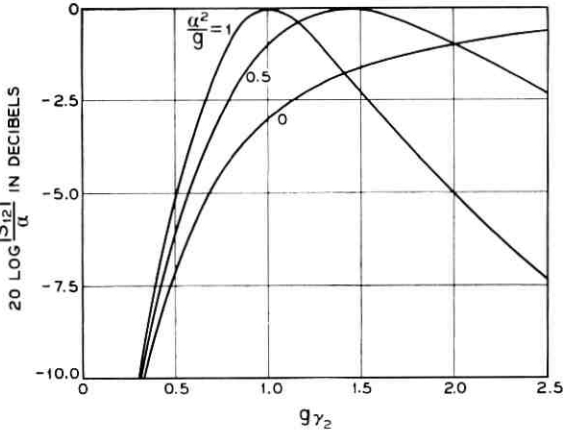


Fig. 5 — Normalized input reflection coefficient for various values of α^2/g vs normalized idler reflection coefficient $g\gamma_2$.

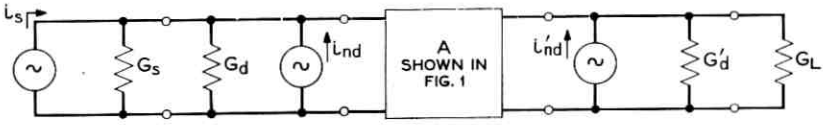


Fig. 6 — Noise sources of the unilateral parametric amplifier.

$$G_s Z_1 = 1 + G_d Z_1 . \tag{30}$$

Note that the condition in (30) does not necessarily mean that the signal source impedance is matched to Z_1.

The maximum power available into the black box A of Fig. 6 is given by

$$| a_1 |^2 = \frac{P_a}{G_s Z_1} = \frac{P_a}{1 + G_d Z_1}, \tag{31}$$

where P_a is the maximum available power from signal source. The power delivered to the load, denoted by P_0, is given by the following equation:

$$P_0 = G_L Z_1 | b_1' |^2. \tag{32}$$

Therefore, from the second equation of (13), and assuming no input signal at the idler frequency, we obtain the following gain expression:

$$PG = \frac{P_0}{P_a} = \frac{G_L Z_1}{1 + G_d Z_1} g^2, \tag{33}$$

where g is given by (20). [In (20) we assumed that the idler loads are not necessarily matched. However, as stated there, we can assume the matched idler loads without loss of generality.] This gain expression is very similar to that of a single-diode reflection-type amplifier, as mentioned in Section II.

Assuming noise sources are thermal and the temperatures of the amplifier, the signal source and the load are the same, denoted by $T°$ Kelvin, noise currents and noise wave are expressed in the following relation:

$$\frac{\overline{i_{nd}^2}}{4G_d} = \frac{\overline{i_{nd}'^2}}{4G_d'} = | a_1' |^2 = | a_2 |^2 = | a_2' |^2 = kT\,\Delta f \tag{34}$$

where $k = 1.38 \times 10^{-23}$ joule/°Kelvin is Boltzmann's constant and Δf is the infinitesimal bandwidth concerned.

Though we have a mismatched signal source with respect to Z_1, the difference between the noise power from a matched signal source and that from the mismatched signal source is exactly the same as the noise

power reflected at the terminal of this mismatched signal source. This reflected noise power is originally from the perfectly matched signal output network through the amplifier without gain nor loss in the reverse direction. Therefore, the resultant incoming noise wave into the amplifier is the same as mentioned in (34)

$$| a_1 |^2 = kT\Delta f. \tag{34'}$$

Substituting (34) and (34′) into the second equation of (13), we obtain

$$| b_1' |^2 = g^2 \left\{ 1 + \frac{\omega_1}{\omega_2} \left(1 - \frac{1}{g^2} \right) \right\} kT\,\Delta f.$$

Therefore, the noise output power due to $| b_1' |^2$ is given by

$$P_{01} = G_L Z_1 g^2 \left\{ 1 + \frac{\omega_1}{\omega_2} \left(1 - \frac{1}{g^2} \right) \right\} kT\,\Delta f.$$

The directly radiated noise power into the load from G_d' is also the other part of the noise output power. We obtain the following equation for this noise output power:

$$P_{02} = G_L Z_1 G_d' Z_1 kT\,\Delta f.$$

Therefore, we obtain the following expression for the noise figure of the amplifier:

$$\begin{aligned} F &= \frac{P_{01} + P_{02}}{(PG)kT\,\Delta f} \\ &= (1 + G_d Z_1) \left\{ 1 + \frac{G_d' Z_1}{g^2} + \frac{\omega_1}{\omega_2} \left(1 - \frac{1}{g^2} \right) \right\}. \end{aligned} \tag{35}$$

It is shown in (35) that $G_d Z_1$ and ω_1/ω_2 should be smaller in order to obtain a better noise figure. And the expression for the noise figure is very similar to that of a single-diode amplifier. Note that (35) was obtained on the assumption that the temperature of the resistive load or isolator after the amplifier is $T°$ Kelvin, and the noise figure is also normalized to $T°$ Kelvin. If the load temperature is 0° Kelvin, we have to subtract $G_d^2 Z_1^2 G_L Z_1/(1 + G_d Z_1)$ from the right-hand side of (35), which comes from the noise power from the load at $T°$ Kelvin.

If the load or isolator temperature, T_I, is higher than T, the noise power generated in the isolator travels toward the input side of the amplifier, reflects back at the input diode, and is amplified. Therefore, the higher the isolator temperature is, the smaller reflection must be

kept at the input diode. If the isolator temperature is much higher than the amplifier temperature, a matched condition to Z_1 at the input diode, i.e.

$$G_s Z_1 + G_d Z_1 \approx 1 \tag{30'}$$

must be held in order to keep the effect of the hot isolator small. In this case, the power gain and the noise figure normalized to $T°$ Kelvin are approximately obtained in the following equations:

$$PG \approx G_{L1} Z_1 G_s Z_1 g^2 \tag{33'}$$

$$F \approx \frac{1}{1 - G_d Z_1}\left\{1 + \frac{G_d' Z_1}{g^2} + \frac{\omega_1}{\omega_2}\left(1 - \frac{1}{g^2}\right) + \frac{T_I}{T}\cdot\frac{G_L Z_1}{4}\left(G_s Z_1 + G_d Z_1 - 1\right)^2\right\} \quad \text{if} \quad \frac{T_I}{T} \gg 1. \tag{35'}$$

The last term in parentheses represents the deterioration of noise figure due to input mismatching.

V. OPTIMUM NOISE FIGURE OF THE AMPLIFIER WITH LOSSY DIODES

To simplify the analysis, we assume that a variable-capacitance diode is represented by an equivalent circuit shown in Fig. 7(a). The junction capacitance $\tilde{C}_j$ is the only variable element and is given by the following equation:

$$\tilde{C}_j = C_j + c \cos \omega_p t \tag{36}$$

where ω_p denotes the pumping angular frequency and equals $\omega_1 + \omega_2$.

The characteristic quantities of the diode, ω_{cr}, ω_0 and Q_0 are defined as follows: the critical angular frequency

$$\omega_{cr} = \frac{1}{2}\frac{1}{C_j R_s}\frac{|c|}{C_j} \approx \omega \tilde{Q}, \tag{37}$$

where $\tilde{Q}$ is a dynamic quality factor of the diode; the self-resonant angular frequency

$$\omega_0 = \frac{1}{\sqrt{L_s C_j}}; \tag{38}$$

and the diode Q at the resonant frequency

$$Q_0 = \frac{\omega_0 L_s}{R_s}. \tag{39}$$

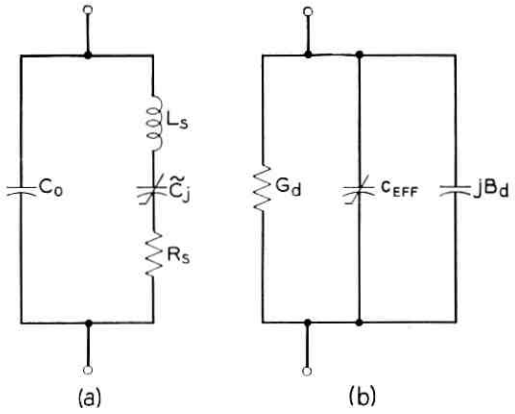


Fig. 7 — Equivalent circuits of a variable capacitance diode.

Transforming the circuit of Fig. 7(a) into the form shown in Fig. 7(b), where c_{eff} is an equivalent sinusoidally varying shunt capacitance, the new circuit parameters at frequencies away from the self-resonant frequency f_0 are given by the following relations:

$$\begin{aligned} B_d &\approx \frac{1}{R_s Q_0}\left\{\frac{C_0}{C_j}\frac{\omega}{\omega_0} - \frac{1}{\dfrac{\omega}{\omega_0} - \dfrac{\omega_0}{\omega}}\right\} \\ G_d &\approx \frac{1}{R_s Q_0^2}\frac{1}{\left(\dfrac{\omega}{\omega_0} - \dfrac{\omega_0}{\omega}\right)^2} \\ c_{\text{eff}} &\approx \frac{\dfrac{\omega_0^2}{\omega_1\omega_2}}{\left(\dfrac{\omega_1}{\omega_0} - \dfrac{\omega_0}{\omega_1}\right)\left(\dfrac{\omega_2}{\omega_0} - \dfrac{\omega_0}{\omega_2}\right)}\, C. \end{aligned} \tag{40}$$

Figs. 8 and 9 show the frequency characteristics of B_d and G_d respectively for various values of a parameter C_0/C_j.

From (40), we obtain the following relation:

$$\frac{\omega_1\omega_2 \mid c_{\text{eff}} \mid^2}{4} = \frac{\omega_{cr}^2}{\omega_1\omega_2} G_{d1}G_{d2}. \tag{41}$$

For a high-gain amplifier, the following condition should be satisfied from (20):

$$G_{T2} = (1 + L_2)G_{d2} \approx \tfrac{1}{8}\omega_1\omega_2 \mid c_{\text{eff}} \mid^2 Z_1 \tag{42}$$

where L_2 is an external idler loading factor of the amplifier.* Substituting (41) into (42), the high-gain condition is expressed as follows:

$$\frac{1}{2}\frac{\omega_{cr}^2}{\omega_1\omega_2}\frac{G_{d1}Z_1}{1+L_2} \approx 1. \tag{42'}$$

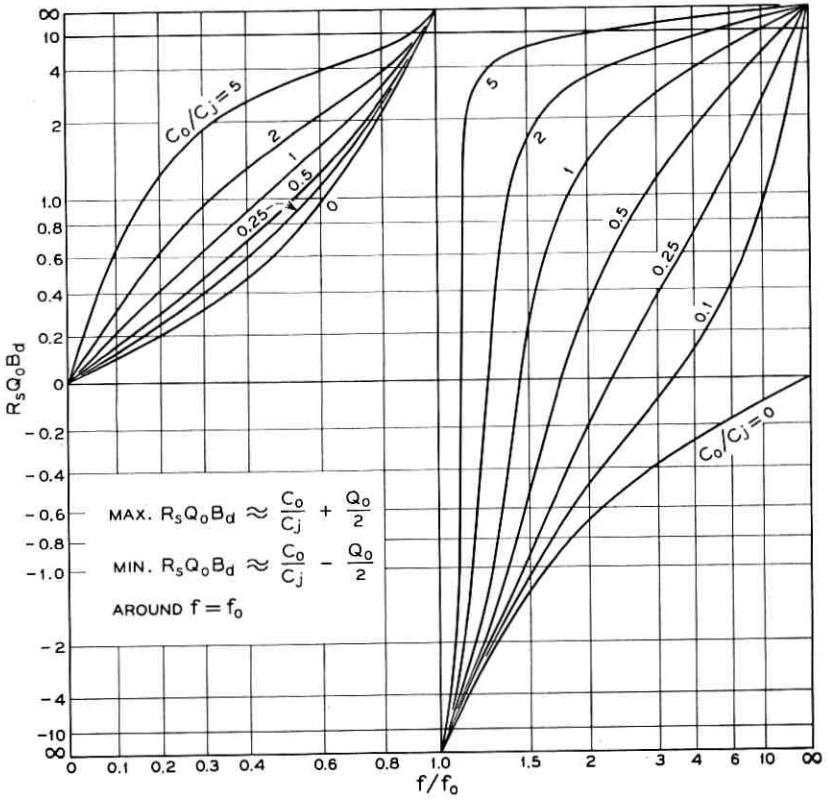


Fig. 8 — Frequency characteristics of an equivalent shunt susceptance of a diode.

From (35), a noise figure expression for a high-gain amplifier is approximately given by the following equation:

$$F \approx (1 + G_{d1}Z_1)\left(1 + \frac{\omega_1}{\omega_2}\right). \tag{35''}$$

Eliminating $G_{d1}Z_1$ from (35″) by using (42′), and differentiating (35″)

* Originally introduced by M. Uenohara.

with respect to ω_2/ω_1, we obtain the optimum ratio of idler-to-signal frequency for the optimum noise figure, and the optimum noise figure for a given signal frequency f_1, a diode critical frequency f_{cr} and an idler loading factor L_2

$$F_{\text{opt}} = \left\{1 + \frac{\omega_1}{\omega_{cr}}\sqrt{2(1 + L_2)}\right\}^2 \approx \left\{1 + \frac{1}{\tilde{Q}_1}\sqrt{2(1 + L_2)}\right\}^2 \quad (43)$$

$$\left(\frac{\omega_2}{\omega_1}\right)_{\text{opt}} = \frac{\frac{\omega_{cr}}{\omega_1}}{\sqrt{2(1 + L_2)}} \approx \frac{\tilde{Q}_1}{\sqrt{2(1 + L_2)}}. \quad (44)$$

Fig. 10 shows the numerical values of $(F)_{\text{opt}}$ and $(\omega_2/\omega_1)_{\text{opt}}$.

It is shown in (43) that the optimum noise figure is fairly good but slightly higher* than that of a single-diode parametric amplifier. The reason for the higher noise figure is attributed to the fact that, to obtain the same amount of gain with a given ratio of reactance swing, the quarter-wave coupled amplifier needs a lower idler frequency than that of a single-diode reflection-type amplifier. However, for reflection-type amplifiers it may not be possible to use the optimum idler frequency in order to obtain a wide bandwidth, and a small difference in thermal noise

* For a single-diode parametric amplifier, the optimum noise figure and the optimum ratio of idler-to-signal frequency are obtained in terms of a critical frequency and idler loading factor, as follows:

$$F_{\text{opt}} = \left\{\sqrt{1 + \left(\frac{\omega_1}{\omega_{cr}}\right)^2(1 + L_2)} + \frac{\omega_1}{\omega_{cr}}\sqrt{1 + L_2}\right\}$$

$$\approx \left(\sqrt{1 + \frac{1 + L_2}{\tilde{Q}_1^2}} + \frac{1}{\tilde{Q}_1}\sqrt{1 + L_2}\right)^2$$

$$\left(\frac{\omega_2}{\omega_1}\right)_{\text{opt}} = \sqrt{\left(\frac{\omega_{cr}}{\omega_1}\right)^2 \frac{1}{1 + L_2} + 1} - 1$$

$$\approx \frac{\tilde{Q}_1}{\sqrt{1 + L_2}} \cdot \frac{1}{\sqrt{1 + \frac{1 + L_2}{\tilde{Q}_1^2}} + \frac{\sqrt{1 + L_2}}{\tilde{Q}_1}}$$

If $\tilde{Q}_1 \gg 1$, these equations yield

$$F_{\text{opt}} \approx \left(1 + \frac{1}{\tilde{Q}_1}\sqrt{1 + L_2}\right)^2$$

$$\left(\frac{\omega_2}{\omega_1}\right)_{\text{opt}} \approx \frac{\tilde{Q}_1}{\sqrt{1 + L_2}}.$$

These are the same shown in (43) and (44) if Q_1 is replaced by $_1/\sqrt{2}$.

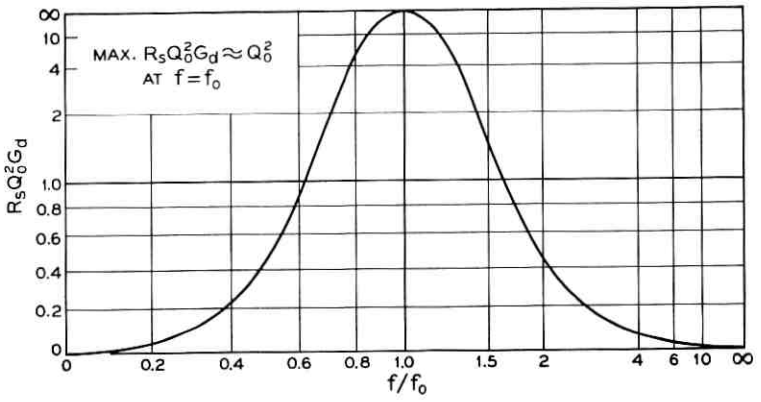


Fig. 9 — Frequency characteristic of an equivalent shunt conductance of a diode.

becomes negligible in comparison with the noise contribution from antenna circuit losses when the amplifier is refrigerated at a very low temperature; therefore, the disadvantage of slightly higher noise performance of the unilateral amplifier seems to be not always a serious problem in practical applications.

VI. DESIGN CONSIDERATION OF THE AMPLIFIER

In previous sections, we discussed the amplifier characteristics in general terms. Now we have to consider what kind of diode is necessary and what image impedance we have to choose in order to obtain a prescribed gain at a prescribed impedance of the system. And also we have to consider what characteristics the amplifier should have without pumping in order to have gain with pumping.

From the equivalent circuit of the diode shown in Fig. 7, the equivalent change in shunt susceptance ΔB due to a static change in the junction capacitance ΔC_j is given as follows:

$$\Delta B \approx \frac{\partial B}{\partial C_j} \Delta C_j = Q_0 G_d \frac{\omega_0}{\omega} \frac{\Delta C_j}{C_j} . \tag{45}$$

If the static change in the junction capacitance ΔC_j equals the amplitude of sinusoidal variation, $|c|$, (45) yields the resultant change in the equivalent susceptance as follows:

$$\Delta B = 2(\omega_{cr}/\omega) G_d . \tag{46}$$

Multiplying (46) by Z_1 and substituting it into (42′), we obtain the following equation:

$$Z_1 \Delta B_1 \approx 4 \frac{\omega_2}{\omega_{cr}} (1 + L_2). \tag{47}$$

It is seen in (47) that in order to obtain a high-gain amplifier, when the diode bias is varied over the entire range for making a passive test of the amplifier, the susceptance change normalized by the image impedance Z_1 at the signal frequency should be twice as much as the amount given in the right-hand side of (47). Also, the amplifier network should be a piece of matched transmission line if it is not pumped, and the two diodes in the amplifier should have the opposite direction of change in susceptance at the signal frequency when the bias voltage is changed in the same direction. If we have external idler ports for loading the amplifier, the two diodes should have the same direction of change in susceptance at the idler frequency. In Fig. 11 are shown required amounts of susceptance, which are twice the amount shown in (47), to fulfill the condition of unilateral amplification, as functions of idler to critical frequency ratio for various idler loading factors.

The bandwidth problem of this amplifier is not so straightforward as the noise figure problem. As mentioned in (21) of Section III, the ap-

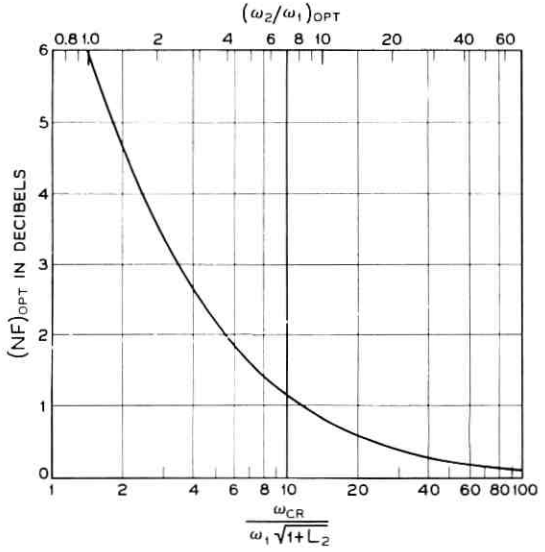


Fig. 10 — Optimum noise figure of the unilateral parametric amplifier.

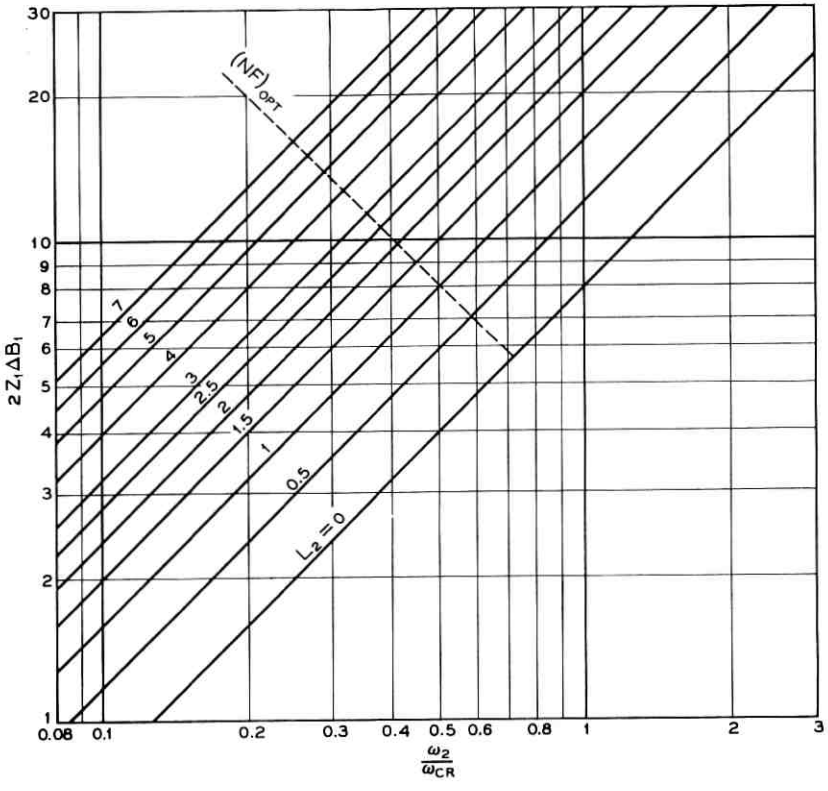


Fig. 11 — Required susceptance change at the signal frequency for high-gain amplification.

proximate bandwidth is determined by the idler circuit that can be improved by a multiple-tuning technique. However, in order to utilize a broadband idler circuit, the signal network should have a broad enough bandwidth to prevent any oscillations. The reason for this is as follows: in order to obtain a unilateral amplification a larger capacitance swing is needed than for a reflection-type amplification. Thus, wherever impedance conditions outside the signal frequency band become favorable for ordinary reflection-type amplification, the amplifier tends to change its mode of operation and breaks into oscillation. For this reason, not only must the image impedance of the network at the signal frequency be flat, but the phase constant also must be less frequency-sensitive in the frequencies where the amplifier has a gain. [For a closer investigation of the stability problem, we have to study (9).]

If both signal and idler frequencies are far from the self-resonant frequency of the diode, an unloaded Q of the diode at a given angular frequency ω, which is approximately given by

$$Q_{ud} \approx \frac{\omega}{2G_d} \frac{\partial B_d}{\partial \omega}, \tag{48}$$

is an important factor in determining the bandwidth of the amplifier. This factor is expressed by the following equation:

$$Q_{ud} \approx \frac{Q_0}{2} \left\{ \frac{\omega}{\omega_0} + \frac{\omega_0}{\omega} + \frac{C_0}{C_j} \frac{\omega}{\omega_0} \left(\frac{\omega}{\omega_0} - \frac{\omega_0}{\omega} \right)^2 \right\}. \tag{48'}$$

Fig. 12 shows the frequency dependence of Q_{ud} for various values of C_0/C_j. In Fig. 12, it is shown that the idler frequency should be close to the self-resonant frequency in order to obtain a large bandwidth. If the idler circuit is a single-tuned circuit with a loaded Q of Q_{L2}, the gain-bandwidth product is roughly given by

$$\left(\frac{\Delta f}{f_1} \right)_{3db} \sqrt{PG} \approx \frac{f_2}{f_1} \frac{2}{Q_{L2}}. \tag{49}$$

6.1 *Numerical example*

Suppose the signal and idler frequencies and diode parameters are given as follows:

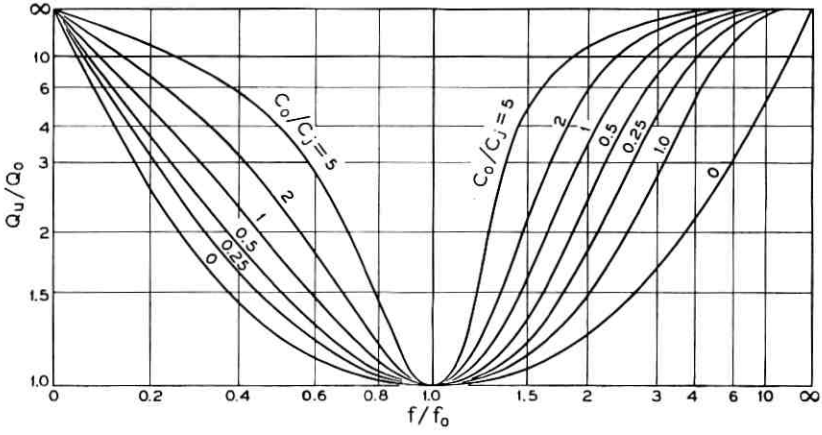


Fig. 12 — Frequency characteristic of a normalized unloaded Q of a diode.

$$f_1 = 4 \text{ gc}$$

$$f_2 = 8 \text{ gc}$$

$$C_j = 0.65 \text{ pf}$$

$$L_s = 0.6 \text{ nh}$$

$$C_0 = 0.15 \text{ pf}$$

$$f_{cr} = 30 \text{ gc}$$

$$\frac{|c|}{C_j} = 0.5.$$

From (38),

$$f_0 = \frac{1}{2\pi\sqrt{L_s C_j}} = 8.07 \text{ gc}.$$

Therefore f_0 is very close to f_2.

From (37),

$$30 \text{ gc} = \frac{1}{2\pi}\frac{1}{2}\frac{1}{C_j R_s}\frac{|c|}{C_j}$$

$$\therefore R_s = 2.05 \text{ ohms}.$$

From (39),

$$Q_0 = 14.8.$$

From the values of C_0 and C_j,

$$C_0/C_j = 0.231.$$

From the values of f_1, f_2 and f_0,

$$f_1/f_0 = 0.496$$

$$f_2/f_0 = 0.992$$

$$f_p/f_0 = (f_1 + f_2)/f_0 = 1.488.$$

From Figs. 8 and 9,

$$R_s Q_0 B_{d1} = 0.80 \quad \therefore B_{d1} = 26.3 \text{ m mhos}$$

$$R_s Q_0 B_{dp} = -0.85 \quad \therefore B_{dp} = -28.0 \text{ m mhos}$$

$$R_s Q_0^2 G_{d1} = 0.46 \quad \therefore G_{d1} = 1.02 \text{ m mhos}$$

$$R_s Q_0^2 G_{dp} = 1.44 \quad \therefore G_{dp} = 3.20 \text{ m mhos},$$

where the subscript p denotes a quantity for the pumping frequency. From (46),

$$\Delta B_1 = 15.3 \text{ m mhos.}$$

Assuming $L_2 = 0$, the right-hand side of (47) yields

$$4(\omega_2/\omega_{cr}) = 1.07.$$

Therefore, in order to have a good gain, the following condition should be held

$$Z_1 \Delta B_1 > 1.07$$

$$\therefore \; Z_1 > 70 \text{ ohms.}$$

Suppose we use 80 ohms for Z_1, which satisfies the above condition. The expected noise figure at room temperature is obtained from (35″) as follows:

$$F = 1.081 \times 1.49 = 1.61 = 2.1 \text{ db.}$$

If the amplifier and isolator are both refrigerated at 78° Kelvin, the noise temperature defined by

$$T_e = (F - 1) \times 290°\text{K}$$

becomes

$$T_e = 48°\text{K}$$

If the amplifier is further refrigerated at 42°K, the amplifier noise temperature is

$$T_e = 2.6°\text{K.}$$

If the input mismatching VSWR normalized to Z_1 is 1.2, the noise temperature of the amplifier increases by 1° Kelvin even with the isolator refrigerated at 78° Kelvin [cf. (35′)].

When we use a single-tuned idler circuit without external loading, $Q_{L2} \approx Q_0$, the percentage gain bandwidth product obtained from (49) is

$$(\Delta f/f_1)_{3\text{db}} \sqrt{PG} \approx 27 \text{ per cent.}$$

If we choose 16 db gain, the bandwidth is calculated as

$$(\Delta f/f_1)_{3\text{db}} \approx 4 \text{ per cent} \quad \text{or} \quad (\Delta f)_{3\text{db}} \approx 160 \text{ mc.}$$

With a double-tuning technique, the bandwidth may be improved up to 300 mc.

VII. CONCLUSION

Performance characteristics of a unilateral parametric amplifier with two diodes separated by a quarter wavelength at the signal frequency have been theoretically investigated on the basis of a scattering matrix representation. It is shown that a broadband signal circuit is essential in order to obtain a unilateral gain; otherwise, the amplifier may easily oscillate. The reason for this is that a unilateral amplifier requires a larger capacitance swing than an ordinary bilateral amplifier. It is shown that it is possible in principle to obtain any amount of low noise amplification without adjusting an input and output matching network of this amplifier, though the optimum noise figure is slightly higher than that of an ordinary reflection-type amplifier. This higher optimum noise figure does not always present a serious problem for practical applications, because the idler frequency is often determined by other factors such as broadbanding or pump availability. The bandwidth is primarily determined by the idler circuit. Broadbanding by introducing a mismatch in the signal circuit is not practical because it deteriorates the unilateral characteristic. A numerical example is given to show the potentiality of building an amplifier at 4 gc which has a 2.1 db noise figure and more than 300 mc bandwidth at 16 db gain. Since this amplifier does not have substantial reverse loss, it usually requires an isolator at its output port.

VIII. ACKNOWLEDGMENT

The author wishes to express his sincere gratitude to M. Uenohara for his stimulating discussions and his thorough review of the manuscript.

APPENDIX

Substituting the following relations

$$\begin{aligned} \theta_1 &= (2m + 1)\frac{\pi}{2} + \delta_1 \\ \theta_2 &= n\pi + \delta_2 \\ \theta_p &= (2m + 2n + 1)\frac{\pi}{2} + \delta_p \end{aligned} \tag{50}$$

into (10), and assuming δ's are small quantities, each coefficient of the scattering matrix is modified as follows:

$$\begin{aligned} \Delta &= 1 - 2\omega_1\omega_2\kappa^2 - j2\omega_1\omega_2\kappa^2(2\omega_1\omega_2\kappa^2\delta_2 + \delta_p) \\ S_{11}\cdot\Delta &= S_{22}\cdot\Delta = j2\omega_1\omega_2\kappa^2(\delta_1 + 2\omega_1\omega_2\kappa^2\delta_2 + \delta_p) \end{aligned} \tag{51}$$

$$S_{12}\cdot\Delta = (-)^{m+1}j[1 - 2\omega_1\omega_2\kappa^2 - j\{\delta_1(1 - 2\omega_1\omega_2\kappa^2) + 2\omega_1\omega_2\kappa^2(\delta_2 + \delta_p)\}]$$

$$S_{21}\cdot\Delta = (-)^{m+1}j[1 + 2\omega_1\omega_2\kappa^2 - j\{\delta_1(1 + 2\omega_1\omega_2\kappa^2) - 2\omega_1\omega_2\kappa^2(\delta_2 - \delta_p)\}]$$

$$S_{13}\cdot\Delta = -(\omega_1/\omega_2)S_{42}\cdot\Delta = -\omega_1\kappa \exp(j\theta_{p0}/2)\{\delta_1 + \delta_2(4\omega_1\omega_2\kappa^2 - 1) + \delta_p\}$$

$$S_{14}\cdot\Delta = -(\omega_1/\omega_2)S_{32}\cdot\Delta = (-)^n\omega_1\kappa \exp(j\theta_{p0}/2)(\delta_1 + \delta_2 + \delta_p) \quad (52)$$

etc., where

$$\theta_{p0} = (2m + 2n + 1)(\pi/2).$$

Equation (52) shows that the unilateral characteristic is deteriorated by the imperfect phase synchronization. And the amount of coupling between the signal wave in the reverse direction and the other three waves is determined by $\delta_1 + \delta_2 + \delta_p$ for a high-gain amplification where $2\omega_1\omega_2\kappa^2 \approx 1$. An imperfect pump synchronization represented by δ_p shifts the center of the amplification band as shown in (51). It is also shown in (51) that the bandwidth is affected mostly by a frequency-dependent idler phase constant if κ is kept constant.

REFERENCES

1. Baldwin, L. D., Nonreciprocal Parametric Amplifier Circuits, Proc. I.R.E., **49,** June, 1961, p. 1075. The same kind of amplifier is also described in Thompson, G. H. B., Unidirectional Lower Sideband Parametric Amplifier without Circulator, Proc. I.R.E., **49,** November, 1961, pp. 1684–1785.
2. Kurokawa, K., and Uenohara, M., Minimum Noise Figure of the Variable Capacitance Amplifier, B.S.T.J., **40,** May, 1961, pp. 695–722.
3. Kurokawa, K., On the Use of Passive Circuit Measurements for the Adjustment of Variable Capacitance Amplifiers, B.S.T.J., **41,** January, 1962, pp. 361–381.

Contributors to This Issue

R. B. Blackman, A.B., 1926, California Institute of Technology; Bell Telephone Laboratories, 1926—. Mr. Blackman was first engaged in physical research in hearing, acoustics, and electromechanical filters. He later worked in applied mathematical research, specializing in linear networks and feedback amplifiers. Since 1940, he has been engaged in the development of data-smoothing and prediction techniques for various military and satellite projects. Member, AAAS and Tau Beta Pi; Fellow, IEEE.

Fang-Shang Chen, B.S., 1951, National Taiwan University; M.S. E.E., 1955, Purdue University; Ph.D., 1959, The Ohio State University; Bell Telephone Laboratories, 1959—. He has engaged in the development of ferrite devices, traveling-wave masers, and more recently in optical modulation. Member, Tau Beta Pi, Sigma Xi and IEEE.

Franz Th. Geyling, B.S. (Civil Eng.), 1950, M.S. (Civil Eng.), 1951, and Ph.D. (Eng. Mechanics), 1954, Stanford University; Bell Telephone Laboratories, 1954—. He was initially engaged in photoelastic stress analysis and shell theory. Since 1958, his work has been concerned with the ballistics of satellites and space vehicles, including analytic perturbation studies, tracking, orbit determination and guidance studies, and the writing of large-scale computer programs for the digital simulation of space flight missions. He is coauthoring a book on the dynamics of space vehicles. His other areas of responsibility have been blast studies, the analysis of large antenna structures, and hypervelocity impact studies. Member, AIAA, ASME, International Association for Bridge and Structural Engineering and Tau Beta Pi.

Richard F. Grantges, B.S., B.E.E., 1953, University of Minnesota; Bell Telephone Laboratories, 1953—. Mr. Grantges was first engaged in systems engineering studies of bandwidth reduction techniques for submarine cable systems and early studies of digital transmission in the exchange area plant which led to the T1 carrier system. Since 1958 he has participated in the systems engineering of No. 1 ESS, particularly the

design and engineering of the switching network. At present he supervises a group working on mechanized engineering, ESS network and peripheral equipment studies. Member, Tau Beta Pi, Eta Kappa Nu and IEEE.

HERMANN K. GUMMEL, Dipl. Phys., 1952, Phillipps University (Germany); M.S., 1952, and Ph.D., 1957, Syracuse University; Bell Telephone Laboratories, 1956—. His work has been in research and development of semiconductor devices. Member, American Physical Society and Sigma Xi.

JOJI HAMASAKI, B.S.E.E., 1953, M.S.E.E., 1955, and D.E.E., 1958, University of Tokyo; Assistant Professor, Institute of Industrial Science, University of Tokyo, 1958–1961 and 1963—; Bell Telephone Laboratories, 1961–1963. At the Institute of Industrial Science, he has conducted research on microwave solid-state devices. At Bell Laboratories he worked on solid-state microwave amplifiers and related components. Member, IEEE, Institute of Electrical Communication Engineers of Japan and Institute of Electrical Engineers of Japan.

GORDON W. MILLS, B.E.E., 1950, University of Dayton; M.S. (Physics), 1962, Ohio State University. Mr. Mills worked with the University of Dayton Research Institute on nuclear weapons tests for the U. S. Air Force, 1952–1959. He joined Bell Laboratories in 1960 and has been principally concerned with connection reliability and contact studies.

T. J. NELSON, B.S., 1961, Iowa State University; M.E.E., 1963, New York University; Bell Telephone Laboratories, 1961—. Mr. Nelson has worked on ultrasonic delay lines and digital light deflection. At present he is on leave of absence to study physics at Iowa State University.

MICHAEL RAPPEPORT, B.S., 1957, Rensselaer Polytechnic Institute; M.E.E. 1958, Yale University; Bell Telephone Laboratories, 1959—. Since joining Bell Laboratories, Mr. Rappeport has been working on various analytic approaches to data transmission systems, stressing simulation approaches to studying such systems. Member, IEEE and Institute of Mathematical Statistics.

IRWIN W. SANDBERG, B.E.E., 1955, M.E.E., 1956, and D.E.E., 1958, Polytechnic Institute of Brooklyn; Bell Telephone Laboratories, 1958—. He has been concerned with analysis of military systems, particu-

larly radar systems, and with synthesis and analysis of active and time-varying networks. He is currently involved in a study of the signal-theoretic properties of nonlinear systems. Member, IEEE, Society for Industrial and Applied Mathematics, Eta Kappa Nu, Sigma Xi and Tau Beta Pi.

Norman R. Sinowitz, B.A., 1957, Yeshiva University; M.S., 1960, New York University; New York University Medical Center, 1958–1960; Bell Telephone Laboratories, 1960—. At the Medical Center he was engaged in research in radiological physics. Since joining Bell Laboratories he has been primarily concerned with electronic switching systems control and network engineering studies. Member, Association for Computing Machinery.

Friedolf M. Smits, Dipl. Phys., 1950, Dr. rer. nat., 1950, University of Freiburg, Germany; research associate, Physikalisches Institute, University of Freiburg, 1950–54; Bell Telephone Laboratories, 1954–62. Mr. Smits went to the Sandia Corporation in May, 1962. His work at Bell Laboratories included studies of solid-state diffusion in germanium and silicon, device feasibility, and process studies, as well as the development of UHF semiconductor devices. He supervised a group that conducted radiation damage studies on components, particularly solar cells, used in the Telstar satellite. Member, American Physical Society and German Physical Society.

William J. Tabor, B.S. in Chemistry, 1953, Renssalaer Polytechnic Institute; A.M. (Physics), 1954 and Ph.D. (Chemical Physics), 1957, Harvard University; Bell Telephone Laboratories, 1959—. Since coming to Bell Laboratories, he has engaged in research and development work on microwave masers, including the maser for the Telstar ground station receiver. Recently he has extended his work into optical masers.

B.S.T.J. BRIEFS

The Use of Wollaston Prisms for a High-Capacity Digital Light Deflector

By W. J. TABOR

Manuscript received March 23, 1964

A digital light deflector was recently proposed by T. J. Nelson[1] in which n optical modulators and n uniaxial crystals were used to provide 2^n positions of the beam. Each uniaxial crystal was used to deflect a beam of light into either of two beams, with the light at the output remaining parallel to the input but displaced by an amount proportional to the thickness of the crystal. For a large number of positions it was found that a lens had to be employed in order to focus the beam into a small spot. The use of the lens requires that converging or diverging light must pass through the uniaxial crystal. In this brief we point out that this converging beam, when passing through the crystal as an extraordinary ray, is subjected to an index of refraction which varies rapidly with angle. The digital light deflector using this type of deflection therefore has appreciable image distortion, and the limiting spot densities that can be achieved are less than would be predicted if only diffraction were important.

A system which has less image distortion than the above can be constructed by using Wollaston prisms[2] (similarly Rochon or Senarmont prisms) and parallel light. (See Fig. 1.) Only the first two prisms are shown in Fig. 1. A Wollaston prism has the property that a collimated beam incident to its first face will be deviated, depending on its polarization, into either of two collimated beams. In this case the output beams will have an angular separation, whereas in the Nelson proposal the two output beams were laterally displaced without a change in angle. In general, n prisms will be used to provide 2^n resolvable angles, and these angles can be displayed as 2^n focused spots of light by means of a lens. This system has several advantages: (1) the parallel bundles of light contain only the angular variation implied by diffraction theory, which can be made small; (2) the light rays are always either nearly parallel or perpendicular to the optic axis where the index varies only slowly with angle and (3) the Wollaston prism contains much less material than the equivalent blocks of uniaxial crystal.

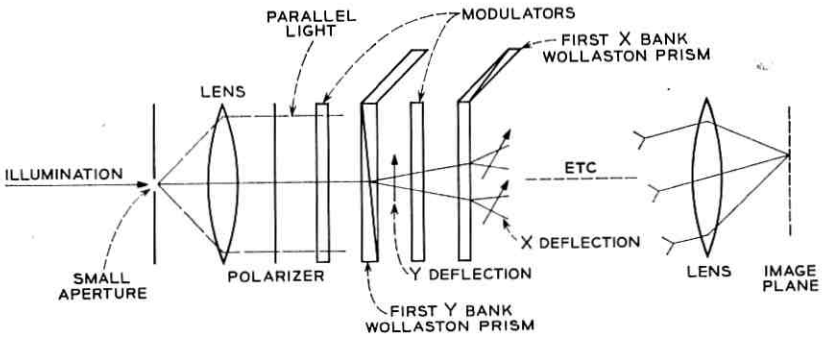


Fig. 1 — Use of Wollaston prisms in a digital light deflector.

One disadvantage of the Wollaston prism is that the deviation angle is not constant as the incident angle of the parallel bundle of light is varied from the perpendicular direction. This may necessitate placing the Wollaston with the smallest deviation first, the next largest second, etc. With this arrangement no Wollaston prism will have an incident angle differing from the perpendicular direction by amounts as large as the deviation angle of that prism. In this way the problem of the varying deviation angle should be minimized.

REFERENCES

1. T. J. Nelson, Digital Light Deflection, B.S.T.J., this issue, p. 821.
2. Jenkins, F. A., and White, H. E., *Fundamentals of Optics*, third ed., McGraw-Hill, New York, 1957, p. 504.