THE BELL SYSTEM
TECHNICAL JOURNAL

Volume 47 December 1968 Number 10

"Mental Holography": Stereograms Portraying Ambiguously Perceivable Surfaces

By BELA JULESZ and STEPHEN C. JOHNSON

(Manuscript received June 28, 1968)

An algorithm has been devised that can generate the same stereogram for two (or more) selected surfaces. Prior to this development the only known ambiguous stereograms have been periodic grid patterns that could be perceived at various parallel depth planes. The new algorithm, however, can portray two (or more) selected surfaces of general shapes and the observer can perceive each of these surfaces, but only one at a time. The technique is an extension of random-dot stereograms and it is shown that in most cases adequate degrees of freedom remain for coloring the random-dot texture. Since these ambiguously perceivable stereograms permit the portrayal of both the visible and the hidden surfaces of objects, they are analogous to holograms. However, in the case of holograms the observer has to inspect them from various positions, while for ambiguous stereograms it is the mind of the observer that wanders around.

I. INTRODUCTION

There are many limitations in portraying our three-dimensional environment on a two-dimensional surface. One of the primary shortcomings is the difficulty of effectively representing the hidden surfaces of objects together with the visible ones. Perspective drawings and even stereoscopic images did not alleviate this limitation. The cubist

artists tried to place the hidden surfaces of their objects side by side with the visible ones, but the results were rather confusing. The usual representation by multiple projections solves the problem geometrically yet it is very difficult to combine these projections into a unified spatial percept. The invention of panoramagrams using lenticular screens by Ives[1] and multiple lens arrays (fly's eye) by Lippmann[2] portrayed the three-dimensional objects within a wide angle, but in order to inspect some hidden surfaces the observer still has to move around the panoramagram. Furthermore, certain hidden surfaces (such as the boundaries of inside cavities of opaque objects) stayed invisible. Since holograms (invented by Gabor[3] and improved by Leith and Upatnieks)[4] are similar to panoramagrams only more simply made, some change in the geometry of the optical rays has to be initiated in order to obtain various stored organizations. This is usually achieved by the observer when inspecting the hologram from various angles.

This article describes a method of generating ambiguously perceivable stereograms. These stereograms contain several predetermined surfaces, out of which only one can be perceived at a time. Moreover, some surfaces might be "internal," hidden from every angle. In order to inspect the various surfaces the viewer can sit still; it is his mind that wanders around the object.

II. AMBIGUOUS STEREOGRAMS

Random-dot stereograms, introduced in this journal in 1960, have shifted interest to the problem of how the visual system resolves ambiguities.[5] Indeed, in random-dot stereograms, hundreds of dots are presented on a horizontal line in the left and right eye's views, and the observer is confronted with ambiguities as of which of the many dots in the left retinal projection corresponds to a given point in the right retinal projection. This problem is further amplified for random-dot stereograms, since no monocular familiarity or depth cues are provided that would aid perception. Findings with random-dot stereograms have shown that the visual system selects that organization which yields dense surfaces, and all other possible organizations pass unnoticed.[5-7]

Would it be possible to create random-dot stereograms which portray simultaneously more than one dense surface? If possible, which organization would be preferentially perceived? That there are ambiguous stereograms is well known in psychology. Grid-like structures

containing vertical bars of constant periodicity (such as wallpapers and old-fashioned radiators) can be fused at multiple depth levels since the binocular disparity can be any integer multiple of the horizontal periodicity. In Ref. 7, such a periodic random-dot stereogram has been demonstrated which could be perceived as a plane in front of or behind the real plane of the printed page. These periodic random-dot patterns have been successfully used in studies of perception[7, 8] yet are limited in their scope, since only parallel planar surfaces can be portrayed by this method.

This report describes a general algorithm which generates a single stereogram portraying two (or more) specified surfaces. That such stereograms can exist is based on the fact that certain areas are seen by one eye only and thus can be freely selected for one surface. Also segments in which the two (or more) surfaces coincide add to the degrees of freedom, since these surfaces can be covered by any random texture at will. In general, there is no restriction on the surfaces to be portrayed simultaneously. Provided that the number of surfaces is restricted to two and the resolution is fine (the number of samples is large) there is enough freedom in choosing the texture elements that the formation of monocularly perceivable short periodicities can be prevented.

Before the general algorithm is discussed in detail, a brief illustration is given of the basic idea in Fig. 1. Fig. 1 (solid line shows how points P_i^1 and P_i^2 (belonging to surface B) have to be colored identically to P_i (the original point of fixation, belonging to surface A), in order to obtain the same retinal projections for both surfaces. Fig. 1 (dotted line) shows how this requirement forces P_i^3 and P_i^4 to be colored the same. Fig. 1 (dashed line) shows the next step P_i^5 and P_i^6. As this procedure is continued the algorithm assigns the color of P_i to a gamut of points belonging to the two surfaces.

For a new point of fixation on either surface A or B, then there are two possibilities. Either this point of fixation already has been assigned a color by the previous algorithm or can be colored freely some brightness value. In the latter case the above described algorithm is continued which assigns this chosen color to another set of points belonging to the two surfaces. This procedure continues until each point on either surface has been colored. It is interesting to notice that the color of any point depends on some *global* relationship between surfaces A and B.

In this algorithm, as long as only two surfaces have to be portrayed each, iterative step assigns one new constraint. In the case of three surfaces, each iterative step generates two new constraints which in

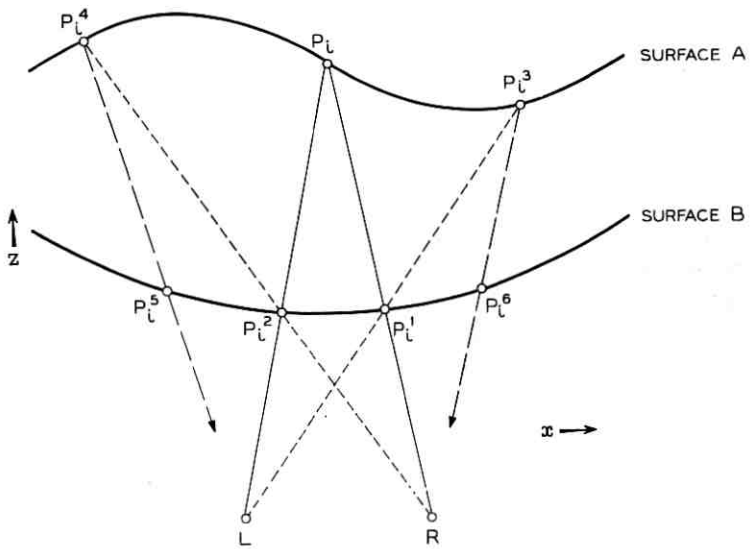


Fig. 1 — Iterative steps in the algorithm that generates the same stereoscopic projections for surfaces A and B.

turn generate 4, 8, 16, . . . new constraints. This exponential proliferation of constraints drastically reduces the degrees of freedom for three or more surfaces. Therefore, in the forthcoming examples we restrict ourselves to two surfaces to be portrayed.

The study of how such ambiguous stereograms are perceived can be of considerable scientific value but also permits some useful applications. It is now possible to portray hidden surfaces of objects together with the visible ones. Since in these ambiguous stereograms only one surface can be seen at a time, multivalued functions of two variables can be portrayed. For instance, if one surface is chosen to be the front view of an object (such as a statue or a machine part) and the other surface is chosen to be the rear view, one can obtain an entire 360 degree impression of an object, since the perception of the two surfaces may be alternated at will.

III. THE GENERAL ALGORITHM

The algorithm is an extension of the technique of random-dot stereograms.[5, 9] The following simple example will give an insight into the workings of the general algorithm. The two surfaces, A and B,

to be portrayed by the stereogram are given in the x-z plane in Fig. 2 and for simplicity are selected as cylindrical; (that is, $z = f_A(x)$ and $z = f_B(x)$, independent of y). Since stereoscopic vision operates on corresponding single rows in the two views, the algorithm given here applies to any surface (not only to cylindrical ones).

In order to construct the stereogram we must specify the textures $T_L(x)$ and $T_R(x)$ for the left and right images, respectively. The right image of the stereogram is selected as the perpendicular projection of Fig. 2, while the left image is viewed from an angle of 45 degrees.

Examine for a moment the case where we have just one figure, $z = f(x)$. We may pick the texture $T_R(x)$ at random; the texture $T_L(x)$ is now basically chosen as follows:

$$T_L[x + f(x)] = T_R(x). \tag{1}$$

This just expresses the fact that a point x seen in the left image is displaced horizontally by a distance $f(x)$ when viewed in the right image.

There are two necessary qualifications to the above rule:

(*i*) If $x + f(x)$ and $x' + f(x')$ are equal for x unequal to x', in fact

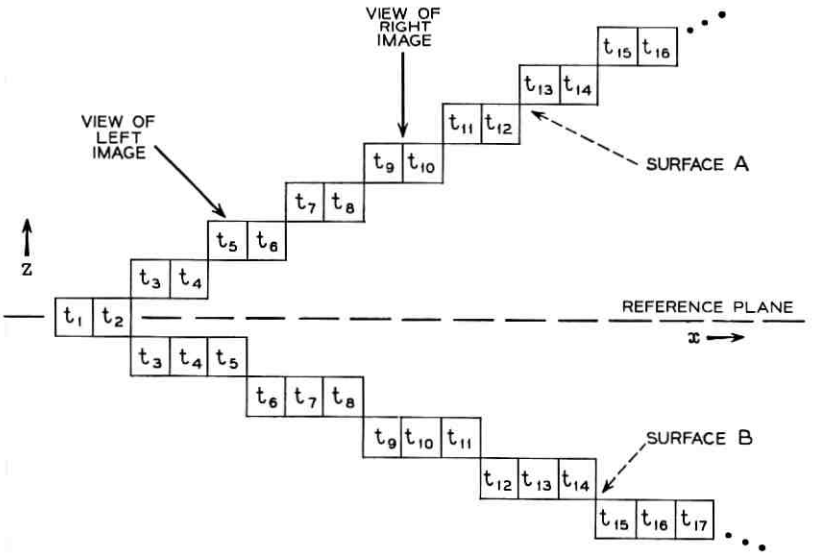


Fig. 2 — A simple example of two surfaces to be portrayed. (The cross section is indicated in the x-z plane.)

only the point corresponding to min (x, x') is seen by the left eye, the other being "in the shadow." The constraint (1) thus does not hold for the larger of x and x'; we say the larger is obscured in this case.
(*ii*) In the event that, after applying all the constraints, there are some values of $T_L(x)$ not determined by T_R, these values may be chosen at random.

Julesz and Miller developed these ideas extensively.[9]

Now to color two figures, $z = f_A(x)$ and $z = f_B(x)$, we apply a similar method; the main difference is that there are more constraints, so that the right image can no longer be chosen at random. The images must represent both A and B; thus, if T_L and T_B represent the left and right textures as above we have basically

$$\begin{aligned} T_L[x + f_A(x)] &= T_R(x) \\ T_L[x + f_B(x)] &= T_R(x) \end{aligned} \tag{2}$$

for all x. Qualification i is still valid, and may serve to eliminate one or both of the above constraints for certain values of x. From this it is also seen that if x and x' are distinct, and neither x nor x' is obscured, then

$$f_A(x) + x = f_B(x') + x' \text{ implies } T_R(x) = T_R(x') = T_L[x + f_A(x)]. \tag{3}$$

These are also easily seen to be the only constraints on T_R.

Once again, qualification *ii* is valid; anything not explicitly constrained may be chosen at random.

A simple example of the computational ease of this algorithm is given in Table I. Notice that x, $f_A(x)$, and $f_B(x)$ are given in the first three rows. The rows labeled L_A and L_B are formed by putting the integer x into positions $x + f_A(x)$ and $x + f_B(x)$, respectively, subject to qualification i. A * indicates a position that has been "uncovered" by the shifting process—the texture here may be chosen at random if not otherwise constrained.

We may read off our constraints directly from Table I; if $L_A(x) \neq L_B(x)$, and neither $L_A(x)$ or $L_B(x)$ is a *, then

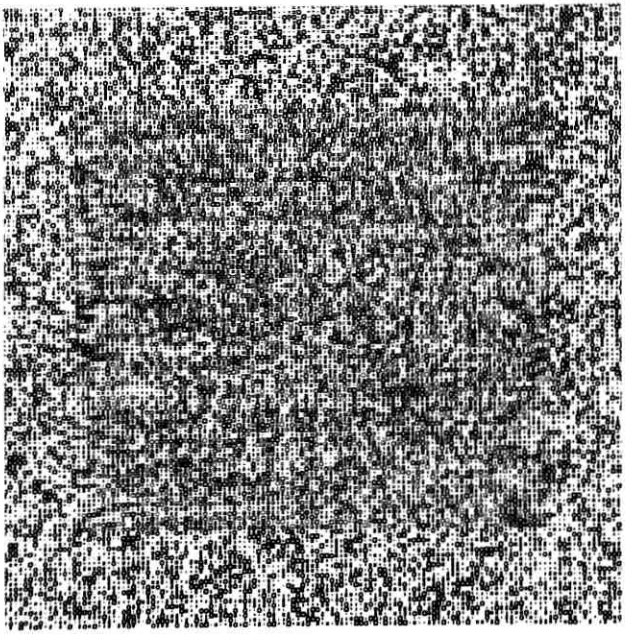
$$T_R[L_A(x)] = T_R[L_B(x)]. \tag{4}$$

After all such constraints have been applied to the right image, the left image can be generated by reference to equation (2) and qualification *ii*. Table I gives final values for T_R and T_L in terms of randomly chosen texture values $t_1, t_2, t_3, \ldots$.

TABLE I PROJECTIONS*

x	1	2	3	4	5	6	7	8	9	10	11	12	13	14
$f_A(x)$	0	0	1	1	2	2	3	3	4	4	5	5	6	6
$f_B(x)$	0	0	−1	−1	−1	−2	−2	−2	−3	−3	−3	−4	−4	−4
$L_A(x)$	1	2	*	3	4	*	5	6	*	7	8	*	9	10
$L_B(x)$	1	2	4	5	7	8	10	11	13	14	16	17	19	20
$T_L(x)$	t_1	t_2	t_4	t_3	t_4	t_6	t_3	t_5	t_9	t_4	t_6	t_{10}	t_7	t_3
$T_R(x)$	t_1	t_2	t_3	t_4	t_3	t_5	t_4	t_6	t_7	t_3	t_5	t_8	t_9	t_4

* The left and right projections of the surfaces given in Fig. 2 before and after the constraints.



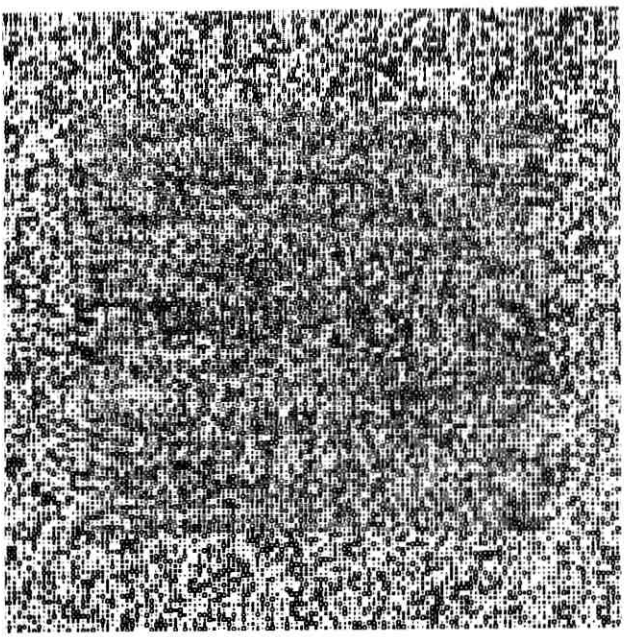


Fig. 3 — Stereogram of unambiguous pyramidal staircase in front of the printed page with a 128 × 128 picture element resolution. There are nine steps, altogether. Use the viewers fastened inside the back cover of this issue to see stereoptical effect. Place the red filter over your left eye.

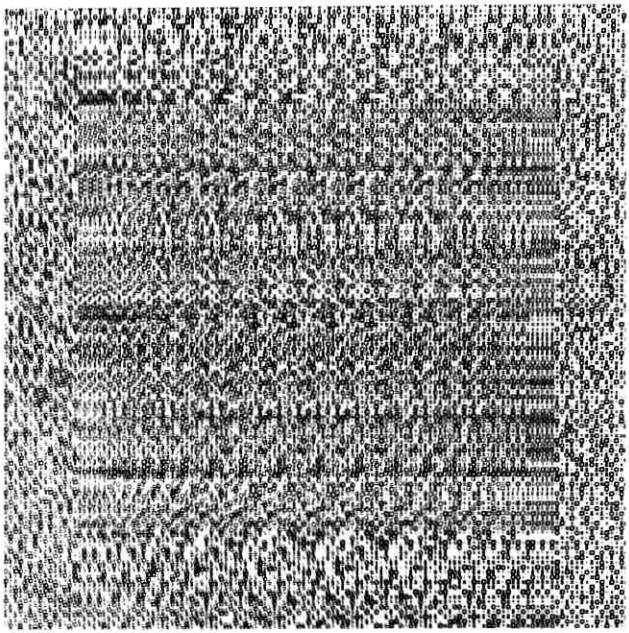
IV. CONCRETE EXAMPLES

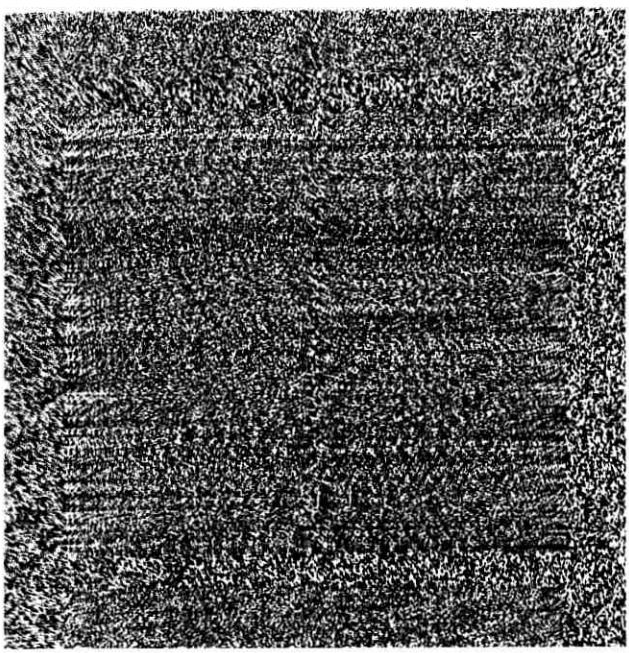
The demonstration will be quite general and the 1000 × 1000 dot resolution permits the portrayal of surfaces having complex shapes. The only restriction on the surfaces will be the use of cylindrical shapes. For this case, the algorithm determines the *same* constraints *for each row* but of course within the degrees of freedom each row is independently colored by a random process. For general surfaces each row would have to be computed separately, which would increase the computation time (now about two minutes on a GE 645 computer) nearly a thousand times. The cylindrical surfaces have another advantage; they permit us to use two unambiguous surfaces at the top and bottom margins of the stereogram respectively, to facilitate perceptual reversals for the unexperienced observer.

Besides the 1000 × 1000 dot resolution there are a few stereograms composed of 128 × 128 picture elements. In the 1000 × 1000 dot array each dot (picture element) can take three different brightness values; for the coarser array, each picture element can take eight brightness values. This is done by using three (eight) characters (blank, period sign, degree sign, asterisk, and so on) of the General Dynamics (Stromberg Carlson) 4060 microfilm printer. For the 128 × 128 array the probability of each of the eight characters is equal (⅛). For the 1000 × 1000 array the probability of using the light and heavy period signs is 0.05 while the probability for the blank is 0.9. Thus the average number of portrayed dots in these stereograms is 10^5, which is within the resolution capabilities of the printing process.

Figure 3 shows an unambiguous pyramidal staircase in front of the real plane of the printed page with a 128 × 128 picture element resolution. In Fig. 4 the left and right images of Fig. 3 have been interchanged and the pyramidal staircase is descending behind the printed plane. To view these illustrations use the anaglyphoscopes fastened inside the back cover of this issue. Put the red filter over your left eye.

Fig. 4 — Stereogram, identical to Fig. 3, except that the left and right images have been interchanged. The unambiguous pyramidal staircase is descending behind the page.





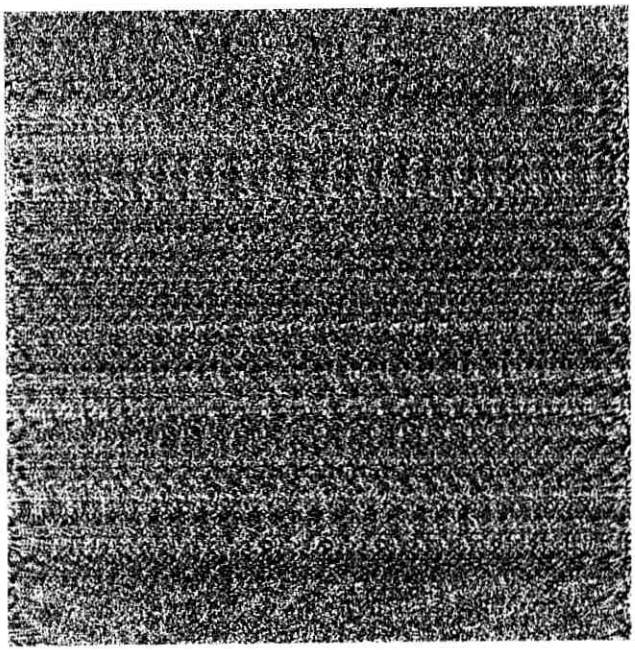
← Fig. 5 — Ambiguous stereogram, that contains both the pyramidal staircases in front of and behind the plane of the printed page as given in Figs. 3 and 4. Both of these organizations can be obtained when sterosopically viewing Fig. 5, but only one at a time.

If you have good stereopsis, the depth should become apparent to you within several seconds.

Figure 5 demonstrates an ambiguous stereogram that contains both the pyramidal staircases above and below the printed plane as given in Figs. 3 and 4. Both of the organizations can be obtained when stereoscopically viewing Fig. 5, but only one at a time. In a brief study Fig. 3 or Fig. 4 have been shown to 21 subjects who have never seen these stimuli before. After viewing one of these unambiguous stimuli for a minute the ambiguous stimulus of Fig. 5 has been shown. Ten subjects perceived that organization in the ambiguous stereogram which corresponded to the previous unambiguous organization. Nine subjects perceived the ambiguous stereogram always as the descending staircase, while two subjects always as the ascending staircase. There was no attempt on our part to train these subjects to learn to reverse the organization. On the other hand, the reader can learn easily the reversal if he alternates between Figs. 3 or 4 prior to viewing Fig. 5. Convergence movements of the eyes can influence the reversals, but it might require careful studies to determine whether reversal could be obtained while the eyes are immobilized. In Fig. 5, the maximum disparity is 8 picture elements. The degrees of freedom are 46 (out of 128).

The degrees of freedom can be greatly increased and the staircasing greatly reduced by increasing the resolution to 1000 × 1000 dots. Such an ambiguous stereogram is shown in Fig. 6, portraying a single wedge behind the printed plane and two wedges in front of the printed plane. The maximum depth is ±60 picture elements (dots), and to facilitate perceptual reversal a 150 picture element wide margin in the upper and lower portions of the images contains the unambiguous surfaces A and B, respectively. A 50 picture element wide gap at zero depth level separates the unambiguous surfaces from the ambiguous organiza-

← Fig. 6 — Ambiguous stereogram of 1000 × 1000 picture element resolution with unambiguous margins in the upper and lower areas. Surface A is a wedge behind the plane of the printed page, while surface B is a double wedge in front of the page. Either one of the two surfaces can be perceived at will when stereosopically viewed, yet reversal can be aided by viewing the upper or lower margins, respectively.



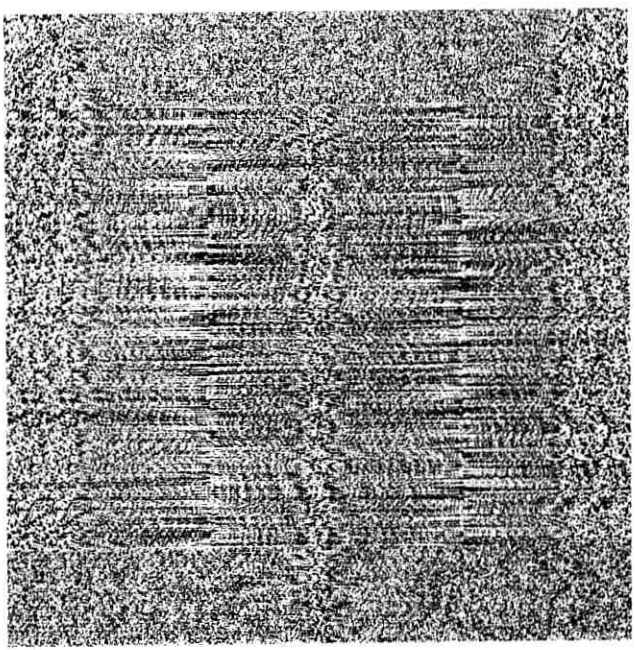


Fig. 7 — Ambiguous stereogram with unambiguous margins in the upper and lower areas. Surface A is a horizontal plane in front of the printed plane, while surface B is a wedge behind the printed plane. For viewing instructions see Fig. 6.

tion. When looking at the upper portion of the fused stereogram the unambiguous ascending wedges usually carry with them the percept of the ambiguous wedges, while when looking down the ambiguous organization reverses. Of course, these unambiguous margins serve only as an aid, and the ambiguous organization can be reversed at will when the margins are covered up. This stereogram is similar to Fig. 5, yet because of increased resolution the degrees of freedom are 359 (out of 1000 samples). This is adequate to portray images without excessive formation of perceivable periodic stripes.

Particularly interesting is Fig. 7, which has 1000 × 1000 resolution. Here the upper margin contains the unambiguous surface A which is a plane with 40 picture element disparity, while the lower margin contains the unambiguous surface B which portrays a descending wedge having a maximum disparity of -60 picture elements. In this example there is no gap between the ambiguous and unambiguous surfaces. Organization A is very strong and our everyday experience would suggest that when each dot of a front plane is seen by both eyes without any hidden areas present, then this plane should be the only percept. Yet, as Fig. 7 demonstrates, it is relatively easy to obtain the other organization too. Here the degrees of freedom are only 99 (out of 1000), yet in spite of this low degree of freedom the image quality is very good.

When we try to portray more than two surfaces, the degrees of freedom rapidly diminish. On the other hand, some of the stereograms with two surfaces yield some additional percepts. For instance, in Fig. 7 after the front plane is perceived, sometimes the percept of an ascending wedge above the plane can be obtained too.

Figure 8 shows a stereogram that can be perceived in many different ways as illustrated in Fig. 9. The unambiguous margins correspond to the two shapes in Fig. 9a, but the reader might obtain the other percepts as well. Obviously the strongest constraints are obtained for

Fig. 8 — Ambiguous stereogram with unambiguous margins in the upper and lower areas. Two slanted planes (above the plane of the printed page) that interest each other are portrayed. Fig. 9 shows the various percepts which can be obtained.

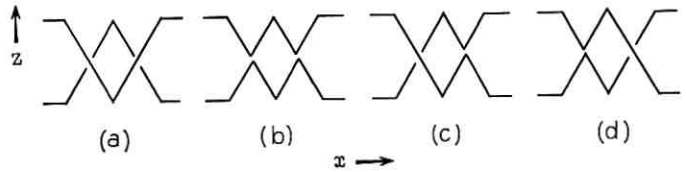


Fig. 9 — Schematic illustration of the various ways Fig. 8 can be perceived. The cross sections in the x-z plane are indicated.

parallel surfaces A and B with a few dots separation in depth. This occurs near the intersection of the two surfaces, yielding visible clusters of dots having the same brightness values.

Figure 10 portrays a cosine function and a cosine function of lesser amplitude and half periodicity. Since both surfaces are in front of the surround and close together, an interesting perceptual phenomenon can be experienced. For the previous demonstrations only one

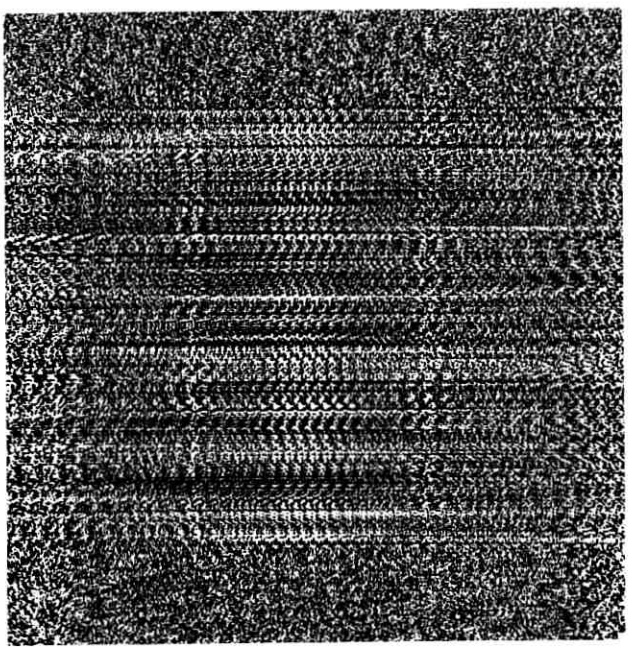


Fig. 10 — Ambiguous stereogram with unambiguous margins in the upper and lower areas. It portrays a cosine function and a cosine function of lesser amplitude (height) and half periodicity. Both surfaces appear in front of the printed plane.

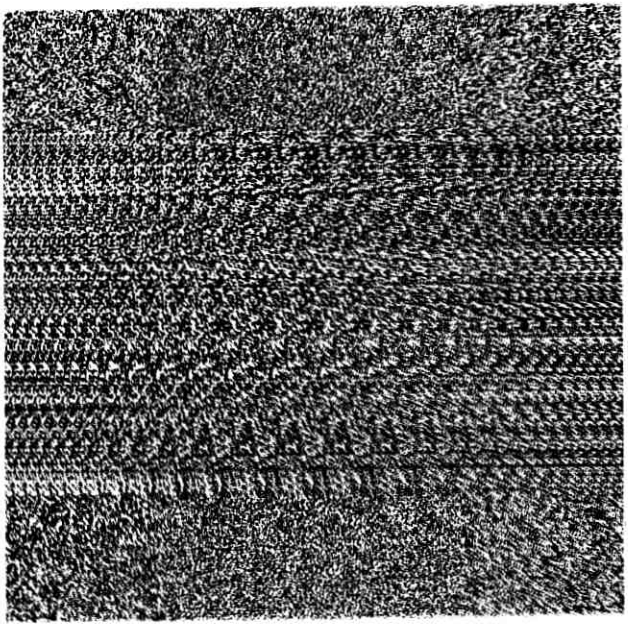
organization could be perceived at a time and considerable amounts of convergence movements had to be initiated in order to get rid of the prevailing organization and to bias the other organization. In Fig. 10 it is possible to retain one organization and meanwhile start to perceive the other organization. The perceptual effect is that of a transparent surface behind which another transparent surface is seen. However, this double perception is not a stable state and it is easier to perceive only one organization at a time. The degrees of freedom are 119 (out of 1000).

Finally Fig. 11 shows a case in which three surfaces are portrayed. Besides a cosine function in front and behind the surround there is a plane with zero disparity. Interestingly enough for this special case the degrees of freeedom are not additionally reduced. As long as the surfaces A and B are each other's mirror images and the third surface is a plane with zero disparity, it is possible to portray three surfaces without additional constraints. The degrees of freedom are 78 (out of 1000). An unambiguous rectangle at the top and bottom margin is presented at ± 70 picture element disparities in order to aid perceptual reversal.

V. CONCLUSIONS

From these demonstrations it is possible to see some of the limitations of the ambiguous random-dot stereograms. In order to avoid short periodicities the surfaces to be portrayed should not intersect or come closer than a few picture elements in the z direction. The worst case is if the two surfaces are one picture element apart. In this case the periodicity is one and the two surfaces are formed of horizontal lines having the *same* color. As we have discussed the two surfaces can coincide but after separation they have to separate in a discontinuous fashion having a jump in depth of several picture elements. In Fig. 12 special care has been taken to separate the two surfaces in depth. Therefore a cosine function (to be seen in front of the plane of the printed page) has been placed on a 10 picture element high pedestal. The other surface is a wedge (behind the printed plane). The degrees of freedom are 128 (out of 1000). Because of this pedestal the shortest periodicity is limited to ten and the resulting stereogram can be easily fused and reversed.

Another limitation is the rapid decrease in the degrees of freedom as the number of surfaces is three or more. By increasing the resolution of the images the absolute degree of freedom increases as well,



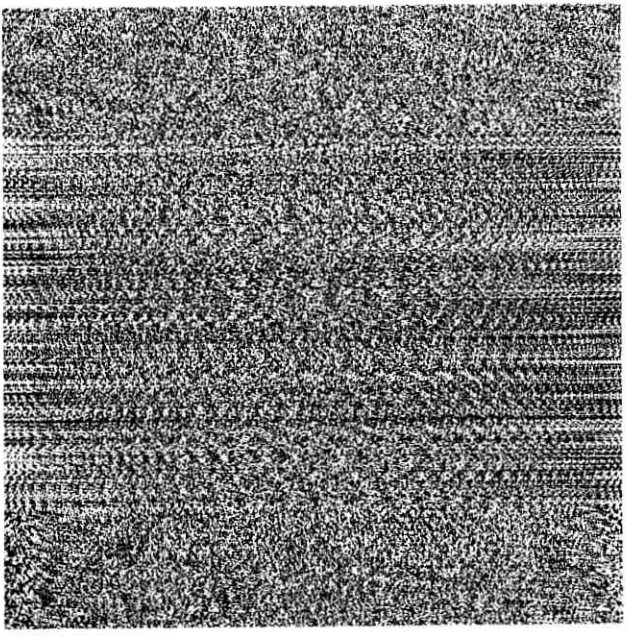


Fig. 11 — Ambiguous stereogram with two unambiguous rectangles in the upper and lower areas. It portrays three surfaces. A cosine function in front of the plane of the printed page, the same function behind the plane, and the printed plane, itself (a plane with zero disparity).

so that three or more surfaces could be adequately portrayed. Unfortunately, the resolving acuities of the eyes limit the size of individual picture elements to about 1 minute of arc. With finer image resolution more than one picture element (of different brightness levels) falls on a single receptor of the retinas and the image contrast rapidly decreases.

A third limitation of ambiguous stereograms is the unavoidable fact that because of the constraints there will be some other dense surfaces perceivable besides the selected two surfaces (as pointed out in the demonstration). Since most of these phantom surfaces are perceived at greater depth than the desired ones, there is a way to eliminate them. If the desired surfaces span the depth limits for fusion, then the phantom surfaces will be outside the region of maximum disparity for stereoscopic fusion.

As long as the surfaces to be portrayed are placed such that they stay separate in distance and the number of surfaces is two, the above technique gives satisfactory results. The obtained results are analogous to holography, but only superficially. After all, holograms contain a vast number of stereoscopic views, while ambiguous stereograms contain only a single one. Holography is based on the diffraction properties of coherent wave optics, while our technique uses plain geometrical optics. For holography the observer has to move around the hologram in order to inspect it from various angles, while for ambiguous stereograms the viewer can stand still. It is his mind that wanders around the object. Furthermore, ambiguous stereograms can portray any mathematical surface, including completely hidden ones, from any view, this could be obtained by computer-aided or computer-generated holograms as well, but would require more effort.

The advantages of holography or lenticular screen panoramagrams could be combined with ambiguous stereoscopy. It might be possible to generate stereograms by computer and place them behind a lentic-

Fig. 12 — Ambiguous stereogram with wide unambiguous margins. It portrays a cosine function in front of the printed plane and a wedge behind the plane.

ular screen such that each separate stereoscopic view is constrained by our algorithm. Since each view is independent from each other, no further reduction of the obtainable degrees of freedom will result. Unfortunately, the generation of ambiguous stereograms for the most general surfaces is at the limit of present computer economies and to compute hundreds of them for a single portrayal is certainly impractical. Yet, with next generation computers these and similar representations can be tried.

The emphasis of this article has been on the reporting of a new tool for pictorial representations and the obvious psychological implications have been only briefly mentioned. We can now study, for example, whether convergence motions of the eyes are necessary to destroy an existing perceptual organization in order to reverse to the other one. Furthermore, each of the ambiguous organizations can be biased by a few randomly introduced unambiguous picture elements in order to counteract natural bias. (This biasing technique has been successfully tried for periodic random-dot patterns in a study of perception time.[7, 8])

A pilot study was reported above which showed that if one of the organizations has been first presented as an unambiguous stereogram, then the perception of the ambiguous stereogram could be influenced accordingly. It remains to be seen whether prior auditory or tactile information would similarly influence perception.

Originally, random-dot stereograms were conceived to remove from the monocular images all the familiarity cues and Gestalt factors that influence perception in uncontrollable ways. It is therefore somewhat unexpected that this further development of the technique in the form of ambiguous random-dot stereograms seem to provide a powerful tool for the study of Gestalt factors. In this instance the shapes are the configurations of the surfaces in depth. Questions exemplified in Figs. 8 and 9 can be studied, such as whether good Gestalt or good continuation outweigh the reduction of disparity. Reversible figures have been frequently used in perceptual psychology such as Necker cubes (two-dimensional outline drawings of a cube), ambiguous staircases, and so on. However, these stimuli have been selected from a small repertoire and exploited certain ambiguities inherent in two-dimensional drawings. That ambiguities can be produced in three-dimensions without practical limitations on the organizations to be portrayed is a result which seems far from trivial.

VI. ACKNOWLEDGMENT

We wish to thank Miss Roseanne Hesse for writing a microfilm plotting routine for portraying our computer programs and her assistance throughout this project.

REFERENCES

1. Ives, H. E., "A Camera for Making Parallax Panoramagrams," J. Opt. Soc. Amer., *17*, No. 6 (December 1928), pp. 435–439.
2. Lippmann, G., "Epreuves Reversibles donnant la Sensation du Relief," J. de Phys., *7*, No. 11, 4th series (November 1908), p. 821.
3. Gabor, D., "A New Microscope Principle," Nature, *161*, No. 4098 (May 15, 1948), pp. 777–778.
4. Leith, E. N. and Upatnieks, J., "Reconstructed Wavefronts and Communication Theory," J. Opt. Soc. Amer., *52*, No. 10 (October 1962), pp. 1123–1130.
5. Julesz, B., "Binocular Depth Perception of Computer-Generated Patterns," B.S.T.J., *39*, No. 5 (September 1960), pp. 1125–1162.
6. Julesz, B. and Spivack, G. J., "Stereopsis Based on Vernier Acuity Cues Alone, *Science, 157*, No. 3788 (August 4, 1967), pp. 563–565.
7. Julesz, B., "Binocular Depth Perception without Familiarity Cues," *Science, 145*, No. 3630 (July 24, 1964), pp. 356–362.
8. Julesz, B., "Texture and Visual Perception," *Scientific American, 212*, No. 2 (February 1965), pp. 38–48.
9. Julesz, B. and Miller, J., "Automatic Stereoscopic Presentation of Functions of Two Variables," B.S.T.J., *41*, No. 2 (March 1962), pp. 663–676.

The Capacity of Multiple Beam Waveguides and Optical Delay Lines

By D. GLOGE and D. WEINER

(Manuscript received July 10, 1968)

The capacity of a beam waveguide can be increased by transmitting a multitude of Gaussian beams in such a way that they are clearly resolved at the receiving end. Various systems with maximum capacity but different crosstalk sensitivity are discussed. Linking the available channels end to end in an optical cavity produces a delay line or storage device. An optimized system is described which has surprisingly large storage capacity. For the analysis of both lens guides and optical cavities, a phase space representation of Gaussian beams is used which avoids cumbersome mathematics.

I. INTRODUCTION

A Fabry-Perot interferometer with curved mirrors can be used as an optical delay line by inserting a laser beam through a small center hole in one mirror.[1] The beam performs many off-axis round trips before leaving the interferometer through the entrance hole.[2] Reference 1 suggests that the injection and retrieval of the beam could be improved by mismatching beam and cavity. A systematic study is carried out here to find the longest folded path that starts and ends in the center hole, thus optimizing the system for maximum storage capacity.

Very similar to this problem is the analysis of a periodic lens guide in which many beams are to be transmitted in such a way that they are clearly resolvable at the receiver end. One such system is a transmission link that forms an image array of modulators in the receiver plane. The possible density of channels is given by the number of resolvable spots in this plane.[3]

The investigation of all possible Gaussian beams transmitted simultaneously in a guide will show that this is only one among many possible systems. All these systems exhibit the maximum theoretical

capacity as given by the classical limit,[4] but may be affected differently by guide imperfections.

The first part of this study will outline a simple geometrical method of describing gaussian beams avoiding the cumbersome mathematics connected with gaussian beam optics.[5] Based on this method, it will be easy to find the optimum storage cavity and to investigate various multiple beam transmission systems.

II. PHASE PLANE AND PHASE SPACE

In continuous or periodic guiding media, the "phase space" representation of paraxial rays is very convenient. Consider, for example, the two-dimensional continuous lens-like medium in Fig. 1a in which the index of refraction is a function of the transverse coordinate only:

$$n(x) = n_0\left(1 - \frac{1}{2}\frac{x^2}{\Delta^2}\right). \tag{1}$$

Call Δ the "focusing parameter." The paraxial ray solutions are sine waves with the period

$$P = 2\pi\Delta \tag{2}$$

as shown in Fig. 1b.[6] Figure 1c shows a "phase plane" in which every ray of Fig. 1b is represented by a point. The coordinates of the points

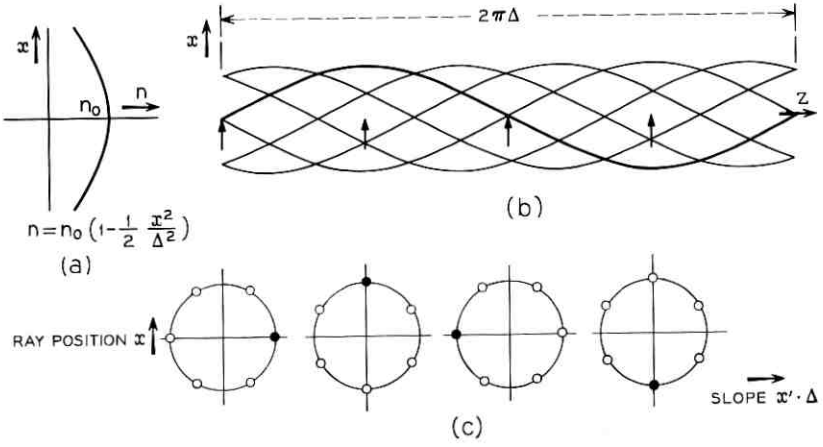


Fig. 1 — Rays in a homogeneous guiding medium. (a) The square-law index profile. (b) Rays oscillating with various phases. (c) The corresponding points in the phase plane.

correspond to the position x and the slope x' of the ray multiplied by Δ. As the rays proceed in the square-law medium, the points orbit around the origin of the phase plane while their position with respect to one another stays the same.

Steier[7] has shown that for every gaussian light beam one can find a packet of rays equivalent to this beam in the sense that the packet envelope gives the beam width and the average ray slope is perpendicular to the beam phase front. The ray packet may be represented by an array of points in the phase plane. Consider, for example, a fundamental gaussian beam propagating in the square-law medium of Fig. 1. The $1/e$-half width of such a beam is

$$w = \left(\frac{\Delta\lambda}{\pi}\right)^{\frac{1}{2}} \tag{3}$$

where λ is the optical wavelength. The equivalent ray packet is basically the one shown in Fig. 1b with ray amplitudes w. The corresponding points in the phase plane occupy a circle with radius w similar to the presentation in Fig. 1c.

Following these arguments, any gaussian beam—varying in position, slope, or width along the guide—may be represented by its array of points in the phase plane. The points form a "phase spot" in the phase plane whose shape and position determine the beam parameters. Once the phase spot is known at one point along the guide, it can be found for any other point by simply rotating the phase plane. The correspondence rules between the phase spot and the beam parameters follow from Steier's ray racket equivalence and are explained in the following examples.

Figure 2 shows a gaussian beam of width w entering the guide with a slope α. The beam phase front is tilted by α and consequently the average slope of all rays in the ray packet must be α. This condition is satisfied by a circular phase spot displaced horizontally by $\alpha\Delta$. As the beam proceeds in the guide, the phase spot orbits around the origin of the phase plane. Projection of the phase spot on the vertical axis yields the beam width and position. The horizontal displacement determines the slope. Notice that Fig. 2 and the following figures are two-dimensional beam representations. The phase spots should not be confused with a cross-sectional view of the beam.

Figure 3 shows a beam that enters the guide with a phase front curved with a radius R. Consequently, the average slopes of the equivalent rays vary linearly across the ray packet. In the phase plane horizontal

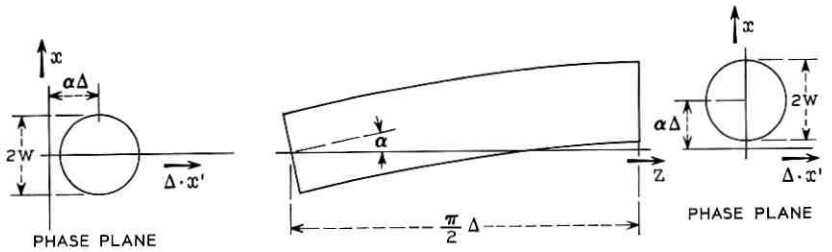


Fig. 2 — The phase spot for a beam entering at an angle α.

slices of the phase spot are displaced horizontally by $x\Delta/R$ according to their position x which distorts the circular phase spot to an ellipse. Notice that the area of the phase spot is not changed by this process. From equation (2) one finds that this area is $\pi w^2 = \Delta\lambda$. It is the same for any gaussian beam of a given wavelength in a given guiding medium. A beam, for example, that enters with a plane phase front and a half width $u \neq w$ has an elliptic phase spot with the principal axes u and

$$v = \Delta\lambda/\pi u. \tag{4}$$

If the guiding medium is not homogeneous along the z axis but a periodic sequence of lenses, the phase plane method is still valuable though, with the same convenience, the beam can only be described in the planes of the lenses and not in the sections between. This, however, is in general sufficient because, no matter what the features of the gaussian beam, it will always be largest at the lenses and therefore it will be this width that determines the aperture of the whole system.

For a periodic sequence of lenses with focal length f, spaced at a distance d, the convergence parameter is[6]

$$\Delta = d/\sin \Phi \tag{5}$$

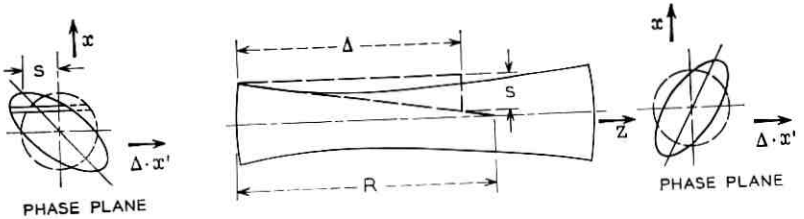


Fig. 3 — The phase spot for a beam entering with a curved phase front.

with

$$\cos \Phi = 1 - d/2f \tag{6}$$

and the equivalent ray period is $2\pi\, d/\Phi$.

For thin biconvex lenses, it is the symmetry planes of the lenses where the Gaussian beam can be defined most conveniently. Beam width and phase front curvature in this plane determine the equivalent phase spot. Counterclockwise rotation of the phase spot by an angle Φ corresponds to passage from one lens to the next.

The phase plane method may also be extended to nonperiodic structures. It is restricted, however, to the paraxial approximation, to square-law guiding profiles (including uniform dielectrics), and to coherent beams with Gaussian intensity profile and spherical phase fronts.

Notice that the phase plane considers only deflection and displacement in x direction and that a similar definition exists for the y coordinate. The two phase planes combined yield the four-dimensional phase space, and the phase spot becomes a four-dimensional structure.

III. SPATIALLY INDEPENDENT CHANNELS

The capacity of a beam waveguide can be increased by transmitting several gaussian beams separated spatially. The tolerable crosstalk determines the separation of the individual beams. For convenience, let us describe around every beam a fictitious tube, k times wider than the $1/e$ width, where k is chosen so that the crosstalk requirement is met when these tubes just touch. In practice, the main source of crosstalk will be beam distortion and scattering rather than the spread of the ideal beam. The factor k, therefore, will vary from guide to guide according to the tolerances of the guiding components.

Figure 4 shows a two-dimensional square-law medium of width $2a_x$ and the corresponding phase plane. In order for the beams to clear the guide walls, the phase spots must stay within the circle $r = a_x$ while orbiting in the phase plane. Considering that the phase spots require an area $k^2\ \Delta\lambda$ to fulfill the crosstalk conditions, it is easy to find the phase spots that make the best use of the available guide (see Fig. 4). The phase spots determine the beam parameters.

As the beams oscillate in the guide, they overlap in certain areas. There are, however, cross sections spaced by distances $\pi\Delta$ at which all beams are separated. One of these cross sections may be chosen as the receiver plane.

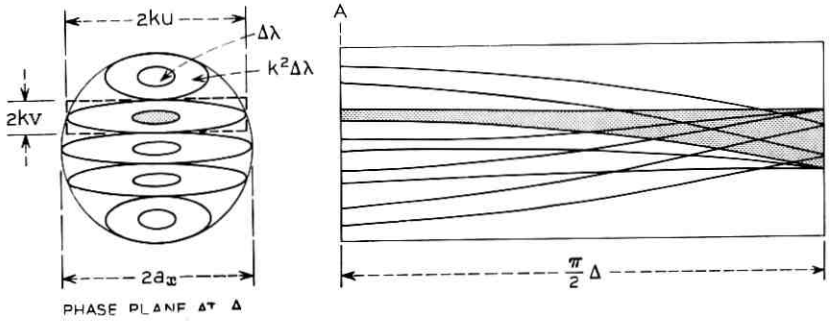


Fig. 4 — A possible distribution of phase spots in the useful phase area.

If the guiding medium is not homogeneous along the z axis, but consists of a periodic sequence of lenses of width $2a_x$, the useful phase area may be different from that in Fig. 4. In this case, discrete apertures have to be considered at the positions of the lenses. (The intermediate guide diameter in a lens guide in general is immaterial, because between the lenses the beams have a smaller cross section and separation than at the lenses). Figure 5 shows the useful phase area and the spot pattern for confocally arranged lenses.

This case surmises that, proceeding from lens to lens, the rotation of the phase pattern is exactly 90°. Even if the tolerances for the focal lengths and lens spacings are very strict, these rotations will eventually, after many lenses, get out of step with respect to the lens positions and aperturing will occur when the phase pattern is rotated at any angle in the phase plane. This situation is shown in Fig. 6. If this happens, the useful phase plane is restricted to a circular area. It seems, therefore, that Fig. 4 represents a more general case for practical applications.

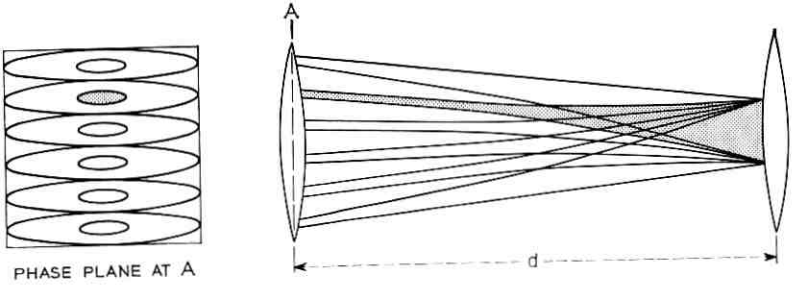


Fig. 5 — The useful phase area for a confocal imaging system.

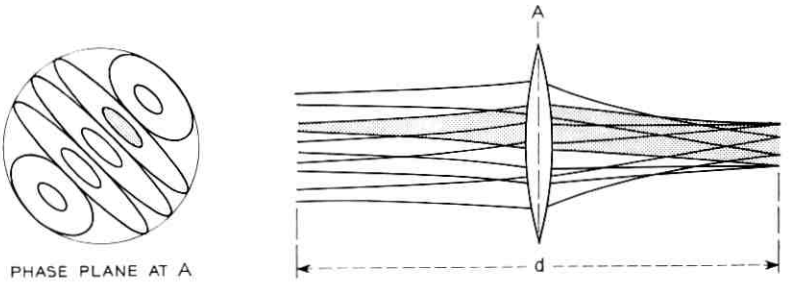


Fig. 6 — The phase pattern is out of step with the confocal lens position.

The number of phase spots that can be fitted into the circular area πa_r^2 of Fig. 4 is approximately

$$n = \frac{\pi a_r^2}{4k^2 uv} = \frac{\pi^2 a_r^2}{4k^2\,\Delta\lambda}, \tag{7}$$

assuming that the area occupied by one spot may be approximated by the dotted rectangle in Fig. 4. For large numbers of beams ($n \geqq 10$), this approximation is satisfactory.

There is no reason why the beams have to be arranged the way they are in Fig. 4. There is no restriction on shape and location of the phase spots in the useful phase plane. Of course, arranged as in Fig. 4, the beams are clearly separated at distinct cross sections, which makes launching and receiving a simple and straightforward matter.

E. A. J. Marcatili of Bell Telephone Laboratories suggested the arrangement shown in Fig. 7 and demonstrated how such beams may be launched: a common lens is used for every overlapping group of beams feeding every member of the group at a different angle. At a distance $\pi\Delta/2$ from this lens, or at multiples of this distance, the

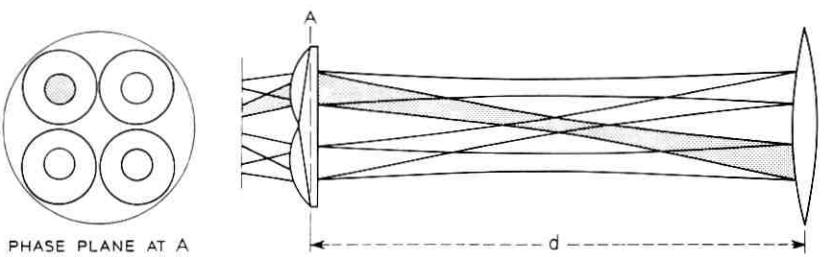


Fig. 7 — A distribution of the phase spots that minimizes distortion.

beams arrange themselves in groups again and could be separated by means similar to the transmitter scheme.

Whereas the beams in Fig. 4 vary considerably in width as they propagate along the guide, the beams in Fig. 7 constantly keep the minimum width w as given by (3). From the nature of the distortions on optical surfaces, it might be expected that large beams will be distorted more than small ones. In this case, the arrangement in Fig. 7 might be less susceptive to mutual interference of beams and consequently permit a smaller k factor and a larger capacity. For equal k there is no difference in capacity of both schemes, at least not within the accuracy of (7) which was derived for large n. Other arrangements as well as combinations of the schemes in Figs. 4 and 7 may be practical for certain cases.

By applying the phase space technique to three-dimensional systems, some of the lucidity is lost, but one can still gain some interesting results. If the beam waveguide has a cylindrical cross section of radius A, a circular area with the radius

$$a_y = (A^2 - a_x^2)^{\frac{1}{2}} \tag{8}$$

in the y-phase plane is available simultaneously with the area πa_x^2. The total useful phase space is consequently

$$S = \int_0^{\pi A^2} \pi a_y^2 \, d(\pi a_x^2). \tag{9}$$

By inserting (8) into (9), one has

$$S = \tfrac{1}{2}\pi^2 A^4. \tag{10}$$

Allowing rectangular areas for the phase spots in both the x- and y-phase plane, as in the two-dimensional example, the total capacity is found to be approximately

$$N = \frac{\tfrac{1}{2}\pi^2 A^4}{(4k^2 \, \Delta\lambda/\pi)^2} = \frac{\pi^4}{32} \frac{A^4}{k^4 \Delta^2 \lambda^2}. \tag{11}$$

If the total number of beams is large ($n \geqq 100$), the rectangular approximation for the area occupied by the phase spots is satisfactory and (11) holds independent of the way in which the beams are arranged in the guide. Figure 8 shows a nomogram based on (11) for an optical wavelength of 1 micron. Given the radius, lens spacing, and filling factor k of a lens guide, one can easily find the possible capacity. Consider, for example, lenses spaced confocally by 100 m. If their useful optical area

has a radius of 10 cm, and the filling factor is $k = 3$, approximately 300 beams could be transmitted in parallel.

Comparing (11) with M. von Laue's formula for the spatial degrees of freeedom of an optical system,[4] one finds that N approaches the classical limit for k about 1, that is, the beams would have to overlap at their $1/e$ amplitudes in order for the capacity of the guide to be fully used. There are many reasons why this limit cannot be reached in practice. Particularly important are the imperfections in the guide itself.

IV. BEAMS IN CAVITIES

The rules of gaussian beam geometry can also be used for optical cavities. Considering the cavity as a folded beam waveguide, possible beam paths can be traced using the phase plane. This way the useful capacity can be found for delay or storage applications.

Figure 9a shows a 2-dimensional square-law medium. The two plane surfaces M_1 and M_2 are highly reflecting mirrors and form an optical cavity. A beam, launched off-axis through the center hole of mirror M_1, would perform several round trips between the mirrors before hitting the entrance hole and leaving the cavity. Figure 9a un-

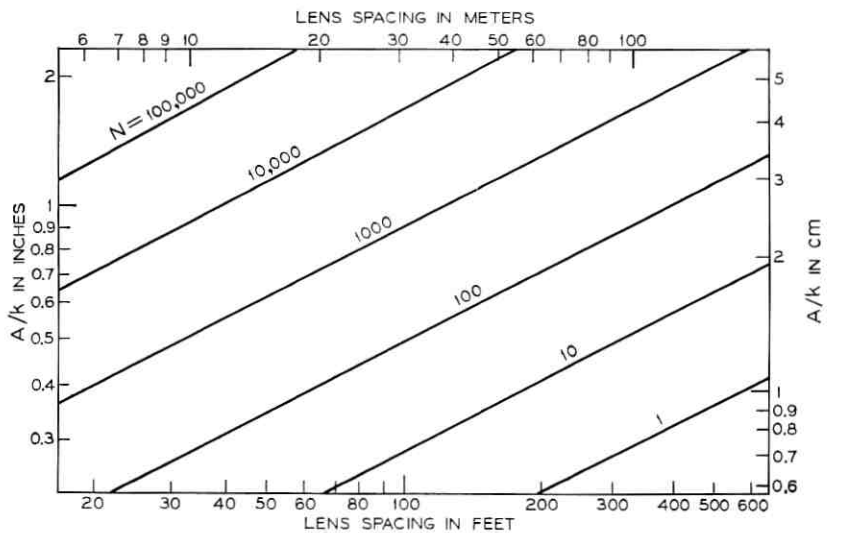


Fig. 8 — Nomogram evaluating the guide capacity for $\lambda = 1$ micron.

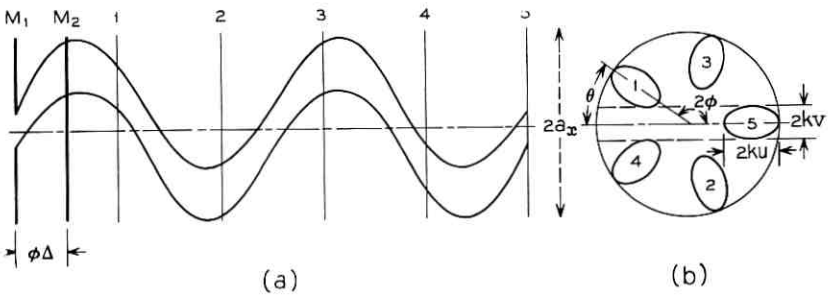


Fig. 9 — The beam path in a storage cavity consisting of a focusing medium between the mirrors M_1 and M_2. (a) The beam path unrolled along the axis. (b) The phase spots at the mirror M_1.

rolls the beam path along the guide axis. Figure 9b shows the phase spots for the cross sections 1 to 5 which correspond to reflections of the mirror M_1. In agreement with the previous considerations, the spots correspond to tubes k times wider than the $1/e$-width of the beam. It is assumed that interference between different round trips and distortion is tolerable if the phase spots do not intersect one another, the boundary of the useful phase area, and the area occupied by the center hole (area between the broken lines in Fig. 9b).

Obviously, the total number of round trips can be increased by decreasing the angle θ shown in Fig. 9b. θ is smallest when the phase spots just touch the broken lines. Also there should be an optimum shape of the phase spots for which θ is a minimum. Though the area $k^2\ \Delta\lambda$ of a phase spot is fixed, the main axes u and v can be chosen. Particularly if the cavity radius a_x is large and a large number of round trips is to be stored in the cavity, the best ellipses will be long and thin, and the center hole diameter 2kv will be small.

It is now a simple matter of geometry to calculate the exact parameters. From the requirement that spot 1 touch the broken line, one finds

$$\sin\theta = \frac{2kv(a_x - ku)}{a_x^2 - 2a_xku + k^2v^2}. \tag{12}$$

As indicated above, v will be much smaller than a_x and u for optimum systems with large capacity. By neglecting v^2 in the denominator and replacing v by (4) in the numerator, one has

$$\sin\theta \cong \frac{2\ \Delta\lambda\ k}{\pi a_x u}\,\frac{a_x - ku}{a_x - 2ku}. \tag{13}$$

The derivative $d \sin \theta/du$ vanishes for the optimum value

$$u_{\text{opt}} = \frac{a_x}{k}\left[1 - \frac{1}{(2)^{\frac{1}{2}}}\right] \tag{14}$$

and it turns out that

$$v_{\text{opt}} = \frac{\Delta\lambda\, k}{\pi a_x} \frac{1}{1 - 1/(2)^{\frac{1}{2}}} \tag{15}$$

is indeed small for large a_x. Under these conditions θ will also be small, and by replacing $\sin \theta$ by θ one finds

$$\theta_{\min} \cong \frac{\Delta\lambda\, k^2}{\pi a_x^2}\left[1 - \frac{1}{(2)^{\frac{1}{2}}}\right]^{-2}. \tag{16}$$

The maximum number of round trips is

$$n = \frac{2\pi}{2\theta_{\min}} = \frac{\pi^2 a_x^2}{\Delta\lambda\, k^2}\left[1 - \frac{1}{(2)^{\frac{1}{2}}}\right]^2. \tag{17}$$

For $n \geqq 10$, this formula gives satisfactory results.

The proper length of the cavity in Figure 8a is

$$\Phi\Delta = \tfrac{1}{2}(\pi - \theta_{\min})\Delta \tag{18}$$

with $\theta_{\min}$ from (16). If, instead of a homogeneously focusing medium, concave mirrors are used, (5) and (6) determine the mirror spacing d and the focal length f. In connection with (17) and (18) one has

$$d = \Delta \cos \frac{\theta_{\min}}{2} \cong \Delta \tag{19}$$

and

$$1 - \frac{d}{2f} = \cos \Phi = \sin \frac{\theta_{\min}}{2} \cong \frac{\pi}{2n}. \tag{20}$$

Notice that for large n the mirrors are almost confocally spaced.

Knowing the solution in the x plane, one would like to solve the three-dimensional problem by just doing the same in the y plane. Figure 10 shows equivalent phase planes for the x- and y-axes of the end mirror. Projection into the mirror plane yields the actual beam cross sections. Though it will be shown later that this arrangement is not quite optimum, Fig. 10 is very useful to calculate the cavity radius A necessary to accommodate this or, later on, an improved beam path. The maximum displacement $a_x/(2)^{\frac{1}{2}} = a_y/(2)^{\frac{1}{2}}$ occurs simultaneously in the x and y directions. The total displacement of the beam

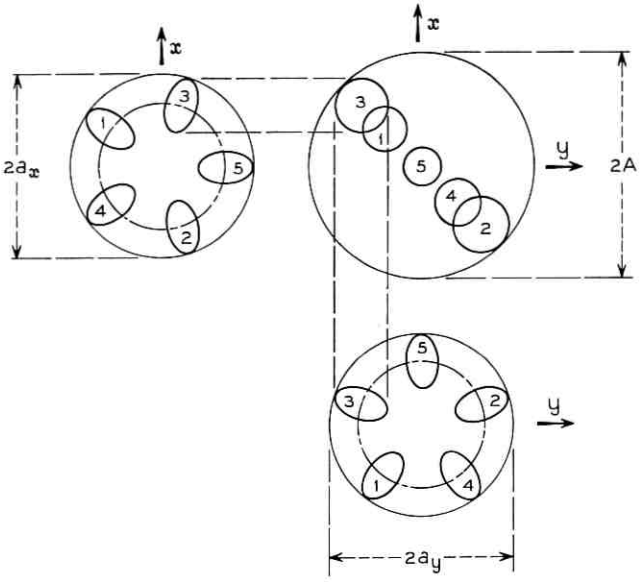


Fig. 10 — Construction of the beam cross sections at the mirror surface to determine the radius A of a cylindrical cavity.

axis is therefore $a_x = a_y$. The maximum displacement coincides with the largest beam cross section whose radius is u_{opt} as given in (14). Consequently, a radius

$$A = a_x + ku_{opt} = a_x\left[2 - \frac{1}{(2)^{\frac{1}{2}}}\right] \tag{21}$$

is necessary for a cylindrical cavity.

The capacity can be increased drastically by deliberately introducing astigmatism as described in Ref. 1. Suppose the beam behaves in the x plane as shown in Fig. 9 but, simultaneously, oscillates in the y plane in such a way that it returns to the center of the y plane already after 4 round trips. It can only leave the cavity when it is displaced neither in the x- nor the y-direction, and that happens for the first time after 20 round trips. Generally speaking, one achieves

$$N = 2n(n - 1) \tag{22}$$

transits by this method. Technically, this can be done by warping one or the other of the mirrors slightly. Writing (20) for both x and y plane and subtracting one from the other yields, by using (22),

$$\frac{d}{f_x} - \frac{d}{f_y} = \frac{\pi}{n} - \frac{\pi}{(n-1)} = \frac{2\pi}{N}. \tag{23}$$

For large n the approximation $N \simeq 2n^2$ can be made and by using (17) and (21) one has

$$N \cong \frac{2\pi^4 A^4}{\Delta^2 \lambda^2 k^4} \frac{1}{[3 + (2)^{\frac{1}{2}}]^4}. \tag{24}$$

This is about a factor of 6 less than the number given in (11), that is, the described method does not fill the total available capacity. Notice that, in Fig. 9b, there is room for exactly one more set of spots between the used spots. This space would be filled by a beam that followed the same path as the described one, but in opposite direction. The additional capacity could be exploited by reflecting the existing beam back on itself. It can be shown that, in the three-dimensional scheme, there is space for an additional totally independent path in the cavity. Both paths could be linked by an outside mirror. By reflecting the two linked beams back into itself, the number of round trips could be quadrupled.

Without considering these sophistications, let us investigate what (24) means in terms of storage capacity. For large N (19) can be used to calculate the total length of the beam path which, with the numerical factors evaluated, is

$$l = Nd = \frac{A^4}{2\, d\lambda^2 k^4}. \tag{25}$$

Surprisingly enough, this path is longest for a small cavity length d. Of course, the number of bounces (and consequently the losses) increase in a short cavity. The best mirrors available introduce a reflection loss of 0.05 percent or 43.5 dB attenuation after 20,000 bounces.[8] This corresponds to 10μs delay in a 15 cm cavity with mirrors 4 cm in diameter. If part of the loss is compensated by an amplifying material in the cavity, the number of round trips is eventually limited by scattering in the system.[9]

Perpetual recirculation of PCM information could be achieved by using an arrangement that amplifies 2π-pulses[10] or a fast saturating absorber in combination with a suitable laser amplifier.[11] In both cases only pulses of a certain length and intensity are amplified, while any other signal is attenuated. It would be sufficient to provide the amplication at a few particular parts of the folded path where it is spatially separated from other round trips. If enough amplification

of this kind is provided to make up for all losses, pulses of the proper kind could circulate in the cavity perpetually without noise building up.

If the mirrors are moved so close to one another that they touch at the circumferences, the paraxial approximation, basis for the previous calculations, loses its validity. By extrapolating (25) into this range, however, one finds some interesting, though speculative, results. The center of the confocal mirrors are now spaced by $d = 2A/(3)^{\frac{1}{2}}$. For a bandwidth b small compared to the light frequency ν, the capacity is

$$c = \frac{bl}{\nu\lambda} = \frac{0.433}{k^4}\frac{b}{\nu}\frac{A^3}{\lambda^3} \tag{26}$$

with l from (25).

It is easy to calculate the volume V of this nut-shaped cavity. It is

$$V = \frac{10}{9(3)^{\frac{1}{2}}}\pi^2 A^3. \tag{27}$$

The number of the degrees of freedom of a cavity whose dimensions are large compared to λ is independent of its shape and has the value*

$$c_{\text{th}} = \frac{8bV}{\nu\lambda^3}. \tag{28}$$

In other words, the maximum number of bits which V can hold in the form of electromagnetic energy is c_{th}. By using (26), (27), and (28), one finds the (extrapolated) efficiency of the beamfolding method to be

$$\frac{c}{c_{\text{th}}} = \frac{1}{(3.3k)^4}. \tag{29}$$

For $k = 3$, this efficiency is only 10^{-4}, but even then a capacity of 16 k bit seems achievable with 1 GHz bandwidth in a cavity with the radius $A = 1$ cm.

V. CONCLUSIONS

Various methods can be used to transmit a multitude of beams through a lens guide in such a way that all beams are clearly resolvable at the receiving end. The number of beams which can be transmitted is proportional to the square of the guide cross section and may be of the order of 300 for a guide of 10 cm radius with lenses

* See, for example, Ref. 4.

spaced confocally by 100 m. In this case, the centers of adjacent beams would be spaced by 6 beamwidths at particular cross sections in the guide.

Linking the available channels end to end in a cavity produces a delay line or storage device. At 1 micron wavelength a 10 μsec delay can be achieved in an optimized cavity, 15 cm long and 4 cm in diameter. The storage capacity is inversely proportional to the cavity length. Hence, the ultimate configuration would consist of two confocal mirrors with their circumferences touching. Extrapolating the paraxial theory to this situation yields a capacity of 16kbit for a bandwidth of 1 GHz if the nut-shaped cavity has a radius of only 1 cm.

ACKNOWLEDGMENTS

We are very thankful for stimulating discussions and fruitful suggestions by R. Kompfner and E. A. J. Marcatili.

REFERENCES

1. Herriott, D. H. and Schulte, H. T., "Folded Optical Delay Lines," Appl. Opt. *4,* No. 8 (August 1965), pp. 883–889.
2. Herriott, D. H., et al., "Off-Axis Paths in Spherical Mirror Interferometers," Appl. Opt. *3,* No. 4 (April 1964), pp. 523–526.
3. Basov, N. G., et al., "Some Properties in the Transmission and Reception of Information Using Laser Oscillators and Amplifiers," Radio Eng. and Elec. Phys. (trans. Radioteknika i Elektronika), *9,* No. 9 (September 1964), pp. 1387–1391.
4. von Laue, M., "Die Freiheitsgrade von Strahlenbündeln," Ann. Physik *44,* No. 16 (August 1914), pp. 1197–1212.
5. Kogelnik, H., "Imaging of Optical Modes—Resonators with Internal Lenses," B.S.T.J., *44,* No. 3 (March 1965), pp. 455–494.
6. Tien, P. K., et al., "Focusing of a Light Beam of Hermite-Gaussian Distribution in Continuous and Periodic Lens-Like Media," Proc. IEEE, *53,* No. 2 (February 1965), pp. 129–136.
7. Steier, W. H., "Ray Packet Equivalent of a Gaussian Light Beam," Appl. Opt., *5,* No. 7 (July 1966), pp. 1229–1233.
8. Gronros, W., Herriott, D. R., Murray, R. G., and Yocom, W. H., unpublished work.
9. Yocom, W. H., Jarzyna, E. S., and Herriott, D R., "Experiments with Active Optical Delay Lines," presented at Conf. Elec. Device Res., U. of Colorado, Boulder, Colo., June 20, 1968.
10. Rivlin, L. A., "Propagation of Light in a Stable Medium with Negative Absorption," Radio Eng. and Elec. Phys. (trans. Radiotechnika i Elektronika), *12,* No. 2 (February 1967), pp. 253–258.
11. Patel, C. K. N. and Slusher, R. E., "Self-Induced Transparency in Gases," Phys. Rev. Letters, *19,* No. 18 (October 1967), pp. 1019–1022.

A Model of a Domestic Satellite Communication System

By LEROY C. TILLOTSON

(Manuscript received July 12, 1968)

A preliminary study of a domestic satellite system is reported. Since the objective was to determine what might ultimately be possible, no attempt is made to relate system capacity to estimated needs; rather an effort has been made to conceive a system to carry the greatest possible amount of traffic. By making full use of modern rocket technology including the Saturn V class propulsion systems, highly directive multibeam antennas operating in the range from 15 to 40 GHz, interference resistant modulation methods, highly stabilized synchronous repeater platforms, and integrated solid state microwave repeater electronics, a very large communication capacity is obtained. For example, using 50 ground stations and 50 satellites operating in bands at 20 and 30 GHz, each 4 GHz wide, a total of 100 million voice circuits, or equivalent, can be provided.

I. INTRODUCTION

Domestic satellite systems can be expected to handle a large amount of traffic compared with that carried by transoceanic systems; this is true even if for various reasons only a fraction of the total domestic traffic goes via satellite. For this reason among others, the presently allocated frequency bands at 4 and 6 GHz are not well suited to domestic use. Frequencies above 10 GHz are attractive in that they are not as heavily loaded as the lower frequency bands. But they are subject to propagation difficulties which generate new problems in their use.* The amount of atmosphere traversed by a ground-to-satellite path can be relatively small if the look angle is restricted to elevations which are not too small. However, attenuation will be large under conditions of excessive rainfall and diverse ground terminals will

* Similar arguments apply to terrestrial systems operating above 10 GHz. These are not discussed here, but some of the propagation studies described here were designed with the needs of terrestrial as well as satellite systems in mind.

be needed when common carrier grade continuity of service is required.

The frequency range from 10 to 40 GHz can offer a unique opportunity if broad continuous bands are allocated for satellite service. Modern rocket technology gives us the opportunity of placing large multichannel satellite repeaters in synchronous (24 hour) equatorial or inclined orbits, thus making it possible to exploit broad communication bands and to use orbit space very efficiently. This would make possible best use of orbit space, frequency space, and the investment in facilities since very large amounts of traffic could be carried by each satellite repeater. The availability of such frequency allocations and adequate orbit space is assumed in what follows. *A basic assumption made here is that frequency space and orbit space are precious and limited resources which must be conserved.*

Radio frequencies above 10 GHz have disadvantages in propagation, but their short wavelength makes possible very narrow beams from antennas of a size suitable for use on a contemporary satellite. If we combine this feature with interference resistant modulation techniques, such as PCM and multiple feed antennas, we can construct multichannel, multibeam satellites which can communicate simultaneously with many ground stations using only one frequency assignment. The idea can also be used in reverse to enable a single ground station to communicate simultaneously with several satellites. Thus if we have N channels per frequency assignment, S satellites, G ground stations, and every satellite "sees" every ground station, and vice versa, we have a total communication capacity of $C = G \times S \times N$ channels. With tens of ground stations and tens of satellites working in bands a few GHz wide, this can results in a truly prodigious capacity.

In order to take full advantage of these possibilities, very reliable, efficient, and small radio repeaters for satellites will be required. Solid-state integrated microwave circuits and devices will help insure reliability and small size, and we have reason to hope that eventually efficient use of dc power can also be obtained. This is important since a significant part of the total in-orbit satellite cost results from the solar power supply.

II. SYSTEM CONCEPT

Designers of communications systems normally have as their objective a facility to meet a fairly well defined need and naturally try to choose a system configuration which is most economical in serving that need. Thus, one of the most important parameters influencing

system design is the present traffic level and the expected growth rate. In addition to the amount, the geographical distribution of traffic is important. If a large part of the traffic terminates at one node, the usable system capacity will be reached when all of the channels from the satellites to that node are full. If the traffic volume continues to increase but the geographical distribution remains unchanged, growth must be handled by other (terrestrial) means.

While much of the present-day telephone traffic is concentrated in large metropolitan centers, an assumption that this trend will continue into the indefinite future may not be warranted. There are many signs that some of these areas have reached a "critical mass" and are beginning to explode; future growth may be spread much more uniformly over the United States. At any rate, we have assumed uniform traffic density when calculating total system capacity; to the extent that this is untrue, part of the system potential will be unrealizable. This approach has the disadvantage of making the study somewhat more abstract, but our goal is to display the potential of a satellite system designed for domestic communication rather than to design a specific system.

To this end we postulate 50 ground stations distributed more or less uniformly over the continental United States working with 50 satellites in synchronous orbit stationed due south of the United States and spaced 1° apart. Each satellite is precisely stabilized in attitude, carries a 10 meter multibeam antenna operable at 20 and 30 GHz, transmits down with a power output of 2 watts at 20 GHz and receives from the ground at 30 GHz. We further assume that two bands each 4 GHz wide and centered at 20 and 30 GHz are assigned to this service. Given these frequency bands, we can design for eight 630 megabit per second two-way channels using four-phase angle modulation. To serve 50 ground stations each satellite will require (eight RF channels per beam) × (50 beams) = 400 repeaters. At 10,000 voice circuits per RF channel (630 megabits per second capacity) this results in 4 million (one-way) voice circuits per satellite. In terms of present day telephone traffic, this is a very large cross-section, but for broadband services which require 100 (*Picturephone*® visual telephone) to 1,000 (television) times more bandwidth, it is not so large.

If the ground stations are equipped with 10 × 17 meter multibeam antennas, 10 watt 30 GHz transmitters and a cooled parametric receiver preamplifier operating at 20 GHz with an equivalent noise temperature of 150°K, one can calculate carrier-to-noise ratios with the result shown in Table I. Several factors are worthy of note. The

ground-to-satellite link is assumed to suffer interference −39 dB relative to the desired carrier; the down link interference is assumed −33 dB relative to the carrier. In both cases the most important interfering signals are admitted by the sidelobe response of the multibeam satellite antenna. Interference at the satellite caused by minor lobes of the ground station antennas can be controlled by spacing of the satellites, and in fact, if the satellites are 1° apart as assumed and the 10 × 17 meter 30 GHz ground station antennas have a response

TABLE I — C/N ON HEAVY ROUTES

System Parameters

d = 23,000 miles = 3.7 × 10^7 m.
f = 30 GHz (up link)
f = 20 GHz (down link)

Width of RF assignments (2)		4000 MHz
Two-way RF channels per beam		8
Pulse rate		315 × 10^6/s
Bit rate (4 φ angle mod.)		630 × 10^6/s
RF channels	per satellite	8(G)
	per ground station	8(S)
	in system	8(G)(S)
Picturephone® circuits per RF channel		100.
Voice circuits per RF channel		10,000.

Rocket class: Saturn V

Satellite

Antenna	
Diameter	10 m
Diameter illuminated by each feed	8.3 m
Allowance for illumination taper (−2 dB)	0.63
Effective area	34.0 m²
Transponder power output	+3 dBw
Receiver noise figure	6 dB
Receiver noise bandwidth	400 MHz

Ground Station

Antenna	
Dimensions	10 × 17 m
Sector illuminated by each feed	10 × 10 m
Allowance for illumination taper	0.80
Other problems (−1.5 dB)	0.70
Effective area	56.0 m²
Transmitter power output	+10 dBw
Receiver noise temperature, T_e	150°K
Receiver noise bandwidth	350 MHz

Carrier-to-Noise Ratio	Up link at 30 GHz	
Net path loss $P_T/P_R = \lambda^2 d^2/A_T A_R$	78.6	dB
Transmitter / Receiver system loss, including filters	1	dB
Receiver power, $P_R = P_T - 80.6$ dB Satellite receiver noise	−70.6 −112	dBw
C/N at satellite C/I for ground stations 264 miles apart C/N + I at satellite	41.4 39 37.0	dB
	Down link at 20 GHz	
Net path loss $P_T/P_R = \lambda^2 d^2/A_T A_R$	82.1	dB
Transmitter / Receiver system loss, including filters	0.5	dB
Received power, $P_R = P_T - 83.1$ dB Ground station receiver noise (no rain)	−80.1 −121.5	dBw
C/N in ground receiver (no rain) C/I for stations 275 miles apart C/N + I (including up link) C/N + I for 10^{-7} error rate	41.4 33 31.1 20.0	dB
Margin (no rain attenuation) C/N + I for 10^{-5} error rate C/N + I (for 10 dB rain attenuation) Margin during moderate rain	+11.1 17.0 25.7 +8.7	

envelope which follows a fourth power law, interference at the nearest neighbor will be so small as to be negligible.

A similar comment applies to interference at a ground receiver admitted by the sidelobe response of the 20 GHz 10 × 17 meter ground station antenna. Under conditions of normal propagation (no rain) the system will be interference limited on both the up and down links. This is unusual in present day radio system design and comes about because we are attempting to get the maximum possible communication capacity. Any further improvement will require better antennas. Notice also that the down link margin for a 10^{-5} error rate is +8.7 dB during a rainstorm which produces 10 dB of attenuation.* For a similar, but independent fade on the up link the margin is 14 dB.

Another important feature of the model satellite system is the fact that a given repeater can be seen from many areas in the United States. This makes possible a flexibility not found in terrestrial links

* Calculation of error rates corresponding to the stated ratios of $C/(N + I)$ indicates much smaller rates than those given in Table I. The stated larger values allow for imperfection in a high bit rate PCM receiver.

where standby or alternate facilities must be provided link by link. In the satellite case we have, at least in principle, the possibility of making good a fault in one part of the network with capacity normally assigned to a distant link and temporarily unused because of time zone or other differences.

On the other hand, if we adopt the now generally accepted rule that only one 24-hour synchronous satellite link is allowed per circuit, this ubiquitousness of satellite system ground terminals is not just a convenience, it is a necessity. If a satellite system is to carry the bulk of the domestic traffic, we require not just circuits from outlying stations into a network node, but circuits which go directly from every place to every other place; thus we require many ground stations for complete coverage. The fact that channels from every ground station pass through each repeater makes available in one place channels serving widely separated parts of the country and provides us with the opportunity to reroute channels on board a satellite in accordance with changing traffic patterns.

To accomplish this, we must provide a suitable switching matrix; this is most conveniently implemented at the intermediate frequency which can be common to all repeaters, and hence channels can be interchanged as desired upon command from the ground. This flexibility also makes it possible to make good a limited amount of failed apparatus on board the satellite.

III. PROPAGATION STUDIES

Since electromagnetic waves with frequencies above 10 GHz are severely attenuated by liquid water, satellite systems using these frequencies must be designed to tolerate a few dB of attenuation which may be experienced for long periods and to switch to a diversity ground station on those rare occasions when a very large attenuation is caused by excessive rain. Also the diversity ground station must be far enough removed so that there is little likelihood that an intense rainstorm will cover both ground stations at the same time.

If the attenuation statistics on the two paths are uncorrelated and enough margin is provided to make both outages small, the system outage time can be reduced to a satisfactory degree. In general, this requires a more detailed and a more quantitative knowledge of rainfall and attenuation caused by rain than presently exists. Accordingly, measurements have been undertaken at Bell Telephone Laboratories to determine for terrestrial paths and for ground-to-satellite paths,

the constraints on system design imposed by this natural phenomenon.[1] These programs are discussed elsewhere; only a summary is attempted here.

Measurements of rain made at the earth's surface are not sufficient to determine attenuation along a satellite-to-ground path, but they can indicate the amount of structure characteristic of a typical rainstorm. A sample contour map of instantaneous rain rate as measured by the Crawford Hill rain gauge network is shown on Figure 1.[2] These data form a snapshot of a specific area at a given moment. It

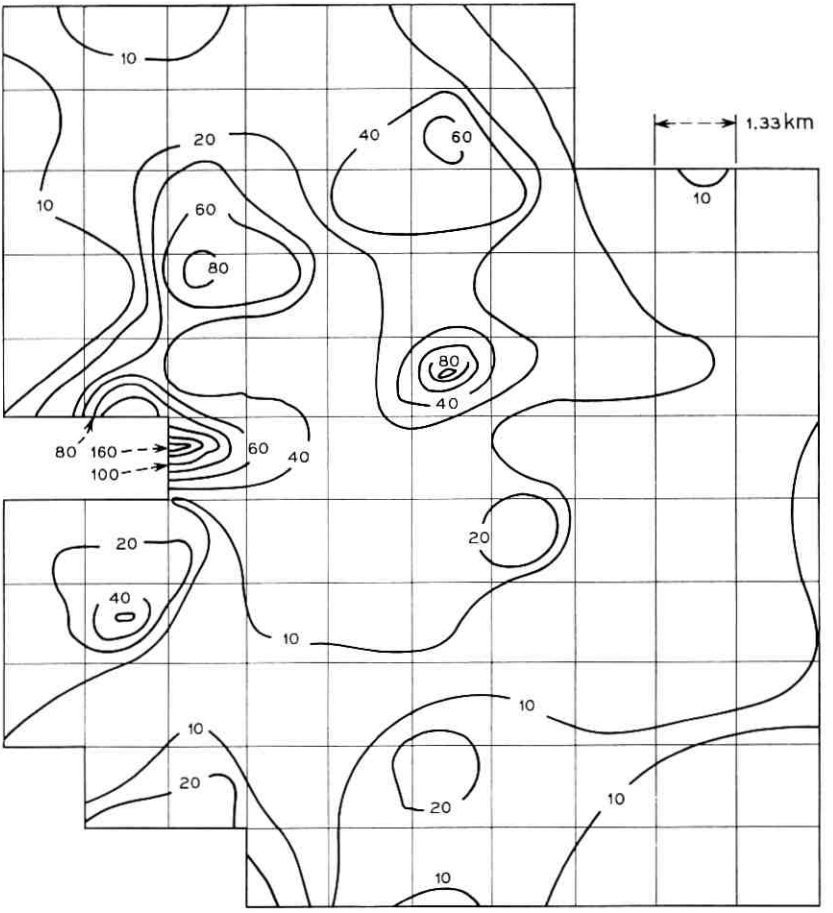


Fig. 1 — Sample scan of Crawford Hill rain gauge network for August 25, 1967 at 20:43 EST, contours in mm per hour.

was selected from many hours of data to illustrate one of the most intense rainstorms observed during a one-year period of measurement. The data illustrate that the most intense rain occurs in very limited cells and that rain which covers large areas (square miles) falls at the rate of one inch per hour or less. These are the considerations behind the proposal to use diversity ground stations separated by several miles.

A more quantitative determination of attenuation statistics is being made with a radiometer which tracks the sun. This installation, which has been described by Wilson,[3] measures the noise received from the sun at 16 and 30 GHz and thus determines the attenuation caused by any intervening precipitation. Figure 2 is a photograph of the Crawford Hill sun tracker. At night the radiometer is pointed to a cold part of the sky and detects precipitation as an increase in effective temperature. Good accuracy is obtained to about 35 dB of attenuation, when tracking the sun, and to about 10 dB when the cold background is used. Thus we are accumulating data which will be helpful in the design of domestic satellite systems, especially those operating above 10 GHz.

Unfortunately, if one is to obtain a statistically meaningful result, operation must include at least a year in order to encompass all of the

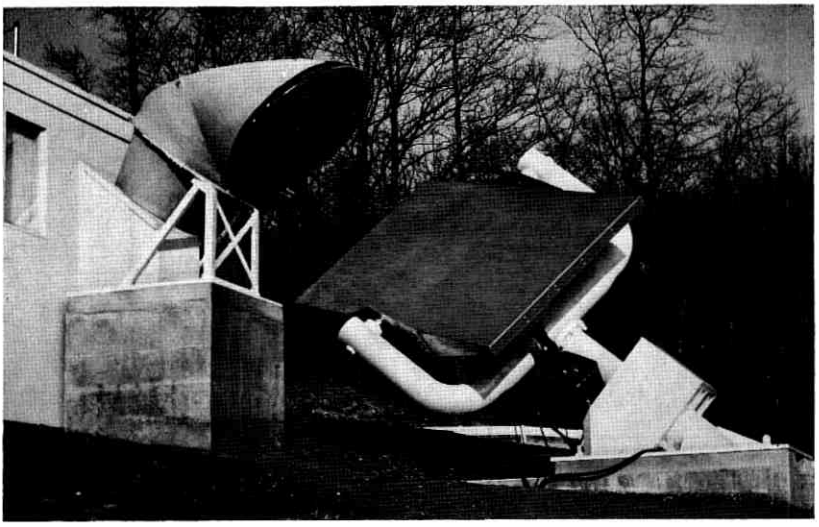


Fig. 2 — Crawford Hill sun tracker.

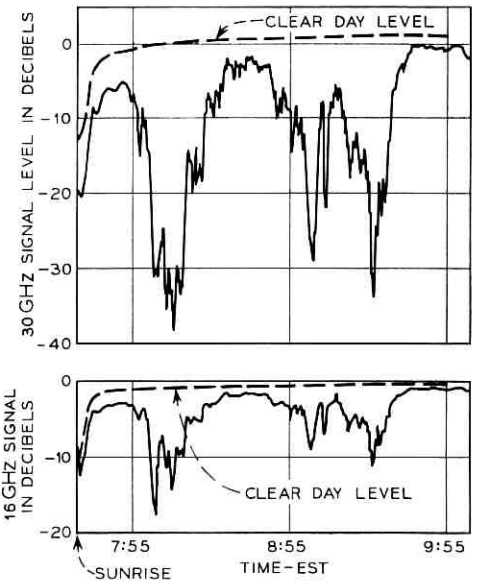


Fig. 3 — Recording by Crawford Hill sun tracker of shower on December 12, 1967.

seasons. And this cannot be hurried. When we have obtained enough data to be worthy of statistical analysis, it will be published; data obtained over a shorter interval must necessarily be regarded as incomplete. With this qualification, we present the following data and observations which, though limited, are relevant:

(*i*) Attenuation exceeding 30 dB at 30 GHz (about 10 dB at 16 GHz) and lasting 6 to 8 minutes has ben observed on several occasions. Figure 3 is a strip chart recording showing such an occurrence along with a plot of the simultaneous attenuations at the two frequencies. Table II summarizes the data obtained for about 10 months through August 9, 1968.

(*ii*) These events are associated with intense rain cells (precipitation exceeding 100 millimeters per hour).

(*iii*) Heavy overcast and even light drizzle produce little attenuation (less than 2 dB at 30 GHz).

(*iv*) Attenuation by fog and snow is no problem at these frequencies.

(*v*) Since a second sun tracker is not yet operating, we have no direct data on diversity advantage. However, data from our rain gauge

TABLE II — DATA FROM CRAWFORD HILL SUN TRACKER

Attenuation (dB) at 30 GHz*	Percent of total observing time 3091 daylight hours
>3	1.93
>6	1.02
>9	0.59
>15	0.32
>21	0.18
>27	0.11
>33	0.08
at 16 GHz†	**2520 daylight hours**
>1	1.29
>2	0.75
>3	0.49
>5	0.28
>7	0.19
>9	0.14
>11	0.12
>13	0.08

* Data from October 14, 1967 through August 9, 1968.
† Data from December 8, 1967 through August 9, 1968.

network indicate that heavy rains are highly structured and are uncorrelated at distances greater than a few miles.

In addition, a variety of locations representative of all of those areas in which ground stations will be located should be investigated. The present sun tracker is located in New Jersey, which is only one of many possible locations, but which does have the advantage of being climatically intermediate between the extremes represented by parts of the gulf coast area and the western desert regions.

Since the sun moves continuously only a small time is spent looking in a given direction; on the other hand, at least some data on many possible satellite paths are being obtained. At best, these data will leave many questions unanswered, but we will have some quantitative data on which to base our system designs. More locations, and particularly the advantage obtained by using two or more diversity ground stations, must be studied.

A beacon in synchronous orbit would be an ideal source for propagation experiments, but if it is to be worth the cost, a satellite experiment must have some advantage over a sun tracker. A stationary satellite with a highly coherent source of adequate power output would make possible simpler and therefore cheaper ground stations. Because it would be stationary, the satellite would not require precise

tracking as does the sun. Tracking the sun is simple but does complicate the installation and costs money. Use of a satellite would make possible more observation points for the same total ground station cost and thus enable us to gather design data more quickly. A highly coherent source, in contrast to the thermal noise emitted by the sun, would ease the measurement of the phase coherence of received signals over very large bandwidths. This is important because the system model requires very broad bandwidths (300 to 400 MHz) and correspondingly short pulses.

Direct evidence that satellite radio channels will support such transmission is scarce and more is needed. Considerable indirect evidence is available, however. A relatively short (12.5 km) terrestrial path has been studied at 12 GHz for over a year with only one instance of moderate multipath activity; a second path 6.5 km long was also observed at 18 GHz for more than a year with no evidence of multipath fading.[4] Very large narrow beam antennas have been used for radio astronomy at 30 GHz with no significant beam broadening observed.[5] Measurement with an array at 30 GHz has shown that the phase front of the incoming wave is not badly distorted.[6]

Scattering from refractive index inhomogeneities in the atmosphere could conceivably introduce time delays which would restrict bandwidth, but this same phenomena would broaden antenna beams and reduce antenna gain. Probably the most sensitive indicators are the wide base interferometers used for radio astronomy; some have obtained resolutions of a fraction of a second of arc. All of the above applies to relatively high elevation angles, say, above 15°. It seems certain that all of the difficulties associated with long terrestrial paths will be experienced as the satellite path approaches grazing incidence. The problems which remains is to make this relationship quantitative; this would be most easily done using a highly coherent source on board a satellite. However, if we are to obtain data useful for system design, the satellite must be continuously available to insure that we do not miss any data, and it must be available over a long enough period of time to insure that statistics derived from the data are reasonably stable, that is, seasonal and other variations are included.

IV. INTERFERENCE CONSIDERATIONS

A radio system designed for maximum efficiency will be interference limited. This is true because doing so makes best use of frequency space, geometry (which in the case of a synchronous satellite system

includes orbit space), and the investment in facilities. An interference limited system is the end result of trying to get the maximum possible communication with a given limited resource. We assume that the required technology has reached the state where thermal noise can be reduced to a small value compared with interference at least during normal propagation conditions. Calculations of carrier-to-noise ratios made in this paper (see Table I) verify that such an assumption is warranted.

4.1 *Antenna Patterns*

The statement that a radio system is interference limited is equivalent to saying that the radiation patterns of the antennas used are of crucial importance; hence, we start with these. As in other areas, we look at current practice, estimate what may be possible in the future, and try to land somewhere between.

Patterns for an antenna of current design are compared with greatly simplified relationships in Fig. 4. We see that the microwave

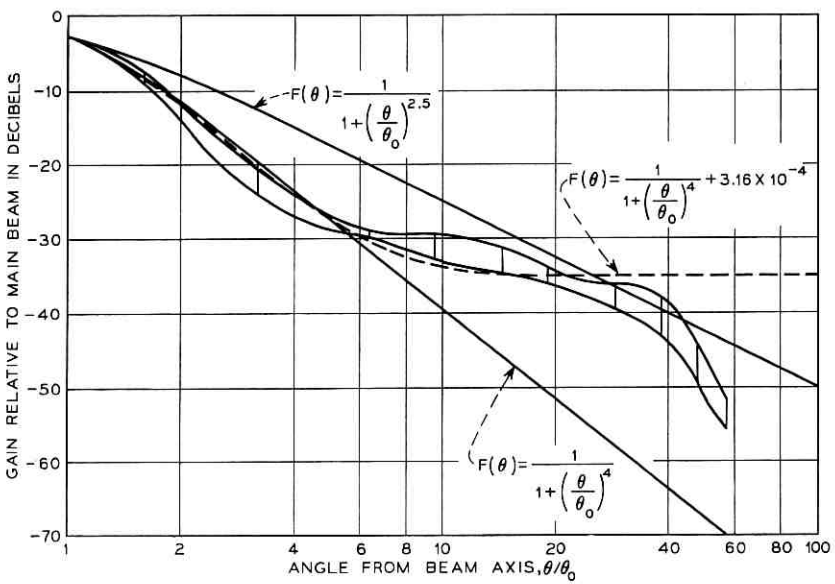


Fig. 4 — Side lobe levels of microwave pole line periscope antenna. Beamwidth = 2.5°, θ_0 = 1.25°. Vertical lines show range of envelope of side lobe levels for both polarizations.

pole line antenna* is almost everywhere better than $F(\theta) = [1 + (\theta/\theta_0)^{2.5}]^{-1}$, where

θ = angle off the peak of main lobe
θ_0 = $\frac{1}{2}$ of the 3 dB beamwidth
$F(\theta)$ = antenna response (power)

which we understand has sometimes been used in studies by the International Consultative Committee on Radio. In fact, the microwave pole line antenna follows the $(\theta/\theta_0)^4$ curve down to about -25 dB. On Fig. 5 we have plotted theoretical response curves for antenna apertures with specified variations in the amplitude of the illumination together with a uniform phase. We see that the envelope of a $\cos^N [(\pi/2)x]$, where $N = 1$, distribution equals or exceeds the discrimination predicted by the $(\theta/\theta_0)^4$ curve and that for all higher values of N the $(\theta/\theta_0)^4$ curve is greatly exceeded. But, of course, these curves are not for real antennas. However, in Fig. 6 we see the envelope of the pattern of a horn-reflector antenna using a special dual mode feed designed by R. H. Turrin.[7] In this case the pattern discrimination provided by a real antenna considerably exceeds the $(\theta/\theta_0)^4$ curve for all levels of discrimination and both polarizations.†

For our present calculations we use the relationship, $F(\theta) = [1 + (\theta/\theta_0)^4]^{-1}$, with the expectation that a multibeam antenna with a reasonable amount of illumination taper designed to reduce or eliminate aperture blockage by the feeds will be able to do at least this well. We think that the data presented above makes this a reasonable assumption.

4.2 *Interference at Satellite*

Our system model assumes that each ground station communicates with each satellite and vice versa. If we consider the situation at the satellite antenna feed corresponding to a desired ground station, we see that it will receive power from the desired station through its main lobe and from all other ground stations through its side lobes. However, with a fourth power variation in antenna response, the nearest neighbors will contribute most of the interference. Thus if we visualize

* This is a well-shielded antenna designed for a specific system. It was suggested by A. B. Crawford and measured by R. H. Turrin. (Unpublished work).

† The problem which remains in this case is that no one knows how to build a multibeam horn-reflector antenna.

a rectangular grid of uniformly spaced ground stations, there will be stations stretching out to the north, south, east, and west with spacings 1, 2, 3, . . . times unit distance. The distance from the desired ground station to another ground station having coordinates (n, k) will be $(n^2 + k^2)^{1/2}$; hence, the total interference is

$$F(\theta) = \sum_{n,k}^{\infty} \left[1 + \left(\frac{\theta_{n,k}}{\theta_0}\right)^4\right]^{-1}.$$

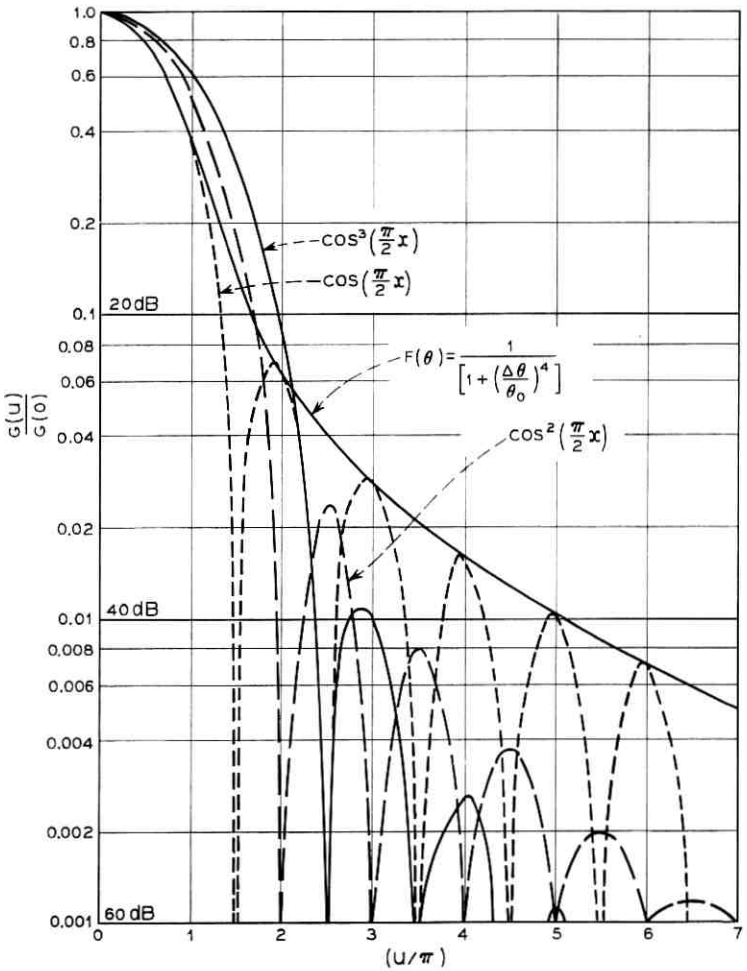


Fig. 5 — Calculated antenna pattern for $\cos^n[(\pi/2)x]$ illumination of square aperture. a is aperture opening. θ is angle with normal to aperture. $u \equiv (\pi/\lambda)\, a \sin \theta$.

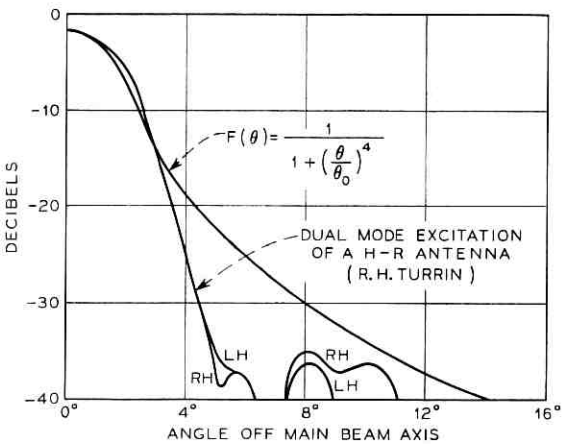


Fig. 6 — Comparison of measured antenna discrimination with assumed value.

When, as in the present case

$$(\theta_{1,1}/\theta_0)^4 \gg 1, \quad \text{and} \quad \theta_{n,k} \doteq \frac{S}{R}(n^2 + k^2)^{\frac{1}{2}}$$

where

S = uniform spacing between ground stations
R = distance to the satellite
$n, k,$ = 1, 2, 3, $\cdots$.

Then

$$F(\theta) = \sum_{n,k}^{\infty} \left[\left(\frac{\theta_{n,k}}{\theta_0}\right)^4\right]^{-1} = \sum_{nk,}^{\infty} \left[\left(\frac{S}{R}\right)^4 \frac{(n^2 + k^2)^2}{\theta_0^4}\right]^{-1}$$

$$F(\theta) = \sum_{n,k}^{\infty} \frac{\theta_0^4}{\left(\frac{S}{R}\right)^4} \frac{1}{(n^2 + k^2)^2} = 5.94\left[\frac{\theta_0}{\left(\frac{S}{R}\right)}\right]^4.$$

Hence, the total interference caused by all of these interfering ground stations will be 5.94 times or about 7.7 dB up on the interference produced by a single nearest neighbor.

As applied to the present case and as assumed for the calculations made in support of Table I where a C/I ratio of 39 dB (net) at the satellite was used, a distance of 264 miles with a 10 meter satellite antenna was obtained. These computations are (see Fig. 7):

$$F(\theta) = \left[\frac{1}{1 + \left(\frac{\Delta\theta}{\theta_0}\right)^4}\right], \text{ or } F(\theta)^{-1} \approx \left(\frac{\Delta\theta}{\theta_0}\right)^4.$$

For C/I = 39 dB (all sources), 39 dB + 7.7 dB = 46.7 dB (one source) thus

$$\left(\frac{\Delta\theta}{\theta_0}\right)^4 = 10^{4.7}, \qquad \frac{\Delta\theta}{\theta_0} = 10^{1.176} = 15.0.$$

For up link $f = 30$ GHz, $\lambda = 1$ cm, and assuming a 10 meter satellite antenna:

$$2\theta_0 = 1.22\frac{\lambda}{a} = 1.22\left(\frac{1 \text{ cm}}{830 \text{ cm}}\right)$$

$$2\theta_0 = 1.47 \times 10^{-3} \text{ rad.}$$

$$\theta_0 = 0.735 \times 10^{-3} \text{ rad.}$$

$$\Delta\theta = 15.0\theta_0 = 11.0 \times 10^{-3}$$

$$S = R(\Delta\theta) = 24 \times 10^3(11.0)10^{-3}$$

$$S = 264 \text{ miles.}$$

4.3 *Interference at a Ground Station*

Again we visualize a grid of uniformly spaced ground stations, each of which communicates with every satellite.* Thus a given ground station will be surrounded by beams from the satellite to all of the other ground stations, and the minor lobes associated with these satellite antenna feeds will crosstalk directly into the main beam of the ground station since it is "looking" at the desired satellite. As before, with a fourth power variation in antenna discrimination versus angle, most of the interference will come from the nearest neighbors. The analysis for this case follows the previous one for the interference at a satellite with nearly the same result except that at the longer wavelength (20 GHz) there is slightly less antenna discrimination. As applied to the present example—a 10 meter antenna on a Saturn launched satellite—we obtain the required spacing for the ground stations of 275 miles, as follows (see Fig. 8):

$$F(\theta) = \left[1 + \left(\frac{\Delta\theta}{\theta_0}\right)^4\right]^{-1}, \text{ or } F(\theta)^{-1} \approx \left(\frac{\Delta\theta}{\theta_0}\right)^4.$$

* Obviously this is an idealized model. Even if the ground stations were located in a regular pattern, which is not at all likely, the pattern would appear distorted to the satellite. An accurate calculation must be based on actual locations of ground stations which are at present unknown and north-south spacings will depend on station latitude; the idealized model is adequate for our purpose. No advantage has been taken of polarization. This is deliberate since a quantitative evaluation is not available at this time. We leave any polarization advantage for use in solving unforeseen problems.

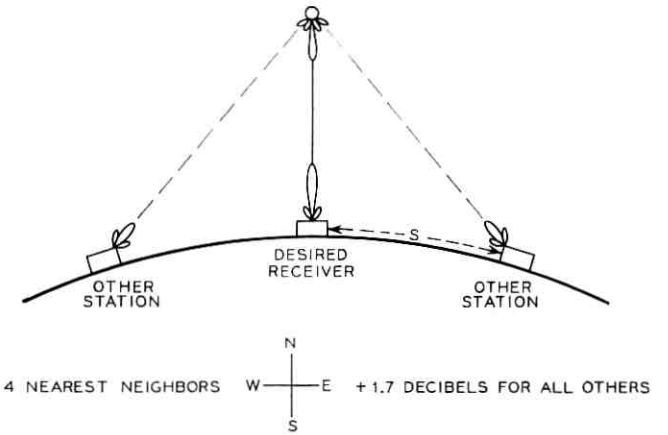


Fig. 7 — Interference at a satellite.

For C/I = 33 dB net, 33 dB + 7.7 dB = 40.7 dB for one source, thus

$$\left(\frac{\Delta\theta}{\theta_0}\right)^4 = 10^{4.07}, \qquad \left(\frac{\Delta\theta}{\theta_0}\right) = 10^{1.02} = 10.4.$$

For the down link $f = 20$ GHz, $\lambda = 1.5$ cm, and assuming a 10 meter satellite antenna:

$$2\theta_0 = 1.22\,\frac{\lambda}{a} = 1.22\left(\frac{1.5\text{ cm}}{830\text{ cm}}\right)$$

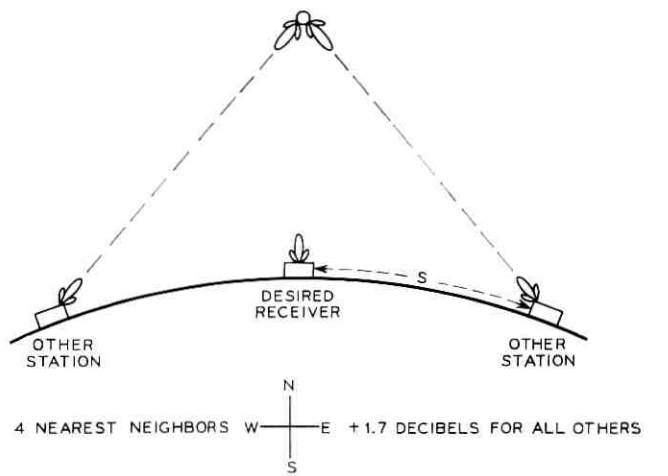


Fig. 8 — Interference at a ground station.

$$\theta_0 = 1.10 \times 10^{-3} \text{ rad.}$$

$$\Delta\theta = 10.4\theta_0 = 11.45 \times 10^{-3} \text{ rad.}$$

$$S = R(\Delta\theta) = 24 \times 10^3(11.45 \times 10^{-3})$$

$$S = 275 \text{ miles.}$$

4.4 *Modulation Techniques*

The use of different look angles to achieve discrimination by antenna pattern is one important means for achieving co-channel operation in a limited geographical area. Another is choice of a modulation technique resistant to interference, particularly co-channel interference. The possibilities here are many, at least in theory. However, when, as in the present case, one is considering channels having bit rates of several hundred megabits per second, complicated encoding schemes—especially those which require appreciable storage—are not worth the cost and complication. In the present case, as is indicated in Table I, we have chosen four-level over binary PCM even though the latter is more resistant to interference.*

While the increased resistance to interference provided by two level modulation is attractive, our studies indicate that the possibility to double the channel capacity in approximately the same bandwidth by using four levels is well worth its cost. It is conceivable that even more levels would be advantageous, but the return diminishes at such a rate that specific situations must be studied to be certain. These studies are being continued.

One of the features of quantized modulation is that the signal can be regenerated at intermediate repeaters in which case any errors made in regeneration accumulate along the system, but thermal noise, interference from crosstalk, and imperfections in repeater characteristics do not. In the present satellite system configuration, the satellite-to-ground or "down-link" is the most difficult from both the standpoint of thermal noise and interference. This comes about because we can use a powerful enough transmitter at the ground station to make thermal noise at the satellite receiver front-end negligible, and we can space the satellites far enough apart (1° in the present example) so that, with highly directive ground station antennas,

* Buried in these statements are assumptions about the manner in which co-channel, adjacent frequency channel, and adjacent time slot interference impair a signal. These problems are being intensively studied and will be reported elsewhere.

interference at adjacent satellites is small. Under these circumstances it is not worthwhile to regenerate on board the satellite.

However, if in the future synchronous orbit space becomes so crowded that we are forced to tolerate the same interference on the up-path as on the down-path, regeneration will be very much worthwhile. If we regenerate, the error rate would increase from, say, 1 in 10^7 to 2 in 10^7 for both links, but if we do not regenerate, the error rate would increase from 1 in 10^7 to 1 in 10^4. However, under actual operating conditions the vagaries of propagation will affect the received signal level and under some conditions can adversely affect the signal-to-interference ratio. Thus an actual system must be designed with some margin for changes in signal level, and even a few dB of margin will make regeneration on board the satellite much less profitable. Regeneration at each intermediate repeater is most useful with very stable media such as coaxial cable or waveguide, where changes in transmission loss are small and systems can be designed to operate continuously at the specified error rate.

V. THE SATELLITE

Satellites of the size and complexity envisaged here have never been built, but there is no particular reason why they cannot be. Once launched, the environment in which the repeater must operate is in many ways more benign than here on earth. In the early days, our ignorance of the space environment clouded our vision and obscured this fact. But experience over the past several years with communication and other satellites has encouraged consideration of larger and more sophisticated spacecraft. Reliability remains a problem, but this is not peculiar to spacecraft since present day terrestrial communication equipment must also be designed to be maintenance free for up to ten years because of cost.

Design of a radio repeater for satellite use has much in common with design of its terrestrial counterpart. This is particularly true where, as in this case, such large high-gain antennas are used on the spacecraft and at the ground station that the loss on a satellite path is not much larger than is encountered on typical terrestrial paths. Using this approach, a satellite repeater can be of rather conventional design, but with special attention to efficiency in the use of dc power and reliability. Integrated microwave circuits and solid-state devices would be used throughout except for the traveling wave tube final RF amplifier.

A repeater of this type is estimated to weigh 3¼ pounds and to require 18 watts of unregulated power from the solar supply for an RF output of 2 watts per channel. To this must be added 2 pounds per ground station (8 RF channels) to allow for diplexers and channel dropping and recombining networks. In addition the *pro rata* share of power supply weight will be at the rate of 2 watts per pound as discussed in Section 5.2 for a total of 9 pounds for 18 watts to power one transponder. Thus 8 transponders to serve one ground station will weigh 8 (3¼ pounds) for electronics + 8 (9 pounds) for power + 2 pounds for filters = 100 pounds.

5.1 *Stabilization*

When very narrow antenna beams are used, as in this design, precise stabilization of the spacecraft is crucial. The first task which must be accomplished is sensing the attitude of the spacecraft with a high degree of accuracy and reliability. In this discussion we assume that crude orientation, that is, within a degree or so about all three axes, is accomplished by present methods. More precise sensing for final attitude adjustment to $\pm 0.01°$ (0.2×10^{-3} radians) can be achieved by radiometric means. For example, at the upper frequency of 30 GHz, antennas spaced a distance $a = 10$ meters will receive signals differing in phase by π radians when the platform tilt $\varphi = \lambda/2a = 10^{-2}$ meters/20 meters $= 5 \times 10^{-4}$ radians. A comparison circuit should be able to detect a difference of $\pi/100$ radians or a tilt angle of 5×10^{-6} radians. This is about two orders of magnitude smaller than the allowable variation and should be adequate for precise attitude control.*

If we use one-meter antennas on the spacecraft for this purpose, reduce the receiver bandwidth to 400 kHz, and keep the ground transmitter power unchanged at 10 watts, the carrier-to-noise ratio in this sensing channel will be 50 dB. This is more than enough, but the consequences of losing attitude control are severe and plenty of margin is in order. Another basic aspect of attitude control is the amount of fuel (weight) required to stabilize a spacecraft for a long period, say, ten years.

This phase of the problem has been studied from two complementary points of view: (*i*) a study of the basic effects which perturb the satellite orbit[9] and (*ii*) a rudimentary station keeping and attitude control subsystem design. The first study indicates that an impulse

* An ingenious means for keeping most of the complication of the attitude sensing and control system on the ground has been given by C. C. Cutler in Ref. 8.

of 157 feet per second per year would correct for changes in the inclination of the orbit of 0.86° per year caused by lunar and solar gravitation plus the lesser effect of a westward acceleration of 1.68×10^{-3} degrees per day resulting from the nonsphericity of the earth's gravitational field. Attitude control, even to the very precise limits assumed here, appears to be primarily a question of the reliability of components to be used in a system with a design lifetime of many years.

In comparison with the amount of fuel required for station keeping, the weight of fuel required for attitude control is negligible, that is, a few pounds per year per ton of satellite weight. Thus in the design study it was assumed that an impulse of 160 feet per second per year would be required for station keeping. Our conclusion from these studies is that the weight penalty imposed on a satellite having a ten year operational lifetime while significant is not prohibitive.

The very difficult engineering problems which remain to be solved concern reliability in the space environment. We estimate that the hardware plus fuel required to correct an initial velocity error of 100 meters per second, keep the satellite on station and stabilize its attitude for ten years invokes a weight penalty of about 16 percent of the total weight of the spacecraft. Thus $W_{\text{total}} = W_{\text{payload}} + 0.164\ W_{\text{total}}$ or $W_{\text{total}} = 1.20\ W_{\text{payload}}$, and this is the relationship used here to calculate the total spacecraft weight once the weight of the communication equipment, solar power supply, and other necessary components has been determined.

5.2 *Solar Power Plant*

Values of 15 to 20 watts per pound have been reported for sun-oriented solar cell arrays, but 10 watts per pound is closer to present performance. We use two watts per pound. This is reasonable for an unoriented array and very conservative if, in fact, it is decided to point the solar power plant at the sun. Since a single two watt transponder will require nearly 18 watts total solar power, power for each transponder will require a weight of nearly 9 pounds as its *pro rata* share of the solar power plant weight.

5.3 *Structure and Antenna*

An allowance of 20 percent of the total payload weight will be assumed for satellite structure and a like amount for the associated communications antenna together with its feeds. Admittedly this is a

guess based on previous experience and would require an innovative design to realize a lightweight antenna operable at 30 to 40 GHz and able to survive the launch environment. However, once the satellite is in orbit and on station, and this is the only condition under which it would be used, the disturbing forces will be very small.

As already mentioned, an attitude and orbit control system capable of keeping a satellite on station and properly oriented for 10 years will require a total in-orbit weight of $W_{\text{total}} = 1.20\ W_{\text{payload}}$. Thus, the payload weight required for eight broadband channels to serve one ground terminal is

$$\begin{aligned} W_{\text{payload}} &= 100 \text{ lbs. for eight two-watt transponders} \\ &\quad + 0.2\ W_{\text{payload}} \text{ for structure} \\ &\quad + 0.2\ W_{\text{payload}} \text{ for antenna.} \end{aligned}$$

thus

$$W_{\text{payload}} = 167 \text{ lbs. per ground station}$$

and

$$W_{\text{total}} = 1.20\ (167) = 200 \text{ lbs. per ground station.}$$

VI. EARTH STATIONS

As already pointed out, a terrestrial terminal will consist of two or more diversity locations each equipped with a multibeam antenna suitable for both transmitting and receiving. The number of diversity locations required, and the distance between them will be determined by the specified system reliability and the characteristics of rainstorms in the particular location being considered. Much more extensive (geographically) and complete data on space diversity will need to be obtained before such installations can be designed with confidence, but data available from the Bell Laboratories rain gauge network and other sources[10] make it reasonable to postulate that two sites separated by 10 to 20 miles will be adequate, and this is the assumption used in the following. Actually a study of data of the sort shown in Fig. 1 leads one to hope that a considerably smaller distance would suffice.

Provision of the first spare transmission facility is no great hardship since this is needed in any case to provide for maintenance and for protection against apparatus failure, but the diversity site must be connected by a broadband facility adequate to carry all of the traffic received or transmitted by either ground station. In the model system

this will vary from about 80,000 voice channels for a station working with a single satellite up to about four million voice channels for one that works with 50 satellites.

Since the most attractive transmission medium for such large volumes of traffic appears to be a millimeter waveguide of the type described by S. E. Miller,[11] a facility of this sort is assumed as the site interconnecting link. A fully implemented terminal handling four million voice circuits would require eight pipes of two-inch inside diameter, each carrying 60 two-way 630 megabits per second channels. It is expected that for distances up to 15 or 20 miles between diversity ground stations no intermediate repeaters would be required when a low-loss waveguide is used. Concentration of the electronics at two sites which presumably would be attended at least part time should simplify maintenance and increase reliability. This site interconnecting link, with its large volume of traffic to be carried over a short distance and without a need to demultiplex the individual voice circuits at either end, is a good candidate for an optical transmission link when technology has progressed far enough to make such an installation feasible.

The proposed ground station antenna is of the type being studied by A. B. Crawford and T. S. Chu.[12] This is a multibeam antenna 10 meters high by 17 meters long designed for up to 10 feeds for use with a like number of satellites. The common reflector is so designed that working in conjunction with the multiple feeds, good antenna patterns can be maintained over a 10° change in the azimuth angle but with no provision for variation in declination angle. The latter is obtained by tilting the entire assembly to the declination angle appropriate to the particular site.

As already mentioned, these systems will be interference limited, hence good antenna discrimination is more important than gain, and the antenna design takes advantage of this fact. In the long run electronically steerable arrays may prove to be the best and most economical solution to the multibeam antenna problem both on board the satellite and at the ground station. In fact, if crowding of the synchronous orbit eventually makes it imperative to use inclined 24-hour orbits as described by H. E. Rowe and A. A. Penzias, steerable arrays appear to be the only workable solution.[13] At the present time, however, multiple feeds used with a focusing reflector appear to be the most attractive solution.

As indicated on Table I, an over-all effective noise temperature of 150°K is assumed for the ground terminal receiver; this can best

be achieved by using a cooled parametric amplifier. Considerably lower effective temperatures are possible. But with an interference limited system operating at 20 GHz where the effective antenna temperature will approach 300°K under conditions of heavy rainfall, little improvement in actual system performance would be obtained. Although Table I specifies a 10 watt 30 GHz transmitter which provides more than 40 dB C/N at the satellite (and this may well be adequate), increased transmitter power is feasible and would not be very expensive if further propagation data makes this appear necessary.

VII. LAUNCH CONSIDERATIONS

The cost of placing a given weight in orbit can be obtained for many different combinations of upper and lower stages. Costs differing by nearly 10 to 1 will be found, depending on the time, scale of production, size of launcher used, and so on. Such a wide variety of combinations of lower and upper stages including future exotic high energy upper stages and injection rockets are technically possible (some are parts of on-going funded and scheduled programs while others exist only on paper), that large variations in load capacity and cost estimates are to be expected.

The time scale being considered is an especially confusing factor; things are always predicted to be much better in the future. In thinking about the present system model, particularly satellite vs ground terminal size and complexity, it has been assumed that launch cost is directly proportional to payload weight. How realistic is this? Rocket capabilities tend to be quantized, and certainly one must pay the entire launch cost if it is made for his benefit. However, if we take a longer range view, a wide variety of combinations of lower and upper stages will become available.

For example, the basic Titan 3-C will place 2,140 pounds in synchronous orbit, but various combinations of lower and upper stages based on Titan have been proposed which would place 650, 2140, 2450, 3450, 5800, 9000, 18,000, and even 23,000 pounds in synchronous orbit. It is true that many of these combinations will require considerable research and development before a reliable launch vehicle results, but at least in concept it is possible to tailor the rocket to the job. In addition, we have the possibility, now amply demonstrated, of combining several small payloads into one for multiple launch. Combinations of these two techniques should eventually make it possible to match the rocket and its payload quite precisely.

VIII. SUMMARY AND CONCLUSIONS

It has been argued that by taking full advantage of modern space and communication technology and assuming that very broad frequency bands in the range between 15 and 40 GHz can be allocated to this service, realization of a domestic satellite system with considerable communication capacity is feasible within the next decade. Some of the parameters of satellites designed for such a system are given in Table III.

Major unresolved technical problems are reliability of the propagation of these very short radio waves through the earth's atmosphere and stabilization of a space platform to $\pm 0.01°$ for up to ten years. Data adequate to provide a solid statistical base for predicting system performance are not available, but a start has been made and more experiments are planned. Several programs are dedicated to improving performance of satellite attitude stabilization systems and achievement of the required value has been predicted. A preliminary study of the basic forces acting on a satellite shows that attitude stabilization even to these very precise limits does not require prohibitive amounts of fuel. However, a considerable advance in technology will be required before an attitude stabilization subsystem of adequate reliability and precision is available. The very broadband radio repeaters which are required to implement this concept are less of a problem since these are already being developed for other purposes and can be adapted for satellite use.

TABLE III — SUMMARY OF SATELLITE PARAMETERS

	Number of ground stations served			
	8	16	32	50
Weight of satellite (pounds)	1640	3280	6560	11,500
DC power (kw)	1.15	2.30	4.60	7.20
Transponders	64	128	196	400
Gigabits through satellite	40	80	160	252
One-way equivalent voice circuits through satellite (thousands)	640	1,280	1,960	4,000
One-way equivalent TV channels through satellite				

In addition to the problems peculiar to the present proposal, there are, of course, those common to all synchronous satellites: an approximately ½-second absolute round-trip delay and the attendant echo suppression requirement at a cost which is not negligible, slow delay variations resulting from path length changes which may amount to several microseconds, service outages caused by interruption of solar power (or provision of storage batteries), and outages caused by sun transit, which will increase the 20-GHz receiver effective noise temperature from the assumed 150°K to about 15,000°K (20 dB).

Assuming a favorable outcome from the propagation measurements, particularly a demonstration that common carrier standards for reliability can be achieved with diversity ground stations, planning for such a system could begin immediately. Other requirements of a nontechnical nature include resolution of the frequency allocation problem and policy decisions regarding system ownership and operation.

IX. ACKNOWLEDGMENTS

This paper has benefited from discussions with R. Kompfner, A. B. Crawford, D. C. Hogg, and C. L. Ruthroff. Mr. Ruthroff was especially helpful with Section 4.2. It has also been improved by comments received from H. W. Evans and others in the Systems Engineering Division of Bell Telephone Laboratories who suggested a more detailed discussion of interference and provision of larger working margins.

REFERENCES

1. Hogg, D. C., "Millimeter-Wave Communication Through the Atmosphere," Science, *159*, No. 3810 (January 5, 1968), pp. 39–46. This paper summarizes our understanding of millimeter wave propagation through the atmosphere at the time the present experimental programs were undertaken.
2. Semplak, R. A., "Gauge for Continuously Measuring Rate of Rainfall," Rev. Sci. Instruments, *37*, No. 11 (November 1966) pp. 1554–1558.
3. Wilson, R. W., "A Millimeter Wave Sun Tracker," International Scientific Radio Union, Commission 2, Washington, D. C., April 9, 1968.
4. Li, T., and Semplak, R. A., unpublished work.
5. Metzger, P., Proceedings International Scientific Radio Union Meeting, Washington, D. C., 1966.
6. Lee, R. W. and Waterman, A. T., Jr., "A Large Antenna Array for Millimeter Wave Propagation Studies," Proc. IEEE, *54*, No. 4 (April 1966), pp. 454–458.
7. Turrin, R. H., unpublished work.

8. Cutler, C. C., "Remote Attitude Control of Earth Satellites," U. S. Patent 3,060,425, issued October 23, 1962, and "Attitude Control for Satellite Vehicles," U. S. Patent 3,088,697, issued May 7, 1963.
9. Morgan, S. P., Burford, T. M., and Sinden, F. W., unpublished work.
10. Marshall, J. S., Haltz, C. D., and Weiss, M., "McGill University Scientific Report MW-48" (1965).
11. Miller, S. E., "Waveguide as a Communication Medium," B.S.T.J., *33*, No. 6 (November 1954), pp. 1209–1265.
12. Crawford, A. B., and Chu, T. S., unpublished work.
13. Rowe, H. E. and Penzias, A. A., "Efficient Spacing of Synchronous Communication Satellites," B.S.T.J., this issue, pp. 2379–2434.

Path Length Variation in a Synchronous Satellite Communications Link

By ANTHONY G. LUBOWE

(Manuscript received April 30, 1968)

The path length and path length rate variations in a communications link connecting earth stations on the East and West coasts of the continental United States by means of an equatorial synchronous satellite are investigated numerically and shown to be less than 50,000 feet and 5 feet per second for periods of at least a day (and likely for a week or more) with no station-keeping during the periods of interest.

I. PROBLEM

The use of a synchronous satellite in a communications link with time division switching for digital channels may introduce problems because of the path length variations resulting from the perturbation of the satellite orbit. For instance, a path length variation of 0.02 percent (10 miles for a typical satellite-tracker configuration) corresponds to a timing change of 0.05 ms. The rate of change of path length (length rate) is also of interest. The object of this paper is not to apply the results to current problems, but to make the results available to prospective users.

Analyses of the synchronous satellite problem[1, 2] have emphasized the long period motion. However, the relatively small short period perturbations have important effects upon path length and length rate. Also we expect the long period motion to be handled by discrete, rather than continuous, station keeping (perhaps by a sequence of impulsive maneuvers every few days) so that the major orbital perturbations will be those of short period.

We assume an equatorial orbit, with the longitude fixed by the communication requirements at 100°W. We must specify the initial altitude and velocity so that the satellite is synchronous. If we assume that the earth's inverse-square gravity results in a circular equatorial orbit with mean motion equal to the earth's rotation rate, then it can

be shown that perturbative forces such as the J_2, J_3, J_4 zonal harmonics and the J_{22} tesseral harmonic introduce radial and tangential forcing terms (e and n) into the equations of motion.[2]

II. ANALYSIS

If the initial satellite longitude is equal to the longitude of the major axis of the earth's elliptical equator, then the initial radius, r_s, can be chosen so e and n are both equal to zero (if higher order terms are neglected). If the initial longitude must be different, then $n = 0$ cannot be obtained and there will be drift away from the desired longitude. However, we can still obtain $e = 0$ by proper choice of r_s (with little effect upon n) and this is the procedure we follow.

We do not include the luni-solar gravity perturbations (which we include in our computer runs) in the choice of initial conditions, since these perturbations are time-varying in an earth-fixed frame and the details of their inclusion depend upon the interval of interest.[1] Their effect is small, and the complexity they introduce into an operational procedure probably outweighs any reduction in the perturbations.

TABLE I—PATH LENGTH AND LENGTH RATE VARIATIONS

Date	$L_{MAX} - L_{MIN}$ (feet)	$\dot{L}_{MAX} - \dot{L}_{MIN}$ (feet per second)
1–01:00	16,200	1.45
1–21:16*	30,700	2.69
4–01:00	27,000	2.58
4–20:21*	32,400	2.88
7–01:00	35,600	3.40
7–18:05*	31,800	3.12
8–16:12*	40,700	4.24
9–14:19*	38,900	4.03
10–01:00†	36,100	3.71
10–08:00	12,600	1.27
10–14:04*	35,900	3.56
10–15:00	41,300	4.22
10–22:00	12,900	1.14
10–29:00‡	38,400	3.85
11–12:14*	45,200	4.15
12–12:03*	29,700	2.75
12–22:00	18,900	1.65

* New moon
† Full moon on 9–29:17
‡ Full moon on 10–29:10

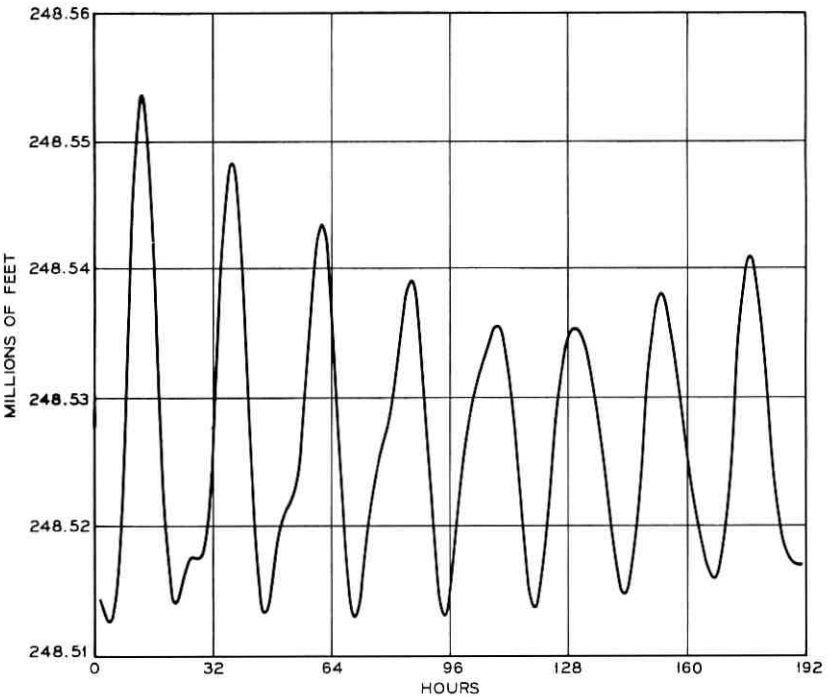


Fig. 1 — Path length variation for eight days starting 10–15:00.

III. RESULTS

The results were obtained using modifications of existing computer programs to integrate the equations of motion of the satellite numerically.[3]

The ground stations are located at 38°N latitude and at 75°W and 125°W longitude. The initial satellite longitude is midway between the tracker longitudes. We list some results in Table I. In each case the satellite has been injected into orbit, or the orbit has been corrected, so that the radial forcing function is zero on the date shown. (The dates are given as "month-day:hour," so that 1–21:16 means January 21 at 4 p.m., Greenwich time. All runs were made for 1966 since simple luni-solar position routines were available for that year; however, they should be quite representative of the variations to be expected.)

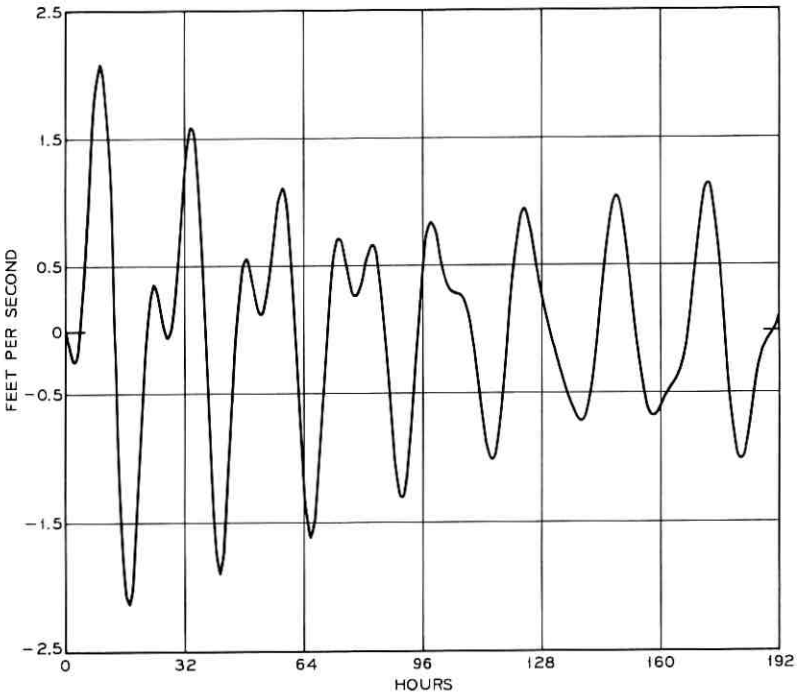


Fig. 2— Path length rate variation for eight days starting 10–15:00.

The maximum path length and length rate variations differ considerably over the solar cycle but even more over a lunar cycle. This is reasonable since the solar perturbation on a synchronous satellite is only 1/4 of the lunar perturbation. Peak-to-peak variations in L and $\dot{L}$ were less than 50,000 feet and 5 feet per second in all cases. The maximum variations during the lunar cycle occur around new moons and to a slightly smaller degree around full moons, and in the Fall of the solar cycle. The length variations shown in Table I result from all satellite position components. We notice that radial, longitudinal, and latitudinal components were in all cases less than 2,500 feet., 0.03°, and 0.005°, respectively, and often considerably less.

To see the effect of omitting daily stationkeeping, the satellite was allowed to move uncorrected for eight days beginning on a day of high and one of low variation. (See Figs. 1 through 4.) An interesting and rather unexpected result is that in both cases, the maxima for the one week runs are the same as for those one day runs with the

larger maxima, whether they occurred at the beginning or end of the week. Thus, one might tolerate stationkeeping once per week rather than daily if the main orbital constraints were peak-to-peak variations in L and $\dot{L}$ of less than 50,000 feet and 5 feet per second.

Another orbital constraint might be a maximum allowable longitudinal drift, say $\Delta\lambda < 0.1°$. The drifts are approximately linear for the first few days, that is, for

$$\text{Oct. 15 to 16,} \quad \Delta\lambda = 0.026°$$

$$\text{Oct. 15 to 23,} \quad \Delta\lambda = 0.199°$$

and for

$$\text{Oct. 22 to 23,} \quad \Delta\lambda = 0.0023°$$

$$\text{Oct. 22 to 30,} \quad \Delta\lambda = 0.0250°.$$

Thus a constraint of $\Delta\lambda < 0.1°$, might require stationkeeping every

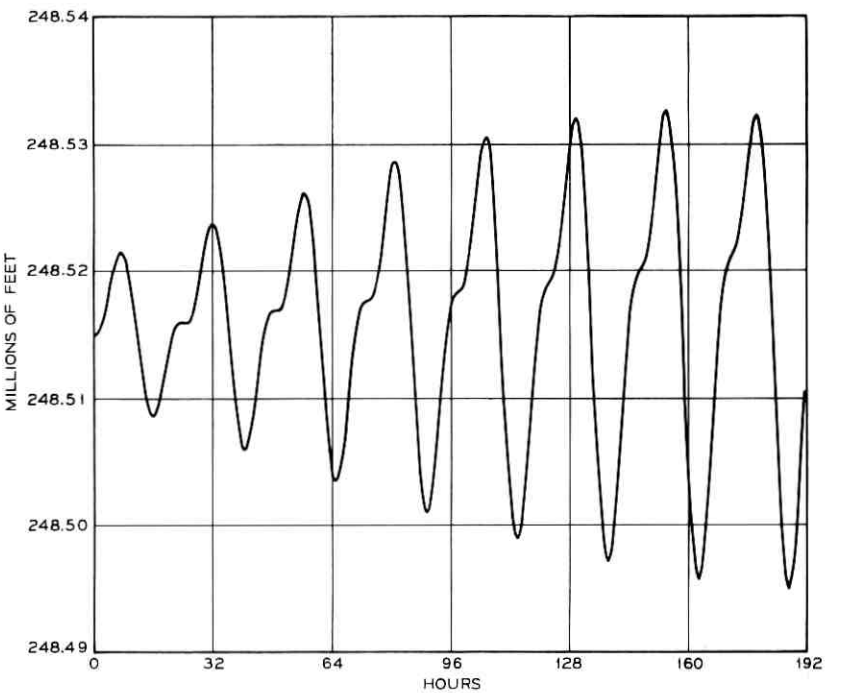


Fig. 3 — Path length variation for eight days starting 10-22:00.

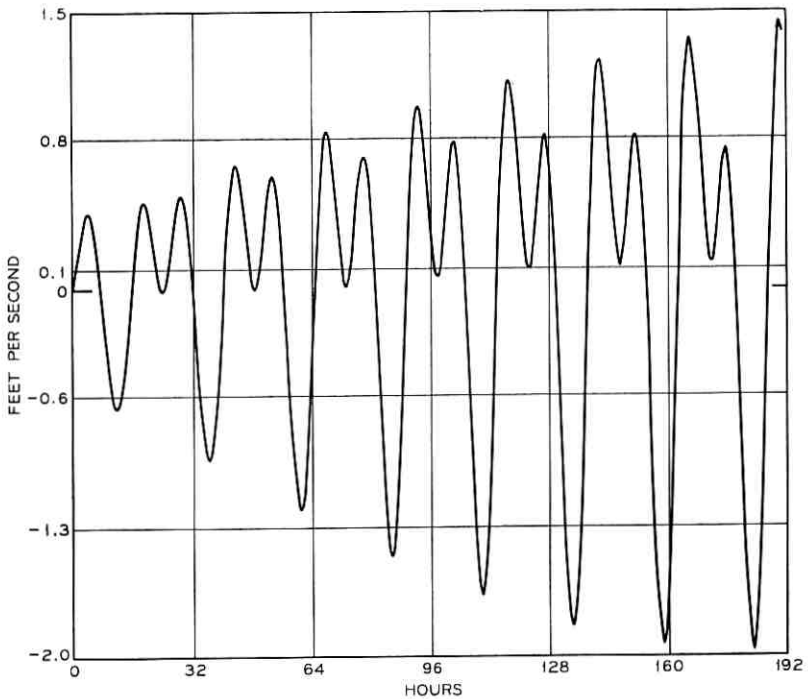


Fig. 4 — Path length rate variation for eight days starting 10–22:00.

3 or 4 days in a period of large variations, although the path length and length rate constaints would not be violated for at least a week.

IV. ACKNOWLEDGMENT

I am grateful to Mrs. C. B. Wood for obtaining the computer results, to D. Berardinelli for several helpful discussions, and to F. Elenbaas and M. P. Wilson for comments on the manuscript.

REFERENCES

1. Frick, R. H. and Garber, T. B., "Perturbations of a Synchronous Satellite," Rand Corporation Report R-399-NASA, May 1962.
2. Wagner, C. A., "The Drift of a 24-Hour Equatorial Satellite Due to an Earth Gravity Field through 4th Order," NASA Technical Note TN D-2103, February 1964.
3. Lubowe, A. G., "Orbital Behavior of Large Synchronous Satellites," Astronautica Acta, *13* (1967) No. 1, pp. 49–61.

Parametric Representation of Ground Antennas for Communication Systems Studies*

By D. L. POPE

(Manuscript received April 9, 1968)

Mathematical models relating the gain, cost, diameter, frequency, and rms surface tolerance of ground antennas are developed for both exposed and radome-enclosed parabolic reflectors. Diameters considered range from 15 to 500 feet, while frequencies vary from 1 to 100 GHz. Data from existing installations are used to develop standard cost vs diameter and rms surface tolerance vs diameter relationships. The standard cost is associated not only with diameter, but also with standard surface tolerance. Quality factors are introduced to relate deviations from the standard rms surface tolerances to expected departures from the standard cost curve. The models are completed by the inclusion of an approximate relation for the gain of parabolic reflectors. Each model comprises two equations among five variables. Although they are relatively complex, these models should be valuable to the communication systems planner in considering the gross features of alternative concepts for ground antenna installations. They are intended as guidelines for the conceptual stages of communication systems development, and are especially useful in terms of the trade-off studies they encompass. Three examples dealing with typical questions of this type illustrate their use.

I. INTRODUCTION

Mathematical models relating five important variables encountered in the consideration of ground antennas for communication systems are developed for both exposed and radome-sheltered structures. The variables are: diameter, cost, gain, frequency, and rms surface tolerance. Often, in communication systems studies, the costs associated

* A preliminary version, "Trade-off Models for Ground Antennas in Microwave and Millimeter Wave Communication Systems," was presented at the Fifth Space Congress, Cocoa Beach, Florida, March 1968.

with a ground antenna are estimated by using a simple relationship based on antenna diameter alone. The results of our study have a similar utility, but are substantially broader in scope and have greater flexibility. They represent some of the major features of ground antennas for use in the preliminary phases of communication systems planning.

In spite of the refinements offered by the present models in comparison with previously available guidelines, parameters which are important in the design and operation of a specific ground antenna system are not included. For example, there is no provision for including antenna noise temperature. Hence, the comparison of concepts on the basis of signal-to-noise ratio is not possible. Similarly, the effect of aperture illumination and the side lobe levels to be expected are not considered. There are other factors which must be accounted for in the design process but do not appear in the present formuation. For any specific application, performance requirements are carefully defined and vary considerably depending on the intended use of the antenna. It remains a challenge to the antenna specialist to optimize his design in order to meet those performance requirements in the best possible way. The present models, which characterize antennas in terms of only a few of their gross features, cannot and should not be expected to apply at such levels of refinement.

In this study, diameters range from 15 to 250 feet for the exposed antennas and 30 to 500 feet for antennas enclosed by a radome. Frequencies vary from 1 to 100 GHz. Only conventional reflectors are included. No consideration is given to actively controlled surfaces, multiple antenna synthetic apertures, or other such concepts which may be important in the future.

The first relation introduced is Ruze's formula relating gain to diameter, frequency, and rms surface tolerance. Two additional equations which relate a standard rms surface tolerance to diameter and a standard cost to diameter are developed from information available on existing and proposed antenna installations. Both the standard rms and the standard cost relationships are based on the same specific data points within each class of antennas. The points used are believed to be consistent and span the diameter range of interest. The result of this interpretation is a standard cost relationship which depends not only on the diameter of the antenna, but also is related to its surface tolerance.

The specific correlation of cost, diameter and rms surface toler-

ance is a novel feature of the present approach. Finally, quality factors are introduced to associate departures from the standard rms surface tolerance with departures from the standard cost curve. The functional relations chosen to represent these factors are justified with largely heuristic arguments because the existing information does not permit a more precise determination. However, when the data that is available on cost and rms surface tolerance is adjusted to a common standard using the functions chosen, the resulting agreement is encouraging.

Each model consists of two equations among the five variables of interest. Although much more complex than the simple power law relationship often used to represent the cost vs diameter of ground antennas, these sets of equations contain considerably more information. Not only do they readily yield information about any specific case, but also they provide a starting point for various optimization studies. Three examples are given, dealing with maximum cost-effective antennas, minimum cost antennas given the gain and the frequency, and the variation of gain for a specified cost and frequency. The first two examples are examined for both exposed and enclosed antennas. These examples suggest others that also could be done.

II. THE PROBLEM

The most elusive part of this study was determining satisfactory rms surface tolerance vs diameter and cost vs diameter relationships for ground antennas. Many reports were studied and several personal contacts made in the attempt to assemble sufficient information to draw the necessary conclusions. In spite of this effort, the functional relations which are ultimately suggested to represent standard cost and standard rms surface tolerance variations with diameter remain substantially empirical. It is worthwhile to discuss some of the reasons.

The original hope was to process the available data on existing and proposed antenna structures by some statistically satisfactory technique in order to determine the most likely functional forms for the relations of interest. This approach was abandoned for two reasons. In the first place, the number of antennas for which there is reliable information available on cost and rms surface tolerance is too small to admit a satisfactory statistical treatment. Secondly, cost and rms surface tolerance are vague and hard-to-define concepts which admit differing interpretations in each case. The lack of any common stand-

ards for reporting these quantities makes it unrealistic to determine functional relations involving such quantities by formal manipulations of the available data.

It is not hard to understand the reasons for the ambiguities in the reported data. The problems involved in measuring the surface tolerance of a large paraboloidal reflector are difficult, and to perform such a measurement is expensive and time-consuming. Often user demands on high performance antennas are sufficient to prohibit measuring the surface tolerance in any sort of a statistically satisfactory way. The measurement techniques themselves can place constraints on the structure, such as zero zenith angle and benign environmental conditions, which are unrealistic in terms of operational requirements. The data reduction process to determine the tolerance figure reported may introduce extraneous variables or unsuspected biases influencing the result. In a few cases, the surface accuracy has been calculated by measuring the gain over a range of frequencies and then using Ruze's gain equation, in which the rms surface tolerance is assumed to be the only unknown. This seems to be a powerful and effective technique, but it supposes a knowledge of the aperture efficiency at each frequency, a quantity that is extremely difficult to determine independently.[1]

Further ambiguity is introduced by neglecting to define the important concepts carefully. The rms surface tolerance can be measured with respect to the best-fit paraboloid or to the original design contour. It often contains a systematic as well as a random component, which may or may not have been eliminated in the published value. It can, of course, be a deviation normal to the reflector surface or normal to the aperture plane, although the distinction is not especially significant for shallow reflectors. In a few cases, it is the maximum peak-to-peak deviations that are reported and an equivalent rms surface tolerance must somehow be found.

The resolution of the cost associated with existing and proposed antennas, while not encumbered with the technical problems of surface tolerance determination, is beset with other difficulties. A high performance antenna is a custom-made item. The price reflects necessary research and development, engineering, tooling, and fabrication costs which are difficult to determine precisely and which cannot be spread over a large number of units. The requirements of each situation must be dealt with separately. There are relatively few companies building such structures, and the competition is fierce. Pricing

information and guidelines are proprietary and are simply not available to an interested outsider.

A different sort of problem arises in the attempt to establish the costs of existing structures. A specific cost figure can generally be found for most of the antennas in operation today, but it is difficult to determine exactly what the reported number of dollars bought. There are numerous ancillary items with a ground antenna that may or may not be included: the electronics, feed structure, servo systems, data readout and transmission, power plants, land acquisition, support buildings, heating, lighting, ventilation, and so on. Seldom is the reported cost broken down in sufficient detail. Meaningful cost comparisons cannot be made without knowing which subsystems are included in the reported cost and which are not.

In view of such uncertainties regarding the costs and surface tolerances of existing structures, and the relatively small numbers of such high performance antennas in operation, it is clear that a statistical approach to determining functional relations among the variables of interest would be illusory. Instead, the standard cost vs diameter and rms surface tolerance vs diameter relations are established by considering only a few data points (3 for exposed antennas and 4 for antennas with a radome) which span the diameter range of interest and seem to form a consistent subset of the data available.

The outcome of this line of reasoning is a pragmatic and qualitative model consisting of various relations among the variables of interest. It is not strictly defensible on grounds of statistical rigor, in spite of the rather satisfying way in which the available data are shown to fit within its structure. It is certainly neither unique nor absolute. Its usefulness lies in its reflection of acknowledged trends and its capacity as a basis for comparisons, trade-offs, and various sorts of optimization studies on ground antenna systems at a relatively coarse level.

III. THE GAIN FORMULA

In 1952, Ruze suggested a formula for the on-axis gain of a reflector antenna.[2] This formula has been generally acclaimed and enjoys wide popularity in spite of the restrictive assumptions it incorporates. These assumptions were clearly restated by Ruze in his 1966 article.[1]

Ruze's formula states:

$$G \cong \eta\left(\frac{\pi D}{\lambda}\right)^2 \exp - \left(\frac{4\pi\epsilon}{\lambda}\right)^2. \tag{1a}$$

Here D is the reflector diameter, λ is the wavelength at the frequency of interest, ϵ is the rms deviation of the reflector surface from the best-fit paraboloid, and η is the aperture efficiency, a measure of the overall electronic properties of the antenna. D, λ and ϵ must be in consistent units.

The leading factor in Ruze's formula expresses the gain for a perfect reflector. The effect of deviations from a perfect paraboloid are contained in the exponential factor. No distinction is made between manufacturing inaccuracies and deflections of the reflecting surface resulting from environment. The gain of a given antenna, with a specified diameter and surface tolerance, first increases as frequency is increased. However, a point is reached at which the exponential factor takes over, and a further increase in frequency results in a decrease of the gain. The point at which the gain is a maximum is called the gain-limit point.

The same phenomenon can be noticed if the frequency is held fixed and the diameter is varied. The cause for a gain-limit point in diameter is not immediately apparent from equation 1a, but it occurs because the rms surface tolerance is also a function of diameter. Stack has pointed this out in his work,[3] and he gives curves of gain vs diameter for a number of frequencies.[4]

The aperture efficiency, η, includes the effect of nonuniform illumination, spill-over, aperture blockage, front-end losses in antenna electronics and other factors which contribute to degradation in performance. It specifically does not include the effects of an imperfect reflecting surface. For a properly engineered antenna, η should lie between 0.65 and 0.75 (see Ref. 5). The aperture efficiency also depends to a certain extent on antenna configuration. In addition, the aperture efficiency depends on operating frequency for a given reflector. As a result of the uncertainty associated with the aperture efficiency, when equation 1a is used in this report, the aperture efficiency is simply taken to be 70 per cent. Refinements of this assumption would require information that is not available.

Equation 1a requires modification in order to use it for antennas with a radome. Experience with operation of radome-enclosed antennas has been generally satisfactory at microwave frequencies. At such frequencies the radome is responsible for approximately 1 dB loss in gain mostly from aperture blockage.[6] It also contributes to system noise temperature. The total system degradation depends to a considerable extent on local weather conditions. Consideration of

these effects is beyond the scope of this discussion. Further details can be found in Ref. 6.

There is little experience with radomes at millimeter frequencies. In this regime, the radome thickness is no longer small with respect to wavelength and special design techniques will clearly be necessary to minimize losses. For the present, we assume that equation 1a can be modified appropriately by means of a multiplicative factor

$$G \cong R(\lambda)\left[\eta\left(\frac{\pi D}{\lambda}\right)^2 \exp -\left(\frac{4\pi\epsilon}{\lambda}\right)^2\right]. \tag{1b}$$

The factor $R(\lambda)$ is chosen to represent the loss in gain caused by the presence of a radome, both because of aperture blockage and path losses in the radome.

The gain calculated using Ruze's formula (1a or 1b) does not include atmospheric effects that can degrade signal strength, such as turbulence or rain. These effects may be extremely important, particularly at high frequencies, but are not explicitly part of the ground antenna considerations.

IV. COST AND DIAMETER

When the information available on the costs of ground antennas is plotted with the diameters, a substantial scatter of the data is evident. Even with logarithmic coordinates, the dispersion precludes a satisfactory straight-line fit to the data points over the entire diameter range. Such a straight-line fit in logarithmic coordinates would correspond to the familiar power law relation, cost = (constant) (diameter)^n. A piecewise-linear cost function, corresponding to an increase in the power law exponent with diameter would be much better, but would introduce troublesome analytic complications.

To establish the standard cost vs diameter relationship, we have selected three antenna structures which span the diameter range of interest, and have fit a three-parameter expression to these three data points. The antennas chosen are:*

(*i*) The 15-foot antenna operated by Aerospace Corporation, El Segundo, California.[7]

(*ii*) The 85-foot antenna operated by the Naval Research Laboratory at Maryland Point, Maryland.[8]

---

* A typical reference is given for each antenna. There are several other sources describing each of the antennas.

(*iii*) The 210-foot antenna operated by the Jet Propulsion Laboratory at Goldstone, California.[9]

All three antennas were manufactured by the same company, and all are exposed and fully steerable, although the first two have polar mounts and the third has an azimuth-elevation mount. Good rms surface tolerance information is available for all three. More importantly, the cost information obtained from user and manufacturer agrees reasonably well for all three structures. The costs cited include the structure, drives, and control, but do not include electronics, readout equipment, or other ancillary costs, insofar as could be determined. The standard cost-diameter relation obtained this way is

$$\$^* = 6.7(10)^5 D^{-\frac{1}{3}} \exp(D/45). \tag{2a}$$

This curve is shown in Fig. 1. In equation 2a the diameter D is in feet. Potter's power law curve[10] for the 85- to 250-foot range is also shown in Fig. 1.

Although the relation 2a fits the three selected points very nicely,

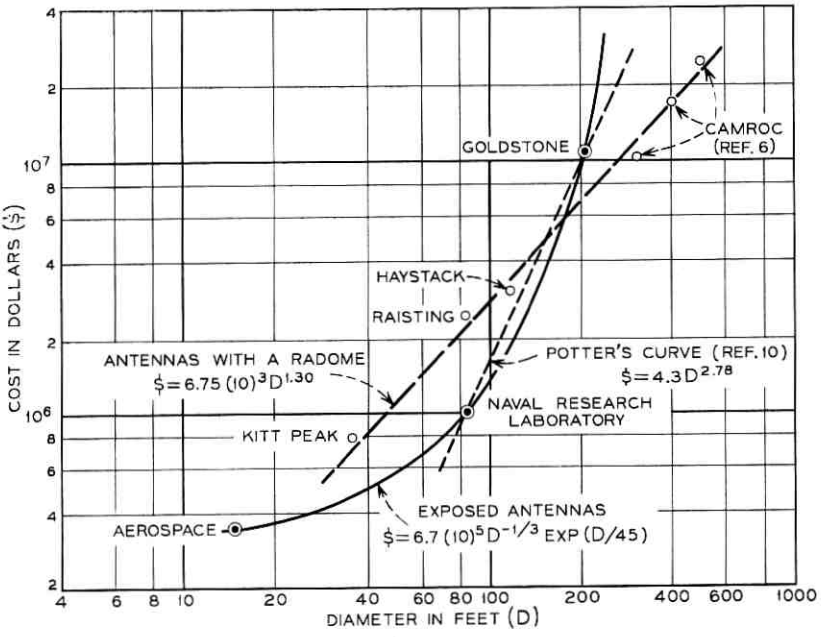


Fig. 1 — Antenna cost vs antenna diameter for both exposed and radome enclosed antennas (standard curves).

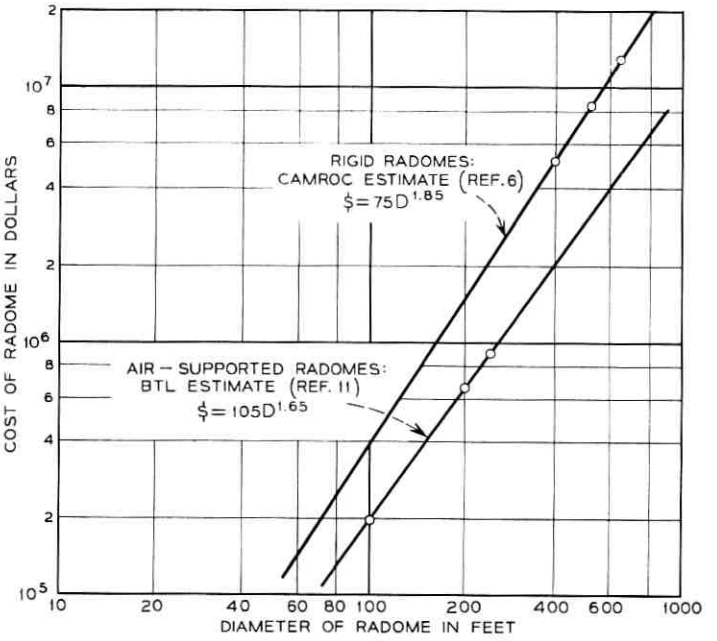


Fig. 2 — Radome cost vs radome diameter for rigid and air-supported radomes.

problems can occur if unconscious extrapolation is attempted. Beyond 210 feet, the costs increase very rapidly with diameter because of the exponential factor. There is a singularity at $D = 0$, and costs again increase as diameter decreases below 15 feet. The exponential increase in cost for very large antennas is probably not entirely unrealistic. However, relation 2a should be used only in a diameter range from 15 to 250 feet.

The cost-diameter relation for antennas with a radome is also shown in Fig. 1. In this case, the items included in, and excluded from, the reported cost are the same as for exposed antennas with one important exception. The cost of the radome is included. A simple power-law relation seems to be satisfactory for these antennas over a 30- to 500-foot diameter range. This relation is

$$\$^* = 6.75(10)^3 D^{1.30}. \tag{2b}$$

Again, the diameter is expressed in feet.

The estimated cost of the radome alone, including foundation and environmental control equipment, is shown in Fig. 2. Both air-sup-

ported and rigid space frame radomes are considered. This information is taken directly from Ref. 6 for the rigid radomes and from Ref. 11 for the air-supported radomes. Considerable extrapolation is required in both cases to cover the entire diameter range of interest. In addition, this data deals entirely with radomes designed to enclose antennas operating at relatively low frequencies. For the higher frequencies, special designs for the radome will have to be found in order to minimize losses and noise. Manufacturing and construction tolerances will be substantially more stringent than those reflected in the prices represented by Fig. 2. Radome cost will clearly depend on the operating frequency as well as on diameter. However, since there is virtually no experience with high frequency radomes, even the approximate nature of this dependence is unknown. In lieu of a more appropriate representation, the relations shown in Fig. 2 will be used in the present model.

The diameter of the radome required to enclose an antenna of diameter D is assumed to be $4/3\ D$. The cost of the antenna alone can now be determined using both Figs. 1 and 2.

V. SURFACE TOLERANCE AND DIAMETER

The data available on rms surface tolerance of existing antennas is plotted in Fig. 3. The ranges shown with many of the data points are attributable to a number of factors. In a few cases, they reflect honest uncertainty. In others, they represent a range of reported values resulting from the different tolerances at different elevation angles, or under different environmental conditions. In some cases the range shown encompasses values reported from different sources for the same antenna. For one or two antennas, the range shown represents the design goal.

The surface tolerance data shown represents the surface accuracy under operating conditions. Thus all factors that combine to produce mechanical deviations from a perfect paraboloid are included. These include manufacturing inaccuracies, as well as surface deflections caused by gravity, wind, and thermal effects. In general, the antennas represented are fully steerable, and operate satisfactorily in steady winds up to about 30 mph and other environmental conditions normally expected for such antennas.

The antennas used as a basis for the cost curves (Fig. 1) also determine the rms tolerance vs diameter curves. These curves are given by two straight lines; one for exposed antennas, and the other for

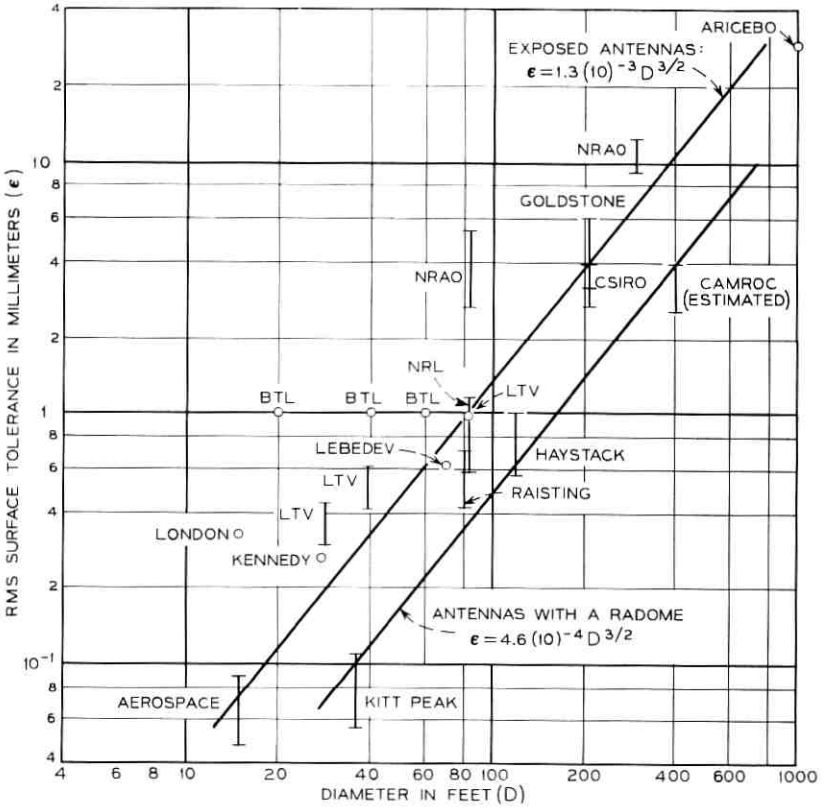


Fig. 3 — Antenna rms surface tolerance vs antenna diameter for both exposed and radome enclosed antennas (standard curves).

antennas operated inside a radome. The appropriate functional relationship is of the form:

$$\epsilon^* = \alpha D^{\frac{3}{2}} \tag{3}$$

where ϵ^* is the standard rms surface tolerance in millimeters, and D is the reflector diameter in feet. Caution: take special note of this rather unusual juxtaposition of units. The constant α is;

$$\alpha = 1.3(10)^{-3} \text{ for exposed antennas,}$$
$$\alpha = 4.6(10)^{-4} \text{ for antennas with a radome.}$$

The curve for reflectors protected by a radome has the same slope as the one for exposed structures but, at any given diameter, the surface

errors are considerably less for the enclosed structure because of the benign environment.

Since the same points have been used to develop the standard expressions, the cost-diameter relations (equations 2) give not just the cost of an antenna of diameter D, but specifically the cost of an antenna of diameter D with a surface tolerance given by equation 3. This correlation between the cost-diameter and rms-diameter curves is extremely important. The two relationships, taken together, express cost in terms of diameter and rms tolerance. The step from rms tolerance to frequency is simple, thus the cost is implicitly related to diameter and frequency. The costs and rms surface tolerances defined by these curves are called the standard values, and are indicated by the symbols ϵ^* and \$*, respectively.

VI. QUALITY

We now turn to a determination of the effects of departures from the standard curves. For example, how much can be saved by relaxing the rms requirement at a given diameter? How much more will it cost to improve the surface tolerance at some given diameter? These questions are typical of many others, such as, is it better to increase diameter or surface tolerance to achieve desired performance?, and the possibilities of an interesting optimization study begin to emerge.

To deal with these questions, we introduce quality factors to relate departures from standard rms surface tolerance to departures from the standard cost. This ingenious approach to the problem was first introduced by Stack.[3, 4] The actual rms (ϵ) and actual cost (\$) can be expressed as

$$\epsilon = f_1\epsilon^*,$$
$$\$ = f_2\$^*$$

where ϵ^* and \$* represent the standard values as shown in Figs. 1 and 3. For antennas with radomes, the cost appearing in these relations is the cost of the antenna alone. The problem now reduces to an appropriate selection of the quality factors f_1 and f_2.

The functions chosen are

$$f_1 = 1/x \tag{4a}$$
$$f_2 = \exp(x - 1). \tag{4b}$$

The parameter x provides the connection between the quality factors. For $x > 1$, the actual rms surface error is less than the standard rms

error as given by equations 2. Conversely for $x < 1$, the surface is less precise than given by the standard relations.

The range of the parameter is $0 \leqq x \leqq \infty$. This range includes the possibility of achieving a nearly perfect reflecting surface by requiring x to be very large. Physically, of course, this is not possible. There exists a limiting tolerance, almost certainly a function of diameter, beyond which the surface accuracy can no longer be improved. Unfortunately, this limit is unknown. Equations 4 represent a compromise with this situation. Although equation 4a admits the possibility of infinite improvement in rms surface tolerance, equation 4b associates an infinite cost with such an improvement. In fact, the cost factor f_2 expressed by equation 4b extracts a very heavy cost penalty for even modest rms surface tolerance improvement. In addition, expression 4b limits the possible reduction of cost to approximately $\frac{1}{3}$ of the standard cost, regardless of the reduction of quality of the reflecting surface.

The factor f_1 is the same as the one proposed by Stack.[4] However, the factor f_2 is significantly different. These factors are based on the realization that the standard curves of Figs. 1 and 3 represent very good reflecting surfaces indeed. It is reasonable to expect it to be extremely expensive to improve the surface quality still further, while some saving should result if the standards of accuracy were relaxed. The particular factors chosen would probably not be applicable if we had selected the three basic antennas from which the standard cost curve is derived nearer the center of the spectrum of available products. However, the three points actually used represent a definite bias toward the excellent, and this bias justifies the present choice of the quality factors.

There is insufficient good raw data on the rms tolerance and cost of existing antennas to establish the quality factors directly. At least two reliable data points would be necessary for each of several different diameters in order to succeed. We would expect that the quality factors should also reflect the influence of diameter. However, the inclusion of this effect could not be justified on the basis of the available information.

It is possible to check approximately the quality factors chosen against the data that is available. Although the cost of each antenna considered is influenced by factors not included here (such as environmental considerations and tracking requirements), the comparison will be carried out on the basis of rms surface tolerance and diameter alone. A value of the parameter x can be determined by comparing

the actual and the standard rms values according to Fig. 3 and using the definition of f_1. The appropriate cost factor f_2 can then be found from equation 4b. If the inverse of the calculated cost factor is applied to the reported cost of the antenna, a revised standard cost is found. This figure represents the expected cost of the antenna, had it been built to the standard rms surface tolerance. The points obtained by performing this exercise for several different antennas are shown in Fig. 4. While agreement is not perfect, it is considerably better than any possible fit to the unmodified raw data.

Of course, the quality factors given in equation 4 are not unique. Other functions can be found which provide the same sort of qualitative trends. Functions can be suggested which permit finite limits to be placed on the possible improvement or degradation of the rms surface tolerance, and finite limits can be established for the associated cost as well. Such functions are more complicated than those actually chosen, and the implications of their use have not been investigated. Presently available information simply does not permit a definite choice to be made among all the possible functions that might be appropriate. The decision was to accept a set of functions that were both qualitatively reasonable and analytically convenient, as given by equation 4.

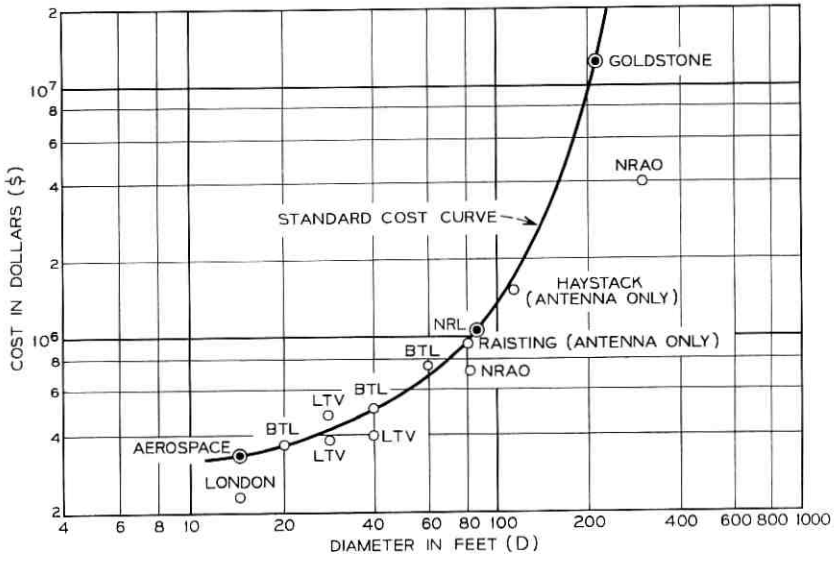


Fig. 4 — Revised standard cost vs diameter for existing antennas.

VII. MATHEMATICAL MODELS

Combining the equations developed above, we find the sets of equations which can be used to study the interrelation of gain, cost, diameter, surface tolerance and frequency, for both exposed and radome-enclosed antennas. For exposed antennas:

$$\epsilon = \frac{\alpha_1 D^{\frac{3}{2}}}{x}, \tag{5a}$$

$$\$ = \alpha_2 D^{-\frac{1}{2}} \exp(\alpha_3 D + x - 1), \tag{5b}$$

$$G = \eta(\alpha_4 D\Omega)^2 \exp -(\alpha_5 \epsilon\Omega)^2. \tag{5c}$$

Here, ϵ is the rms surface tolerance in millimeters, D the antenna diameter in feet, G the gain in absolute units, \$ the cost in dollars, η the aperture efficiency, and Ω the frequency in GHz, corresponding to a wavelength λ. Appropriate values for the constants are:

$$\alpha_1 = 1.3 \times 10^{-3}; \quad \alpha_2 = 6.7 \times 10^{5}; \quad \alpha_3 = 2.22 \times 10^{-2};$$
$$\alpha_4 = 3.20; \quad \alpha_5 = 4.19 \times 10^{-2}; \quad \eta = 0.70.$$

Expressions 5 are appropriate for a diameter range of approximately 15 to 250 feet. For antennas with a radome:

$$\epsilon = \frac{\beta_1 D^{\frac{3}{2}}}{x}, \tag{6a}$$

$$\$ = \exp(x - 1)[\beta_2 D^{\beta_3} - \beta_4 D^{\beta_5}] + \beta_4 D^{\beta_5} \tag{6b}$$

$$G = R(\lambda)[\eta(\beta_6 D\Omega)^2 \exp -(\beta_7 \epsilon\Omega)^2] \tag{6c}$$

The terms have the same meaning as for exposed antennas. In the present model, the cost factor is applied only to the cost of the antenna. The total cost, however, includes the cost of the radome. The expression for gain has been modified by the factor $R(\lambda)$, which accounts for losses resulting from the radome. The system effect of the noise temperature contribution from the radome is not considered. Appropriate values for the constants in equation 6 are:

$$\beta_1 = 4.6 \times 10^{-4}; \quad \beta_2 = 6.75 \times 10^{3}; \quad \beta_3 = 1.30;$$
$$\beta_6 = 3.20; \quad \beta_7 = 4.19 \times 10^{-2}; \quad \eta = 0.70.$$

For rigid radomes: $\beta_4 = 1.28 \times 10^{2}$; $\beta_5 = 1.85$

For air-supported radomes: $\beta_4 = 1.69 \times 10^{2}$; $\beta_5 = 1.65$.

Expressions 6 are appropriate for a diameter range of approximately 30 to 500 feet.

VIII. EXAMPLES

Much information can be gleaned from the models expressed by equations 5 and 6. Each set is comprised of two expressions among five variables. (The parameter x can easily be eliminated between the first two equations of each set.) Thus if any three are specified the other two can be found directly. There are so many possible combinations that no general solution curves can be given. It is simpler to enter the appropriate equations for each specific case and work out the result.

Of more interest are the various optimization studies that can be carried out. We give the results of three specific studies here. These are obviously not the only such studies that could be done. The details of the necessary algebraic manipulations are omitted, since they are generally straightforward but often tedious. However, in each case the procedure is indicated. The discussion is phrased in terms of equation 5 for exposed antennas. However, the procedure described is also appropriate for equation 6, with obvious changes.

For all examples which include a radome, a rigid radome is assumed. In addition the attenuation factor, $R(\lambda)$, is set equal to 0.793. This corresponds to the assumption of a 1 dB loss caused by the radome, independent of frequency. While this is probably a reasonably good number to use for low frequencies, it is certainly an oversimplication for the higher frequencies in the range of interest. Results displayed as a function of frequency for antennas with radomes incorporate the implicit assumption that the radome is designed to match the operating frequency at each point. In other words, the results do not indicate the performance of a specific system as frequency varies, but imply that the radome design also varies to match the operating frequency.

8.1 *Example 1: Maximum Cost-Effective Antenna*

A maximum cost-effective antenna is defined as one which provides the most gain per dollar. By eliminating either the parameter x or the rms surface tolerance ϵ in the appropriate equations 5, both the cost and the gain can be expressed as functions of the diameter and the remaining variable, ϵ or x. The ratio $G/\$$ is formed, and maxima of this expression, considered as a function of two variables, are sought using standard techniques. $G/\$$ exhibits a single maximum in the frequency range of interest, and the location of the maximum depends on the frequency, as expected. The results of this example are

plotted in Figs. 5 and 6 for exposed and radome-enclosed antennas, respectively. The ordinates represent the diameter and the cost, plotted against a common abscissa, frequency. The gain at the point of maximum cost-effectiveness is indicated on the diameter curve.

There is a much greater variation of gain with frequency for exposed antennas than for antennas with radomes. This is directly attributable to the diameters involved. There is relatively little variation in cost for exposed antennas over the entire frequency range of interest. Such antennas cost about $500,000 regardless of the operating frequency. It is also notable that the maximum cost-effective antennas found in this example all operate well below their gain-limit point.

8.2 *Example 2: Minimum Cost Antenna for a Specified Gain and Frequency*

In this example, the operating frequency and the gain are specified, perhaps as a consequence of other system constraints. The first step

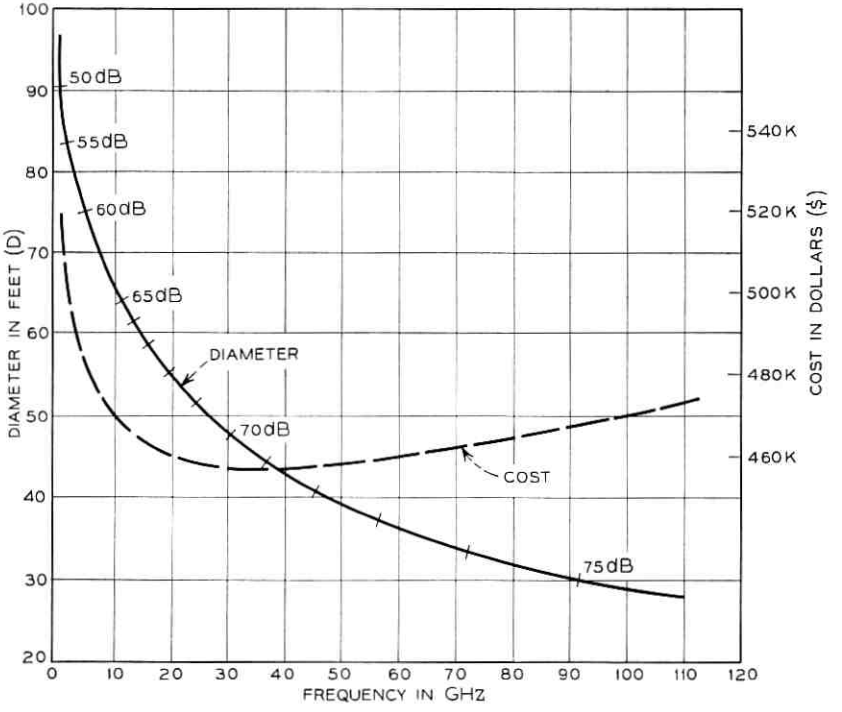


Fig. 5 — Cost and diameter of exposed maximum cost-effective antennas vs frequency (Example 1).

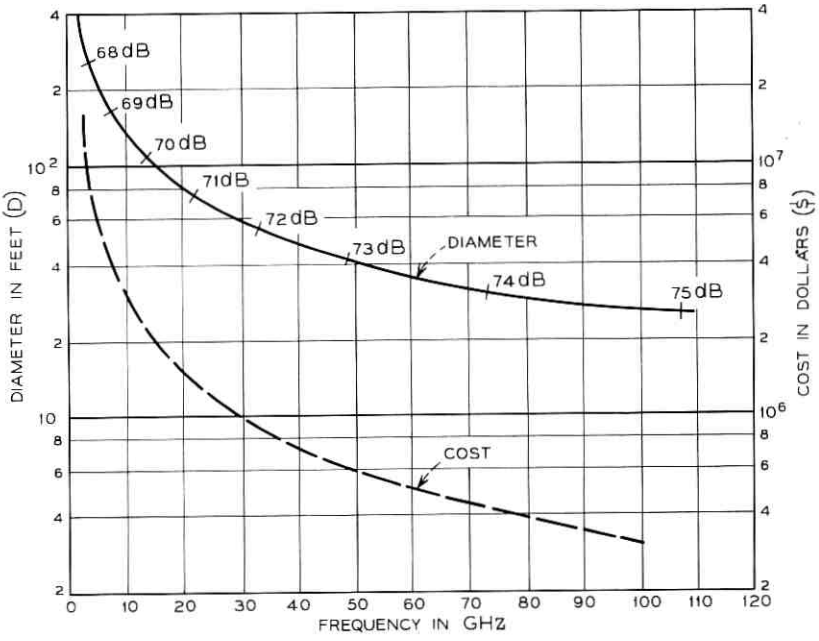


Fig. 6 — Cost and diameter of radome enclosed maximum cost-effective antennas vs frequency (Example 1).

is to substitute equation 5a into equation 5c with G and Ω specified. The resulting expression can then be solved for the parameter $x(D)$. This expression for x is inserted in equation 5b, yielding cost as a function of diameter. The diameter which minimizes this cost is then found by differentiation. The algebra involved in this example is unpleasantly heavy and numerical search techniques were used to determine the minimum cost diameter for both the exposed and the radome-enclosed cases. These results appear in Figs. 7 and 8, respectively. Figure 7a shows diameter vs frequency, and 7b gives cost vs frequency, for several different values of gain. The same pattern is followed in Fig. 8.

For the radome-enclosed antennas, the results of this example are represented by straight lines in logarithmic coordinates. The diameter vs frequency relations for exposed antennas are also straight lines in logarithmic coordinates at the lower frequencies, but exhibit a definite curvature at higher frequencies, particularly for the lower gains. The cost vs. frequency curves for exposed antennas illustrate the excep-

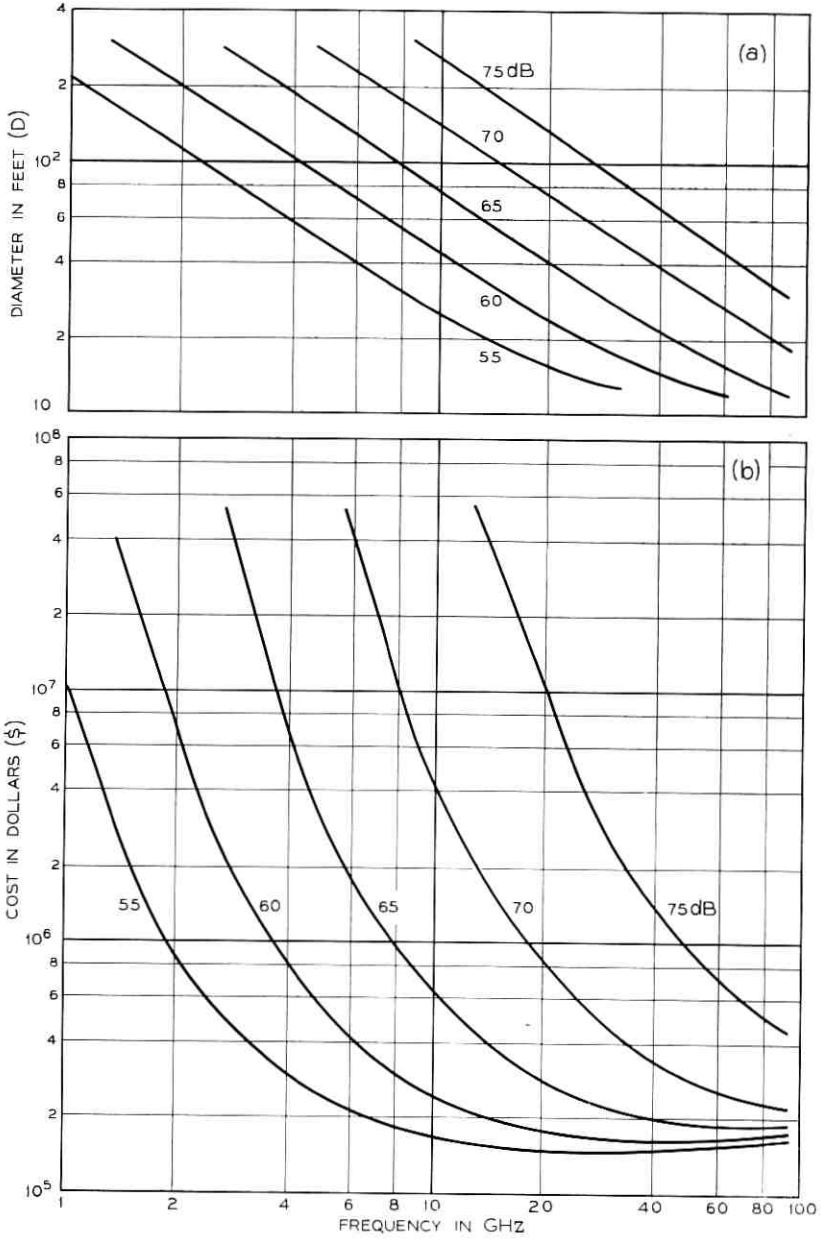


Fig. 7 — (a) Diameter and (b) cost of minimum cost exposed antennas for fixed gain and frequency vs frequency (Example 2).

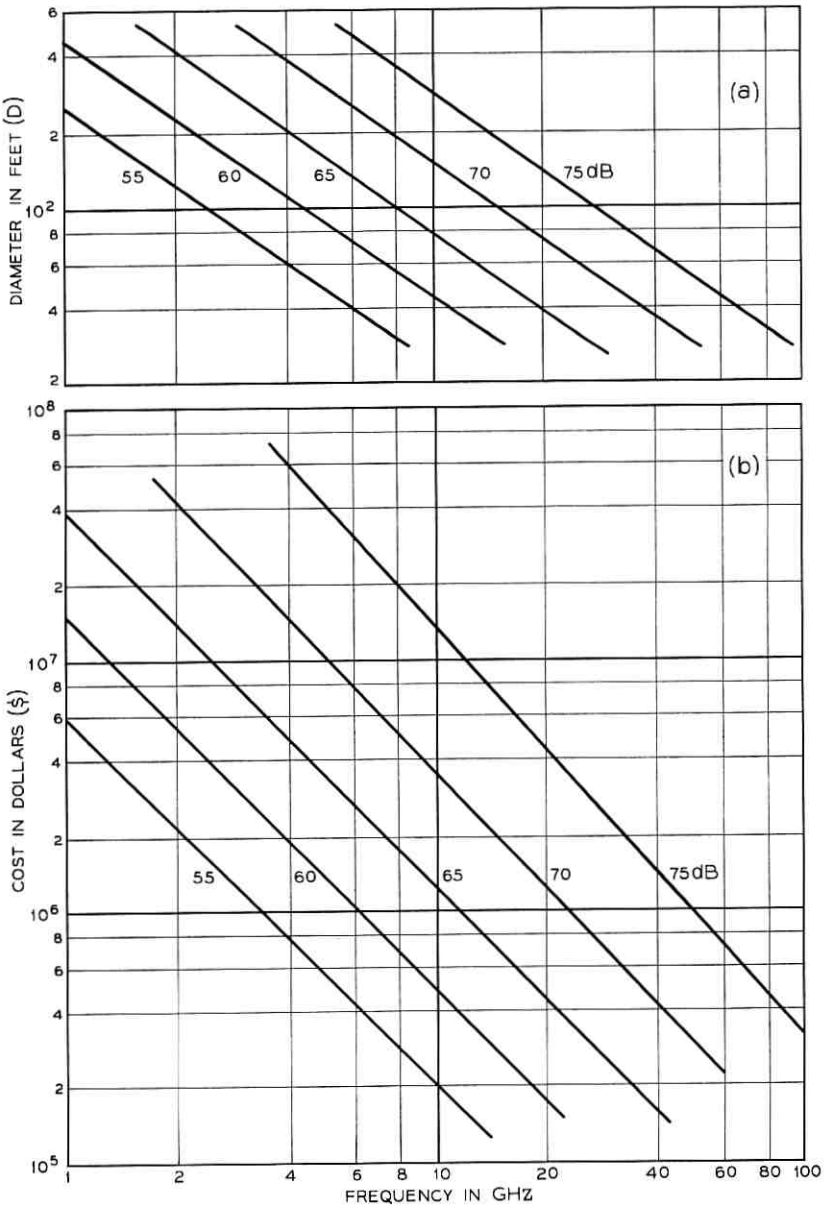


Fig. 8 — (a) Diameter and (b) cost of minimum cost radome enclosed antennas for fixed gain and frequency vs frequency (Example 2).

tionally high cost of gain at low frequencies because of the large diameter antennas required.

An interesting comparison can be made between the costs of equivalent exposed and enclosed systems by using Figs. 7b and 8b. For a specified gain, there is a range of diameters (or frequencies) within which an exposed antenna is less expensive than one enclosed in a radome. This range varies with gain. For other diameters (or frequencies) the radome enclosed system is a better buy for a specified performance.

8.3 *Example 3: Gain as a Function of Diameter for Fixed Cost and Frequency*

In this example, we do not consider the optimization possibilities directly, but we are interested in the gain as a function of diameter when the cost and the frequency are specified. For a given cost, x can be expressed in terms of the diameter according to equation 5b. In solving the model this way, we discovered that it was possible for x to be negative for certain combinations of cost and diameter. To avoid such a meaningless outcome, the restriction $x \geqq 0.1$ was imposed in this example. This is equivalent to restricting attention to antennas with rms surface tolerance no worse than 10 times the standard value. With $x(D)$ determined, equation 5a establishes the rms surface tolerance and the gain can be found from equation 5c with no difficulty.

Figure 9 shows the results of this exercise for two different fixed costs and several different frequencies. This figure shows the expected trends quite clearly. Notice that the optimum or gain-limit point for each set of conditions can be identified readily from the figure. This point could have been found directly, of course, by a procedure similar to that used in Example 2.

This example was not carried out for antennas with a radome.

IX. SUMMARY

Mathematical models relating cost, diameter, gain, rms surface tolerance and frequency have been developed for both exposed antennas and antennas with a radome. The form of the model in each case is a set of two equations among the five variables of interest. This model is much more complicated than others that have been suggested to relate the cost and diameter of ground antennas. However, it is also considerably more general and can be used to study a variety of possible trade-off situations. The major features of these models are:

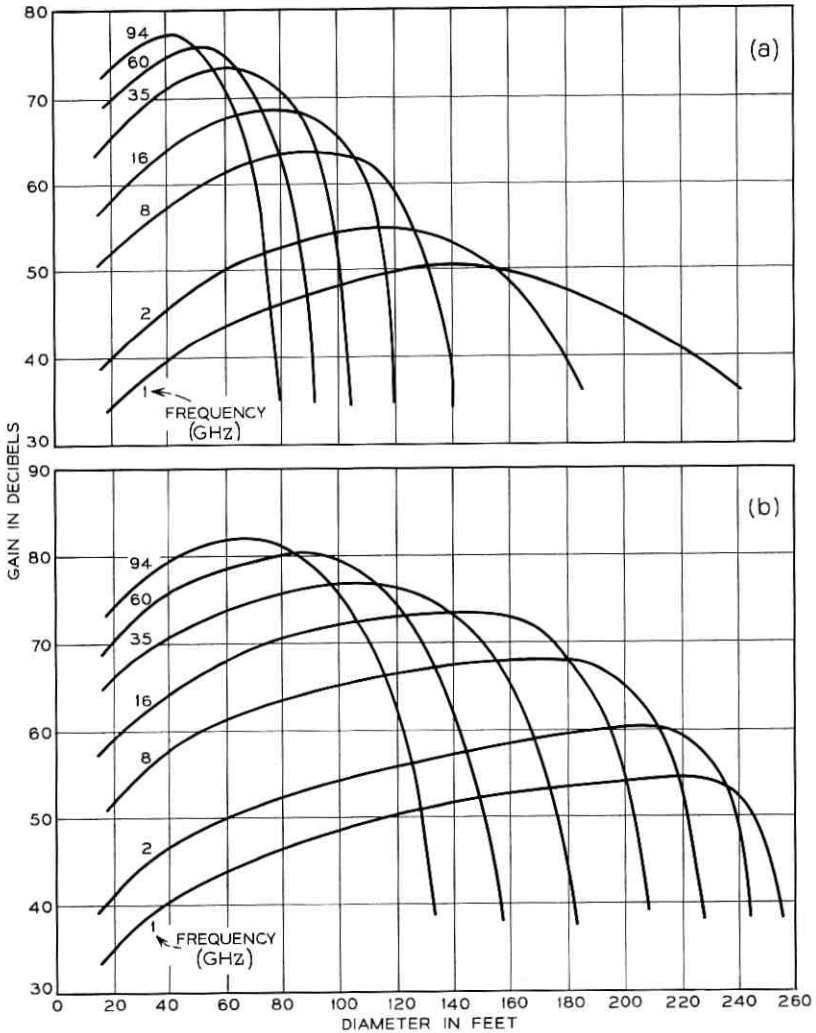


Fig. 9 — Gain vs diameter at several different frequencies with a fixed cost of (a) $1,000,000, and (b) $10,000,000 (Example 3).

(*i*) The inclusion of an exponential factor in the cost vs diameter relation for exposed antennas. This reflects, at least qualitatively, the exceptionally high cost associated with large, high performance exposed antennas.

(*ii*) The specific correlation of the cost vs diameter and rms sur-

face tolerance vs diameter relations. As a result, costs are associated not only with diameter, but also with rms surface tolerance.

(*iii*) The introduction of the quality factors. These factors relate deviations from the standard rms surface tolerance to expected departures from the standard cost curve. Although qualitative in nature, these factors reflect acknowledged trends. Skeptical readers who cannot accept the form of the quality factors used in the present models are encouraged to supply their own.

The two equations comprising each model can be supplemented by additional relations among the variables such as an rms surface tolerance-wavelength relation, for example. There is virtually no limit to the kinds of optimization studies and trade-off investigations that can be carried out within the framework of the suggested models. Examples of three such studies have been included. Through the credibility of the results, these examples further demonstrate the qualitative validity of the models.

These models have proved valuable in the preliminary phases of communications system planning, where competing concepts can be compared in terms of the relatively gross features of the system. They are not intended to usurp the responsibilities of the antenna designer in any specific application, and it would be erroneous to extrapolate their utility to such levels of refinement. Minor revisions in the constants of these models, as a result of new information or even different interpretations of present data, are to be expected and encouraged. However, such refinements should not invalidate the general applicability of the present models nor the qualitative conclusions drawn from their use.

REFERENCES

1. Ruze, J., "Antenna Tolerance Theory—A Review," Proc. IEEE, *54*, No. 4 (April 1966), pp. 633–40.
2. Ruze, J., "The Effect of Aperture Error on the Antenna Radiation Pattern," Suppl. al Nuovo Cimento, Series IX; *9*, No. 3 (1952), pp. 364–80.
3. Stack, B. R., unpublished work.
4. Stack, B. R., "An Approximate Expression for the Cost-Gain Relationship in Large Parabolic Antennas," Stanford Research Institute, Menlo Park, California, December 1967.
5. Zucker, H., unpublished work.
6. The Cambridge Radio Observatory Committee, "A Large Radio-Radar Telescope—CAMROC Design Concepts, Vol. I and II," Cambridge, Massachusetts, January 15, 1967.
7. Jacobs, E. and King, H. E., "Large-Aperture Millimeter-Wave Antenna with High Pointing Accuracy," Proc. Conf. Design and Construction of Large

Steerable Aerials, IEE Conf. Publ. No. 21, London (June 1966), pp. 218–31.
8. McClain, E. F., "Highly Directive Antennas Used in NRL's Radio Astronomy Program, Part III. The 85-foot Radio Telescope," NRL Progress (August 1966), pp. 5–10.
9. Potter, P. D., "Design and Performance of the NASA/JPL 210′ Steerable Paraboloid," Proc. Conf. Design and Construction of Large Steerable Aerials, IEE Conf. Publ. No. 21, London (June 1966), pp. 378–98.
10. Potter, P. D., Merrick, W. D., and Ludwig, A. C., "Big Antenna Systems for Deep-Space Communications," Astronautics and Aeronautics, *4*, No. 10, (October 1966), pp. 84–95.
11. "Design of Earth Terminals for Satellite Communications," *2*, Section II, Antennas; Report to Communications Satellite Corporation prepared by Bell Telephone Laboratories, Whippany, New Jersey, April 1, 1965.

Radiating Properties of Dielectric Covered Apertures

By ELLIOTT R. NAGELBERG

(Manuscript received May 21, 1968)

This paper studies the effect of a dielectric sheath on the far-zone radiation characteristics of an aperture bounded by a perfectly conducting ground plane. An examination of the appropriate Green's function predicts a significant broadening of the radiation pattern in a narrow band centered about that frequency at which a surface wave would begin to propagate along a grounded dielectric slab with the same thickness and permittivity as the sheath. The differences between this phenomenon and what is commonly referred to as a surface wave are discussed. Experimental results are then presented for the broadside and end-fire (perpendicular to the aperture) radiation by a waveguide aperture, from which it is found that the theory does predict the essential character of the observations.

I. INTRODUCTION

In designing phased arrays for radar and other types of communication systems, it is generally necessary to provide some degree of environmental protection. One configuration which suggests itself is the flush-mounted "radome" (Fig. 1), made of a suitable refractory ceramic such as beryllia, alumina, or boron nitride. However, since these materials typically have relative dielectric constants greater than unity, such a covering sheath can be expected to influence the electrical performance of the antenna and hence of the system to which it belongs. The two antenna characteristics of greatest interest are: (*i*) the individual element input impedance and (*ii*) the radiation pattern corresponding to a given aperture illumination.

The effects of a covering sheath on the input characteristics of infinite rectangular waveguide arrays have been discussed by Galindo and Wu, using a technique of spectral analysis.[1] Although the radiation pattern of a single element in an array environment can also be obtained as a by-product of this analysis, it is still useful to study

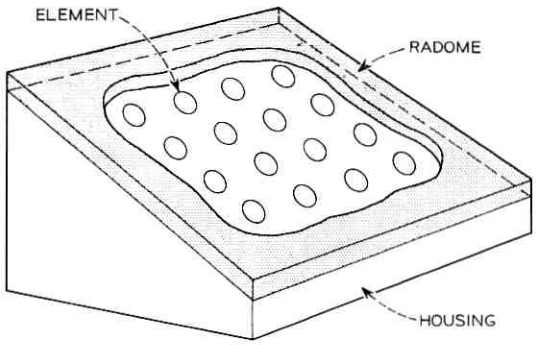


Fig. 1 — Segment of a phased array antenna with protective radome.

a single isolated element, particularly in order to distinguish characteristics which are associated with the element itself from those attributable to the periodicity of the structure. If the individual elements are identical, then the far-field radiation pattern of the total array can be represened as the product of the element pattern (in the array environment so that mutual coupling effects are included) and the array factor, which depends on the periodicity of the aperture and the overall amplitude and phase taper. The sheath does not alter this separability feature and can be regarded as affecting only the element pattern.

As an example, we have chosen to study the radiating properties of a single continuous aperture in a ground plane, covered by a dielectric slab of known thickness and permittivity. A study of the appropriate Green's function shows that the principal effect of the sheath is a significant broadening of the radiation pattern, over a narrow band centered about the frequency at which a surface wave would begin to propagate along the grounded dielectric sheath alone. (It follows that under matched conditions there must also be a dip in the broadside radiation, as required by energy conservation.) The result of this broadening is to produce a sharp increase in the power radiated near the end-fire direction. It should be pointed out, however, that despite the mathematical connection, this radiation differs from a "surface wave"[2] in two respects: (*i*) It belongs to the contiuous spectrum of the radiation field and thus should be regarded physically as a distortion of the far-zone pattern rather than a separate mode of propagation; it thus exhibits the characteristic inverse power law variation with distance from the aperture. (*ii*) The end-fire signal amplitude is roughly symmetrical about the surface wave cut-

off frequency and thus does not itself exhibit a cut-off characteristic.

First we outline the formal solution to the boundary value problem, which is carried out by a Fourier transform analysis. Then we obtain the far-zone field by a saddle point approximation to the inversion integral, from which the broadening is clearly shown by direct evaluation. Finally, we discuss the associated experimental work and present results which serve to corroborate the theoretical analysis. We use rationalized MKS units and the (suppressed) harmonic time dependence $e^{-i\omega t}$ throughout

There is a mathematical question regarding the validity of the saddle point approximation under the conditions $\theta \simeq \pi/2$ and $k_0 a = (k_0 a)_n$. For this combination of parameters, the denominator of the function $F(\alpha)$ in equation (11) vanishes at the saddle point. Notice that the presence of the cos α factor causes the integrand to be finite under these circumstances so that the procedure is still valid. Another point of view is that when the saddle point and "pole" coalesce, the residue vanishes and with it the necessary correction term to the steepest descent formula (see Ref. 3, p. 503).

II. THEORETICAL ANALYSIS

The problem is to determine the far-zone radiating properties of an aperture covered by a dielectric sheath, as shown in Fig. 2. It is assumed that the electric field is in the y direction and that all quantities are independent of y. In addition to its mathematical simplicity,

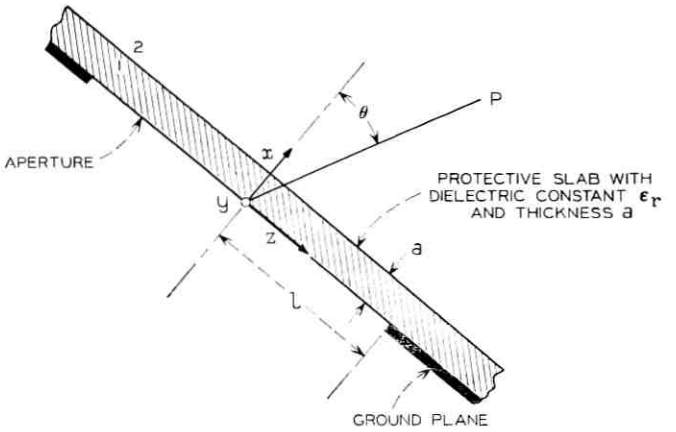


Fig. 2 — Aperture of width $2l$ bounded by ground plane and covered by protective slab of thickness a and dielectric constant ϵ_r.

off frequency and thus does not itself exhibit a cut-off characteristic.

First we outline the formal solution to the boundary value problem, which is carried out by a Fourier transform analysis. Then we obtain the far-zone field by a saddle point approximation to the inversion integral, from which the broadening is clearly shown by direct evaluation. Finally, we discuss the associated experimental work and present results which serve to corroborate the theoretical analysis. We use rationalized MKS units and the (suppressed) harmonic time dependence $e^{-i\omega t}$ throughout

There is a mathematical question regarding the validity of the saddle point approximation under the conditions $\theta \simeq \pi/2$ and $k_0 a = (k_0 a)_n$. For this combination of parameters, the denominator of the function $F(\alpha)$ in equation (11) vanishes at the saddle point. Notice that the presence of the cos α factor causes the integrand to be finite under these circumstances so that the procedure is still valid. Another point of view is that when the saddle point and "pole" coalesce, the residue vanishes and with it the necessary correction term to the steepest descent formula (see Ref. 3, p. 503).

II. THEORETICAL ANALYSIS

The problem is to determine the far-zone radiating properties of an aperture covered by a dielectric sheath, as shown in Fig. 2. It is assumed that the electric field is in the y direction and that all quantities are independent of y. In addition to its mathematical simplicity,

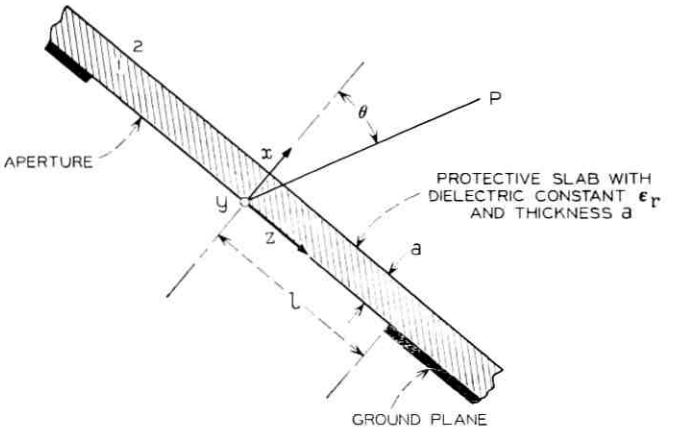


Fig. 2 — Aperture of width $2l$ bounded by ground plane and covered by protective slab of thickness a and dielectric constant ϵ_r.

this two-dimensional situation can be realized experimentally by erecting parallel conducting planes perpendicular to the electric field.

Using a plane wave (Fourier) spectrum representation of the fields, E_y is given in terms of the transform pair

$$E_y(x, z) = \frac{1}{2\pi} \int_{h=-\infty}^{+\infty} G(h, z) e^{ihz}\, dh \tag{1}$$

$$G(h, x) = \int_{z=-\infty}^{+\infty} E_y(x, z) e^{-ihz}\, dz \tag{2}$$

where h, the transform variable, denotes the z-directed propagation constant of the particular plane wave component.

The function $G(h, x)$ thus satisfies the one-dimensional wave equation

$$\frac{d^2G(h, x)}{dx^2} + \beta^2 G(h, x) = 0 \tag{3}$$

where

$$\beta = (k^2 - h^2)^{\frac{1}{2}} \text{ inside the sheath}$$
$$\beta = \beta_0 = (k_0^2 - h^2)^{\frac{1}{2}} \text{ outside the sheath,} \tag{4}$$

k or k_0 being the respective wave number. The proper Riemann surface, shown in Fig. 3, defines the square root such that

$$\beta = +(k^2 - h^2)^{\frac{1}{2}} \; h \text{ real}, \quad |h| < k \tag{5a}$$

$$\beta = +i(h^2 - k^2)^{\frac{1}{2}} \; h \text{ real}, \quad |h| > k. \tag{5b}$$

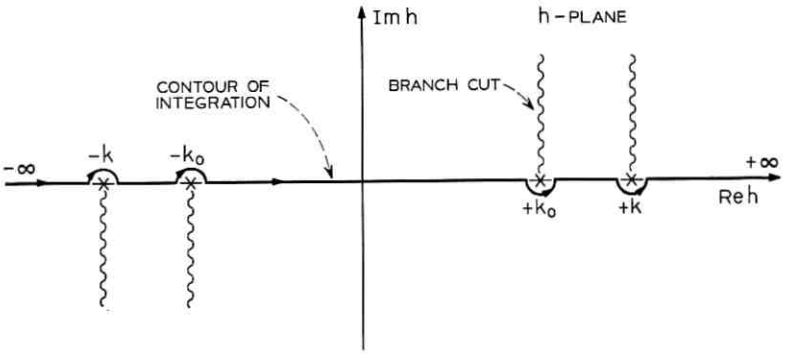


Fig. 3 — Definition of proper Riemann surface for evaluation of $\beta_0 = (k_0^2 - h^2)^{1/2}$ and $\beta = (k^2 - h^2)^{1/2}$.

Conditions given by equations (5a) and (5b) are a result of the physical requirements that: (*i*) waves radiated far from the aperature travel in the $+x$ direction and (*ii*) slow waves propagating in the $\pm z$ direction decrease exponentially in amplitude with distance from the sheath.

The boundary conditions are that

$$G(h, 0) = G_0(h) \tag{6a}$$

$$G, \frac{\partial G}{\partial x} \text{ continuous at } x = a \tag{6b}$$

where $G_0(h)$ is the Fourier transform of the aperture illumination.

The principal interest is in the field in region 2, outside the sheath, whose transform function is denoted by

$$G_2(h, x) = A(h)e^{i\beta_0 x}. \tag{7a}$$

In a similar manner we represent the field in region 1, inside the sheath as,

$$G_1(h, x) = B(h)e^{i\beta x} + C(h)e^{-i\beta x} \tag{7b}$$

and by direct substitution into (6) we find that

$$A(h) = \frac{G_0(h)e^{-i\beta_0 a}}{\cos \beta a - i\dfrac{\beta_0}{\beta} \sin \beta a}. \tag{8}$$

This leads directly to the expression for E_y

$$E_y(x, z) = \frac{1}{2\pi} \int_{h=-\infty}^{+\infty} \frac{G_0(h)e^{-i\beta_0 a}}{\cos \beta a - i\dfrac{\beta_0}{\beta} \sin \beta a} e^{i\beta_0 x} e^{ihz}\, dh \tag{9}$$

The determination of the far-zone radiation pattern from integral representations of the form given in equation 9 generally proceeds by saddle point integration. We first transform into the polar coordinates r, θ shown in Fig. 2, which are related to x, z by

$$x = r \cos \theta \tag{10}$$

$$z = r \sin \theta.$$

Next, using the transformation $h = k_0 \sin \alpha$, we find that equation 9 can be written in the form

$$E_y(r, \theta) = \frac{1}{2\pi} \int_C F(\alpha) e^{ik_0 r \cos(\alpha - \theta)} \cos \alpha\, d\alpha \tag{11}$$

where the new contour is that given in Fig. 4. Notice also that in transforming from the h- to the α-plane, the integral is independent

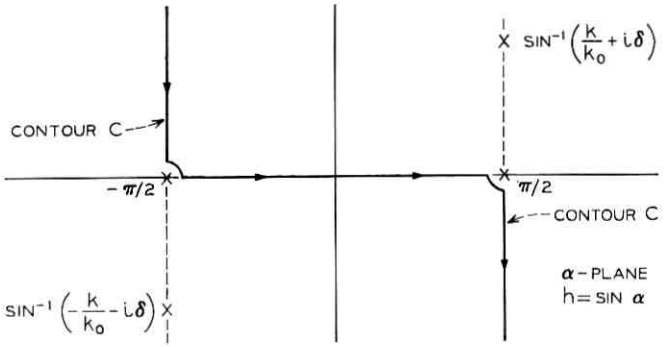


Fig. 4 — Transformation $h = \sin\alpha$ showing new contour.

of the sign of the square root and hence the branch cuts disappear. In order to compute the far-field we let $k_0 r \to \infty$ and use the saddle point method to determine an asymptotic formula for the integral.[3] For purposes of finding the radiation pattern only, it is sufficient to observe that the saddle point occurs at $\alpha = \theta$ and that the *angular variation* of the field can therefore be specified in terms of the (unnormalized) function

$$G(\theta) = G_0(k_0 \sin\theta) \cos\theta \cdot T(\theta) \tag{12}$$

where $T(\theta)$ is the sheath transmission pattern given by

$$T(\theta) = \frac{e^{-ik_0 a \cos\theta}}{\cos\left[k_0 a(\epsilon_r - \sin^2\theta)^{\frac{1}{2}}\right] - \dfrac{i\cos\theta}{(\epsilon_r - \sin^2\theta)^{\frac{1}{2}}} \sin\left[k_0 a(\epsilon_r - \sin^2\theta)^{\frac{1}{2}}\right]} \tag{13}$$

$T(\theta)$ is, of course, dependent on θ and can therefore be expected to alter, at least to some extent, the radiation pattern of the aperture for all angles. However, the most significant changes occur at angles near $\theta \simeq \pi/2$, and at frequencies in the vicinity of cut off for a TE_n surface wave on a dielectric slab covering a ground plane.[2] These waves begin to propagate for respective values of $k_0 a$ given by

$$(k_0 a)_n = \frac{(2n-1)}{(\epsilon_r - 1)^{\frac{1}{2}}} \frac{\pi}{2} \tag{14}$$

Under the conditions $\theta = \pi/2$ and $k_0 a = (k_0 a)_n$, the denominator of equation 13 vanishes, and although this behavior must be weighted by the fact that in equation 12, for $\theta \to \pi/2$, $\cos\theta \to 0$, the overall

effect is still to produce a broadening of the pattern and therefore a resonant component of radiation near the end-fire direction. Typical curves for $G(\theta)$ illustrating this effect are shown in Fig. 5 for values of k_0a centered about $(k_0a)_1$, with $G_0(k_0\sin\theta) \equiv 1$, $\epsilon_r = 6.0$. The results therefore pertain to the appropriate Green's function. With regard to interpretation of these curves, it should be noted that, strictly speaking, the results cannot be applied at angles arbitrarily close to 90°. This is because eventually one reaches the outer boundary of the dielectric sheath in which a different solution must be used. The curves have been drawn up to 90° under the assumption that in the far field the sheath occupies a negligibly small angular sector.

While the distortion of the radiation pattern has the physical appearance of a surface wave, in the sense that its energy is most significant near the dielectric interface, it is actually quite different in several respects. First, the radiation discussed here belongs to the continuous, rather than to the discrete spectrum of the aperture radiation field. Its amplitude thus decays inversely with distance from the

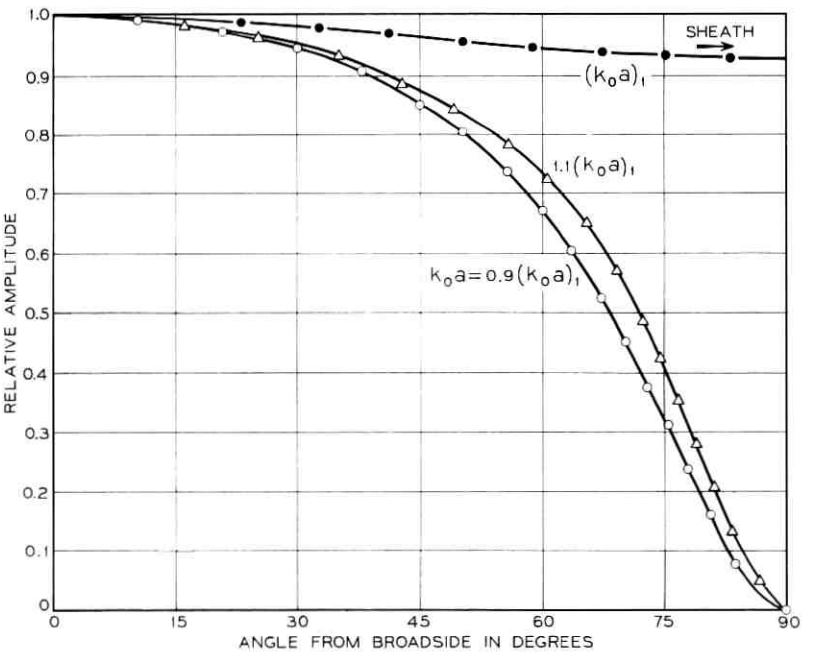


Fig. 5 — $G(\theta)$ (normalized to unity at $\theta = 0°$) for values of $k_0\alpha$ near $(k_0a)_1$, $G_0(k_0 \sin\theta) \equiv 1$, $\epsilon_r = 6.0$.

aperture, rather than being independent of distance, as for an unattenuated mode guided by the slab-ground combination. Furthermore, the end-fire radiation does not exhibit a cutoff characteristic but instead has essentially a symmetrical amplitude variation about the frequencies given by equation 14.

The surface wave modes excited by the aperture are, on the other hand, associated with the discrete spectrum of the radiation field, and are derived mathematically from the residues of the inversion integral with respect to its poles. These poles correspond to solutions of the equation

$$\cos \beta a + \frac{\alpha_0}{\beta} \sin \beta a = 0 \tag{15}$$

where $\alpha_0 = (h_0^2 - k_0^2)^{\frac{1}{2}}$, h_0 being the surface wave propagation constant. There is no detailed discussion of surface wave excitation here because not enough information is yet available to accurately estimate the appropriate excitation coefficient. This observation is made, to some extent, *a posteriori*, from the fact that the "reasonable" assumption of an unperturbed TE_{10} waveguide mode seems to yield a surface wave component considerably higher than we have observed experimentally.[4] For example, theoretical calculations reported in the literature lead to the conclusion that the surface wave extracts an much as 50 percent of the total radiated power over a wide range of frequencies. If this were generally true, the effects would be visible not only as a substantial increase in end-fire radiation but as a corresponding loss in the broadside direction. Section III shows that neither of these effects was observed in the expected manner.

A possible explanation for this discrepancy can be found in the fact that the surface wave excitation coefficient is proportional to $G_0(h_0)$, the Fourier transform of the aperture field distribution, evaluated for the surface wave propagation constant h_0. Since we are interested in dielectric constants greater than unity, $h_0 \geq k_0$, and it is therefore appropriate to cast $G_0(h_0)$ in the form (assuming E_0 is symmetric with respect to z),

$$G_0(h_0) = \sum_{m=0}^{\infty} C_m \frac{E_0^{(m)}(z = l)}{(h_0/k_0)^{m+1}} \tag{16}$$

where the C_m are constants and $E_0^{(m)}(z = l)$ represents the mth derivatime of the aperture field, evaluated at the edge.[5] Equation (16) is different from the analogous formula for the radiation field, which would be a series in direct powers of h/k_0 in which the coefficients would

be weighted integrals of the aperture field. This dependence of surface wave excitation on conditions at the edge of the aperture makes the problem very difficult to analyze, expecially since measurements have shown the field in this region to change significantly when a sheath is present.[4] A physical interpretation would be that surface wave excitation depends on diffraction rather than direct radiation.

III. EXPERIMENTAL RESULTS

The apparatus, illustrated in Fig. 6, consists of a network feeding an open X-band (WR-90) waveguide flanged to a ground plane. So as to simulate the two-dimensional situation analyzed in the previous section, parallel conducting walls in the H-plane were used, as shown in Fig. 7. With this configuration, no additional field components are induced (at least theoretically), and the fields are essentially independent of position in the E-plane. The ground plane and aperture were covered by a uniform slab of stycast material with dielectric constant of 6.0 and thickness such that the $(TE)_1$ mode cut-off frequency was 10.0 GHz.

The measurements determined the individual signals received by the end-fire and broadside detectors, which consisted of open-ended waveguides terminated with broadband matched crystals. These were

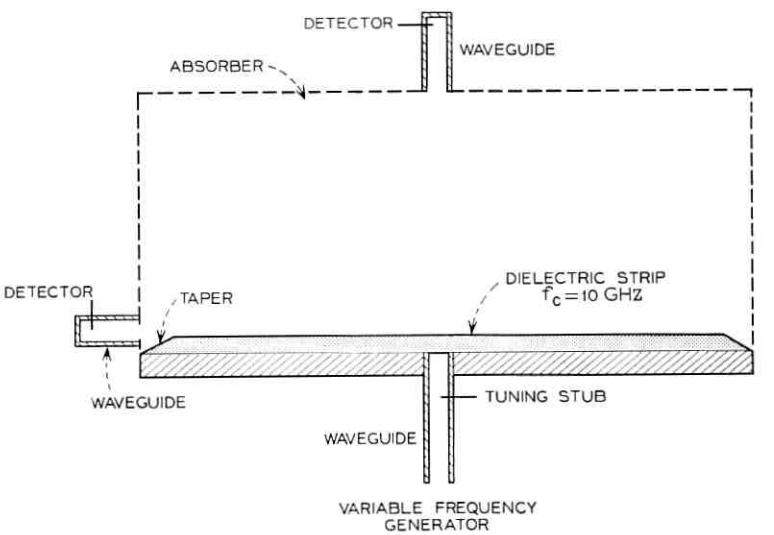


Fig. 6 — Apparatus for measuring broadside and end-fire radiation.

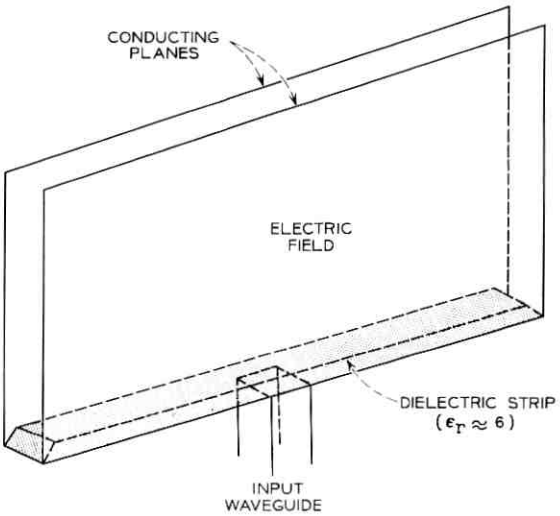


Fig. 7 — Use of parallel conducting planes to simulate two-dimensional geometry.

compared with the power incident on the aperture by a ratio meter having one terminal connected to the input waveguide via a directional coupler. At each frequency the tuning stub was adjusted to keep the return loss below −30 dB; the results therefore pertain to a matched element.

In order to avoid mutual coupling between waveguide detectors, respective measurements of broadside and end-fire signals were made with the other detector removed. Since the waveguides were mounted in brackets which could be rigidly connected to the conducting planes, repeatability was not considered to be a serious problem.

Figure 8 shows the experimental results, which demonstrate the resonant end-fire radiation phenomenon. The peak of the end-fire signal occurs at 10.0 GHz, as expected, and the general shape of the curve, down to approximately 6 dB below the peak value (normalized to 0 dB maximum), accurately follows the theoretical behavior. The latter was determined by calculating $\cos \theta T(\theta)$ for a value of $\theta = 88°$. In this regard, it was found by direct calculation that the shape of the relative variation with frequency is not sensitive to small deviations from 90°. The result given in Fig. 8 should therefore be a reasonable estimate for the behavior of the waveguide aperture receiver. Notice that 0 dB is defined as the measured end-fire signal

at 10 GHz and that the theoretical curve is also normalized at this frequency. Thus no attempt has been made to arrive at an absolute corroboration, which would require both the solution inside the dielectric slab and a detailed description of radiation at the end taper. The deviations in the end-fire radiation in Fig. 8 below 6 dB probably result from these latter aspects.

Under matched conditions, conservation of energy requires that an increase in radiation in any direction must be accompanied by a corresponding reduction in other directions. This is clearly evidenced by the decrease in broadside radiation in the vicinity of 10.0 GHz. The reduction amounts to approximately 3–4 dB, which is quite significant, perhaps even more so in terms of antenna performance than the increase in the end-fire signal. The displacement of the minimal frequency to the slightly lower value of 9.7 GHz is, in all likelihood, caused by the generally increasing nature of the broadside signal before the dip occurs, attributed to the slow increasing gain

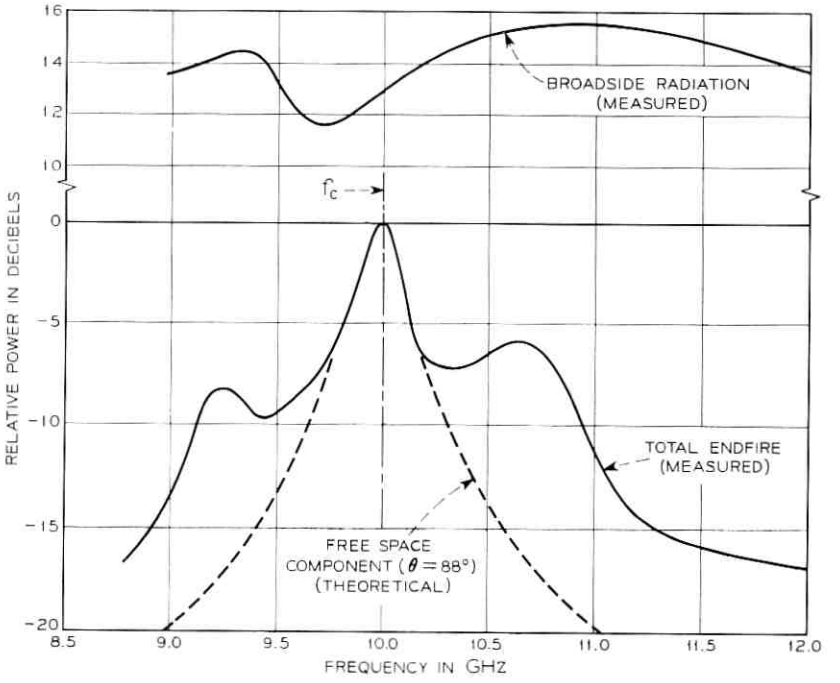


Fig. 8 — Observed end-fire and broadside radiation under matched conditions.

of the receiving aperture with frequency. This would tend to skew the curve slightly to the left.*

In interpreting these experimental results, it is worth noting that strong coupling to a surface wave in this experiment would be seen in two ways. First, the end-fire radiation would exhibit a cut-off characteristic in which the amplitude would be relatively low below the cut-off frequency, in this case 10.0 gHz, at which point it would increase to a higher value. Second, at approximately the same frequency the broadside radiation would correspondingly decrease. Since the experimental results do not behave in this way, but rather in a manner predictable by a far field analysis, it is concluded that strong coupling to the surface wave component did not occur for this particular set of parameters.

IV. SUMMARY AND CONCLUSIONS

The purpose of this study has been to examine, both theoretically and experimentally, the radiating characteristics of an aperture-type antenna covered by a dielectric sheath. Such an antenna might represent, for example, an element of a phased array which has been protected against a high temperature environment.

The first principal effect of the sheath on the pattern of a matched element is to introduce a broadening of the radiation pattern and a resulting component of radiation in the end-fire direction at frequencies near the cut-off frequency of a surface wave on the dielectric slab. However, it is important to distinguish this phenomenon, which is properly viewed as a distortion of the far-zone radiation pattern, from what is commonly referred to as a surface wave, which is a separate mode of propagation belonging to the discrete spectrum of the radiation field. No effects attributable to the latter were observed in the experiment, showing that the excitation was negligibly small over the measured frequency range.

Under matched conditions, there must be a corresponding dip in the broadside radiation over the same frequency range. This requirement follows directly from energy conservation.

It is concluded that a matched phased array element is not sufficient to guarantee that the antenna is functioning as required. Dielectrics can have a pronounced effect on the directivity and efficiency

* Recent computational studies by C. P. Wu corroborate this slight frequency shift of the dip in broadside radiation.

of aperture radiators, especially near frequencies which are characteristic of surface wave propagation along some part of the structure.

V. ACKNOWLEDGMENT

The author would like to thank Mr. P. J. Puglis, who actively contributed to all aspects of this work. Conversations with Dr. C. P. Wu were also most helpful.

REFERENCES

1. Galindo, V., and Wu, C. P., "Dielectric Loaded and Covered Rectangular Waveguide Phased Arrays," B.S.T.J., *47*, No. 1 (January 1968), pp. 93–116.
2. Collin, R. E., *Field Theory of Guided Waves*, New York: McGraw-Hill, 1960, p. 470.
3. Mathews, J., and Walker, R. L., *Mathematical Methods of Physics*, New York: W. A. Benjamin, 1964, p. 78.
4. Croswell, W. F., Rudduck, R. C., and Hatcher, D. M., "The Admittance of a Rectangular Waveguide Radiating into a Dielectric Slab," IEEE Trans. Antennas and Propagation, *AP-15* (September 1967), pp. 627–633.
5. Erdélyi, A., *Asymptotic Expansions*, New York: Dover Publications, 1965, p. 46.

A New Reference Frequency Standard for the L Multiplex System

By W. A. KESTER

(Manuscript received July 17, 1968)

Four precision quartz crystal oscillators form the heart of a new solid-state precision frequency standard, designed to meet the requirements of the present and next generation of L multiplex carrier equipment. The output frequencies of 64 and 512 kHz are derived from the 4.096-MHz oscillator outputs by binary frequency dividers. A master and standby divider chain form a redundant path linking two of the oscillators and the passive output distribution circuits. Intentional transfers between master and standby channels can be made without introducing significant phase hits at the output. Automatic transfer between channels occurs upon a catastrophic loss of the output signal. Two additional oscillators and dividers are provided as "hot spares" which can be manually patched into service. Critical points in the system are monitored by major and minor alarm circuits, and a very low frequency receiver-comparator provides a means for maintaining the oscillator frequencies to within one part in 10^9 of a vlf standard frequency broadcast.

I. INTRODUCTION

The L multiplex system is a single-sideband suppressed carrier system in which the carrier frequency supplies in the various offices are synchronized by the transmission of pilot frequencies over a tree-like network. At each office the carrier frequencies necessary to perform the modulation and demodulation steps are derived from the outputs of an "office master" supply which is phase-locked to the incoming synchronizing pilot tone. This "office master" supply, named the primary frequency supply in L multiplex terminology, has been described in previous literature.[1]

The tree-like synchronization network originates at a single primary frequency supply called the "system master." While relative frequency accuracy between offices depends on the phase-locked

synchronization network, the absolute frequency accuracy of the pilot tones is dependent upon the absolute frequency accuracy of the system master. Although the free-running accuracy of a primary frequency supply is such that fairly appreciable periods of pilot outage can be tolerated, a highly accurate reference frequency standard is still required to ultimately control the system master.

In the past this frequency standardization service has been provided by Bell Telephone Laboratories, Murray Hill, New Jersey, where the Bell System Primary Standard of Frequency has been maintained to an absolute frequency accuracy of one part in 10^9 by periodic corrections. In recent years the need has arisen to replace the Murray Hill frequency standard with a more reliable and rugged solid-state version which would be located and maintained in a hardened telephone office. The new Bell System Reference Frequency Standard described in this article has been designed to meet these immediate requirements of the present L multiplex system, including recent requirements of the new L-4 coaxial system. In addition, system objectives and requirements for the new standard have been selected in order to anticipate the needs of at least the next generation of multiplex equipment after L-4.

II. SYSTEM OBJECTIVES AND REQUIREMENTS

2.1 *Frequency Accuracy*

The absolute frequency accuracy requirement for the reference frequency standard has been established as one part in 10^9. This requirement meets the needs of the present L multiplex system (including L-4) and should be adequate for any future multiplex systems which may be developed over the useful life of the new standard.

A survey of recent developments in the field of frequency standards and frequency measurement techniques reveals that the above requirement can be met economically and reliably by the use of a solid-state double-oven quartz crystal oscillator as the frequency source and the very low frequency broadcasts of the National Bureau of Standards as a comparison for periodic corrections of the oscillator frequency. The technique of measuring frequency offsets using a vlf receiver-comparator is relatively simple, and accuracies of a few parts in 10^{10} can be obtained in a measuring interval of several hours (see Section IV).

The accumulated oscillator frequency drift during the time between corrections (maintenance interval) is equal to the oscillator drift

rate multiplied by the maintenance interval. This implies that for a minimum maintenance interval of one month, the oscillator drift rate must be less than approximately 5 parts in 10^{10} per week in order to maintain the accuracy requirement of one part in 10^{9}. This assumes that at the time of maintenance the oscillator is corrected to the lower limit if the drift rate is positive and to the upper limit if the drift rate is negative.

2.2 *System Reliability*

The overall reliability objective for the new standard is that it be able to provide continuous and accurate output signals to the L multiplex system. Although short periods of outage can be tolerated, the new standard has been designed to minimize their probability and duration. These reliability objectives have been met in the following manner:

(*i*) Solid-state circuits are used wherever possible.

(*ii*) Redundant circuits have been provided which are automatically switched into service in the event of catastrophic failures. Other redundant circuits can be manually patched into service in case of double failures or failures in the automatic switching circuits themselves.

(*iii*) Special circuits have been designed which allow routine maintenance checks of the redundant switching equipment to be performed without causing either interruption of service or phase perturbations in the L multiplex synchronization network.

(*iv*) Critical areas in the system are monitored by major and minor alarm circuits. Visual and audible indications of trouble conditions are provided.

(*v*) The equipment comprising the system has been designed to facilitate ease of replacement and repair.

2.3 *L Multiplex Interface*

The interface betweeen the reference frequency standard and the L multiplex synchronization network is the primary frequency supply which has been designated "system master." The requirements of the primary frequency supply thus determine the following requirements of the new standard:

(*i*) Output frequencies should be 64 and 512 kHz to meet the requirements of two existing versions of the primary frequency supply.

(*ii*) The 64- and 512-kHz output signals should be square waves

at an approximate level of −23 dBm (fundamental component) and an impedance of 135 ohms, balanced.

(*iii*) Phase hits introduced at the output because of intentional transfers between master and standby channels should be less than 20 nanoseconds. This requirement allows maintenance transfers to be performed without significantly perturbing the phase-lock circuits in the primary frequency supplies.

III. SYSTEM DESCRIPTION

3.1 *General*

Figure 1 is a block diagram of the circuits comprising the primary signal paths in the reference frequency standard. The output frequency of the precision oscillators is 4.096 MHz, hence binary frequency dividers are used in deriving the output signals of 512 and 64 kHz. There are four precision oscillators (1, 2, 3, and 4) and four frequency dividing channels (A, B, C, and D). A and B are master and standby channels in normal system operation. The outputs of these channels pass through gates which allow either channel A or channel B (but not both) to drive conjugate ports on the 64- and 512-kHz combiner hybrids. The output ports of the hybrids drive resistive distribution buses, and the conjugate ports drive the major alarm transfer circuit which detects a loss of signal at the hybrid output. The gates are under the control of the automatic switch circuit which initiates a channel transfer upon receipt of a command from the major alarm transfer circuit. Provisions are also made for manually reversing the state of the gates.

The phase align and channel transfer circuit in conjunction with the capacitive phase shifters and the magnetic clutch assembly allow intentional manual transfers to be performed which introduce phase hits of less than 20 nanoseconds (less than 4° at 512 kHz) at the output distribution buses. This type of "hitless" transfer is initiated by a pushbutton switch on the front panel which enables the phase align and channel transfer circuit and the magnetic clutch assembly. The shaft of the phase shifter in the off-line channel is then connected to the shaft of a knob on the front panel by energizing the appropriate magnetic clutch. The knob is rotated, and when phase coherence (less than 20 nanoseconds phase difference) of the 4.096-MHz square waves in the two frequency dividers is detected, the counters in the off-line channel frequency divider are reset to zero

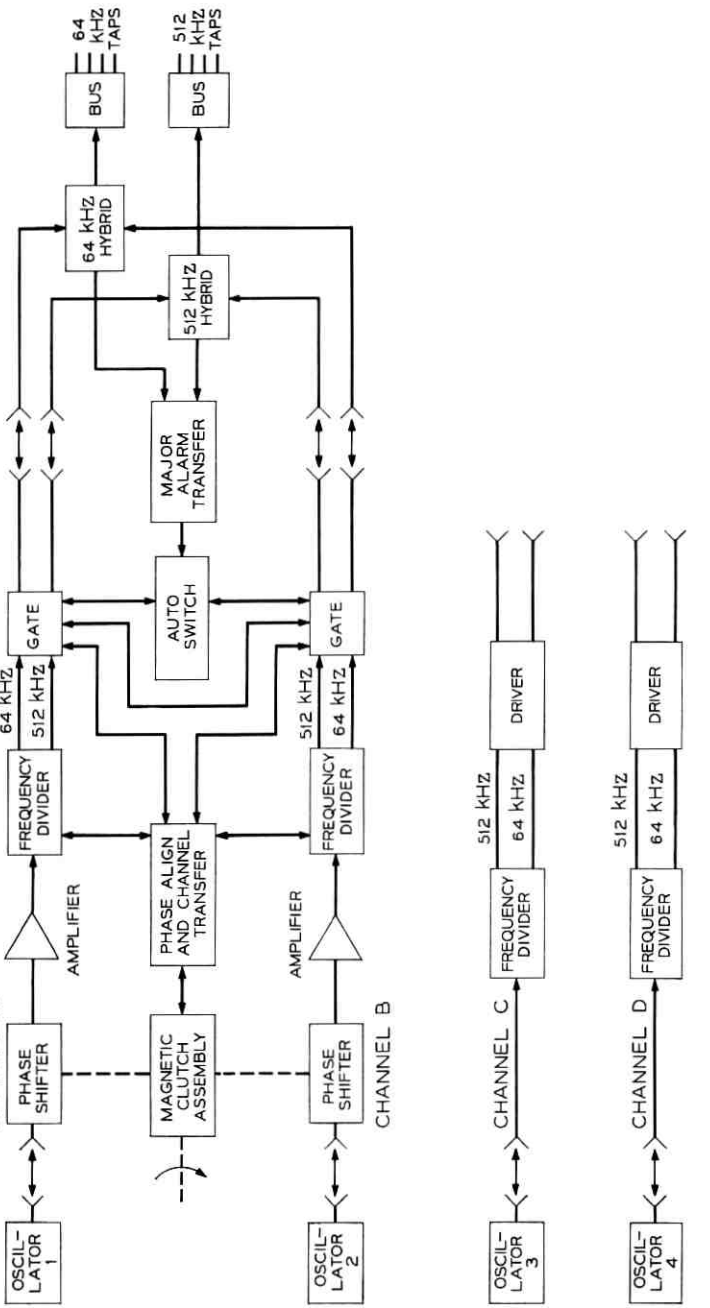


Fig. 1 — Block diagram of principal circuits in reference frequency standard.

upon detection of the all-zero code in the on-line divider. At this instant a command pulse changes the state of the channel gates, thereby effecting the "hitless" transfer. Upon completion of the above sequence, the magnetic clutch is de-energized, and the phase align and channel transfer circuit is returned to its normal state.

For additional redundancy, two "hot spare" channels have been included, either of which can be manually patched into the combiner hybrids should a failure occur in both channel A and B or the common control circuits. These hot spare channels, C and D, are complete with oscillator, divider, and driver circuits.

Manual patching allows any of the four oscillators to be associated with any of the four channels. Under normal conditions, the two most stable oscillators (determined by vlf measurements) would be patched into channels A and B, and the outputs of channels A and B would be connected to the combiner hybrids. The flexibility provided by the manual patches at both the oscillator outputs and the hybrid inputs allows frequency service to be restored under a wide variety of failure conditions.

In addition to the primary circuit functions there are a number of secondary functional blocks which are described in greater depth in sections to follow. They are: power supply circuits (Section 3.5), major alarm circuits (3.6), minor alarm circuits (3.7), vlf receiver-comparator circuits (3.8), and digital time-of-day clock (3.9).

3.2 *Precision Oscillators*

The solid-state precision oscillators used in the reference frequency standard are the double-oven quartz-crystal type.[2, 3] Figure 2 is a block diagram of the oscillator circuit. The crystal is an AT-cut polished plano-convex quartz plate designed to operate on its fifth mechanical overtone in the thickness-shear mode. The crystal is mounted in an evacuated enclosure and is housed in an inner temperature-controlled oven. A vernier frequency adjustment is provided which allows a variation of several parts in 10^7 about a nominal value of 4.096 MHz by changing the bias on a varactor diode. The circuits comprising the oscillator (including the inner oven assembly) are surrounded by an outer temperature-controlled oven for greater temperature stability.

The drift rate for the precision oscillators is a few parts in 10^{10} per week. Figure 3 shows a plot of frequency offset as a function of time over a period of several months for one of the oscillators. Drift rates such as these can only be realized after an initial warm-up

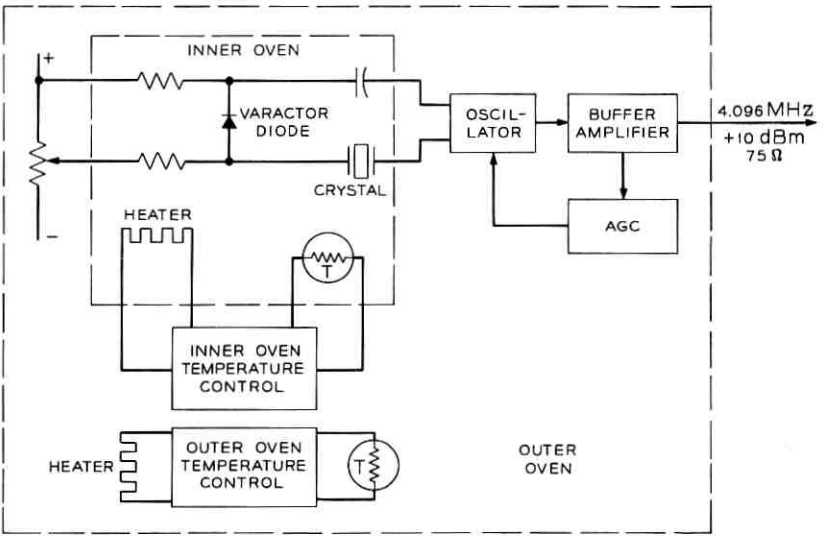


Fig. 2 — Precision oscillator block diagram.

period of approximately one month, during which the frequency gradually stabilizes. Any disruption of power after this time requires another period of stabilization before low drift rates can again be realized. For this reason considerable redundancy has been designed into the oscillator power supply arrangement (see Section 3.5).

3.3 *Binary Frequency Dividers*

The frequency dividers used in the reference frequency standard are six-stage binary counters with output taps at 512 and 64 kHz (see Fig. 4). The input to the counter chain is a 4.096-MHz square wave derived from the oscillator output sine wave signal. The input to the 512-kHz output flip-flop is the result of gating the 4096-, 2048-, and 1024-kHz signals. This technique reduces jitter and eliminates the cumulative delay uncertainty of several counter stages. A similar gating arrangement is provided at the input to the 64-kHz flip-flop. High speed switching transistors with storage times of less than 5 nanoseconds are used in the counter circuits to further reduce delay uncertainty. These techniques are required to insure accurate operation of the phase align and channel transfer circuit and to reduce jitter in the output signals.

Figure 5 is a schematic diagram of a single stage in the counter chain. The "toggle" input is triggered by negative-going pulses, and

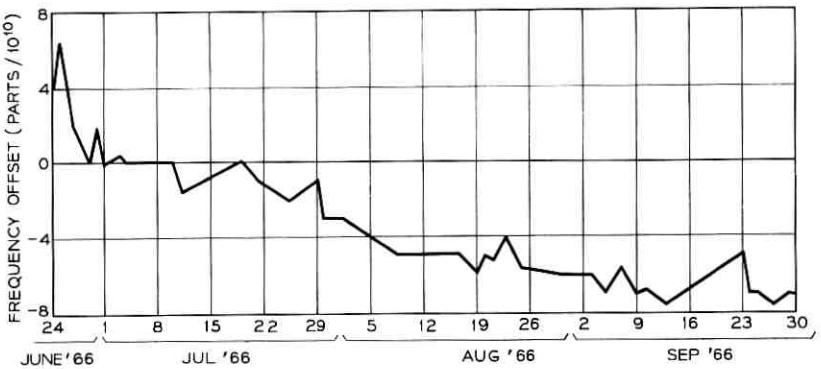


Fig. 3 — Typical precision oscillator drift characteristics.

a positive pulse at the "reset" input forces the output to a logic zero, corresponding to the saturation voltage of the transistor.

3.4 *Phase Align and Channel Transfer Circuits*

The purpose of the phase align and channel transfer circuits is to allow the initiation of an intentional transfer between channels A and B which introduces a phase hit of less than 20 nanoseconds at the output buses. The basic transfer sequence is described in Section 3.1. Fundamentally, the circuits must detect the phase coincidence of the 4.096-MHz square waves in both channels and the all-zero

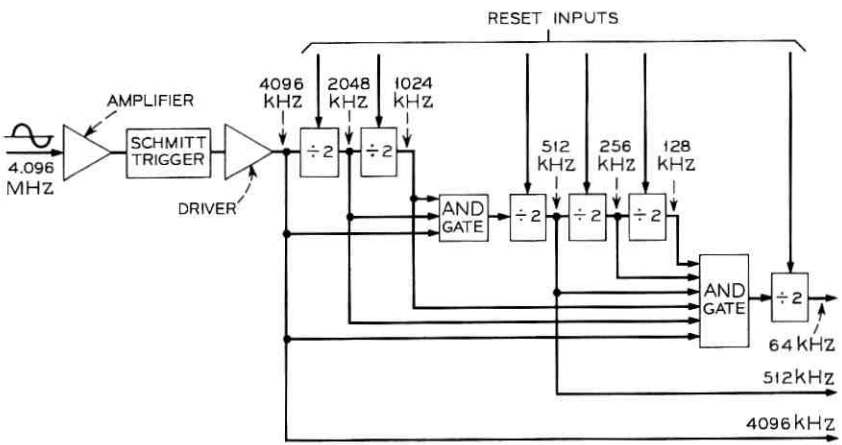


Fig. 4 — Frequency divider circuit.

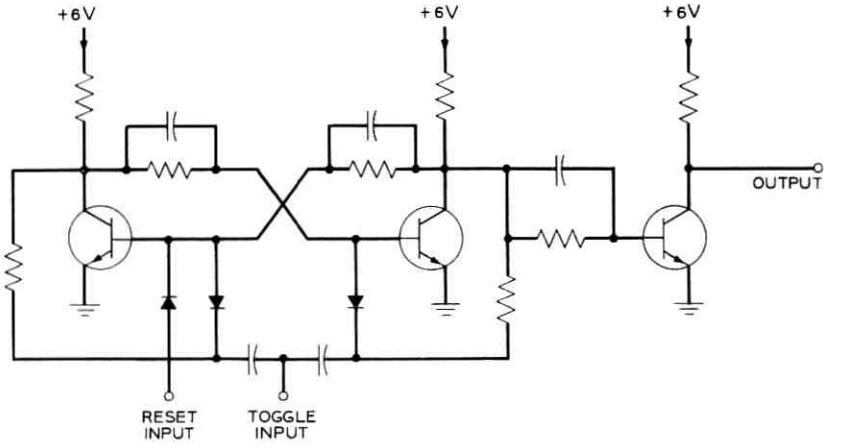


Fig. 5 — Binary flip-flop circuit.

code in the on-line channel in order to generate properly timed reset pulses for the off-line divider and channel transfer commands for the automatic switch circuit. Figure 6 is a block diagram of the detection and pulse generating circuit.

The shapers (see Fig. 7) generate narrow pulses of less than 10

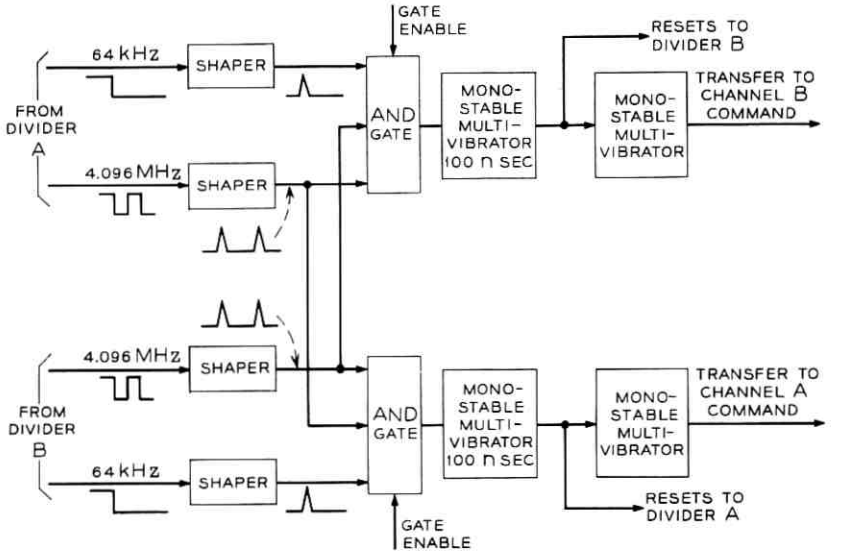


Fig. 6 — Phase align and channel transfer, phase coincidence circuit.

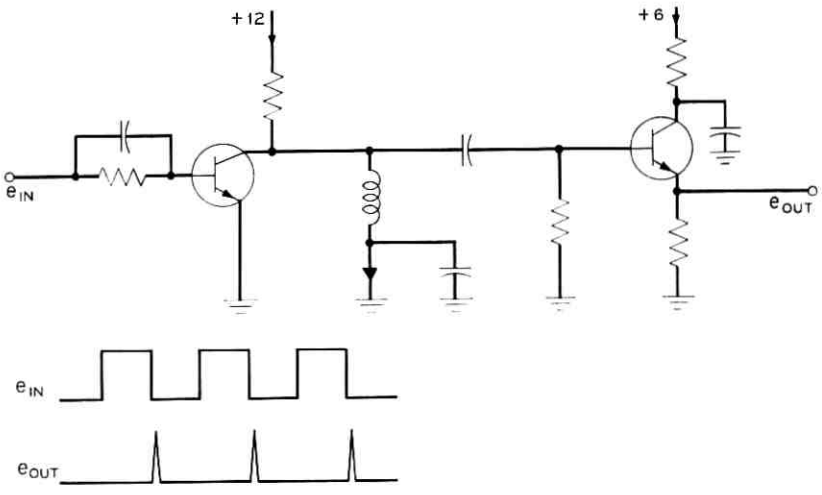


Fig. 7 — Shaper circuit.

nanoseconds basewidth from the negative-going portions of the input square waves. When the phase align and transfer circuit is activated, the appropriate diode AND gate is opened and the output of the gate is a properly timed trigger pulse which occurs only when the 4.096-MHz signals are in phase and at the precise instant of the all-zero code in the on-line channel frequency divider. A 100 nanosecond monostable multivibrator, which is triggered by the pulse, generates reset pulses for the off-line divider before the next 4.096-MHz cycle and inhibits the diode AND gate. A second monstable multivibrator which is triggered by the reset pulse generates the transfer signals for the channel gates. A special mode is provided for tests which allows the off-line counters to be reset but inhibits the output of the second monostable multivibrator, thus preventing an actual channel transfer.

3.5 *Power Supply Circuits*

In the frequency standard, dc-to-dc converters have been used to isolate critical circuits from noise and voltage fluctuations associated with the central office battery. A total of eight converters are used in the system. Four converters supply power to the oscillators and four to the digital circuits. The modular plug-in construction of the converters facilitates replacement.

A redundancy scheme is followed in the fusing and distribution of

the regulated voltages so that no single converter failure will cause a catastrophic failure at the output of the distribution buses. Figure 8 illustrates the redundant power feed arrangement for the precision oscillators. Office battery voltage is brought to the frequency standard cabinet over two independent fused paths from the battery bus. Two high-power silicon diodes form an OR gate which allows either path to fail (either open or short) without disrupting power to the dc-to-dc converters. The outputs of each pair of converters are also passed through OR gates so that any one of the four converters may fail without disrupting power to any oscillator. The fusing arrangement insures that no single failure in either a diode, a converter, or a fuse to the left of the dotted line in Fig. 8 disrupts power to any oscillator.[4] This fusing scheme also allows power to be temporarily removed from faulty diodes or converters, thereby allowing repairs to be made without disrupting power to the parts of the system not directly affected by the failure. A fuse alarm contact is provided on each fuse. This contact is connected to the fuse input if the fuse should blow, thereby notifying the appropriate minor alarm circuits.

3.6 *Major Alarm Transfer Circuits*

The major alarm transfer circuits monitor the output signal levels and generate a central office major alarm should the levels fall outside preset limits. Figure 9 is a block diagram of the circuit. The alarm circuits are driven from the ports of the 64- and 512-kHz hybrids which are conjugate to the ports driving the distribution buses. Two

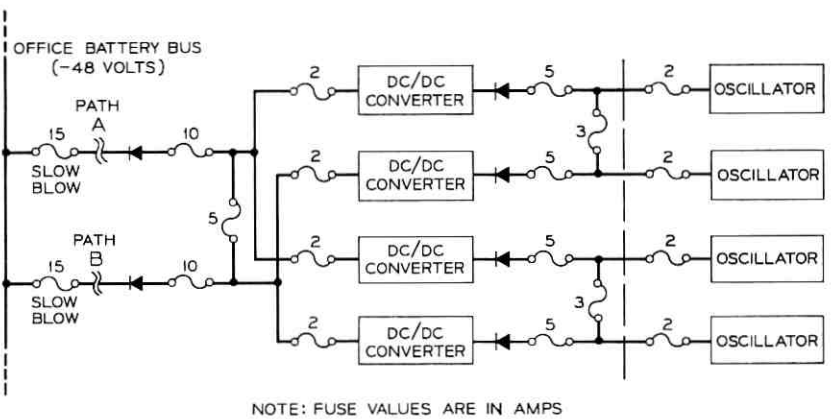


Fig. 8 — Oscillator power supply circuit.

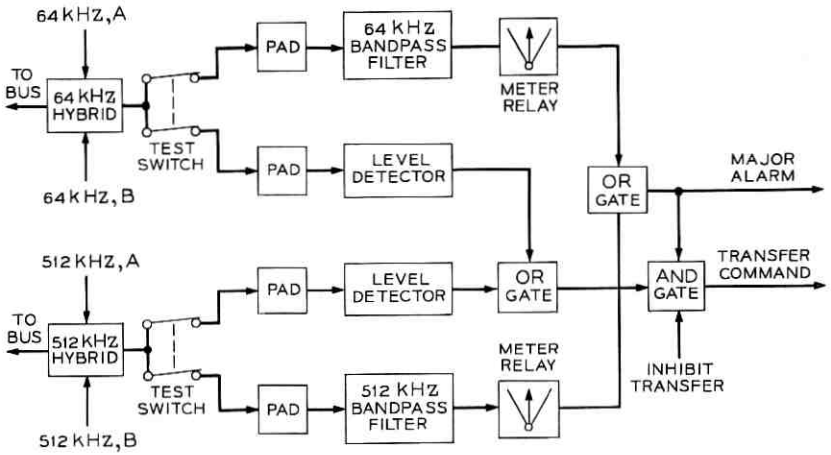


Fig. 9 — Major alarm transfer circuit.

monitoring channels are driven from each hybrid. One channel drives a meter-relay with adjustable upper and lower limits. The bandpass filter in series with the meter allows this channel to detect gross frequency shifts as well as level variations. The second channel contains a circuit which detects a catastrophic loss of level. A central office major alarm is generated if either or both the meter indications drift outside the present limits of a few dB above or below the nominal level of −23 dBm. A command which initiates an actual channel transfer is generated only if both monitoring channels detect a fault. This logic insures that a single failure in either monitoring channel does not generate a needless transfer and a possible phase hit at the output.

An input from the minor alarm circuits prevents an automatic transfer if a minor alarm condition exists in the off-line channel. Test switches are provided which allow a failure to be simulated, thereby verifying the proper operation of the major alarm circuits and the redundancy switch. This test can be performed without introducing a phase hit at the output by first using the phase align and channel transfer procedure to bring the two signals into phase coincidence.

3.7 *Minor Alarm Circuits*

The minor alarm circuits monitor critical points associated with the oscillators, frequency dividers, and power supplies. In case of a failure in any of the monitored circuits, the central office alarm is

activated and lamps on the front of the cabinet allow the trouble to be quickly isolated.

In order to insure a prompt indication of any significant shift in the output frequency of the oscillators driving channels A and B, the circuit shown in Fig. 10 constantly monitors the frequency offset between channels A and B. A similar circuit monitors the frequency offset between channels A and C. Two meter-relays with adjustable upper limits provide the alarm initiation. The meters are calibrated to read 12 parts in 10^9 full scale. Assuming that the frequency of only one of the three channels drifts out of limits, simple relay logic determines which of the three channels contains the fault. The monitoring circuit (Fig. 10) uses a balanced modulator followed by a low-pass filter to generate the difference frequency. Zero-crossings of the difference frequency are counted by a six-stage counter for a 10-minute interval, and the count is read into a register. The binary output of the register is translated by a digital-to-analog converter into a signal which is proportional to the number of zero-crossings of the difference frequency in a 10-minute interval. Each count represents a phase shift of 180° between the two 4.096-MHz signals, or 122 nanoseconds. This corresponds to a fractional offset of 2.04 parts in 10^{10} per count. With a six-stage counter, 63 discrete nonzero levels exist, hence a full count of 63 represents 12.9 parts in 10^9 frequency offset between channels. The signals which control the gates and reset the counters at 10-minute intervals are derived from the binary coded decimal time code outputs of the digital time-of-day clock by a decoder circuit.

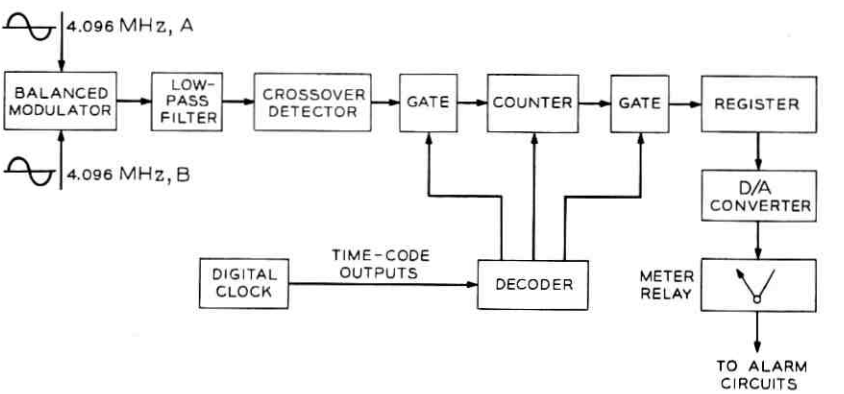


Fig. 10 — Difference frequency detector.

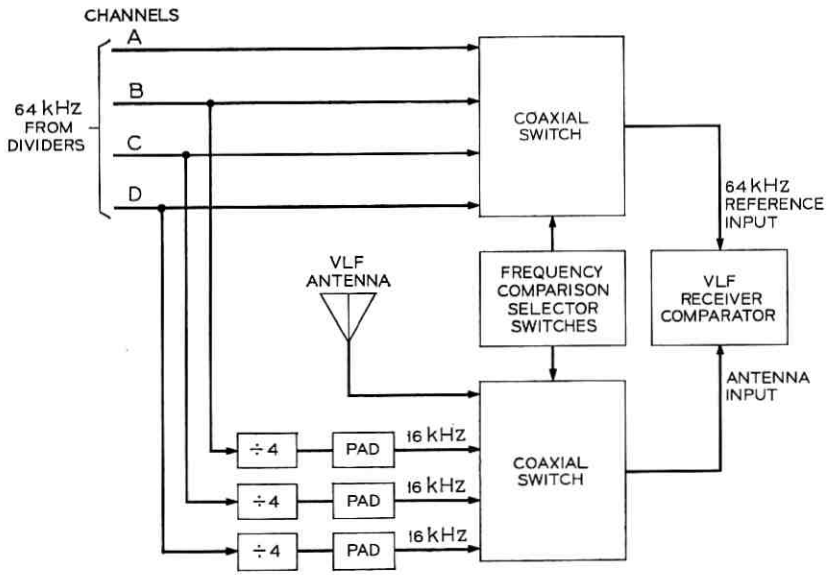


Fig. 11 — VLF receiver-comparator mode-select circuits.

Several other points in the system are monitored by various minor alarm circuits. They are:

(*i*) Inner oven heater voltage for each of the four oscillators (deviations about a nominal level of more than one volt generate a minor alarm).

(*ii*) Output of each frequency divider (loss of output of any off-line divider generates a minor alarm).

(*iii*) Oscillator power supplies (deviations of more than one volt generate a minor alarm).

(*iv*) Digital power supplies (loss of any supply generates a minor alarm).

(*v*) Fuses (any blown fuse generates a minor alarm).

3.8 *VLF Receiver-Comparator Circuits*

The vlf receiver-comparator circuits allow accurate frequency comparisons to be made between any of the four channels and any receivable vlf station. Frequency comparisons between any two channels are also possible. This feature allows frequency standardization to continue using any of the four oscillators as a local standard should

normal vlf receptions be impaired for significant periods of time, such as during a national disaster.

Figure 11 is a block diagram of the vlf receiver-comparator circuit. The selector switches on the front panel (see Fig. 17) allow any one of the ten possible frequency comparison modes to be selected. Intercomparisons between channels A and B, for example, are made by connecting the 64-kHz signal from channel A to the reference frequency input of the receiver and a low-level 16-kHz signal derived from channel B to the antenna input of the receiver. This frequency division to 16 kHz is necessary since neither 64 kHz or 32 kHz are valid received frequencies for the vlf receiver-comparator.

The vlf receiver-comparator accepts a 64-kHz reference signal input. The receiver then synthesizes a 1.1-kHz signal from this 64-kHz

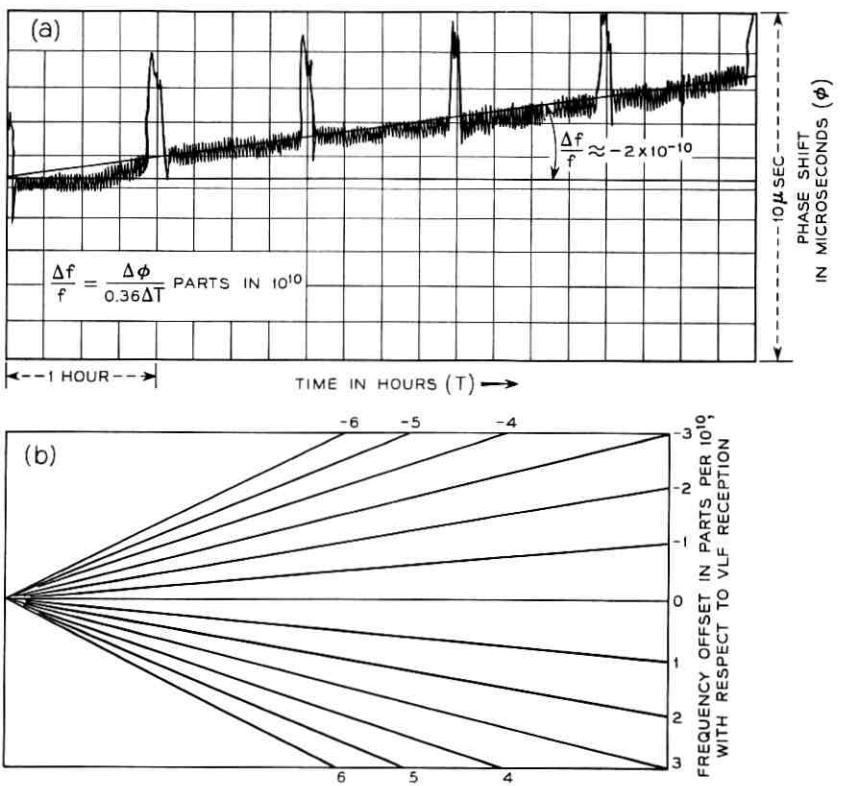


Fig. 12 — (a) VLF strip-chart recording and (b) frequency offset template.

input and phase corrects it to a 1.1-kHz signal derived from the antenna input. The accumulated phase difference between the two 1.1-kHz signals is plotted on a strip-chart recorder, from which accurate frequency offsets can be easily computed (see Section IV).

3.9 *Time-of-Day Clock*

The digital time-of-day clock is driven from the alarm port of the 64-kHz combiner hybrid (see Fig. 1). Days, hours, minutes, and seconds are displayed on the panel by indicator tubes, and controls are provided which allow the clock readout to be synchronized with a suitable time-standard broadcast. The digital clock also generates a time code (binary coded decimal) which is used by the difference frequency detecting circuits (see Section 3.7) in deriving the periodic control pulses.

IV. FREQUENCY STANDARDIZATION

The technique of accurate frequency offset measurements using a vlf receiver-comparator is relatively simple and has attained widespread use in recent years.[5, 6] Because of the effects of propagation anomolies during sunrise, sunset, and darkness, vlf measurements usually are made only when both the vlf transmitter and receiver are in daylight. A typical strip chart recording is shown in Fig. 12a. The plot shows the phase difference between the 1.1-kHz signal derived from the 64-kHz source and the 1.1-kHz signal derived from the vlf

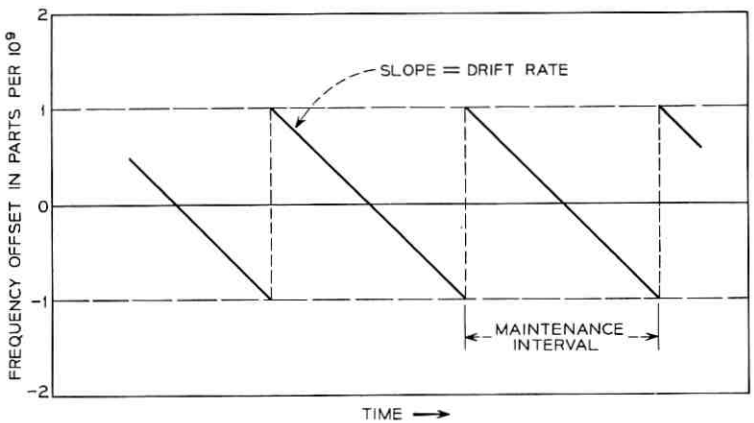


Fig. 13 — Maintenance of oscillator frequency accuracy by periodic corrections.

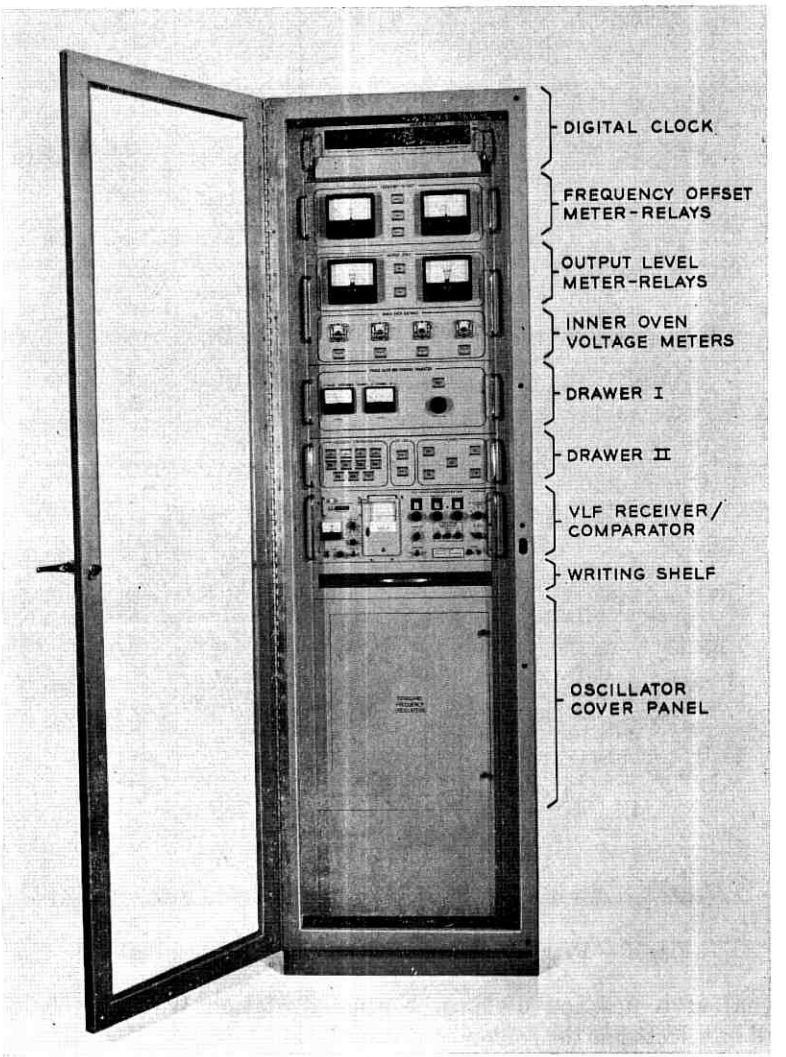


Fig. 14 — Front view of reference frequency standard cabinet.

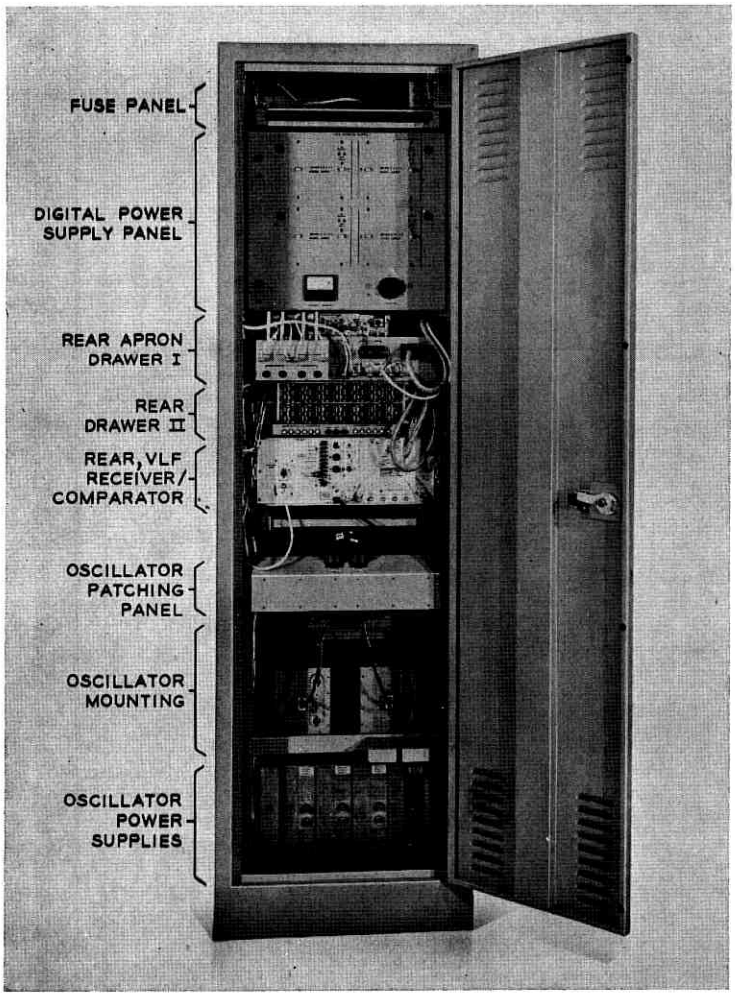


Fig. 15 — Rear view of reference frequency standard cabinet.

signal as a function of time. Normally, station WWVB (60 kHz) will be selected as the reference station.*

The phase offsets which occur every hour are introduced by WWVB to serve as identification. The fractional frequency offset is simply

* Station WWVB, Fort Collins, Colorado is operated by the National Bureau of Standards Time and Frequency Division at Boulder, Colorado. The 60-kHz transmitted signal is based on the atomic second which is defined in terms of a specified transition between electron energy levels of Cesium-133.

the ratio of the accumulated phase shift to the time interval over which the phase shift occurs. For the example shown in Fig. 12a this ratio is approximately 2 parts in 10^{10}, the sign of the slope indicating that the oscillator frequency is low with respect to WWVB. Measurements of this type can be quickly made with the aid of a special transparent template which is calibrated directly in fractional parts in 10^{10} as shown in Fig. 12b.

The procedure for maintaining an oscillator to within one part in 10^9 of WWVB transmissions is as follows. A series of weekly frequency offset measurements are made in order to establish the oscillator's approximate drift rate. The maximum allowable maintenance interval is then determined based on the drift rate and the accuracy requirement of one part in 10^9. A plot of frequency offset versus time for an oscillator being maintained to the above accuracy is shown in Fig. 13. The periodic correction of 2 parts in 10^9 is made using the calibrated vernier frequency adjustment on the oscillator.

V. EQUIPMENT DESIGN

The equipment comprising the reference frequency standard is housed in a special seven-foot high steel cabinet which is designed

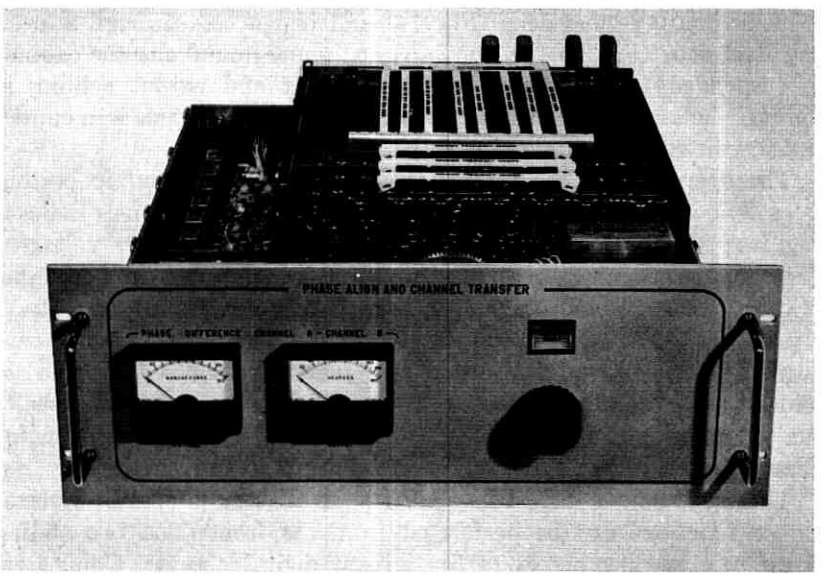


Fig. 16 — Drawer I.

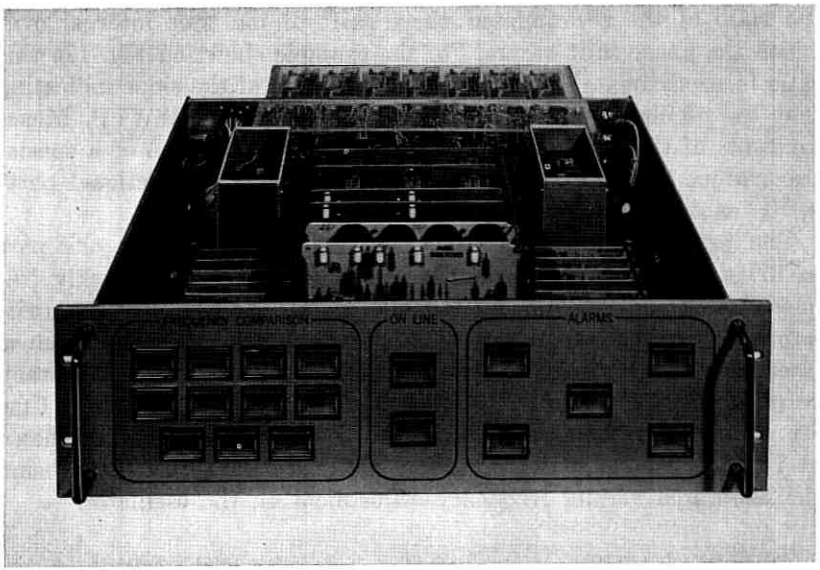


Fig. 17 — Drawer II.

to mount standard 19-inch panels or drawers. Front access is through a locking safety-glass door (Fig. 14) and rear access through a locking steel door (Fig. 15). In the hardened underground site the cabinet is suspended from shock mounts. All input and output cabling is through holes in the top of the cabinet. The vlf antenna is mounted outside the entrance to the site.

The overall front cabinet panel arrangement from top to bottom is as follows: digital clock (mounted on slides); meter panels indicating frequency offset, output levels, and inner oven heater voltages; drawer I containing frequency dividers, phase align and channel transfer circuits, and certain alarm circuits (mounted on slides); drawer II containing frequency comparison selector switches and logic, difference frequency circuits, and alarm circuits (mounted on slides); vlf receiver-comparator (mounted on slides); writing shelf; and oscillator panel which swings open for access to precision oscillators.

The front panels have been designed so that most normal maintenance functions can be performed from the front. The two sliding drawers give easy access to the critical digital circuits. Connectors at the rear of each drawer allow the entire drawer to be easily

removed from the cabinet should extensive maintenance be required.

Drawers I and II are shown in Figs. 16 and 17, respectively. The meters on the front panel of drawer I display the phase difference between the 4.096-MHz signals in dividers A and B and between the 64-kHz signals in dividers A and B. The pushbutton initiates the operation of the phase align and channel transfer circuits, and the knob allows the off-line phase shifter to be rotated. The meters indicate proper operation of the circuit when they read zero after the channel transfer has occurred.

The front panel of drawer II contains the frequency comparison selector switches which allow any possible frequency comparison to be made by the vlf receiver. On-line indicator lamps and various major and minor alarm lamps are also on this panel as well as the alarm cutoff pushbutton switch which silences the office alarm.

A major portion of the circuits in the frequency standard are mounted on printed wiring plug-in cards for ease and speed in replacement of defective units. The cards on which high-speed logic is performed have a solder-plated ground plane on the component side as shown in Fig. 18. This ground plane is returned to chassis ground via the connector. In this manner a good sink is provided for the large transient ground currents which are generated by the high-speed switching circuits. Other low-speed circuits are mounted on a more standard card of the type shown in Fig. 19.

The units which can be reached from the rear of the cabinet (see

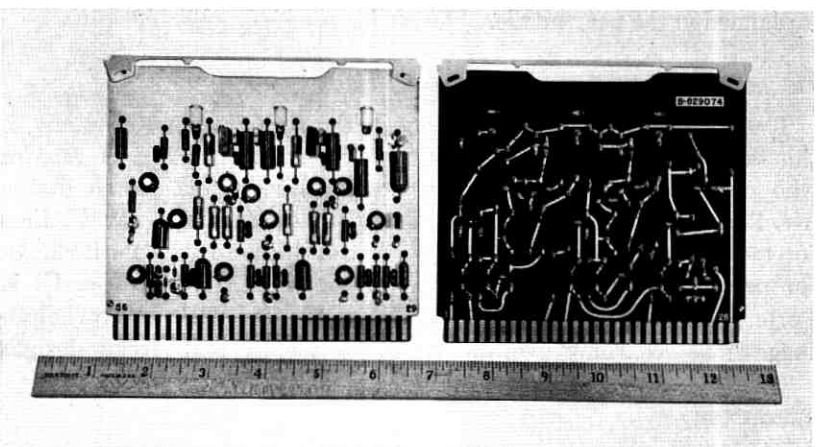


Fig. 18 — High speed plug-in card: (a) component side, (b) printed wiring side.

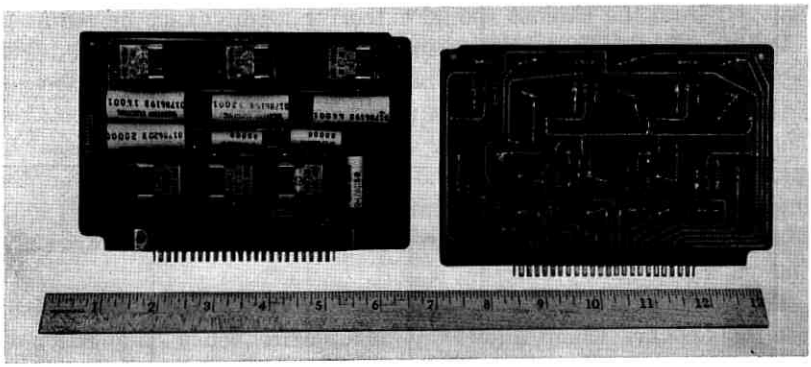


Fig. 19 — Standard plug-in card: (a) component side, (b) printed wiring side.

Fig. 15) are (from top to bottom): fuse panel; digital power supply panel; rear panels of Drawers I and II; rear panel of vlf receiver; oscillator patching panel; rear of oscillators; and oscillator power supply panel.

VI. CONCLUSIONS

The new Bell System Reference Frequency Standard has been designed to meet the frequency accuracy, reliability, and maintainability requirements of the present L multiplex system as well as the next generation of carrier equipment. The system has been installed and tested in the hardened underground telephone office, and a cutover is planned in the near future.

VII. ACKNOWLEDGMENTS

The author wishes to acknowledge the contributions of many in this development effort. Recognition is especially due Mr. H. B. Goff who participated with the author in the system and circuit design, Mr. P. E. Riley for assistance in circuit design and tests, Mr. D. Lane for mechanical design, Messrs. W. L. Smith and H. S. Pustarfi, Jr., for precision oscillator design and development, and Messrs. G. R. Porter, D. C. Weller, and R. E. Benjamin for their many helpful suggestions regarding overall system philosophy and circuit designs.

REFERENCES

1. Clark, O. P., Drazy, E. J., and Weller, D. C., "A Phase-Locked Primary Frequency Supply for the L Multiplex," B.S.T.J., *42*, No. 2 (March 1963), pp. 319–340.

2. Smith, W. L., "Miniature Transistorized Crystal-Controlled Precision Oscillators," I.R.E. Trans. Instrumentation, *I-9,* No. 2 (September 1960), pp. 141–148.
3. Smith, W. L., "Precision Quartz Crystal Controlled Oscillators Using Transistor Circuits," Bell Laboratories Record, *42,* No. 8 (September 1964), pp 273–279.
4. Benjamin, R. E., unpublished work.
5. Stone, R. R., Jr., Markowitz, W., and Hall, R. G., "Time and Frequency Synchronization of Navy VLF Transmissions," I.R.E. Trans. Instrumentation, *I-9,* No. 2 (September 1960), pp. 155–161
6. Stone, R. R., Jr., "Synchronization of Local Frequency Standards with VLF Transmission," Proc. 16th Annual Frequency Control Symposium, 1962, pp 227–240.

Analysis and Synthesis of a Digital Phase-Locked Loop for FM Demodulation

By G. PASTERNACK and R. L. WHALIN

(Manuscript received July 12, 1968)

A method of synthesizing a general nth order phase-locked loop is presented. In contrast to conventional phase-locked loops, the circuitry is digital rather than analog. The general circuit consists of an assembly of logic blocks (gates and storage elements) which, when driven by external clock signals, exhibits phase-locked loop properties. These properties, along with high stability and the absence of adjustments, make the digital phase-locked loop ideally suited for use in large systems which use monolithic integrated circuits for microminiaturization. Analysis and synthesis techniques make use of Z-transform methods in achieving the desired frequency response as the realization of an nth order difference equation. A general technique is developed and two specific cases, $n = 1$ and $n = 2$, are considered in detail. Analytic results relating to the phase-locked loop's static and dynamic performance are derived and found to correlate well with laboratory results for actual circuits.

I. INTRODUCTION

A new phase-locked loop (PLL) with interesting properties has been developed for potential application in large multiple data set installations which provide low speed serial data communications for time-shared computers. An objective for such data set arrangements is to minimize cost per channel by putting the major part of the required circuitry into a common section where it may be shared by all channels. This objective is achieved by using a digital PLL as an FM demodulator with low cost logic circuits located in the channel units and clocks with their associated driving amplifiers located in the common circuits. PLL's which use analog circuits have received considerable attention and analysis and synthesis methods

are available.[1, 2] However, the circuits covered here are digital, and the approach is similar to that of digital (or sampled-data) filters.[3, 4] By using a digital PLL, no low-pass filter or voltage-controlled oscillator, generally associated with the feedback loop of conventional PLL's is required.* This property, along with high stability and the absence of adjustments makes the digital PLL ideal for microminiaturization using monolithic integrated circuits.

This paper presents snythesis procedures for an nth order digital PLL. The PLL realized by such a procedure possesses a response which obeys a linear nth order difference equation. Analysis is performed using Z transform methods commonly encountered in sampled-data control systems.[5] The technique is that of establishing a mapping between the s plane and the z plane so that a correspondence between the coefficients of the controlling nth order difference equation and the desired s plane poles may be established. An iterative circuit is presented so that once the coefficients are determined, the nth order loop may be realized.

The remainder of the paper is devoted to the synthesis, realization, and analysis of two specific examples, $n = 1$ and $n = 2$. Both systems are analyzed to determine static and dynamic performance, and the results of data transmission tests are given. The out-of-lock behavior as well as internal noise resulting from jitter is characterized for the first order PLL. The second order PLL is representative of higher order systems and the analysis is easily extended. The "capture phenomenon" associated with underdamped systems is encountered. In both of the examples considered, experimental results are found to correlate well with theory.

II. THE DIGITAL PHASE-LOCKED LOOP

As a prelude to the synthesis procedure we show that the loop response of the PLL can be expressed as an nth order difference equation. Figure 1 is a block diagram of the loop. Among its basic components are an exclusive OR comparator which develops an output gating function dependent upon the phase relation of its inputs, and a transmission gate acting on several clock signals f_i, g_i to provide inputs to register circuitry. When the loop is locked, a shift circuit periodically transfers the contents of the $(i - 1)$th register to the ith register; the period is one half that of the input signal. The shift

* Although no internal loop filter is needed, a low-pass filter is required to recover the demodulated baseband signal.

TABLE I — SYMBOL

α_j	Real part of z-plane pole, z_j.
β_j	Imaginary part of z-plane pole, z_j.
γ, δ	Parameters pertaining to stability analysis of second order PLL.
Δ	Normalized total input frequency deviation.
Δf	Total input frequency deviation.
Δf_L	Lock range of PLL.
Δv	Voltage step.
Δv_N	Peak-to-peak noise voltage.
Δv_{sig}	Peak-to-peak signal voltage.
ϵ_0, ϵ_0', ϵ_1	Time errors resulting from phase discontinuity when switching between clocks.
μ	Real and imaginary parts of Butterworth characteristic with carrier frequency input.
$\xi(k)$	Duration of kth positive (logical "1") level of feedback (flip-flop) signal.
$\rho(k)$	Interval between kth and $(k + 1)$th zero crossing of input signal.
σ_j	Real part of s-plane pole, s_j.
$\tau(k)$	Duration of kth positive (logical "1") level at exclusive- OR output.
ω_c	Radial cutoff frequency of low-pass filter.
ω_j	Imaginary part of s-plane pole, s_j.
C	Counting capacity of N-stage counter.
E	Input voltage.
E_f	Feedback voltage.
E_0	PLL output voltage.
f	Frequency of input signal.
f_a, f_b	Discrete input frequency.
f_c	Cutoff frequency of low-pass filter.
f_m	Average input (carrier) frequency.
f_r	Rest frequency of PLL.
f_u, f_l	End point of frequency lock range.
f_i, g_i	Clock frequencies ($i = 1, 2, \cdots, n$); also denotes input to registers.
F_i, G_i	Normalized clock frequencies ($i = 1, 2, \cdots, n$).
$H(s)$	Transfer function (Butterworth).
i	$(-1)^{\frac{1}{2}}$.
i, j, k, l	Counting indices.
M	Threshold of final register.
n	Order of system; number of registers.
N	Number of counter stages per register.
N_m	Minimum N for stable operation.
s_j	jth s-plane pole.
t	Time.
T_d	Period of baseband data signal.
T_1, T_2, T_3, T_4	Signal durations for stability analysis.
$v(k)$	Average output voltage during kth interval
$v(t)$	Continuous time function.
v_2	Average output voltage for parasitic mode.
z	Argument of z-transform.
$Z[v(k)]$, $V(z)$	Z-transform of $v(k)$.

(and reset of the first register) is controlled by the nth register which provides an output pulse after the Mth clock pulse is counted. A flip-flop converts the output pulse train to a square wave and provides an input to the comparator.

The difference equation describing operation of this circuit is developed with the aid of waveforms shown in Fig. 2, which gives the

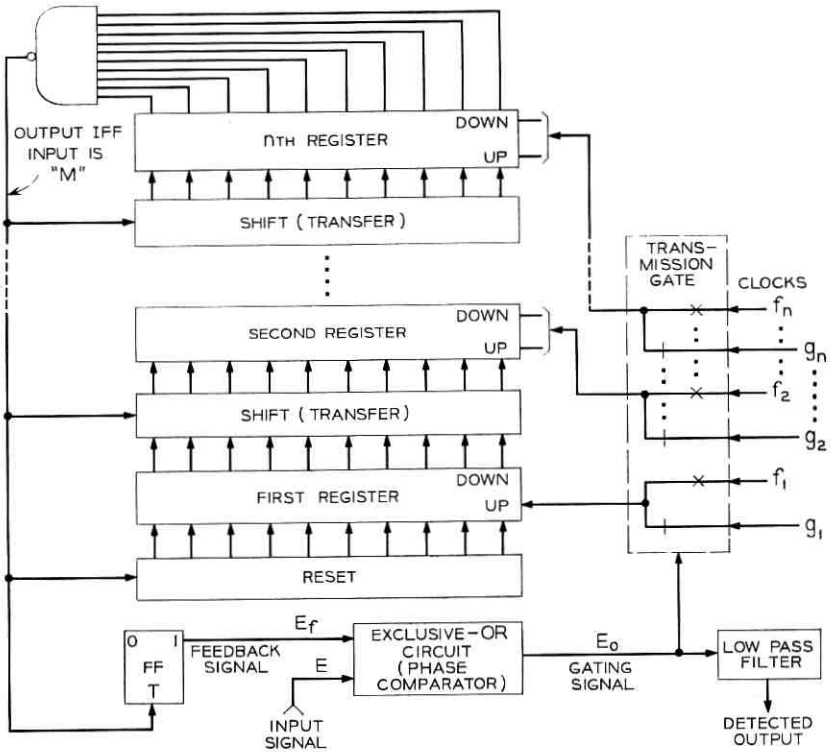


Fig. 1 — Block diagram of general digital PLL.

steady-state terminal signals for the comparator. The output signal is the gating function. Assume that clocks $g_1, g_2, g_3, \ldots, g_n$ are enabled during the "0" level, and all other clocks $f_1, f_2, \ldots, f_n$ are enabled during the "1" level. With the first register initially cleared, its count at the conclusion of the $(k+1)$th period of gating is $g_1[\rho(k) - \tau(k)] + f_1\tau(k+1)$. During each successive period, this count is shifted into the ith register and augmented by a count of $g_i[\rho(k+i-1) - \tau(k+i-1)] + f_i\tau(k+i)$ for $i = 2, 3, \ldots, n$. The count propagates through the n registers where, upon reaching the number M at the nth register, the count is reinitiated. Although n periods are required for a complete count cycle, the process may be thought of as the interleaving of cycles initiated ρ seconds apart.

III. GENERAL FORM OF THE DIFFERENCE EQUATION

It can be seen from Fig. 2 that the count of the nth register is the sum of pulses counted during the intervals $\rho(k) - \tau(k)$, $\rho(k+1) - \tau(k+1), \ldots, \rho(k+n-1) - \tau(k+n-1), \tau(k+1), \tau(k+2), \ldots, \tau(k+n)$. Summing the counts for the sources $f_1, f_2, \ldots, f_n$, $g_1, g_2, \ldots, g_n$ and equating the sum to M gives

$$\begin{aligned} g_1[\rho(k) - \tau(k)] + g_2[\rho(k+1) - \tau(k+1)] + \cdots \\ + g_n[\rho(k+n-1) - \tau(k+n-1)] \\ + f_1\tau(k+1) + \cdots + f_n\tau(k+n) = M, \end{aligned} \tag{1a}$$

or rewriting,

$$\begin{aligned} f_n\tau(k+n) + (f_{n-1} - g_n)\tau(k+n-1) + \cdots - g_1\tau(k) \\ = M - [g_n\rho(k+n-1) + g_{n-1}\rho(k+n-2) + \cdots \\ + g_2\rho(k+1) + g_1\rho(k)]. \end{aligned} \tag{1b}$$

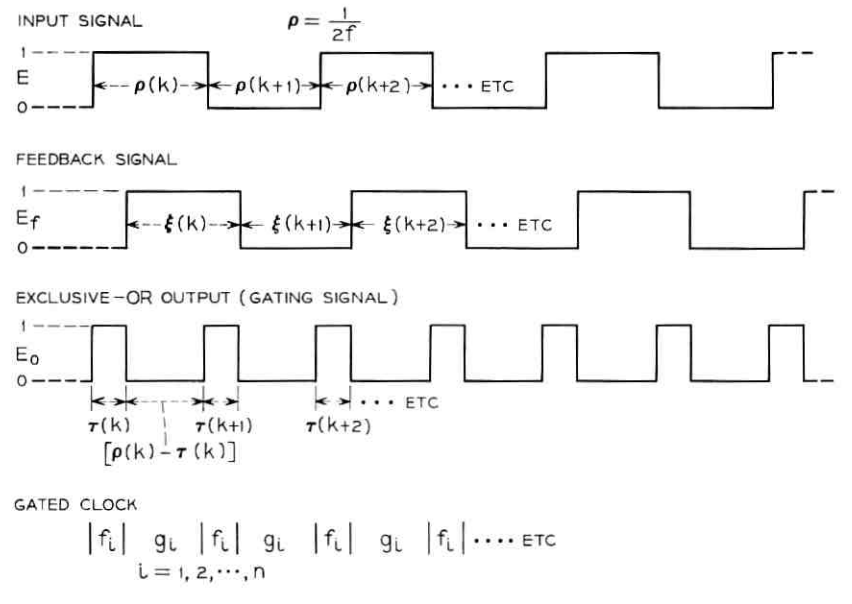


Fig. 2 — Input, feedback, and output waveforms for PLL which is assumed to be locked with constant input frequency. Gated clock signals are shown symbolically.

This is the system difference equation relating the response $\tau(k + i)$ to an excitation $\rho(k + i)$. In its general form this equation, when Z-transformed, is analogous to a Laplace transformed system equation in which polynomials in s multiplying the response and excitation functions result in poles and zeros, respectively. Since there is a one-to-one correspondence between the s and z planes with regard to poles and zeros, it is expected that by properly choosing coefficients, equation 1b may be synthesized to provide a desired frequency response (high-pass, low-pass, bandpass, and the like) possessing specified critical frequencies. As we show, it is the objective of this paper to exploit the low-pass properties of equation 1b. To achieve maximum high frequency attenuation, the coefficients g_i, $i = 2, 3, \ldots, n$ will be set equal to zero thereby locating all z-plane zeros at infinity.

$$f_n\tau(k + n) + f_{n-1}\tau(k + n - 1) + \cdots - g_1\tau(k) = M - g_1\rho(k). \quad (1c)$$

Equation 1c is normalized by letting $F_j = f_j/f_n$, $j = 1, \ldots, n - 1$ and $G_1 = g_1/f_n$. Also, notice that the cycle-by-cycle average voltage of the $\tau(k + j)$ interval expressed as a fraction of the maximum possible voltage is given by

$$v(k + j) = \frac{\tau(k + j)}{\rho(k + j)}. \quad (2)$$

Incorporating these substitutions gives

$$v(k + n) + F_{n-1}v(k + n - 1) + \cdots + F_1v(k + 1) - G_1v(k) = 2MF(k) - G_1, \quad (3)$$

where

$$F(k) = \frac{1}{2f_n\rho(k)}.$$

If it is assumed that $\rho(k)$ changes very little with k and that $\rho(k)$ is small with respect to the system response time (which it is), the $v(k)$ can be represented as a sampled continuous function of time, $v(t)$, letting $t = k\rho$. That is, $v(t)$ is a function whose value at the kth zero crossing of the input signal is $v(k)$.

This type of equation is best solved using z transform methods assuming that the input frequency is constant. In particular, the following transform pairs are noted:

$$Z[v(t)] = Z[v(k)] \equiv V(z) \equiv \sum_{j=0}^{\infty} v(j\rho)z^j. \quad (4a)$$

$$Z[A \exp(\alpha k)] = \frac{Az}{z - \exp(\alpha)}. \tag{4b}$$

$$Z[v(k + j)] = z^j[V(z)] - \sum_{l=1}^{j} z^l v(j - l). \tag{4c}$$

Accordingly, equation 3 is transformed as follows:

$$\begin{aligned} &z^n[V(z)] - \sum_{l=1}^{n} z^l v(n - l) \\ &\quad + F_{n-1}z^{n-1}[V(z)] - F_{n-1}\sum_{l=1}^{n-1} z^l v(n - 1 - l) + \cdots \\ &\quad + F_1 z[V(z) - v(0)] - G_1 V(z) = (2MF - G_1)\left[\frac{z}{z - 1}\right]. \end{aligned} \tag{5}$$

Combining terms,

$$\begin{aligned} &[z^n + F_{n-1}z^{n-1} + \cdots + F_1 z - G_1]V(z) \\ &\quad = (2MF - G_1)\left(\frac{z}{z - 1}\right) + \sum_{j=1}^{n}\sum_{l=1}^{j} F_j z^l v(j - l), \end{aligned} \tag{6}$$

where

$$F_n = 1.$$

IV. GENERAL SYNTHESIS PROCEDURE

Before developing a synthesis technique for the digital PLL, it is of interest to review a specific application, that of FM demodulation. A "lock range" may be defined for the PLL whereby steady state input signals having constant frequencies lying within this range will cause a steady-state output, $v(k + j) = v(k + j + 1)$, all j, such that this output is linearly related to the input frequency $f = 1/2\rho$. This is the relation required for demodulation.

For dynamic behavior one may consider a binary baseband signal in which each of the two states is assigned a discrete frequency f_a, f_b within the lock range. If the baseband signal switches randomly between states with a maximum rate $1/T_d$ a new phenomenon is introduced. In this context the PLL may be regarded as a low-pass filter which should possess a bandwidth, ω_c, equal to or greater than π/T_d. By proper choice of the coefficients in equation 6 the desired filter shaping may be achieved. It is expected that the low pass filter characteristics will be a function of the input frequencies f_a, f_b. The resulting complication is

conveniently eliminated by making use of the narrowband approximation $f_a, f_b \approx f_m$ where $f_m = \frac{1}{2}(f_a + f_b)$. This is not unrealistic, because the PLL was developed for just such a narrowband system.* With this in mind, we now give a synthesis procedure. In using Z transform techniques it is assumed that the input frequency is approximately constant so that samples are taken at equal time intervals.

The coefficient of $V(z)$ in equation 6 is the "characteristic polynomial" of the system and using it, the desired low pass filtering properties of the loop can be synthesized. For example, assume that a transfer function with poles in the s plane at $s_1, s_2, \ldots, s_n$ is to be synthesized. Its characteristic polynominal is given by

$$\prod_{i=1}^{n} (s - s_i). \tag{7}$$

A conformal mapping between the s plane and z plane is given by the transformation

$$z = \exp(\rho s), \tag{8}$$

by which the pole $s_i = \sigma_i + i\omega_i$ is mapped into

$$\begin{aligned} z_i &= \exp[(\rho)(\sigma_i + i\omega_i)] \\ &= [\exp(\rho\sigma_i)][\cos(\omega_i\rho) + i \sin(\omega_i\rho)]. \end{aligned} \tag{9}$$

Since the complex s plane poles occur in conjugate pairs, and $\sin(\omega_i\rho)$ is an odd function, the corresponding z plane poles also occur in conjugate pairs, and the desired characteristic equation is transformed into

$$\prod_{i=1}^{n} (z - z_i) = z^n + A_{n-1}z^{n-1} + \cdots + A_1 z + A_0 , \tag{10}$$

where each of the coefficients are real. Equating coefficients in equation 6 to those in equation 10 gives the required clock frequencies $g_1, f_1, \ldots, f_n$. A negative value implies an associated register which counts "down" while a positive value suggests a register which counts "up."

V. FIRST ORDER DIGITAL PLL

5.1 *Static and Dynamic Behavior*

The difference equation for an $n = 1$ PLL is easily written from equation 3:

* Full duplex transmission with $f_a = 1070$ Hz, $f_b = 1270$ Hz in one band, and $f_a = 2025$ Hz, $f_b = 2225$ Hz in the other band.

$$v(k+1) - G_1 v(k) = 2MF - G_1, \tag{11}$$

in which $M = 2^{N-1}$ and N is the number of counter stages in the register. For a steady state condition to exist, $v(k+1) = v(k)$ so that

$$v(k) = \frac{2MF - G_1}{1 - G_1} = \frac{2Mf - g_1}{f_1 - g_1}. \tag{12}$$

Since $0 \leqq v(k) \leqq 1$ the end points of the lock range are given by

$$f \mid_{v(k)=0} \equiv f_l = g_1/2M; \qquad f \mid_{v(k)=1} \equiv f_u = f_1/2M. \tag{13}$$

And so the lock range is

$$\Delta f_L = \mid f_u - f_l \mid = \frac{1}{2M} \mid g_1 - f_1 \mid. \tag{14}$$

The static response is diagrammed in Fig. 3.

The dynamic response to a step change in frequency is given by equation 6 for $n = 1$.

$$[z - G_1]V(z) = [2MF - G_1]\left[\frac{z}{z-1}\right] + zv(0). \tag{15}$$

A partial fraction expansion yields

$$V(z) = \left[\frac{2MF - G_1}{1 - G_1}\right]\left[\frac{z}{z-1}\right] + \left[v(0) - \frac{2MF - G_1}{1 - G_1}\right]\left[\frac{z}{z - G_1}\right]. \tag{16}$$

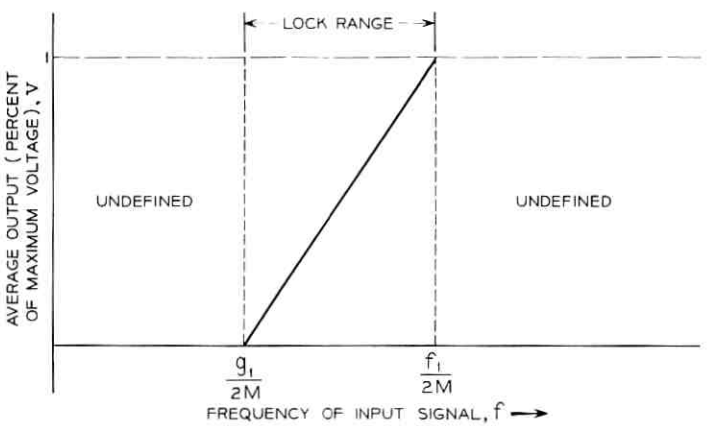


Fig. 3 — Average output voltage versus input frequency for first order PLL.

Assume that for time $t < 0$ the normalized input frequency is F_a. At $t = 0$ the input steps to F_b. Then from equation 12:

$$v(0) = \frac{2MF_a - G_1}{1 - G_1}. \tag{17}$$

Substitution into equation 16 gives the z transform of the response for $t \geqq 0$

$$V(z) = \left[\frac{2MF_b - G_1}{1 - G_1}\right]\left[\frac{z}{z - 1}\right] + \left[\frac{2M(F_b - F_a)}{1 - G_1}\right]\left[\frac{z}{z - G_1}\right]. \tag{18}$$

The inverse transform is easily found with the aid of equation 4a.

$$\begin{aligned} v(k) &= \frac{2MF_b - G_1}{1 - G_1} - \frac{2M(F_b - F_a)}{1 - G_1}(G_1)^k \\ &= \frac{2Mf_b - g_1}{f_1 - g_1} - \frac{2M(f_b - f_a)}{f_1 - g_1}\left(\frac{g_1}{f_1}\right)^k \quad \text{for} \quad k \geqq 0. \end{aligned} \tag{19}$$

Assuming that $v(k)$ is a continuous function of time, we let $v(k) = v(t)$ and $t = k\rho = k/2f_b$. This gives

$$v(t) = \frac{2Mf_b - g_1}{f_1 - g_1} - \frac{2M(f_b - f_a)}{f_1 - g_1} \exp\left[-2f_b t\right][\ln (f_1/g_1)]. \tag{20}$$

The resulting time constant is

$$T = \frac{1}{2f_b \ln (f_1/g_1)}, \tag{21}$$

and thus the half-power bandwidth is

$$f_c = \frac{1}{\pi} f_b \ln (f_1/g_1). \tag{22}$$

Notice the dependence of filter shaping upon input f_b. It follows that there is a somewhat different time constant for input frequency changes from f_b to f_a.

A simplified first order digital PLL has been built as shown in Fig. 4. It is a special case of the general system of Fig. 1. The FM input was composed of the discrete frequencies $f_a = 1070$ Hz and $f_b = 1270$ Hz. Clock frequencies g_1 and f_1 were chosen as $(2M)(970)$ Hz and $(2M)(1370)$ Hz, respectively, to provide a lock range of twice the total input frequency deviation. M was chosen to be 128 (see Section 5.3). The above values gave a time constant, T, of 1.13 ms. Interchanging f_a and f_b resulted in an increased time constant of 1.34 ms.

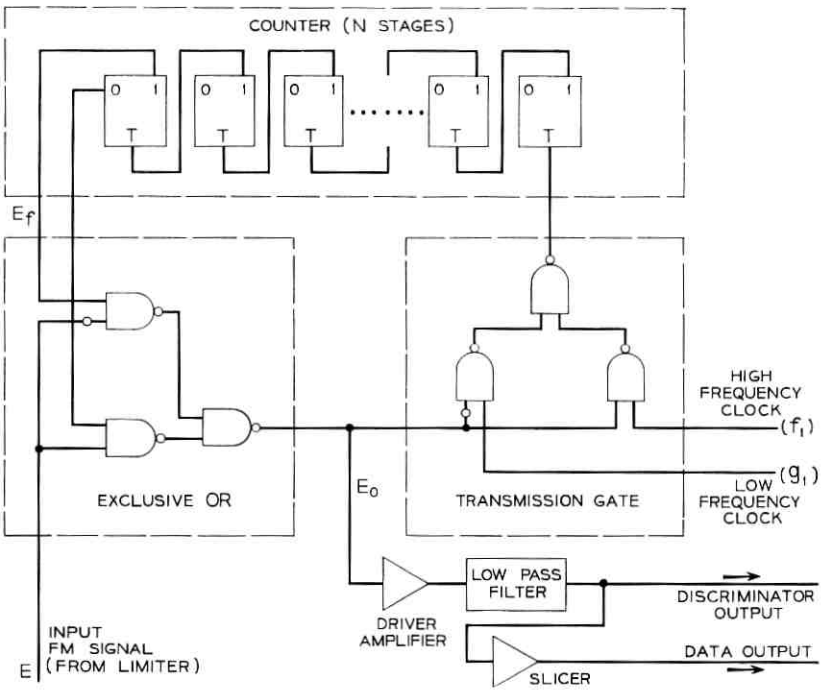


Fig. 4 — Block diagram of simplified first order PLL FM discriminator.

Oscilloscope photographs showing the two step responses of the actual loop are shown in Fig. 5.

5.2 *Out-of-Lock Oscillation*

If there is no input to the PLL, it will run at a rest frequency of

$$f_r = \left(\frac{M}{f_1} + \frac{M}{g_1}\right)^{-1} = \frac{1}{M}\frac{g_1 f_1}{g_1 + f_1}. \tag{23}$$

Notice, however, that one half cycle will be at $g_1/2M$ and the next at $f_1/2M$.

If an input to the PLL is present, but its frequency lies outside the lock range, that is, if $f < f_1/2M$ or $f > g_1/2M$, the output voltage $v(t)$ will oscillate between 0 and 1. Figure 6 shows waveforms for a loop which is out of lock. For time to the left of the dotted line, the loop is attempting to lock and the frequency of the feedback signal, E_f, is approaching that of the out-of-lock input E. Within this region,

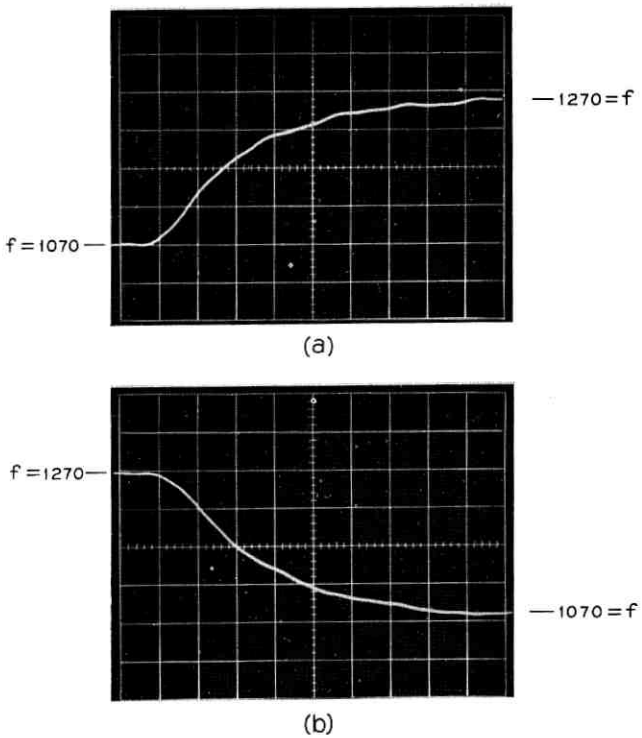


Fig. 5 — Output voltage response of first order PLL to input step frequency of (a) 1070 to 1270 Hz and (b) 1270 to 1070 Hz. Horizontal scale: 0.5 ms per cm.

equation 16 applies and is rewritten in the time domain as

$$v(k) = \left(\frac{2Mf - g_1}{f_1 - g_1}\right) + \left[v(0) - \frac{2Mf - g_1}{f_1 - g_1}\right]\left(\frac{g_1}{f_1}\right)^k. \tag{24}$$

For time to the right of the dotted line in Fig. 6, a new difference equation applies. It is written as

$$g_1[\rho(j) - \tau(j+1)] + f_1\tau(j) = M \tag{25}$$

so that a derivation similar to that previously used results in

$$v(j) = \left[\frac{g_1 - 2Mf}{g_1 - f_1}\right]\left[\left(\frac{f_1}{g_1}\right)^j - 1\right] + v'(0)\left(\frac{f_1}{g_1}\right)^j. \tag{26}$$

Equations 24 and 26 are represented graphically in Fig. 7. M, g_1, and f_1 take on the values of the previous example and an out-of-lock

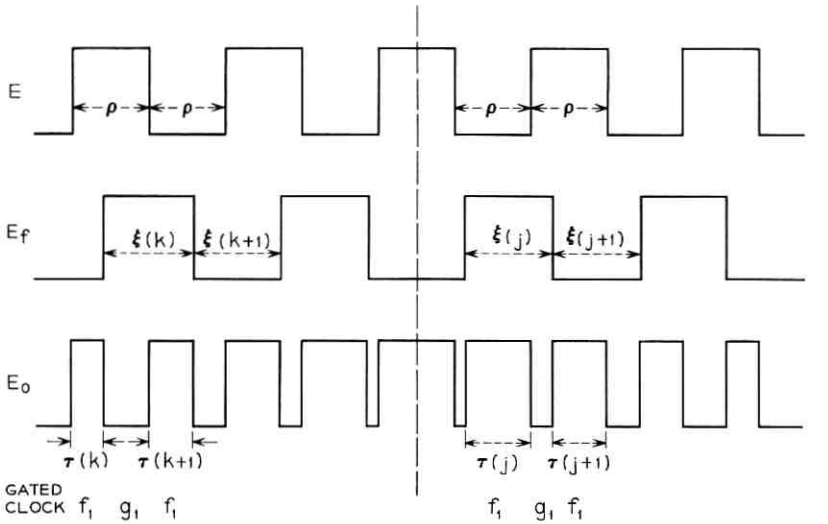


Fig. 6 — Input, feedback, and output waveforms of nonlocked first order PLL.

input frequency $f = 1400$ Hz is chosen as the input. The positive slope segment to the left corresponds to equation 24 with an initial voltage $v(0)$ assumed to be zero. When $v(k)$ reaches unity, the next segment is governed by $v(j)$ with an initial condition $v'(0) = 1$. When $v(j)$ reaches zero, the response is again given by $v(k)$ and with

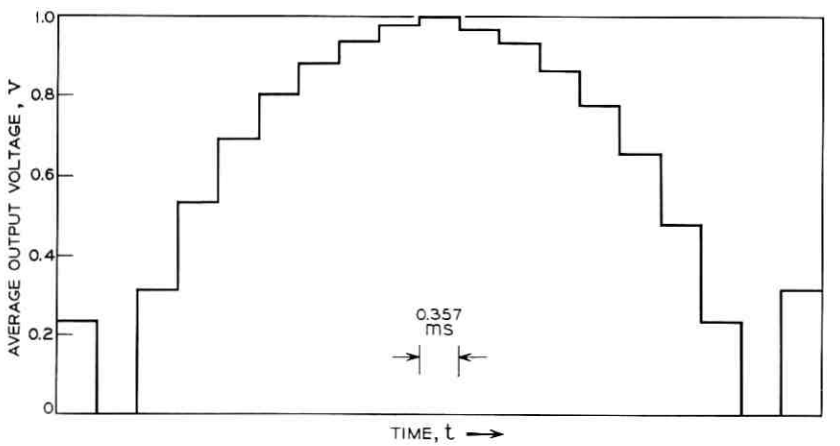


Fig. 7 — Single cycle of output average voltage oscillation for nonlocked first order PLL.

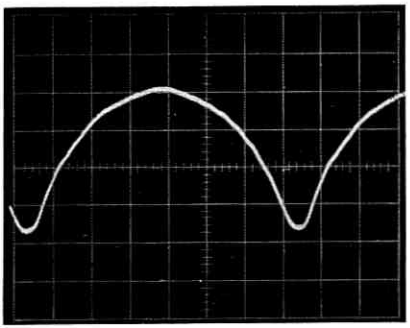


Fig. 8 — Actual output voltage of experimental first order PLL showing out-of-lock oscillation. Horizontal scale: 1 ms per cm.

initial condition adjusted accordingly. The actual waveform for the experimental PLL using this example is shown in Fig. 8.

5.3 *Internal Noise From Phase Discontinuity*

The clock signals are unsynchronized, and thus a random phase discontinuity results at the time of gating. This results in an internal noise voltage which exhibits itself as a fluctuation in the output voltage. This effect can be thought of as a quantization noise of the phase. A portion of Fig. 2 is redrawn in Fig. 9 to show the time errors ϵ_0, ϵ_0', and ϵ_1 at the instant of gating caused by differences in clock phases. It is assumed that the frequency of the ripple is low enough so as not to be attenuated by the PLL filtering property. With this in mind, difference equation 11 may be amended to account for the phase error.

$$[\tau(k+1) - \epsilon_1(k+1)]f_1 + [\rho(k) - \tau(k) - \epsilon_0(k) - \epsilon_0'(k)]g_1 + 2 = M. \quad (27)$$

For a constant input frequency and very slowly varying $\tau(k)$, $\tau(k+1) \cong \tau(k)$ so that

$$\tau(k)\,[f_1 - g_1] = M - g_1\rho + f_1\epsilon_1(k+1) + g_1[\epsilon_0(k) + \epsilon_0'(k)] - 2, \quad (28)$$

or

$$[v(k)]\,[f_1 - g_1] = 2Mf - g_1 + 2ff_1\epsilon_1(k+1) + 2fg_1[\epsilon_0(k) + \epsilon_0'(k)] - 4f. \quad (29)$$

Δv_N is defined as the difference in voltage for maximum and minimum (zero) phase errors. Thus

$$\Delta v_N = \left[\frac{2f}{f_1 - g_1}\right][f_1\epsilon_1(k+1)\,|_{\max} + g_1(\epsilon_0(k)\,|_{\max} + \epsilon_0'(k)\,|_{\max})]. \quad (30)$$

Since

$$\epsilon_1(k+1)\,|_{\max} = \frac{1}{f_1} \quad \text{and} \quad \epsilon_0(k)\,|_{\max} = \epsilon_0'(k)\,|_{\max} = \frac{1}{g_1},$$

then

$$\Delta v_N = \frac{6f}{f_1 - g_1} = \frac{6f_m}{f_1 - g_1}, \quad (31)$$

where f is approximated by the carrier f_m for the narrowband case.

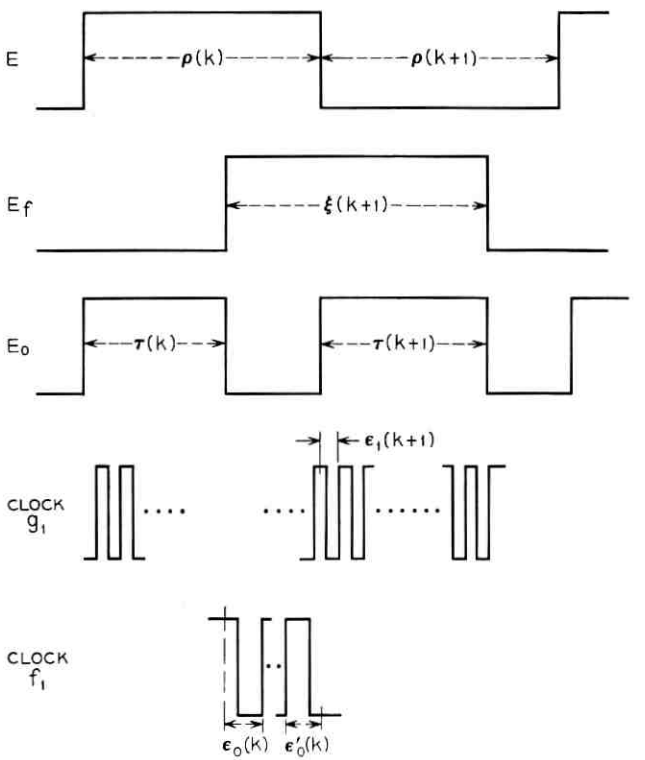


Fig. 9 — Waveforms for locked first order PLL showing how time jitter is caused by phase differences in clock signals.

It may be seen that the peak-to-peak output for an incoming FM wave with phase continuity in the clock signals is given by equation 12 as

$$\Delta v_{sig} = \left| \frac{2Mf_a - g_1}{f_1 - g_1} - \frac{2Mf_b - g_1}{f_1 - g_1} \right| = 2M \frac{\Delta f}{f_1 - g_1}, \tag{32}$$

where Δf is the total frequency deviation $| f_a - f_b |$. Thus the minimum peak-to-peak voltage signal-to-noise ratio within the loop is

$$S/N = \frac{\Delta v_{sig}}{\Delta v_N} = \frac{2M(\Delta f)}{6f_m} = 2^N \frac{\Delta f}{6f_m}. \tag{33}$$

Continuing the previous example with $\Delta f = 200$Hz and $f_m = 1170$ Hz, it is desired that the internal noise be 20 dB below the signal so that $2^N \geqq 361$. This suggests a nine stage counter ($2^9 = 512$); however, it should be pointed out that a worst case of phase jitter has been assumed and the average phase jitter is less than this. Laboratory experiments have shown satisfactory results with an eight stage counter thus giving a minimum signal-to-noise ratio of 17.2 dB.

It is apparent from equation 33 that the internal noise may be made as low as desired by choosing a sufficiently large value of N. Thus a direct relationship exists between equipment cost (number of counters) and performance (jitter distortion).

VI. SECOND ORDER DIGITAL PLL

The location of the desired s-plane poles must be specified and mapped into the z-plane. The following example is concerned with synthesizing the familiar Butterworth response, but the procedure is certainly applicable to other filter classes.

A second order PLL is to be synthesized so that its response is that of a Butterworth low-pass filter with cutoff at ω_c. The $2n = 4$ poles of $H(s)H(-s)$ lie on a circle of radius ω_c and subtend equal arcs such that the filter's characteristic equation is

$$s^2 + (2)^{\frac{1}{2}}\omega_c s + \omega_c^2 = 0, \tag{34}$$

with poles at

$$s_{1,2} = \frac{\omega_c}{(2)^{\frac{1}{2}}}(-1 \pm i). \tag{35}$$

The corresponding poles in the z plane are

$$z_{1,2} = \exp(s_{1,2}\rho) = [\exp(-\alpha)][\cos(\alpha) \pm i \sin(\alpha)]$$

where

$$\alpha = \frac{\omega_c \tau}{(2)^{\frac{1}{2}}} = \frac{\pi}{(2)^{\frac{1}{2}}} \frac{f_c}{f}. \tag{36}$$

Digressing for a moment, it is of academic interest to map the entire Butterworth circle into the z plane. This is easily done by letting s be a circle of radius ω_c, that is, $s = \omega_c \exp(i\phi)$, $0 \leqq \phi \leqq 2\pi$. Substituting $z = \exp(s\rho)$ and approximating f by f_m gives

$$z = \left\{\exp\left[\frac{\omega_c}{2f_m}\cos(\phi)\right]\right\}\left\{\cos\left(\frac{\omega_c}{2f_m}\sin(\phi)\right) + i\sin\left(\frac{\omega_c}{2f_m}\sin(\phi)\right)\right\}. \tag{37}$$

The resulting cardioid-like shape is shown in Fig. 10 with f_c/f_m as a parameter. The angle ϕ is also shown so that portions of the Butterworth circle may be conveniently translated into the z plane.

Returning to the specific example, the characteristic polynominal is written using equation 36 as

$$(z - z_1)(z - z_2) = z^2 - [2\exp(-\alpha)][\cos(\alpha)]z + \exp(-2\alpha). \tag{38}$$

The resulting coefficients are equated to the respective coefficients in equation 6, namely $z^2 + F_1 z - G_1$ so that:

$$F_1 = f_1/f_2 = -2[\exp(-\alpha)][\cos\alpha] \tag{39}$$

and

$$G_1 = g_1/f_2 = -\exp(-2\alpha).$$

A third equation is arrived at by fixing the steady-state voltage $v(k)$ for a specific input frequency, f. The choice is arbitrary, and since the system's input spectrum is symmetrical about the carrier frequency, f_m, it is reasonable to fix the corresponding output voltage at 0.5. The equation is obtained from equation 6 by noting that steady-state implies periodicity so that $v(k) = v(k+j)$. Thus:

$$v(k)[1 + F_1 - G_1] = 2M\frac{f_m}{f_2} - G_1 . \tag{40}$$

Solving equations 39 and 40 gives

$$f_2 = \frac{2Mf_m \exp(\alpha)}{\cos(\alpha) - \sinh(\alpha)}, \tag{41}$$

$$f_1 = \frac{4Mf_m \cos(\alpha)}{\cos(\alpha) - \sinh(\alpha)}, \tag{42}$$

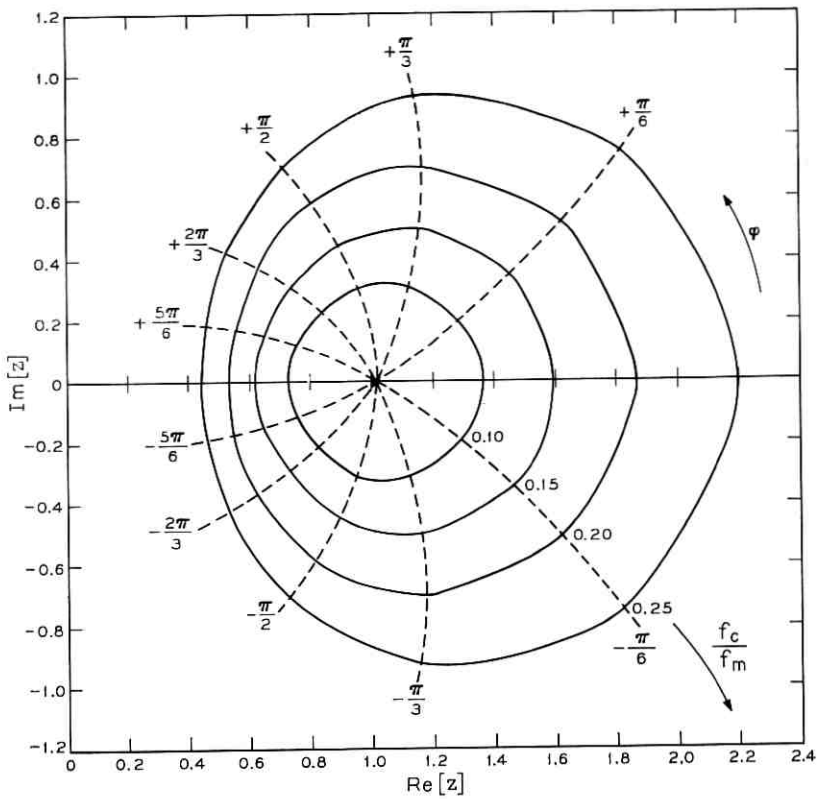


Fig. 10 — Mapping of the s-plane Butterworth circle into the z-plane.

and

$$g_1 = \frac{2Mf_m \exp(-\alpha)}{\cos(\alpha) - \sinh(\alpha)}. \tag{43}$$

As a practical example, consider the demodulation of a narrowband FSK wave whose spectrum is centered about a carrier of 2125 Hz. The signaling rate is limited to $\leqq 300$ baud so that a cutoff frequency of 250 Hz should prove adequate. As previously noted, the choice of M is dependent upon the maximum baseband jitter (quantizing noise) which can be tolerated. This must be weighed against the added circuitry and higher clock frequencies imposed by large values of M. A convenient choice is $M = 128$. Thus

$$\alpha = \frac{\pi}{(2)^{\frac{1}{2}}} \frac{f_c}{f_m} = 0.262$$

so that

$$g_1 = 598\ \text{kHz},$$

$$f_1 = 1504\ \text{kHz},$$

and

$$f_2 = -1012\ \text{kHz}.$$

As previously mentioned, the negative value for f_2 requires a "down counter" for the second register. This example is continued in subsequent sections.

6.1 *Step Response of the System*

Once the coefficients $g_1, f_1, \ldots, f_n$ are determined, the characteristic equation is uniquely specified and the system's time response to a step in input frequency is easily found. The step response of a second order system is now derived in the interest of providing further insight into the characteristics of the loop.

Equation 6 is written for $n = 2$ as:

$$[z^2 + F_1 z - G_1]V(z) = (2MF - G_1)\left(\frac{z}{z-1}\right) + v(1)z + v(0)z^2 + zF_1 v(0). \quad (44)$$

Assume that the input frequency steps from F_a to F at $t = 0$ (as before, capitalization denotes normalization so that $F_i = f_i/f_2$, $G_i = g_i/f_2$). Thus for $t \geqq 0$ we have

$$V(z) = (2MF - G_1)\left[\frac{z}{(z^2 + F_1 z - G_1)(z-1)}\right] + v(0)\left[\frac{z^2}{z^2 + F_1 z - G_1}\right] + [F_1 v(0) + v(1)]\left[\frac{z}{z^2 + F_1 z - G_1}\right]. \quad (45)$$

The initial conditions $v(1)$ and $v(0)$ are determined from the difference equation

$$v(k+2) + F_1 v(k+1) - G_1 v(k) = 2MF - G_1. \quad (46)$$

Assume that the system is in steady state for $t \leqq 0$ so that $v(k) = v(k-i)$, $k \leqq 0$. For $k = -2$

$$v(0)[1 + F_1 - G_1] = 2MF_a - G_1, \quad (47)$$

so that

$$v(0) = \frac{2MF_a - G_1}{1 + F_1 - G_1}. \tag{48}$$

Now let $k = -1$:

$$v(1) = 2MF - G_1 - [F_1 - G_1]\, v(0), \tag{49}$$

which may be written as

$$v(1) = v(0) + 2M\Delta, \tag{50}$$

where

$$\Delta = [F - F_a]. \tag{51}$$

Substituting equations 50 and 51 into 45 and simplifying leads to

$$\begin{aligned} V(z) = &\left[v(0) + \frac{2M\Delta}{1 + F_1 - G_1}\right]\left[\frac{z}{z - 1}\right] \\ &- \left[\frac{2M\Delta}{1 + F_1 - G_1}\right]\left[\frac{z^2}{z^2 + F_1 z - G_1}\right] \\ &- \left[\frac{2M\Delta G_1}{1 + F_1 - G_1}\right]\left[\frac{z}{z^2 + F_1 z - G_1}\right]. \end{aligned} \tag{52}$$

The following substitutions are made so that $V(z)$ may be easily transformed. Let

$$F_1 = -2[\cos \beta] \exp(-\alpha); \qquad G_1 = -\exp(-2\alpha). \tag{53}$$

Substituting and rearranging gives

$$\begin{aligned} V(z) = &\left[v(0) + \frac{2M\Delta}{1 + F_1 - G_1}\right]\left[\frac{z}{z - 1}\right] \\ &- \left[\frac{2M\Delta}{1 + F_1 - G_1}\right]\left[\frac{z^2 - z(\cos \beta) \exp(-\alpha)}{z^2 - 2z(\cos \beta) \exp(-\alpha) + \exp(-2\alpha)}\right] \\ &- \left[\left(\frac{4M\Delta G_1}{1 + F_1 - G_1} - F_1\right)(-4G_1 - F_1)^{-\frac{1}{2}}\right] \\ &\cdot\left[\frac{z(\sin \beta) \exp(-\alpha)}{z^2 - 2z(\cos \beta) \exp(-\alpha) + \exp(-2\alpha)}\right]. \end{aligned} \tag{54}$$

This is transformed into

$$v(k) = \left[v(0) + \frac{2M\Delta}{1 + F_1 - G_1}\right]u(k)$$

$$- \exp(-\alpha k)\left[\left(\frac{2M\Delta}{1 + F_1 - G_1}\right)\cos(\beta k) + \left(\frac{4M\Delta G_1}{1 + F_1 - G_1} - F_1\right)(-4G_1 - F_1)^{-\frac{1}{2}}\sin(\beta k)\right], \quad (55)$$

where $u(k)$ is the unit step. The time response is found by again assuming that $v(k)$ is a continuous function of time $v(t)$, $k \approx 2ft$. Substitution gives the familiar response of an underdamped second order system:

$$v(t) = \left[v(0) + \frac{2M\Delta}{1 + F_1 - G_1}\right]u(t) - \exp(-2\alpha ft)\left[\left(\frac{2M\Delta}{1 + F_1 - G_1}\right)\cos(2\beta ft) + \left(\frac{4M\Delta G_1}{1 + F_1 - G_1} - F_1\right)(-4G_1 - F_1)^{-\frac{1}{2}}\sin(2\beta ft)\right]. \quad (56)$$

The experimentally determined step response of the actual $n = 2$ PLL closely approximates the Butterworth response and is shown in Fig. 11.

6.2 *Lock Range for the Second Order System*

As was the case for the $n = 1$ PLL, the loop is said to be in "lock" if, for a steady input frequency, the output $v(k)$ is constant. Although there are many frequency ranges for which the loop exhibits this property, there is a "fundamental" lock range over which $v(k)$ varies linearly with frequency between its defined limits, and the output frequency equals the input frequency. In steady state, $v(k+2) = v(k+1) = v(k)$ so that from equation 40:

$$v(k) = \frac{2MF - G_1}{1 + F_1 - G_1} = \frac{2Mf - g_1}{f_2 + f_1 - g_1} \quad \text{if} \quad 0 \leqq v(k) \leqq 1. \quad (57)$$

The end points of the lock range are

$$f\,|_{v(k)=0} \equiv f_u = \frac{g_1}{2M} \quad (58)$$

and

$$f\,|_{v(k)=1} \equiv f_l = \frac{f_1 + f_2}{2M}. \quad (59)$$

The lock range is given by

$$\Delta f_L = |f_u - f_l| = \frac{1}{2M}|g_1 - f_1 - f_2|. \quad (60)$$

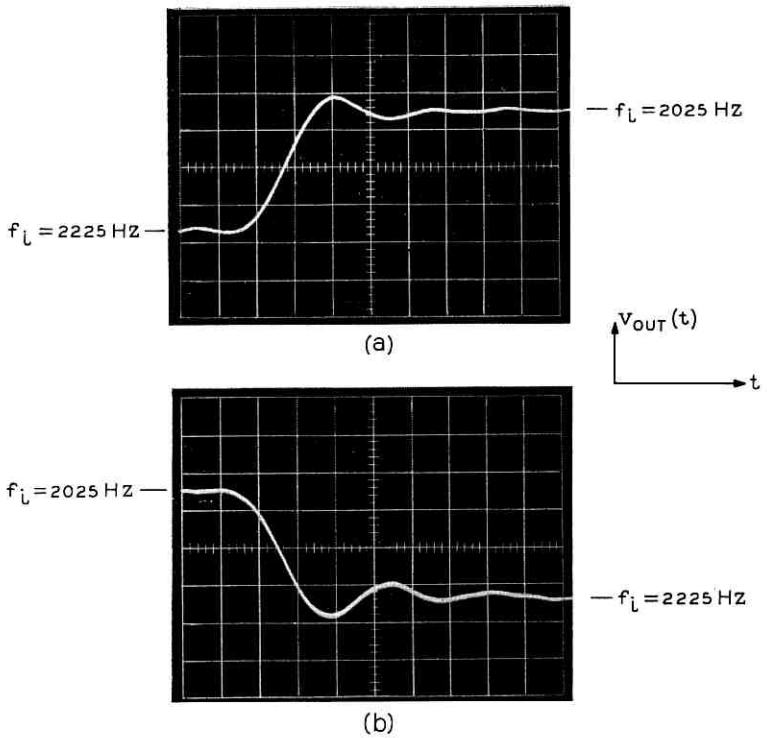


Fig. 11 — Output voltage response of second order PLL to input step frequency of (a) 2225 to 2025 Hz and (b) 2025 to 2225 Hz. Horizontal scale: 1 ms per cm.

Substituting the values calculated previously gives a lock range of 414 Hz which is approximately twice the bandwidth required. The useful bandwidth is reduced still further as will be shown.

6.3 *Bound on Cutoff Frequency*

For a step in frequency $f_a - f_b$ we have a corresponding step in voltage,

$$\Delta v = \frac{2Mf_a - g_1}{f_2 + f_1 - g_1} - \frac{2Mf_b - g_1}{f_2 + f_1 - g_1} = \frac{2M(f_a - f_b)}{f_2 + f_1 - g_1}. \tag{61}$$

Δv is defined to have a maximum value of unity so

$$f_2 + f_1 - g_1 \geqq 2M(f_a - f_b) \tag{62}$$

and

$$1 + F_1 - G_1 \leqq \frac{2M}{f_2}(f_a - f_b) \quad \text{for} \quad f_2 < 0. \tag{63}$$

If f_a and f_b are symmetrically distributed about the carrier frequency, f_c, we can let $v = \frac{1}{2}$ at $f_1 = f_m$ so that from equation 40

$$f_2 = \frac{4Mf_m}{1 + F_1 + G_1}. \tag{64}$$

Substituting equation 64 into 63 gives

$$1 + F_1 - G_1 \leqq (1 + F_1 + G_1)\left(\frac{f_a - f_b}{2f_m}\right). \tag{65}$$

Substituting equation 53 for F_1 and G_1 gives

$$1 - 2[\exp(-\alpha)]\cos\beta + \exp(-2\alpha)$$
$$\leqq \left(\frac{f_a - f_b}{2f_m}\right)[1 - 2(\cos\beta)\exp(-\alpha) - \exp(-2\alpha)]. \tag{66}$$

The constants α and β are related to the filter shaping desired and are a function of the input frequency for $t > 0$. For the case of a Butterworth response ($n = 2$)

$$\alpha = \beta = \frac{\pi}{(2)^{\frac{1}{2}}}\left(\frac{f_c}{f}\right).$$

For a narrowband system the input frequency may be approximated by the carrier f_m. Thus

$$1 - 2(\cos\mu)\exp(-\mu) + \exp(-2\mu)$$
$$\leqq [1 - 2(\cos\mu)\exp(-\mu) - \exp(-2\mu)]\left[\frac{\Delta f}{2f_m}\right], \tag{67}$$

where

$$\mu = \frac{\pi}{(2)^{\frac{1}{2}}}\frac{f_c}{f_m}.$$

This may be written as

$$\cos\mu \geqq \frac{\exp(\mu) - \left[\dfrac{f_m + \dfrac{\Delta f}{2}}{f_m - \dfrac{\Delta f}{2}}\right]\exp(-\mu)}{2}. \tag{68}$$

Thus for a given f_a, f_b and f_m, the cutoff frequency f_c is bounded. For example let f_a and f_b be 2025 Hz and 2225 Hz, respectively, and be symmetrically distributed about the carrier f_m. Then

$$\cos \mu \geqq \frac{\exp(\mu) + (0.91) \exp(-\mu)}{2}. \tag{69}$$

Solving this transcendental equation gives

$$\mu \geqq 0.19 \tag{70}$$

and

$$f_c \geqq 182 \text{ Hz}.$$

An upper bound on f_c is found by requiring that the end points of the lock range be positive frequencies. The lower edge of the lock range is required to be greater than zero so that from equations 59, 41 and 42

$$\frac{f_1 + f_2}{2M} \geq 0 \quad \text{or} \quad 2 \cos \mu - \exp(\mu) \geqq 0. \tag{71}$$

This is satisfied if $\mu \leqq 0.54$ which gives positive value for equation 71 as well as for the upper edge of the lock range $g_1/2M$. Thus $0.19 \leqq \mu \leqq 0.54$ so that for $f_m = 2125$ Hz

$$182 \leqq f_c \leqq 542 \text{ Hz}. \tag{72}$$

The bounds on f_c/f_m are loose in the sense that it is assumed that the full lock range is available to input frequencies. This is not the case for underdamped systems, where a hysteresis effect known as "capture" reduces the effective lock range available, and so the bound might be tightened accordingly.

6.4 *The Capture Phenomonen*

When the PLL is "out of lock" its output voltage oscillates between 1 and 0. The "pull-in" frequencies are those frequencies furthest removed from the carrier for which the PLL will ultimately lock. In general, the PLL will exhibit a hysteresis so that the pull-in range will be smaller than the lock range; also, the upper and lower pull-in frequencies will generally not be symmetrically distributed about the carrier f_m. In general, solution of the capture range (for conventional phase-locked-loops) results in a nonlinear integrodifferential equation. Solution for conventional PLL's requires phase plane techniques

and is documented in Ref. 2. An analysis to determine capture range for the digital PLL has not been performed.

6.5 *Stability of the Second Order System Within the Lock Range*

Thus far there has been no restriction placed on N, the number of counters required per shift register. For N less than some critical value N_m it is possible for stable modes of operation, different from that of Fig. 2, to exist. These modes do not exploit the PLL to its fullest advantage and thus in previous derivations it has been assumed that $N \geqq N_m$.

A possible parasitic mode is shown in Fig. 12. The input frequency is constant, as is the frequency of the flip-flop signal, E_f, but the waveform of E_f is no longer square. This operation is caused by the limited length of the registers. Assume that the lengths (number of counter stages) for the $n = 2$ system are equal, and of value N. Each time the feedback flip-flop (see Fig. 1) is triggered, register 1 is reset to zero. During the period T_1+T_2 (see Fig. 12) this register is advanced to a count of $[T_1g_1+T_2f_1]$ modulo C where $C = 2^N$. This number is shifted into the second register where one of two conditions may exist:

$$T_1g_1 + T_2f_1 \geqq M,$$

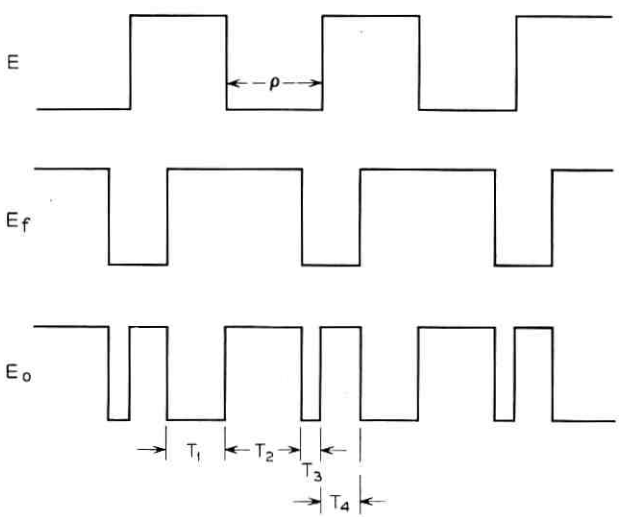


Fig. 12 — Waveforms for parasitic locked mode of second order PLL.

or

$$T_1 g_1 + T_2 f_1 \leqq M.$$

The first equation results in normal operation, analyzed in previous sections. During T_4 the count is reduced by a negative f_2 so that

$$[T_1 g_1 + T_2 f_1 + T_4 f_2] \bmod C = M.$$

The second equation requires that the second register count down to zero (which is congruent to C) and then further reduce the count by $C - M$. Thus

$$[T_1 g_1 + T_2 f_1 + T_4 f_2] \bmod C = M - C.$$

These equations may be combined by introducing the parameter $\delta = -1, 0$ depending upon the mode of operation:

$$T_1 g_1 + T_2 f_1 + T_4 f_2 = M - \delta C. \tag{73}$$

Similarly, an equation similar in form to equation 73 covers the intervals T_2, T_3, and T_4:

$$T_3 g_1 + T_4 f_1 + T_2 f_2 = M - \gamma C, \tag{74}$$

where $\gamma = 0, 1$.

The period, T_2, may be explicitly solved for by noting that

$$T_1 + T_4 = T_2 + T_3 = \rho. \tag{75}$$

Thus

$$v_2 = \frac{T_2}{\rho} = -\frac{g_1}{f_1 + f_2 - g_1} + \frac{(-\delta C + M) f_1 - (-\gamma C + M)(f_2 - g_1)}{f_1^2 - (f_2 - g_1)^2} 2f. \tag{76}$$

Under normal conditions $\delta = \gamma = 0$ so that

$$v_2 = \frac{2Mf - g_1}{f_1 + f_2 - g_1},$$

which is identical to equation 57. The parasitic mode of operation is given by $\delta = -1$, $\gamma = 1$ so that equation 76 may be rewritten as

$$v_2 = \frac{2Mf - g_1}{f_1 + f_2 - g_1} + \frac{2Cf}{f_1 + f_2 - g_1}. \tag{77}$$

Graphical examination (see Fig. 13) of equations 57 and 77 shows that

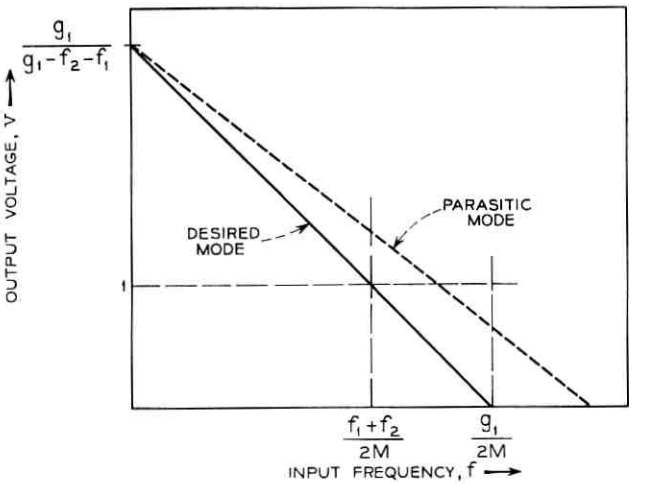


Fig. 13 — Comparison of parasitic and normal mode voltage versus frequency characteristic for second order PLL.

two modes of operation may exist within the lock range. To insure that this will not happen, C must be chosen so that the frequency range for which $0 \leqq v_2 \leqq 1$ lies outside the normal lock range. Thus at the upper lock range edge, $f = g_1/2M$, it is required that $v_2 > 1$ so that from equation 77,

$$\frac{2C\left(\frac{g_1}{2M}\right)}{f_1 + g_1 - f_2} > 1.$$

Hence

$$\frac{C}{M} > \frac{f_1 + g_1 - f_2}{g_1}$$

and thus the minimum number of counters required is

$$N_m = 1 + \text{integer value}\left[\log_2\left(\frac{f_1 + g_1 - f_2}{g_1}\right)M\right]^* \tag{78}$$

VII. PERFORMANCE

In a laboratory test, an FM digital signal with mark and space frequencies of 1270 and 1070 Hz was fed into an $n = 1$ PLL which had

* That is, the integer value of 2.3 is 2.

an 8-stage counter and clock frequencies for which $g_1 = 2M(970)$Hz and $f_1 = 2M(1370)$Hz. The output of the exclusive-OR was fed to an amplifier which clipped the signal to precise levels and provided a constant output impedance. The signal then went to a Butterworth low-pass filter with $n = 3$ and cutoff of 200 Hz. The filtered signal was sliced to give a digital output which could be compared with the original digital input. The FM modulator used in these tests was from a Bell System Data Set 103E1. The eye pattern for this circuit with an input at 300 baud is given in Fig. 14a. When the lock range was increased to $\pm$300 Hz about the carrier, the eye pattern improved, as shown in Fig. 14b. This effect results from the increased bandwidth, f_e.

Start-stop distortion, as measured with a Bell System 911A data test set for bit rates of 150 to 300 baud, was 2 and 5 percent, respectively. Performance with additive gaussian noise is as shown in Fig. 15,

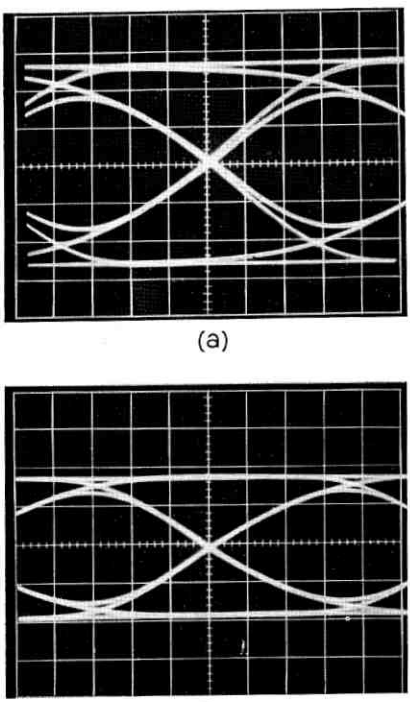


Fig. 14 — First order PLL eye patterns for 300 baud random digital data. Lock range of (a) 400 Hz, (b) 600 Hz. Horizontal scale: 0.5 ms per cm.

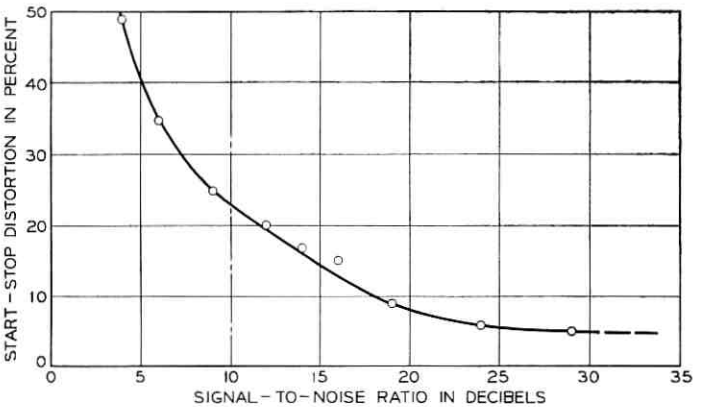
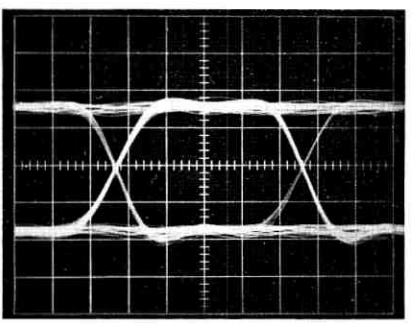
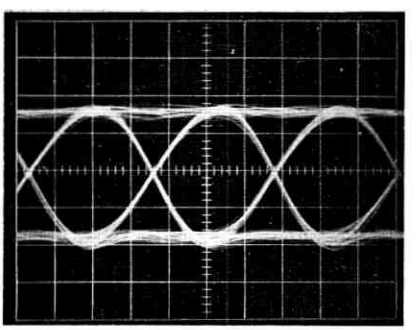


Fig. 15 — Distortion performance of first order PLL with input signal degraded by 3 kHz band-limited white noise.



(a)



(b)

Fig. 16 — Second order PLL eye patterns for (a) 150 baud and (b) 300 baud random digital data. Horizontal scale: 1 ms per cm.

where start-stop distortion is plotted versus signal-to-noise ratio. The noise is 3 kHz band-limited white noise and the bit rate is 300 baud. This performance is similar to that of the receiver in Data Set 103E1. Start-stop distortion for a single frequency (2025 Hz) interference was measured and, as might be expected, single frequency interference is a strong function of frequency, the worst values being at odd multiples of the channel frequency.

Laboratory tests indicated that the second order PLL, with characteristics described in Section VI, performed in accordance with theoretical expectation. The eye patterns for an input of a modulated random data signal at 150 and 300 baud are shown in Fig. 16. It was found that noise performance of the second order system was not much better than that of the first order system. This probably resulted from the counter length falling just short of the value indicated in equation 78, thereby allowing noise perturbations to randomly shift operation between the stable and parasitic modes of operation.

VIII. CONCLUSIONS

A synthesis procedure for the realization of an nth order digital phase-locked loop has been described. Such systems find application in FM demodulators, filters, and in extremely stable locked oscillators. Various loop properties have been theoretically derived and experimental performance has been found to be consistent with these results. The advantages of such circuits for use in large multichannel data sets are:

(*i*) Filtering property—(6n) dB per octive.

(*ii*) Small size—completely integrable using one or more monolithic chips.

(*iii*) Requires no adjustment—permitting lower manufacture and repair costs.

(*iv*) Excellent stability and reliability—the PLL circuit either works or does not, since it is completely digital. Stability of the entire system is dependent upon clock stability which may be as good as required.

(*v*) Multichannel economy—accurate clocks can be used to drive many channel circuits.

In addition, this circuit has two inherent advantages over other types of phase-locked loops. First, it requires no low-pass filter generally found between the phase comparator (multiplier) and voltage-

controlled oscillator in conventional phase-locked loops. Second, it includes, in effect, an ideal voltage-controlled oscillator, the frequency of which is linearly related to voltage.

REFERENCES

1. Byrne, C. J., "Properties and Design of the Phase-Controlled Oscillator with a Sawtooth Comparator," B.S.T.J., *41*, No. 2 (March 1962), pp. 559–602.
2. Moschytz, G. S., "Miniaturized RC Filters Using Phase-Locked Loop," B.S.T.J., *44*, No. 5 (May–June 1965), pp. 823–870.
3. Rader, C. M. and Gold B., "Digital Filter Design Techniques in the Frequency Domain," Proc. IEEE, *55*, No. 2 (February 1967), pp. 149–171.
4. Kaiser, J. F. and Golden, R. M., "Design of Wideband Sampled Data Filters," B.S.TJ, *43*, No 4, part 2 (July 1964), pp 1533–1546.
5. Jury, E. I., *Theory and Application of the Z-Transform Method*, New York: John Wiley, 1964.

First and Second Passage Times of Sine Wave Plus Noise

By A. J. RAINAL

(Manuscript received June 25, 1968)

This paper is concerned with the first and second passage times of a stationary random process, $I(t, a)$, consisting of a sinusoidal signal of amplitude $(2a)^{\frac{1}{2}}$ plus stationary Gaussian noise with a finite expected number of zeros per unit time. This type of random process is present at the output of the IF amplifier of a radio or radar receiver during the reception of a sinusoidal signal immersed in Gaussian noise. Approximate integral equations are developed whose solutions yield approximate probability densities concerning the first and second passage times of $I(t, a)$. The resulting probability functions are presented in graphs for the case when the frequency of the sine wave is located in the center of a band of noise. Related results concerning the approximate distribution function of the absolute minimum or absolute maximum of $I(t, a)$ in the closed interval $[0, \tau]$ are also presented.

I. INTRODUCTION

Exact, explicit, results concerning the first passage times of a Markov or "Markov-like" random process have been given by many authors.[1-7] But very little is known about the first passage times of a random process consisting of a sinusoidal signal plus stationary gaussian noise. This random process is of interest because it serves as a realistic model for the output of the IF amplifier of a typical radio or radar receiver during the reception of a sinusoidal signal immersed in Gaussian noise.

Let $I(t, a)$ denote the stationary random process consisting of a sinusoidal signal of amplitude $(2a)^{\frac{1}{2}}$ and angular frequency q plus stationary gaussian noise, $I_N(t)$, of zero mean and unit variance. Thus,

$$I(t, a) = (2a)^{\frac{1}{2}} \cos (qt + \theta_0) + I_N(t). \tag{1}$$

θ_0 denotes a random phase angle which is distributed uniformly in the interval $(-\pi, \pi)$. a denotes the signal-to-noise power ratio.

The first and second passage times of $I(t, a)$ are indicated in Fig. 1 and are defined as:

(*i*) τ^+ represents the time $I(t, a)$ takes in going from an upcrossing of the level I_1 to the first crossing of the level $I_2 < I_1$.

(*ii*) τ^- represents the time $I(t, a)$ takes in going from a downcrossing of the level I_1 to the first crossing of the level $I_2 < I_1$.

(*iii*) τ_1^+ represents the time $I(t, a)$ takes in going from an upcrossing of the level I_1 to the second crossing of the level $I_2 < I_1$.

(*iv*) τ_1^- represents the time $I(t, a)$ takes in going from a downcrossing of the level I_1 to the second crossing of the level $I_2 < I_1$.

For fixed I_1, $I_2 < I_1$, and a we denote the probability densities of τ^+, τ^-, τ_1^+, and τ_1^- by $W^+(\tau, I_1, I_2, a)$, $W^-(\tau, I_1, I_2, a)$, $W_1^+(\tau, I_1, I_2, a)$, and $W_1^-(\tau, I_1, I_2, a)$, respectively. These four probability densities arise in many branches of science and technology.

Because the random process $I(t, a)$ is symmetrical about its mean value of zero, we need only discuss the case when $I_1 > I_2$ as is indicated in Fig. 1. The case when $I_1 < I_2$ can always be converted to the case under discussion by considering the random process $-I(t, a)$.

For the general process $I(t, a)$, exact, explicit expressions for the four probability densities are unknown. However, as already mentioned, exact, explicit results concerning the first passage times of $I(t, a)$ are known for the special cases when $a = 0$ and $I_N(t)$ is a Markov or Markov-like random process.

The purpose of this paper is to present theoretical approximations for the four probability densities of the first and second passage times of

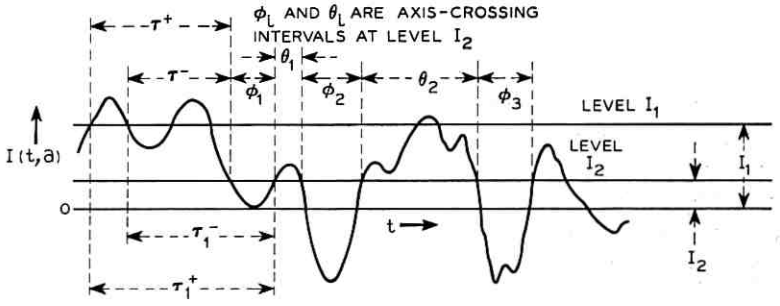


Fig. 1 — τ^+ and τ^- are the first passage times of $I(t, a) = (2a)^{1/2} \cos(qt + \theta_0) + I_N(t)$ defined by the levels I_1 and I_2. Similarly, τ_1^+ and τ_1^- are the second passage times.

$I(t, a)$ for the cases when $a \geq 0$ and $I_N(t)$ is a stationary, gaussian process having a finite expected number of zeros per unit time.

II. AUXILIARY PROBABILITY FUNCTIONS

Using a notation consistent with Ref. 8, we define the following auxiliary probability functions concerning the stationary random process $I(t, a)$:

(*i*) $P_2^{+-}(\tau, I_1, I_2, a)d\tau$, the conditional probability that a downward crossing of the level I_2 occurs between $t + \tau$ and $t + \tau + d\tau$ given an upward crossing of the level I_1 at t.

(*ii*) $P_2^{-+}(\tau, I_1, I_2, a)d\tau$, the conditional probability that an upward crossing of the level I_2 occurs between $t + \tau$ and $t + \tau + d\tau$ given a downward crossing of the level I_1 at t.

(*iii*) $P_2^{--}(\tau, I_1, I_2, a)d\tau$, the conditional probability that a downward crossing of the level I_2 occurs between $t + \tau$ and $t + \tau + d\tau$ given a downward crossing of the level I_1 at t.

(*iv*) $P_2^{++}(\tau, I_1, I_2, a)d\tau$, the conditional probability that an upward crossing of the level I_2 occurs between $t + \tau$ and $t + \tau + d\tau$ given an upward crossing of the level I_1 at t.

These auxiliary probability functions were given in Ref. 8 for the case when $I_1 = I_2$. Here, we need to merely generalize to the case when $I_1 \neq I_2$. The reader should see Rice's work for the definition of all notation which is not defined in this paper.[4] When $a \geq 0$ and $I_1 \neq I_2$, Rice's equation 38 generalizes to:

$$P_2^{+-}(\tau, I_1, I_2, a) = -[2\pi N_{I_1}]^{-1} \int_{-\pi}^{\pi} d\theta \int_0^{\infty} dI_1' \int_{-\infty}^{0} dI_2' I_1' I_2' \, p(I_1, I_1', I_2', I_2) \qquad (2)$$

where

N_{I_1} = Rice's equation 2.7 of Ref. 9 for the expected number of upcrossings (or downcrossings) of the level I_1 per second

$$= \frac{(\beta)^{\frac{1}{2}} \exp(-I_1^2/2)}{2\pi} \sum_{n=0}^{\infty} \frac{(-1)^n (2n)!}{2^n (n!)^3} \left(\frac{a}{2}\right)^n \cdot {}_1F_1\left(-n; \frac{1}{2}; \frac{I_1^2}{2}\right) {}_1F_1\left(-\frac{1}{2}; n + 1; -\frac{aq^2}{\beta}\right)$$

${}_1F_1$ = the confluent hypergeometric function

$$p(I_1, I_1', I_2', I_2) = (2\pi)^{-2}M^{-\frac{1}{2}}$$

$$\cdot \exp\left\{-\frac{1}{2M}\left[M_{22}(I_1'^2 + I_2'^2) + 2M_{22}r_1 I_1' I_2' + 2D_2 I_1' + 2E_2 I_2' + F_2\right]\right\}$$

$$r_1 = \frac{M_{23}}{M_{22}} \qquad Q = (2a)^{\frac{1}{2}}$$

$$D_2 = M_{12}(I_1 - Q\cos\theta) + M_{13}[Q\cos(q\tau + \theta) - I_2] + M_{22}Qq\sin\theta + M_{23}Qq\sin(q\tau + \theta)$$

$$E_2 = M_{12}[Q\cos(q\tau + \theta) - I_2] + M_{13}(I_1 - Q\cos\theta) + M_{22}Qq\sin(q\tau + \theta) + M_{23}Qq\sin\theta$$

$$F_2 = M_{11}\{I_1^2 + I_2^2 - 2Q[I_1\cos\theta + I_2\cos(q\tau + \theta)] + Q^2[\cos^2\theta + \cos^2(q\tau + \theta)]\}$$
$$+ 2M_{12}Qq\{[I_1 - Q\cos\theta]\sin\theta + [Q\cos(q\tau + \theta) - I_2]\sin(q\tau + \theta)\}$$
$$+ 2M_{13}Qq\{[I_1 - Q\cos\theta]\sin(q\tau + \theta) + [Q\cos(q\tau + \theta) - I_2]\sin\theta\}$$
$$+ 2M_{14}[I_1 - Q\cos\theta][I_2 - Q\cos(q\tau + \theta)]$$
$$+ M_{22}(Qq)^2[\sin^2\theta + \sin^2(q\tau + \theta)] + 2M_{23}(Qq)^2\sin\theta\sin(q\tau + \theta).$$

The M's are given in Ref. 4, Appendix I with

$$m(\tau) = \int_0^\infty W(f)\cos 2\pi f\tau\, df, \tag{3}$$

where $W(f)$ = one-sided power spectral density of $I_N(t)$. Also, $\beta = -m''(0)$. The primes denote differentiations.

Equation 2 can be put in the form:

$$P_2^{+-}(\tau, I_1, I_2, a) = [4\pi^2 N_{I_1}]^{-1}M_{22}(1 - m^2)^{-\frac{1}{2}} \cdot \int_{-\pi}^{\pi} \exp(-G_2/2M)J(r_1, h_3, k_3)\, d\theta \tag{4}$$

where

$$J(r_1, h_3, k_3) \equiv \frac{1}{2\pi(1 - r_1^2)^{\frac{1}{2}}}\int_{h_3}^\infty dx \int_{k_3}^\infty dy(x - h_3)(y - k_3)e^z$$

$$z = -\frac{x^2 + y^2 - 2r_1xy}{2(1 - r_1^2)}$$

$$h_3 = M_{22}^{-1}[1 - r_1^2]^{-1}[D_2 - r_1E_2]\left[\frac{1 - m^2}{M_{22}}\right]^{\frac{1}{2}}$$

$$k_3 = -M_{22}^{-1}[1 - r_1^2]^{-1}[E_2 - r_1 D_2]\left[\frac{1 - m^2}{M_{22}}\right]^{\frac{1}{2}}$$

$$G_2 = M_{22}^{-1}[1 - r_1^2]^{-1}[2r_1 D_2 E_2 - D_2^2 - E_2^2] + F_2 .$$

$P_2^{-+}(\tau, I_1, I_2, a)$ is obtained from equation 2 by changing the signs of the ∞'s in the limits of integration. We find that $P_2^{-+}(\tau, I_1, I_2, a)$ is equal to the right side of equation 4 with h_3, k_3 replaced by $-h_3$, $-k_3$.

$P_2^{--}(\tau, I_1, I_2, a)$ is obtained from equation 2 by changing the upper limit of integration of I_1' to $-\infty$. We find that $P_2^{--}(\tau, I_1, I_2, a)$ is equal to the right side of equation 4 with the function $J(r_1, h_3, k_3)$ replaced by the function $J_1(r_1, h_3, k_3)$, where

$$J_1(r_1, h_3, k_3) \equiv \frac{1}{2\pi(1 - r_1^2)^{\frac{1}{2}}} \int_{h_3}^{-\infty} dx \int_{k_3}^{\infty} dy\, (x - h_3)(y - k_3)e^z. \qquad (5)$$

$P_2^{++}(\tau, I_1, I_2, a)$ is obtained from equation 2 by changing the lower limit of integration of I_2' to $+\infty$. We find that $P_2^{++}(\tau, I_1, I_2, a)$ is equal to the right side of equation 4 with the function $J(r_1, h_3, k_3)$ replaced by the function $J_1(r_1, -h_3, -k_3)$.

The functions $J(r_1, h_3, k_3)$ and $J_1(r_1, h_3, k_3)$ are expressed in terms of Karl Pearson's well-known tabulated function (d/N) in Ref. 10.

A considerable simplification occurs when $a = 0$. For this case we find

$$P_2^{+-}(\tau, I_1, I_2, 0) = M_{22}\beta^{-\frac{1}{2}}(1 - m^2)^{-\frac{3}{2}} \cdot \exp\left[\frac{I_1^2}{2} - \frac{(I_1^2 + I_2^2 - 2mI_1I_2)}{2(1 - m^2)}\right] J(r_1, h_3, k_3) \qquad (6)$$

where

$$h_3 = \frac{-m'(mI_1 - I_2)}{[M_{22}(1 - m^2)]^{\frac{1}{2}}}$$

$$k_3 = \frac{m'(I_1 - mI_2)}{[M_{22}(1 - m^2)]^{\frac{1}{2}}}.$$

Equation 6 reduces to Rice's[4] equation 47 when $I_1 = I_2 = I$.

III. APPROXIMATE RESULTS VIA INTEGRAL EQUATIONS

3.1 *Probability Densities*

Let us assume that each of the random variables τ^+, τ^-, τ_1^+, and τ_1^- (see Fig. 1), is statistically independent of the sum of the following $(2N + 2)$ axis-crossing intervals at level I_2 when $N = 0, 1, 2, \cdots$. Under this "quasi-independence" assumption, approximate theoretical results

for the probability densities of τ^+, τ^-, τ_1^+, and τ_1^-, namely $W^+(\tau, I_1, I_2, a)$, $W^-(\tau, I_1, I_2, a)$, $W_1^+(\tau, I_1, I_2, a)$, and $W_1^-(\tau, I_1, I_2, a)$ are given by the following integral equations:

$$P_2^{+-}(\tau, I_1, I_2, a) = W^+(\tau, I_1, I_2, a) + W^+(\tau, I_1, I_2, a) * P_2^{--}(\tau, I_2, I_2, a) \quad (7)$$

$$P_2^{--}(\tau, I_1, I_2, a) = W^-(\tau, I_1, I_2, a) + W^-(\tau, I_1, I_2, a) * P_2^{--}(\tau, I_2, I_2, a) \quad (8)$$

$$P_2^{++}(\tau, I_1, I_2, a) = W_1^+(\tau, I_1, I_2, a) + W_1^+(\tau, I_1, I_2, a) * P_2^{++}(\tau, I_2, I_2, a) \quad (9)$$

$$P_2^{-+}(\tau, I_1, I_2, a) = W_1^-(\tau, I_1, I_2, a) + W_1^-(\tau, I_1, I_2, a) * P_2^{++}(\tau, I_2, I_2, a). \quad (10)$$

The P_2's are the auxiliary probability functions presented in Section II, and the symbol * denotes the convolution operator, that is,

$$f * g \equiv \int_{-\infty}^{\infty} f(t)g(\tau - t)\, dt. \quad (11)$$

From symmetry we have

$$P_2^{++}(\tau, I_2, I_2, a) = P_2^{--}(\tau, I_2, I_2, a). \quad (12)$$

Integral equations 7 through 10 are analogous to McFadden's equation 47 and Rice's equation 84 in Refs. 11 and 4, respectively.

Let us define two additional probability densities defined by the first and second passage times of $I(t, a)$ between the levels I_1 and I_2:

(*i*) $W(\tau, I_1, I_2, a)\,d\tau$, the conditional probability that the first crossing of the level I_2 occurs between $t + \tau$ and $t + \tau + d\tau$ given a crossing of the level $I_1 > I_2$ at t.

(*ii*) $W_1(\tau, I_1, I_2, a)\,d\tau$, the conditional probability that the second crossing of the level I_2 occurs between $t + \tau$ and $t + \tau + d\tau$ given a crossing of the level $I_1 > I_2$ at t.

Clearly, we have

$$W(\tau, I_1, I_2, a) = \tfrac{1}{2}W^+(\tau, I_1, I_2, a) + \tfrac{1}{2}W^-(\tau, I_1, I_2, a) \quad (13)$$

$$W_1(\tau, I_1, I_2, a) = \tfrac{1}{2}W_1^+(\tau, I_1, I_2, a) + \tfrac{1}{2}W_1^-(\tau, I_1, I_2, a). \quad (14)$$

3.2 *Absolute Minimum In a Closed Interval*

Let us define the following probability functions concerning the absolute minimum of $I(t, a)$ in a closed interval $[0, \tau]$:

$$F^+(\tau, I_1, I_2, a) \equiv \Pr\{\min_{0\le t\le\tau} I(t, a) > I_2 \mid I(0, a) = I_1 > I_2, I'(0, a) > 0\} \quad (15)$$

$$F^-(\tau, I_1, I_2, a) \equiv \Pr\{\min_{0\le t\le\tau} I(t, a) > I_2 \mid I(0, a) = I_1 > I_2, I'(0, a) < 0\} \quad (16)$$

$$F(\tau, I_1, I_2, a) \equiv \Pr\{\min_{0\le t\le\tau} I(t, a) > I_2 \mid I(0, a) = I_1 > I_2\} \quad (17)$$

where $\Pr\{\cdot\}$ denotes the probability of the event inside the brace. Clearly, we have

$$\begin{aligned} F^+(\tau, I_1, I_2, a) &= \int_\tau^\infty W^+(\tau, I_1, I_2, a)\, d\tau \\ &= 1 - \int_0^\tau W^+(\tau, I_1, I_2, a)\, d\tau \end{aligned} \quad (18)$$

$$\begin{aligned} F^-(\tau, I_1, I_2, a) &= \int_\tau^\infty W^-(\tau, I_1, I_2, a)\, d\tau \\ &= 1 - \int_0^\tau W^-(\tau, I_1, I_2, a)\, d\tau \end{aligned} \quad (19)$$

$$\begin{aligned} F(\tau, I_1, I_2, a) &= \int_\tau^\infty W(\tau, I_1, I_2, a)\, d\tau \\ &= 1 - \int_0^\tau W(\tau, I_1, I_2, a)\, d\tau. \end{aligned} \quad (20)$$

Because we are discussing only the case when $I_1 > I_2$ as is indicated in Fig. 1, we discuss only the probability functions concerning the absolute minimum of $I(t, a)$ in a closed interval $[0, \tau]$. The corresponding probability functions concerning the absolute maximum of $I(t, a)$ in a closed interval $[0, \tau]$ are associated with the case when $I_1 < I_2$, and they can be obtained from symmetry by considering the random process $-I(t, a)$.

IV. RESULTS FOR SINUSOIDAL SIGNAL CENTERED IN LOW-PASS NOISE

For purposes of computation we set the angular frequency, q, of the sinusoidal signal in the center of a band of gaussian noise with an ideal low-pass power spectral density of cutoff frequency f_0. Thus,

$$q = \pi f_0 \quad (21)$$

and

$$W(f) = \begin{cases} f_0^{-1} & 0 \leqq f \leqq f_0 \\ 0 & \text{otherwise.} \end{cases} \tag{22}$$

Accordingly, from equation 3,

$$m(\tau) = \frac{\sin 2\pi f_0 \tau}{2\pi f_0 \tau}. \tag{23}$$

From equation 23 we see that it is convenient to define normalized time as $u_0 = 2\pi f_0 \tau$. All our results are plotted with respect to normalized time u_0.

4.1 *Experimental Verification of Auxiliary Probability Functions*

The auxiliary probability functions for the case when $I_1 = I_2 = 0$ are useful for studying the zero-crossing intervals, the axis-crossing intervals defined by the level $I_1 = I_2 = 0$, of $I(t, a)$. Figures 2, 3, and 4 present P_2^{+-}, P_2^{-+}, P_2^{++}, and P_2^{--} for the case when $I_1 = I_2 = 0$ and $a = 0$, 1, and 4. These results were computed by using Simpson's rule. The results compare satisfactorily with the initial behavior of the experimental probability densities presented in Figs. 34, 35, 42, 43, 46, and 47 of Ref. 12. Notice that the experimental probability densities pertain to the case when the power spectral density of the noise is

$$W_0(f) = \frac{1}{1 + \left(\frac{f}{f_0}\right)^{14}} \tag{24}$$

rather than the power spectral density defined by equation 22.

4.2 *Results When* a = *0 and* a = *1*

Figures 5 through 13 present the results when $a = 0$, signal absent, and $I_1 = 1$, $I_2 = 0$; $I_1 = 1$, $I_2 = -1$; and $I_1 = 0$, $I_2 = -1$. The P_2's were computed by using Simpson's rule, the integral equations defining the W's were solved numerically by using the trapezoidal rule, and the F's were computed by using Simpson's rule.

Similarly, Figs. 14 through 22 present the results when $a = 1$, signal present, and $I_1 = 1$, $I_2 = 0$; $I_1 = 1$, $I_2 = -1$; and $I_1 = 0$, $I_2 = -1$.

As an example of the interpretation of these results, we see from Fig. 16 that the median time, τ_m, for the random process $I(t, 1)$ to go from the level $I_1 = 1$ to the level $I_2 = 0$ for the first time is given by

$$u_0 = 2\pi f_0 \tau_m \doteq \pi \tag{25}$$

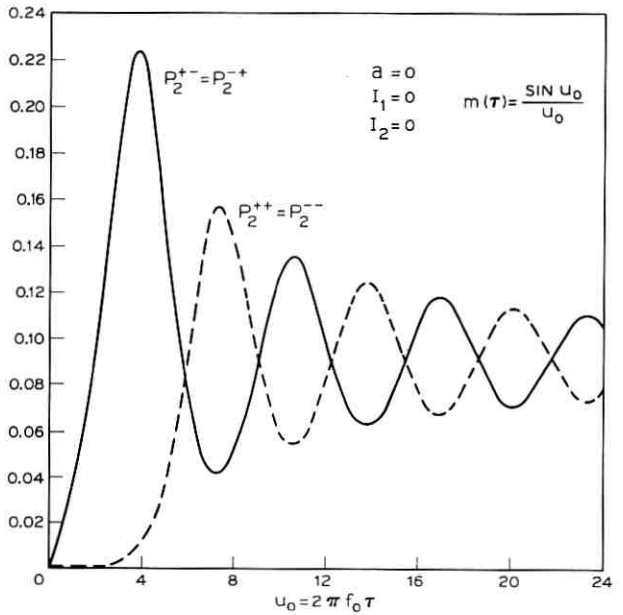


Fig. 2 (also Figs. 3 and 4) — Plots of the probability functions $P_2^{+-}(u_0, I_1, I_2, a)$, $P_2^{-+}(u_0, I_1, I_2, a)$, $P_2^{++}(u_0, I_1, I_2, a)$, and $P_2^{--}(u_0, I_1, I_2, a)$ associated with the crossings of the levels I_1 and I_2 by a stationary random process consisting of a sinusoidal signal of frequency $f_0/2$ plus stationary gaussian noise with autocorrelation function $m(\tau)$. a denotes the signal-to-noise power ratio.

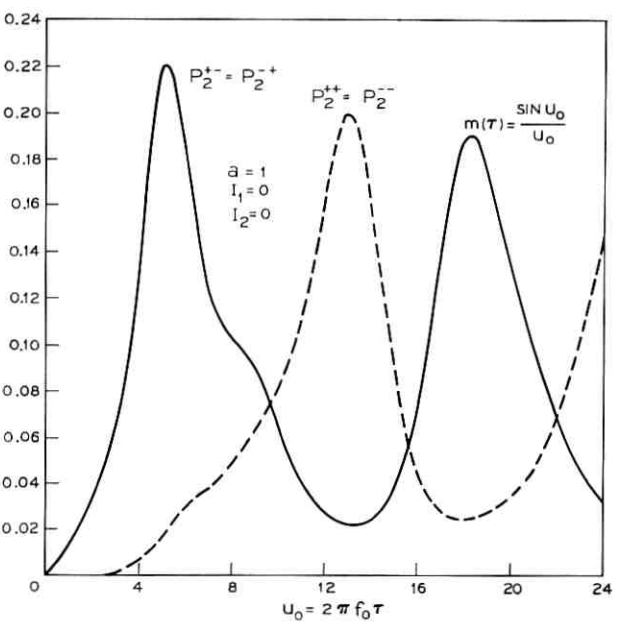


Fig. 3 — (See Fig. 2.)

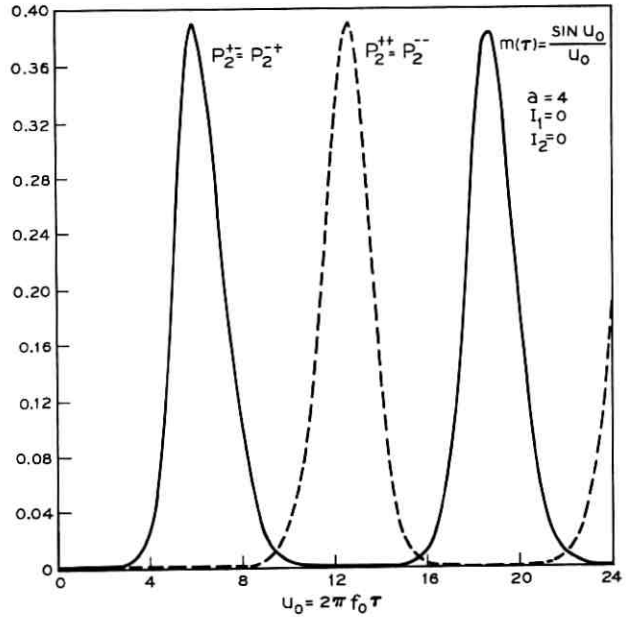


Fig. 4 — (See Fig. 2.)

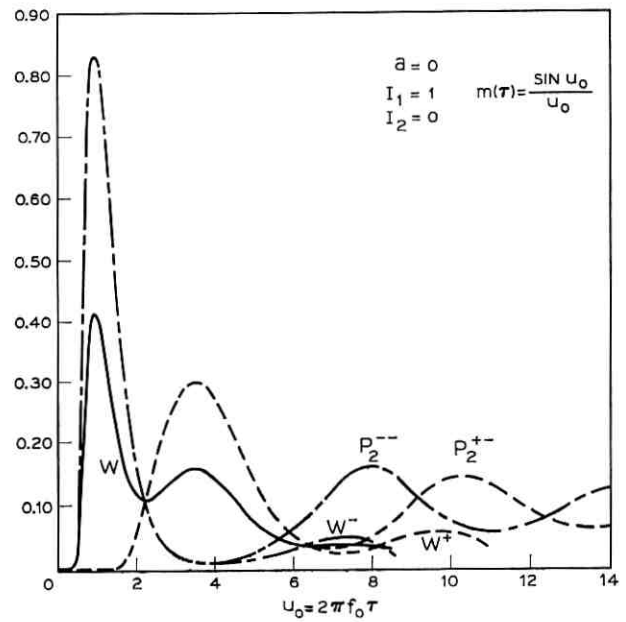


Fig. 5 (also Figs. 8, 11, 14, 17, and 20) — Plots of the probability functions $P_2^-(u_0, I_1, I_2, a)$, $W^-(u_0, I_1, I_2, a)$, $P_2^+(u_0, I_1, I_2, a)$, $W^+(u_0, I_1, I_2, a)$, and $W(u_0, I_1, I_2, a)$ associated with the crossings of the levels I_1 and I_2 by a stationary random process consisting of a sinusoidal signal of frequency $f_0/2$ plus stationary gaussian noise with autocorrelation function $m(\tau)$. a denotes the signal-to-noise power ratio.

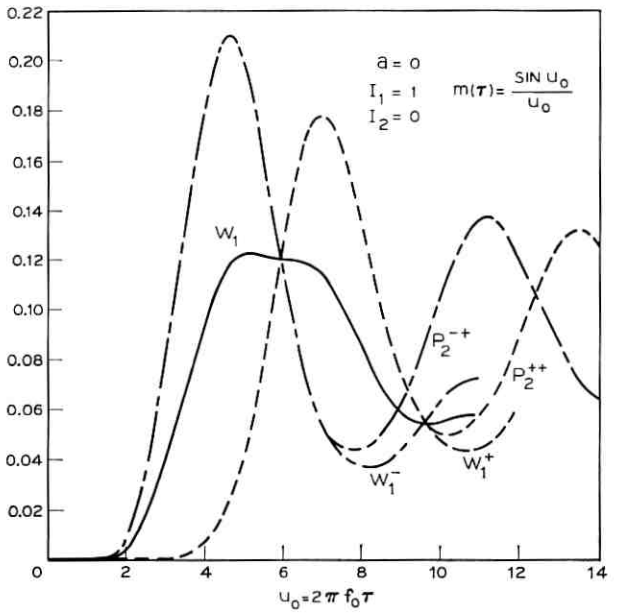


Fig. 6 (also Figs. 9, 12, 15, 18, and 21) — Plots of the probability functions $P_2^{-+}(u_0, I_1, I_2, a)$, $W_1^{-}(u_0, I_1, I_2, a)$, $P_2^{++}(u_0, I_1, I_2, a)$, $W_1^{+}(u_0, I_1, I_2, a)$ and $W_1(u_0, I_1, I_2, a)$ associated with the crossings of the levels I_1 and I_2 by a stationary random process consisting of a sinusoidal signal of frequency $f_0/2$ plus stationary gaussian noise with autocorrelation function $m(\tau)$. a denotes the signal-to-noise power ratio.

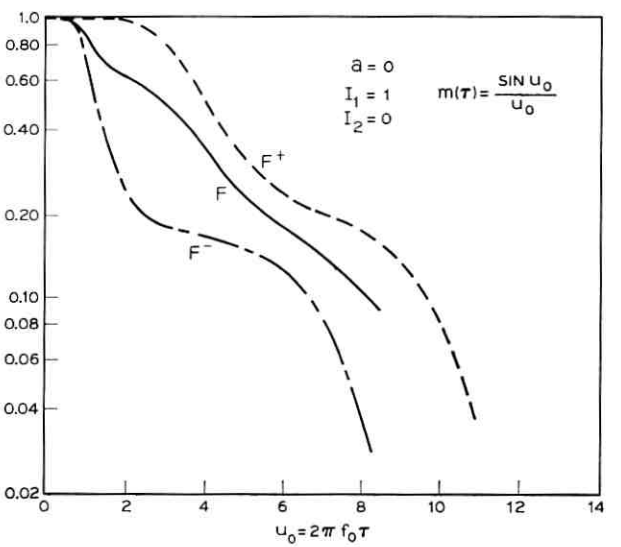


Fig. 7 (also Figs. 10, 13, 16, 19, and 22) — Plots of the probability functions $F^{+}(u_0, I_1, I_2, a)$, $F^{-}(u_0, I_1, I_2, a)$, and $F(u_0, I_1, I_2, a)$ associated with the crossings of the levels I_1 and I_2 by a stationary random process consisting of a sinusoidal signal of frequency $f_0/2$ plus stationary gaussian noise with autocorrelation function $m(\tau)$. a denotes the signal-to-noise power ratio.

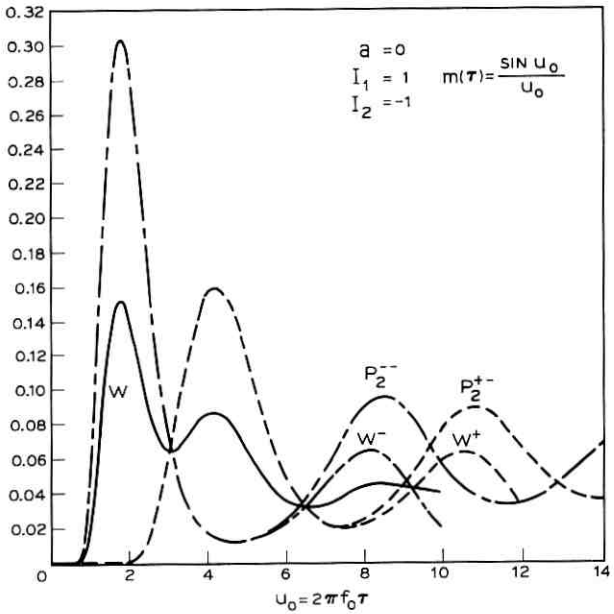


Fig. 8 — (See Fig. 5.)

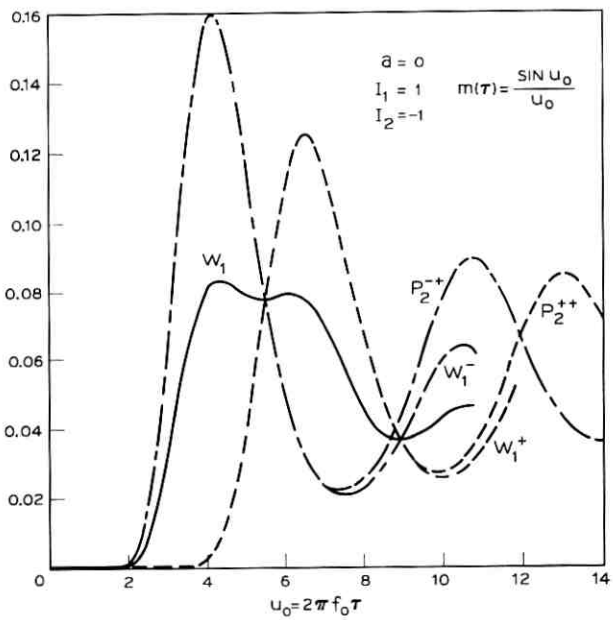


Fig. 9 — (See Fig. 6.)

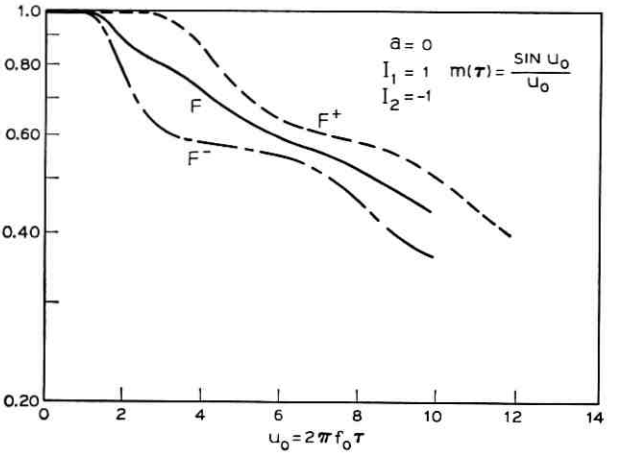


Fig. 10 — (See Fig. 7.)

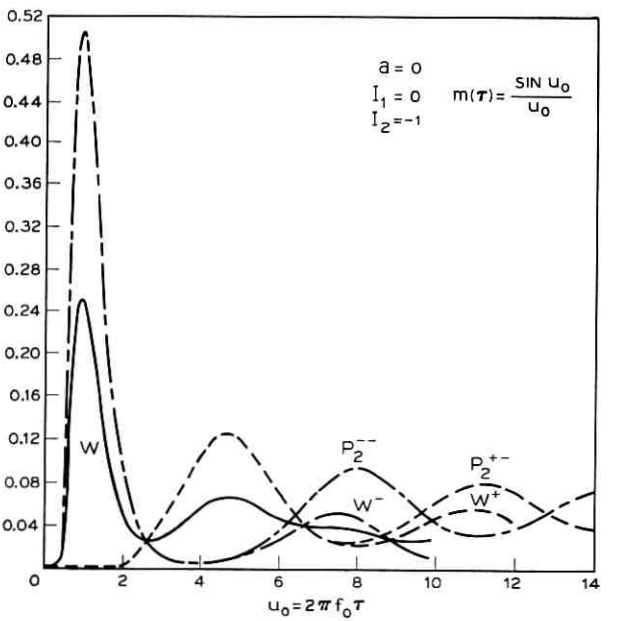


Fig. 11 — (See Fig. 5.)

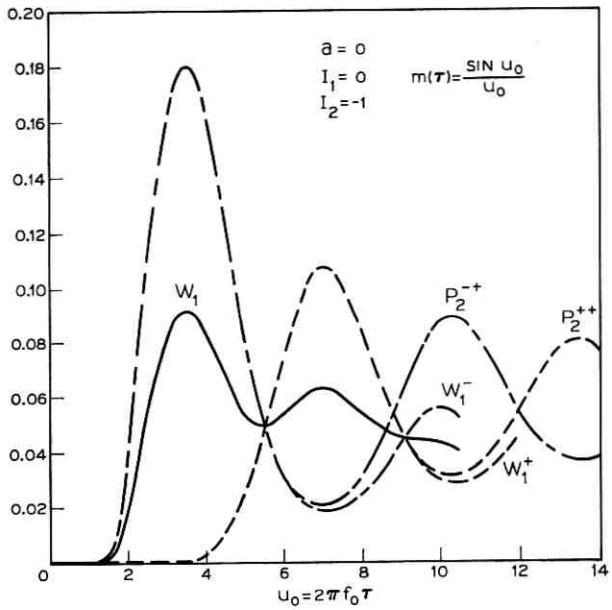


Fig. 12 — (See Fig. 6.)

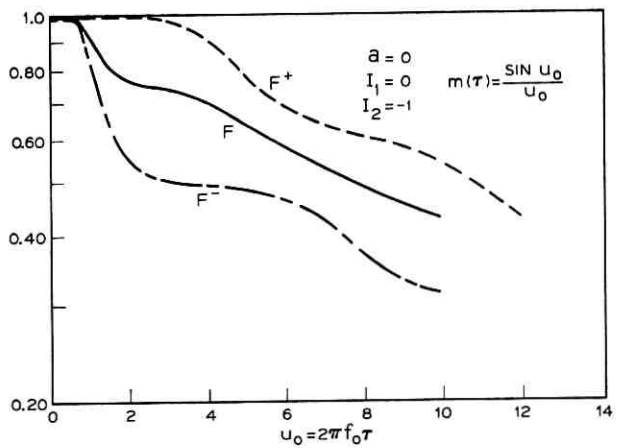


Fig. 13 — (See Fig. 7.)

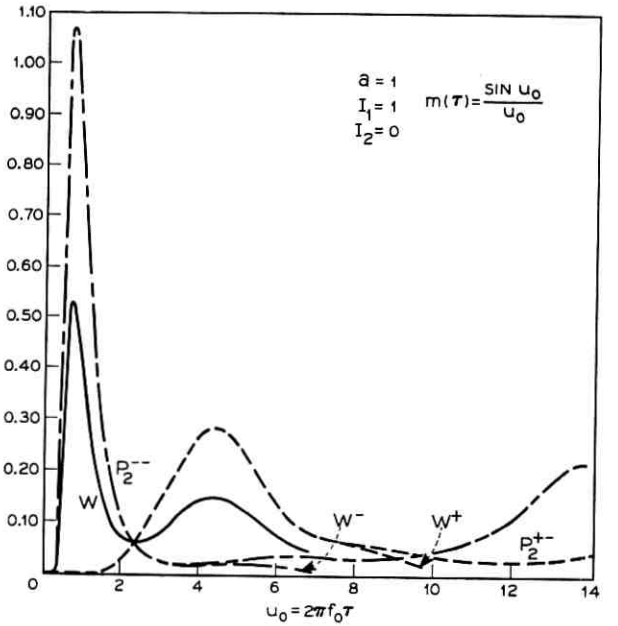


Fig. 14 — (See Fig. 5.)

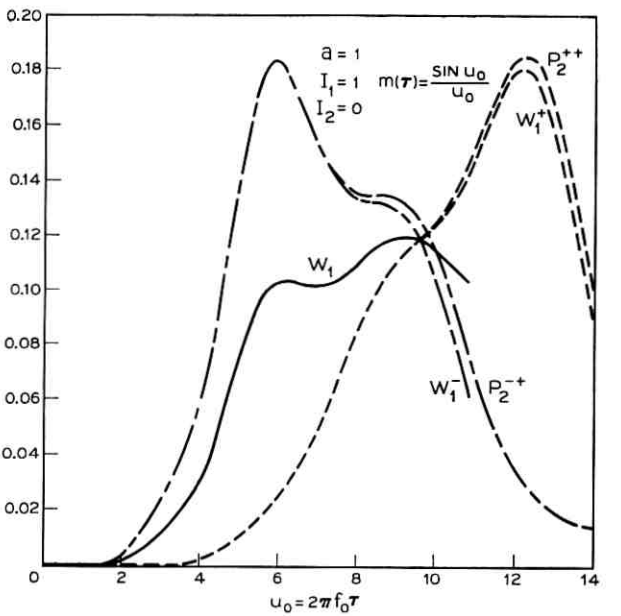


Fig. 15 — (See Fig. 6.)

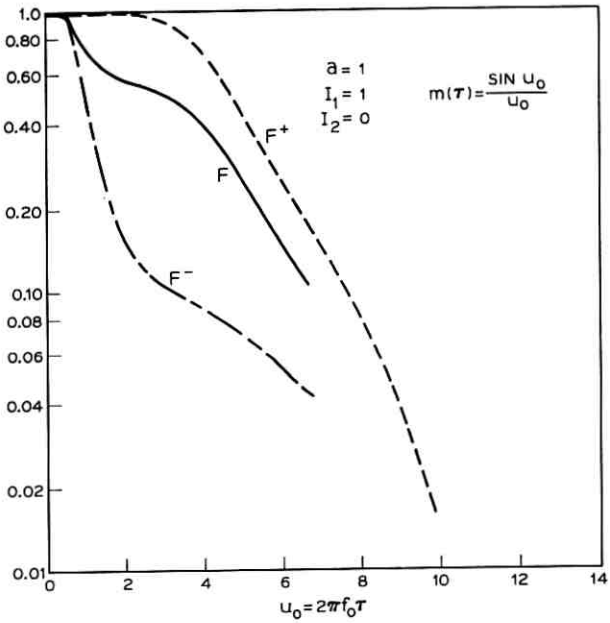


Fig. 16 — (See Fig. 7.)

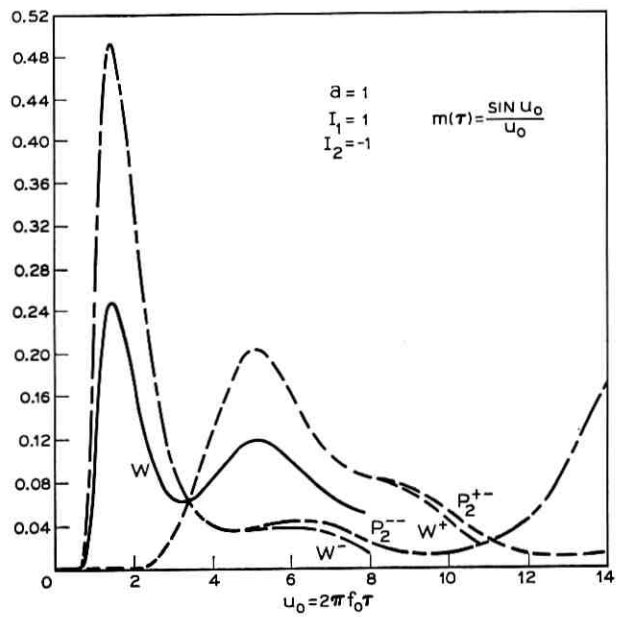


Fig. 17 — (See Fig. 5.)

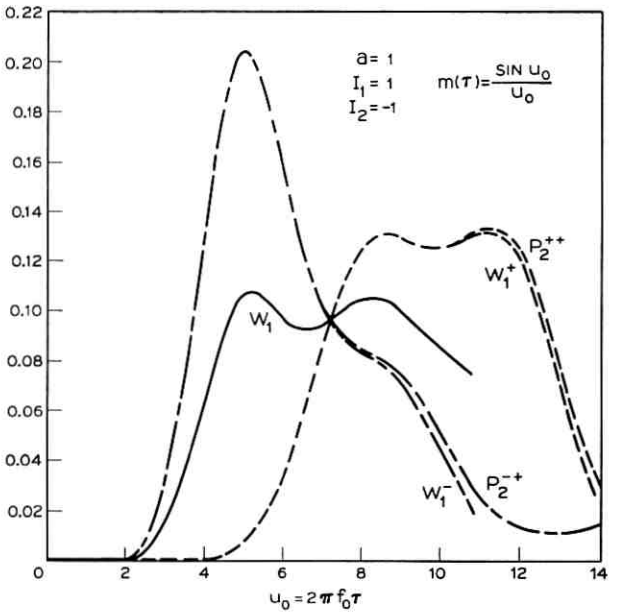


Fig. 18 — (See Fig. 6.)

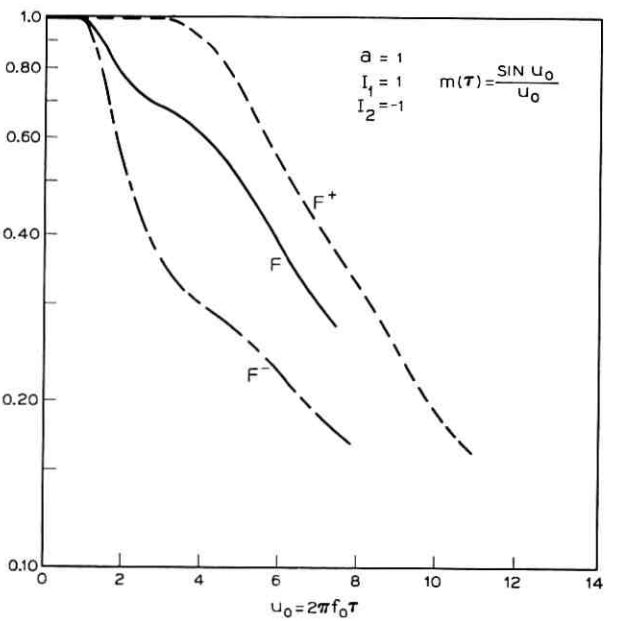


Fig. 19 — (See Fig. 7.)

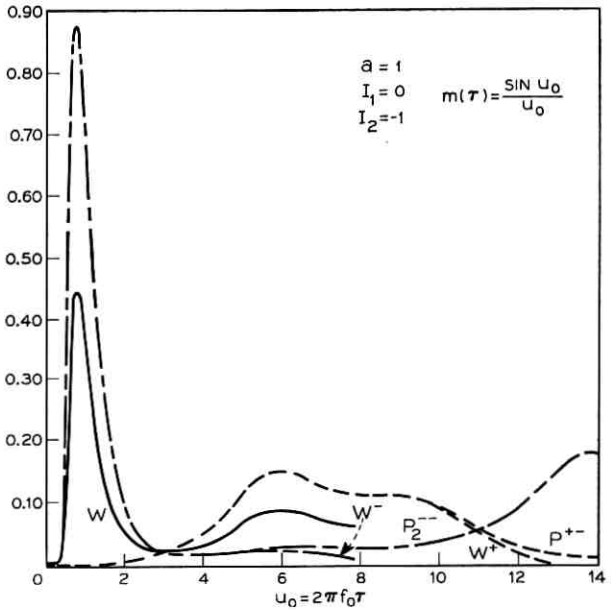


Fig. 20 — (See Fig. 5.)

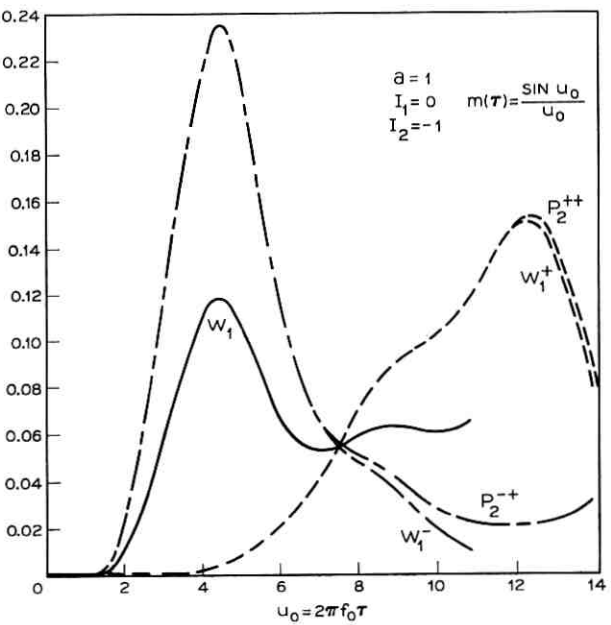


Fig. 21 — (See Fig. 6.)

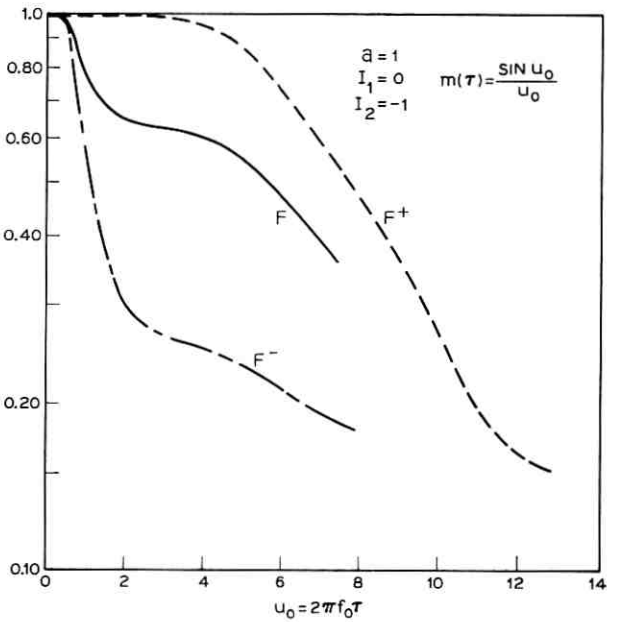


Fig. 22 — (See Fig. 7.)

or

$$\tau_m \doteq (2f_0)^{-1}. \tag{26}$$

τ_m also represents the median time during which the random process $I(t, 1)$ remains continually above the level $I_2 = 0$ when it starts at the level $I_1 = 1$. From symmetry, τ_m also represents the median time during which the random process $I(t, 1)$ remains continually below the level $I_2 = 0$ when it starts at the level $I_1 = -1$.

V. CONCLUSIONS

The exact auxiliary probability functions can be used in approximate integral equations in order to deduce approximate probability densities of the first and second passage times of sine wave plus stationary, gaussian noise with a finite expected number of zeros per unit time. These approximate probability densities can be used to deduce the approximate median times associated with the first and second passage times. Also, the approximate probability densities of the first passage times can be used to deduce approximate distribu-

tion functions for the absolute minimum or absolute maximum of sine wave plus noise in a closed interval.

The corresponding exact results are not yet known.

VI. ACKNOWLEDGMENT

The author is indebted to Miss A. T. Seery for programming a digital computer to produce the graphical results.

REFERENCES

1. Wang, Ming Chen, and Uhlenbeck, G. E., "On the Theory of the Brownian Motion II," Rev. Modern Phys., *17* (April–July 1945), pp. 323–342.
2. Siegert, A. J. F., "On the First Passage Time Probability Problem," Phys. Rev., *81* (February 15, 1951), pp. 617–623.
3. Darling, D. A. and Siegert, A. J. F., "The First Passage Problem for a Continuous Markov Process," Ann. Math. Stat., *24* (December 1953), pp. 624–639.
4. Rice, S. O. "Distribution of the Duration of Fades in Radio Transmission," B.S.T.J. *37*, No. 3(May 1958), pp. 581–635.
5. Slepian, D., "First Passage Time for a Particular Gaussian Process, Ann. Math. Stat., *32* (June 1961), pp. 610–612.
6. Kac, M., "Probability Theory as a Mathematical Discipline and as a Tool in Engineering and Science," *Proc. First Symp. on Eng. Applications of Random Function Theory and Prob.*, ed. J. L. Bogdanoff and F. Kozin, New York: Wiley 1963, pp. 31–67.
7. Mehr, C. B. and McFadden, J. A., "Certain Properties of Gaussian Processes and Their First-Passage Times," J. Royal Stat. Soc., Series B., *27*, No. 3 (1965), pp. 505–522.
8. Rainal, A. J. "Axis-Crossing Intervals of Sine Wave Plus Noise," B.S.T.J., *46*, No. 7 (September 1967), pp. 1655–1658.
9. Rice, S. O., "Statistical Properties of a Sine Wave Plus Random Noise," B.S.T.J., *27*, No. 1 (January 1948), pp. 109–157.
10. Rainal, A. J., "Axis-Crossing Intervals of Rayleigh Processes," B.S.T.J., *44*, No. 6 (July–August 1965), pp. 1219–1224
11. McFadden, J. A., "The Axis-Crossing Intervals of Random Functions-II," IRE Trans. Inform. Theory, *IT-4* (March 1958), pp. 14–23.
12. Rainal, A. J., "Zero-Crossing Intervals of Random Processes," Technical Report AF-102, DDC No. AD-401-148, Baltimore, Maryland: Johns Hopkins University, Carlyle Barton Laboratory, April 1963. Abstracted in IEEE Trans. Inform. Theory., *IT-9* (October 1963), p. 295.

The Spectrum of a Simple Nonlinear System

By S. C. LIU

(Manuscript received July 3, 1968)

The random motion of a particle with nonlinear damping is investigated. The spectrum of the velocity of the particle is obtained by solving the associated nonstationary Fokker–Planck equation and also by using the equivalent-linearization technique. The first procedure yields an exact solution in terms of Laguerre polynomials. The second leads to simple, approximate results which are valid for cases where the small nonlinearity assumption holds. Results obtained by these two methods are compared and good agreement is observed over a large frequency range.

I. INTRODUCTION

The recent advance in space and communicational technologies has led engineers to numerous difficult but fascinating problems in regard to the structural dynamics in random environments. For example, Hempstead and Lax have investigated noise in self-sustained oscillation;[1, 2] Ariaratnam and Sanker have studied the dynamic snap-through of shallow, arch-type aircraft components under stochastic pressure.[3] In this paper the random vibration of a simple mass with nonlinear damping is studied. The nonlinearity of the system is introduced to linear viscous damping by adding to it an extra term which is inversely proportional to the first power of the current velocity. Emphasis of the analysis is placed on finding the power spectral density of the random motion.

Two different approaches are used to obtain the desired solution. First, the exact spectrum is found by solving the associated nonstationary Fokker–Planck equation in terms of the eigenfunction expansion of the degenerate ordinary differential equation. Second, approximate solutions are obtained by using the equivalent linearization technique by which the original nonlinear system is converted to an equivalent linear one. The equivalent linear system, constructed by the least mean square error criterion and based on the small nonlinearity assumption, is then solved by standard linear theory.

II. NONSTATIONARY FOKKER–PLANCK EQUATION

Consider the first-order nonlinear system described by the following differential equation:

$$\dot{x} + F(x) = f(t), \tag{1}$$

which may be thought of as the velocity equation of a unit mass with nonlinear damping $F(x)$ subject to a force $f(t)$.

Let us discuss the problem of obtaining the power spectral density of $x(t)$ when $f(t)$ is a random process. We limit the discussion to the case where $f(t)$ is a stationary white gaussian process with the first two moments defined as

$$\langle f(t) \rangle = 0 \tag{2}$$

and

$$\langle f(t_1) f(t_2) \rangle = 2\pi s_o \, \delta(r_1 - t_2) \tag{3}$$

where s_o is a constant, the symbol $\langle \ \rangle$ indicates the ensemble average and δ indicates the Dirac delta function.

Caughey and Dienes[4] have investigated a similar problem for $F(x) = k \operatorname{sgn} x$. We shall however consider a different case in which

$$F(x) = \beta x - \frac{\gamma}{x} \tag{4}$$

$$0 < x < \infty$$

where β is a constant and γ is a smaller nonlinear coefficient. In case $\gamma = 0$, equation 1 becomes the familiar linear differential equation.

The Fokker–Planck equation which governs the transition probability $p(x_o \mid x, \tau)$ with given initial velocity $x_o = x(t_o)$ for the velocity $x(t)$ at time t is

$$\dot{p} = \frac{\partial}{\partial x}\left[\left(\beta x - \frac{\gamma}{x}\right)p\right] + \pi s_o \frac{\partial^2 p}{\partial x^2} \tag{5}$$

where $\tau = t - t_o$. The initial and boundary conditions for equation 5 are

$$p_{t_o} = \delta(x - x_o) \tag{6}$$

and

$$p(0, t) = p(\infty, t) = 0, \tag{7}$$

respectively. As the time of passage t becomes sufficiently large, $p(x_o \mid x, \tau)$ in equation 5 approaches a stationary value $p_{st}(x)$ independent of t and initial condition 6. Setting $\dot{p} = 0$ in 5, such a stationary density can be found by solving the degenerate stationary equation. The result is:

$$p_{st} = C \exp\left[-\frac{1}{\pi s_o}\left(\frac{\beta x^2}{2} - \gamma \log x\right)\right] \tag{8}$$

where C is the normalization factor determined by

$$\int_0^\infty p_{st}\, dx = 1. \tag{9}$$

The power spectrum is the Fourier transform of the autocorrelation function which is determined by the joint probability density $p(x_o, x, \tau)$. Thus from the relation

$$p(x_o, x, \tau) = p_{st}(x_o)p(x_o \mid x, \tau), \tag{10}$$

we need to find the transition probability density, that is, the nonstationary solution of equation 5. Let $p(x_o \mid x, \tau) = T(t)X(x)$ in 5. It follows that

$$T(t) + \lambda T(t) = 0 \tag{11}$$

and

$$\sigma^2 \frac{\partial^2 X}{\partial x^2} + \frac{\partial}{\partial x}(xX) - \frac{\sigma^2\gamma}{\pi s_o}\frac{\partial}{\partial x}\left(\frac{X}{x}\right) + \frac{\lambda}{\beta}X = 0 \tag{12}$$

where $\sigma^2 = \pi s_o/\beta$. If $X_m(x)$ is the eigenfunction and λ_m the corresponding eigenvalues satisfying equation 12 and the prescribed boundary condition 7, it can be shown that[5]

$$p(x_o \mid x, \tau) = \sum_{m=0}^{\infty} \frac{X_m(x)X_m(x_o)}{p_{st}(x_o)} e^{-\lambda_m(t-t_o)}. \tag{13}$$

In deriving 13, the following orthogonality condition has been used:

$$\int X_m(x)X_n(x)\frac{dx}{p_{st}(x)} = \delta_{mn}. \tag{14}$$

Following the transformations adopted by Stratonovich, let $\mu = \frac{1}{4}(\gamma/\pi s_o - 1)$, $z = x^2/2\sigma^2$, and $u = z^{-\mu}X$. Equation 12 becomes

$$\frac{\partial^2 u}{\partial z^2} + \frac{\partial u}{\partial z} + \left[\frac{(1/2) + u + (\lambda/2\beta)}{z} + \frac{(1/4) - u^2}{z^2}\right]u = 0. \tag{15}$$

Equation 15 is a degenerate hypergeometric differential equation which has eigenfunctions $U_n(z)$ with corresponding eigenvalues $\lambda_n = 2n\beta$:

$$U_n(z) = z^{\mu+(1/2)}e^{-z}L_n^{(2\mu)}(z) \tag{16}$$

where

$$L_n^{\alpha}(z) = \frac{1}{n!}\, e^z z^{-\alpha}\, \frac{d^n}{dz^n}\,(e^{-z}z^{n+\alpha}) \tag{17}$$

is the Laguerre polynomial of degree n.

Transforming back to the original variables and applying equation 14, we obtain the following normalized eigenfunctions:

$$X_n(x) = \frac{(2)^{\frac{1}{2}}}{\sigma}\, \frac{z^{2\mu+(1/2)}e^{-z}L_n^{(2\mu)}(z)}{[n!\,\Gamma(n+2\mu+1)\Gamma(2\mu+1)]^{\frac{1}{2}}}. \tag{18}$$

From equations 13 and 10 the transition density and jointly density, respectively, can be found as

$$p(x_o \mid x, \tau) = \frac{(2)^{\frac{1}{2}}}{\sigma}\, \frac{z^{2\mu+(1/2)}e^{-z}}{L_o^{(2\mu)}(z_o)} \sum_{n=0}^{\infty} \frac{L_n^{(2\mu)}(z)L_n^{(2\mu)}(z_o)}{n!\,\Gamma(n+2\mu+1)}\, e^{-2n\beta|\tau|} \tag{19}$$

and

$$p(x_o\,,\, x,\, \tau) = \frac{2}{\sigma^2}\,(zz_o)^{2\mu+(1/2)}e^{-(z+z_o)} \cdot \sum_{n=0}^{\infty} \frac{L_n^{(2\mu)}(z)L_n^{(2\mu)}(z_o)}{n!\,\Gamma(n+2\mu+1)\Gamma(2\mu+1)}\, e^{-2n\beta|\tau|} \tag{20}$$

where

$$z = \frac{x^2}{2\sigma^2} \quad \text{and} \quad z_o = \frac{x_o^2}{2\sigma^2}.$$

From the above the autocorrelation function $R_x(\tau)$ of $x(t)$, where $\tau = t - t_o$, is

$$R_x(\tau) = \int_0^{\infty}\int_0^{\infty} p(x_o\,,\, x,\, \tau)x_o x\, dx\, dx_o$$

$$= 2\sigma^2 \sum_{n=0}^{\infty} \frac{e^{-2n\beta|\tau|}}{\Gamma(n+2\mu+1)\Gamma(2\mu+1)n!}\, I(\mu, n) \tag{21}$$

where

$$I(\mu, n) = \int_0^{\infty}\int_0^{\infty} (z)^{2\mu+(1/2)}e^{-z}(z_o)^{2\mu+(1/2)}e^{-z_o}L_n^{(2\mu)}(z)L_n^{(2\mu)}(z_o)\, dz\, dz_o\,. \tag{22}$$

In the appendix we show that

$$\int_0^\infty z^{2\mu+(1/2)}e^{-z}L_n^{(2\mu)}(z)\,dz = \frac{-\Gamma[2\mu+(3/2)]\Gamma[n-(1/2)]}{2(\pi)^{\frac{1}{2}}n!}, \tag{23}$$

from which it follows that

$$I(\mu, n) = \frac{1}{4\pi}\left\{\frac{\Gamma[2\mu+(3/2)]\Gamma[n-(1/2)]}{n!}\right\}^2. \tag{24}$$

Substituting 24 into 21, we obtain the following expression for the autocorrelation function $R_x(\tau)$:

$$R_x(\tau) = \frac{\sigma^2}{2\pi}\frac{\Gamma^2[2\mu+(3/2)]}{\Gamma(2\mu+1)}\sum_{n=0}^{\infty}\frac{\Gamma^2[n-(1/2)]e^{-2n\beta|\tau|}}{(n!)^3\Gamma(n+2\mu+1)}, \tag{25}$$

which is a monotonically decreasing function of τ.

The power spectral density of $x(t)$ can be derived from equation 25 as the following:

$$S_x(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty}R_x(\tau)e^{-i\omega\tau}\,d\tau = 2\sigma^2\frac{\Gamma^2[2\mu+(3/2)]}{\Gamma(2\mu+1)}\,\delta(\omega)$$
$$+\left(\frac{\sigma}{2\pi}\right)^2\frac{\Gamma^2[2\mu+(3/2)]}{\Gamma(2\mu+1)}\sum_{n=1}^{\infty}\frac{\Gamma^2[n-(1/2)]}{(n!)^3\Gamma(n+2\mu+1)}\frac{4n\beta}{4n^2\beta^2+\omega^2}. \tag{26}$$

Notice that $S_x(\omega)$ is again a monotonically decreasing function of ω and has a spike at $\omega = 0$.

The nonstationary mean value of $x(t)$ is given by

$$\langle x(\tau)\rangle = \int_0^\infty xp(x_o \mid x, \tau)\,dx. \tag{27}$$

Using equations 19 and 23 it can easily be shown that

$$\langle x(\tau)\rangle = \frac{\sigma\Gamma[2\mu+(3/2)]}{-\sqrt{2\pi}\,L_o^{(2\mu)}(z_o)}\sum_{n=0}^{\infty}\frac{\Gamma[n-(1/2)]L_n^{(2\mu)}(z_o)}{(n!)^2\Gamma(n+2\mu+1)}e^{-2n\beta|\tau|}. \tag{28}$$

Because $x_o = (2\sigma^2 z_o)^{\frac{1}{2}}$, we notice that $\langle x(\tau)\rangle$ depends on the initial velocity x_o. As $\tau \to \infty$, $\langle x(\tau)\rangle$ in equation 28 approaches its stationary value $\langle x\rangle_{st}$, which is given by

$$\langle x\rangle_{st} = \frac{(2)^{\frac{1}{2}}\Gamma[2\mu+(3/2)]}{\Gamma(2\mu+1)}\sigma. \tag{29}$$

$\langle x\rangle_{st}$ is independent of the initial condition x_o. This stationary mean velocity, $\langle x\rangle_{st}$, can also be found by using the stationary density

$p_{st}(x)$ as follows:

$$\begin{aligned}\langle x\rangle_{st} &= \int_0^\infty x p_{st}(x)\,dx \\ &= \int_0^\infty x X_o(x)\,dx \\ &= \frac{(2)^{\frac{1}{2}}\sigma}{\Gamma(2\mu+1)}\int_0^\infty (z)^{2\mu+(1/2)}e^{-z}L_o^{2\mu}(z)\,dz \\ &= \frac{(2)^{\frac{1}{2}}\Gamma[2\mu+(3/2)]}{\Gamma(2\mu+1)}\,\sigma.\end{aligned}$$

The variation of $\langle x(\tau)\rangle$ is illustrated in Fig. 1. It has a maximum value x_o at $\tau = 0$ and decreases exponentially to the stationary value $\langle x\rangle_{st}$ as given in equation 29. The nonstationary mean square value of $x(t)$, given its initial value x_0^2, is difficult to evaluate explicitly, but its stationary value $\langle x^2\rangle_{st}$ can be found by integrating $S_x(\omega)$ in equation 25 over the entire frequency range from $-\infty$ to $+\infty$. By this procedure,

$$\langle x^2\rangle_{st} = \int_{-\infty}^{\infty} S_x(\omega)\,d\omega = \frac{\sigma^2}{2\pi}\,\frac{\Gamma^2[2\mu+(3/2)]}{\Gamma(2\mu+1)}\sum_{n=0}^{\infty}\frac{\Gamma^2[n-(1/2)]}{(n!)^3\Gamma(n+2\mu+1)}. \tag{30}$$

As expected, $\langle x^2\rangle_{st}$ is also independent of the initial condition x_o. The variations of $\langle x^2\rangle_{st}$ with the nonlinear coefficient $k = \gamma/\pi s_o$ are shown in Fig. 2.

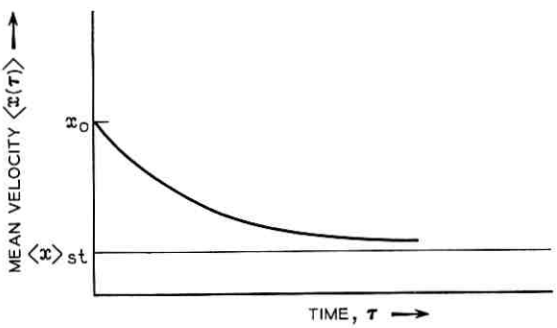


Fig. 1 — Nonstationary mean velocity of nonlinear system subjected to random force.

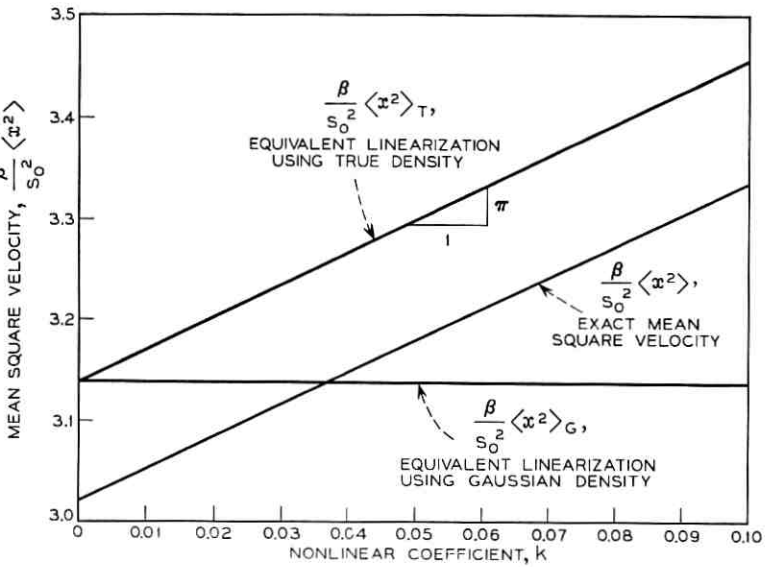


Fig. 2 — Comparison of mean square velocity nonlinear system subjected to random noise.

III. EQUIVALENT LINEARIZATION TECHNIQUE

The eigenfunction expansion of the degenerate ordinary differential equation of the governing Fokker–Planck equation is often difficult to find. In such cases it is sometimes convenient to use the perturbation[6,7] or the equivalent linearization[8,9] techniques to obtain desired quantitative results. If the nonlinearity of the system is small (that is, $\gamma/\beta \ll 1$), these techniques provide the simplest means for obtaining approximate results. In the following, an equivalent linearization procedure is used to derive the power spectral densities of the nonlinear velocity $x(t)$ in equations 1 through 4. Let

$$\dot{x} + \beta_e x + \epsilon(x) = f(t) \tag{31}$$

be the equation of motion of a system equivalent to that described by equations 1 and 4 in which β_e is the equivalent linear stiffness and

$$\epsilon(x) = \beta x + \frac{\gamma}{x} - \beta_e x \tag{32}$$

is the error function. If $\epsilon(x)$ is small and may be ignored, equation 31 becomes a linear differential equation, and its spectrum can be

solved by standard technique. The equivalent linear stiffness β_e can be determined by the criterion that the mean square error is minimized. The mean square error is

$$\langle \epsilon(x)^2 \rangle = \int_0^\infty \left[(\beta - \beta_e)^2 x^2 + \frac{\gamma^2}{x^2} + 2(\beta - \beta_e)\gamma \right] p(x)\, dx.$$

Setting $\partial \langle \epsilon(x)^2 \rangle / \partial \beta_e = 0$ we obtain

$$\int_0^\infty \beta x^2 p(x)\, dx = \int_0^\infty \beta_e x^2 p(x)\, dx - \int_0^\infty \gamma p(x)\, dx;$$

therefore

$$\beta_e = \frac{\beta \int_0^\infty x^2 p(x)\, dx + \gamma}{\int_0^\infty x^2 p(x)\, dx}$$

$$= \beta + \frac{\gamma}{\langle x^2 \rangle}. \tag{33}$$

The following two cases are considered:

(*i*) Assuming $p(x)$ is gaussian, that is,

$$p(x) = \left(\frac{2\pi^2 s_o}{\beta} \right)^{-\frac{1}{2}} \exp\left(-\frac{\beta x^2}{2\pi s_o} \right),$$

then

$$\langle x^2 \rangle_G = \sigma_x^2 = \frac{\pi s_o}{\beta}. \tag{34}$$

Substitution into equation 33 yields

$$\beta_{e,G} = \beta(1 + \gamma/\pi s_o). \tag{35}$$

(*ii*) Using true (stationary) distribution, $p_{st}(x)$ given by equations 8 and 9 becomes:

$$p_{st}(x) = \frac{x^{(\gamma/\pi s_o)} \exp\left[-(\beta x^2/2\pi s_o) \right]}{A} \tag{36}$$

where

$$A = \int_0^\infty x^{(\gamma/\pi s_o)} \exp\left(\frac{-\beta x^2}{2\pi s_o} \right) dx$$

$$= \frac{1}{2} \left(\frac{2\pi s_o}{\beta} \right)^{\frac{1}{2}[1 + (\gamma/\pi s_o)]} \Gamma\left[\frac{1 + (\gamma/\pi s_o)}{2} \right].$$

Therefore the mean square value of $x(t)$, using true distribution equation 36, is

$$\langle x^2 \rangle_T = \int_0^\infty x^2 p_{st}(x)\, dx$$

$$= \frac{2}{A}\left[\int_0^\infty (x)^{2+(\gamma/\pi s_o)} \exp\left(-\frac{\beta}{2\pi s_o} x^2\right) dx\right]$$

$$= \frac{1}{\beta}(\pi s_o + \gamma). \tag{37}$$

Substitution into equation 33 yields

$$\beta_{e,T} = \beta\left(1 + \frac{\gamma}{\pi s_o + \gamma}\right) = \beta\left(1 + \frac{\gamma}{\pi s_o} - \frac{\gamma^2}{\pi^2 s_o^2} + \cdots\right). \tag{38}$$

Comparison of equation 38 with equation 35 indicates that $\beta_{e,G}$ is the first-order approximation of $\beta_{e,T}$.

Now let us consider the simple linear system

$$\dot{x} + \beta_e x = f(t) \tag{39}$$

whose transfer function is given as

$$H(i\omega) = \frac{1}{\beta_e + i\omega}. \tag{40}$$

According to the familiar linear theory of random processes, the power spectrum of $x(t)$ in equation 39 is given by

$$S_x(\omega) = \frac{s_o}{\beta_e^2 + \omega^2} \tag{41}$$

where s_o is defined in equation 3.

Substituting $\beta_{e,G}$ as given in equation 35 into equation 41, we obtain

$$S_{x,G}(\omega) = \frac{s_o}{\beta^2[1 + (\gamma/\pi s_o)]^2 + \omega^2}. \tag{42}$$

Substituting $\beta_{e,T}$ as given in equation 38 into equation 41, we obtain

$$S_{x,T}(\omega) = \frac{s_o}{[\beta(\pi s_o + 2\gamma)/(\pi s_o + \gamma)]^2 + \omega^2}. \tag{43}$$

By setting $\gamma = 0$, both equations 42 and 43 give

$$S_x(\omega) = \frac{s_o}{\beta^2 + \omega^2}$$

which is the spectrum of the corresponding linear system. The mean square value and the power spectral density of the velocity response $x(t)$, obtained by solving the Fokker–Planck equation, and the equivalent linear equations using gaussian and true distributions, respectively, are summarized in Table I. In this table $k = \gamma/\pi s_o$ is the nonlinear coefficient.

The mean square velocities $\langle x^2 \rangle$, $\langle x^2 \rangle_G$, and $\langle x^2 \rangle_T$ are compared in Fig. 2. It is seen that for $k < 0.035$, that is, in a very small nonlinearity range, both linearization cases give higher mean square velocities than the exact solutions. For $k > 0.035$, equivalent linearization methods give larger results when using true distribution and smaller results when using gaussian distribution than the exact solutions.

The power spectral density functions $S_{x,G}(\omega)$ and $S_{x,T}(\omega)$ obtained by the equivalent linearization procedure, using gaussian distribution and true distribution of $x(t)$, respectively, are compared in Fig. 3 in which $B_1 = (\beta^2/s_o)S_{x,G}(\omega)$ and $B_2 = (\beta^2/s_o)S_{x,T}(\omega)$. Notice that both $S_{x,G}(\omega)$ and $S_{x,T}(\omega)$ are monotonically decreasing functions, and the differences between them are negligible for small nonlinearity.

The exact power spectral densities are compared with the equivalent linearized solutions in Fig. 4. For the exact solution of $S_x(\omega)$ as given by equation 25, the spike at $\omega = 0$ is evaluated by normalizing $S_x(\omega)$ to an area equal to that given by Fig. 2, that is,

$$\langle x^2 \rangle = \int_0^\infty S_x(\omega)\, d\omega = 3.053 \quad \text{for} \quad k = 0.01.$$

Curves shown in Fig. 4 are seen to be monotonically decreasing. The equivalent linearized systems have higher power spectra of $x(t)$ in the low-frequency region and lower power spectra of $x(t)$ in the high-frequency region than the actual nonlinear system has.

IV. CONCLUSION

It has been shown that the exact expression for the two-dimensional nonstationary probability distribution of a class of simple nonlinear systems can be found in terms of the spatial eigenfunction expansion of the governing Fokker–Planck equation. Equivalent linearization techniques can be very useful in generating approximate response statistics for certain systems having small nonlinearities. In Figs. 2, 3, and 4 good agreement has been achieved in the comparison of the exact and approximate mean square values and of the power spectral densities of the nonlinear random response.

TABLE I—COMPARISON OF RESPONSE STATISTICS OF THE VELOCITY RESPONSE $x(t)$

	Mean square velocity, $\frac{\beta}{s_o}\langle x^2\rangle$	Power spectral density of velocity, $\frac{\beta^2}{s_o}S_x(\omega)$
Exact solution	$\frac{1}{2}\frac{\Gamma^2[2\mu+(3/2)]}{\Gamma(2\mu+1)}\sum_{n=0}^{\infty}\frac{\Gamma^2[n-(1/2)]}{(n!)^3\Gamma(n+2\mu+1)}$	$\frac{\Gamma^2[2\mu+(3/2)]}{\Gamma(2\mu+1)}\left[\frac{2\pi s_0\delta(\omega)}{\Gamma(2\mu+1)}+\frac{1}{4\pi}\sum_{n=1}^{\infty}\frac{4n\,\Gamma^2[n-(1/2)]}{(n!)^3\Gamma(n+2\mu+1)\,[4n^2+(\omega/\beta)^2]}\right]$
Equivalent linearization using gaussian density	π	$\frac{1}{(1+k)^2+(\omega/\beta)^2}$
Equivalent linearization using true density	$\pi(1+k)^*$	$\frac{1}{[(1+2k)/(1+k)]^2+(\omega/\beta)^2}$

* $k=\gamma/\pi s_o$.

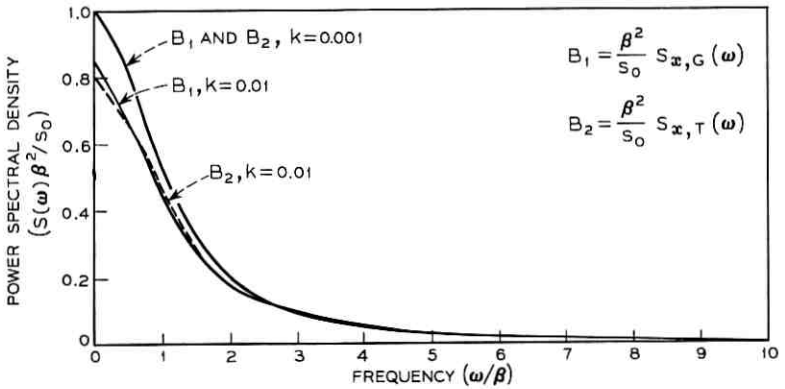


Fig. 3 — Power spectral density by equivalent linearization.

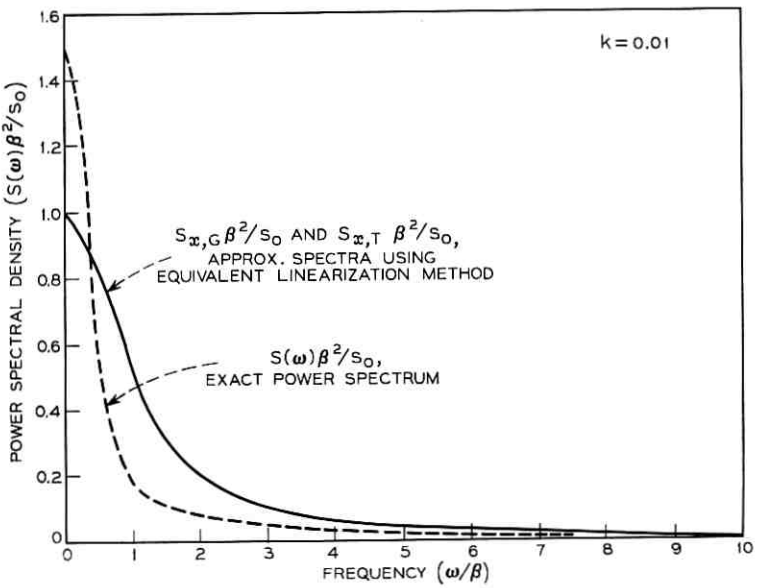


Fig. 4 — Comparison of the exact and approximate power spectra.

V. ACKNOWLEDGMENT

The author wishes both to acknowledge that Miss M. A. Wamp of Bell Telephone Laboratories, Whippany, New Jersey, generated the numerical results for Figs. 2, 3, and 4, and to express his appreciation.

APPENDIX

Derivation of Equation 23

The following formulae are used in the derivation of equation 23:

$$\int_0^\infty e^{-st} t^\beta L_n^\alpha(t)\, dt$$

$$= \frac{\Gamma(\beta+1)\Gamma(\alpha+n+1)}{n!\,\Gamma(\alpha+1)}\, s^{-\beta-1} F\left(-n, \beta+1; \alpha+1; \frac{1}{s}\right)$$

$$(\operatorname{Re}\beta > -1, \operatorname{Re} s > 0) \qquad (44)$$

where

$$F(\alpha, \beta; \gamma; z) = {}_2F_1(\alpha, \beta; \gamma; z) \qquad (45)$$

is a generalized hypergeometric series which is defined as

$${}_pF_q(\alpha_1, \alpha_2, \cdots \alpha_p; \beta_1, \beta_2 \cdots \beta_q; z) = \sum_{k=0}^{\infty} \frac{(\alpha_1)_k(\alpha_2)_k \cdots (\alpha_p)_k}{(\beta_1)_k(\beta_2)_k \cdots (\beta_q)_k} \frac{z^k}{k!}$$

in which

$$(\alpha)_k = \frac{\Gamma(\alpha+k)}{\Gamma(\alpha)}$$

and

$$(\beta)_k = \frac{\Gamma(\beta+k)}{\Gamma(\beta)}. \qquad (46)$$

A special case for equation 45 is when $z = 1$:

$$F(\alpha, \beta; \gamma; 1) = \frac{\Gamma(\gamma)\Gamma(\gamma-\alpha-\beta)}{\Gamma(\gamma-\alpha)\Gamma(\gamma-\beta)}$$

$$(\operatorname{Re}\gamma > \operatorname{Re}(\alpha+\beta), \operatorname{Re}\gamma > \operatorname{Re}\beta > 0). \qquad (47)$$

Using equations 44 through 47, the integral involved in equation 22 can be evaluated as follows:

$$\int_0^\infty z^{2\mu+(1/2)} e^{-z} L_n^{(2\mu)}(z)\, dz$$

$$= \frac{\Gamma[2\mu+(1/2)+1]\Gamma(2\mu+n+1)}{n!\,\Gamma(2\mu+1)} F(-n, 2\mu+\tfrac{3}{2}; 2\mu+1; 1)$$

$$= \frac{\Gamma[2\mu + (3/2)]\Gamma(2\mu + n + 1)}{n!\,\Gamma(2\mu + 1)} \frac{\Gamma(2\mu + 1)\Gamma[2\mu + 1 + n - 2\mu - (3/2)]}{\Gamma(2\mu + 1 + n)\Gamma[2\mu + 1 - 2\mu - (3/2)]}$$

$$= \frac{\Gamma(2\mu + (3/2)]\Gamma[n - (1/2)]}{n!\,\Gamma[-(1/2)]} = \frac{-\Gamma[2\mu + (3/2)]}{2(\pi)^{\frac{1}{2}}} \frac{\Gamma[n - (1/2)]}{n!}$$

which is equation 23.

REFERENCES

1. Hempstead, R. D. and Lax, M., "Classical Noise VI. Noise in Self-Sustained Oscillators Near Threshold," Phys. Rev., *161*, No. 2 (September 1967), pp. 350–366.
2. Lax, M., "Classical Noise V. Nose in Self-Sustained Oscillators," Phys. Rev., *160*, No. 2 (August 1967), pp. 290–307.
3. Ariaratnam, S. T. and Sanker, T. S., "Dynamic Snap-Through of Shallow Arches under Stochastic Loads," J. of Amer. Instit. of Aeronautics and Astronautics, *6*, No. 5 (May 1968), pp. 798–802.
4. Caughey, T. K. and Dienes, J. K., "Analysis of a Nonlinear First-Order System With a White Noise Input," J. Appl. Phys., *32*, No. 11 (November 1961), pp. 2476–2479.
5. Stratonovich, R. L., *Topics in the Theory of Random Noise*, vol. I, New York: Gordon and Breach, Science Publishers, Inc., 1963, chapter 4.
6. Crandall, S. T., "Perturbation Techniques for Random Vibration of Nonlinear Systems," J. Acoust. Soc. Amer., *35*, No. 11 (November 1963), pp. 1700–1705.
7. Khabbaz, G. R., "Power Spectral Density of the Response of a Nonlinear System to Random Excitation," J. Acoust. Soc. Amer., *38*, No. 5 (November 1965), pp. 847–850.
8. Booton, R. C., "The Analysis of Nonlinear Control Systems With Random Inputs," Proc. Symp. Nonlinear Circuit Anal., Polytechnic Inst., Brooklyn, New York, *2* (1953).
9. Caughey, T. K., "Equivalent Linearization Techniques," J. Acoust. Soc. Amer., *35*, No. 11 (November 1963), pp. 1706–1711.

Statistical Analysis and Stochastic Simulation of Ground-Motion Data

By S. C. LIU

(Manuscript received July 31, 1968)

The time variations of the root-mean-square accelerations, the autocorrelation functions, and the power spectral density functions of 12 strong-motion earthquake accelerograms are analyzed. The results indicate that: (i) strong-motion accelerograms of sufficiently long duration are stationary in the rms sense, (ii) the stationary rms acceleration is a good measurement of earthquake intensity, and (iii) the autocorrelation and power spectral density functions of strong-motion accelerograms resemble those of a narrowband process. Based upon these results, a method of determining the transfer characteristics of a site is introduced. A procedure for generating a filtered, gaussian stationary process to simulate ground motions is developed, and two applications of this simulation procedure illustrate its significance.

I. INTRODUCTION

For many years structural engineers have been concerned with the dynamic response of structural systems subjected to seismic excitations. Ground motions may be caused by natural earthquakes, by underground explosions, or by nuclear air blasts. Structures such as high-rise buildings, nuclear reactor facilities, or sensitive equipment in the vicinity of such events are vulnerable to induced random-type disturbances. Traditionally, deterministic methods of analysis relying on the known earthquake response spectra have been used.[1] These methods have provided valuable information regarding the behavior of structures during earthquakes. However, this procedure has a serious restriction in that only a few strong-motion accelerograms exist which provide ground-motion input. An earthquake is usually initiated by a series of irregular slippages along faults, followed by

innumerable random reflections, refractions, dispersions, and attenuations of the seismic waves within the complex ground formations through which they travel.

Since ground motions are generally random, a probabilistic method of analysis appears to be more appropriate than the traditional method of establishing a reliable design basis for structures subjected to ground motions. The simulation of ground motion is undoubtedly a necessary step in performing such a probabilistic analysis. Because earthquakes are unpredictable, some researchers in structural and earthquake engineering have attempted in recent years to simulate earthquakes by using stochastic processes. Both stationary and nonstationary models have been investigated.[2-8]

It is the purpose of this paper to investigate the best characterization of ground motions and to establish a valid basis for the stochastic simulation of seismic records. For this purpose, 12 commonly used strong-motion earthquake accelerograms are analyzed. The time variations of the rms accelerations, the autocorrelation functions, and the power spectral density functions of these accelerograms are investigated. The stationary rms accelerations are used as a measure of earthquake intensities and are compared with those found by Housner.[9] From the power spectral density analysis of existing ground-motion records, a linear filter can be determined to represent the transfer characteristics of the ground layers at a particular site. This filter is used in developing a method of generating a gaussian stationary process to simulate ground-motion accelerations.

In Section III the generation of synthetic ground acceleration records using a digital computer is discussed. The synthetic records are generated from existing records and from estimated response spectra. Two practical examples, of importance to structural engineering, are illustrated in Section IV.

II. STATISTICAL ANALYSIS OF GROUND-MOTION DATA

In general, no individual record is representative of any other record in an ensemble. However, if the data are stationary, valuable statistics may be derived by averaging the existing records. If an ergodic process is considered, a single record will be sufficient to represent the entire process.

The mean value of a given time-history record $x(t)$ of duration T is defined by

$$\langle x \rangle_{av} = \frac{1}{T} \int_0^T x(t)\, dt. \tag{1}$$

The mean square value of $x(t)$ is defined by

$$\langle x^2 \rangle_{av} = \frac{1}{T} \int_0^T x^2(t)\, dt. \tag{2}$$

Following this definition, the root-mean-square or the rms value of $x(t)$ is the positive square root of the mean square value $\langle x^2 \rangle_{av}$. This is given by

$$\text{rms of } x(t) = \left[\frac{1}{T} \int_0^T x^2(t)\, dt \right]^{\frac{1}{2}} \tag{3}$$

If $x(t)$ is a stationary random process with zero mean, its autocorrelation function $R_x(\tau)$ and power spectral density function $S_x(\omega)$ are given by the following transform pair:

$$R_x(\tau) = \lim_{T \to \infty} \frac{1}{T} \int_0^T x(t)x(t + \tau)\, dt. \tag{4}$$

$$S_x(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} R_x(\tau) e^{-i\omega\tau}\, d\tau. \tag{5}$$

Notice that $R_x(\tau)$ is always a real-value even function with a maximum at $\tau = 0$ and that, if $\langle x \rangle_{av} = 0$,

$$R_x(0) = \int_{-\infty}^{\infty} S_x(\omega)\, d\omega = \lim_{T \to \infty} \langle x^2 \rangle_{av}\ , \tag{6}$$

that is, the maximum autocorrelation represents the mean square value of a stationary random process.

These simple statistical quantities defined in equations 1 through 5 are useful in applying random vibration theory in earthquake engineering. The mean value, mean square value, and root-mean-square value represent the time-average strength of the input function. The timt variations of $\langle x \rangle_{av}$, $\langle x^2 \rangle_{av}$, or rms of $x(t)$ can be used to test the restricted sense stationarity of a time series. The autocorrelation function and the power spectral density function are closely related to the second-order properties and are generally used as characterization functions of a stationary process. Since strong-motion earthquake accelerograms are, in general, gaussian,[5] either the autocorrelation functions or the power spectral density functions will be sufficient to provide a complete statistical description. These two functions also provide a mathematical basis for the random response analysis of linear structural systems.

If $x(t)$ is given in digitized form, equations 1 through 5 can be

written respectively as follows:

$$\langle x \rangle_{av} = \frac{1}{N} \sum_{k=1}^{N} x_k \,, \tag{7}$$

$$\langle x^2 \rangle_{av} = \frac{1}{N} \sum_{k=1}^{N} x_k^2 \,, \tag{8}$$

$$\text{rms of } x(t) = \left[\frac{1}{N} \sum_{k=1}^{N} x_k^2 \right]^{\frac{1}{2}}, \tag{9}$$

$$R_k = R_x(k\,\Delta t) = \frac{1}{N-k} \sum_{i=1}^{N-k} x_i x_{i+k} \,, \qquad k = 0, 1, 2, \cdots, m, \tag{10}$$

and

$$S_k = S_x(\omega) = \frac{\Delta t}{\pi} \left[R_o + 2 \sum_{j=1}^{m-1} R_j \cos \frac{\pi jk}{m} + (-1)^k R_m \right], \qquad k = 0, 1, 2, \cdots, m, \tag{11}$$

in which $N = T/\Delta t$ equals the digitization time interval Δt, $x_k = x(k\Delta t)$, and m represents the maximum lag number.

2.1 *Six Strong-Motion Earthquakes*

Using equations 7 through 11, six commonly used strong-motion earthquakes each with two horizontal components are analyzed. They are:

A. El Centro, California, December 30, 1934, N-S.
B. The same, but E-W.
C. El Centro, California, May 18, 1940, N-S.
D. The same, but E-W.
E. Olympia, Washington, April 13, 1949, S10E.
F. The same, but S80W.
G. Taft, California, July 21, 1952, S21W.
H. The same, but N69W.
I. Golden Gate Park, San Francisco, March 22, 1957, N10E.
J. The same, but S80E.
K. Alameda Park, Mexico City, May 11, 1962, N10°46′W.
L. The same, but N79°14′E.

The accelerograms for these earthquakes are presented in Fig. 1. Upon first inspecting these existing strong-motion accelerograms, one might conclude that they are nonstationary. However, this conclusion

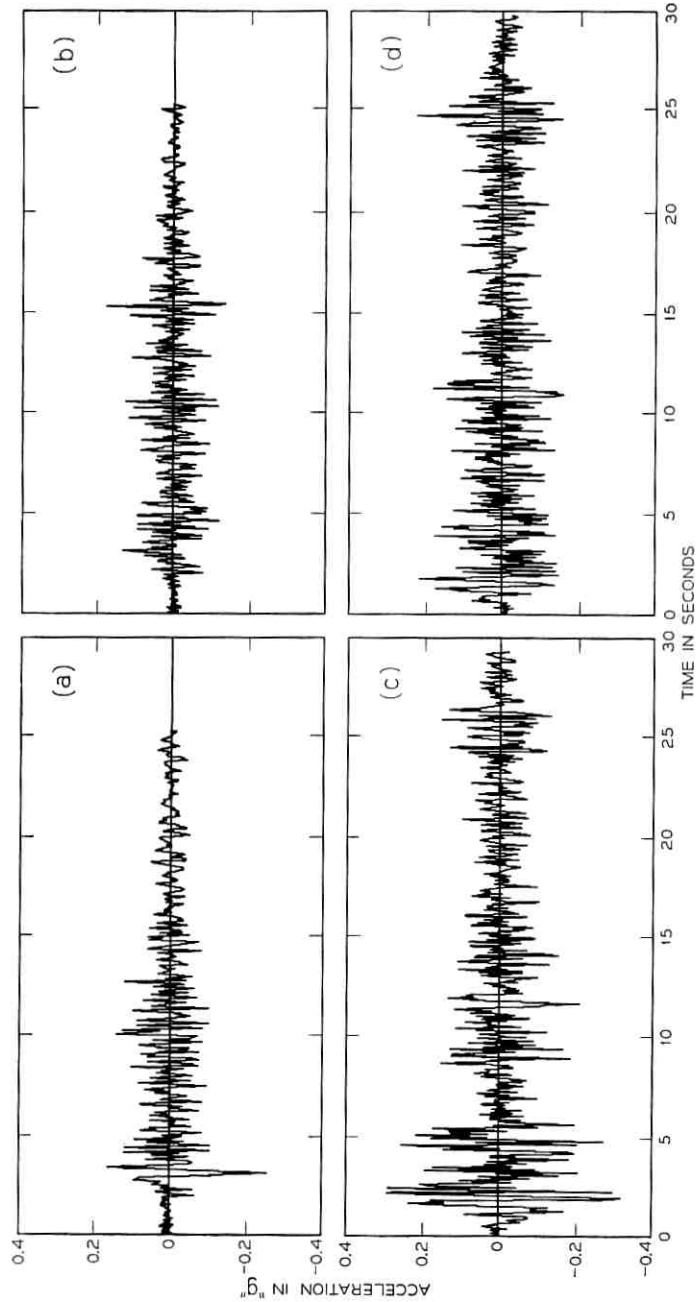
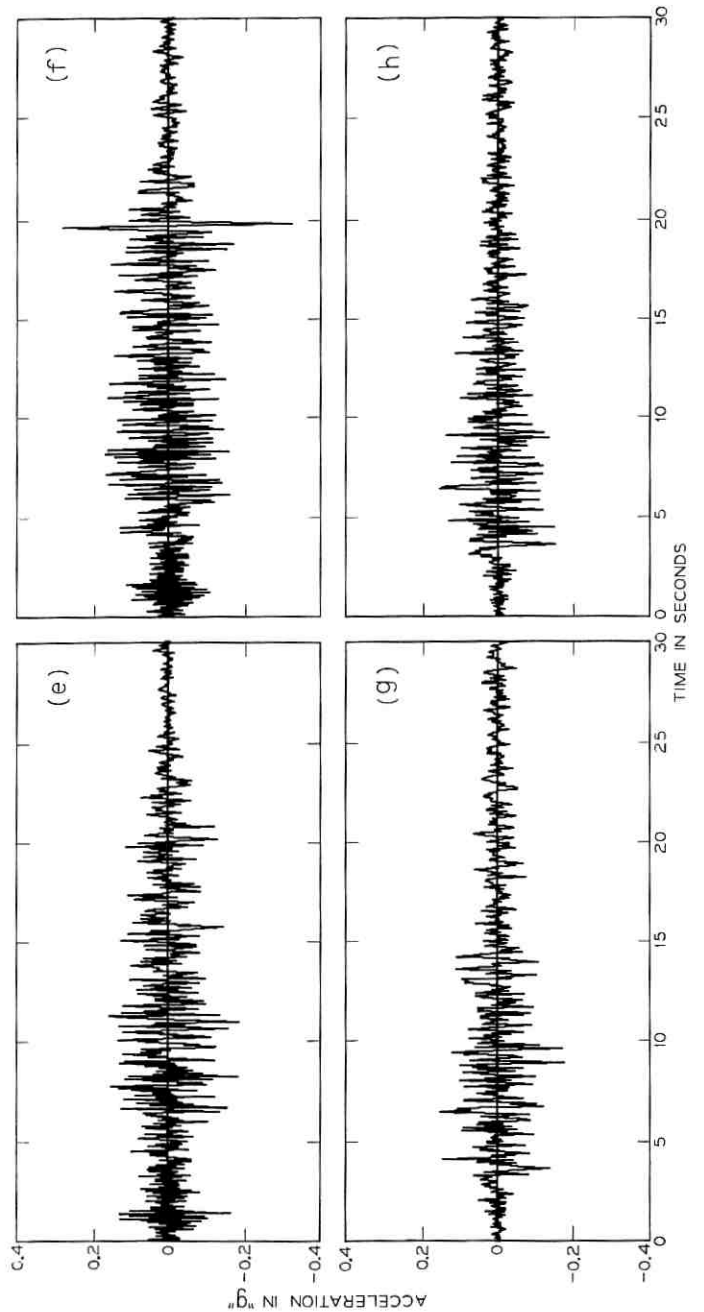
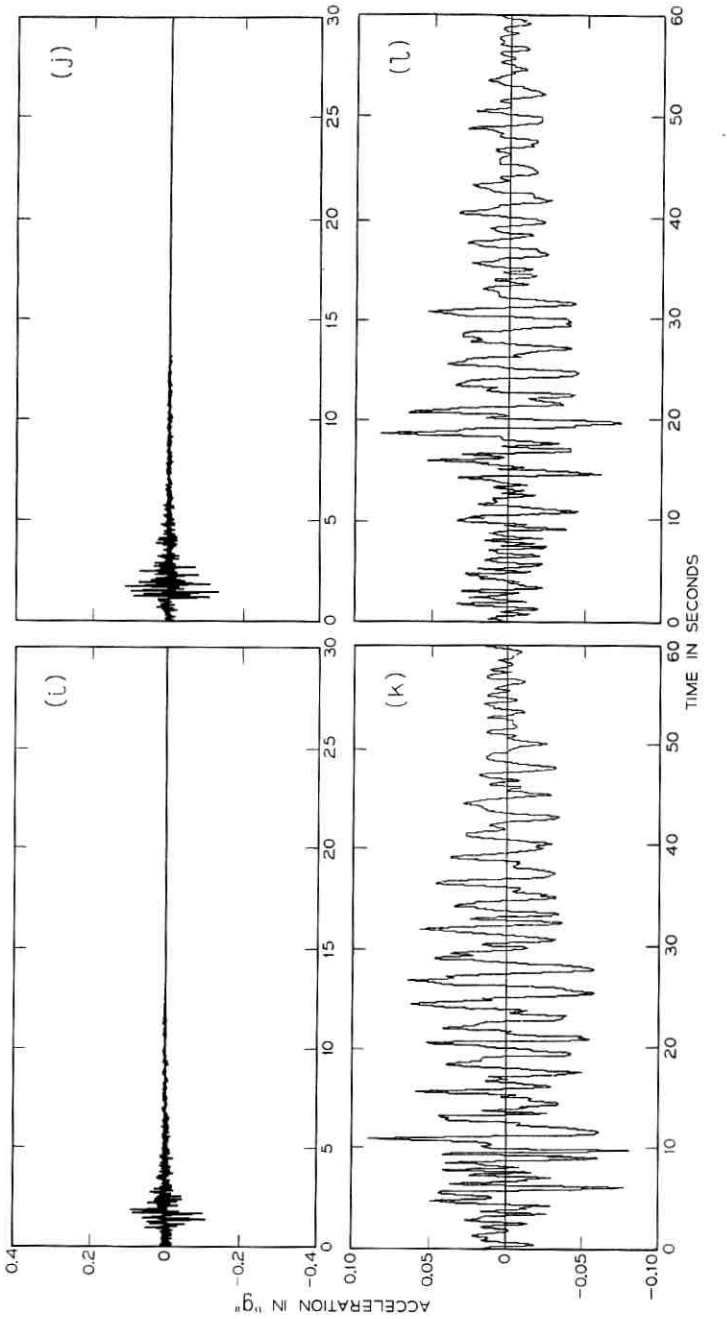


Fig. 1—Strong motion earthquake accelerograms. El Centro, Calif., Dec. 30, 1934, (a) N-S, (b) E-W; El Centro, Calif., May 18, 1940, (c) N-S, (d) E-W; Olympia, Wash., Apr. 13, 1949, (e) S10E, (f) S80W; Taft, Calif., July 21, 1952, (g) S21W, (h) N69W; Golden Gate Park, San Francisco, Calif., March 22, 1957, (i) N10E, (j) S80E; Alameda Park, Mexico City, May 11, 1962, (k) N10°46'W, (l) N79°14'E.


(e)
(f)
(g)
(h)
0.4
0.2
0
-0.2
-0.4
0
5
10
15
20
25
30
TIME IN SECONDS
ACCELERATION IN "g"


(i)
(j)
(k)
(l)
ACCELERATION IN "g"
TIME IN SECONDS
0.4
0.2
0
-0.2
-0.4
0.10
0.05
0
-0.05
-0.10
0
5
10
15
20
25
30
40
50
60

is debatable, considering the lack of sufficient ground-motion data for statistical studies, the difficulties of establishing valid nonstationary characteristics, and the ultimate objectives in developing a useful stochastic model for earthquake-induced ground accelerations.

The time variations of the rms amplitude of these strong-motion accelerograms are obtained and presented in Fig. 2, from which it is seen that the earthquake rms acceleration in general approaches a stationary value as the duration of the accelerogram is increased. This phenomenon indicates that strong-motion accelerograms are stationary in their rms amplitude at long duration ($T > 20$ seconds, approximately).

The longest and shortest periods existing in a digitized time-history record are $2T$ and $2\Delta t$, respectively. These two extremes constitute an effective period range for the record. Any analysis of the record beyond its effective period range will be of no significance. A Δt of 0.01 second was used in this study.

The stationary rms acceleration can be used to measure the intensity of an earthquake if it is accompanied by the corresponding effective period range. The rms intensities for all earthquakes, assuming $T = 20$ seconds (corresponding to a period range of 0 to 40 seconds), are listed in Table I. This table shows that the N-S component of the El Centro 1940 earthquake (C) has the strongest accelerogram.

It is interesting to compare earthquake rms intensities with the traditionally used Housner intensity,[9] which is defined as

$$SI_\lambda = \int_{0.1}^{2.5} S_v(\lambda, T_o)\, dT_o\,, \tag{12}$$

where S_v is the solution of the following equations:

$$\ddot{u} + \frac{4\pi}{T_o}\lambda\dot{u} + \frac{4\pi^2}{T_o^2}u = -\ddot{x}_g(t) \tag{13}$$

$$S_v(\lambda, T_o) = \max |\dot{u}| \tag{14}$$

for constant coefficients λ and T_o.

Physically, equation 13 represents the equation of motion of a basic single-degree-of-freedom linear system with mass m, viscous damping λ, and stiffness k subjected to earthquake excitation $\ddot{x}_g(t)$ at the support, as shown in Fig. 3. Correspondingly, $T_o = 2\pi\ (m/k)^{1/2}$ is the natural period of the system, and S_v in equation 14 is the maximum relative velocity reached by this system during the earthquake excitation. Therefore, Housner's intensity definition actually represents an

earthquake's potential peak velocity of structures in the period range 0.1 to 2.5 seconds. In Table II the rms (normalized to a factor of 2.7/2.20) and Housner's intensities are compared and good agreements are observed. The basic difference between these two intensity definitions is that the rms intensity is independent of the transfer

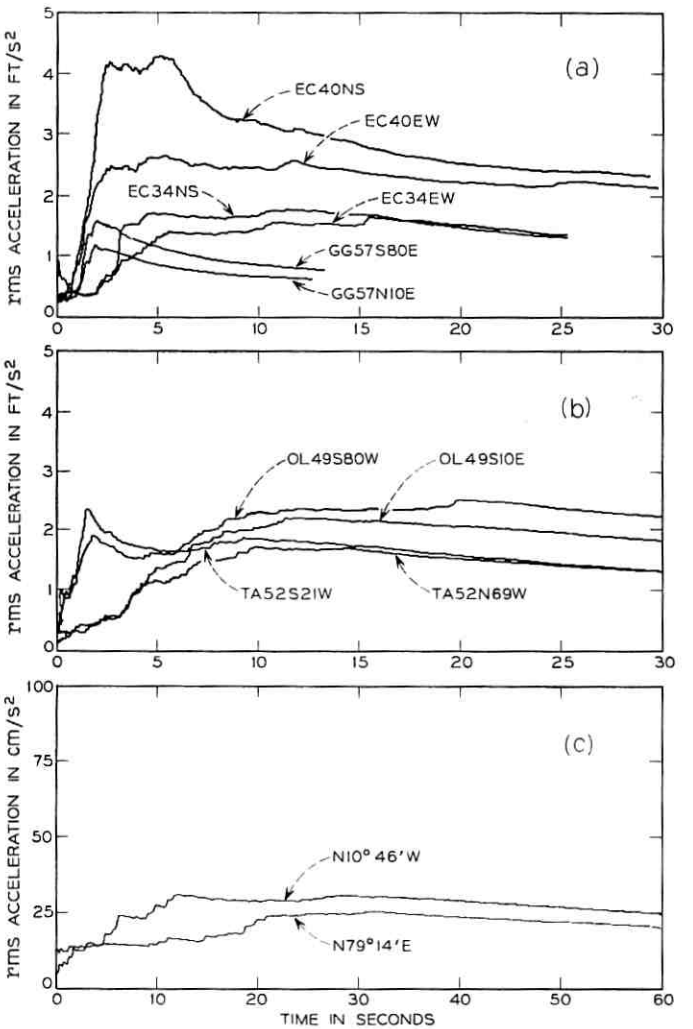


Fig. 2 — Root-mean-square acceleration of strong-motion earthquakes. (a) El Centro and Golden Gate, (b) Olympia and Taft, (c) Alameda Park.

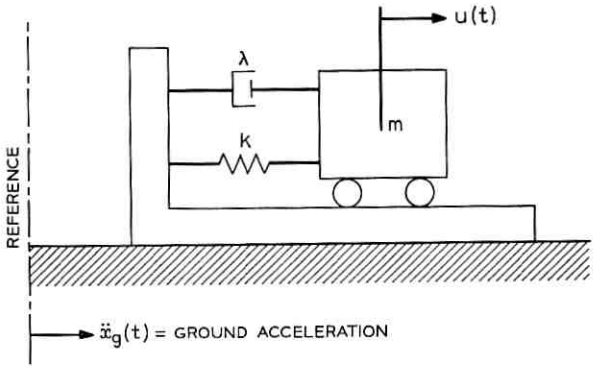


Fig. 3 — Linear mechanical system subjected to ground excitation.

characteristics of the structure while Housner's intensity depends on both the input and the transfer characteristics of the structure. It is much simpler to calculate the rms value for a given input time- history function than to find the SI_λ values, which require numerous mathematical integrations on the computer. However, the rms intensity should be used with care; one should consider the associated effective period range and the predominant period contained in the waveform.

2.2 *Determination of the Predominant Period*

The autocorrelation and power spectral density for all accelerograms considered are found on a digital computer by using a Fortran program

TABLE I — RMS INTENSITIES OF STRONG-MOTION EARTHQUAKES

Case	Identification	RMS acceleration at T = 20 s (ft/s²)	Average rms (ft/s²)
A	EC34NS	1.4	1.4
B	EC34EW	1.4	
C	EC40NS	2.4	2.20
D	EC40EW	2.0	
E	OL49S10E	1.7	1.85
F	OL49S80W	2.0	
G	TA52S21W	1.4	1.4
H	TA52N69W	1.4	
I	GG57S80E	0.8*	0.7*
J	GG57N10E	0.6*	
K	AL62N10°46′E	0.75	0.725
L	AL62N70°14′E	0.7	

* Taken at T = 10.0 s.

based on equations 10 and 11. Those functions, $R(\tau)$ and $S(\omega)$, for the El Centro 1940 earthquake are shown in Figs. 4 and 5 for the N-S and E-W components, respectively. The smooth curves in the power spectral density diagrams (Figs. 4b and 5b) were obtained by introducing a Hanning or smoothing procedure[10] to the raw estimates given by equation 11. The autocorrelation functions (Figs. 4a and 5a) all have a maximum at $\tau = 0$ and diminish rapidly at large correlation time. The ordinates of the power spectral density functions (Figs. 4b and 5b) generally increase with increasing frequencies to a maximum value at some frequency which may be considered a predominant or characteristic ground frequency, and then decrease rather rapidly toward zero in an asymptotic manner. Also of interest is the fact that this general rise and fall of the power spectral density function is accompanied by local random fluctuations.

The above results indicate that the autocorrelation and power spectral density of strong-motion earthquake accelerograms resemble those of a narrowband process. This implies that earthquake acceleration may be simulated by passing a wideband process through a linear filter which reflects the local geological conditions.

III. STOCHASTIC SIMULATION OF GROUND-MOTION ACCELEROGRAMS

3.1 *Basic Requirements for Ground-Motion Model*

There are many occasions when a ground-motion model is required. Examples are the prediction of ground motion at a certain site where no past records are available, and statistical analyses of structural responses based upon very limited actual ground-motion records. The hypothesized models must, of course, possess the pertinent characteristics of real ground motions and must be supported by existing data. More importantly, these models must properly reflect the damage (or response) potential of future ground motions to a wide range of structures.

TABLE II — COMPARISON OF EARTHQUAKE INTENSITIES

Earthquake case	A & B	C & D	E & F	G & H	I & J	K & L
Normalized rms intensity	1.7	2.7	2.2	1.7	0.84	0.9
Housner intensity	1.9	2.7	1.9	1.6	unavailable	

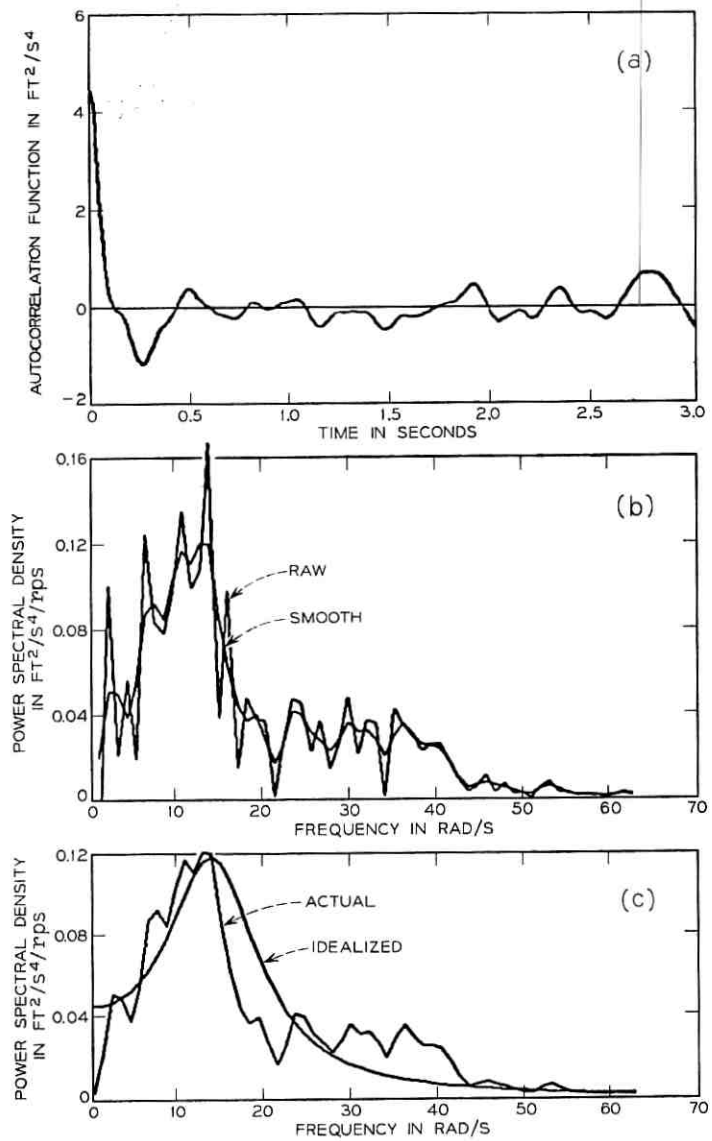


Fig. 4 — N-S component of earthquake at El Centro, Calif., May 18, 1940. (a) Autocorrelation function. (b) Power spectral density function. (c) Comparison of actual and idealized power spectral density functions.

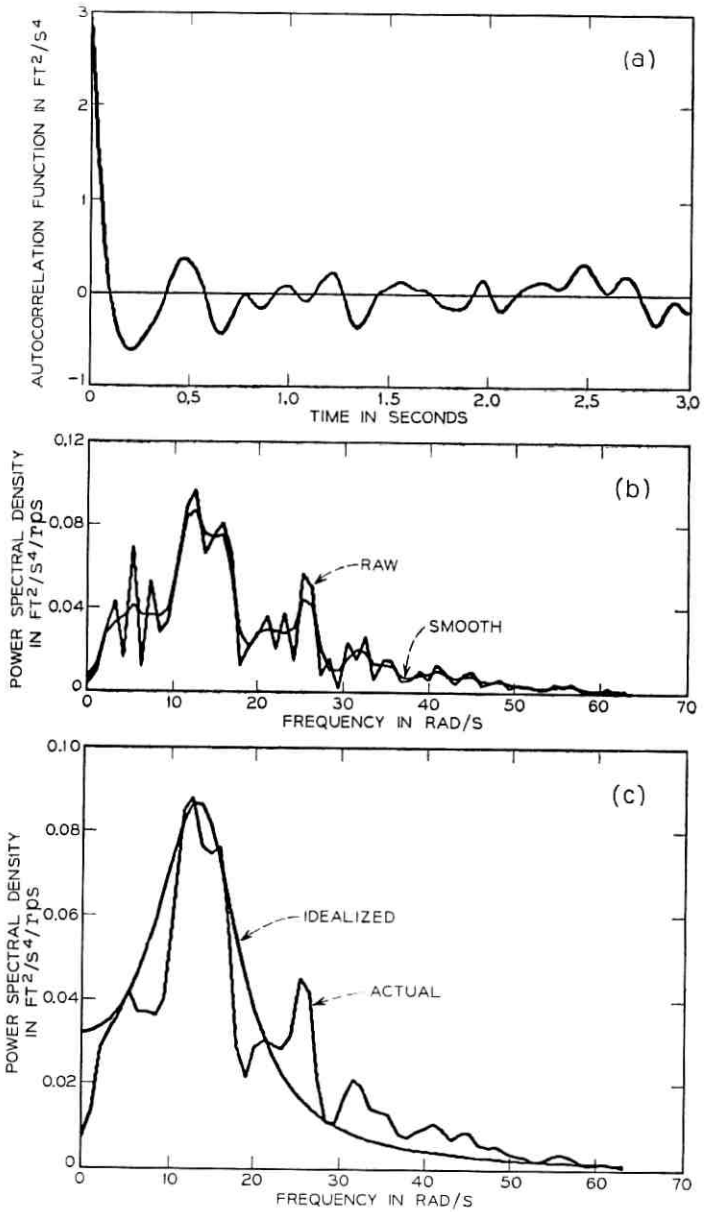


Fig. 5 — E-W component of earthquake, El Centro, Calif., May 18, 1940. (a) Autocorrelation function. (b) Power spectral density function. (c) Comparison of actual and idealized power spectral density functions.

The random phenomenon observed in earthquake ground motions suggests that one can reasonably assume that the resultant seismic wave arriving at the surface of the ground will contain a random assemblage of velocity impulses having gaussian distribution. Furthermore, the analyses of the earthquake accelerograms in the previous sections show that they have characteristics resembling those of a narrowband process. It is therefore postulated that the use of stochastic processes would be appropriate in modeling earthquake ground motions.

Basic criteria for establishing such a stochastic model can be specifically stated as follows:

(*i*) The model must have the basic properties reflected by the past recorded data such as the intensity, duration, general physical appearance, and all important characteristics resulting from local geological conditions.

(*ii*) The response statistics of the stochastic model must be equivalent to those produced by the real ground motion or to those predicated on strong theoretical or empirical bases.

Some simplifying assumptions required in the development of our stochastic model for earthquakes are:

(*i*) The input seismic wave transmited at bedrock by an earthquake is represented by stationary, white noise.

(*ii*) The ground layers of a seismic station during the shock are represented by a single-degree-of-freedom system with linear behavior. (A more sophisticated ground-layer filter may also be used.)

(*iii*) Local random fluctuations appearing in the power spectral density function of the real earthquake accelerogram are neglected when modeling the transfer characteristics of the site.

The representation of ground layers by a single-degree-of-freedom system is shown in Fig. 6. Such a simple system is characterized by its transfer function $h(\tau)$ in the time domain or $H(i\omega)$ in the frequency domain, with $h(\tau)$ and $H(i\omega)$ known as the unit impulse response and the complex frequency response of a linear system, respectively.

If the simple mechanical system shown in Fig. 3 is used as the linear filter, the power spectral density of the total acceleration of the mass, when the acceleration at the support is taken as the input, can be found in terms of the corresponding transfer function.

$$S_o(\omega) = | H(i\omega) |^2 S_i = \frac{S_i[1 + 4\lambda_g^2(\omega/\omega_g)^2]}{[1 - (\omega/\omega_g)^2]^2 + 4\lambda_g^2(\omega/\omega_g)^2} \tag{15}$$

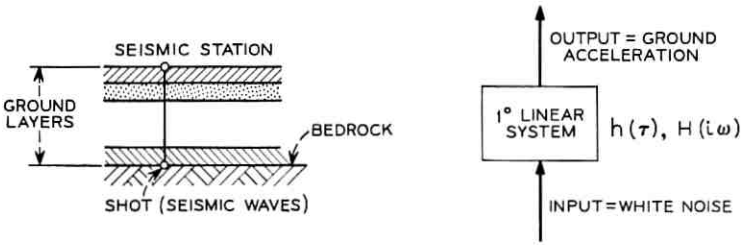


Fig. 6 — Ground layers represented by a linear system.

where S_i is the constant power spectral density of the input acceleration by assumption (*i*); λ_g and ω_g are the ground characteristic damping and frequency, respectively.

For known S_i , λ_g , and ω_g , equation 15 gives a smooth curve for $S_o(\omega)$. The location of the peak amplitude of $S_o(\omega)$ is determined by ω_g . The shape or the rate of the rise and fall of $S_o(\omega)$ is governed by λ_g , and the relative amplitude of $S_o(\omega)$ depends on S_i . For any power spectral density derived from a real ground-motion record, an idealized, smooth equivalent power spectral density can be found by using equation 15 and by properly adjusting the three-characteristic constant λ_g , ω_g , and S_i in this equation. The idealized power spectral density function for both horizontal components of the El Centro, California 1940 earthquake were found by the prescribed procedure. As shown in Figs. 4c and 5c for both cases, the idealized power spectral density function covers the same area as the actual power spectral curve in the frequency range 0 through 10 cps. The ω_g , λ_g , and S_i values used for the N-S and E-W components of the El Centro 1940 earthquake are 15.5 rad/s, 0.42, 0.046 ft^2/s^4/rps, and 14.7 rads/s, 0.41, 0.033 ft^2/s^4/rps, respectively.

Once the representative ground-layer filter for a given ground motion is determined, the artificial earthquake can be generated by passing white noise through this filter and measuring the output time history.

3.2 *Generation of a Gaussian Stationary Process*

This procedure starts by sampling a sequence of pairs of statistically independent random numbers $x_1, x_2; x_3, x_4; \cdots; x_{n-1}, x_n$ all of which have a uniform probability distribution over the range $0 < x < 1$. A new sequence of pairs of statistically independent random numbers $y_1, y_2; y_3, y_4; \cdots; y_{n-1}, y_n$ are then generated using the relations

$$y_i = (-2 \log_e x_i)^{\frac{1}{2}} \cos 2\pi x_{i+1} \qquad i = 1, 3, \cdots, n-1 \qquad (16)$$

$$y_{i+1} = (-2 \log_e x_i)^{\frac{1}{2}} \sin 2\pi x_{i+1}$$

which have been shown to have a gaussian distribution with a mean of zero and a variance of unity.[11-13]

A single waveform $y(t)$ can now be established by assigning the values $y_1, y_2 \ldots, y_n$ to n successive ordinates spaced at equal intervals Δt along a time abscissa and by assuming a linear variation of ordinates over each interval. To define a time origin, assume that the initial ordinate y_o, which is taken equal to zero, is located at $t = t_o$ where t_o is a random variable having a uniform probability density function of intensity $1/\Delta t$ over the interval $o < t_o < \Delta t$.

For practical reasons Δt must, of course, be taken as finite; however, its value should be set sufficiently small so that the true power spectral density function is reasonably constant at intensity S_i over the lower range of frequencies which are to be properly represented in the process. A value of 0.025 second or smaller is recommended.

To establish the desired stationary process $a(t)$, each member $y_r(t)$ $(r = 1, 2, \cdots, N)$ of the normalized process $y(t)$ must be filtered in accordance with equation 15. This step can be accomplished by assuming that a simple single-degree-of-freedom system having an undamped circular frequency ω_o and a damping ratio λ_o is subjected separately to the N support accelerations $y_r(t)$ and then by calculating the corresponding absolute or total acceleration functions $a_r(t)$ of the mass. In mathematical form this statement is equivalent to saying that one must solve the differential equations

$$\ddot{Z}_r(t) + 2\omega_o\lambda_o\dot{Z}_r(t) + \omega_o^2 Z_r(t) = -y_r(t) \qquad r = 1, 2, \cdots, N \qquad (17)$$

for the functions $\ddot{Z}_r(t)$ and then evaluate the desired family of acceleration functions $a_r(t)$ using the relation

$$a_r(t) = \ddot{Z}_r(t) + y_r(t) \qquad r = 1, 2, \cdots, N. \qquad (18)$$

Equation 17 can be solved numerically on a digital computer using the standard linear acceleration method. Based on the assumption that the input earthquake duration is divided into the very short equal time intervals Δt, and the response acceleration varies linearly over each interval, this method gives the following simple relations between the response displacement Z_r and its derivatives at step n and $n+1$:

$$(\dot{Z}_r)_{n+1} = (\dot{Z}_r)_n + \frac{(\ddot{Z}_r)_n}{2}\,\Delta t + \frac{(\ddot{Z}_r)_{n+1}}{2}\,\Delta t$$

$$(Z_r)_{n+1} = (Z_r)_n + (\dot{Z}_r)_n\,\Delta t + \frac{(\ddot{Z}_r)_n}{3}\,\Delta t^2 + \frac{(\ddot{Z}_r)_{n+1}}{6}\,\Delta t^2. \qquad (19)$$

IV. ILLUSTRATIONS OF GENERATION OF ARTIFICIAL ACCELERATION

4.1 *By Known Power Spectral Density*

As the first example, artificial earthquakes are generated using the prescribed approach, to simulate the average of U. S. strong-motion earthquakes observed at seismic stations having a firm soil foundation. Values of 15.6 rad/s for ω_g and 0.6 for λ_g, representative of such soil conditions, were used. A total of 50 artificial accelerograms ($N = 50$) were generated for process $a(t)$ with a duration of 30 seconds which corresponds to $\Delta t = 0.025$ second and $n = 1200$. The intensity S_i of the unfiltered "white noise" was set at 0.00614 ft²/s³ so that the mean velocity response spectrum curves for the filtered process $a(t)$ would give a "best fit" with the standard response spectrum curves published by G. Housner.[9] This intensity is slightly less than the value of 0.0063 ft²/s³ used by J. Penzien[14] in a previous investigation to correlate the mean velocity response spectrum curves for "white noise" with Housner's standard curves.

A typical sample member of artificial accelerograms generated is shown in Fig. 7. It is interesting that this accelerogram appears to be very similar to the real accelerograms in Fig. 1 except for the general

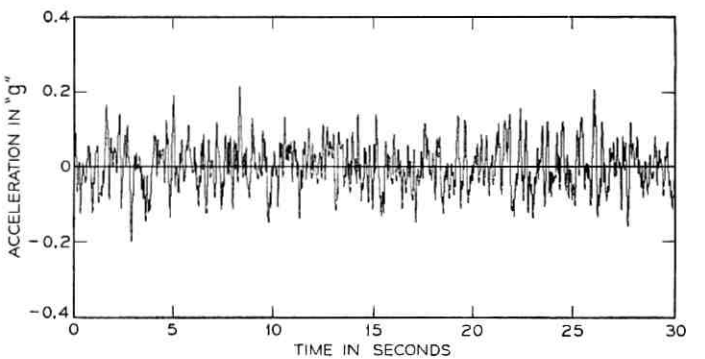


Fig. 7 — Sample member of artificial accelerogram.

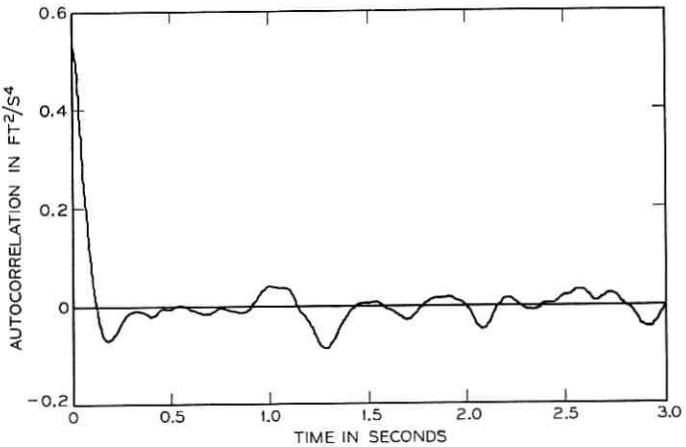


Fig. 8 — Autocorrelation function of the sample member of artificial accelerogram.

stationary apparance of the artificial accelerograms versus the nonstationary appearance of the real accelerograms.

The autocorrelation and corresponding power spectral density function for this sample artificial earthquake are shown in Figs. 8 and 9, respectively. It is of particular significance that, while these autocorrelations and power spectral densities are similar to those for the real earthquake (Figs. 4 and 5), local random fluctuations also appear in

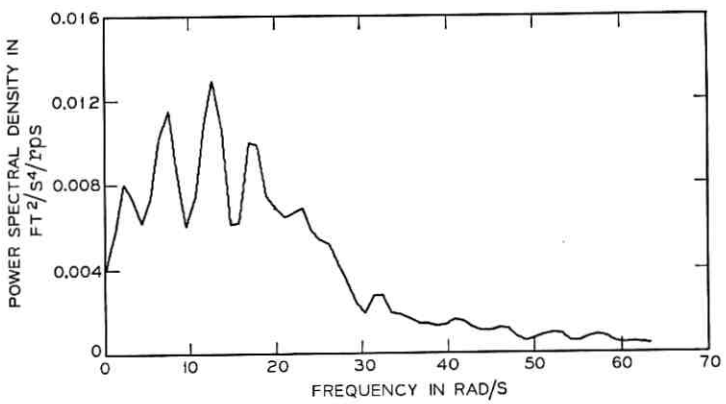


Fig. 9 — Power spectral density function of the sample member of artificial accelerogram.

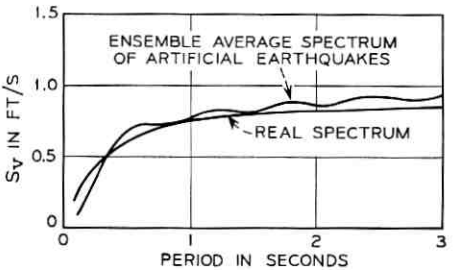


Fig. 10 — Comparison of velocity spectra of earthquakes, $\lambda = 0.02$.

the individual power spectral density function. It is expected that these local fluctuations and the variations from one individual power spectral density function to another will be eliminated by the averaging procedure. It has been shown that the average power spectral density function of artificial earthquakes is quite close to the prescribed function.[15]

The average response spectra ($\lambda = 2\%$) of 50 artificial earthquakes are compared with the real spectra given by Housner[9] in Fig. 10. Good agreements are observed over the significant period range. The confidence limits for these average spectra are given in Fig. 11 by extending 3σ (standard deviation) of S_v above and below the mean. Since the distribution of the response spectra is no longer gaussian, these limits correspond to an approximate 89 percent confidence band according to Chebyshev inequality.[16, 17]

The above results (Figs. 7 through 11) show that the basic require-

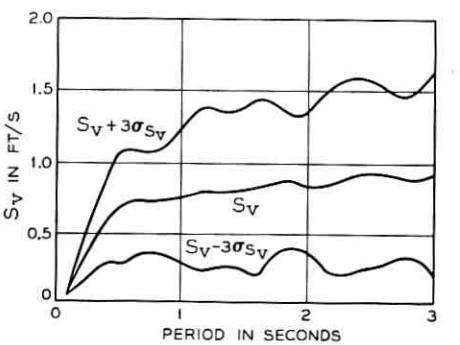


Fig. 11 — Variations and confidence limits for average velocity spectra of artificial earthquakes, $\lambda = 0.02$.

ments for ground-motion simulation have been satisfied. It takes approximately 10 seconds to generate one sample function on a CDC 6600 digital computer. These facts indicate that a good, reasonably economical ground-motion simulation has been achieved.

4.2 *By Known Response Spectrum*

In the second example, we have a situation somewhat different from the first. Engineers are asked to generate representative ground-motion accelerations from an estimated pseudovelocity spectra as given in Fig. 12. There is no past recorded ground-motion data available. The generated artificial earthquakes will be used to evaluate the probable damage of structures resulting from possible seismic events tak-

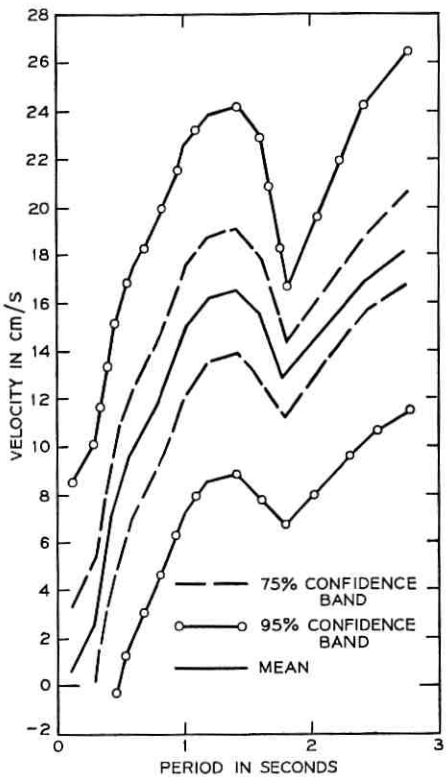


Fig. 12 — Pseudovelocity response spectra, $\lambda = 0.02$.

ing place in the vicinity of a proposed site. The estimated response spectra were obtained from a careful examination of earthquake activities in a certain seismic area and from an analogous prediction using records taken at stations having geological conditions similar or comparable to those of the given site. If the event refers to an underground explosion, the estimated response spectra might be predicted from past events under the same physical conditions (yield level and depth of the explosion, epicenter distance, and so on) and similar geological environments.

The given pseudoresponse spectrum curves (Fig. 12) show peaks at period $T_o \cong 1.25$ seconds if the long-period portion ($T_o > 2.5$ seconds) or the low-frequency portion ($\omega_0 < 2.5$ rad/s) is neglected. The corresponding frequency value where the peak spectrum amplitudes occur is approximately 5 rad/s. This value can be used as an approximate predominant ground frequency ω_g. The proposed site has a firm soil foundation, therefore a value of 0.6 for the characteristic ground damping λ_g is used.

Using these two characteristic values and letting S_i be unity in equation 15, the same procedure as used in the first example can be followed again. Five typical artificial accelerograms of 30-second duration are generated and shown in Fig. 13. The individual response spectra and the mean response spectra are obtained (Fig. 14). The general shape of the pseudoresponse spectra using $S_i = 1.0$ is similar to the estimated spectra. By matching the area covered by the estimated and pseudospectra curves, the amplitudes of the pseudospectra are normalized to have the same order of magnitude as the estimated spectra. The accelerations shown in Fig. 13 which have been modified by these same normalization procedures, represent the final form of the desired artificial accelerograms.

The comparisons of individual and mean pseudoresponse spectra with the given response spectra (Fig. 14) are satisfactory except for the high-period range ($T_o > 2.5$ seconds). If only those general building structures are considered which have a natural period range of 0.3 to 2.5 seconds, the large discrepancy of response spectra in the high-period range is not significant. The mean pseudoresponse spectrum (Fig. 14f) gives better results, as is expected. Overall results of the simulation may be improved with more accurate determination of values for ω_g, λ_g, and the larger sample size N of the simulating process. Methods for such improvements are presented in a separate paper.[18]

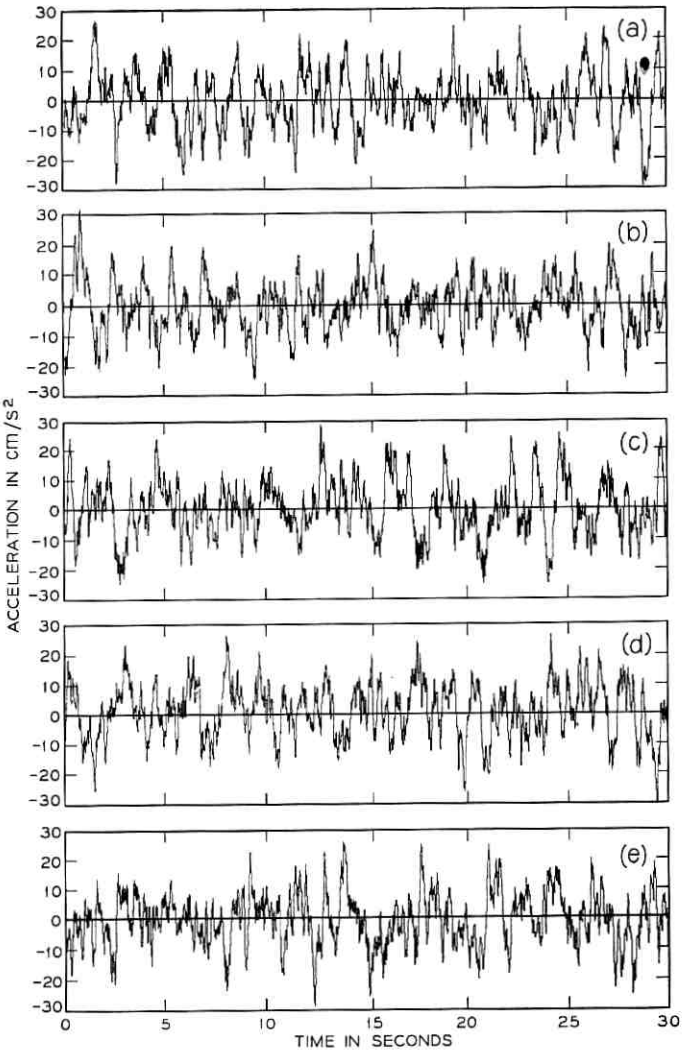


Fig. 13 — Pseudoearthquake accelerograms.

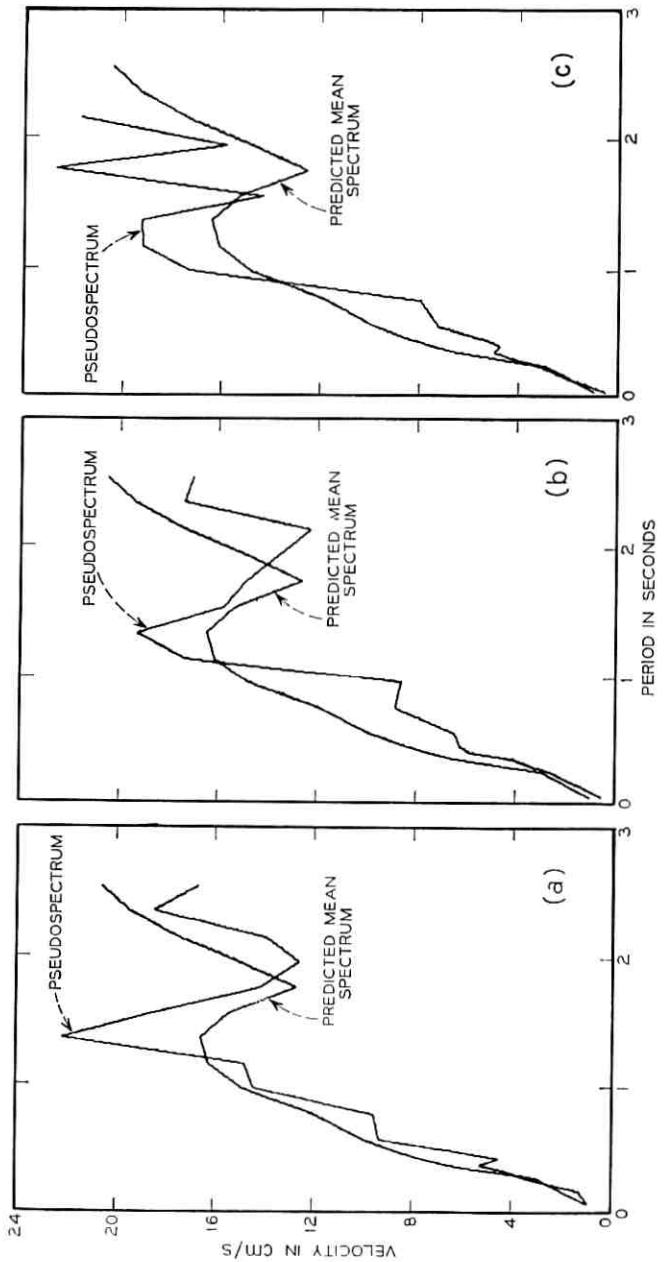
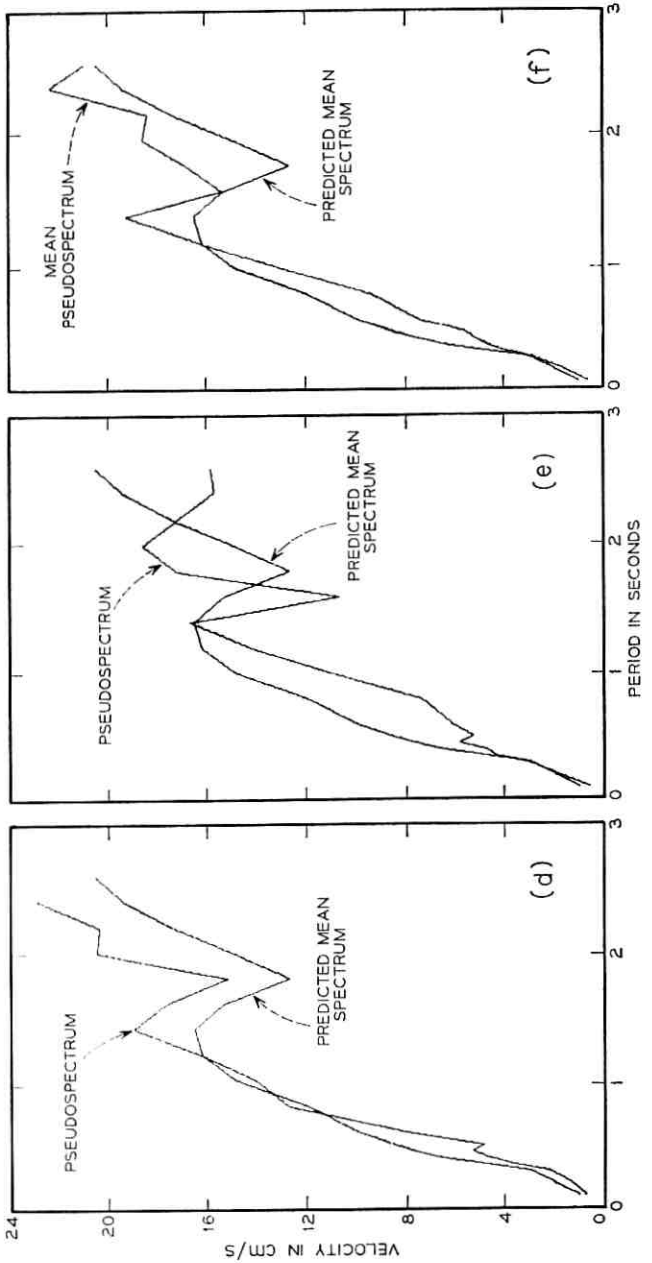


Fig. 14 — Comparisons of velocity response spectra, $\lambda = 0.02$.


VELOCITY IN CM/S
PERIOD IN SECONDS
(d)
(e)
(f)
PSEUDOSPECTRUM
PREDICTED MEAN SPECTRUM
MEAN PSEUDOSPECTRUM

V. CONCLUSIONS

From the results presented in this investigation, the following conclusions may be deduced:

(*i*) In general, strong-motion earthquake accelerograms of sufficiently long duration ($T > 20$ seconds, approximately) are stationary in the rms sense.

(*ii*) The rms acceleration of an earthquake, along with the effective period range and the predominant period of the associated accelerogram, can be used to determine the intensity of the earthquake.

(*iii*) The autocorrelation and the power spectral density of a strong-motion earthquake resemble those of a narrowband process.

(*iv*) The power spectral density analysis of existing earthquake records suggests that the transfer characteristics of the given site can be represented by a simple, linear system.

(*v*) A filtered, gaussian, stationary process generated on a digital computer proves to be successful in modeling ground motions induced by earthquakes or by nuclear underground explosions.

VI. ACKNOWLEDGMENTS

The author wishes to express his appreciation for the valuable suggestions given by Professor J. Penzien of The University of California at Berkeley, and for the careful review of the manuscript given by M. Oien, L. W. Fagel, and S. E. Wisnewski of Bell Telephone Laboratories.

REFERENCES

1. Alford, J. L. and Housner, G. W., "Spectrum Analyses of Strong-Motion Earthquakes," Research Rep., Earthquake Eng. Res. Laboratory, California Inst. of Tech., Pasadena, Calif., August 1951.
2. Bycroft, G. N., "White Noise Representation of Earthquakes," Proc. Amer. Soc. Civil Engineers, *86*, No. EM2 (April 1960), pp. 1–16.
3. Housner, G. W. and Jennings, P. C., Jr., "Generation of Artificial Earthquakes," Proc. Amer. Soc. Civil Engineers, *90*, No. EM1 (February 1964), pp. 113–150.
4. Bogdanoff, J. L., Goldberg, J. E., and Bernard, M. C., "Response of a Simple Structure to a Random Earthquake-Type Disturbance," Bull. Seismological Soc. Amer., *54*, No. 1 (February 1964), pp. 263–276.
5. Caughey, T. K. and Stumpf, H. J., "Transient Response of a Dynamic System Under Random Excitation," J. Appl. Mechanics, Trans. Amer. Soc. Mechanical Eng., *28*, No. E-4 (December 1961), pp. 563–566.
6. Amin, M. and Ang, A. H. S., "A Nonstationary Stochastic Model for Strong Motion Earthquakes," Technical Rep. *306*, Civil Engineering Studies, University of Illinois, April 1966.

7. Ward, H. S., "Analog Simulations of Earthquake Motions," Proc. Amer. Soc. Civil Engineers, *91*, No. EM5 (October 1965), pp. 173–190.
8. Shinozuka, M. and Sato, Y., "Simulation of Nonstationary Random Process," Proc. Amer. Soc. Civil Engineers, *93*, No. EM1 (February 1967), pp. 11–40.
9. Housner, G. W., "Behavior of Structures During Earthquakes," Proc. Amer. Soc. Civil Engineers, *85*, No. EM4 (October 1969), pp. 109–129.
10. Blackman, R. B. and Tukey, J. W., *The Measurement of Power Spectra*, New York: Dover Publications, Inc., 1958, pp. 14–37.
11. Box, G. E. P. and Miller, M. E., "A Note on the Generation of Random Normal Deviates," Ann. Math. Stat., *29*, No. 2 (June 1958), pp. 610–611.
12. Franklin, J. N., "Deterministic Simulation of Random Processes," Math. of Computation, *17*, No. 81 (January 1963), pp. 28–59.
13. Franklin, J. N., "Numerical Simulation of Stationary and Nonstationary Gaussian Random Processes," SIAM Review, *7*, No. 1 (January 1965), pp. 68–80.
14. Penzien, J., "Applications of Random Vibration Theory in Earthquake Engineering," Bull. Int. Inst. Seismology and Earthquake Eng., *2* (1965), pp. 47–69.
15. Liu, S. C., "Nondeterministic Analysis of Nonlinear Structures Subjected to Earthquake Excitations," Ph.D. dissertation, University of California, Berkeley, 1967.
16. Bendat, J. S., "Mathematical Analysis of Average Response-Values for Nonstationary Data." IEEE Trans. BioMedical Eng., *BME-11*, No. 3 (July 1964), pp. 72–81.
17. Bendat, J. S. and Thrall, G. P., "A Summary of Methods for Analyzing Nonstationary Data," NASA Technical Report 32–744, Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California, September 1, 1965, p. 4.
18. Liu, S. C. and Jhaveri, D. P., unpublished work.

Reliable Information Storage in Memories Designed from Unreliable Components*

By MICHAEL G. TAYLOR

(Manuscript received April 10, 1968)

This is the first of two papers which consider the theoretical capabilities of computing systems designed from unreliable components. This paper discusses the capabilities of memories; the second paper discusses the capabilities of entire computing systems. Both present existence theorems analogous to the existence theorems of information theory. The fundamental result of information theory is that communication channels have a capacity, C, such that for all information rates less than C, arbitrarily reliable communication can be achieved. In analogy with this result, it is shown that each type of memory has an information storage capacity, $\mathcal{C}$, such that for all memory redundancies greater than $1/\mathcal{C}$ arbitrarily reliable information storage can be achieved. Since memory components malfunction in many different ways, two representative models for component malfunctions are considered. The first is based on the assumption that malfunctions of a particular component are statistically independent from one use to another. The second is based on the assumption that components fail permanently but that bad components are periodically replaced with good ones. In both cases, malfunctions in different components are assumed to be independent. For both models it is shown that there exist memories, constructed entirely from unreliable components of the assumed type, which have nonzero information storage capacities.

I. INTRODUCTION

The problem of designing systems which operate reliably even though their components are unreliable has been formulated in many different ways. In a typical formulation, one considers some particular

* This work, which is based on part of a doctoral thesis submitted to the Department of Electrical Engineering, M.I.T., September 1966, was supported by the National Aeronautics and Space Administration (Grant NsG-334).

system which performs a computation with a nonzero probability of error. The problem is to design some other "reliable" system which performs the same computation using the same types of components but with a smaller probability of error. In fact, the ultimate objective is to show that it is possible to design systems, using only unreliable components, which perform computations with an arbitrarily small probability of error. Unfortunately, there is no standard terminology for describing these systems; therefore, the following section introduces the terminology to be used throughout this paper.

1.1 *Definitions*

The computations performed by the computing systems are described in terms of *elementary operations* where an elementary operation is any Boolean function of two binary operands. There are sixteen different elementary operations, each one of which can be represented by a binary matrix of the type shown in Fig. 1. Typical elementary operations are AND, OR, and modulo-2 addition. The computing systems to be considered are constructed from *components* which are devices that either perform one elementary operation or store one binary digit. The *complexity* of a system is defined to be equal to the number of components within the system.

In an irredundant computing system, the *amount of computation* performed by the system equals the number of elementary operations which are executed. Corresponding to each irredundant computing

AND	0	1
0	0	0
1	0	1

Fig. 1 — Binary matrix for AND operation. There are $2^4 = 16$ ways of filling this table, each one of which describes one of the 16 allowed Boolean functions.

system, there are many redundant computing systems which perform equivalent computations. These redundant systems are more complex than the equivalent irredundant one but, hopefully, they are also more reliable. The amount of computation performed by any one of these redundant computing systems is defined to be equal to the amount of computation performed by the corresponding irredundant system. Finally, the *redundancy* of a computing system equals the ratio of the complexity of the system to the amount of computation performed by the system.

To illustrate the use of these terms, consider a system that computes, in parallel, the modulo-2 sums of the digits in two k-digit sequences. The system first encodes each sequence of digits into a code word from an (n, k)* group code, then forms the modulo-2 sums of the digits in these code words, and finally decodes the result. The *amount of computation* equals k since an equivalent irredundant computer would simply perform k *elementary operations* each consisting of a module-2 sum. If the *complexities* of the encoder and decoder within the redundant system are C_E and C_D, respectively, the *complexity* of the entire system equals $C_E + C_D + n$, where the last term arises from the n modulo-2 adders required to perform the desired operation. The *redundancy* of this system equals $(C_E + C_D + n)/k$.

1.2 *Historical Background*

Von Neumann was one of the first to propose a system which uses redundancy to gain reliability.[1] He considered systems consisting of interconnections of identical elements† where all the elements compute either the majority function or the Sheffer-stroke function. The form (network topology) of the redundancy network is similar to that of the original irredundant network, the precursor. Specifically, each element in the precursor is replaced by a set of $3n$ elements of the same type in the redundant network (redundancy $= 3n$), and each interconnection is replaced by a bundle of n interconnections. The $3n$ elements in each set are interconnected in such a way that there are n outputs. It is assumed that a malfunction occurs in a particular set of elements whenever more than a certain fraction, θ, of the n outputs are in error; θ is chosen to minimize the probability of a malfunction within the entire system. Von Neumann showed that, for large n, the

* n is the length of each code word; k is the number of information digits in each word.

† The terms "element" and "network element" are used to indicate devices consisting of several components (some finite number).

probability that one set of $3n$ Sheffer-stroke elements malfunctions on one particular use is

$$\Pr(\text{malfunction in one set of elts.}) \cong 6.4/(n)^{\frac{1}{2}} \cdot 10^{-8.6n/10,000}$$

where it is assumed that the probability of error for each use of each element is 5.10^{-3}. Therefore, for this system, the probability of malfunction decreases exponentially with the redundancy, provided that the redundancy is sufficiently large. Other approaches involving the use of more complex modules have led to results similar to those of von Neumann.[2] In some cases the resulting network is more efficient (less redundant for a given probability of system failure) than von Neumann's network. However, in all cases, to achieve an arbitrarily small probability of system failure it is necessary to make the redundancy arbitrarily large.

It is interesting to compare von Neumann's results with those obtained by Shannon concerning the reliability of communication systems.[3] Both show that the probability of error within the system can be made arbitrarily small. In the case of communication systems, this can be achieved for certain nonzero information rates by choosing the constraint length of the code arbitrarily large. The largest information rate for which the probability of error can be made arbitrarily small is called the *capacity of the communication channel.* By making an analogy with this result, one might expect that, in the case of a computing system, it should be possible to achieve an arbitrarily small probability of error for certain bounded values of the redundancy by choosing the complexity sufficiently large. Extending this analogy, the reciprocal of the minimum redundancy for which the probability of error can be made arbitrarily small is called the *computing capacity* of the computing system. For the systems proposed by von Neumann, there is no finite redundancy for which the probability of error can be made arbitrarily small; therefore, these systems have a computing capacity of zero.

The question remained whether there existed any method for designing a "reliable" system with a nonzero computing capacity. Since it was well known that by using suitable coders and decoders it is possible to obtain a "reliable" communication system, it was natural to attempt to apply coding techniques to the problem of designing a "reliable" computing system. One approach is to consider coding the inputs to each computing component and decoding the output. Elias set out to show that this method could not be used to design a general computing system with a nonzero computing

capacity.[4] Since Elias desired a negative result, he permitted all coders and decoders to be noiseless, and since a general computing system must be capable of performing all 16 elementary operations, Elias considered 16 types of computing components, each capable of performing one of the elementary operations.

The computing components were divided into two classes. The operations performed by components in the first class were those represented by matrices containing an odd number of ones and zeros and in the second class were those containing an even number. Elias showed that components in the first class have a computing capacity equal to zero; but that it is possible for components in the second class to have a nonzero computing capacity.*

Unfortunately, the only component in the second class which performs a nontrivial operation is the modulo-2 adder; furthermore, there is no combination of class two components that can perform class one operations such as AND and OR. Therefore Elias concluded that this coding technique could not be used to design a general computing system with a nonzero computing capacity.

More recently Winograd and Cowan proposed another scheme for designing a "reliable" computing system.[8] Their approach was very similar to von Neumann's. However, instead of considering a single irredundant network as the precursor for the redundant network, Winograd and Cowan considered a composite network consisting of k copies of this irredundant network, each computing independently, to be their precursor. If the original irredundant network has a complexity σ then this composite network has complexity $k\sigma$; but its redundancy is still one since each network is capable of performing independent operations on different inputs.

To introduce redundancy into this composite network, one considers sets of k network elements, the members of each set being the corresponding elements in each of the k precursors. The redundant network is formed by replacing each set of k elements with n modules. These modules have the property that each of their inputs is encoded according to some (n, k) block code, thus allowing each one to perform an error correction operation on each of its inputs. Every set of n modules performs the appropriate operations on the corrected inputs so that the n binary outputs from the n modules form the

* To achieve this nonzero computing capacity, it is necessary to assume that the complexity of the encoders and decoders grows only linearly with the block length of the code as in the case of convolutional coders and sequential decoders.[5-7]

code word corresponding to the desired result, namely the code word whose information digits are equal to the corresponding k outputs in the original composite network.

Winograd and Cowan assumed that modular malfunctions are statistically independent from one module to another; furthermore, they also assumed that for all modules except the "output modules" the probability of a modular malfunction is p_0, independent of the operations performed by that module. The "output modules," those modules whose output constitutes the output of the computing system, were assumed to be noiseless. To compute a bound on the probability of error for this system, each set of n noisy modules can be modeled by a set of n noiseless modules, where the output from each module is passed through a binary symmetric channel (BSC) with crossover probability p_0. A failure occurs whenever the output from any set of n BSC's is such that a noiseless decoder would be unable to decode this word correctly. According to the noisy channel coding theorem,[9] there exist codes for which the probability of such a failure is bounded by

$$\Pr\text{[modular failure]} \leqq e^{-nE(R)}$$

where $E(R) > 0$ for codes with information rates less than the capacity of a BSC with crossover probability p_0. Since there are at most $n\sigma$ modules within the network and since each module is used only once during the computation, the probability of a malfunction anywhere within the network during the computation is bounded by

$$\Pr\text{[failure in network]} \leqq n\sigma \cdot e^{-nE(R)}$$

which can be made arbitrarily small by making n sufficiently large.

If the complexity of the modules were fixed, this result would imply that the probability of error can be made arbitrarily small by making the complexity sufficiently large, while keeping the redundancy bounded. Unfortunately, each module must perform encoding and decoding which requires a number of operations that grows at least linearly with n. Therefore, the complexity of the modules must grow at least linearly with n which implies that the redundancy of the overall network must grow at least linearly with n rather than being bounded as one would have hoped. Therefore, the probability of error for the system can be made to approach zero only in the limit of infinite redundancy. Hence, Winograd and Cowan's system also has a computing capacity equal to zero.

1.3 *Error Criteria*

Although the term "probability of error" is used in connection with each of the three systems discussed in the previous section, the error criterion was different for the different systems. Elias, Winograd, and Cowan assumed that each input to the computing system is uncoded and that the output from the system is also uncoded. They assumed that for each set of inputs, the system has one correct output defined as the output which would be obtained if the system were noiseless. If the actual output differs from the correct output, the system has made an error. To obtain a probability of error for the system that is not lower bounded by the probability of error for the output components, they required that the output components be noiseless.

The necessity for using noiseless output devices arises because of the requirement that the output be uncoded. Von Neumann avoided this problem by assuming that all inputs and outputs are repeated n times. He assumed that the result is "correct" provided that the fraction of the outputs which are in error is small. Von Neumann's assumption that all inputs and outputs must be repeated is a special case of the assumption that all inputs and outputs must be coded according to some error correcting code. The latter, more general assumption is made in this paper. The only condition that is imposed is that both the inputs and the output must be coded according to the same code so that computing systems are compatible with each other. Since the outputs are coded, there must exist classes of outputs corresponding to the decoding equivalence classes of the code. A result is considered to be correct provided that it is within the class which contains the code word corresponding to the desired result.

The concept of coded inputs and outputs might, at first, seem unrealistic since it implies that the user is capable of performing error correcting coding and decoding. However, if we consider the case of two people communicating with each other, we observe that a very complicated process of coding and decoding takes place. Appropriate redundancy is introduced not only through the inherent redundancy in the language but also through the use of "diversity channels" which, in this case, correspond to facial expressions, hand and arm movements, voice inflections, and so on. Therefore, since there is always an appropriate coding used in the transmission of information between individuals, it is unrealistic to expect that a machine and user could communicate without the use of some type of

error correcting procedure. In fact, the condition that all information must be coded is a necessary requirement for the reliability of a computing system in which every component is noisy.

1.4 *Synopsis*

It is our ultimate objective to show that it is possible to construct from unreliable components a reliable computing system where a *reliable system* is defined as a type of system which has a nonzero computing capacity. For the first part of the analysis, it is assumed that component errors are statistically independent from one component to another and from one use of a particular component to another use. A computing system constructed from components of this type is called a *noisy computing system.* A virtually identical analysis and similar results apply in the case where components within the system fail permanently but where periodic maintenance is performed on the system; that is, at regularly spaced times, components which have failed are replaced by good ones. In this paper we restrict our attention to memories. It is shown that information can be stored reliably within a "stable memory," a device constructed entirely from unreliable components. The paper following this shows that it is possible to design a computing system, using unreliable components, that performs operations reliably on information stored within stable memories.

II. STABILITY

The remainder of this paper is concerned with reliable information storage in memories constructed from unreliable components. A memory is a device in which information is stored at one time and recovered at some later time. If a memory is to be useful, it must have two important characteristics. First, it should be possible both to store information in the memory and to read information from the memory at any time specified by the user, or at least at any one of a set of discrete times where the members of the set are closely spaced. All memories considered in this paper can have information stored in them at any time and retrieved from them at any time. Second, the information read out of the memory must be identical to or at least equivalent to the information originally stored. With memories constructed entirely from unreliable components, it is unreasonable to expect that the word read out of the memory will be identical to the word stored; however, we can hope that the informa-

tion will be preserved. To clarify this idea of information preservation we introduce the concept of stability.

2.1 *The Concept*

To illustrate the meaning of stability, let us consider a simple memory consisting of one noisy register and a correcting network. The *state* of this memory is defined by the word contained within the register. It is assumed that initially ($t = 0$) a code word from some error correcting code is read into the register, thus defining the memory's initial state. Since the register is noisy, errors occur in the stored digits; hence, the state of the memory is perturbed away from the original one. The correcting network monitors the contents of the register, performs error correction, and inserts into the register an estimate of the original code word. If there were no correcting network, the state of the memory would "wander away" from the original one; however, the correcting network provides a "restoring force" which tends to bring the state of the device back to the original. The noise may perturb the state of the device beyond the error correcting capability of the correcting network. If this happens, we say that a *memory failure* has occurred. To define a memory failure more precisely, it is necessary to associate with the different input code words disjoint classes of states. As long as the state of the memory remains within the appropriate class, no memory failure has occurred.

The redundancy of a memory is defined to be the ratio of the complexity of the memory to the complexity of an irredundant memory which has the same information storage capability. The inputs to the memory being considered are code words from an (n, k) block code; hence, the memory has a storage capability of k bits. This memory is denoted by M_k. An irredundant memory with the same information storage capability as M_k would have a complexity equal to k, since it would consist of k one-bit information storage components. If we consider k to be a variable, we can speak of a sequence of memories where M_k is a typical member of the sequence. If the complexity of every memory in this sequence is less than $\alpha \cdot k$, the redundancy of any memory in the sequence must be less than α which, by assumption, is independent of k. Hopefully, for any T, it is possible to make the probability of a memory failure during the time interval $0 \leqq t \leqq T$ arbitrarily small by choosing k sufficiently large while keeping α bounded. If the sequence of memories has this property, we say that the sequence is *stable*. For convenience we refer to a typical member of a stable sequence of memories as a *stable memory*. Here is a more concise definition of stability.

2.2 *The Definition*

Consider a sequence of memories denoted by $\{M_i\}$. The memories in this sequence are ordered according to their information storage capability; that is, the kth memory, M_k, has a storage capability of k bits. The sequence $\{M_i\}$ is called *stable* if it satisfies the following conditions:

(*i*) For any k, M_k must have 2^k allowed inputs which are denoted by $\{I_{ki}\}$, $0 < i \leqq 2^k$.

(*ii*) With each input there must be associated a class of states of M_k. The classes associated with different inputs must be disjoint. For any k and i, the class of states corresponding to I_{ki} is denoted by $C(I_{ki})$.

(*iii*) The complexity of M_k must be bounded by $\alpha \cdot k$ where α is a fixed parameter for any particular sequence.

(*iv*) Suppose that at $t = 0$ any one of the allowed inputs is transmitted to each memory in the sequence. Let there be no inputs for all $t > 0$. Consider a typical memory M_k. Denote the particular input that was transmitted to M_k by I_{ki}. The probability that the state of M_k does not belong to $C(I_{ki})$ at $t = T$ is denoted by $p_{ki}(T)$ and $\max_i [p_{ki}(T)]$ is denoted by $p_k(T)$. If the sequence is stable then, for any $T > 0$ and $\delta > 0$, there must be a k such that $p_k(T) < \delta$.

2.3 *Examples of Memories*

To further clarify the meaning of stability, consider two types of memories. The first consists of a noisy register without any correcting network and the second consists of a noisy register with a noiseless correcting network. In light of the discussion in Section 2.1, we do not expect that memories of the first type can be stable whereas we do expect that memories of the second type can be. These expectations are correct.

Consider first a sequence of noisy registers. A typical member of this sequence is a register containing n binary digits which define the register's state. Let p denote the probability that any particular digit stored in the register is changed during a time interval τ. If one is interested in the contents of the register only at the times $t = 0, \tau, 2\tau, 3\tau, \cdots$, a model for the noisy register is the noiseless register together with the n binary symmetric channels shown in Fig. 2. The allowed inputs to this noisy register, $\mathfrak{M}_k$, are the 2^k code words from some (n, k) code. The class of states corresponding to a particular input is the decoding equivalence class corresponding to that code word. The complexity

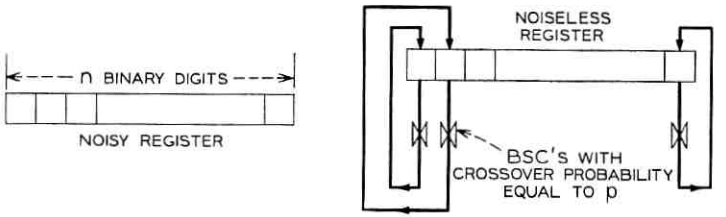


Fig. 2 — Model for a noisy register. The digits in the noiseless register are transmitted through the BSC's once every τ seconds.

of $\mathfrak{M}_k$ is $n = k/R$, where R is the information rate of the code; therefore, the appropriate value of α for this sequence is $1/R$.

Suppose that at $t = 0$ one of the allowed inputs is transmitted to $\mathfrak{M}_k$. At $t = \tau$ the probability of error per digit is p. The probability of a memory failure at $t = \tau$ equals the probability that the word in the register at $t = \tau$ does not belong to the decoding equivalence class containing the original input. Provided that the code is at least as "good" as an average random code, the noisy channel coding theorem[9] states that $p_k(\tau)$ is bounded by

$$p_k(\tau) \leqq \exp - [(k/R)\ E(R)]$$

where $E(R)$ is positive for $R < 1 - H(p)$. $[1 - H(p)]$ is the capacity of a BSC with crossover probability p. Since the probability of failure is independent of the particular input it is unnecessary to perform the maximization over all inputs.

Next consider the state of the register at time $t = T = L\tau$. According to the model in Fig. 2, to determine the state of the noisy register at $t = L\tau$ we must transmit the n binary digits through their respective BSC's L times. This is equivalent to transmitting each binary digit through L BSC's in series. Since the overall capacity of these channels in series decreases asymptotically to zero as L increases, for any fixed rate there must exist some L for which the capacity of the L channels in series is less than R. Therefore, for any sequence of noisy registers with bounded redundancy (fixed α or R), there must exist some L (or T) such that there is no register in the sequence which has a sufficiently small probability of failure to satisfy the requirements for stability. Thus, as would be expected, noisy registers are not stable.

As a second example, consider the same sequence of noisy registers but this time associate with each register a noiseless correcting net-

work. Within each time interval of length τ, this correcting network takes the output from the n BSC's and maps this vector onto a code word which is then inserted into the register to define its new state. This type of device is shown in Fig. 3. The operation performed by the correcting network is equivalent to that performed by a noiseless decoder followed by a noiseless encoder. This operation can be performed by a correcting network whose complexity is proportional to k; for example, a noiseless sequential decoder followed by a noiseless convolutional encoder.[5-7] If such a correcting network is used, the redundancy is independent of k and therefore can be bounded for all k. The probability of a memory failure at $t = \tau$ is again upper bounded by $\exp - [(k/R)\ E(R)]$ but the probability of a memory failure at $t = L\tau$ is now bounded by

$$p_k(L\tau) < L \cdot \exp - [(k/R)\ E(R)]$$

according to the union bound. Therefore, for any finite L, $p_k(L\tau)$ approaches zero as k approaches infinity provided that $R < 1 - H(p)$. This shows that a noisy register with a noiseless correcting network can be stable.

III. THE STABILITY OF MEMORIES CONSTRUCTED ENTIRELY FROM UNRELIABLE COMPONENTS

We now show that a noisy register with a noisy correcting network can be stable. This proof of stability is extended to memories in which components fail permanently but where the components which have failed are periodically replaced.

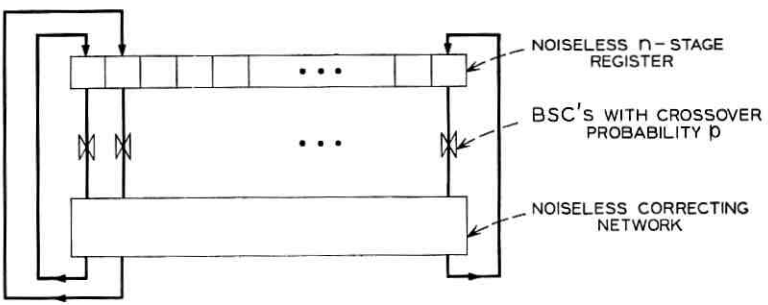


Fig. 3 — Noisy register with noiseless correcting network.

3.1 *The Importance of Low-Density Parity-Check Codes*

The memories to be considered store information in the form of code words from an error correcting code. Each memory consists of one or more noisy registers and a noisy correcting network that performs operations which are very similar to those performed by a decoder. It is our objective to show that noisy memories of this type can be stable. Since the complexity of a stable memory is required to be bounded by $\alpha \cdot k$, where k is the information storage capability and α is a proportionality factor which does not depend on k, the complexity of the correcting network in any stable memory can grow only linearly with k. There are only two kinds of correcting networks (decoders) known to have this property. One is a correcting network based on a sequential decoder[6, 7] and the other is a correcting network based on a low-density parity-check decoder.[10]

In deciding whether a particular correcting (decoding) procedure is suitable for use in a noisy correcting network, one must consider whether there are any essential steps in the procedure which could not be performed with a small probability of error by a noisy device. For example, almost all parity-check decoders are required to compute the modulo-2 sum of a set of digits where the number of digits in the set is proportional to the constraint length, N, of the code.

To compute the probability that a noisy device makes an error in performing this operation, consider the noisy addition network shown in Fig. 4. This network, consisting of $N - 1$ adders (modulo-2), computes the modulo-2 sum of N binary inputs. Let us assume that each adder in the network has a probability of error p_a and that adder errors are statistically independent from one adder to another and from one use of a particular adder to another. The output of the noisy addition network will be in error if an odd number of adders make errors. It can be shown[10] that the probability of this event equals

$$\text{Prob of error} = \frac{1 - (1 - 2p_a)^{N-1}}{2}.$$

This probability approaches $\frac{1}{2}$, exponentially with N, as N approaches infinity.

The memories under consideration store coded information. To make the probability of a memory failure arbitrarily small (as required for stability), one would expect that it would be necessary to make the constraint length of the code arbitrarily large. Therefore, the noisy correcting network must be able to perform error correction

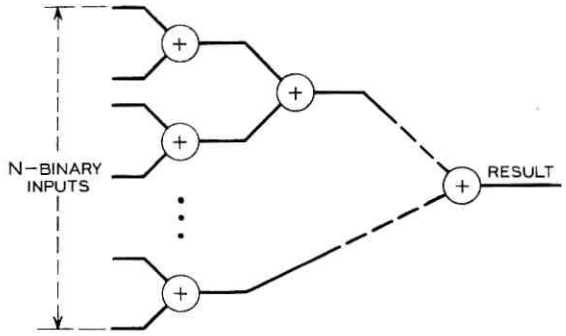


Fig. 4 — Modulo-2 addition network. ⊕ represents one adder (modulo-2).

even if the constraint length of the code is very long. This implies that no correcting procedure involving a modulo-2 addition operation of the type just described is suitable for use in a noisy correcting network. For example, the correcting network based on a sequential decoder must generate hypothesized branches of a code tree. Each digit on one of these branches is computed by forming the modulo-2 sum of a set of digits where the number of digits in the set is proportional to the constraint length of the code. The probability that a noiseless sequential decoder makes an error can be made arbitrarily small only if this constraint length is made arbitrarily large. But making the constraint length arbitrarily large makes the probability of an addition error within the noisy decoder arbitrarily close to ½. Therefore, a noisy correcting network should not be based on a sequential decoder.

Fortunately the correcting network based on a low-density parity-check decoder does not have this problem. A low-density parity-check decoder does evaluate parity checks, modulo-2 sums of the digits in parity-check sets; however, the number of digits in each parity-check set is not a function of the block length of the code.

3.2 *The Correcting Algorithm*

The memories to be considered consist of several registers and a correcting network and they store information in the form of code words from a low-density parity-check code. It is our objective to show that memories of this type can be stable. The first step is to consider the details of the correcting algorithm. This requires a brief explanation of low-density parity-check codes.

An (N, J, K) low-density parity-check code is defined to be a code with block length N such that there are K digits in each parity-check set and J parity-check sets containing any particular digit. Such a code can be represented by a parity-check matrix which has K ones in each row and J ones in each column. For example, an $(N = 20, J = 4, K = 5)$* low-density parity-check matrix is shown in Fig. 5.

$$\begin{bmatrix}
1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\
1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\
0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0
\end{bmatrix}$$

$$d_{21}\ d_{31}\ d_{41}\ d_1\ d_{42}\ d_{22}\ d_{32}\ d_{43}\ d_2\ d_{33}\ d_0\ d_{11}\ d_{12}\ d_{13}\ d_{14}\ d_{23}\ d_{34}\ d_{44}\ d_{24}\ d_3$$

Fig. 5 — The parity-check matrix for a (20, 4, 5) low-density parity-check code. The digit positions, denoted by d with appropriate subscripts, are numbered in the way described in Fig. 6.

Gallager has described two schemes for decoding low-density parity-check codes, both of which are iterative.[10] However, we shall only be concerned with the simpler one. Each iteration of this scheme consists of first computing all the parity checks and then changing the value of any digit that is contained in more than a certain fixed number of unsatisfied parity-check constraints (if a parity check equals one, the corresponding parity-check constraint is unsatisfied). Provided that there were not too many errors initially, each successive iteration decreases the number of digits in error and, eventually, all parity check constraints are satisfied indicating that the resulting word is a code word.

To illustrate how this method works, let us suppose that the digit d_0 is in error but that all other digits are correct. In this case, all parity-check constraints involving d_0 will be violated, whereas at most one parity-check constraint involving any other digit will be violated. Therefore, d_0 will be changed whereas all other digits will be unchanged. In this way the digit d_0 will be corrected. If there are

* Notice that K is always greater than J.

errors among the digits used to check d_0, this digit may not be corrected during the first iteration. However, after one or more iterations sufficiently many of these errors may have been corrected to allow d_0 to be corrected also.

This simple decoding algorithm does have one problem. Recall that for any digit d_0, the original values of all the other digits in the J parity-check sets containing d_0 are involved in the determination of the new value of d_0 and similarly that the original value of d_0 is used in computing the new value of each digit in these J parity-check sets. This means that on successive iterations the values of the digits involved in making a new estimate of d_0 depend on the previous estimate of d_0. This leads to a complex interrelation between the errors which degrades the decoder's performance and, needless to say, greatly complicates its analysis. Fortunately Gallager has suggested a way to modify the algorithm which at least partially solves this problem. Using this modified algorithm, J estimates of each digit are made, each one being computed using a different combination of $J - 1$ out of the J parity-check sets containing the digit to be estimated. The value of a particular estimate is changed if $J/2$ or more out of the $J - 1$ parity checks are equal to one.

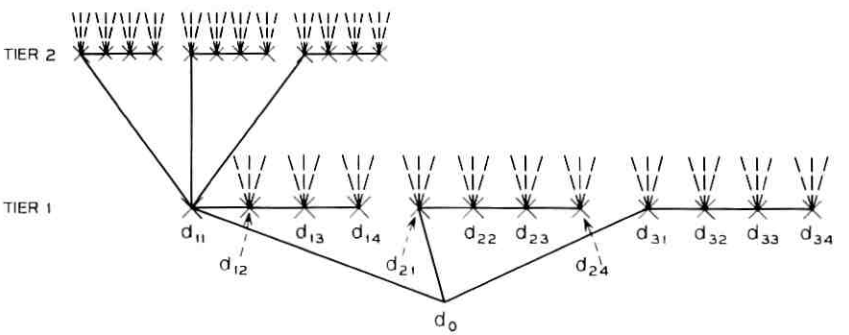
To construct a correcting network based on this algorithm, one starts with J registers of length N. Although the assignment of digit estimates to register locations can be arbitrary, one would probably choose to assign one estimate of each of the N digits in a code word to the corresponding N locations in each register. Since each estimate of a digit, say d_0, is to be made on the basis of $J - 1$ parity-check sets, each containing $K - 1$ digits other than d_0, there must be $(J - 1) \cdot (K - 1)$ other digits interconnected with the input to d_0. Using the modified algorithm, it is necessary to specify not only the digits to be interconnected but also the appropriate estimate of each digit. For example, consider the parity check set $(d_0, d_{11}, d_{12}, d_{13}, d_{14})$. The appropriate estimate of the digit d_{11} to be used in correcting d_0 is that estimate based on the $J - 1$ parity check sets which omit the one containing d_0. This estimate is used so that the values of the digits involved in computing the new estimate of d_0 do not, themselves, depend on a previous estimate of d_0.

This correcting network performs many iterations. During the first iteration each digit is estimated on the basis of the $(J - 1) \cdot (K - 1)$ digits to which it is interconnected. Since, during the second iteration, the same operations are performed, the resulting second estimate of each digit depends on the first estimates of these $(J - 1)$

$\cdot(K - 1)$ interconnected digits which, in turn, depend on the initial estimates of a much larger set of digits. The sets of digits involved in making successive estimates of some digit, say d_0, can be represented by means of parity-check set trees of the type shown in Fig. 6. The branches rising from d_0 represent one set of $J - 1$ parity-checks containing this digit. The interconnected nodes on the first tier of this tree represent the digits, other than d_0, in one of these parity-check sets. Each digit on the first tier of the tree is also contained in $J - 1$ other parity-check sets. These other parity-check sets are represented by the branches rising from the first tier to the second tier of the tree. This parity-check set tree can be extended indefinitely. The structure of the tree, beyond the first tier, is completely specified by the parity-check sets of the codes which, in turn, are specified by the parity-check matrix.

To see how the tree represents the digits involved in estimating d_0, let us suppose that the decoder has performed i iterations. During each iteration the digits on each tier of the tree are used to estimate the digits on the tier immediately below. Hence, after i iterations, the value of each digit, particularly d_0, has been influenced by the values of the digits on the first i tiers above it.

If a parity-check set tree is extended for many tiers, eventually some digit will appear in two different places in the tree. If the first repetition within any of the $J \cdot N$ trees occurs on the $(m + 1)$th tier,



$$d_0 + d_{11} + d_{12} + d_{13} + d_{14} = c_1$$
$$d_0 + d_{21} + d_{22} + d_{23} + d_{24} = c_2$$
$$d_0 + d_{31} + d_{32} + d_{33} + d_{34} = c_3$$
$$d_0 + d_{41} + d_{42} + d_{43} + d_{44} = c_4$$

Fig. 6 — Two representations for the parity-check constraints involving d_0. At the top is one of the J parity-check set trees rising from d_0. Beneath it are the parity-check equations containing d_0.

then, according to Gallager's nomenclature, the code has m *independent iterations*. Notice that this number, m, is a parameter of the code which is unrelated either to the statistics of the noise or to the particular decoding algorithm. This number plays a particularly important part in the analysis of the correcting procedure as it is shown that errors in specific sets of digits are statistically independent during these first m iterations.

The physical configuration of the memory is shown in Fig. 7. Within the correcting network there are $J \cdot N$ identical sets of components. Each set of components performs the operations required to estimate one particular digit. Let us consider a set of components which computes estimates of the digit d_0. The first operation that must be performed by this set of components, the computation of the $J-1$ parity checks rising from d_0, requires $(J-1) \cdot (K-1)$ two-input binary modulo-2 adders. The second operation, deciding whether the digit d_0 should be changed, can be performed by a decision device (threshold device) the output of which is a 1 if d_0 is to be changed and a 0 if d_0 is not to be changed. Finally, the output of this decision device must be added modulo-2 to the previous estimate of d_0, the operation requiring one binary adder (modulo-2). Similar operations are performed to obtain estimates of all $J \cdot N$ digits. These operations

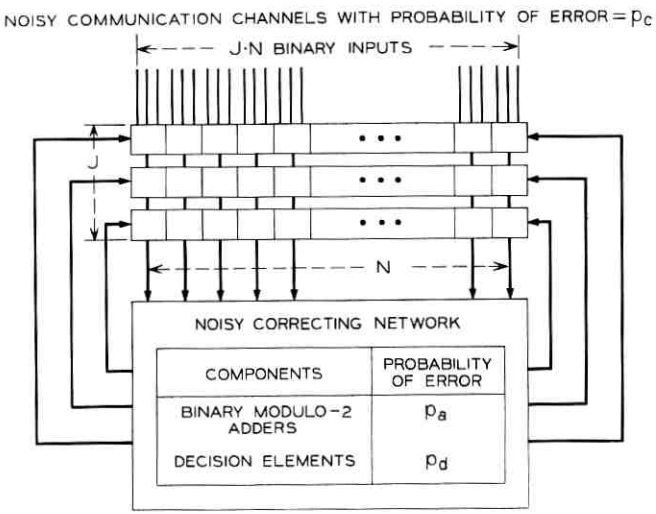


Fig. 7 — Physical configuration of a memory based on a low-density parity-check decoder.

constitute one correcting cycle (iteration). The correcting network performs correcting cycles continuously once information has been read into the registers.

Finally, let us consider the situations which cause the estimate of the digit d_0 to be in error after a particular iteration. Since any digit d_0 which is in error will be corrected only if a sufficient number of the parity checks containing d_0 are equal to 1, ideally, we would like every parity check containing d_0 to equal 1 if d_0 is in error but to equal 0 if d_0 is correct. If a parity check does not equal the desired value we say that the parity check is in error. Thus a parity check error depends on errors made by the adders and errors in the digits involved in estimating d_0, but not on the value of d_0 itself.

To simplify the mathematical analysis, we restrict our attention to the class of low-density parity-check codes with $J = 2l$, $l = 2, 3, 4, \ldots$ and the correcting algorithm stated previously. A set of events each one of which alone leads to an error (indicated by ϵ) in the estimate of the digit d_0 after a particular iteration are:

(*i*) $d_0 = \epsilon$ after previous iteration and $J/2$ or more parity checks are in error.

(*ii*) $d_0 \neq \epsilon$ after previous iteration and $J/2$ or more parity checks are in error.

(*iii*) The decision (threshold) device makes an error.

The first two conditions demonstrate an interesting property of this class of low-density parity-check codes. For this class of codes the conditions leading to an error in any digit after any iteration are the same whether or not the digit was in error before the iteration. This property will help to simplify the following mathematical analysis.

3.3 *The Stability Theorem for Noisy Memories*

Theorem 1: There is a stable sequence of noisy memories where every component in every memory has a fixed, nonzero probability of error per use.

Proof: The memories under consideration consist of J registers of length N, a noisy correcting network based on a low-density parity-check decoder, and a set of communication channels over which the inputs are transmitted. The definition of stability given in Section 2.2 includes specific conditions that must be satisfied by stable memories. To prove that the memories under consideration are stable, we must show that they satisfy all these conditions.

The definition of stability requires that there be a set of allowed

inputs for each memory. Each allowed input for a memory of the type under consideration consists of J copies of a code word from some (N, J, K) low-density parity-check code. If R equals the information rate of the code, the memory having registers of length N has $2^{RN} \triangleq 2^k$ allowed inputs. This memory is denoted by M_k. To define a sequence of memories, we allow k(and N) to be a variable but keep the parameters of the code, J and K, fixed.* Let the 2^k allowed inputs to M_k be denoted by $\{I_{ki}\}$ $0 < i \leqq 2^k$. With each input we must associate a class of states of M_k. The classes to be used are essentially the decoding equivalence classes of the code. To be more precise, the classes can be described in terms of the equilibrium states of M_k, denoted by $\{E_{ki}\}$ $0 < i \leqq 2^k$, where an equilibrium state is one in which every register contains one and the same code word. With each input, I_{ki}, there is a corresponding equilibrium state, E_{ki}, in which the registers contain the code words represented by I_{ki}. The state of M_k belongs to the class $C(I_{ki})$ if a noiseless correcting network (that is, a noiseless low-density parity-check decoder) could correct all the errors; in other words, if its final state would be the equilibrium state, E_{ki}.

The definition of stability requires that for each sequence of memories there must be an α such that for every k the complexity of M_k is bounded by $\alpha \cdot k$. The complexity of the correcting network in M_k is computed in Section 3.2. If each of the J registers in M_k is of length N, the correcting network must contain $[1 + (J-1)(K-1)]\, J \cdot N$ binary adders (modulo-2) and $J \cdot N$ decision devices. Since a decision device must determine whether $J/2$ or more out of the $J-1$ parity checks are in error, the complexity of each decision device depends only on the code parameter, J, which is fixed for any particular sequence of memories. Thus the complexity of each decision device, denoted by D, is independent of k. Finally, the noisy registers within the memory must contain $J \cdot N$ storage components. Therefore the complexity of M_k is:

$$\begin{aligned} \text{Complexity of } M_k &= [2 + D + (J - 1)(K - 1)]J \cdot N \\ &= \frac{[2 + D + (J - 1)(K - 1)]J}{R} \cdot k. \end{aligned}$$

Since Gallager[10] has shown that $R \geqq 1 - J/K$, the complexity proportionality factor for the sequence of memories under consideration is upper bounded by

* There are low-density parity-check codes for all values of N which are integer multiples of K.[10]

$$\alpha \leqq \frac{[2 + D + (J - 1)(K - 1)]J}{1 - J/K}.$$

Every component has a nonzero probability of making an error. The error probabilities are denoted:

p_r = probability that one particular binary digit within the register changes during a time interval τ, the time required for one correcting cycle.
p_a = probability that an adder makes an error on any one use.
p_d = probability that a decision device makes an error on any one use.
p_c = probability of an error in transmitting a digit across a communication channel.

It is assumed that errors are statistically independent from one component to another and from one use of a particular component to another use.

To prove that the information storage devices in question are stable, it must be shown that for any $\mathcal{L}$, the probability of a memory failure in M_k during the time interval $0 \leqq t \leqq \mathcal{L}\tau$ can be made arbitrarily small by choosing k sufficiently large. In the remainder of this section, the probability of a memory failure for the typical device, M_k, is upper bounded. To determine whether a memory failure has occurred at some particular time we use the following algorithm.

Imagine that M_k becomes noiseless at that time and that the noiseless correcting network within M_k performs m more iterations where m is the "number of independent iterations" described in Section 3.2. These are precisely the same operations that a noiseless low-density parity-check decoder would perform. If the final state of the hypothetical noiseless memory is error free (that is, equals the equilibrium state corresponding to the original input) then, by definition, no memory failure occurred. Thus, to bound the probability of a memory failure we must bound the probability that the final error pattern is not error free.

In general, the error pattern within the registers of M_k depends on all the component errors that have occurred since the original input was transmitted to the memory. However, if certain conditions are satisfied, the error pattern is a function of only the component errors that occurred during the most recent m iterations. If no component errors occur during these m iterations and if the required conditions are satisfied, the final error pattern will be error free. On the other hand, if these conditions are not satisfied we say that a *propagation*

failure has occurred. Therefore, to bound (upper bound) the probability of a memory failure, we bound the probability that a propagation failure occurs on the mth noiseless iteration of the hypothetical noiseless memory.

The first step is to show that, in the absence of a propagation failure, the probability of error for digits stored in the memory (the probability of error per digit) has an upper bound, denoted by p_0, such that $p_0 \ll \frac{1}{2}$. This can be seen intuitively by comparing the performance of a noisy corrector with that of a noiseless corrector, that is, a noiseless low-density parity-check decoder. Provided that the initial probability of error is not too large, a noiseless corrector decreases the probability of error per digit with each iteration until this probability reaches zero. If the probability of error for the components within the noisy corrector is small, compared with the initial probability of error per digit, for the first few iterations the noisy corrector decreases the probability of error per digit just as the noiseless one did. However, eventually this probability reaches the same order of magnitude as the probability that the noisy corrector itself makes an error at which time the probability of error per digit reaches an equilibrium value. Notice that although such an equilibrium value is attained, it is still possible for errors to occur in sufficient digits so that a memory failure results. If a memory failure does occur, a propagation failure will also occur and the bound on the probability of error per digit will no longer be valid.

In light of this intuitive argument, we expect that the time at which the probability of error per digit is at its maximum value is just before the end of the first correcting cycle. In Appendix A upper bounds are computed on this probability evaluated just before the end of the first and successive correcting cycles. It is shown that, provided no propagation failure occurs, these bounds form a monotonically decreasing sequence; hence, the bound on this probability evaluated just before the end of the first correcting cycle, denoted by p_0, is the desired bound.

The next step is to bound the probability that the initial propagation failure occurs at some particular time. A propagation failure occurs whenever the error pattern in the memory is related to the component errors that occurred m or more iterations previously. In most cases the error pattern depends only on component errors that occurred in the last few iterations since any digit errors caused by previous component errors would have already been corrected. In order for a propagation failure to occur, the effect of component errors must have propagated from one iteration to the next for at least

m iterations. Thus we expect that the probability of such a propagation decreases as m is increased and, since increasing k increases m, that the probability of a propagation failure can be made arbitrarily small by making k sufficiently large. The explicit relationship between the probability of the initial propagation failure and k is derived in Appendix B. The result is

$$\Pr[\text{initial propagation failure}] < Ck^{-\beta+1}$$

where C and β depend only on the parameters of the code (J and K) and the bound on the probability of error per digit, p_0; hence, C and β are constants for any particular sequence of memories.

Some typical values for β are shown in Table I. For example, if $J = 14, K = 15$, and $p_0 = 10^{-8}$ then $\beta = 7.55$; therefore, in this case, the probability that the first propagation failure occurs at some particular time is upper bounded by a function that decreases as the sixth power of the information storage capability of the memory. By choosing the information storage capability sufficiently large, this probability can be made arbitrarily small.

For the moment, let us assume that neither a memory failure nor a propagation failure has occurred within M_k during the time interval $0 \leqq t < \mathcal{L}'\tau$. We now use the bound on the probability of the initial propagation failure to find an upper bound on the probability that either the initial memory failure or the initial propagation failure occurs at $t = \mathcal{L}'\tau$. To determine whether the initial memory failure occurs at $t = \mathcal{L}'\tau$ we must imagine that M_k becomes noiseless at $t = \mathcal{L}'\tau$ and that the noiseless correcting network within M_k performs m more iterations. As explained previously, the initial memory failure can occur at $t = \mathcal{L}'\tau$ only if the initial propagation failure occurs at $t = \mathcal{L}'\tau$ or during the m noiseless iterations performed after $t = \mathcal{L}'\tau$. Thus the sum of the

TABLE I—TYPICAL VALUES OF $\beta'(J, K, p_0)$ AND $\beta(J, K, p_0)$.

J	K	R (lower bound)	p_0	β	β'
4	5	0.20	10^{-8}	2.66	0.66
6	7	0.14	10^{-8}	3.91	1.91
8	9	0.11	10^{-8}	4.95	2.95
10	11	0.09	10^{-8}	5.89	3.89
12	13	0.07	10^{-8}	6.75	4.75
14	15	0.06	10^{-8}	7.55	5.55

probabilities that the initial propagation failure occurs at $t = \mathfrak{L}'\tau$, $(\mathfrak{L}' + 1)\tau, \cdots, (\mathfrak{L}' + m)\tau$ is an upper bound on the desired probability. Therefore,

$$\Pr[\text{initial memory failure or initial propagation failure at } t = \mathfrak{L}'\tau] < (m + 1)\cdot C\cdot k^{-\beta+1}.$$

Gallager[10] has shown that

$$m < \frac{\log N}{\log[(J-1)(K-1)]}.$$

Therefore,

$$\begin{aligned} m + 1 &< \frac{\log N}{\log[(J-1)(K-1)]} + 1 \\ &< \frac{\log\left[\frac{k}{1 - J/K}\right]}{\log[(J-1)(K-1)]} + 1 \\ &< \log\left[\frac{k}{1 - J/K}\right] \quad \text{if} \quad k > 2. \end{aligned}$$

(Note : $K > J \geqq 4$).

The probability that the initial memory failure occurs during the time interval $0 \leqq t \leqq \mathfrak{L}\tau$ is upper bounded by the sum of the probabilities that either the initial memory failure or the initial propagation failure occurs at $t = 0, \tau, 2\tau, \cdots, \mathfrak{L}\tau$. This bound equals

$$\begin{aligned} &\Pr[\text{failure during time interval } 0 \leqq t \leqq \mathfrak{L}\tau] \\ &\quad < (\mathfrak{L} + 1)\cdot C\cdot \log\left[\frac{k}{1 - J/K}\right]\cdot k^{-\beta+1} \\ &\quad < (\mathfrak{L} + 1)\cdot \frac{C}{1 - J/K}\cdot k^{-\beta+2} \\ &\quad = (\mathfrak{L} + 1)\cdot C'\cdot k^{-\beta'} \end{aligned}$$

where

$$C' \triangleq \frac{C}{1 - J/K} \quad \text{and} \quad \beta' \triangleq \beta - 2.$$

To show that there is a stable sequence of noisy memories, we must show that it is possible to choose J, K, and p_0 such that $\beta' > 0$. Recall that when we speak of a particular sequence of memories, $\{M_k\}$,

we mean that k is a variable whereas the values of J, K, and the probabilities of component errors are all fixed. Certain conditions have already been imposed on J, K and p_0. These conditions are:

(i) $\quad J = 2l, \qquad l = 2, 3, 4 \cdots$

(ii) $\quad K > J$

(iii) $\quad p_0 > 2p_r + p_c$

(iv) $\quad p_0 > \binom{J-1}{J/2}[(K-1)(p_0 + p_a)]^{J/2} + p_d + p_r$

(v) $\quad p_0 > p_a$

where p_a, p_d, p_r and p_c must all be fixed and greater than zero. To demonstrate that there are values of J, K and p_0 which satisfy these conditions and for which $\beta' > 0$, consider an example where $p_a = p_d = p_r = p_c = 10^{-9}$ and where $p_0 = 10^{-8}$. For this example conditions (iii), (iv), and (v) are satisfied for all reasonable values of J and K (that is, where $J < K \ll p_0^{-1}$). The values of $\beta'(J, K, p_0 = 10^{-8})$, which correspond to some typical values of J and K, are shown in Table I. For all the values of J and K which are considered, the value of β' is greater than zero. Therefore, in all these cases, the probability of a memory failure in M_k at $t = \mathcal{L}\tau$ can be made arbitrarily small by making k sufficiently large. This proves that there are stable sequences of noisy memories.

3.4 *Stability of Memories Constructed from Failure-Prone Components*

Thus far we have restricted our attention to memories in which component malfunctions are assumed to be statistically independent both from one component to another and from one use of a particular component to another use. These assumptions form the basis of a mathematical model for the component malfunctions that are commonly attributed to "noise" in the system. Unfortunately, the model does not adequately represent the most common type of component malfunction that one finds in computing systems: malfunctions where individual components fail permanently.

To see whether memories of the type considered in the previous section can be stable, if their components fail permanently, let us recall the proof of the stability theorem for noisy memories. In carrying out this proof, it is necessary to show that there are types of memories for which the probability of error per digit can be bounded (see Appendix A). If one attempts to find memories in which the

components fail permanently and for which a similar bound on the probability of error per digit exists, it becomes clear that memories constructed from these components cannot have such a time independent bound of value less than ½. This is because the probability that any particular component has failed increases with time given that components fail permanently; hence, if one hypothesizes any bound on the probability of error per digit which is less than ½, it is always possible to find a time at which the hypothesized bound is violated, showing that such a bound cannot exist. By using arguments such as those in Section 2.3, one can obtain a statistical model for the errors after each correcting cycle in terms of an equivalent channel whose capacity decreases with time. Since the capacity decreases, it is always possible to find a time at which the capacity of the equivalent channel is below the information rate of the code used in storing the information in the memory, thus precluding the possibility of effective error correction and hence the possibility of stability.

Fortunately, in most "nonspace" applications, regular maintenance is performed on computing systems, that is, components which have failed are periodically replaced with good ones. Numerous specific failure probability distributions and maintenance schemes could be considered individually; however, for the purpose of this analysis we consider instead a general case which includes many of the common probability distributions and maintenance schemes. This general case is the one for which it is possible to upper bound, during each correcting cycle, the probability that each component has failed up to or during that correcting cycle. For example, suppose that each component is replaced every T seconds and that p_f represents the probability that any particular component initially fails during any particular correcting cycle. For this example, the desired upper bound on the probability of component failure equals $T \cdot p_f$ which, by appropriate choice of T and p_f, can be made less than ½. Notice that we are free to choose both T and p_f since, as before, we are only trying to show that there exists some memory of the type under consideration which is stable.

One can now perform an analysis identical to that performed in the previous section. Since the technique used to prove the stability theorem for noisy memories does not rely upon the assumption that component errors are statistically independent from use to use, precisely the same technique used previously can be used here to prove that periodically maintained memories can be stable. In fact, in most

cases, the changes required to make this proof apply when components fail permanently merely involve replacing the words "component error" with the words "component failure."

One change which requires some reinterpretation of terms involves the concept of a propagation failure. This concept was introduced for the purpose of establishing a condition under which the parity checks used to estimate any particular digit would be conditionally independent. To facilitate an intuitive discussion of propagation failures, the original definition of a propagation failure was made more general than necessary for the mathematical analysis. This analysis, in Appendix B, uses the fact that undesirable statistical dependencies can occur only if the effects of previous component malfunctions form a Δ propagation path in some parity check set tree. It is now desirable to redefine a propagation failure in terms of the formation of such a Δ propagation path. This is because permanent component failures can result in recurrent errors in a particular digit during m or more iterations thus causing a propagation failure, according to the original definition. However, since these recurrent errors do not lead to the undesirable statistical dependencies unless they also correspond to an undesirable Δ propagation path, the original definition of a propagation failure should be changed to exclude recurrent errors.

Once these changes have been made, if one represents the bounds on the probabilities of component failures by the same symbols that were used previously to represent the actual probabilities of component errors during each iteration, not only is the method of proving the stability theorem identical to that used previously but so are the forms of all the results. Thus one proves the following theorem.

Theorem 2: There is a stable sequence of memories where every component in each of the memories in the sequence has a nonzero probability of permanent failure but where components which have failed are periodically replaced with good ones.

IV. CONCLUSIONS

In Section I we compare the results obtained by Shannon concerning the reliability of communication systems with the results obtained by von Neumann, Elias, Winograd, and Cowan concerning the reliability of computing systems. Shannon's results were basically different from the other results considered. Shannon was able to show that it is possible to design arbitrarily reliable communication systems through which information can be transmitted at a nonzero

information rate. The maximum rate for which the probability of error can be made arbitrarily small is called the capacity of the communication channel. In analogy with this result, one might expect that it should be possible to design arbitrarily reliable computing systems which have a bounded redundancy. Unfortunately, none of the computing systems proposed previously has this property; therefore, none of these computing systems has a nonzero "computing capacity."

In Sections II and III we restrict our attention to one part of a computing system, namely the memory. It is shown that there are noisy memories and periodically repaired memories constructed from failure-prone components which have the property that the probability of failure can be made arbitrarily small for certain bounded values of the redundancy. The memories which have this property are called "stable memories." This result is analogous to Shannon's result and is basically different from the other results obtained thus far concerning the reliability of computing systems. The fact that it is possible to make a memory arbitrarily reliable while keeping its redundancy bounded indicates that a memory has an "information storage capacity" analogous to the capacity of a communication channel. The information storage capacity of a particular memory equals the reciprocal of the minimum redundancy for which the memory is stable; hence it can be expressed in bits per component. It is a function of the probabilities of error for the components within the memory. The method used to prove the stability theorem does not allow one to compute an explicit value for the information storage capacity of the memories which were considered; however, the fact that these memories can be stable indicates that they do have a nonzero information storage capacity.

V. ACKNOWLEDGMENTS

I would like to sincerely thank Professor Robert M. Fano, who supervised this work, for suggesting the approach to the problem and supplying guidance and encouragement throughout the research. I also wish to thank Professor Robert G. Gallager who was extremely helpful in connection with this work. Many of the results presented here are based on results originally obtained by Professor Gallager. Finally, I wish to thank Professors Peter Elias and Claude E. Shannon who helped to formulate the problem and contributed valuable suggestions and constructive criticism.

APPENDIX A

A Bound on the Probability of Error Per Digit

The first step in computing a bound on the probability of a propagation failure is to bound the probability of error for any digit stored in the registers within the memory. We assume that initially one set of J code words is transmitted across the noisy channels and inserted into these noisy registers. Some time during the next τ seconds the correcting network reads the contents of the registers and starts to perform the first correcting cycle on the newly inserted digits. The time at which this first correcting cycle starts is denoted by $t = 0$ and successive correcting cycles start at $t = \tau, 2\tau, 3\tau, \ldots$. We denote the instant just before the end of the first correcting cycle by $t = \tau - \delta$. If a digit is in error at $t = \tau - \delta$, at least one of the following events must have occurred:

(*i*) An error was made in transmitting the digit across the BSC. The probability of this event is p_c.

(*ii*) The digit was changed because of a component error in the register which occurred either during the time interval $-\tau < t \leqq 0$ or $0 < t \leqq \tau - \delta$. The probability of this event is less than $2p_r$.

Therefore, by the union bound, the probability of error per digit at $t = \tau - \delta$ is bounded by

$$\Pr[\text{digit} = \epsilon \text{ at } t = \tau - \delta] < 2p_r + p_c < p_0$$

where the parameter p_0 has been introduced to simplify the form of the results. Other conditions on p_0 will be imposed later.

Next let us compute a bound on the probability of error per digit at $t = 2\tau - \delta, 3\tau - \delta, \ldots$. If the digit d_0 is in error at any one of these times, at least one of the following events must have occurred:

(*i*) A set of $J/2$ parity checks used to estimate d_0 were in error during the last correcting cycle performed on d_0.

(*ii*) The decision device made an error during the last correcting cycle.

(*iii*) An error occurred while d_0 was stored in the register.

There are $\binom{J-1}{J/2}$ possible events of the first type. We now compute the probability that any one of these events occurred. If the ith parity check used to estimate d_0, c_i, were in error there must have been at least one error among the $K - 1$ adders used to evaluate this parity check, or at least one error among the $K - 1$ digits denoted by d_{i1},

d_{i2} , $\cdots$, d_{iK-1} . According to the union bound, the probability that c_i (for any $0 < i \leqq J - 1$) is in error is bounded by

$$\Pr [c_i = \epsilon] < \sum_{j=1}^{K-1} \Pr [d_{ij} = \epsilon] + (K - 1)p_a .$$

During the first correcting cycle $\Pr [d_{ij} = \epsilon] < p_0$ for all $0 < j \leqq K - 1$; therefore, for this iteration

$$\Pr [c_i = \epsilon \text{ during first correcting cycle}] < (K - 1)p_0 + (K - 1)p_a .$$

To compute the probability of error per digit just before the end of the second iteration, we use the fact that, during the first m iterations, the structure of the code guarantees that errors in the parity checks used to estimate any particular digit are statistically independent. To see this, consider the parity-check set tree rising from the digit d_0 as shown in Fig. 6. Each node on this tree represents a particular digit. With each digit there is associated a set of components used to compute each estimate of the digit. Just as the tree represents a history of the digits which have been involved in the computation of the successive estimates of d_0 , it also represents a history of the components which have been involved in the computation of these estimates. The later interpretation is more useful for our purposes since it is the components which cause the errors. If the code has m independent iterations, all the digits on the first m tiers of the tree must be different and all the components associated with these digits must be different also. There are $(K - 1)(J - 1)$ digits on the first tier of this tree. Provided that $m \geqq 1$, the errors in these digits after the first iteration must be statistically independent since the digits are all different, and hence the components used to compute the estimates of these digits are all different. (It is assumed that errors in different components are statistically independent.) In general, the errors in the digits on the first tier of the tree, and hence the $J - 1$ parity checks, are statistically independent provided that the sets of components used to compute the estimates of these digits are disjoint. The structure of the tree guarantees that this condition will be satisfied for the first m iterations.

Since, during the first m iterations, the errors in the parity checks used to estimate any particular digit are statistically independent, the probability that a set of $J/2$ parity checks is in error equals the product of the probabilities that each one of these $J/2$ parity checks is in error. Thus the probability of error per digit at $t = 2\tau - \delta$ is

bounded by

$$\Pr[\text{digit} = \epsilon \text{ at } t = 2\tau - \delta] < \binom{J-1}{J/2}[(K-1)(p_0 + p_a)]^{J/2} + p_d + p_r \triangleq p_1 .$$

Since we are not attempting to show that all memories of the type under consideration are stable but merely that there exist some memories of this type which are stable, we shall restrict our interest to those memories for which it is possible to make $p_1 < p_0$. This is not a serious restriction since in most cases there is no difficulty in bounding p_1 by p_0. For example, if

p_a, p_r and $p_d = 10^{-9}$, $p_0 = 10^{-8}$, $J = 14$ and $K = 15$, then $p_1 \approx 2.10^{-9}$ illustrating one case where $p_1 < p_0$.

Precisely the same argument can be used to obtain a bound on the probability of error per digit at $t = 3\tau-\delta, 4\tau-\delta, \ldots, (m+1)\tau-\delta$. The results are:

$$\Pr[\text{digit} = \epsilon \text{ at } t = 3\tau - \delta] < \binom{J-1}{J/2}[(K-1)(p_1 + p_a)]^{J/2} + p_d + p_r \triangleq p_2 < p_1 < p_0 .$$

Similarly

$$\Pr[\text{pigit} = \epsilon \text{ at } t = (m+1)\tau - \delta] < \binom{J-1}{J/2}[(K-1)(p_{m-1} + p_a)]^{J/2} + p_d + p_r \triangleq p_m < p_{m-1} < \cdots < p_1 < p_0 .$$

If no propagation failure occurs at $t = m\tau$, the error pattern evaluated at that time depends on the component errors that occurred during the previous m iterations, but not on the original errors that were present at $t = 0$. Imposing this condition changes the probability of error per digit at $t = (m+1)\tau-\delta$; however, as we now show, the probability computed above is an upper bound on this conditional probability. Using Bayes rule,

$$\Pr[\text{digit} = \epsilon \mid \text{no propagation failure}] = \frac{\Pr[\text{no propagation failure} \mid \text{digit} = \epsilon] \cdot \Pr[\text{digit} = \epsilon]}{\Pr[\text{no propagation failure}]}.$$

It is easily shown, using techniques similar to those in Appendix B, that

$$\Pr[\text{no propagation failure} \mid \text{digit} = \epsilon] \leqq \Pr[\text{no propagation failure}],$$

therefore

$$\Pr[\text{digit} = \epsilon \mid \text{no propagation failure}] \leqq \Pr[\text{digit} = \epsilon].$$

If no propagation failure has occurred, the errors in the set of parity checks used to estimate any particular digit are conditionally independent. This is because, if there is no propagation failure, the errors in each set of parity checks depend on errors in unrelated, disjoint sets of components. Thus the technique used to bound the probability of error per digit during the first m iterations can be applied after the mth iteration provided that no propagation failure has occurred. The general result is

$$\Pr[\text{digit} = \epsilon \text{ at } t = (i+1)\tau - \delta \mid \text{no propagation failure}] < \binom{J-1}{J/2}[(K-1)(p_{i-1} + p_a)]^{J/2} + p_d + p_r \triangleq p_i < p_{i-1} < \cdots < p_i < p_0 .$$

This monotonically decreasing sequence of bounds shows that the probability of error per digit is upper bounded by p_0 provided that it is possible to choose p_0, J, and K such that $p_1 < p_0$ and provided that no propagation failure occurs. In general, the probability of error per digit is upper bounded by p_i provided that the memory has performed i or more correcting cycles and provided that the conditions stated above are satisfied.

APPENDIX B

The Probability of a Propagation Failure

Intuitively, the concept of a propagation failure is very simple. A propagation failure occurs whenever the present error pattern in the registers is in some way related to the component errors that occurred more than m iterations before, where m is the number of in-

dependent iterations. This intuitive concept can be made more precise by defining "error configurations" which are hypothetical error patterns that are functions of some subset of the actual set of component errors. In particular, a *type i error configuration*, for any $i > 0$, is a $J \cdot N$-tuple corresponding to the error pattern that would be present if there had been no component errors before the last i iterations. If no propagation failure has occurred, all error configurations of type m and higher must be identical, thus we are really interested in the difference between error configurations. For this reason, it is more convenient to restate the definition of a propagation failure in terms of "Δ configurations" where, for all $i > 0$, the *type i Δ configuration* is the difference between the type i and type $i+1$ error configurations. The type 0 Δ configuration is defined to be equal to the type 1 error configuration. A propagation failure occurs if there are one or more 1's in any Δ configuration of type m or higher.

Let us consider the situations which, for any i, lead to a 1 in the type i Δ configuration evaluated at some particular time, say $t = L\tau$. Suppose that there is a 1 in the position corresponding to the digit d_0 in this Δ configuration. This means that the value of the digit d_0 in the type i error configuration is different from that in the type $i+1$ error configuration where both configurations are evaluated at $t = L\tau$. This change in the value of d_0 must be related to component errors that occurred during the time interval $(L-i-1)\tau < t < (L-i)\tau$ since all the other component errors upon which the type i and type $i+1$ error configurations are based are the same. We refer to the component errors that occurred during the time interval $(L-i-1)\tau < t < (L-i)\tau$ as the *controlling errors* for the type i Δ configuration evaluated at $t = L\tau$ and we say that the value of d_0 at $t = L\tau$ has been changed by these controlling errors.

If the value of d_0 is changed at $t = L\tau$, the controlling errors must have changed at least one of the digits used to estimate d_0. These are the digits on the first tier of the parity-check set tree rising from d_0, and changes in them are represented by 1's in the appropriate positions in the type $i-1$ Δ configuration evaluated at $t = (L-1)\tau$. In general these controlling errors must have caused changes in some digits on the lth tier of the parity-check set tree, the changes being represented by 1's in the appropriate positions in the type $i-l$ Δ configuration evaluated at $t = (L-l)\tau$. These changed digits define at least one continuous path in the parity-check set tree rising i tiers from d_0. We call these paths Δ *propagation paths*. The particular Δ propagation path which has just been described is referred to as the i-tier

Δ propagation path rising from d_0 at $t = L\tau$ (see Fig. 8). Every 1 in a type i Δ configuration must have at least one i-tier Δ propagation path associated with it. To bound the probability that there is a 1 in the entry corresponding to the digit d_0 in the type i Δ configuration evaluated at $t = L\tau$, we shall bound the probability that one or more i-tier Δ propagation paths rise from the digit d_0 at $t = L\tau$.

Since Δ propagation paths must be continuous, for each 1 in a type i Δ configuration evaluated at $t = L\tau$ there must have been at least one 1 in a type $i-1$ Δ configuration evaluated at $t = (L-1)\tau$. If no propagation failure occurred before $t = L\tau$, the Δ configurations of type m and higher must have beeen all zero for all $t < L\tau$. This implies that the Δ configurations of type $m+1$ and higher must be all zero at $t = L\tau$. Hence, the only way that the initial propagation failure can occur at $t = L\tau$ is if there is a 1 in the type m Δ configuration evaluated at that time. Therefore, to bound the probability that the first propagation failure occurs at some particular time, we need

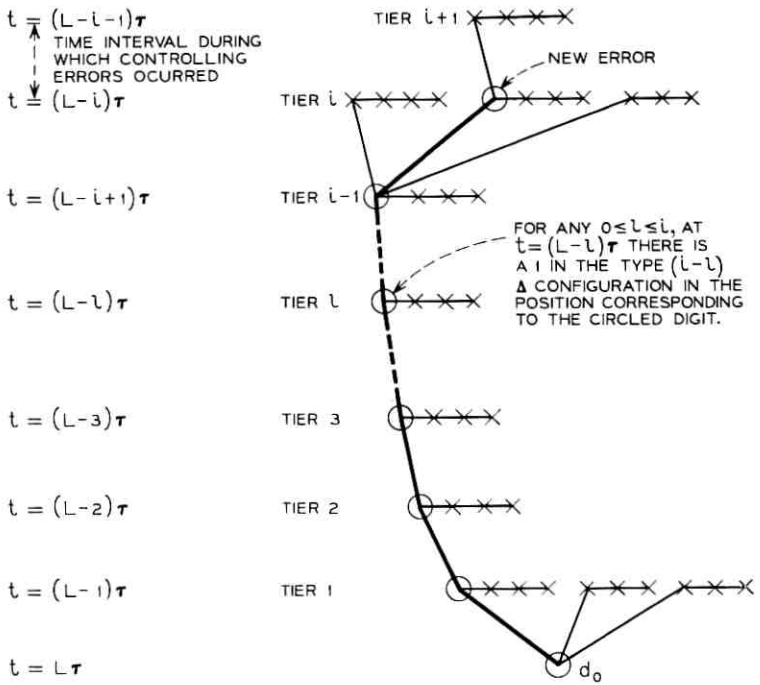


Fig. 18 — An example of an i-tier Δ propagation path rising from d_0 at $t = L\tau$.

only bound the probability of one or more 1's in the type m Δ configuration evaluated at that time. Since we assume that the memory fails whenever the first propagation failure occurs, we shall never be concerned with computing the probability of a 1 in any Δ configuration of type $m+1$ or higher. It is important that we can restrict our attention to the type m Δ configuration since this Δ configuration can be computed for any particular time by considering only the component errors that occurred during the previous $m+1$ correcting cycles. The first of these correcting cycle results in statistically independent digit errors and, as explained in Appendix A, the structure of the code guarantees that during the next m correcting cycles the errors in the parity checks used to estimate any digit are statistically independent.

To bound the probability that the initial propagation failure occurs at $t = L\tau$, we must compute a bound on the probability of one or more 1's in the type m Δ configuration evaluated at $t = L\tau$. This is done by bounding the probability that an m-tier Δ propagation path rises from one or more of the $J \cdot N$ digits in the registers within the memory at $t = L\tau$. At first we restrict our attention to one particular digit d_0. A bound is computed on the probability that an m-tier Δ propagation path rises from d_0 at $t = L\tau$. The first step in this computation is to bound the probability that the component errors that occurred during the time interval $(L-m-1)\tau < t < (L-m)\tau$ would cause any particular digit on the mth tier of the parity-check set tree rising from d_0 to be in error at $t = (L-m)\tau$. Any m-tier Δ propagation path rising from d_0 at $t = L\tau$ must terminate on one of these errors which we call *new errors*. The next step is to bound the probability that at $t = (L-m+i)\tau$ an i-tier Δ propagation path rises from any particular digit on the $(m-i)$th tier of the parity-check set tree rising from d_0. This probability is denoted by $Pr[d_{m-i}=\Delta_i]$. By substituting m for i, we obtain a bound on the probability that at $t = L\tau$ an m-tier Δ propagation path rises from the digit d_0 (that is, $Pr[d_0 = \Delta_m]$). The probability that an m-tier Δ propagation path rises from one or more of the $J \cdot N$ digits at $t = L\tau$ is upper bounded by $J \cdot N \cdot Pr[d_0 = \Delta_m]$.

The first step, namely bounding the probability of a new error at $t = (L-m)\tau$, is particularly simply. In Appendix A we computed a bound on the probability of error per digit which was denoted by p_0. Since this bound was computed by considering all possible errors that could exist at a particular time, it must certainly be a bound on the

probability of a new error at some particular time. Therefore, p_0 is a bound on the probability that a new error occurs at $t = (L-m)\tau$.

The next step is to bound Pr $[d_{m-i} = \Delta_i]$ for all $1 \leqq i \leqq m$. Let us consider a particular digit on the $(m - i)$th tier of the parity-check set tree rising from d_0 and denote this digit by d_{m-i}. Now consider the conditions which must be satisfied if the value of d_{m-i} is changed by the controlling errors; that is, if an i-tier Δ propagation path rises from d_{m-i} at $t = (L - m + i)\tau$.

To describe this "change" in more detail, we must consider two sets of component errors. One is the set of component errors that occurred since $t = (L - m - 1)\tau$ and the other is the set of component errors that occurred since $t = (L - m)\tau$ [assume that no component errors occurred before $t = (L - m - 1)\tau$ and $t = (L - m)\tau$, respectively]. When we say that the value of d_{m-i} has been changed by the controlling errors, we mean that the value of d_{m-i} at $t = (L - m + i)\tau$ is correct when it is computed under the assumption that one of these sets of component errors actually occurred, whereas it is incorrect when it is computed under the assumption that the other set of errors actually occurred. There are two necessary conditions for this change. Assume that the value of d_{m-i} was changed by the controlling errors. Denote the set of component errors for which $d_{m-i} = \epsilon$ by $\mathcal{S}_\epsilon$ and the set for which $d_{m-i} \neq \epsilon$ by $\mathcal{S}_{\text{correct}}$. If the value of d_{m-i} is changed by the controlling errors, both of the following conditions must be satisfied:

(*i*) There must have been at least one parity check used to estimate d_{m-i} which was wrong [at $t = (L - m + i)\tau$] under the assumption that $\mathcal{S}_\epsilon$ occurred but which was correct under the assumption that $\mathcal{S}_{\text{correct}}$ occurred. Denote one of these parity checks by c_Δ.

(*ii*) On the basis of the errors in the set $\mathcal{S}_\epsilon$, $J/2 - 1$ or more parity checks other than c_Δ must have been wrong at $t = (L - m + i)\tau$.

In order for condition i to be satisfied, the value of at least one of the $(J-1)(K-1)$ digits immediately above d_{m-i} in the parity-check set tree must have been changed [at $t = (L-m+i-1)\tau$] by the controlling errors. The probability of such a change has been denoted by $Pr[d_{m-i+1} = \Delta_{i-1}]$. Therefore, the probability that the value of one or more of these digits was changed is upper bounded by

Pr [condition i is satisfied]

$<$ Pr [value of any digit immediately above d_{mi} is changed by controlling errors]

$<$ $(J - 1)(K - 1)$ Pr $[d_{m-i+1} = \Delta_{i-1}]$.

A bound on the probability that condition *ii* is satisfied was derived in Appendix A. This bound equals

$$\Pr[\text{condition } (ii) \text{ is satisfied}] < \binom{J-2}{J/2-1}[(K-1)(p_0+p_a)]^{J/2-1}.$$

As explained previously, the structure of the code guarantees that parity-check errors are independent; hence, during the m iterations of interest, these two conditions are independent. Therefore, the probability that both conditions are satisfied, which is a bound on $Pr[d_{m-i} = \Delta_i]$, is given by

$$\Pr[d_{m-i} = \Delta_i] < (J-1)(K-1)\Pr[d_{m-i+1} = \Delta_{i-1}] \cdot \binom{J-2}{J/2-1}[(K-1)(p_0+p_a)]^{J/2-1}.$$

Substituting $i = 1, 2 \ldots m$, we obtain

$$\Pr[d_{m-1} = \Delta_1] < (J-1)(K-1)p_0\binom{J-2}{J/2-1}[(K-1)(p_0+p_a)]^{J/2-1}$$

$$\Pr[d_{m-2} = \Delta_2] < (J-1)(K-1)\Pr[d_{m-1} = \Delta_1] \cdot \binom{J-2}{J/2-1}[(K-1)(p_0+p_a)]^{J/2-1}$$

$$< p_0\left\{(J-1)(K-1) \cdot \binom{J-2}{J/2-1}[(K-1)(p_0+p_a)]^{J/2-1}\right\}^2$$

$$\vdots$$

$$\Pr[d_0 = \Delta_m] < p_0\left\{(J-1)(K-1) \cdot \binom{J-2}{J/2-1}[(K-1)(p_0+p_a)]^{J/2-1}\right\}^m.$$

Gallager has found a technique for constructing low-density parity-check codes[10] with m, the number of independent iterations, bounded by

$$\frac{\log\left[\frac{N}{2K} - \frac{N}{2J(K-1)}\right]}{2\log[(J-1)(K-1)]} \leqq m \leqq \frac{\log N}{\log[(J-1)(K-1)]}.$$

Substituting this lower bound into the equation for $Pr[d_0 = \Delta_m]$ gives

$$\Pr[d_0 = \Delta_m] < p_0\left\{(J-1)(K-1)\binom{J-2}{J/2-1} \cdot [(K-1)(2p_0)]^{J/2-1}\right\}^{\log[(N/2K)-[N/2J(K-1)]]/2\log[(J-1)(K-1)]}$$

$$= p_0\left\{\left[\frac{1}{2K} - \frac{1}{2J(K-1)}\right]N\right\}^{-\beta}$$

where

$$\beta \triangleq -\frac{\log\left\{(J-1)(K-1)\binom{J-2}{J/2-1}[(K-1)(2p_0)]^{J/2-1}\right\}}{2\log(J-1)(K-1)}.$$

We have assumed that p_0 has been chosen such that $p_0 > p_a$. For any L, probability that the initial propagation failure occurs at $t = L\tau$ is bounded by

$$\Pr[\text{initial propagation failure occurs at } t = L\tau]$$

$$= 0 \qquad \text{for } L < m$$

$$< J \cdot N \cdot \Pr[d_0 = \Delta_m] \qquad \text{for } L \geqq m$$

since, by definition, a propagation failure cannot occur before $t = m\tau$.

We have chosen to number the memories according to their information storage capability, k. We can express this result in terms of k by noting that $N = k/R$ and $1 - J/K \leqq R \leqq 1$; therefore,

$$\Pr[\text{initial propagation failure occurs at } t = L\tau]$$

$$= 0 \qquad \text{for } L < m$$

$$< J \cdot \left[\frac{k}{1 - \frac{J}{K}}\right] \cdot p_0 \cdot \left\{\left[\frac{1}{2K} - \frac{1}{2J(K-1)}\right]k\right\}^{-\beta} \qquad \text{for } L \geqq m$$

$$= C \cdot k^{-\beta+1}$$

where

$$C \triangleq \frac{J}{1 - J/K} \cdot p_0 \cdot \left[\frac{1}{2K} - \frac{1}{2J(K-1)}\right]^{-\beta}.$$

Both C and β are functions of J, K and p_0. For any particular sequence of memories, J, K and p_0 will all be constants. For example, if

$J = 14$, $K = 15$, and $p_0 = 10^{-8}$, then $\beta = 7.55$; therefore, in this case, the probability that the first propagation failure occurs at $t = L\tau$ (for any L) is bounded by a function that decreases as the sixth power of the information storage capability of the memory. By choosing the information storage capability sufficiently large, this probability can be made arbitrarily small.

REFERENCES

1. von Neumann, J., "Probabilistic Logics and the Synthesis of Reliable Organisms from Unreliable Components," in *Automata Studies*, ed. C. E. Shannon and J. McCarthy, Princeton, New Jersey: Princeton University Press, 1956, pp. 43–98.
2. Allanson, J. T., "The Reliability of Neurons" Proc. First Congress on Cybernetics, 1956, Paris: Gauthier-Villars, 1959, pp. 687–694.
3. Shannon, C. E., "A Mathematical Theory of Communication," B.S.T.J., *27*, Nos. 3 and 4 (July and October 1948), pp. 379–423, and 623–656.
4. Elias, P., "Computation in the Presence of Noise," IBM J. Research and Development, *2* (October 1958), pp. 346–353.
5. Elias, P., "Coding for Two Noisy Channels" in *Information Theory*, ed. Colin Cherry, New York: Academic Press, 1956, pp. 61–74.
6. Wozencraft, J. M., and Reiffen, B., *Sequential Decoding*, New York: Technology Press and John Wiley and Sons, Inc., 1961.
7. Fano, R. M., "A Heuristic Discussion of Probabilistic Decoding," IEEE Trans. Inform. Theory, *IT-9* (April 1963), pp. 64–74.
8. Winograd, S. and Cowan, J. C., *Reliable Computation in the Presence of Noise*, Cambridge, Massachusetts: MIT Press, 1963.
9. Gallager, R. G., "A Simple Derivation of the Coding Theorem and Some Applications," IEEE Trans. Inform. Theory, *IT-11* (January 1965), pp. 3–18.
10. Gallager, R. G., *Low-Density Parity-Check Codes*, Cambridge, Massachusetts: MIT Press, 1963.

Reliable Computation in Computing Systems Designed from Unreliable Components*

By MICHAEL G. TAYLOR

(Manuscript received April 10, 1968)

This is the second of two papers which present information-theory-type results pertaining to the reliability of computing systems designed from unreliable components. Two models for component malfunctions are considered. The first is based on the assumption that malfunctions of a particular component are statistically independent from one use to another. The second is based on the assumption that components fail permanently but that the components which have failed are periodically replaced with good ones. In both cases, malfunctions in different components are assumed to be independent. Just as a channel capacity is defined for communication channels, a computing capacity is defined for computing systems. For both component failure models, it is shown that there are computing systems, designed entirely from unreliable components of the assumed type, which have nonzero computing capacities.

I. INTRODUCTION

The objective of this paper is to show that it is possible for a computer, designed entirely from unreliable components, to perform operations reliably on information stored in stable memories. The concept of a stable memory is introduced in the paper preceding this, where it is shown that it is possible to store information reliably in memories constructed entirely from unreliable components.[1] Two different models for component malfunctions are considered. The first is based on the assumption that component malfunctions are statistically independent from one use of a particular component to another use. The second is based on the assumption that components

* This work, which is based on part of a doctoral thesis submitted to the Department of Electrical Engineering, M.I.T., September 1966, was supported by the National Aeronautics and Space Administration (Grant NsG-334).

fail permanently but that the components which have failed are periodically replaced with good ones. In both cases component malfunctions are assumed to be statistically independent from one component to another. For both component malfunction models it is shown that there are types of memories, called "stable memories," that have a nonzero information storage capacity; that is, for certain fixed values of the memory's redundancy, the probability of a memory failure can be made arbitrarily small. A particular type of stable memory is considered in some detail.

Since the operation of the computers to be described in this article is closely related to the operation of these stable memories, let us look briefly at these memories. They consist of several registers and a correcting network as shown in Fig. 7 of Ref. 1. The registers store information which is coded according to a low-density parity-check code[2] and the correcting network periodically monitors the contents of the registers, corrects errors and reinserts the corrected words into the registers. The correcting network is very similar to a low-density parity-check decoder. Such a decoder, if constructed from reliable components, decreases the probability of error for digits stored in the registers with each successive correcting cycle (iteration) provided that the initial probability of error is not too large.

In order to understand the operation of a low-density parity-check corrector constructed from unreliable components, let us suppose that the initial probability or error for stored digits is somewhat higher than the probability of malfunction for any component in the corrector. For the first few interations the corrector decreases the probability of error per digit almost as much as the reliable decoder does. However, eventually this probability becomes comparable to the probability that a new error is made by the correcting network itself. Thus an equilibrium probability of error per digit is established such that the probability of error per digit before and after each correcting cycle is the same. Figure 1 is typical graph of this probability.

Although the probability of error per digit approaches an asymptotic value, it is still possible for a memory failure to occur. A memory failure occurs whenever the configuration of errors within the registers of the machine is such that a noiseless decoder would be unable to correct all the errors. The probability of a memory failure during the time interval $0 \leqq t \leqq \mathcal{L}\tau$ is calculated in Ref. 1 and it is shown to have the same form for both component malfunction models considered. For any k and $\mathcal{L}$, the probability of a memory failure in M_k, a memory of the type under

consideration which can store k bits of information, is given by

$$\Pr[\text{failure of } M_k \text{ during } 0 \leqq t \leqq \mathfrak{L}\tau] < (\mathfrak{L} + 1) \cdot C' \cdot k^{-\beta'}$$

where C' and β' are constants for any particular type of memory, or to be more mathematically precise, for any particular sequence of memories, $\{M_i\}$, where the members of the sequence are ordered according to their information storage capabilities. It also is shown that it is possible to make $\beta' > 0$; in fact, a numerical example is presented where $\beta' = 5.55$. Therefore, for this example, it is possible to make the probability of a memory failure arbitrarily small by choosing k, the number of bits stored, sufficiently large. Furthermore, it is shown that increasing k does not increase the redundancy of the memories being considered, thus completing the proof of stability for these memories.

Our objective is to show that it is possible to design a reliable computer which performs arithmetic operations on operands stored in stable memories and which presents the result in a form that can itself be stored in a stable memory. The latter condition assures that successive arithmetic operations can be performed by computers of this type. Just as stability is defined to be a property of sequences of memories, reliability will be defined to be a property of sequences of computing systems.

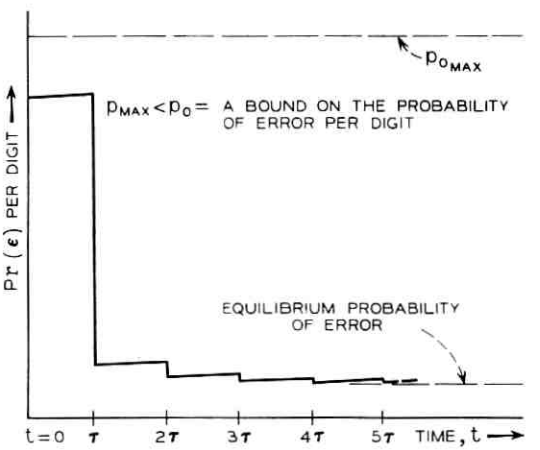


Fig. 1 — The probability of error per digit.

II. DEFINITION OF RELIABILITY

The computing systems we are considering consist of a number of stable memories for information storage and some logic circuits which perform the arithmetic operations. The operations are performed on operands which have been stored in stable memories and the results of the operations are stored in stable memories. Before it is possible to define "reliability," it is necessary to establish a criterion for determining whether a particular result is correct. We consider a result (coded) to be correct if a noiseless low-density parity-check decoder could correct all the errors; that is, if it could obtain the desired uncoded result. Thus, the class of correct results is precisely the decoding equivalence class which contains the code word whose information digits correspond to the desired uncoded result.

Now consider a sequence of computing systems $\{S_i\}$ where, for any k, the computing system S_k contains memories having an information storage capability of k binary digits. We require that all the computing systems in $\{S_i\}$ be able to perform the same set of operations. The sequence $\{S_i\}$ is called *reliable* if it satisfies the following conditions:

(*i*) For all k, the redundancy of S_k must be less than α where α is a constant independent of k. The redundancy of a computing system is defined as the ratio of the complexity of the system to the amount of computation performed by the system.*

(*ii*) For any $\mathfrak{I} > 0$ and $\delta > 0$, there must be a member of $\{S_i\}$ for which the probability that the result of any sequence of $\mathfrak{I}$ operations (within the allowed set) will be in error is less than δ.

The reciprocal of the minimum redundancy for which a particular sequence of computing systems is reliable is called the *computing capacity* for these systems.

III. OPERATION OF VECTOR ADDITION MODULO-2

The operation of vector (bit-by-bit) addition modulo-2 is considered first because low-density parity-check codes have the property that when this operation is performed on two code words, the result is always another code word. This is the only nontrival operation that can be performed without using some elaborate procedure for generating the check digits required to form the coded result. For

* Ref. 1 gives precise definitions of "complexity" and "amount of computation."

this reason, the operation of modulo-2 addition is particularly easy to perform.

The computing system to be considered consists of stable memories which store the operands and the results, and logic circuits which perform the arithmetic operation. It is assumed that all components within the computing system have a nonzero probability of malfunction. Let the stable memories containing the operands for one particular operation be denoted by M_k' and M_k'' . At $t = T$, the operation of vector addition modulo-2 is performed on the contents of these memories and the result is stored in another stable memory. It is assumed that all the memories within a particular computing system are physically identical. This means that all memories must have the same set of possible states; furthermore, since the states of a stable memory are divided into classes of equivalent states, the classes of states must be the same for all of these stable memories. Let us suppose that at $t = T$, the state of M_k' belongs to $C(I_{ki_1})$, the class of states containing the code word I_{ki_1} , and that the state of M_k'' belongs to $C(I_{ki_2})$. If the operation is performed correctly, the state of the stable memory containing the result will belong to $C(I_{ki_1} \oplus I_{ki_2})$.

To show that computing systems of this type are reliable, we must show that a sequence of $\mathfrak{I}$ operations can be performed with an overall probability of error that can be made arbitrarily small by choosing k sufficiently large while keeping the redundancy fixed. To compute the redundancy we evaluate the ratio of the complexity of the system to the amount of computation performed by the system. Each modulo-2 vector addition performed by the system involves three stable memories, each with an information storage capability of k, and a number of modulo-2 adders equal to the number of information storage components in one stable memory.

In Ref. 1 it was shown that the complexity of a stable memory is proportional to k; hence the complexity of these three memories and the associated modulo-2 adders must also be proportional to k. The amount of computation, that is, the number of two-input binary operations that an equivalent irredundant computer would perform, equals k since this irredundant computer would perform k additions (modulo-2) on its two k-bit operands. Since both the complexity of the systems being considered and the amount of computation that they perform are proportional to k, their ratio, the redundancy, is independent of k as required.

Let us start by considering just one operation performed by the computing system. We assume that before the operation was performed

both M_k' and M_k'' performed at least one correcting cycle on the stored operands. Thus, according to Fig. 1, the probability of error for digits stored in each of these memories is upper bounded by p_1 which, in general, is very small compared with the maximum allowable probability of error per digit. After the operation has been performed, the probability of error (ϵ) for digits in the memory storing the result is upper bounded by

$$\Pr[\text{digit} = \epsilon] < 2p_1 + p_a + 2p_r$$

since any one of the following events could lead to an error in one particular digit:

(*i*) The corresponding digit was in error in M_k'(probability $\leqq p_1$).

(*ii*) The corresponding digit was in error in M_k''(probability $\leqq p_1$).

(*iii*) The adder that performed the operation on these digits made an error (probability denoted by p_a).

(*iv*) An error occurred in this particular digit position in the result memory between the time the result was stored and the end of the first correcting cycle performed on the result (probability $\leqq 2p_r$).

Provided that the resulting probability of error per digit is less than the maximum allowable value, successive iterations of the result memory decrease this probability as shown in Fig. 1. After the result memory has performed one iteration, the contents of this memory can be used as an operand in a second operation. Successive operations can be performed provided that at least one correcting cycle is performed on each intermediate result.

An error is made on one of these operations if the state of the result memory does not belong to the desired class of states, or equivalently, if a memory failure occurs within the result memory immediately following the operation. The method for computing the probability of such an error is almost identical to that used in Ref. 1 to compute the probability of a memory failure. The result, which is obtained in the Appendix, is:

$$\Pr[\text{any } \oplus \text{ operation is performed incorrectly} \mid \text{all previous } \oplus \text{ operations were performed correctly}] < \frac{C}{(1 - J/K)^2} \cdot k^{-\beta+3}$$

where C and β are functions of the parameters of the code (J and K) and p_0, the bound on the probability of error per digit. For any par-

ticular sequences of computing systems, J, K, and p_0, and hence C and β, are all constants. This result applies for either of the component malfunction models discussed in the introduction.

Finally we must bound the probability that an error occurs on any one of a sequence of $\mathfrak{z}$ operations. It is assumed that neither a memory failure nor a propagation failure occurred in any stable memory between the time when an operand was originally stored and the time when the first operation was performed on that operand. We need not concern ourselves about the concept of a propagation failure except to notice that the bounds on the probability of a memory failure were actually derived by bounding the probability of either a memory failure or a propagation failure. Hence, imposing the condition of no propagation failure does not make the requirements any different from those which have already been used. The probability of an error during $\mathfrak{z}$ operations is upper bounded by the sum of the probabilities that the initial error occurs on any one of these operations; that is,

$$\Pr[\text{error during a sequence of } \mathfrak{z} \oplus \text{ operations} \mid \text{no memory failure or propagation failure in memories containing the original operands}] < \mathfrak{z} \cdot \frac{C}{(1 - J/K)^2} \cdot k^{-\beta+3}.$$

If $J = 14$, $K = 15$, and $p_0 = 10^{-8}$ then $\beta = 7.55$; therefore, for this sequence of computing systems, the probability of an error in $\mathfrak{z}$ vector modulo-2 additions can be made arbitrarily small by choosing k sufficiently large, thus providing that this sequence of computing systems is reliable.

IV. GENERAL VECTOR OPERATIONS

Consider a sequence of computing systems, $\{S_i\}$, in which each system is capable of performing many different operations. The inputs to each of these systems are stored in stable memories, as before; however, the inputs must specify not only the operands but also the desired operations. The digits stored in each memory are coded according to a low-density parity-check code, but now it is assumed that the code is in systematic form; that is, the information digits appear first in each code word. When an operation is performed on two code words, the desired result is the code word whose information digits are computed by performing the desired operation on the information digits of the two operands.

The allowed operations for the computing systems under consideration consist of any vector (bit-by-bit) operation on the information digits of the operands. That is, the operation performed on a particular pair of information digits in the operands can be any one of the 16 Boolean functions of two variables. Each of these operations is assigned a four-bit operation code. Since different operations can be performed on different pairs of information digits in the operands, the computing system S_k, which contains memories having a storage capability of k bits, is able to perform $(16)^k$ different operations on any pair of operands. The desired operation is specified by means of four code words. For all $0 < i \leqq k$, the set of four digits in the ith digit position in each of these four code words gives the operation code for the operation to be performed on the ith information digit in each operand. For each pair of information digits in the operands, the computing system first selects the operation specified by the operation code and then performs this operation on the appropriate pair of digits in the operands. In this way all the information digits in the result are computed.

Since the desired result is a code word, it is necessary for the computing system to generate the appropriate check digits to go with the desired information digits of the result. The operation of vector addition modulo-2 is particularly easy to perform on the information digits of two code words because the appropriate check digits can be generated by performing the operation of vector addition modulo-2 on the check digits of the two operands. In general, it is more difficult to generate the appropriate check digits.

We consider first a method by which a noiseless system could generate the check digits, then show that by making a rather simple modification to this method, it is suitable for use in a noisy system. Finally we bound the probability that an error occurs in a sequence of 3 operations performed using this modified method. To simplify the terminology throughout this and subsequent discussions, we contrast noiseless and noisy systems. A noiseless system is one in which there are no component malfunctions whereas, for these discussions, a noisy system is any one in which the components have nonzero probabilities of malfunction. All results presented apply for either of the component malfunction models described in the introduction.

Let us consider how a noiseless system consisting of memories, operation selectors, and computing devices might compute the desired result. Since there are 16 operations that could be performed on any pair of information digits in the operands, there must be 16 types of computing devices. One device of each type and one operation selector

is associated with each pair of information digits in the operands. Each operation selector decodes the operation code and selects the appropriate computing device. This device then performs the desired computation on one pair of information digits. In this way all the information digits in the desired result are computed. These resulting information digits are represented by the row vector $\mathbf{v}$.

The second step in computing the desired result is to compute the check digits. Let us assume that, within the computing system, k noiseless memories are used to store a set of k basis vectors for the code. This set of k basis vectors is represented by the generator matrix, $\mathbf{G}$. It is assumed that the basis vectors have been chosen in such a way that $\mathbf{G}$ is in reduced-echelon form. (That is, $\mathbf{G} = [\mathbf{I}_k\mathbf{P}]$ where $\mathbf{I}_k$ is a $k \times k$ identity matrix and $\mathbf{P}$ is a $k \times (N - k)$ matrix. The notation $\mathbf{D} = [\mathbf{AB}]$ means that the matrix $\mathbf{D}$ can be partitioned into two submatrices, $\mathbf{A}$ and $\mathbf{B}$.)[3] The desired result is obtained by performing the operation $\mathbf{v} \odot \mathbf{G}$ where $\odot$ represents the operation of matrix multiplication in which all additions are performed modulo-2.

This matrix multiplication operation is performed in two steps. The first step consists of performing the bit-by-bit AND of the digits represented $\mathbf{v}^T$ with the digits represented by each column of $\mathbf{G}$ (see Table I). The resulting row vectors, represented by the matrix $\mathbf{G}'$, are stored in another set of k noiseless memories. A row of $\mathbf{G}'$ is all zeros if the corresponding digit in $\mathbf{v}^T$ is 0 but it is identical to the same row of $\mathbf{G}$ if the corresponding digit in $\mathbf{v}^T$ is 1. Therefore, the nonzero rows of $\mathbf{G}'$ are the generators of the desired result. The final step in performing the matrix multiplication operation consists of adding modulo-2 the rows of $\mathbf{G}'$.

Let us see if a noisy computing system, containing stable memories rather than noiseless memories, can obtain the desired result in the

TABLE I—AN EXAMPLE OF THE OPERATION $\mathbf{v} \odot \mathbf{G}$.

$$\mathbf{v} = [1 \quad 0 \quad 1]$$

$$\mathbf{G} = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \end{bmatrix}$$

$\mathbf{v}^T$ ANDed with columns of $\mathbf{G}$

$$= \begin{bmatrix} 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 \end{bmatrix}$$

$$\triangleq \mathbf{G}'$$

Vector sum modulo-2 of the rows of $\mathbf{G}'$

$$= [1 \quad 0 \quad 1 \quad 1 \quad 0]$$

$$\triangleq \mathbf{v} \odot \mathbf{G}$$

same way. The information digits of the result are computed by the method described above. This time, however, there is a nonzero probability that any particular information digit is in error. There is now a problem when we try to use these digits to find the generators of the desired result. If there is any error in these information digits, the generators of the result will be specified incorrectly, in which case the resulting state of the stable result memory will almost certainly be outside the desired class of states. Since there is no way to eliminate the errors in the information digits, the computing method as described is not suitable for use in a noisy computing system.

The problem with the computing method that has just been described is that there is only one copy of the information digits of the result and there is no way to guarantee that this copy is error free. Suppose that we had many copies of these information digits such that the errors were statistically independent from one copy to another. We shall show later that it is actually possible to obtain copies with these properties. These copies of the desired information digits can be represented by means of a matrix $\mathbf{V}$ in which each column represents one of these copies. Each row of $\mathbf{V}$ is almost all zeros or almost all ones and the errors across any row of $\mathbf{V}$ are statistically independent. Each row of $\mathbf{V}$ that is almost all ones indicates that the corresponding row of the generator matrix is one of the generators of the desired result. Suppose that the generators of the code are stored in stable memories, the generator set. Since each stable memory containing k bits of information actually contains $J \cdot N$ binary digits (N is the block length of the code and J is a parameter of the code described previously), the contents of these stable memories in the generator set can be represented by a $k \times JN$ matrix $\mathcal{G}$. If we require that $\mathbf{V}$ be a $k \times JN$ matrix too (that is, that we have JN copies of the information digits), an operation can be performed on the memories equivalent to ANDing corresponding entries in the matrices $\mathcal{G}$ and $\mathbf{V}$, the result being represented by a $k \times JN$ matrix $\mathcal{G}'$. This operation leaves the desired generators essentially unchanged but it replaces the undesired generators with vectors which are equivalent to the zero vector.

In this case an error in one copy of the information digits causes at most one error in the digits represented by $\mathcal{G}'$. This is much better than the result of the previous method where one error in the information digits causes errors in an entire row of $\mathbf{G}'$. If the probability of error per digit in $\mathcal{G}'$ is not too high, it would be hoped that each stable memory corresponding to a row of $\mathcal{G}'$ would be able to reduce this probability of error by performing one correcting cycle. The operation of vector addition modulo-2 could then be performed to obtain the desired result. Since

this method looks promising, it will be considered in more detail. The first step is to make sure that it is possible to obtain copies of the information digits with the desired properties. Then a sequence of computing systems which perform operations by this method is analyzed to determine if the sequence can be made reliable.

When the computing system is ready to operate on a particular operand it copies the operand JN times and each copy is stored in a stable memory. A set of JN stable memories corresponding to one operand is called an operand set (see Fig. 2). An operand matrix, $\mathbf{A}_i$ $i = 1, 2$, is defined in the following way. Each column of $\mathbf{A}_i$ is associated with one stable memory in the corresponding operand set. The digits in a particular column of $\mathbf{A}_i$ equal the digits in the first noisy register within the stable memory corresponding to that column of $\mathbf{A}_i$. Recall that each register within a stable memory contains a noisy approximation to the same code word. Thus the digits in each row of an operand matrix are almost all ones or almost all zeros. The digits different from the dominant one in each row correspond to errors in different stable memories within the operand set. Let each operand memory perform m (m = the number of independent iterations[2]) iterations on the digits contained in it. If no memory failure or propagation failure occurs in any stable memory within an operand set, the errors in the digits stored

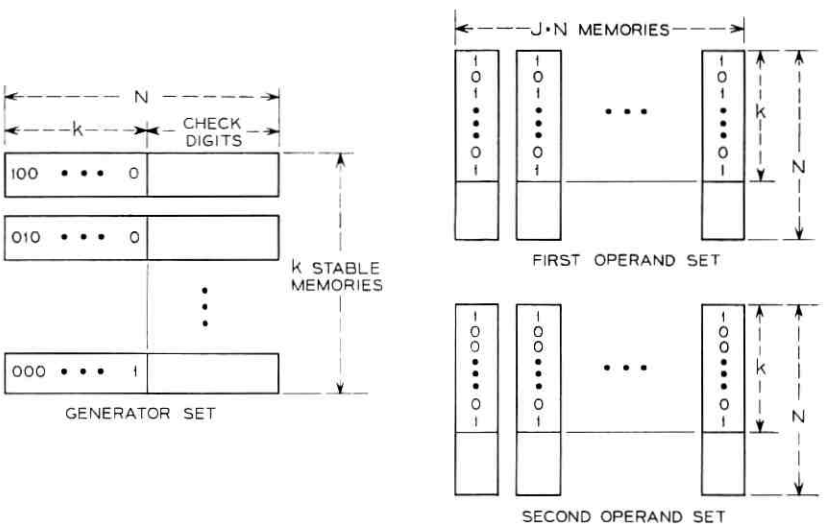


Fig. 2 — The generator set and typical operand sets. The digits are typical of the digits in each of the J noisy registers within the stable memory.

within a stable memory depend only on the component errors that occurred in that memory during the last m iterations,[1] and these component errors are assumed to be statistically independent from one stable memory to another. Since the contents of these stable memories are represented by the columns of the operand matrix $\mathbf{A}_i$, the errors in one column of this matrix are statistically independent of the errors in any other column. Therefore, after m iterations, the errors across any *row* of the operand matrix are statistically independent.

The four stable memories which contain the code words that specify the operation code are called the "operation code memories." The contents of each of these memories are copied JN times to form an operation code set (see Fig. 3). Each memory in each operation code set performs m iterations just as the memories in each operand set do. An operation code matrix $\mathbf{\Theta}_i$, $0 < i \leqq 4$, is defined for each operation code set; the definition is analogous to the definition of each operand matrix as stated previously.

Let us consider a particular digit position in each of the operation code matrices and each of the operand matrices. The four digits in this position in the four operation code matrices form the operation code for the operation to be performed on the two digits in this position in the two operand matrices. After each memory in each set of memories has performed m iterations, the operations specified by the operation code matrices are performed. The digits which are the results of these opera-

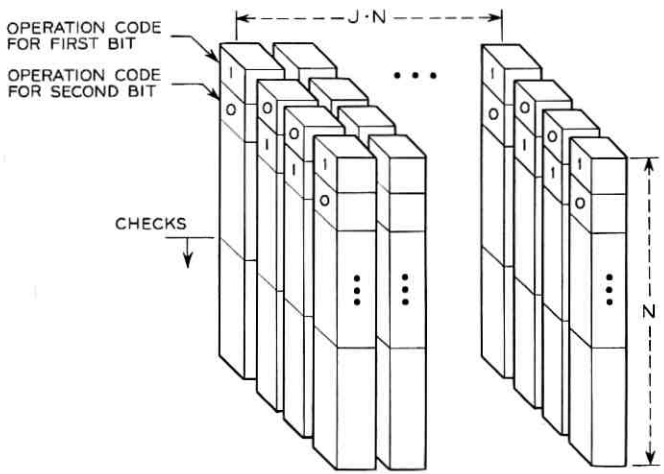


Fig. 3 — The operation code sets.

tions can be represented by $k \times JN$ matrix, each column of this matrix being an approximation to the information digits of the desired result.

Since the errors in each row of each operand matrix and each operation code matrix are statistically independent, and since the errors made by the components which perform the operations are statistically independent, the errors along any row of this resulting matrix must be statistically independent. Therefore, this is precisely the matrix $\mathbf{V}$ described previously. The digits represented by this matrix are ANDed with the digits represented by $\mathfrak{G}$ to form the matrix $\mathfrak{G}'$. The operation of vector addition modulo-2 is then performed on the digits in the registers represented by the rows of the matrix $\mathfrak{G}'$ to obtain the desired result (see Fig. 4).

Finally we must compute the probability that all of the operations are performed in such a way that the state of the result memory belongs to the desired class of states. The first step required that m iterations be performed by each memory in the operand sets and the operation code sets. There are a total of $6 \cdot J \cdot N$ stable memories in these sets. The memories represented by the generator matrix $\mathfrak{G}$ must also perform m iterations. There are k memories in this generator set. It is assumed that neither a memory failure nor a propagation failure occurred in any stable memory before the m iterations were started. We wish to bound the probability that a memory failure or a propagation failure occurs in any memory during these m iterations.

In Ref. 1 the probability that either a memory failure or a propagation failure occurs in any one memory on any particular iteration, given that no memory failure or propagation failure occurred previously, was upper bounded. This bound equals

$$\Pr[\text{memory failure or propagation failure} \mid \text{no previous memory failure or propagation failure}] < \frac{C}{1 - J/K} k^{-\beta+2}$$

where C and β depend on the parameters of the code and the probabilities of component errors but not on k. This bound has the same form for either of the two component malfunction models discussed in the introduction.

Since

$$N \leqq \frac{k}{1 - J/K} \quad \text{and} \quad m < \log \left[\frac{k}{1 - J/K} \right] \quad \text{for} \quad k > 2,$$

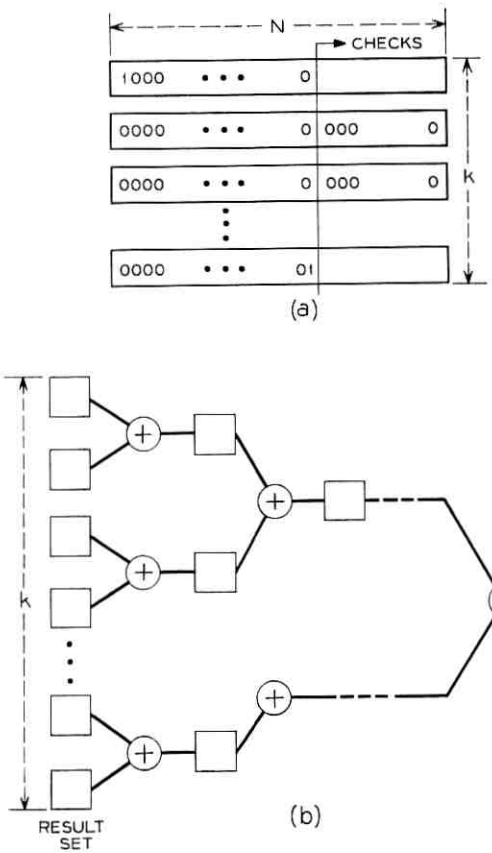


Fig. 4 — (a) The result set corresponding to matrix $\mathcal{G}$ (illustrates the vector AND of the two operands). (b) The addition operation.

the probability that a memory failure or a propagation failure occurs during the m iterations is bounded by

Pr[memory failure or propagation failure during m iterations]

$$< \left\{6\left[\frac{Jk}{1 - J/K}\right] + k\right\} \cdot \left\{\log\left[\frac{k}{1 - J/K}\right]\right\} \cdot \frac{C}{1 - J/K} \cdot k^{-\beta+2}$$

for $k > 2$.

The operation is performed following these m iterations. The result of this operation should be the generators of the desired result, each generator being stored in a stable memory. These memories are called the result set and are represented by the matrix $\mathcal{G}'$. Let us suppose that

no memory failure or propagation failure occurred during the m iterations. We now bound the probability that a digit within a memory in the result set is in error just before the time when the memory has completed the first correcting cycle on the newly inserted digits. If one of these digits is in error, at least one of the following events must have occurred:

(*i*) The corresponding digit was in error in either one of the operand sets. The probability of this event is bounded by $2p_1$ (see Fig. 1).

(*ii*) The corresponding digit was in error in the generator set. The probability of this event is bounded by p_1.

(*iii*) An error was made in determining the operation to be performed. This probability is denoted by p_{control}.

(*iv*) An error was made in performing the desired operation. This probability is denoted by $p_{\text{operation}}$.

(*v*) An error was made in performing the AND operation. This probability is denoted by p_{AND}.

(*vi*) An error occurred within the memory in the result set. The probability of this event is bounded by $2p_r$.

Therefore, by the union bound

$$\Pr[\text{error in any digit within the registers in the result set}] < 3p_1 + p_{\text{control}} + p_{\text{operation}} + p_{\text{AND}} + 2p_r$$

and we require that this sum of probabilities be less than p_0. In order to show that it is possible to satisfy this inequality, let us consider a numerical example where

$$p_{\text{control}} = p_{\text{operation}} = p_{\text{AND}} = p_a = p_r = p_d = 10^{-10},$$
$$p_0 = 10^{-8},\ J = 14 \text{ and } K = 15.$$

For this example $p_1 \approx 2 \times 10^{-10}$, therefore

$$3p_1 + p_{\text{control}} + p_{\text{operation}} + p_{\text{AND}} + 2p_r \approx 1.1 \times 10^{-9} < 10^{-8} = p_0$$

showing that the condition on p_0 is satisfied.

The probability that a memory failure or a propagation failure occurs in any one of the memories in the result set can be computed by exactly the same method used in the previous section to compute the probability of a memory failure in the stable result memory following the operation of vector addition modulo-2. In fact the result is exactly the same as that obtained in the previous section since introducing the bound p_0 makes all the relevant probabilities iden-

tical. The result is as follows

Pr[memory failure or propagation failure in one memory in result set]

$$< \frac{C}{[1 - J/K]^2} \cdot k^{-\beta+3}.$$

Therefore, the probability that any memory failure or propagation failure occurs anywhere in the result set is bounded by

Pr[memory failure or propagation failure in result set]

$$< \frac{C}{[1 - J/K]^2} \cdot k^{-\beta+4}.$$

The final step consists of performing the operation of vector addition modulo-2 on the contents of the memories in the result set. The adder network, shown in Fig. 4, performs $k-1$ vector additions modulo-2. According to the results of the previous section, the probability that any one of these operations is performed incorrectly is bounded by

Pr[error in any of $k - 1 \oplus$ operations]

$$< (k - 1) \frac{C}{[1 - J/K]^2} \cdot k^{-\beta+3}$$

$$< \frac{C}{[1 - J/K]^2} \cdot k^{-\beta+4}.$$

The result is in error only if a memory failure or propagation failure occurs during the first m iteration, during the computation, or during the vector additions modulo-2. Therefore, the probability that the result is in error is bounded by

Pr[result is in error]

$$< 6\left\{\left[\frac{J}{1 - J/K}\right] + 1\right\} \cdot \left\{\frac{C}{1 - J/K} \cdot \log \left[\frac{k}{1 - J/K}\right]\right\} k^{-\beta+3} + \frac{2C}{[1 - J/K]^2} \cdot k^{-\beta+4} \qquad \text{for} \quad k > 2$$

$$\triangleq P_{\text{result}}.$$

Finally, we must bound the probability that an error occurs during a sequence of 3 operations. This probability is bounded by the sum of the probabilities that an error occurs on any one of the 3 operations. Therefore,

Pr[error during a sequence of $\mathfrak{z}$ operations] $< \mathfrak{z} \cdot P_{\text{result}}$.

The dominant term in this bound is proportional to $k^{-\beta+4}$ where β is a constant for any particular sequence of computing systems. Ref. 1 gives a numerical example where β equals 7.55. For this example, the probability that an error occurs in a sequence of $\mathfrak{z}$ operations is bounded by a function which decreases as k^{-3}. Therefore, for the sequence of computing systems for which $\beta = 7.55$, the probability of an error during $\mathfrak{z}$ operations can be made arbitrarily small by making k sufficiently large.

Thus far we have not considered the complexity of these computing systems. The "basic processor," that is, the machine that performs one operation on each of the k pairs of digits has a complexity which is proportional to k^2. Let us suppose that the system S_k is capable of performing a sequence of $\mathfrak{z}$ operations on each of the k digits. In general, this requires $\mathfrak{z}$ basic processors. Thus the complexity of the system is proportional to $\mathfrak{z} \cdot k^2$. The amount of computation, the number of operations performed on pairs of digits, is equal to $\mathfrak{z} \cdot k$. Thus, for S_k, the ratio of the amount of computation to the complexity, is proportional to k^{-1}. The computing capacity equals the maximum value of this ratio for which the probability of error can be made arbitrarily small. Since the probability of error for S_k is proportional to $\mathfrak{z} \cdot k^{-\beta+4}$, this probability of error can be made arbitrarily small only in the limit as k approaches infinity; but in this limit the ratio described above approaches zero. Thus the computing capacity for these systems equals zero.

V. ARITHMETIC OPERATIONS ON OPERANDS OF BOUNDED MAGNITUDE

The systems described in the previous section have two major shortcomings. The first is that the computing capacity equals zero and the second is that the systems are restricted to performing operations on corresponding bits in the operands. Let us consider the first of these shortcomings. We would like to show that it is possible to modify these systems in such a way that the amount of computation per component is independent of k, whereas the probability of error decreases with increasing k. In order to obtain such a result, we must reuse the basic processor. This means that we must be able to "program" the computing system. We have already shown that it is possible to "program" the basic processor to perform different operations. We must now show that it is possible to store the program and the operands in memories which can be located when the contents are needed.

In the previous section we described a method for locating desired

generators which were stored within the generator set. The jth generator was found by first setting up an "address set" consisting of $J \cdot N$ stable memories each of which has a one in the jth information symbol position and zeros in all other information symbol positions. After all stable memories had performed m iterations, the generators were ANDed with the address set and the result was propagated through a modulo-2 addition network as shown in Fig. 4. We can store the program and operands in a "program set" consisting of k memories and use precisely the same method to locate a memory in this set. Thus, the "address" of the jth memory in the program set is a 1 in the jth information position of a code word.

In order to keep track of the next operation to be performed, we need to use one memory as an "instruction counter." Initially this memory contains a code word with a 1 in the first information position and a zero in all other information positions. After each operation, we shift the information digits one position to the right. A simple modification of the basic processor allows it to perform this shift operation. In order to specify whether a shift is desired we add one more operation-code set. The information digits in the memories within this additional set are all 1's if a shift is desired and all 0's otherwise. (Notice that we are really wasting $(k-1)$ of the information digits in each memory in this additional set since only one digit is required to specify whether a shift is desired).

The processor checks the "shift bit" before performing any operation to see if a shift is desired. If the shift bit equals 1, the shift is performed on the first operand and no other operations are performed. If the shift bit is 0, the operations specified by the other operation code sets are performed. In either case, the result is computed as before by adding the appropriate generators.

The second major shortcoming of the systems described in the previous section is that only operations on corresponding bits in the operands can be performed. In order to perform more general operations we might consider permuting the digits in one of the operands before the operation is performed. Unfortunately, the probability that a particular digit is permuted incorrectly depends on k and, in fact, approaches 1/2 as k approaches infinity. This problem arises because we have attempted to perform one operation that involves all k information digits, but it can be avoided by dividing the k digits into a number of segments where the number of digits in each segment does not depend on k; that is, the number of segments must grow with k. We can then treat each segment as a separate operand. This

means that we must restrict our attention to operations performed on operands of bounded magnitude, which is certainly not a severe restriction.

Let us consider the digits in one particular segment. If it were possible to permute these digits before each operation, we could compute the sum or product of the digits in this segment or, in fact, any finite sequence of arithmetic operations on these digits by performing a sequence of bit-by-bit operations on the appropriately permuted digits. Consider the additional modifications that must be made to the basic processor in order to allow it to perform these permutations. For the purpose of specifying the desired permutation we must again increase the number of operation-code sets. If the longest segment contains s digits, the permutation can be described by the contents of the memories in $\log s$ additional operation-code sets. The permutation of a particular segment of digits is specified by the corresponding segments of the memories in these additional operation-code sets. (Notice that a permutation of s digits can be described by $s \cdot \log s$ digits.) The operation is performed according to the method described previously. However, in this case the information bits within each segment of one operand are permuted just before the actual operation is performed.

Let us consider how these modifications affect the bound on the probability of error for the basic processor. This bound is given in Section IV. The first term bounds the probability that either a memory failure or a propagation failure occurs anywhere within the operand sets, the operation code sets, or the generator set during the first m iterations. Since there are now $[\log s] + 1$ additional operation-code sets, the coefficient of this first term must be changed from

$$\left[6\left(\frac{J}{1 - J/K}\right) + 1\right] \quad \text{to} \quad \left[(7 + \log s)\left(\frac{J}{1 - J/K}\right) + 1\right].$$

However, the dependence on k of the first term is unchanged.

The second term bounds the probability that a memory failure occurs either in the result set or in one of the memories required for the modulo-2 addition operation. In deriving this second term, it was necessary to bound the probability of error for digits in the result set. We considered a set of events at least one of which must have occurred if a particular digit is in error. There are now two additional events which could lead to such an error. The first is that the shift instruction was interpreted incorrectly and the second is that the permutation operation was performed incorrectly. In both

cases, the probability of error per digit is independent of k. Therefore, it is possible to find examples where p_0 bounds the probability of error for digits in the result set.

For example, consider the numerical values presented in Section IV. If the probability of a permutation error equals 10^{-10} and the probability of a shift error also equals 10^{-10}, then $p_0 = 10^{-8}$ is still a bound on the probability of error per digit. Thus the second term is not changed by the system modification. Therefore the dominant term in this probability of error is still proportional to $k^{-\beta+4}$. The equipment required to select the appropriate memory from the program set is similar to a basic processor. Therefore, the probability of making an error in one selection operation is of the same form as the probability of error for the basic processor. In particular the dominant term in this probability of a selection error is proportional to $k^{-\beta+4}$.

Finally, let us consider the number of operations that can be performed by this modified computing system. We have allowed k memories for storing the program and the operands. This means that the number of steps in the computation can be at least proportional to k where each step consists of one set of k operations performed by the basic processor. Therefore the amount of computation can be proportional to k^2. Furthermore, the complexity of the system is also proportional to k^2 since only one basic processor and one operation selector is needed. Therefore, the amount of computation per component is independent of k and in general is nonzero.

The probabibity of an error during the k steps in the computation is upper bounded by k times the probability that an error occurs during any one step. Therefore, this overall probability of error is proportional to $k^{-\beta+5}$. Since we have already presented an example where $\beta = 7.55$, in this case the probability of error can be made arbitrarily small by making k sufficiently large while keeping the redundancy fixed. This shows that the computing capacity for systems of this type is nonzero.

VI. CONCLUSIONS

There are basic processors, designed entirely from unreliable components, which can perform arbitrary binary operations on the corresponding information bit of two operands stored in stable memories. These information bits can be shifted, by one bit, or permuted, within segments of bounded length, before performing the operation. The probability of error for the basic processor can be made to vanish

as $k^{-\beta+4}$ where k is the number of information bits in each operand and β is a function of the component malfunction probabilities and the information rate of the code used in storing information in the memories.

A program, including the necessary operands, can be stored in k stable memories and the operations specified by the program can be performed in sequence by a single basic processor. It is possible to perform arbitrary k-step computations on numbers of bounded magnitude. Furthermore, the number of these computations that can be performed simultaneously is proportional to k. The complexity of the equipment required to perform the computations and the amount of computation are both proportional to k^2. Thus the amount of computation per component and hence the computing capacity is nonzero.

These results apply to systems designed from either of two types of unreliable components. The first type is one which malfunctions because it is perturbed by random noise in the system. The malfunctions of these components are modeled mathematically by assuming that they are statistically independent from one component to another and from one use of a particular component to another use. The second type of component is one which fails permanently; however, it is assumed that components which have failed are regularly replaced by good ones. In this case the mathematical model is based on the assumptions that components fail independently of each other and that there is an upper bound, of value less than $\frac{1}{2}$, on the probability of malfunction of any particular component on any single use.

All the results are existence proofs. Therefore, the criterion for choosing the particular systems to be considered was simplicity of analysis rather than efficiency of operation. Furthermore, the emphasis was not placed on obtaining the tightest possible bounds but rather on obtaining simple bounds that were sufficient to prove the desired results. In particular, one would hope that the probability of error would decrease exponentially with k rather than algebraically with k. However, the particular techniques used here do not lead to such a bound.

There are many questions still to be considered concerning the design of practical systems constructed from unreliable components. There are also many theoretical questions concerning the derivation of bounds which are tighter than those derived here. It is hoped that the development of practical, reliable computing systems will follow the presentation of existence theorems like those presented here just

as the development of practical, reliable communication systems followed the presentation of the existence theorems of information theory.

VII. ACKNOWLEDGMENTS

I would like to sincerely thank Professor Robert M. Fano, who supervised this work, for suggesting the approach to the problem and supplying guidance and encouragement throughout the research. I also wish to thank Professor Robert G. Gallager for his help. Many of the results presented here are based on results originally obtained by him. Finally, I wish to thank Professors Peter Elias and Claude E. Shannon who contributed valuable suggestions and constructive criticism.

APPENDIX

A Bound on the Probability of Error

In this appendix we upper bound the probability of error for one vector addition modulo-2. The method for deriving this bound is based on the one used in Ref. 1 to upper bound the probability of a memory failure. Since the method is long and rather involved, it is not reviewed here. Please see the proof in Ref. 1, because here we discuss only those places where the two proofs differ. Both the terminology and the notation here is the same as in Ref. 1.

The computing systems being considered are described in Section III. It is assumed that the operation was performed at $t = T$. The operation takes t_{op} seconds to be performed. Some time during the τ seconds following $t = T + t_{op}$, the stable result memory starts to perform the first correcting cycle on the newly inserted digits. We denote the time at which this correcting cycle starts by $t = T + \sigma$ (see Fig. 5).

It is assumed that neither a memory failure nor a propagation failure has occurred in either operand memory up to $t = T$. A bound has already been derived on the probability of error per digit in the result memory. We must now bound the probability of a propagation failure occurring in the result memory. This involves relating the Δ configurations for the result memory to the Δ configurations for the two operand memories. Finally, the probability of a propagation failure will be used to obtain the desired bound on the probability of making an error in performing the operation.

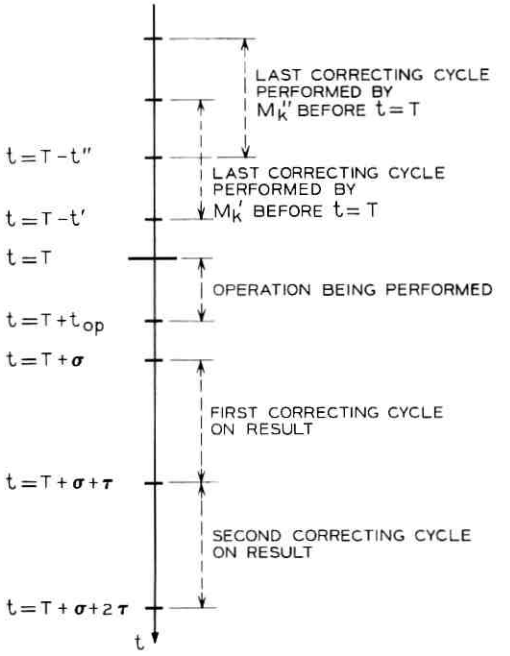


Fig. 5 — Sequence of events pertaining to $\oplus$ operation.

To relate the Δ configurations for the result memory to those for the two operand memories, let us consider the type $i(i > 0)$ Δ configuration for the result memory evaluated at $t = T + \sigma$. If there is a 1 in the entry corresponding to the digit d_0, the controlling errors must have changed the value of d_0 in the result memory; but this change can occur only if the controlling errors changed the value of one but not both digits in position d_0 in M_k' and M_k''. If both these digits in position d_0 had been changed, the two changes would cancel each other when the operation of vector addition modulo-2 was performed. Therefore, the type $i(i > 0)$ Δ configuration for the result memory evaluated at $t = T + \sigma$ is equal to the vector sum modulo-2 of the type i Δ configurations for M_k' and M_k'' where the Δ configuration for M_k' (or M_k'') is evaluated at the end of the last correcting cycle performed before $t = T$ (that is, at $t = T - t'$ or $t = T - t''$).

The type 0 Δ configuration evaluated at $t = T + \sigma$ is computed in a different way. A 1 in this type 0 Δ configuration represents a new error in the stable result memory at $t = T + \sigma$. To compute the configuration of new errors, we imagine that the computing system was

noiseless up to the beginning of the most recent correcting cycle performed on the digits stored in the result memory at $t = T + \sigma$. In this case, the "most recent correcting cycle" consists of the last correcting cycle performed by M_k' before $t = T$ and the last correcting cycle performed by M_k'' before $t = T$ (see Fig. 5).

Since the probability of error per digit at $t = T + \sigma$ can be bounded by p_0, the probability of a new error at $t = T + \sigma$ can also be bounded by p_0. Therefore the probability of a 1 in the type 0 Δ configuration evaluated at $t = T + \sigma$ must be less than p_0. Since the type $i (i > 0)$ Δ configuration for the result memory evaluated at $t = T + \sigma$ is the modulo-2 sum of the type i Δ configurations for M_k' and M_k'', the probability of a 1 in the type i Δ configuration evaluated at $t = T + \sigma$ is bounded by twice the probability of a 1 in the type i Δ configuration for M_k' (or M_k'') evaluated at $t = T - t'$ (or $t = T - t''$).

The method for computing a bound on the probability of a memory failure at $t = T + \sigma$ is exactly the same as that used to compute a bound on the probability of a memory failure at any other time for any stable memory. This method consists of bounding the probability that a noiseless correcting network could correct all errors present at $t = T + \sigma$ by performing m iterations. This probability is bounded by the probability that the initial propagation failure occurs at $t = T + \sigma$ or during the m noiseless iterations performed after $t = T + \sigma$.

The initial propagation failure occurs whenever there are one or more 1's in the type m Δ configuration. Since, by assumption, no propagation failure occurred in either M_k' or M_k'' before $t = T$, the type m Δ configuration for M_k' and M_k'' must be all zero for all $t < T$. The type m Δ configuration for the result memory evaluated at $t = T + \sigma$ is the vector sum modulo-2 of two of these type m Δ configurations for M_k' and M_k''. Therefore, the type m Δ configuration evaluated at $t = T + \sigma$ must be all zero. Hence, no propagation failure can occur at $t = T + \sigma$.

A bound is derived in Ref. 1 on the probability of a 1 in the entry corresponding to the digit d_0 in the type m Δ configuration evaluated at $t = L\tau$. The equation used to compute this bound is;

$$\begin{aligned}\Pr[&1 \text{ in one particular entry in type } m \ \Delta \text{ configuration} \\ &\qquad \text{evaluated after } L\text{th iteration}] \\ &< (J-1)(K-1)\binom{J-2}{J/2-1}[(K-1)(2p_0)]^{J/2-1} \\ &\quad \cdot \Pr[1 \text{ in one particular entry in type } m-1 \ \Delta \text{ configuration} \\ &\qquad \text{evaluated after } (L-1)\text{th iteration}] \quad (1)\end{aligned}$$

where p_0 is both a bound on the probability of error for digits in the registers of M_k and a bound on the probability of a 1 in the type 0 Δ configuration. To compute a bound on the probability that the initial propagation failure occurs at $t = T + \sigma + \tau$, we apply this recurrence relation. By applying it $m - 1$ times, we obtain the probability of a 1 in the type $(m - 1)$ Δ configuration for M_k' evaluated at $t = T - t'$, namely

$$\Pr[1 \text{ in one particular entry in type } (m-1)\ \Delta \text{ configuration evaluated at } t = T - t']$$

$$< p_0\left\{(J-1)(K-1)\binom{J-2}{J/2-1}[(K-1)(2p_0)]^{J/2-1}\right\}^{m-1}$$

$$\triangleq P_{m-1}\ .$$

The probability of a 1 in the type $(m - 1)$ Δ configuration evaluated at $t = T + \sigma$ is bounded by $2 \cdot P_{m-1}$. One final application of the recurrence relation leads to

$$\Pr[1 \text{ in one particular entry in type } m\ \Delta \text{ configuration evaluated at } t = T + \sigma + \tau]$$

$$< 2p_0\left\{(J-1)(K-1)\binom{J-2}{J/2-1}[(K-1)(2p_0)]^{J/2-1}\right\}^{m}.$$

Gallager has shown that there are low-density parity-check codes[2] for which

$$\frac{\log\left[\frac{k}{2K} - \frac{k}{2J(K-1)}\right]}{2\log(J-1)(K-1)} < m < \frac{\log\left[\frac{k}{1-J/K}\right]}{\log(J-1)(K-1)}.$$

Therefore,

$$\Pr[1 \text{ in one particular entry in type } m\ \Delta \text{ configuration evaluated at } t = T + \sigma + \tau]$$

$$< 2p_0\left\{(J-1)(K-1)\binom{J-2}{J/2-1}\cdot[(K-1)(2p_0)]^{J/2-1}\right\}^{\log\left[(k/2K)-[k/2J(K-1)]\right]/2\log(J-1)(K-1)}$$

$$= 2p_0\left\{\left[\frac{1}{2K} - \frac{1}{2J(K-1)}\right]k\right\}^{-\beta}$$

where

$$\beta \triangleq -\frac{\log\left\{(J-1)(K-1)\binom{J-2}{J/2-1}[(K-1)(2p_0)]^{J/2-1}\right\}}{2\log(J-1)(K-1)}.$$

The probability that the initial propagation failure occurs at $t = T + \sigma + \tau$ is bounded by the probability of one or more 1's in the type m Δ configuration evaluated at $t = T + \sigma + \tau$. Since there are $J \cdot N$ entries in this Δ configuration, and since $N \leq k/(1 - J/K)$, by the union bound,

$$\begin{aligned}\Pr[\text{initial propagation failure occurs at } t = T + \sigma + \tau] &< J \cdot \left[\frac{k}{1 - J/K}\right] \cdot 2p_0 \left\{\left[\frac{1}{2K} - \frac{1}{2J(K-1)}\right]k\right\}^{-\beta} \\ &\triangleq 2Ck^{-\beta+1}.\end{aligned}$$

This same bound applies to the probability that the initial propagation failure occurs on any one of the m iterations performed after $t = T + \sigma$.

A memory failure can occur at $t = T + \sigma$ only if the initial propagation failure occurs on one of the m iterations performed after $t = T + \sigma$. Therefore, by the union bound,

$$\begin{aligned}&\Pr[\text{memory failure at } t = T + \sigma \mid \text{no memory failure or propagation failure for } t < T] \\ &= \Pr[\oplus \text{ operation is performed incorrectly}] \\ &< 2 \cdot C \cdot m \cdot k^{-\beta+1} \\ &< 2 \cdot C \cdot \log\left[\frac{k}{1 - J/K}\right] \cdot k^{-\beta+1} \\ &< 2 \cdot \frac{C}{1 - J/K} \cdot k^{-\beta+2}.\end{aligned}$$

To show that the computing systems being considered are reliable, we must show that it is possible to perform a sequence of 3 operations with an overall probability of error that can be made arbitrarily small by choosing k sufficiently large. After one iteration of the stable result memory (that is for $t \geqq T + \sigma + \tau$), the result stored in this memory can be used as an operand for another vector addition modulo-2. Let us suppose that another addition operation is performed at $t = T + \sigma + \tau \triangleq T'$ and that the first correcting cycle on this second result starts at

$t = T' + \sigma'$. We now bound the probability that this second operation is performed incorrectly given that the first operation was performed correctly (that is, given that no memory failure or propagation failure occurred at $t = T + \sigma$). The method for deriving this bound is identical to the one used above. We bound the probability that the initial propagation failure occurs at $t = T' + \sigma'$ or on any one of the next m iterations. In this case, the probability that the initial propagation failure occurs at $t = T' + \sigma'$ is not zero since a propagation failure could have occurred at $t = T'$.

Equation 1 relates the probability of a 1 in the type i Δ configuration evaluated after one particular correcting cycle, to the probability of a 1 in the type $(i - 1)$ Δ configuration evaluated after the previous correcting cycle. The application of this equation is simple except in cases where an addition operation has been performed between successive correcting cycles. It has been shown already that in each case where an addition operation has been performed, we must double the probabilities that otherwise would have been substituted into the equation. Since, in this case, there were two addition operations performed within the m correcting cycles of interest, the value of this bound must be twice the value of the bound derived above.

Therefore,

$$\begin{aligned}\Pr[\text{second } \oplus \text{ operation is performed incorrectly} \mid \text{first } \oplus \\ \text{operation was performed correctly}] \\ < 2^2 \cdot [m + 1] \cdot C \cdot k^{-\beta+1} \\ < 2^2 \cdot C \cdot \log \left[\frac{k}{1 - J/K}\right] \cdot k^{-\beta+1} \\ < 2^2 \cdot \frac{C}{1 - J/K} \cdot k^{-\beta+2}\end{aligned}$$

where we have used the bound

$$m + 1 < \frac{\log \left[\frac{k}{1 - J/K}\right]}{\log (J - 1)(K - 1)} + 1 < \log \left[\frac{k}{1 - J/K}\right] \quad \text{for } k > 2.$$

This result can be extended to any number of vector additions modulo-2 performed in series. However, since only m iterations are considered in deriving this bound, the largest value that this bound can have corresponds to the case where an addition operation has

been performed between each iteration for the last m iterations. For this "worst case" the bound on the probability that the mth vector addition modulo-2 is performed incorrectly, given that all previous additions were performed correctly, is

$$\begin{aligned}\Pr[&\text{any } \oplus \text{ operation is performed incorrectly} \mid \text{all previous} \\ &\oplus \text{ operations were performed correctly}] \\ &< 2^m \cdot C \cdot \log\left[\frac{k}{1 - J/K}\right] \cdot k^{-\beta+1} \\ &< \frac{C}{1 - J/K} \cdot \log\left[\frac{k}{1 - J/K}\right] \cdot k^{-\beta+2} \\ &< \frac{C}{[1 - J/K]^2} \cdot k^{-\beta+3}.\end{aligned}$$

In this Appendix we have not made any reference to the underlying assumptions concerning the models for component malfunctions. We are interested in two models as explained in the introduction. Since equation 1 applies for either of these models and since the mathematical development based on equation 1 also applies for either model, the final result, namely the probability of the initial modulo-2 operation error, has the form given above for either component malfunction model.

REFERENCES

1. Taylor, M. G., "Reliable Information Storage in Memories Designed from Unreliable Components," B.S.T.J., *47*, No. 10 (December 1968), pp. 2299–2337.
2. Gallager, R. G., *Low-Density Parity-Check Codes*, Cambridge, Massachusetts: MIT Press, 1963.
3. Peterson, W. W., *Error Correcting Codes*, New York: The MIT Press and John Wiley and Sons, Inc., 1961.

A Self-Healing Control

By BRUCE E. BRILEY

(Manuscript received April 30, 1968)

This article describes an electronic computer self-repair technique which does not require system duplication. It identifies the properties demanded of control hardware to effect this end and considers their practicability. It also decribes a prototype computer constructed to test the feasibility of this form of self-repair, and discusses test results.

I. INTRODUCTION

In electronic computing machines, where a high degree of availability is a requisite, some form of self-repair is provided. The rational in most such machines is that a given piece of hardware must be backed up by an identical part which may be switched in upon failure of the first.

Work in nonreplicative self-repair has chiefly centered about the arithmetic unit and other controlled entities where the solutions, though not trivial, are relatively straightforward. The control portion, however, has consistently been avoided as unmanageable short of replication.[1] This paper describes a technique for what will be called "self-healing" the control of a machine which differs radically from conventional self-repair, and displays potential for considerable economic savings.

II. DEFINITIONS

Terms used in describing the technique and their definitions are:

Order denotes that portion of an instruction conventionally used to identify the instruction type, not including the address and other fields of the instruction, even when used for augmentation rather than addressing.

An order structure is called *closed* (closure under imitation) if to each order there corresponds at least one program of finite length using only a subset of the order structure diminished by the object

order such that the program imitates the salient actions of the object order in every detail except timing.

A control is called *failure autonomous* if the circuitry associated with the sequencing of an order is unique to it and does not propagate the effects of an internal failure beyond the associated circuitry's geographic bounds.

A control unit is called *self-diagnosing* if, under failure, (*i*) the control immediately ceases activity, (*ii*) the identity of the offending order is immediately available, and (*iii*) the control is failure autonomous.

A control unit is called *entropic* if the circuitry associated with each order assumes a state under failure such that any subsequent attempt to execute that order will cause a summary hang-up without any of the order's control signals having become active.

Self-diagnosing means immediate "incidental" diagnosis as a basic property, as opposed to "self-diagnosable," which means lending itself to easy but finite diagnostic processing, and contains the former term.

The fact that all order structures are not closed may not be apparent, but examination of the repertoire of a typical computer will bring to light orders which violate the definition.

Proof of the existence of a nonclosed order structure capable of all the conventional operations requires only observation of the effect of a failure upon Van der Poel's limiting case machine.[2] (Van der Poel shows that a computer with only a specialized subtract instruction can perform all essential operations.) This machine is left with the empty set upon diminishing its repertoire by one. However, existence in the repertoire of any order which performs a unique function such as setting a flip-flop accessible to no other order is sufficient to render the structure, of which it is a member, outside the closure definition.

A seemingly trivial example of a closed order structure is the case of a repertoire with each instruction duplicated; the imitating routine for a given instruction, in this case, is of length one. This extreme appears to be, at best, equivalent to conventional duplication approaches, and certainly worse than any other self-healing case. However, H. Y. Chang has pointed out that no conventional checking circuits would be necessary in the control and imitation would be performed with no reduction in efficiency. This form of duplication, then, may be viewed as the boundary between self-healing and conventional self-repair, with some of the advantages of both.

III. BACKGROUND

The definitions are chosen advisedly because a control which is self-repairing in the sense described must be the realization of an order structure which is closed, and must be entropic and self-diagnosing (which implies failure autonomy).

The practical possibility of building a self-healing control came to light as the result of study of a (picoprogramming) control with self-diagnosing properties.*

Self-diagnosing is a required attribute because (*i*) the identity of the offending order must be immediately available without processing (the control cannot be trusted to perform diagnostic processing), (*ii*) control activity must cease until remedial action is taken or insane processing may be performed, and (*iii*) failure autonomy must be realized or several orders may be rendered imperfect by a single failure.

Entropic behavior is required so that (*i*) the repairing entity may be called automatically into play each time the defunct order is to be executed and (*ii*) the defunct order's circuitry will remain inactive and not interfere with the repairing entity's action.

The order structure realized must be closed to permit the technique to function; this requirement is central to the technique's operation.

IV. PRACTICABILITY

These requirements may seem artificial; indeed, unattainable. However, the self-diagnosing control has the very properties demanded above, so that the only otherwise artificial requirement to meet is that of order-structure closure.

We conjecture that the addition of about 10 percent to the repertoire of a typical machine would effect closure. Recalcitrant instructions, the length of whose imitating routines threaten to increase without bound, can be handled by the compromise of duplicating those instructions only. Their imitating routines then shrink to one instruction in length. (A machine designed specifically to be self-healing would, of course, try to avoid such instructions.)

V. THE TECHNIQUE

The technique is relatively simple. An inductively coupled detection lead monitoring all order circuits observes a continuous sequence of pulses during normal operation. However, because of the second

* See the Appendix and Refs. 3 through 6.

property of the self-diagnosing feature, when the circuitry associated with an order fails, the monitoring lead will notice a cessation of activity. A time-out will take place, and the machine will be placed in the remedial mode.

In the remedial mode, the entire (failed) instruction is stored, then control is transferred to a location in memory whose address is generated from the order field of the instruction, and normal operation is resumed. The transferred-to location in memory contains the first word of a routine which imitates the action of the failed order using only other orders. This is possible because of the closed order structure.

The transition from imitation to a continuation of the program being run is smooth, requiring no intervention. The next time execution of the failed instruction is called for the machine hangs up again, without any control pulses being generated, because of the control's entropic character. The same procedure as before is then followed.

A simple example is logical left shift (shift the contents of the accumulator logically one binary place to the left), because its imitating routine is brief. Figure 1 shows the effective mechanization of self-healing for the shift left instruction. The action of sensing the health

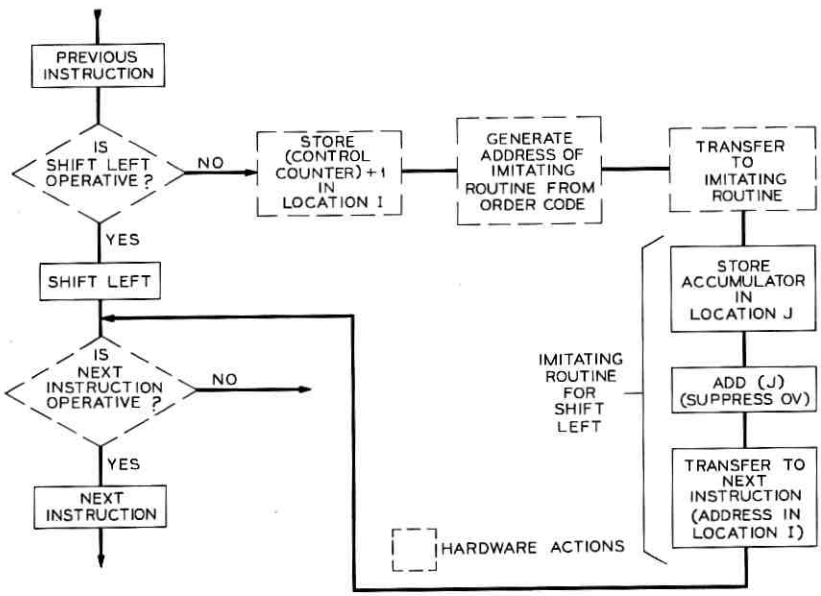


Fig. 1 — Example of self-healing.

of an instruction is indicated by the test preceding each instruction; actually, the test takes place throughout the duration of an instruction's execution, and is much less specific than the illustration implies, asking only the general question, "Is the machine still running?" rather than, "Is shift left operative?"

The net result is that the machine is, in effect, healed; software is used as a patch over the "crack" in the hardware. Execution of the failed instruction will be, in general, very inefficient, but if the time-out is short, the overall speed of the machine generally will be only slightly affected.

If fast memory is considered too expensive or limited in size to contain all of the imitation routines, they may reside in auxiliary storage, and the one selected under failure would be called into a standard block in fast memory. Subsequent uses of the defunct order would execute its imitation routine out of this block. Such "calling" would require more elaborate remedial hardware.

VI. REALIZATION

The picopragrammed control lends itself to self-healing because of certain properties peculiar to its implementation:

(*i*) The circuitry implementing each order is segregated to a single card and is unique to it; magnetic coupling to the control leads prevents propagation of failure effects.

(*ii*) Because the control is autochronous[3] (self-timed), a failure in a sequencer causes an asynchronous hang-up, leaving the order register in one-to-one correspondence with the failed sequencer.

(*iii*) The nature of the ferrite disk[7] sequencer is such that any failure in the active sequencing mode will leave the disk in a partially switched condition (in a state representing one element of a continuum of states) which prevents the success of any subsequent attempts at switching, and results in a hang-up. No control signals are generated by attempts at switching subsequent to failure.

To a "conventional" picoprogrammed control must be added intelligence to perform remedial action. The remedial tasks illustrated in Fig. 1 are:

(*i*) Store in an appropriate location the nonorder fields of the instruction (which usually change each time the order is executed).

(*ii*) Store in an appropriate location the current address (or the current address +1).

(*iii*) Derive from the order field the location of its imitating routine.
(*iv*) Transfer to the imitating routine.

The imitating routine alters its instructions according to the non-order fields of the defunct instruction, performs the imitation, and transfers to the defunct instruction's successor in the interrupted program.

VII. PROTOTYPE IMPLEMENTATION

A small prototype computer was built at Bell Laboratories to test the feasibility of self-healing (see Fig. 2). Picoprogramming was used, providing the necessary properties for self-healing.

7.1 *Duplication*

The special (limiting) case of duplication is the simplest to implement, and was the first tried. In effect, dual repertoires are installed whose order codes differ by one bit (are adjacent). Any number of techniques could be used in a full scale system to guarantee that both halves of the extended repertoire are exercised regularly, such as assembler insertion or internal bit complementing. The equivalent of the second technique was used in the prototype because

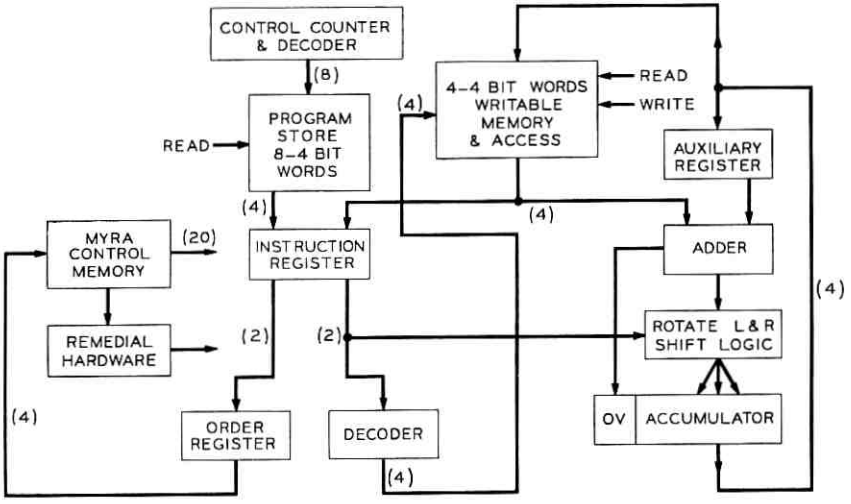


Fig. 2 — Prototype self-healing processor.

it does not require an additional bit per order in main memory; the internal order code was augmented with a bit which served to choose between twin instructions.

The mechanism used to recognize failure (cessation of activity) was a simple integrating circuit, which monitored the completion signal common to all instructions, and a level discriminator.

The remedial action consisted simply of complementing the "twin" bit and injecting a bogus completion signal to start the system again.

Physically simulated failures were experimentally shown to yield to self-healing independent of the instruction class, and the implementation of this grade of self-healing was definitely proved possible.

7.2 *Nonduplication*

A test off the boundary line was necessary to show that a nonduplicating grade of self-healing is possible. The limited size of the prototype's repertoire, however, precluded a complete closed repertoire from being tested. Instead the properties of self-healing were studied "in the small"; repertoires were chosen which could imitate a subset of themselves, but which did not include duplicates.

The procedure followed was to place the machine in a convenient program loop performing calculations easily checked for accuracy, such as a first order Markov chain. In another memory area was stored a routine that could imitate the action of an order used in the calculating routine.

A manual switch was provided to either disable or invoke self-healing capability, and another to disable the object order.

With the machine running in the calculation loop, and with self-healing disabled, it was easily shown that disabling (simulating a failure) the object order would bring the machine to a halt, leaving on display its location, its identity, and its nonorder fields.

Invoking self-healing capability, and running in the calculation routine, disabling the object order had no effect upon the calculations' accuracy in most cases; however, occasionally an error would occur. The errors (which will be explained) occurred only at the instant of simulated failure; calculations after that (while self-healing) were properly performed. When the simulated failure was removed, the machine automatically reverted to its original mode of operation.

An order is always assumed innocent until proven guilty. That is, whether or not an order failed the last time it was to be used, an attempt is made to use it the next time the program calls for it. The

innate properties of a failed order thus provide the necessary memory that the order is inoperative, and prevent these attempts from causing trouble. Should the order spring back to life again, however, either because the "failure" was caused by noise or because the failed card is replaced, the order "takes over" again, and self-healing ceases.

This technique does not solve all problems. One difficulty is that a fragment of an instruction may be executed before failure occurs (this was the cause of the occasionally observed errors at the instant of failure), and it may be impossible to reconstruct the status of the system before execution of this instruction began. Two solutions are:

(*i*) Defensive programming, a technique which is recommended for all real time programming applications, but difficult to implement as a complete solution. It consists of programming in such a fashion that errors (the outward manifestation of a partial instruction execution) do not appreciably disturb operation.

(*ii*) Picoinstruction counting, which forces the picoinstructions to count themselves as they are executed (the counting register is reset upon the successful completion of each instruction). Under failure the contents of this register are digested by the imitating routine, which (the first time) performs only those operations not yet performed by the defunct instruction.

VIII. HARDCORE

8.1 *Control Hardcore*

The hardcore (that portion whose failures cannot be healed) specifically associated with a self-healing control consists of a fixed and a variable component. The fixed component includes the instruction register, an order register which copies the order field of the instruction register, and their associated circuitry; in addition, a relatively simple integrating and level sensing circuit to recognize and react to failure, and a flip-flop and timing circuit to store the control mode and provide interinstruction timing, are required.

The variable component depends upon the degree of duplication. For full duplication a flip-flop is needed to differentiate between twin instructions. Zero duplication requires a means for storing the nonorder field of the failed instruction and its address for use by the imitating routine, and a means for generating the address of the first instruction of the imitating routine and placing it in the instruction counter register (assuming a single address format).

8.2 *Noncontrol Hardcore*

The noncontrol portions of the machine are not directly aided by self-healing, but the indirect help is far reaching. For automatic diagnosis, the hardcore of the machine does not include the entire control, but rather the control's relatively small hardcore. This means that much more of the machine may be automatically diagnosed. For extension beyond diagnosis to self-repair (and possibly self-healing) in the noncontrol portions of the machine, the more effective diagnosis and more trustworthy control are substantial aids.

Among the more interesting possibilities is that of building an all memory machine (including the control) which would use tables in a manner reminiscent of the IBM 1620 to replace the arithmetic unit and reduce the diagnostic problem to one of handling memory (which, because of its relatively homogeneous nature, tends to be tractable).

IX. ANALOGIZING VIEW

Suppose it were feasible to store a set of remedial routines for each order, which could imitate the order's action for all possible combinations of order failures. Then as the machine grew old, and one by one the orders failed, the control would survive but become increasingly slower until a minimum critical subset (possibly one order) remained. Under these conditions the control would be acting in a manner analogous to the compilation of a high-level language, written in itself. It is necessary in the latter case (the compilation) to (software) implement a critical subset of the language which can bootstrap the remainder of the language, just as it is necessary to have a healthy (hardware) implementation of a critical subset of the instructions in the former case (the aged machine). Extending the analogy further, a correspondence may be seen between the computer's microinstruction language (if it exists) for the former case, and the machine language of the computer used in the latter case.

X. CONCLUSIONS

Self-healing is a practicable and potentially useful variation of self-repair.

(Comment: Although picoprogramming was used in this work, it was a vehicle of convenience, not an essential constituent. Other implementations are possible.)

XI. ACKNOWLEDGMENTS

I wish to acknowledge helpful discussions with H. Y. Chang, and the excellent experimental assistance of D. J. Matter.

APPENDIX

Picoprogramming

This is a very brief description of picoprogramming; the serious student should see Ref. 3.

Picoprogramming may be viewed as a logical extension of microprogramming wherein the control memory, which in a practical microprogrammed machine constitutes some fraction of the control, is allowed to expand until it is identical with the control. Figures 3 through 5 depict the evolution in oversimplified form.

Figure 3 illustrates the typical pre-third-generation machine control, with the order portion of the instruction used to identify a sequence implemented, in general, with ill partitioned circuitry denoted by crosshatching.

Figure 4 shows the microprogrammed control where the order field identifies the starting address of a microprogram in the control memory, but a conglomerate of circuitry is still required to implement the microinstructions, count through them, and the like.

Figure 5 shows a picoprogrammed control consisting only of a memory with the order field now choosing a single memory element.

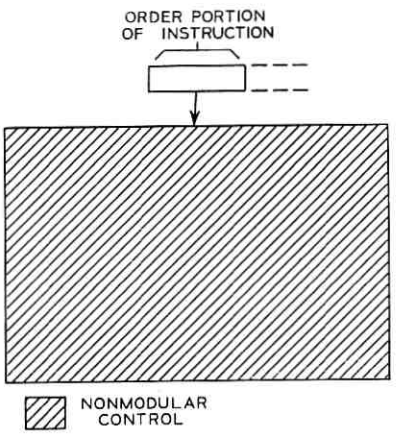


Fig. 3 — Conventional control.

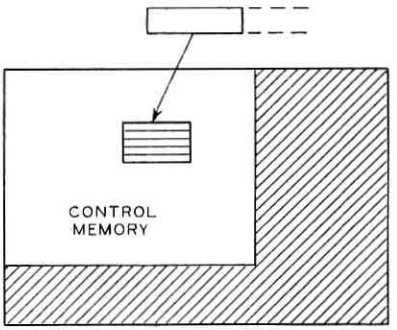


Fig. 4 — Microprogrammed control.

The memory element in Fig. 5 is a square loop ferrite disk called a myra (for myria-apertured) disk. This disk has the property that, when selected, it can spill out many sequential strings of control pulses temporally juxtaposed in almost any desired fashion and at voltage and driving point impedance levels capable of driving most logic circuits directly. Most instructions can thus be implemented with a single disk.

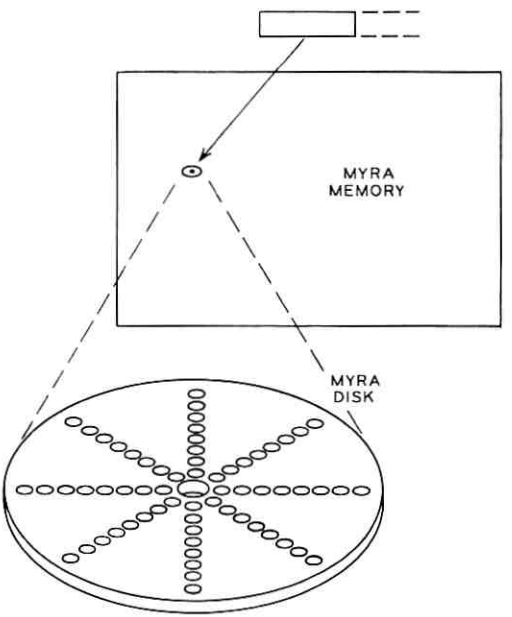


Fig. 5 — Picoprogrammed control.

When these disks are mounted separately with their drivers, they exhibit failure autonomy. Further, since no clock is used, the disk corresponding to an instruction under execution is in complete and independent control, and the machine stops if it fails. In fact, an unequipped card can be used to implement a HALT instruction.

The instruction cards can, in general, be interchanged to provide a variable repertoire.

REFERENCES

1. Avizienis, A., "Design of Fault Tolerant Computers," Proc. Fall Joint Computer Conf., *31* (November 14–16, 1967, Anaheim, Calif.) pp. 733–743.
2. Van der Poel, W. L., "The Essential Types of Operations in an Automatic Computer," Nachrichtentechnische Fachberichte, *4* (1956), pp. 144–145.
3. Briley, B. E., "Picoprogramming: A New Approach to Internal Computer Control," Proc. Fall Joint Computer Conf., *27* (Las Vegas, Nev., November 30–December 2, 1965), pp. 93–98.
4. Valassis, J. G., Macrander, M. C., Pacer, T. A., and Rekiere, R. J., "An Integrated-Circuit MYRA Picoprogrammed Computer," Automatic Electrical Technical Journal, *10*, No. 8 (October 1967) pp. 326–336.
5. Valassis, J. G., Mehta, M. A., and Holden, J. R., "Analysis of the MYRA Picoprogramming Control Technique," Automatic Electric Technical Journal, *10*, No. 8 (October 1967), pp. 327–348.
6. Valassis, J. G., "Modular Computer Design with Picoprogrammed Control," Proc. Fall Joint Computer Conf., *31* (November 14–16, 1967, Anaheim, Calif.) pp. 611–619.
7. Briley, B. E., "MYRA: A New Memory Element and System," IEEE Intermag. Conf. Proc., Washington, D. C. (April 21–23, 1965), pp. 14.8–1—14.8–6.

Efficient Spacing of Synchronous Communication Satellites

By HARRISON E. ROWE and ARNO A. PENZIAS

(Manuscript received April 10, 1968)

Satellites in equatorial circular synchronous orbit remain fixed with respect to an observer on the earth, while those in a circular synchronous orbit inclined with respect to the equator appear to move in a figure 8. A number of satellites can move on a given 8. Optimum packing for a single 8 is determined; the number of satellites per 8 increases as the 8 becomes larger and as the allowable closest approach between satellites decreases. Several packing schemes using multiple 8's, regularly spaced, are presented. For potential systems using frequencies above 12 GHz, with perhaps 1° closest approach between satellites, the best scheme described here permits approximately six times as many satellites as an equatorial system. An illustrative system serving North America has space for a total of 477 satellites, compared with 95 satellites if only equatorial orbits are allowed.

I. INTRODUCTION

When a satellite is in an orbit whose period is equal to that of the earth's rotation, it is said to be synchronous. A satellite in a synchronous, circular, equatorial (in the plane of the equator) orbit will appear stationary when viewed from the earth. The number of such satellites that can be used in a communication system in the same frequency band is limited by the directivity of the ground-based antennas.[1] The maximum number would be obtained by spacing them equally, at an angular separation just sufficient to keep crosstalk down to a tolerable level.

If the synchronous orbit is inclined to the equator, a satellite will no longer appear stationary when viewed from the earth, but will appear to move in a figure 8. By placing several satellites in a single 8, and spacing different 8's regularly along the equator, a greater number of synchronous communication satellites can be used than with an equatorial orbit alone.

A price must be paid for increasing the number of synchronous satellites by using inclined circular orbits:

(*i*) The earth-based antennas must track the satellites.

(*ii*) The satellite-based antennas must track the earth stations.

(*iii*) As the angle of inclination of the orbits is increased (in order to increase the number of satellites), the range of latitudes on earth within which the satellites will always be visible decreases.

In this paper we study optimum ways of packing satellites on inclined circular synchronous orbits, and determine the maximum number of such satellites that can be used under various conditions, assuming:

(*i*) Strictly circular, synchronous orbits, with consequent uniform satellite velocity.

(*ii*) Each satellite is usable at all times; that is, each satellite remains in view of ground stations located as far north as the Canadian-United States border, over a strip of substantial east-west width.

(*iii*) If the angular separation between any two satellites exceeds a certain minimum value, the crosstalk requirements will be satisfied. This minimum value is a parameter in the analysis.

(*iv*) The angular separation between satellites is computed as seen from the center of the earth.

These assumptions have the following corresponding consequences for this analysis:

(*i*) The problem is a kinematic or geometric one, rather than one in mechanics as if elliptical orbits, the oblateness of the earth, the effects of the moon, and so on, were considered. (Notice that we do not consider synchronous elliptical orbits.) This assumption implies that the satellites have sufficient on-board fuel to correct their orbits for various small perturbations (such as those mentioned above) over the expected life of the satellites, in addition to any fuel needed to track the earth stations if this is done by mechanically rotating the satellites (rather than by electronic beam-steering of the satellite antennas).

(*ii*) Assumption *ii* guarantees that there need be no switching from one satellite to another in such a communication system; this establishes a convenient bound on the present investigation. Once switching from one satellite to another is allowed, the system complexity appears to have no natural bound; one could theoretically fill the sky with satellites, and find a convenient one somewhere.

(*iii*) The minimum separation assumption avoids a detailed analysis of crosstalk. Such an analysis would have to include, for example, the type of modulation (AM, FM, PCM, and so on), the (angular) distances to all the satellites in the neighborhood of the one under study, the location of the ground stations, ground- and satellite- based antenna patterns, and the like.

(*iv*) By computing separation relative to the center of the earth, we neglect the small corrections which depend on the location of the ground stations; these corrections in satellite separation can vary between zero and as much as +18 percent in the worst case, as shown in Appendix C (see also Fig. 1 or 2, discussed below). However, assumption *iv* leads to simple, exact expressions for almost all of the interesting results.

Figures 1 and 2 show to scale the earth and a synchronous satellite, at declinations of 25° and 30° south latitude. The latitudes of repre-

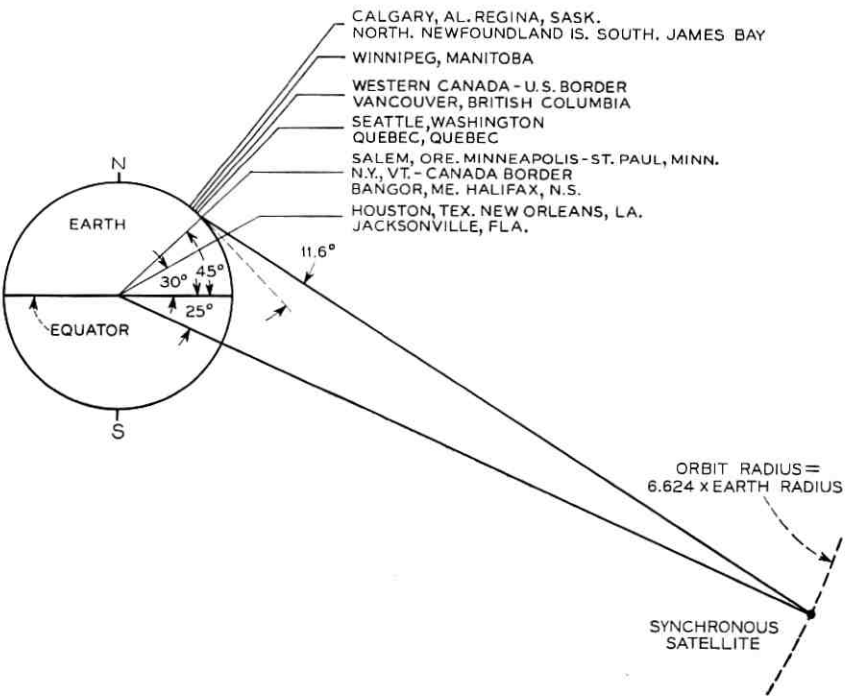


Fig. 1 — Synchronous satellite on an inclined orbit; southernmost declination = inclination = 25°.

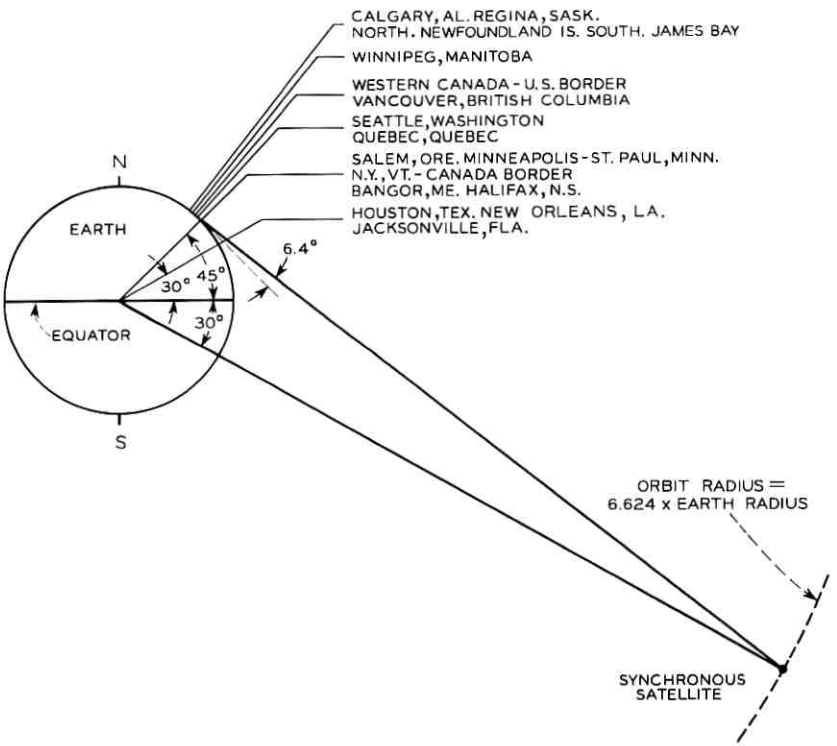


Fig. 2 — Synchronous satellite on an inclined orbit; southernmost declination = inclination = 30°.

sentative points in the United States and Canada are shown. Much of the United States is south of 45° north latitude; a ground station at this latitude will observe a synchronous satellite at the same longitude (as the ground station) at an elevation of 11.6° (Fig. 1) or 6.4° (Fig. 2) above the horizon. The smaller the satellite elevation the greater the path length lying within the atmosphere, with attendant greater rain attenuation, wavefront distortion, and other atmospheric effects.

Section II indicates that the extreme value of latitude (north and south) attained by a synchronous satellite is equal to its orbit's angle of inclination from the equator. Figures 1 and 2 suggest that the maximum angle of inclination of interest for the United States is about 30°. Such large angles will probably rule out major parts of Canada. However, in the systems described below, 8's containing several satel-

lites each will be regularly spaced along the equator, separated by a smaller number of satellites in equatorial orbit equally spaced between the 8's. These equatorial satellites could probably be used to serve the lighter traffic to the more northern regions of North America or the more southern regions of South America, while both the figure 8 satellites and the equatorial satellites serve the heavy traffic within the continental United States.

II. GEOMETRY OF INCLINED, SYNCHRONOUS ORBITS

Let us compute the angular separation between satellites as seen from the center of the earth (assumption *iv*, Section I). This permits the calculation of all desired quantities by the straightforward use of the results of spherical trigonometry. By convention in spherical trigonometry, great circle distances on a sphere are measured by the angle they subtend at the center of the sphere; we use this convention throughout this article without further comment. By convention, with one principal exception, all lines drawn on the surface of a sphere in any figure of this article are great circles; the exception is the 8 described in the Introduction. Thus, longitude lines are allowed, while latitude lines are not.

Figure 3 shows the earth with north and south poles, the equator, and the projection on the earth of a (circular) synchronous orbit, inclined at an angle α to the equatorial plane, passing through the

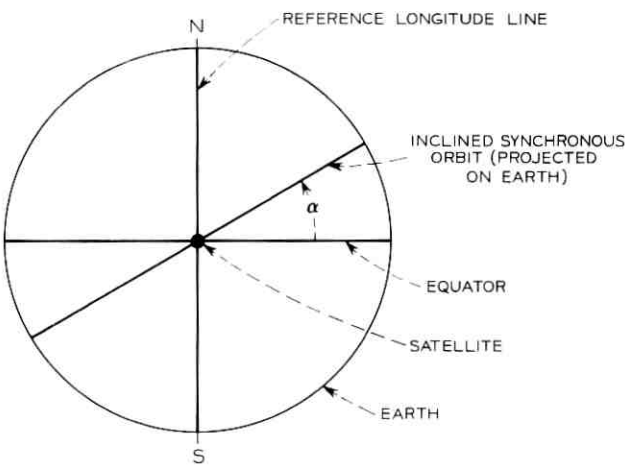


Fig. 3 — Reference point for satellite. $c = 0°$.

equator at its midpoint as seen by the observer. Since the orbit is viewed edge-on, it appears as a straight line rather than as an ellipse. A satellite is shown on the orbit line as a heavy dot. At the instant shown in Fig. 3, the satellite is at the intersection of the orbit with the equator; this configuration provides a convenient reference from which to measure time (or, equivalently, rotation). Finally, a longitude line is shown as a convenient reference on the earth, passing through the point on the equator which the satellite crosses; this also appears as a straight line in Fig. 3, since the plane of this longitude line contains the observer at this particular instant.

Figure 4 shows the satellite and the earth after some time has evolved: after the earth and the satellite have travelled an (angular) distance c. Of course, c will vary linearly with time in all of the following, increasing by 2π radians or 360° in one day. In Fig. 4, $c = \pi/4$ radians $= 45°$. Since both the satellite and the point on the equator that coincided with it when the satellite crossed the equator have travelled a distance $c < 90°$, the satellite will lie to the left to the reference longitude line, on the dashed longitude line shown in Fig. 4. When $c = 90°$ the satellite will have caught up with the longitude line and have attained its maximum north latitude (at the right-hand edge of the earth in Fig. 4). As c increases further the satellite will get ahead of the longitude line and start south; at $c = 180°$ the satellite crosses the equator and the longitude line simultaneously, the situation being as shown on Figure 3 except that satellite and longitude line lie on the far side of the earth. For $180° < c < 360°$ the above sequence is simply repeated symmetrically below the equator.

It is clear from the above description that such a satellite will move in a figure 8 as seen from the earth. Figures 5a and b illustrate this 8 pattern superimposed on Figs. 3 and 4.

The geometric parameters that describe the 8 are defined in Fig. 4, where we recall that all (great circle) distances, namely, c, l, a, e, and ψ, are measured by the angle they subtend at the center of the sphere. l is the latitude of the satellite, e the longitude of the satellite measured with respect to a fixed observer located at the point on the orbit where it crosses the equator, ψ is the relative longitude of the satellite measured with respect to an observer on the earth at the equator and the reference longitude line, a is the great circle distance from this earth observer to the satellite, and ϵ the relative angle (azimuth) between the earth observer's north and the satellite; a and ϵ correspond to local polar coordinates fixed to the surface of the earth.

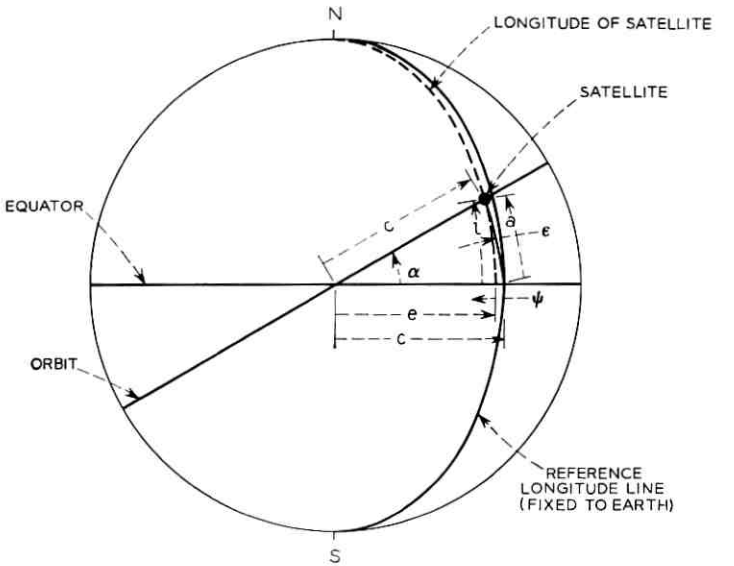


Fig. 4 — Satellite at some later time than Fig. 3. $c = 45°$.

In Fig. 4 and subsequent figures we adopt the convention that dimensions marked with two arrows are regarded as positive quantities; those with only a single arrow are positive in the direction of the arrow, and negative in the opposite direction.

Figure 6 shows an enlarged drawing of the 8 of Fig. 5b viewed from directly overhead, indicating those parameters of Fig. 4 that are measured relative to the earth observer. In addition, x measures the great circle distance from the satellite to the reference longitude, and

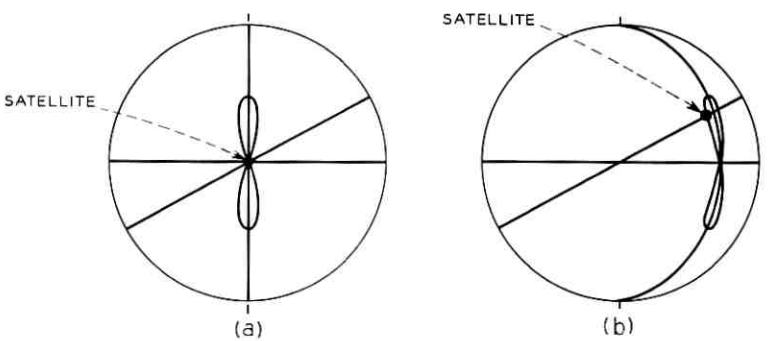


Fig. 5 — Figure 8 pattern. (a) $c = 0°$ as in Fig. 3; (b) $c = 45°$ as in Fig. 4.

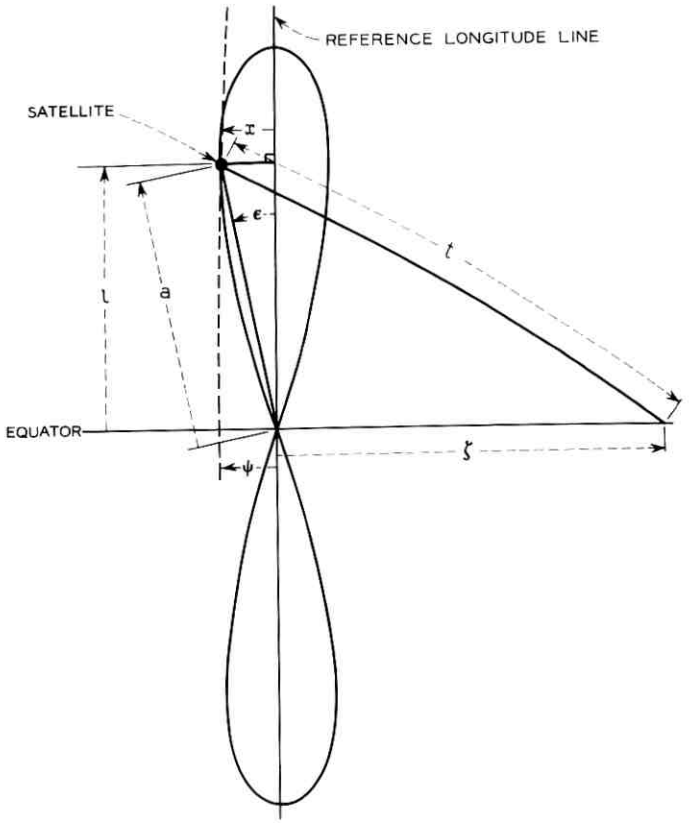


Fig. 6 — Enlarged 8 with relative geometric parameters. $c = 45°$.

t the (great circle) distance from the satellite to a point on the equator a distance ζ from the "cross" of the 8.*

The following results hold true for the parameters of Fig. 4 and 6:

$$\sin \frac{a}{2} = \sin \frac{\alpha}{2} \cdot \sin c. \tag{1}$$

$$\sin \epsilon = \frac{\sin \frac{\alpha}{2} \cdot \cos c}{\left(1 - \sin^2 \frac{\alpha}{2} \cdot \sin^2 c\right)^{\frac{1}{2}}}. \tag{2}$$

* The instant illustrated in Fig. 6 ($c = 45°$) is rather special, in that x has its maximum value and the dashed longitude line passing through the satellite is almost tangent to the 8. These conditions obviously do not hold in general.

$$\sin l = \sin \alpha \cdot \sin c. \tag{3}$$

$$\sin \psi = \frac{\sin^2 \frac{\alpha}{2} \cdot \sin 2c}{(1 - \sin^2 \alpha \cdot \sin^2 c)^{\frac{1}{2}}}. \tag{4}$$

$$\sin e = \frac{\cos \alpha \cdot \sin c}{(1 - \sin^2 \alpha \cdot \sin^2 c)^{\frac{1}{2}}}. \tag{5}$$

$$e + \psi = c. \tag{6}$$

$$\sin x = \sin^2 \frac{\alpha}{2} \cdot \sin 2c. \tag{7}$$

$$\cos t = \cos^2 \frac{\alpha}{2} \cdot \cos \zeta + \sin^2 \frac{\alpha}{2} \cdot \cos (2c + \zeta). \tag{8}$$

Next, we consider the (great circle) separation of two satellites travelling on the same 8. While these satellites lie on the same 8, they lie on different orbits. The geometry of this situation is illustrated in Fig. 7. One orbit is viewed edge-on, as in Figs. 4 and 5; the other orbit's plane is inclined, and so it appears as an ellipse in this figure. The

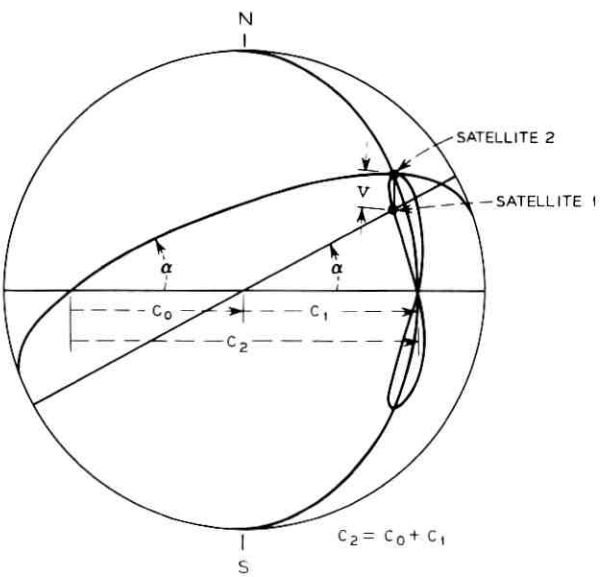


Fig. 7 — Two satellites on the same 8. $c_1 = 45°$ $c_2 = 90°$.

distance between the two satellites is denoted by v. The two satellites have phases c_1 and c_2, with phase difference c_0; c_1 and c_2 of course have the same linear variation with time, while c_0 is fixed. Then the distance between the two satellites is given by

$$\sin^2 \frac{v}{2} = \sin^4 \frac{\alpha}{2} \cdot \sin^2 c_0 + \sin^2 \alpha \cdot \sin^2 \frac{c_0}{2} \cdot \cos^2 \left(c_1 + \frac{c_0}{2} \right),$$

$$c_0 = c_2 - c_1 . \qquad (9)$$

Equation 9 yields equation 1 by setting $v = a$, $c_1 = c$, $c_2 = 0$.

Consider next Fig. 8, which shows two identical 8's spaced by a distance ζ along the equator. The parameter u denotes the (great circle) distance between symmetrically located points at the same latitude. u is illustrated at two representative times, indicated as u_1 and u_2. The corresponding ranges for c_2 are shown on Figure 8, where c_2 is the phase parameter on the right 8. Then

$$\sin \frac{u}{2} = \cos^2 \frac{\alpha}{2} \cdot \sin \frac{\zeta}{2} - \sin^2 \frac{\alpha}{2} \cdot \sin \left(2c_2 - \frac{\zeta}{2} \right). \qquad (10)$$

Our final result contains some of the preceding results as special cases. Consider two 8's of differing size, characterized by parameters α_1 and α_2, spaced by a distance ζ along the equator. This situation is illustrated in Fig. 9. While for clarity the two 8's in Fig. 9 have been drawn as not overlapping, this is not a necessary restriction in the

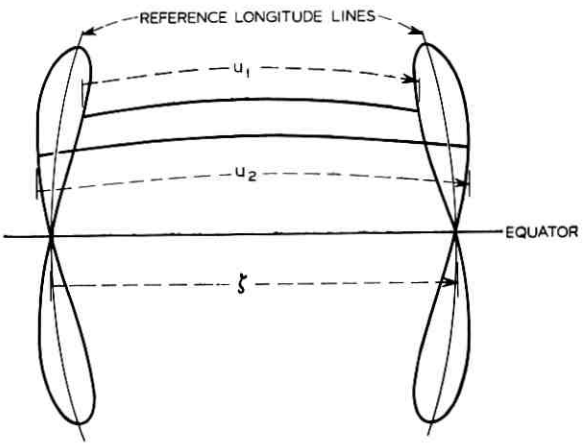


Fig. 8 — Distance between identical 8's at the same latitude. c_2 = phase parameter on right-hand 8. u_1: $0 < c_2 < 90°$; u_2: $90° < c_2 < 180°$.

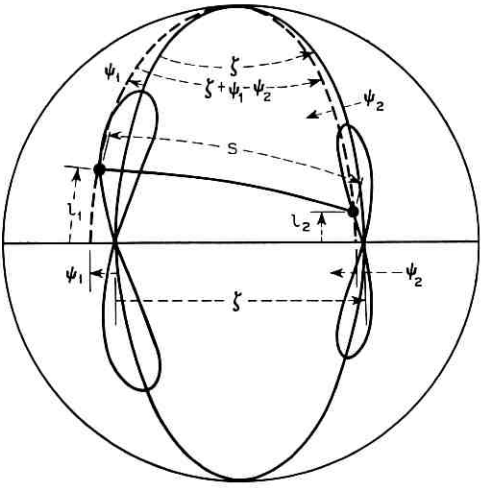


Fig. 9 — Distance between two satellites with arbitrary phases, located on different 8's of arbitrary sizes.

following. Consider a satellite on each 8, having phases c_1 and c_2. Then the great-circle separation s between these two satellites is given by

$$\begin{aligned}\cos s = {} & \cos \zeta \cdot \cos^2 \frac{\alpha_1}{2} \cdot \cos^2 \frac{\alpha_2}{2} \\ & + \frac{\sin \alpha_1 \cdot \sin \alpha_2}{2} [\cos (c_2 - c_1) - \cos (c_2 + c_1)] \\ & + \sin^2 \frac{\alpha_1}{2} \cdot \cos^2 \frac{\alpha_2}{2} \cdot \cos (2c_1 + \zeta) \\ & + \sin^2 \frac{\alpha_2}{2} \cdot \cos^2 \frac{\alpha_1}{2} \cdot \cos (2c_2 - \zeta) \\ & + \sin^2 \frac{\alpha_1}{2} \cdot \sin^2 \frac{\alpha_2}{2} \cdot \cos (2c_2 - 2c_1 - \zeta). \end{aligned} \tag{11}$$

The following specializations of equation 11 yield the indicated above results:

Equation	*Parameters in Equation 11*
1	$s = a : \alpha_1 = \alpha_2 = \alpha,\ c_1 = 0,\ c_2 = c,\ \zeta = 0.$
8	$s = t : \alpha_1 = \alpha,\ c_1 = c,\ c_2 = 0$ or $\alpha_2 = 0.$
9	$s = v : \alpha_1 = \alpha_2 = \alpha,\ \zeta = 0.$
10	$s = u : \alpha_1 = \alpha_2 = \alpha,\ c_1 = \pi - c_2.$

One final specialization of this result is of interest, in which the two 8's have the same size. For convenience we denote the phase difference for the two satellites (each on its separate 8) by c_0, as in equation 9. Thus setting

$$\begin{aligned} \alpha_1 &= \alpha_2 = \alpha, \\ c_2 &= c_1 + c_0 , \end{aligned} \tag{12}$$

in equation 11, yields

$$\begin{aligned} \sin^2 \frac{s}{2} = &\left[\cos^2 \frac{\alpha}{2} \cdot \sin \frac{\zeta}{2} + \sin^2 \frac{\alpha}{2} \cdot \sin \left(c_0 - \frac{\zeta}{2} \right) \right]^2 \\ &+ \left[\sin \alpha \cdot \sin \left(\frac{c_0 - \zeta}{2} \right) \cdot \cos \left(c_1 + \frac{c_0}{2} \right) \right]^2 . \end{aligned} \tag{13}$$

This relation also reduces to equation 9 for $\zeta = 0$, in an obvious way.

All of the above results are exact; no small-angle or any other approximations have been made. Their derivation is sketched in Appendix A. They are necessary for calculating minimum satellite separation in various packing schemes, discussed in Sections III through V.

III. OPTIMUM PACKING ON A SINGLE 8

Equation 9 shows that the separation between two satellites on the same 8 varies periodically with time in a simple way. This relation, together with Fig. 7, readily permits the exact calculation of the closest approach or minimum separation $v_{\min}$ and the corresponding satellite phases $c_{1\,\min}$ and $c_{2\,\min}$ at which it occurs.

$$\begin{aligned} c_{1\,\min} &= \pm \frac{\pi}{2} - \frac{c_0}{2} , \qquad c_{2\,\min} = \pm \frac{\pi}{2} + \frac{c_0}{2} , \\ \sin \frac{v_{\min}}{2} &= \sin^2 \frac{\alpha}{2} \cdot | \sin c_0 | , \qquad c_0 = c_2 - c_1 , \end{aligned} \tag{14}$$

where all c's are measured in radians.

Figure 10 illustrates the satellite positions at minimum separation for two representative cases. Minimum separation for any two satellites occurs when the two lie at the same latitude. (A similar study for maximum separation is easily performed; maximum separation is illustrated in Fig. 11 for the same satellite distributions shown in Fig. 10.)

Consider N satellites placed on an 8, with phases $c_1, c_2, \cdots c_N$, all

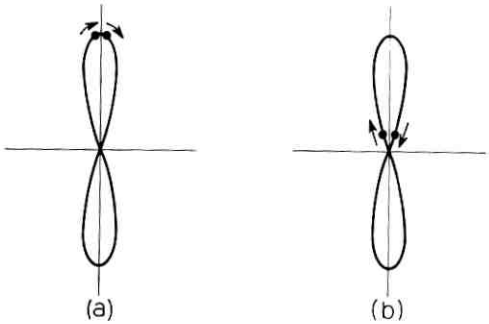


Fig. 10 — Satellite positions at minimum separation. Identical (minimum) separation in both figures. (a) $c_0 = 20°$, (b) $c_0 = 160°$.

having the same linear variation with time t (but different values at $t = 0$); each c increases by 2π radians in one day. We represent the phases of these satellites by points on a circle at angles $c_1, c_2, \cdots c_N$. This pattern of points will rotate uniformly counterclockwise at one revolution per day, and the relative angular positions of these representative points will remain constant with time. One such point is shown in Fig. 12 at angle c_1. Suppose the desired minimum separation between any two satellites on this 8 is specified as $v_{\min}$. The separation between the satellite with phase c_1 in Fig. 12 and any other satellite will equal or exceed $v_{\min}$ if (equation 14)

$$|\sin c_0| \geqq \frac{\sin \frac{v_{\min}}{2}}{\sin^2 \frac{\alpha}{2}} \equiv \sin c_{0\,\min}\,, \tag{15}$$

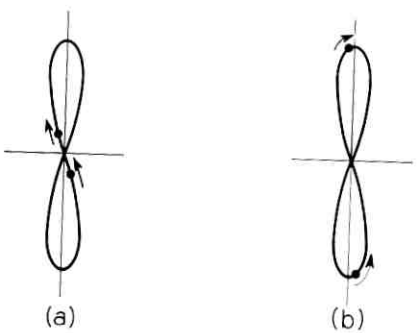


Fig. 11 — Satellite positions at maximum separation. (a) $c_0 = 20°$, (b) $c_0 = 160°$.

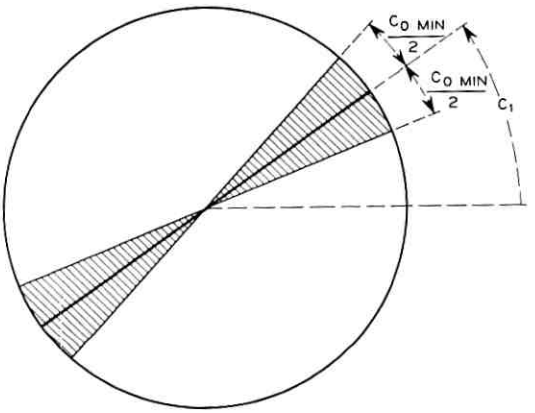


Fig. 12 — Phase of a satellite with associated sector. $c_{0\,min}$ given by equation 15.

where c_0 is now regarded as the phase separation between the satellite under study (No. 1) and any other satellite on this 8.

Suppose we associate with the point c_1 on Fig. 12 a shaded sector of angle $c_{0\,min}$ centered on c_1, and the image of this sector reflected through the axis. Next imagine a second representative point placed on the circle at angle c_2, with a similar pair of shaded sectors. If the shaded regions corresponding to the two representative points do not overlap, equation 15 will be satisfied and $v \geqq v_{min}$. Additional points c_3, c_4, $\cdots$ may be similarly added such that none of the shaded areas overlap, and the separation v between any pair of satellites at any time will be guaranteed to exceed v_{min}.

It is clear that efficient packing requires adjacent shaded areas to just touch; optimum packing will be attained if the entire area of the circle is filled with shaded sections, that is, if

$$N = \frac{\pi}{c_{0\,min}}, \tag{16}$$

where N is the total number of satellites ($c_{0\ min}$ is in radians). The minimum separation with optimum packing for N satellites on the 8 is therefore given by

$$\sin \frac{v_{min}}{2} = \sin^2 \frac{\alpha}{2} \cdot \sin \frac{\pi}{N}. \tag{17}$$

Equation 17 is illustrated in Fig. 13, where the pertinent angles, v_{min} and α, have been expressed in degrees rather than in radians.

It remains only to illustrate the possible geometric distributions of satellites corresponding to the optimum packing of equation 17. Figure 14 illustrates the geometry for an odd number of satellites, $N = 3$. The time has been arbitrarily selected so that one satellite is located at the "cross" of the 8, at zero phase. In Fig. 14a the satellites have been uniformly distributed over half the circle; this may be regarded as the canonical optimum distribution, with others derived from it. The discussion in connection with Fig. 12 makes it clear that an equivalent optimum distribution results from shifting the phase of any of the satellites by 180°.

Figure 14b shows a possible alternative. In general, if the even numbered satellites for N odd are shifted by 180°, then the satellites will be equally-spaced in phase over the entire circle, as illustrated in Fig. 14b. This may be regarded as a second canonical distribution, of subsequent interest for an odd total number of satellites on the 8. Other distributions are clearly possible, although not for $N = 3$. Figure

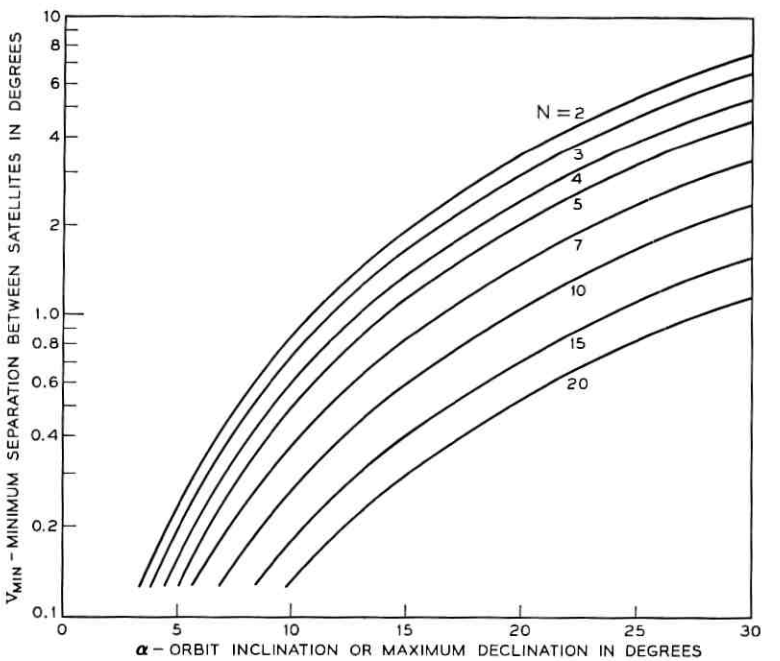


Fig. 13 — Closest approach of satellites on a single 8 with optimum packing. See equation 17. N = number of satellites on a single 8.

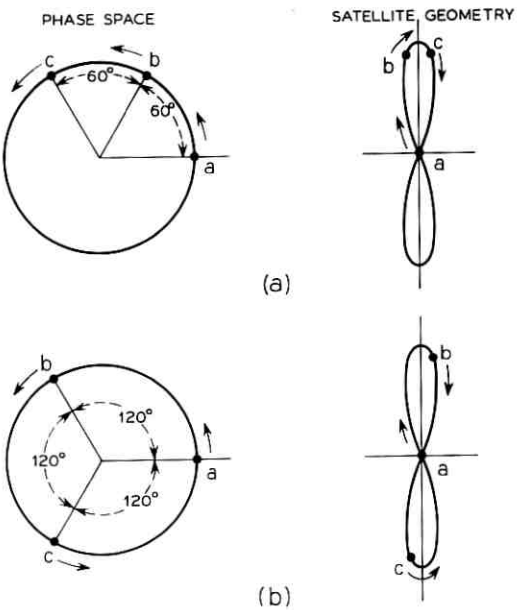


Fig. 14 — Optimum packing for $N = 3$. Uniform phase spacing over (a) 180° and (b) 360°.

15 shows the satellites at different times for the two situations in Fig. 14; the instants of closest approach indicated in Fig. 15 have identical minimum separation.

In general, the optimum uniform distribution over 180° is given by

$$c_p = c_1 + (p - 1)\frac{\pi}{N}, \qquad p = 1, 2, \cdots N;$$

uniform phase spacing over 180° (18)

For an odd number of satellites, we have alternatively

$$c_p = c_1 + 2(p - 1)\frac{\pi}{N}, \qquad p = 1, 2, \cdots N, \qquad N \text{ odd};$$

uniform phase spacing over 360°. (19)

(The c_p are measured in radians in equations 18 and 19.)

All equivalent optimum packing schemes, for a given N and α, result in the identical $v_{\min}$, that is, minimum separation or closest approach between any pair of satellites at any time. However, other geometric properties (such as the average spacing) may vary widely for the

various packing schemes. The minimum separation is the only parameter considered here by assumption *iii* of Section I. The systems implications of different packing schemes are discussed briefly in Section VII.

The results in this section, giving optimum packing of satellites on a single 8, are exact (subject to the assumptions of Section I). They are used in Sections IV and V to determine the best way to stack 8's under various conditions and so determine the maximum number of synchronous satellites that can be used in an extended portion of the sky, subject to different constraints.

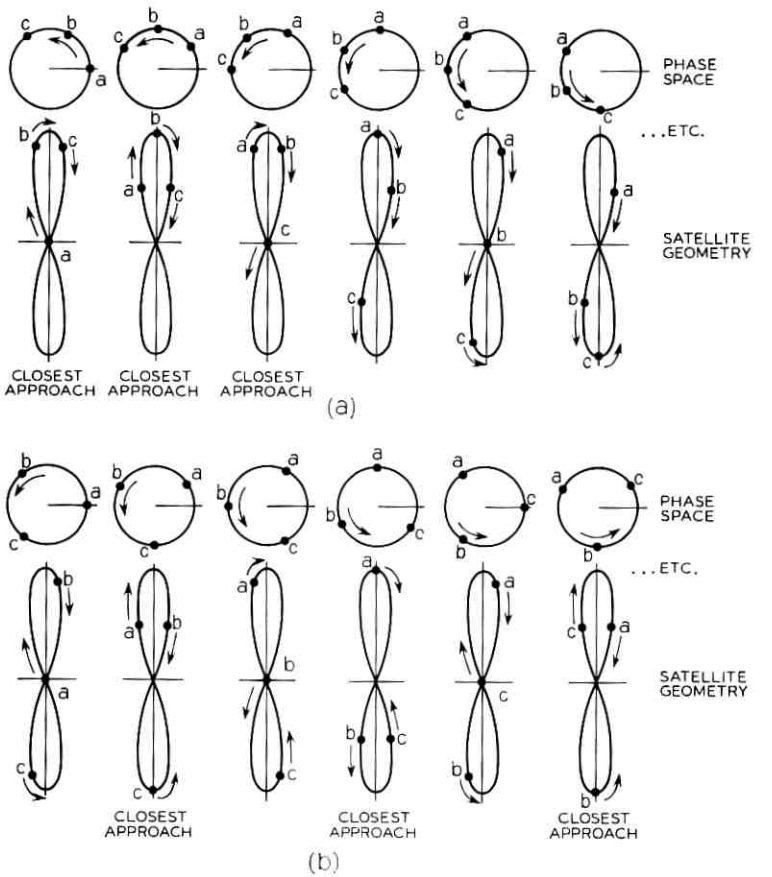


Fig. 15 — Satellite motion for $N = 3$, optimum packing. Phase increment = 30°. Uniform phase spacing over (a) 180° and (b) 360°.

IV. OPTIMUM PACKING—UNSYNCHRONIZED, SEPARATED 8's

Assume that we are given:

(*i*) The minimum allowable separation (that is, the closest approach) between any two satellites $= v_{\min}$.

(*ii*) The maximum allowable orbit inclination (equal to the maximum declination of the satellites) $= \alpha_{\max}$.

Optimum packing on a single 8 has been studied in Section III. We seek to increase the number of synchronous satellites (over the number available in an equatorial system, that is, using a synchronous equatorial orbit only) by using such 8's spaced equally along the equator.

The canonical distribution for N satellites on an 8 is given by equation 18 for general N, or for odd N by equation 19 (with other distributions possible as mentioned in Section III). For a given N and a given $v_{\min}$, it is obviously desirable to use the minimum α allowed, given by equation 17 or Fig. 13, to minimize the space occupied by these N satellites. This equation and figure yield the maximum number of satellites $N_{\max}$ that can be placed on a single 8 for a specified $v_{\min}$, $\alpha_{\max}$. We therefore consider the improvement possible for $N = 2, 3, \cdots N_{\max}$, with corresponding $\alpha = \alpha_1, \alpha_2, \cdots \alpha_{N_{\max}} \leqq \alpha_{\max}$ given by equation 17 or Fig. 13.

It might have been anticipated for specified $v_{\min}$, $\alpha_{\max}$ that the largest number of satellites per 8 (that is, $N_{\max}$) will yield the greatest improvement. This is indeed true for the smaller values of $N_{\max}$ (perhaps $N_{\max} < \sim 10$). Although cases exist for $N_{\max} > 13$ where $N_{\max} - 1$ or $N_{\max} - 2$ are slightly better, the difference is not significant. Consequently, it seems likely that the configuration corresponding to $N_{\max}$, $\alpha_{N\max}$ will either be optimum or very close to it.

Assume that 8's each with N optimally packed satellites are spaced equally along the equator. The N satellites on each 8 must of course be carefully synchronized; however, we assume in this section that no synchronization will be required between the relative phase of groups of satellites on different 8's. Two such 8's are shown in Fig. 16. The minimum (great circle) distance between these two 8's, $u_{\min}$, must equal or exceed $v_{\min}$ in order that the closest approach between satellites on different 8's equals or exceeds $v_{\min}$. The distance between 8's along the equator is denoted by ζ; its minimum value, $\zeta_{\min}$, corresponds to $u_{\min} = v_{\min}$, that is, the minimum distance between 8's equal to $v_{\min}$. Two points are shown on the equator, whose minimum distance to their respective 8's, $t_{\min}$, is set equal to $v_{\min}$; these points are a distance ζ' from each 8 measured along the equator. There is a

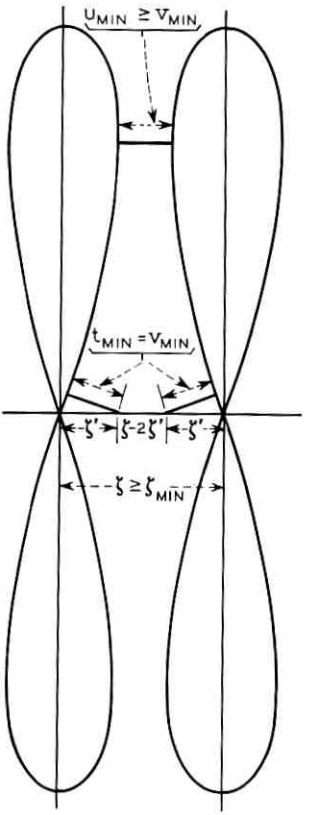


Fig. 16 — Geometry of adjacent 8's. $\xi = \xi_{min}$ for $u_{min} = v_{min}$. v_{min} = closest approach among N optimally packed satellites on each 8.

portion of equator $\zeta - 2\zeta'$ long, available for additional synchronous equatorial satellites, separated from each other by a distance v_{min}.*

We now define an improvement factor I that measures the effective-

* The possibility of using smaller 8's, rather than only equatorial satellites between the major 8's, has also been considered. However, a crude analysis that replaces sines of small angles by their arguments and effectively neglects the curvature of the sphere shows that for $N < 16$ there is not enough room for a small 8 containing even two satellites between the two main 8's spaced at the minimum distance ζ_{min}. This suggests that small 8's will be of no value for the range of interest in this paper ($N < \sim 25$—see Fig. 13), although the packing can obviously be improved for very large N ($N \gg 25$) by this means. A precise investigation of this problem requires a detailed study based on equation 11.

ness of our satellite packing schemes as follows:

$$I = \frac{\begin{matrix}\text{number of satellites per unit of longitude for} \\ \text{packing scheme under consideration}\end{matrix}}{\begin{matrix}\text{number of satellites per unit of longitude} \\ \text{using equatorial orbit only (that is, without 8's)}\end{matrix}}, \tag{20}$$

with closest approach the same in numerator and denominator. If E denotes the number of equatorial satellites between adjacent 8's, then from Fig. 16

$$E = \left\{\frac{\zeta - 2\zeta'}{v_{\min}}\right\} + 1 \tag{21}$$

where the { } denote "the largest integer contained in" the enclosed expression.* For the packing scheme under consideration $N + E$ satellites occupy a longitude interval ζ. For an equatorial system one satellite occupies a longitude interval $v_{\min}$, so that effectively $\zeta/v_{\min}$ satellites occupy a longitude interval ζ. Therefore from equation 20 and 21 the improvement factor is

$$I = \frac{N + 1 + \left\{\frac{\zeta - 2\zeta'}{v_{\min}}\right\}}{\frac{\zeta}{v_{\min}}}, \qquad \zeta \geqq \zeta_{\min}. \tag{22}$$

(Notice that the denominator of equation 22 need not be an integer, as must the numerator.)

I of equation 22 is a function of ζ. Suppose we first set $\zeta = \zeta_{\min}$, that is, place adjacent 8's as close as possible without causing (asynchronously phased) satellites on the two 8's to approach each other closer than $v_{\min}$. In general the E equatorial satellites of equation 21 will not exactly fill up the available space on the equator of length $\zeta_{\min} - 2\zeta'$; that is, the quantity $(\zeta_{\min} - 2\zeta')/v_{\min}$ will not be exactly an integer.

We now inquire if increasing the spacing ζ between adjacent 8's beyond its minimum value can improve I of equation 22. The answer is sometimes yes, sometimes no. Intuitively, if $(\zeta_{\min} - 2\zeta')/v_{\min}$ in the numerator of equation 22 is only slightly greater than an integer, that is, if there is only a little extra room on the equator at minimum distance between 8's, then $\zeta = \zeta_{\min}$ gives the largest improvement

* Equation 21 is obviously valid only when it predicts $E \geqq 0$; it will appear below that $E \geqq 1$.

factor I. However, if $(\zeta_{\min} - 2\zeta')/v_{\min}$ is only slightly less than an integer, in other words, if there is almost enough room on the equator for an extra satellite for $\zeta = \zeta_{\min}$, it pays to increase ζ a little beyond $\zeta_{\min}$ to just accommodate the extra equatorial satellite. It clearly never pays to increase ζ further.

Consequently, we evaluate I and E of equation 22 and 21 for the following two values of ζ:

$$\zeta_1 = \zeta_{\min}, \tag{23}$$

$$\zeta_2 = 2\zeta' + v_{\min} + v_{\min}\left\{\frac{\zeta_{\min} - 2\zeta'}{v_{\min}}\right\}. \tag{24}$$

The corresponding improvement factors I_1 and I_2, with corresponding numbers of equatorial satellites E_1 and E_2, are:

$$I_1 = \frac{N + 1 + \left\{\frac{\zeta_{\min} - 2\zeta'}{v_{\min}}\right\}}{\frac{\zeta_{\min}}{v_{\min}}}; \qquad E_1 = \left\{\frac{\zeta_{\min} - 2\zeta'}{v_{\min}}\right\} + 1. \tag{25}$$

$$I_2 = \frac{N + 2 + \left\{\frac{\zeta_{\min} - 2\zeta'}{v_{\min}}\right\}}{\left\{\frac{\zeta_{\min} - 2\zeta'}{v_{\min}}\right\} + \frac{2\zeta'}{v_{\min}} + 1}; \qquad E_2 = \left\{\frac{\zeta_{\min} - 2\zeta'}{v_{\min}}\right\} + 2. \tag{26}$$

The optimum improvement factor is the greater of I_1 and I_2. It remains only to evaluate equation 25 and 26 from the results of Sections II and III.

From equation 17

$$\sin\frac{v_{\min}}{2} = \sin^2\frac{\alpha}{2}\cdot\sin\frac{\pi}{N}. \tag{27}$$

From equation 10 and Fig. 8

$$\sin\frac{u_{\min}}{2} = \cos^2\frac{\alpha}{2}\cdot\sin\frac{\zeta}{2} - \sin^2\frac{\alpha}{2}; \tag{28}$$

consequently,

$$\sin\frac{\zeta}{2} = \frac{\sin\frac{u_{\min}}{2}}{\cos^2\frac{\alpha}{2}} + \tan^2\frac{\alpha}{2}. \tag{29}$$

We set $u_{\min} \rightarrow v_{\min}$ and $\zeta \rightarrow \zeta_{\min}$ in equation 29 to obtain

$$\sin \frac{\zeta_{\min}}{2} = \tan^2 \frac{\alpha}{2} \left(1 + \sin \frac{\pi}{N}\right). \tag{30}$$

From equation 8 and Fig. 6, with $\zeta \rightarrow \zeta'$ to conform to the notation in this section and Fig. 16,

$$\cos t_{\min} = \cos^2 \frac{\alpha}{2} \cdot \cos \zeta' + \sin^2 \frac{\alpha}{2}. \tag{31}$$

Using the double-angle formula for the two cosines, this becomes

$$\sin \frac{t_{\min}}{2} = \cos \frac{\alpha}{2} \cdot \sin \frac{\zeta'}{2}\ ; \tag{32}$$

finally setting $t_{\min} \rightarrow v_{\min}$ in equation 32,

$$\sin \frac{\zeta'}{2} = \frac{\sin^2 \frac{\alpha}{2} \cdot \sin \frac{\pi}{N}}{\cos \frac{\alpha}{2}}. \tag{33}$$

Equations 27, 30, and 33 substituted in equations 25 and 26 permit evaluation of I (the improvement factor) and E (the corresponding number of satellites on the equator between 8's), as a function of N (the number of satellites per 8) and α (the orbit inclination or maximum declination of the satellites on the 8's), for the two interesting values of spacing (along the equator) between 8's given in equation 23 and 24. These results are plotted on Figure 17; equation 25 is shown by solid lines, and equation 26 by dashed lines.

Figures 13 and 17 used in conjunction determine the improvement factor for the present packing scheme. (Some values of N have been omitted in these figures for clarity, but may of course be supplied by the equations from which these figures were obtained.) Given $v_{\min}$ and $\alpha_{\max}$:

(*i*) On Fig. 13, find the largest value of N corresponding to $v_{\min}$ and $\alpha \leqq \alpha_{\max}$. Denote this value of N by $N_{\max}$, the associated value of α by $\alpha_{N_{\max}}$; $\alpha_{N_{\max}} \leqq \alpha_{\max}$.

(*ii*) On Fig. 17, find the two possible improvement factors (solid and dashed) corresponding to $N_{\max}$, $\alpha_{N_{\max}}$ and choose the larger. Read the associated number of equatorial satellites from the appropriate curve.*

* As mentioned at the beginning of this section, it is possible that a smaller value of $N < N_{\max}$ may yield a slightly, although not significantly, larger improvement factor. Fig. 17 makes it clear that this is possible only for large N, and of course for intermediate N not plotted in this figure.

A variety of possible types of behavior is illustrated in Fig. 17. For some N the solid curve always exceeds the dotted curve (for example, $N = 2, 3, 7, 10$) and the 8's should be spaced as closely as possible; for other N the reverse is true (as when $N = 5$) and the distance between 8's should be increased just enough beyond its minimum value to permit one additional satellite on the equator between each pair of 8's. In both these cases the number of equatorial satellites remains constant for all values of α shown.

There are still other cases where both types of the above behavior hold for a given N; for small values of α the 8's should be close-

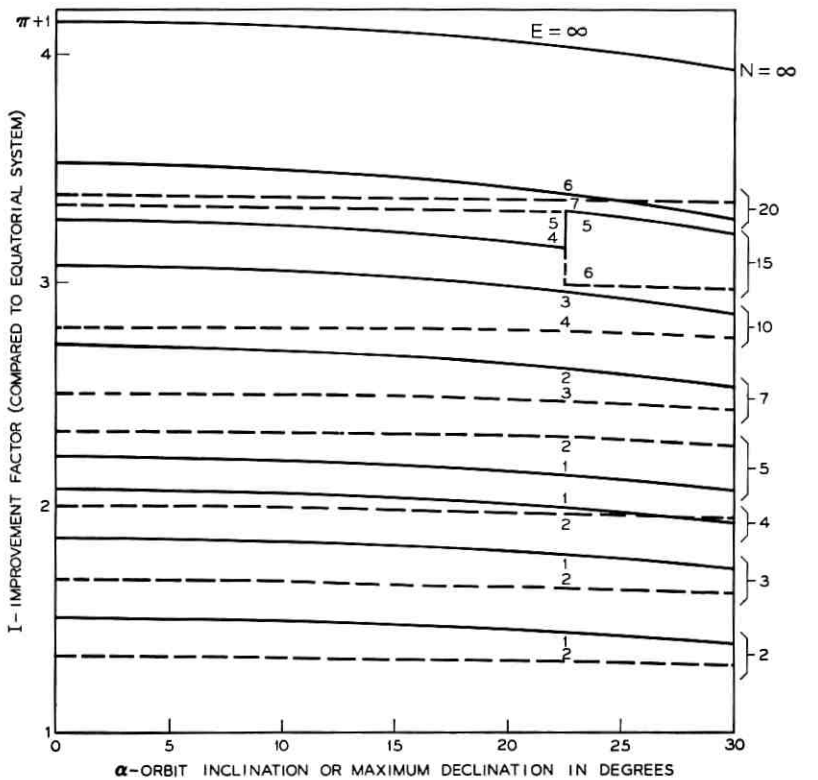


Fig. 17 — Improvement factor for optimum satellite packing on unsynchronized, separated 8's. Use this figure in conjunction with Fig. 13. E = number of satellites on equator between two adjacent 8's. N = number of satellites on a single 8. ——— indicates minimum spacing between adjacent 8's; closest approach for adjacent 8's = closest approach for a single 8. (See fig. 13 and equation 25). - - - - indicates increased spacing between adjacent 8's that just permits 1 additional equatorial satellite. See equation 26.

spaced, for large α there should be an extra equatorial satellite (such as, $N = 4, 20$). Here the number of equatorial satellites increases by 1 at a critical value of α.

Finally, for $N = 15$ the behavior changes at a critical value of α in a different way. Here the number of equatorial satellites remains constant for all α shown. For small α the spacing between 8's should be increased to allow for an extra equatorial satellite, while for large α the 8's are close-spaced.

The above relations simplify as $\alpha \to 0$ or as $N \to \infty$. We have:

$$\underline{\alpha = 0}$$

$$I_1 = \frac{N + \left\{\dfrac{1}{\sin \dfrac{\pi}{N}}\right\}}{1 + \dfrac{1}{\sin \dfrac{\pi}{N}}}; \qquad E_1 = \left\{\frac{1}{\sin \frac{\pi}{N}}\right\}, \tag{34}$$

$$I_2 = \frac{N + 1 + \left\{\dfrac{1}{\sin \dfrac{\pi}{N}}\right\}}{2 + \left\{\dfrac{1}{\sin \dfrac{\pi}{N}}\right\}}; \qquad E_2 = \left\{\frac{1}{\sin \frac{\pi}{N}}\right\} + 1. \tag{35}$$

$$\underline{N = \infty}$$

$$I = I_1 = I_2 = \frac{\pi \sin^2 \frac{\alpha}{2}}{\sin^{-1} \tan^2 \frac{\alpha}{2}} + 1; \qquad \frac{E}{N} = \frac{\sin^{-1} \tan^2 \frac{\alpha}{2}}{\pi \sin^2 \frac{\alpha}{2}}. \tag{36}$$

It is clear from equation 36 that an upper bound on the improvement factor for the present packing scheme is $I = \pi + 1$, with $E/N = 1/\pi$. Figure 13 shows that this may be approached only for extremely small closest approach $v_{\min}$; hence practical improvement factors will be smaller.

V. OPTIMUM PACKING—SYNCHRONIZED, OVERLAPPING 8'S

In the packing schemes of Section IV, optimally-packed 8's are spaced far enough apart so that closest approach for satellites on

adjacent 8's equals or exceeds closest approach for satellites on the same 8, for arbitrary relative phasing on different 8's. We now inquire if the number of satellites can be increased (for a given closest approach and orbit inclination or maximum declination) by careful relative phasing on 8's in such a way that they can lie closer together, and even overlap (with proper interleaving of satellites on different 8's). We restrict the present treatment to 8's of identical size (that is, identical orbit inclination α) for simplicity, with equatorial satellites between 8's as permitted.

The basic relation needed for this study is equation 13 (and equation 12), which shows that the separation between two satellites on different 8's (of the same size) varies periodically with time in a simple way. We denote this minimum separation by $s_{\min}$, occurring at satellite phases $c_{1\,\min}$ and $c_{2\,\min}$ on the left and right 8's (spaced by ζ along the equator), respectively. Then

$$c_{1\,\min} = \pm\frac{\pi}{2} - \frac{c_0}{2}, \qquad c_{2\,\min} = \pm\frac{\pi}{2} + \frac{c_0}{2},$$

$$\sin\frac{s_{\min}}{2} = \sin^2\frac{\alpha}{2}\cdot\left|\sin\left(c_0 - \frac{\zeta}{2}\right) + \cot^2\frac{\alpha}{2}\cdot\sin\frac{\zeta}{2}\right|, \tag{37}$$

$$c_0 = c_2 - c_1 .$$

We see that closest approach for satellites on two different 8's of the same size occurs when the two satellites attain the same latitude; this behavior was observed in Section III for two satellites on the same 8 (compare the top lines of equations 14 and 37). $s_{\min}$ of equation 37 reduces to $v_{\min}$ of equation 14 for $\zeta = 0$, the degenerate case in which the two 8's become identical.

Assume now that $s_{\min}$ is given. Equation 37 then determines two forbidden regions for c_0; for every satellite on the left (No. 1) 8, with phase c_1, two forbidden regions in phase space on the right (No. 2) 8 are established, that is, two ranges of c_2 within which no satellites may be placed. Thus, in addition to the self-phase-space of Section III (determined by equation 14), which governs packing on each 8 considered independently, a mutual-phase-space (determined by equation 37) must be considered for each pair of 8's lying closer than $\zeta_{\min}$ of Section IV (equation 30).

Figure 18 and 19 give a convenient geometric picture of this pair of forbidden regions. We assume each 8 of the pair considered to contain N optimally packed satellites, according to Section III. Then

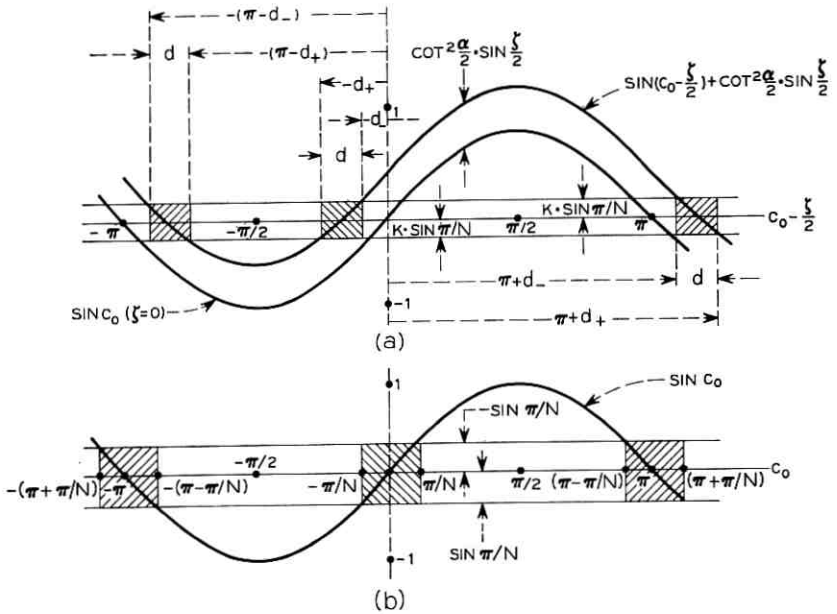


Fig. 18 — Forbidden region for relative phase between satellites on different 8's (equation 37). $N = 9$. (a) General case. $\cot^2 \alpha/2 \cdot \sin \zeta/2 = 0.5$ (separation about half that for two 8's just touching), $K = 0.6$. Enlarged portion of upper curve, containing forbidden region to left of origin, shown in Fig. 19. (b) Limiting case (single 8). $\zeta = 0$, $K = 1$. This curve is the same as the lower curve of (a).

closest approach on each 8 is $v_{\min}$, given by equation 17 as

$$\sin \frac{v_{\min}}{2} = \sin^2 \frac{\alpha}{2} \cdot \sin \frac{\pi}{N}. \tag{38}$$

We assume that $s_{\min}$, closest approach between 8's, is less than $v_{\min}$, according to the following relation for later convenience.

$$\sin \frac{s_{\min}}{2} = K \cdot \sin \frac{v_{\min}}{2}, \qquad 0 < K < 1. \tag{39}$$

Thus,

$$\sin \frac{s_{\min}}{2} = \sin^2 \frac{\alpha}{2} \cdot K \sin \frac{\pi}{N}, \qquad 0 < K < 1. \tag{40}$$

For zero spacing, $\zeta = 0$ (Fig. 18b), the principal forbidden regions are located symmetrically about $c_0 = -\pi, 0$. In this case for $K = 1$ their width is $d = 2\pi/N$. As the distance between 8's increases (Fig. 18a) the centers of the forbidden regions approach each other symmetrically and

their width d increases, slowly at first. This pattern of forbidden regions repeats periodically with period 2π. Figure 19 shows an enlarged portion of the principal forbidden region closest to the origin. The boundaries of the forbidden regions are defined in Fig. 18 and 19, the center of one of them in Fig. 19, at $-D$. Dimensions with a single arrow have a sign (for example, d_+ , d_- , D, **D**), those with two arrows are positive (d, **d**/2), as before.

From Figs. 18 and 19 and equations 37 through 39 the parameters

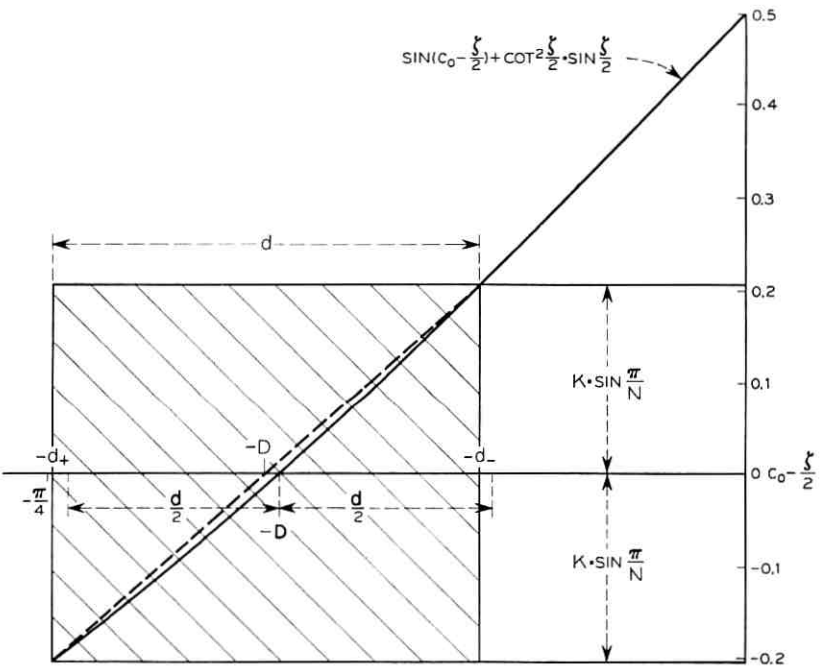


Fig. 19 — Enlarged portion of upper curve of Fig. 18a, showing forbidden region closest to origin.

Forbidden Region

Exact	*Approximate*
$D = 30.92°$	$\mathbf{D} = 30°$
$d = 27.55°$	$\mathbf{d} = 27.2°$
$d_- = 17.15°$	$\mathbf{D} - \frac{\mathbf{d}}{2} = 16.4°$
$d_+ = 44.7°$	$\mathbf{D} + \frac{\mathbf{d}}{2} = 43.6°$

defining the forbidden regions are:

$$d_{\pm} = \sin^{-1}\left(\cot^2\frac{\alpha}{2}\cdot\sin\frac{\zeta}{2} \pm K\sin\frac{\pi}{N}\right). \tag{41}$$

$$D = \frac{d_+ + d_-}{2} \tag{42}$$

$$d = d_+ - d_- . \tag{43}$$

The centers of the principal forbidden regions are located at $-D$, $-(\pi - D)$. In phase space they are located as follows:

$$\begin{aligned} \frac{\zeta}{2} - D - \frac{d}{2} < c_0 < \frac{\zeta}{2} - D + \frac{d}{2}. \\ \frac{\zeta}{2} - \pi + D - \frac{d}{2} < c_0 < \frac{\zeta}{2} - \pi + D + \frac{d}{2}. \end{aligned} \tag{44}$$

The parameter $\cot^2 \alpha/2 \cdot \sin \zeta/2$, which appears throughout the above relations, is a normalized spacing parameter. For such a ζ that the two 8's just touch, $\cot^2 \alpha/2 \cdot \sin \zeta/2 = 1$; therefore, for overlapping 8's this parameter lies between 0 and 1, and for some of the following work where overlapping 8's lie very close together $\cot^2 \alpha/2 \cdot \sin \zeta/2 \ll 1$. Since $\alpha < 30°$, $\zeta/2 < 4.12°$ for overlapping 8's; consequently the approximation $\sin \zeta/2 \to \zeta/2$ is valid for most purposes.

In the large N case certain approximations are useful. D and d of equations 42 and 43 are approximated by $\mathbf{D}$ and $\mathbf{d}$, given by equations 45 and 46 and illustrated in Fig. 19.

$$\mathbf{D} \equiv \sin^{-1}\left(\cot^2\frac{\alpha}{2}\cdot\sin\frac{\zeta}{2}\right) \approx D, \qquad N \gg 1. \tag{45}$$

$$\mathbf{d} \equiv \frac{2K\sin\dfrac{\pi}{N}}{\cos \mathbf{D}} \approx d, \qquad N \gg 1. \tag{46}$$

While these results are not sufficiently precise for all purposes, they serve at least as a rough guide for initial thinking about the problem. For the case illustrated in Fig. 19 ($N = 9$, $K = 0.6$, $\cot^2 \alpha/2 \cdot \sin \zeta/2 = 0.5$) the approximate results of equations 45 and 46 yield a first forbidden region 0.35° too narrow (out of a total width of 27.55°) and 0.92° too far toward the origin. This case has large spacing, and the error will decrease for smaller ζ. In the approximation of equations 45 and 46 the width of the forbidden regions varies linearly with the

secant of the displacement of their centers and with the separation reduction factor K, and the location of their centers is independent of K.

A phase-plane picture of the forbidden regions of Figs. 18 and 19 is illustrated in Fig. 20; for later convenience different shading and edges are used to denote the two forbidden regions. For zero spacing, $\zeta = 0$, the forbidden regions have minimum width d and are located symmetrically about 0 and π in the $(c_0 - \zeta/2)$ plane; as ζ increases the forbidden regions approach each other and their width d increases, slowly at first. In Fig. 20a, $K = 1$ and $d > 2\pi/N$ for $\zeta > 0$; in Fig. 20b $K = 0.6$, and for the range of ζ shown $d < 2\pi/N$. It is clear that given an upper bound on ζ, K can be chosen small enough so that $d < 2\pi/N$.

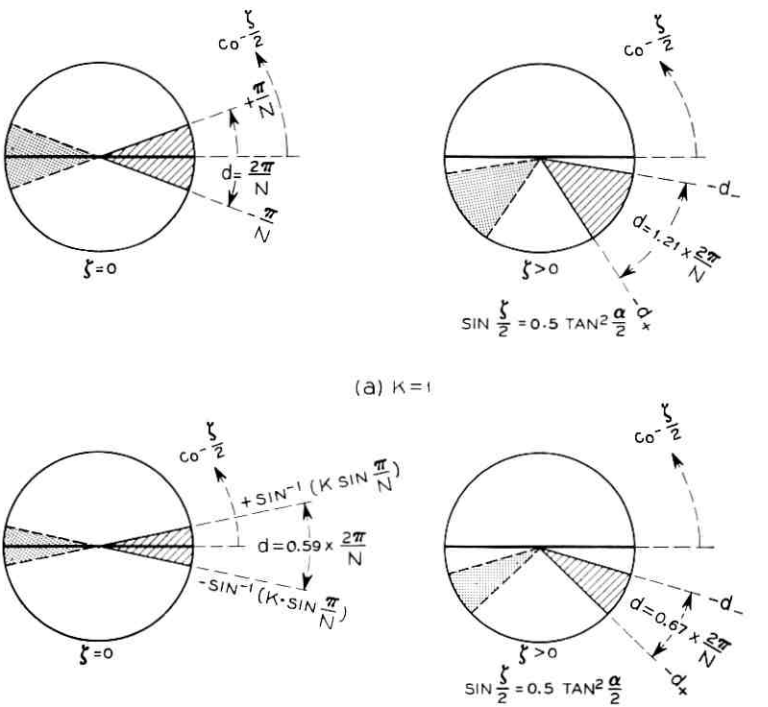


Fig. 20 — Forbidden regions of Figs. 18 and 19 in the $(c_0 - \zeta/2)$ phase plane. $N = 9$. Figures are symmetrical around the vertical axis $(c_0 - \zeta/2 = \pm \pi/2)$.

We now make the additional assumption that each optimally-packed 8 possesses uniform phase spacing over 360°, as in equation 19; this of course implies that N is odd. We inquire whether two or more such 8's can overlap in a useful way, that is, so that improvement factors greater than those of Fig. 17 can be obtained (for the same minimum separation or closest approach). This assumption simplifies the following discussion by rendering all phase-space diagrams N-fold rotationally symmetric, and does yield significant increases in the improvement factor. While superficial study of 180° packing on the individual 8's seems to indicate lower improvement factors in certain simple cases, we have not made a careful study of this alternative case to determine what, if any, advantages it might have. Furthermore, we have not considered other possible (non-180° or non-360° single-8) packing schemes of Section III at all.

Consider the case of two interleaved 8's. The satellite phases on these 8's are (equation 19):

$$c_{1(p)} = c_{1(1)} + 2(p-1)\frac{\pi}{N}, \qquad p = 1, 2, \cdots N;$$
$$c_{2(q)} = c_{2(1)} + 2(q-1)\frac{\pi}{N}, \qquad q = 1, 2, \cdots N; \qquad N \text{ odd.} \tag{47}$$

We define

$$c_0 \equiv c_{2(p)} - c_{1(p)}, \qquad p = 1, 2, \cdots N. \tag{48}$$

All of the c_1's and c_2's vary linearly with time, increasing by 2π in one day. Figure 21a shows the corresponding phase-space plots for each 8, as determined in Section III. In this figure we assume without subsequent restriction that the phase of the $p = 1$ satellite on the left 8 is 0, $c_{1(1)} = 0$. We now inquire what is the optimum value for c_0 (the relative phase between corresponding satellites on the two 8's) to maximize $s_{\min}$, (the closest approach between satellites on different 8's), and thus maximize the packing improvement factor I.

We investigate this question by a mutual phase-space picture derived from Fig. 20. Associated with each of the N satellites on the left 8 (No. 1) there are two forbidden regions in phase space c_2 (or equivalently $c_2 - \zeta/2$ where this is more convenient) determined by a figure such as one of those in Fig. 20. The details depend on the two parameters K (the separation reduction factor of equation 39) and ζ (the equatorial spacing between the two 8's). Figure 20 gives directly the forbidden regions in $c_2 - \zeta/2$ phase space corresponding to

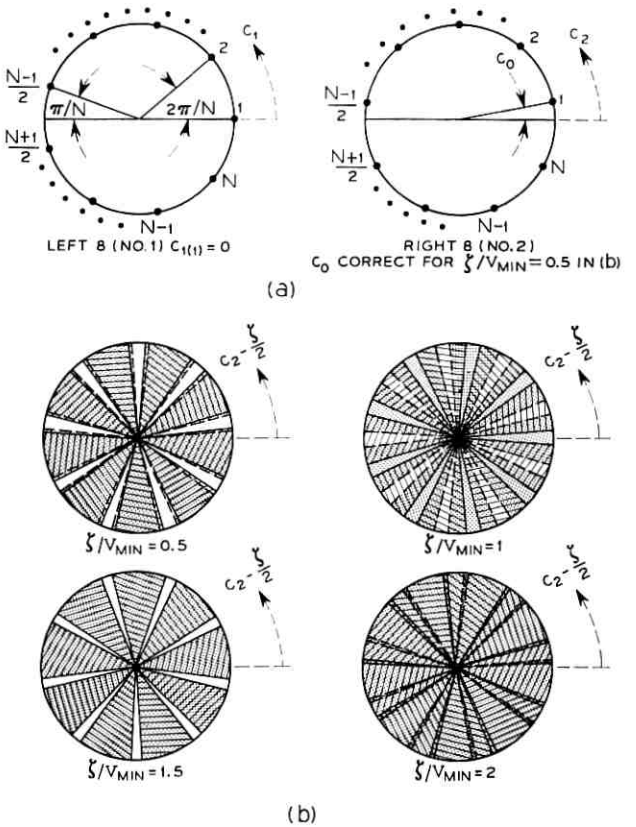


Fig. 21 — Self and mutual phase-space plots for 360° packing on each 8. N is odd, $N = 9$, $\alpha = 30°$, $K = 0.75$. (a) Self phase-space plots for two overlapping 8's. Phase of satellite 1 = 0 on left 8. (b) Mutual phase-space plot: forbidden regions in ($c_2 - \zeta/2$) plane determined by satellites in left 8 with phases c_1 of left figure in (a).

the No. 1 satellite on the left 8 in Fig. 21a; the corresponding forbidden regions for the other $(N - 1)$ satellites are determined by rotating such a figure $(N - 1)$ times by an angle $2\pi/N$.

The resulting mutual phase-space picture, shown in Fig. 21b, gives the forbidden and permitted regions for satellites on the right (No. 2)8.* The values of parameters have been chosen in this figure for convenience of illustration, and not necessarily for optimum packing.

* The phase-space plots of Fig. 21 correspond to the particular time at which $c_{1(1)} = 0$; they rotate uniformly counterclockwise, at one revolution per day.

We may imagine that these figures are generated by two spoked wheels rotating in opposite directions on a common shaft. The clockwise rotating wheel has spokes indicated by shading lines, the counter-clockwise wheel by dots, corresponding to the two forbidden regions indicated in the drawings of Fig. 20. Clearly $K < 1$ in order that there be space between spokes for any ζ. As each wheel rotates the spokes widen, and the space between the spokes decreases, finally vanishing ($\zeta/v_{\min} = 2$ in Fig. 21b) as the wheels become solid. Before this happens there are certain angles ($\zeta/v_{\min} = 0.5, 1.5$ in Fig. 21b) at which the spaces between the spokes on the two counter-rotating wheels line up, offering permitted regions for satellites on the second (right) 8.

As indicated in this figure, it proves convenient to normalize the equatorial spacing ζ between 8's to $v_{\min}$, the minimum separation or closest approach on each 8 (equation 38 or 17, or Fig. 13). Figure 21 shows that there are a number of actual or potential "windows" for interleaving two 8's, at spacings of approximately $\zeta/v_{\min} \approx 0.5, 1.5, \cdots$. It is clear that the width of the corresponding permitted regions in c_2-space decreases (and eventually disappears) as the order of the "window" increases, for a fixed K, since the spokes increase in width as ζ increases. Stated differently, the first window (corresponding to the closest spacing for the two 8's) has the highest permitted value of K (such that the width of the permitted regions in c_2-space just approaches zero). Since the highest packing improvement factor corresponds to the largest K, it is clear that best packing (considering only closest approach) is obtained at the first window, that is, with $\zeta/v_{\min} \approx 0.5$ in Fig. 21. It is further clear that a larger value of K than indicated in Fig. 21 can be used.*

Let ζ'' denote the separation between 8's at the first window with the largest possible value of K, such that the permitted regions in c_2-space approach zero width. Then from Figs. 20 and 21

$$d_- = -\frac{\pi}{2N}, \qquad d_+ = +\frac{3\pi}{2N}. \tag{49}$$

From equation 41

$$\sin\frac{\zeta''}{2} = \tan^2\frac{\alpha}{2}\cdot\left[\frac{\sin\dfrac{3\pi}{2N} - \sin\dfrac{\pi}{2N}}{2}\right] \tag{50}$$

* Although not shown in Fig. 21, it is clear that for $\zeta = 0$ there are no permitted regions (in c_2 - space) for $K > \frac{1}{2}$. This is the degenerate case in which the two 8's coincide, so that only single-8 packing is effectively considered. The present treatment provides an alternative phase-space derivation to that of Fig. 12 and Section III for the optimum single-8 packing results.

$$K = \frac{\sin \frac{\pi}{2N} + \sin \frac{3\pi}{2N}}{2 \sin \frac{\pi}{N}}. \tag{51}$$

Noting equation 38, an equivalent form of equation 50 that is sometimes useful is

$$\frac{\sin \frac{\zeta''}{2}}{\sin \frac{v_{\min}}{2}} = \sec^2 \frac{\alpha}{2} \cdot \left[\frac{\sin \frac{3\pi}{2N} - \sin \frac{\pi}{2N}}{2 \sin \frac{\pi}{N}} \right]. \tag{52}$$

Finally, the relative phase between corresponding satellites on the two 8's (equation 48) is

$$c_0 = \frac{\pi}{2N} + \frac{\zeta''}{2}. \tag{53}$$

N is odd, of course, throughout equations 49 through 53 (see equation 47). We denote the above geometry by the term "close-spaced interleaved 8's".

Figures 22 and 23 illustrate the spacing between interleaved 8's and the corresponding reduction in minimum separation or closest approach (for satellites on different 8's) for the above conditions. Except for the smallest N, the spacing between 8's ζ'' is close to $v_{\min}/2$, one half the closest approach for satellites on each individual 8, as in the first drawing of Fig. 21b. Similarly, the closest approach $s_{\min}$ for satellites on different 8's is only slightly less than $v_{\min}$ except for the smallest N.* Similar results may be derived for the higher-order windows, but as already mentioned they will have smaller values of K, and hence of the packing improvement factor.

Figure 24 shows the geometric distribution of satellites for two close-spaced interleaved 8's in the region of the equator, at three successive times, for large N. Corresponding satellites on the two interleaved 8's have about the same longitude near the equator, neglecting the factor proportional to $\tan^2 (\alpha/2)$ in satellite phase on the right (dashed) 8. In addition to the two intersections shown near the equator in Fig. 24, the two 8's have two other intersections, near the ends of the 8's, where the satellites must also properly interleave, not

---

* $s_{\min}/v_{\min}$ is approximately equal to K, as in Fig. 23, to one part in 10^{-4} for the worst case $N = 3$, and to five parts in 10^{-6} for $N \geqq 7$.

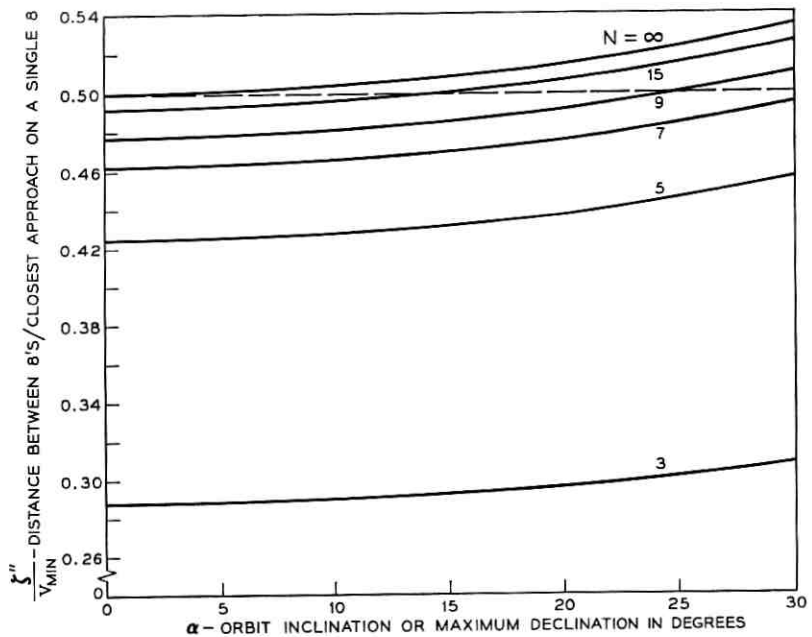


Fig. 22 — Equatorial distance between two close-spaced interleaved 8's. See equations 50 and 52. N = number of satellites on a single 8. N is odd.

shown in Fig. 24. Figure 25 illustrates motion over the complete 8's for a moderate value of N, showing the general nature of closest approaches near all four intersections. Each satellite has 8 closest approaches per day with four different satellites on the other 8 (in addition to four closest approaches per day, with two different satellites on its own 8).

We calculate the improvement factor for such interleaved pairs (close-spaced—that is, at the first "window," $\zeta/v_{\min} \approx 0.5$ for large N) spaced equally along the equator. The $2N$ satellites on each interleaved pair must of course be carefully synchronized; no relative synchronization will be assumed between different pairs.* Two such interleaved pairs are shown in Fig. 26.

The following discussion is quite similar to that of Section IV and

* By synchronizing all of the different pairs they could be placed slightly closer together; indeed, the same is true for the (nonoverlapping) 8's of the packing scheme of Section IV. The phase-space picture given earlier in the present section can easily deal with this problem; however, the possible increase in improvement factor is so slight, particularly for the larger N, that we do not pursue this additional complication here.

Fig. 16; however, the symbols have slightly different definitions here. The closest approach for satellites on different 8's of an interleaved pair, s_{min}, is a little less than v_{min}, the closest approach on each individual 8 (see equation 39 and Fig. 23). Hence the equatorial satellites are assumed to be separated from each other and from the adjacent 8's by s_{min}; the minimum distance between adjacent 8's, u_{min}, must equal or exceed s_{min}; and the equatorial system used for comparison in computing the improvement factor is assumed to have satellites separated by s_{min}. With these changes, and otherwise following the discussion of Section IV, we have:

$$E(\zeta) = \left\{\frac{\zeta - 2\zeta'}{s_{min}}\right\} + 1, \quad \begin{matrix}\text{number of equatorial satellites} \\ \text{between interleaved pairs}\end{matrix} \tag{54}$$

where the { } denote "the largest integer contained in" the enclosed

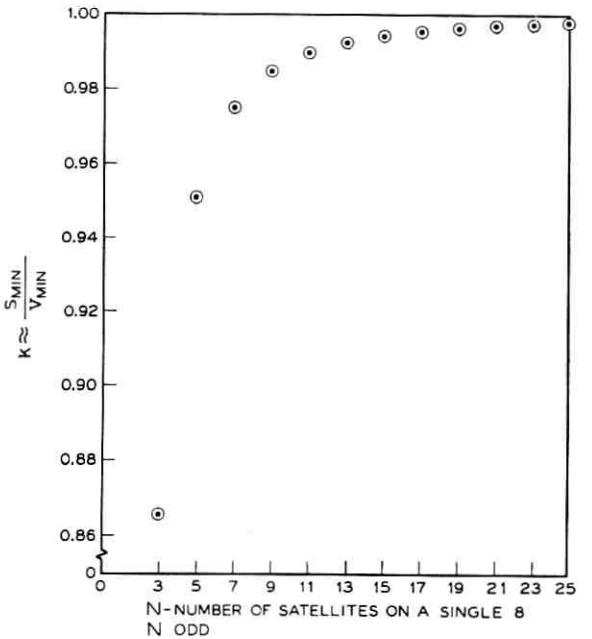


Fig. 23 — Separation reduction factor for two close-spaced interleaved 8's. See equations 51 and 38 through 40.

$\frac{s_{min}}{v_{min}} \approx K$ to graphical accuracy

s_{min} = closest approach for satellites on different 8's

v_{min} = closest approach for satellites on the same 8

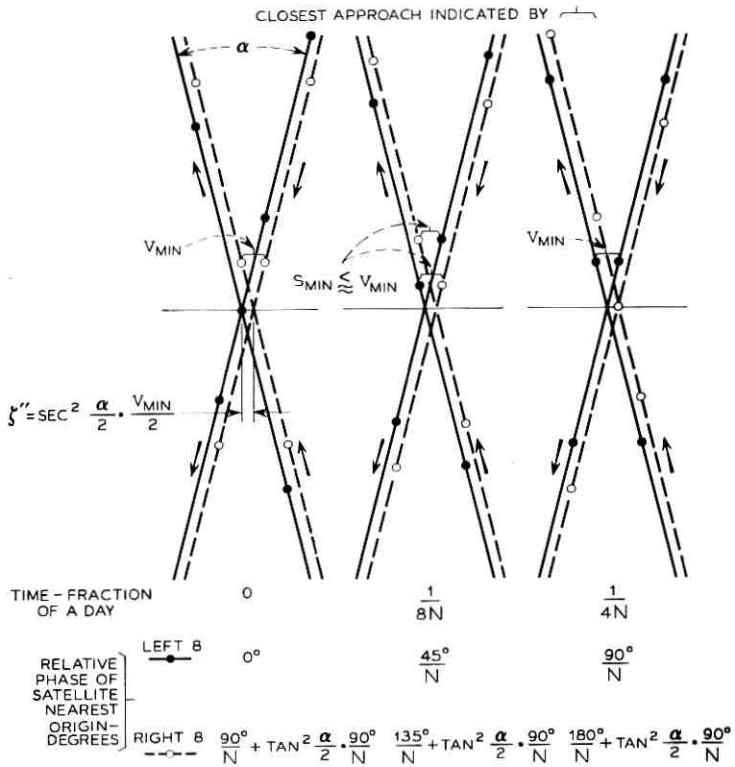


Fig. 24 — Central portion of two close-spaced interleaved 8's. Uniform phase spacing over 360° on each 8. N is odd, N large, $\alpha = 28°$. $\mathrm{Tan}^2\ \alpha/2$ neglected in plotting satellites on right 8.

expression.

$$I(\zeta) = \frac{2N + E}{(\zeta + \zeta'')/s_{\min}} = \frac{(2N + E)s_{\min}}{\zeta + \zeta''}, \qquad \zeta \geqq \zeta_{\min},$$

improvement factor. (55)

$$I_1 = I(\zeta_1), \qquad I_2 = I(\zeta_2); \tag{56}$$

$$\zeta_1 = \zeta_{\min}, \qquad \zeta_2 = 2\zeta' + s_{\min} \cdot E(\zeta_{\min}).$$

I_1 corresponds to minimum spacing between adjacent interleaved pairs, I_2 corresponds to just enough extra spacing to accommodate exactly one more equatorial satellite than with minimum spacing.

$$\sin \frac{\zeta'}{2} = \frac{K \sin^2 \frac{\alpha}{2} \cdot \sin \frac{\pi}{N}}{\cos \frac{\alpha}{2}}. \tag{57}$$

$$\sin \frac{\zeta_{\min}}{2} = \tan^2 \frac{\alpha}{2} \left[1 + K \sin \frac{\pi}{N} \right]. \tag{58}$$

$s_{\min}$ is given by equation 38 and 39 or Fig. 13. In the case of close-spaced interleaved pairs of 8's (that is, using the first "window"), ζ'' is given by equation 50 or 52 or Fig. 22, and K by equation 51 or Fig. 23.

Figure 27 shows the improvement factor in this case in much the same way as Fig. 17 for nonoverlapping 8's. I_1 (corresponding to

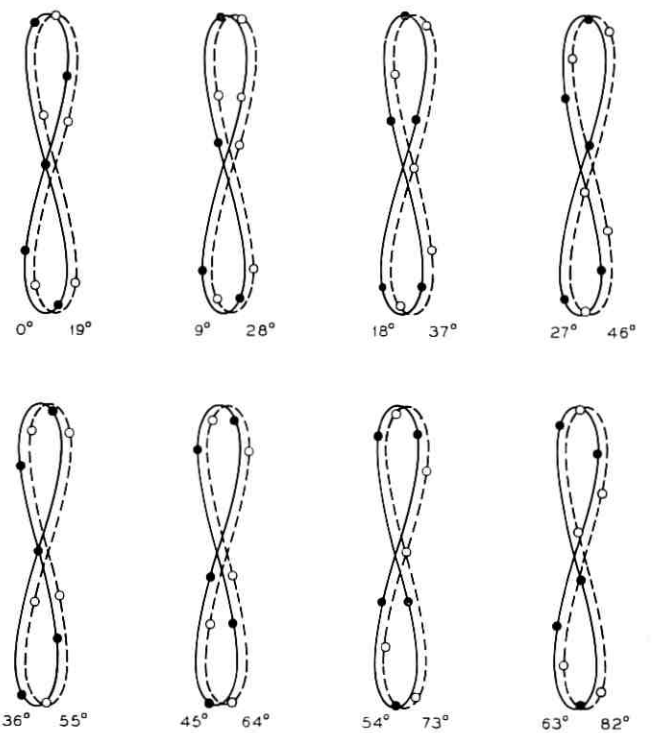


Fig. 25 — Satellite motion for two close-spaced, interleaved 8's. Phases indicated on respective 8's. $N = 5$, $\alpha = 30°$, phase increment = 9°, relative phase between two 8's = 19.03°.

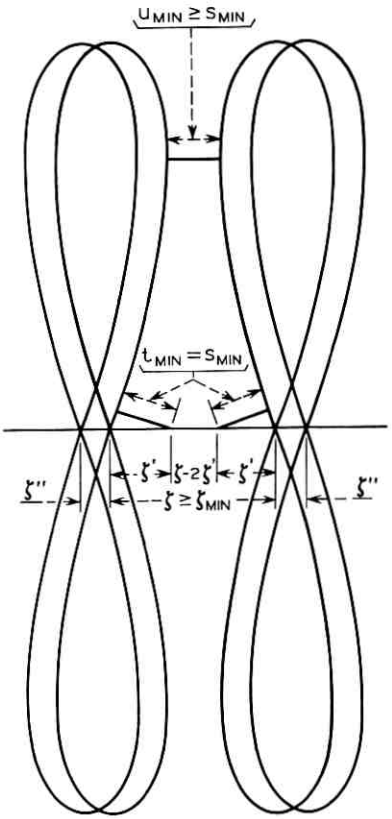


Fig. 26 — Geometry of adjacent interleaved pairs of 8's. $\zeta = \zeta_{min}$ for $u_{min} = s_{min}$. s_{min} = closest approach among $2N$ optimally packed satellites on each interleaved pair of 8's.

minimum separation between adjacent interleaved pairs) is shown by solid lines and I_2 by dashed lines in Fig. 27. However, in contrast to Fig. 17, only the greater of I_1 or I_2, corresponding to the largest improvement factor, is shown. As with the case of Section IV, simpler analytic forms are readily written for $\alpha \to 0$ and $N \to \infty$. We omit the former; the latter yields:

$$\underline{N = \infty}$$

$$I = I_1 = I_2 = 2\pi \frac{\sin^2 \frac{\alpha}{2}}{\sin^{-1} \tan^2 \frac{\alpha}{2}} + 1. \tag{59}$$

Equation 59 shows that $2\pi + 1$ is an upper bound on the improvement factor for the present packing scheme, as compared with an upper bound of $\pi + 1$ for the scheme of Section IV (compare equation 36). Figure 27 shows that this upper bound, like that of Fig. 17 for the prior scheme, may be approached only for extremely small closest approach (very large N); consequently, practical improvement factors will be somewhat smaller.

The packing scheme of Fig. 27 uses the first "window" for interleaving the overlapping 8's. For N satellites per 8, there are $(N-1)/2$ different windows; expressions analogous to those of equations 49 through 53 (for the first window) are readily written for the higher order windows. There are many different possible packing schemes

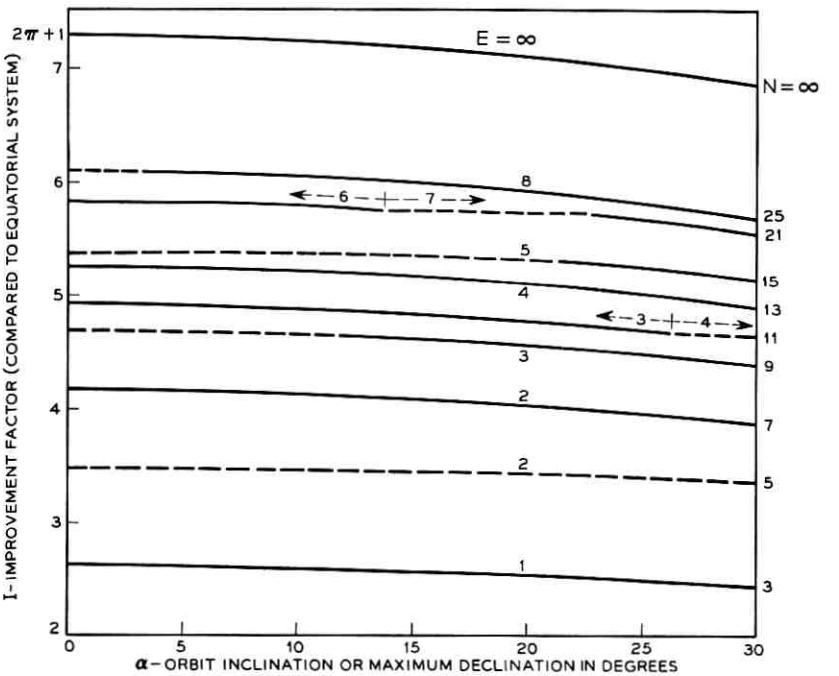


Fig. 27—Improvement factor for optimum satellite packing on separated close-spaced interleaved pairs of 8's. Use this figure in conjunction with Figs. 13, 22, and 23. E = number of satellites on equator between two adjacent interleaved pairs. N = number of satellites on a single 8. N is odd. ——— indicates minimum spacing between adjacent interleaved pairs; closest approach for adjacent pairs = closest approach for a single pair (I_1 of equation 56). – – – indicates increased spacing between adjacent interleaved pairs that just permits 1 additional equatorial satellite (I_2 of equation 56).

using the higher order windows; while it is clear that these have lower improvement factors, some of them may be of interest for other reasons. The improvement factor of this paper is based entirely on closest approach—minimum separation between any pair of satellites. It may pay to degrade this improvement factor in order to increase the average separation between all pairs of satellites in some particular system, for example. Figure 28 shows one possible geometry that will have a somewhat lower improvement factor than the scheme of Fig. 26. The basic relations of this paper permit the study of any of these packing schemes, but we do not pursue this matter further.

In the above examples (Figs. 26 and 27, or 28) only two 8's are effectively interleaved in any given region; consequently, the fairly simple phase-space description of Fig. 21 suffices. We now inquire whether

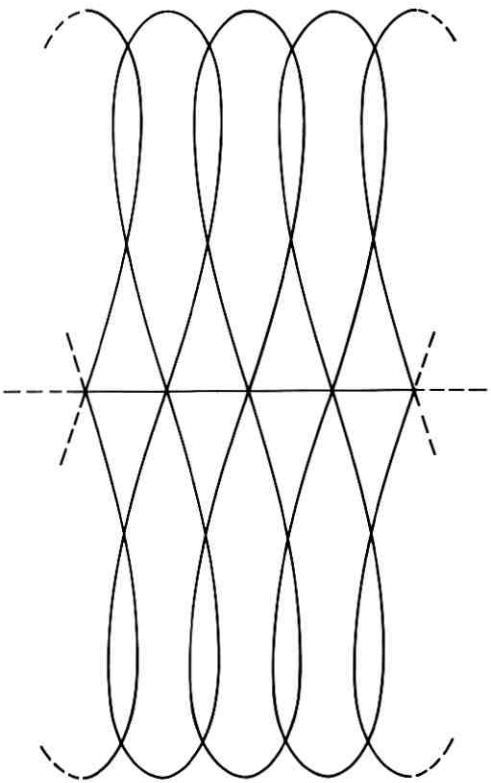


Fig. 28 — Geometry of symmetrically interleaved 8's. Satellites on all 8's must have correct relative phase.

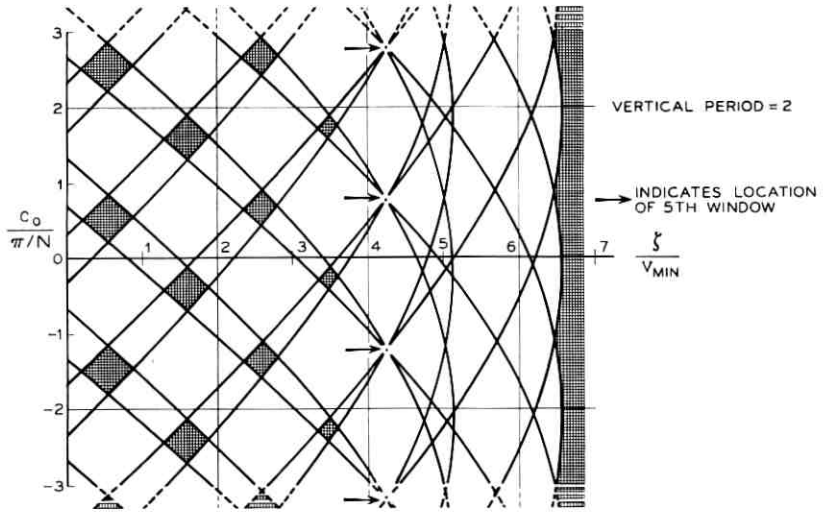


Fig. 29—Windows in the $\zeta - c_0$ plane. $N = 17$, $\alpha = 30°$, $K = 0.674$. K is chosen such that the fifth window is just beginning to open.

	Initial Appearance of Window	
	Window K	$\frac{\zeta}{v_{min}}$
	1 0.996	0.529
	2 0.962	1.569
	3 0.895	2.556
	4 0.798	3.456
Present case →	5 0.674	4.239
	6 0.526	4.878
	7 0.361	5.350
	8 0.184	5.640

greater improvement factors than those of Fig. 27 may be attained by using three or more interleaved 8's. We do not know any general technique for investigating this problem, and investigation of a few special cases chosen at random has not produced any significantly greater improvement factor for parameters in the general region of interest ($N \lesssim 25$). Figure 29 is an alternative picture that casts additional light on the general problem of interleaving 8's, even though its use has not yet led to higher improvement factors than those of Fig. 27.

Figure 29 shows a plot of the "windows," or permitted regions of spacing and relative phase between 8's, for a given value of K, the separation reduction factor (and of course given values $N = 17$ and $\alpha = 30°$). The assumptions here are the same as before in this Sec-

tion, in particular, 360° uniform phase spacing on 8's of identical size (equation 47). Such an 8 may be interleaved with an existing 8 if the equatorial spacing ζ and the relative satellite phase c_0 (equation 48) between the two 8's are such that the corresponding point lies in a shaded region of Fig. 29. For $K = 1$ there are no windows; as K decreases successive windows open up and become larger. As shown by the Fig. 29 table, the first window appears at $K = 0.996$, the second at $K = 0.962, \ldots$, until the fifth appears at $K = 0.674$, corresponding to Fig. 29; as K decreases further, additional windows open up, until all eight are open for $K < 0.184$.

For $\zeta > \zeta_{\min}$ (equation 58) in Fig. 29 for $\zeta/v_{\min} > \zeta_{\min}/v_{\min} = 6.56$, a (nonoverlapping) 8 may be placed with arbitrary relative phase c_0. The border of the shaded region at the right edge of this chart is slightly scalloped; by choosing $c_0/(\pi/N) = 0.85$ the adjacent 8 may be placed a little closer, $\zeta/v_{\min} = 6.45$, but as mentioned in the footnote on page 2412, the potential improvement is so slight (if not nonexistent because it may pay to increase the spacing slightly to accommodate one more equatorial satellite) that we ignore it throughout.

The curves of Fig. 29 are readily obtained from equations 41 through 44 and associated discussion. Four sets of curves must be plotted:

$$c_0 = \begin{bmatrix} \frac{\zeta}{2} - d_- \\ \frac{\zeta}{2} + d_- + \frac{\pi}{N} \\ \frac{\zeta}{2} - d_+ \\ \frac{\zeta}{2} + d_+ + \frac{\pi}{N} \end{bmatrix} + 2(p-1), \quad p = 1, 2, \cdots N; \quad N \text{ odd.} \tag{60}$$

This figure is periodic in the vertical direction, with period $2\pi/N$ for c_0, as a consequence of the 360° uniform phase spacing assumed on each 8.

A chart such as Fig. 29 may be used to consider the interleaving of a number of 8's in the following manner. Take a number of identical charts, cut out the windows, and cut away the paper to the left of the vertical axis. Lay the first chart down on a light-box. Subsequent charts are now laid down on top, with corresponding axes parallel, in such a way that the origin of each chart always lies on a lighted

area. Each chart corresponds to one interleaved 8, so that the total number of charts that can be so laid down equals the number of 8's that can be interleaved for the particular value of K (and of N and α) chosen. The charts must all be translated (with axes parallel), subject to the above constraints, to obtain the maximum total number; the vertical periodicity of these charts greatly reduces the number of possibilities.

After this has been done, if more 8's must be interleaved it will be necessary to repeat the process for a smaller value of K. Since the resulting increase in the number of satellites per unit longitude is accompanied by a decrease in the closest approach $s_{\min}$ (and hence an increase in the denominator of equation 20 defining the improvement factor), it is not obvious, without carrying out this process, whether the improvement factor will be increased or decreased.

It is clear that the process of the preceding two paragraphs does not provide an orderly approach to this problem. A few cases in the region of interest, $N \lesssim 25$, have been tried at random without improving the packing. We recall that since other geometric properties than simply closest approach or minimum satellite separation may be of interest, alternative packing schemes with comparable improvement factors may be of interest; however, we have not pursued these possibilities.

VI. ILLUSTRATIVE SYSTEM SERVING NORTH AMERICA

Figure 30 shows a representative satellite system designed to serve the United States and Canada. While this system has not been optimized, it should serve to illustrate the utility of the above results.

In addition to the general assumptions of Section I, the following general restrictions guided this illustrative design:

(*i*) The most efficient packing scheme that we have devised is used, close-spaced interleaved pairs of 8's (Section V, equation 54 through 59, Figs. 26 and 27).

(*ii*) Closest approach, or minimum separation between any pair of satellites, is about 1°. This would require ground antennas with beamwidths of a fraction of a degree. This might be achieved at frequencies above 12 GHz, possibly in the millimeter band. Such a proposal is somewhat speculative because of the lack of sufficient propagation data.

(*iii*) Minimum elevation above the horizon = 6.4°.

(*iv*) Space in orbit is roughly allocated to what seems the best use.

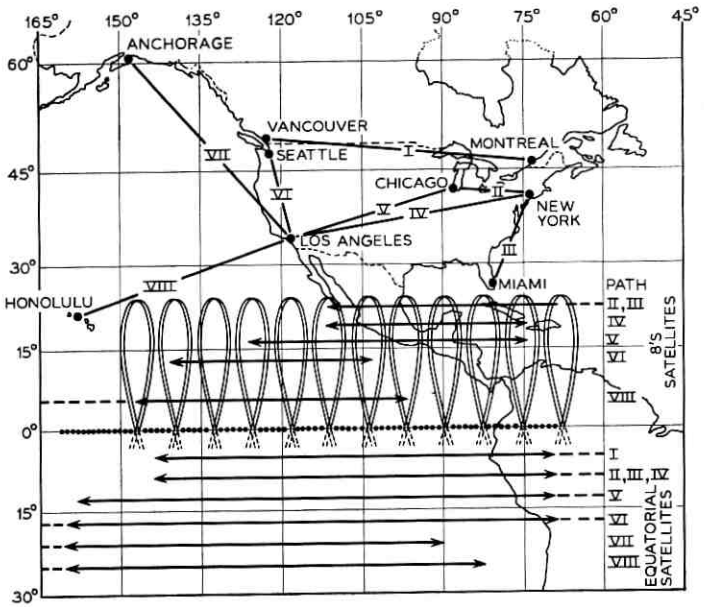


Fig. 30 — Satellite system for North American continent. Total satellites = 477; total in equatorial system occupying same longitude = 95. Packing scheme: close-spaced interleaved pairs of 8's, at minimum spacing ($\zeta = \zeta_{\min}$ in Fig. 26). Design parameters: $\alpha = 25°$, orbit inclination. $v_{\min} = 0.986°$, $s_{\min} = 0.982°$, closest approach. $N = 17$, number of satellites on a single 8 (individual satellites not indicated). $E = 5$, number of equatorial satellites between adjacent pairs of 8's. $I_1 = 5.34$, improvement factor (Fig. 27). Minimum elevation = 6.4°. 8's and equatorial satellites that can serve representative paths without allowing elevation to decrease below 6.4° shown by horizontal arrows, determined from Fig. 32.

Thus, space suitable for trans-Atlantic or trans-Pacific communication has generally not been used. Satellites have been placed farther from the United States coast over the Pacific than over the Atlantic because the Pacific is wider.

(*v*) 8's have been omitted where they would not appear needed (over Hawaii), on the basis of a crude estimate about relative traffic density.

(*vi*) In such a system each satellite might serve every ground station that it can see, within limits imposed by the resolution of the satellite antennas.[1] Referring to Fig. 2, for the minimum elevation chosen for this example (6.4°), an orbit inclination $\alpha = 30°$ allows no tolerance at all in longitude for a ground station at 45° north latitude, somewhat south of the western United States-Canada border.

Consequently, a somewhat smaller inclination ($\alpha = 25°$) has been chosen in Fig. 30.

Figure 31 shows the region of visibility of a satellite from ground stations at several different north latitudes to the same scale as Fig. 30. Equatorial satellites and satellites at 25° south latitude, of inter-

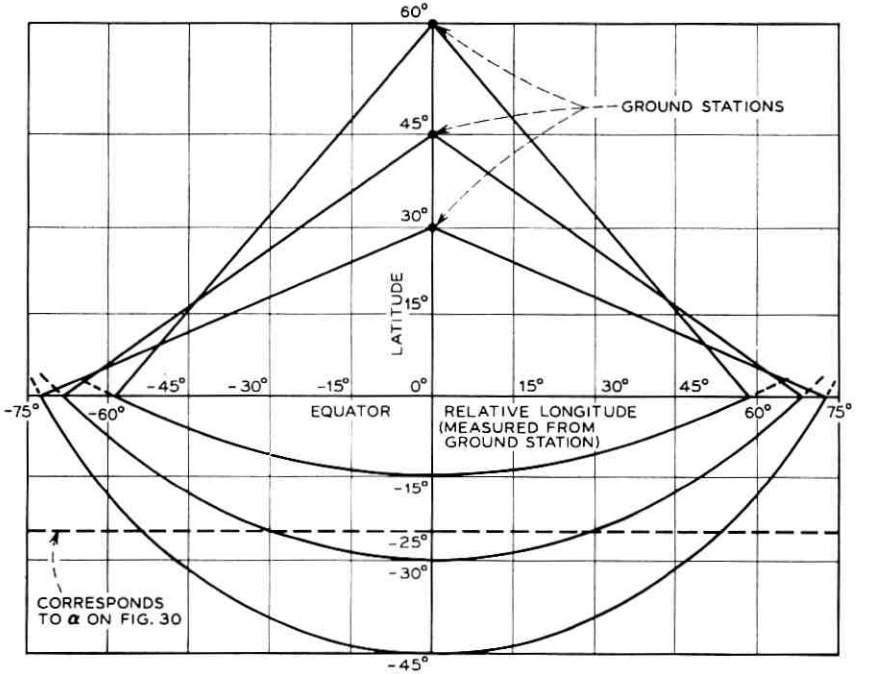


Fig. 31 — Region of visibility of a satellite from ground stations at representative north latitudes, assuming minimum elevation of 6.4° above the horizon. Same scale as Fig. 30, so that these curves give directly permitted 8's and equatorial satellites for representative paths on Fig. 30.

est in Fig. 30, are shown in more detail on Fig. 32.* Using these data, the ranges of 8's and of equatorial satellites that can serve eight representative paths have been shown by the horizontal arrows on Fig. 30.

Two general conclusions are readily apparent:

(*i*) The equatorial satellites are more versatile than the 8's satel-

* The derivation of these results, which is elementary, is given in Appendix B for convenience.

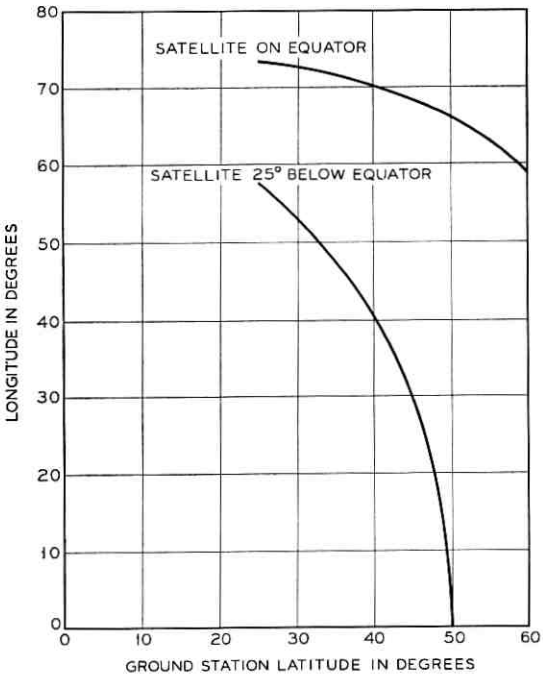


Fig. 32 — Maximum relative longitude from ground station for which satellite is visible 6.4° above the horizon.

lites; only the former can serve paths north of the United States-Canada border.

(*ii*) The most versatile of the 8's satellites lie over the central United States; the further from the center, the smaller the variety of paths served.

Assuming that each satellite carries transponders to serve all ground stations that it can see, the equatorial satellites will be the heaviest, the central 8's satellites next heaviest, and the edge 8's satellites the lightest.

It is obvious that the more easterly satellites of Fig. 30 can also serve Central and South America (with additional transponders for this purpose).

We re-emphasize that this proposed system is only for illustration. Much detailed study would be required to choose the best parameters, even within the assumptions imposed here.

VII. DISCUSSION

The best packing scheme in this paper yields an improvement factor of roughly six over a purely equatorial system in the range of probable interest (see Fig. 27). This result is based on a great many assumptions, already stated. We believe that the relaxation of all of these assumptions should be investigated further. While we have tried to make reasonable choices, we have no assurance either that the improvement factors of Fig. 27 are the best that can be obtained, or that further work will yield better packing schemes (or alternative packing schemes of interest for other reasons).

The treatments of optimum packing a single 8 (Section III) and separated 8's (Section IV) are virtually complete.* However, the treatment of overlapping 8's in Section V, which yields the largest improvement factor, is far from complete. We do not know how to approach this latter problem in a general way. As a start, it seems possible that extending the present treatment of Section V to 8's of moderately different sizes (α's) might allow the interleaving of more than two 8's with significantly greater improvement factors.

All of the discussion has been based on the assumption that each satellite remains in view of every ground station that it serves and no switching is required. This represents one extreme; the other extreme is a low-altitude system, with frequency switching. Intermediate systems are also possible, with switching several times a day. If the assumption is relaxed to allow this, other quite different systems become possible. Two considerations are:

(*i*) Larger 8's (with a larger inclination α) may be used; whether this would lead to any advantage for the United States remains to be studied. However, Europe is enough farther north than the United States that figure 8 systems may not be useful unless occasional switching is allowed.

(*ii*) The antenna scanning and beam shaping problems may be eased. For example with 360° packing on each individual 8, as assumed in the interleaved packing schemes of Section V, each beam would need to scan a much smaller angle.

Once occasional switching is allowed, lower-than-synchronous orbits may be considered. All of the packing schemes discussed above still apply, being based only on circular orbits and not on how fast the

* Not all items of interest have been investigated, but we believe the basic relations given here are sufficient for most purposes.

earth happens to be rotating. However, now the satellite would have to extend all the way around the earth, rather than say one quarter of the way around as in Fig. 30 for synchronous orbits. Further, the mutual visibility problem becomes worse as the altitude decreases; additional study would be required to determine the resulting improvement factors.

In Sections IV and V different 8's and different interleaved pairs of 8's, respectively, were allowed to have arbitrary relative phase. We now inquire what use can be made of this additional parameter. As already mentioned, slightly closer packing and hence slightly greater improvement factors can be obtained in some cases by the correct choice of relative phase, although this increase is small and will yield only a few additional satellites in the illustrative system of Fig. 30.

A more interesting possibility is to use this parameter to reduce the number of different orbits required, so that one vehicle can launch many different satellites lying on the same orbit but on different 8's. From the basic definitions of Fig. 4, it is obvious that two satellites lying on the same orbit but on different 8's of identical size (α), spaced along the equator by ζ, have a relative phase of ζ. The $2 \times 12 \times 17 = 408$ 8's satellites of Fig. 30 generally require 408 orbits. By proper choice of relative phase ($\zeta + \zeta'' =$ phase difference between adjacent pairs—see Fig. 26), only 34 different orbits are required; thus only 34 launch vehicles, each carrying 12 satellites, need be used.

A study of the fuel requirements to maintain desired tolerances in satellite positions for the different systems is needed. It is not obvious whether this requirement is greatly different for the different systems.

VIII. ACKNOWLEDGMENT

The authors thank L. C. Tillotson for suggesting this problem and for many helpful discussions, and Mrs. Evelyn Kerschbaumer for programming computer movies illustrating satellite motion.

APPENDIX A

Derivation of Fundamental Results of Section II

Consider on Fig. 4 the isosceles (spherical) triangle composed of the equator, the orbit, and the great circle connecting the satellite to the earth reference point (at the intersection of the reference longitude and the equator). This triangle has two equal sides of (great circle) length

c (measured by the angle subtended at the center of the sphere) with included angle α, and a third side of length a; let the remaining two equal angles be denoted by γ (not indicated on Fig. 4). Noting that $\gamma + \epsilon = \pi/2$, the law of sines yields

$$\frac{\sin \alpha}{\sin a} = \frac{\cos \epsilon}{\sin c}. \tag{61}$$

Now bisect the above triangle into two equal right triangles by bisecting the angle α (with a great circle). From the law of sines

$$\frac{\sin \frac{\alpha}{2}}{\sin \frac{a}{2}} = \frac{1}{\sin c}, \tag{62}$$

which yields equation 1. We now use equation 1 to eliminate the parameter a from equations 61 by means of the double-angle formula, yielding

$$\cos \epsilon = \left[\frac{1 - \sin^2 \frac{\alpha}{2}}{1 - \sin^2 \frac{\alpha}{2} \cdot \sin^2 c} \right]^{\frac{1}{2}}. \tag{63}$$

Equation 2 now follows directly.

Equation 3 follows from the law of sines applied to the right triangle of Fig. 4 whose sides are the orbit, the equator, and the dotted longitude passing through the satellite. The law of cosines applied to this triangle yields

$$\cos c = \cos l \cdot \cos e. \tag{64}$$

This becomes

$$\sin e = \left(\frac{\sin^2 c - \sin^2 l}{1 - \sin^2 l} \right)^{\frac{1}{2}}; \tag{65}$$

substituting equation 3 into equation 65 yields equation 5.

Consider next the right triangle formed by the equator, the dotted longitude, and the great circle joining the satellite to the earth reference point (defined in the first sentence of this appendix) in either Fig. 4 or 6. The law of cosines yields

$$\cos a = \cos l \cdot \cos \psi. \tag{66}$$

Thus

$$\sin \psi = \left(\frac{\sin^2 a - \sin^2 l}{1 - \sin^2 l}\right)^{\frac{1}{2}}. \tag{67}$$

Using the double-angle formula, substituting equations 1 and 3 into equation 67 yields equation 4.

Consider now the right triangle of Fig. 6 formed by the reference longitude, the great circle connecting the satellite to the earth reference point (see above), and the great circle of length x passing through the satellite and normal to the reference longitude. From the law of sines

$$\sin x = \sin a \cdot \sin \epsilon. \tag{68}$$

Using equations 1 and 2 to eliminate α and ϵ, together with the double angle formula, equation 7 is readily obtained.

We next derive the general result of equation 11, obtaining all the remaining results of Section II by specializing this relation. Applying the law of cosines to the spherical triangle of Fig. 9 whose vertices are the north pole and the satellites on each of the two 8's,

$$\cos s = \sin l_1 \cdot \sin l_2 + \cos l_1 \cdot \cos l_2 \cdot \cos (\zeta + \psi_1 - \psi_2). \tag{69}$$

Make the following substitutions in equation 69:

(*i*) equation 3.

(*ii*) Expand $\cos (\zeta + \psi_1 - \psi_2)$.

(*iii*) From equations 66 and 1, and the double-angle formula,

$$\cos l \cdot \cos \psi = 1 - 2 \cdot \sin^2 \frac{\alpha}{2} \cdot \sin^2 c$$

$$= \cos^2 \frac{\alpha}{2} + \sin^2 \frac{\alpha}{2} \cdot \cos 2c$$

(*iv*) From equations 3 and 4

$$\cos l \cdot \sin \psi = \sin^2 \frac{\alpha}{2} \cdot \sin 2c.$$

After lengthy but straightforward transformations equation 11 is obtained. We readily obtain equation 8 (see Fig. 6) by the second transformation in the table following equation 11. For equation 10, the fourth transformation of this table yields

$$\cos u = \cos \zeta \cdot \cos^4 \frac{\alpha}{2} + \frac{\sin^2 \alpha}{2} [\cos (2c_2 - \zeta) - \cos 2c_2] + \frac{\sin^2 \alpha}{2} + \sin^4 \frac{\alpha}{2} \cdot \cos (4c_2 - \zeta). \quad (70)$$

Make the following substitutions in equation 70:

(*i*) $\cos (2c_2 - \zeta) - \cos 2c_2 = 2 \sin \left(2c_2 - \frac{\zeta}{2}\right) \cdot \sin \frac{\zeta}{2}$.

(*ii*) $\cos \sigma = 1 - 2 \sin^2 \frac{\sigma}{2}$; $\sigma = u, \zeta, 4c_2 - \zeta$.

(*iii*) $\frac{\sin^2 \alpha}{2} = 2 \sin^2 \frac{\alpha}{2} \cdot \cos^2 \frac{\alpha}{2}$.

We have after combining terms

$$\sin^2 \frac{u}{2} = \left[\cos^2 \frac{\alpha}{2} \cdot \sin \frac{\zeta}{2} - \sin^2 \frac{\alpha}{2} \cdot \sin \left(2c_2 - \frac{\zeta}{2}\right)\right]^2, \quad (71)$$

simply the square of equation 10.

To derive equation 13, substitute equation 12 into equation 11 and use the double-angle formula on the α-dependent factors of the third and fourth terms to yield, after some minor rearrangement,

$$\cos s = \cos \zeta \cdot \cos^4 \frac{\alpha}{2} + \sin^4 \frac{\alpha}{2} \cdot \cos (2c_0 - \zeta) + \frac{\sin^2 \alpha}{2} \cdot \cos c_0 + \frac{\sin^2 \alpha}{4} [\cos (2c_1 + c_0 - c_0 + \zeta) + \cos (2c_1 + c_0 + c_0 - \zeta) - 2 \cos (2c_1 + c_0)]. \quad (72)$$

The three terms inside [ ] may be regarded as an AM wave and so are readily combined, to yield

$$\cos s = \cos^4 \frac{\alpha}{2} \cdot \cos \zeta + \sin^4 \frac{\alpha}{2} \cdot \cos (2c_0 - \zeta) + \frac{\sin^2 \alpha}{2} \cdot \cos c_0 - \sin^2 \alpha \cdot \sin^2 \left(\frac{c_0 - \zeta}{2}\right) \cdot \cos (2c_1 + c_0). \quad (73)$$

Make the following substitutions in equation 73:

(*i*) $\cos \sigma = 1 - 2 \sin^2 \frac{\sigma}{2}$; $\sigma = s, \zeta, 2c_0 - \zeta, c_0$.

$$(ii) \qquad \cos(2c_1 + c_0) = 2\cos^2\left(c_1 + \frac{c_0}{2}\right) - 1.$$

$$(iii) \qquad \frac{\sin^2\alpha}{2} = 2\sin^2\frac{\alpha}{2}\cdot\cos^2\frac{\alpha}{2}.$$

Equation 13 is obtained directly. Equation 9 follows immediately, as pointed out in the text.

There are of course a great many alternate ways of deriving these various results.

APPENDIX B

Visibility of a Satellite from a Ground Station

Figure 33 shows the earth, of unit radius, concentric with a sphere of radius R, equal to the radius of a circular satellite orbit. For syn-

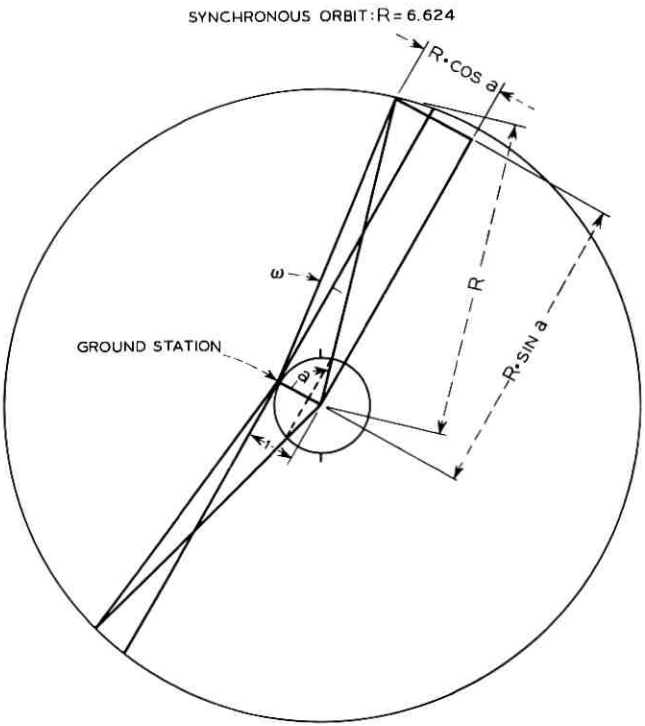


Fig. 33 — Region of visibility of a satellite. Synchronous orbit: $R = 6.624$.

chronous orbits $R = 6.624$. The elevation of a satellite viewed from a ground station must exceed the angle ω; the satellite is restricted to a region of angle from the ground station (measured at the center of the spheres) less than a. Projecting the satellite on a sphere (the earth in Fig. 33), the satellite is restricted to a region within a circle centered on the ground station (shown dashed in Fig. 33) of (great circle) radius a, with the usual convention of spherical trigonometry that great circle distances are measured by the angle they subtend at the center of the sphere. From Fig. 33,

$$\tan \omega = \frac{R \cdot \cos a - 1}{R \cdot \sin a} = \cot a - \frac{\csc a}{R}. \tag{74}$$

Solving for a,

$$\sin a = \cos \omega \left[1 - \left(\frac{\cos \omega}{R} \right)^2 \right]^{\frac{1}{2}} - \frac{\sin 2\omega}{2R}. \tag{75}$$

At synchronous radius,

$$\sin a = \cos \omega (1 - .0228 \cos^2 \omega)^{\frac{1}{2}} - 0.0755 \sin 2\omega,$$
$$R = 6.624. \tag{76}$$

The relation between latitude l and relative longitude ψ on the dotted circle is found from Fig. 34. The ground station is located at north latitude l_0. From the law of sines,

$$\cos l \cdot \sin \psi = \sin a \cdot \sin \epsilon. \tag{77}$$

From the law of cosines

$$\sin l = \cos a \cdot \sin l_0 + \sin a \cdot \cos l_0 \cdot \cos \epsilon. \tag{78}$$

Substituting equation 76 into equations 77 and 78, the results of Figs. 31 and 32 are readily computed parametrically in terms of the azimuth ϵ.

APPENDIX C

Error in Satellite Separation

The error in separation between satellites resulting from displacement from the center of the earth is readily calculated. Fig. 35 shows the earth and a pair of satellites in circular orbits of radius R, separated at this instant by angle s subtended at the center of the earth. The side view shows a ground station; the plane of the paper includes

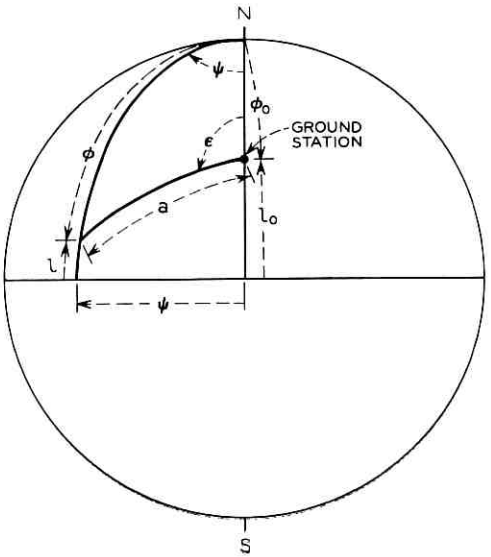


Fig. 34 — Latitude and longitude of region of visibility.

the center of the earth, the ground station, and the upper satellite. The poles of the earth are not necessarily vertical in this figure. The lower satellite does not in general lie in the plane of the paper; the end view shows a line joining the two satellites, making an angle τ with the plane of the upper satellite, ground station, and center of the earth.

In the case of interest here the satellites lie close together. Con-

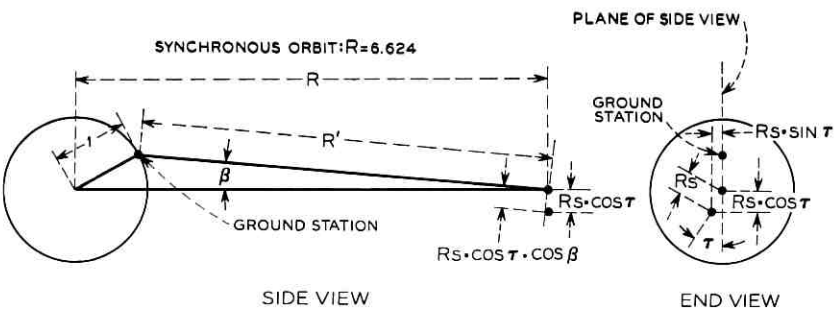


Fig. 35 — Satellite separation as a function of ground station location. s = angular separation between satellites at center of earth. Distance between satellites (Rs) and its projections are true (not angular) lengths.

sequently we assume throughout this appendix that

$$s \ll 1. \tag{79}$$

The ground station is restricted to the region of the earth where the satellite remains visible, the limiting case corresponding to zero elevation.* Consequently the angle β of Fig. 35 is bounded by

$$\beta < \beta_{max}, \qquad \sin \beta_{max} = 1/R. \tag{80}$$

The distance R' from the ground station to the satellite has maximum and minimum values corresponding to maximum and minimum values of β. Thus

$$1 - \frac{1}{R} < \frac{R'}{R} < \cos \beta_{max} = \left(1 - \frac{1}{R^2}\right)^{\frac{1}{2}}. \tag{81}$$

We seek the angular separation between the two satellites as seen from the ground station; let this quantity be denoted by s'. We project the line segment joining the two satellites onto the plane perpendicular to the line R' joining the satellites and the ground station as shown. Then

$$s'/s = (R/R')[1 - (\sin \beta \cdot \cos \tau)^2]^{\frac{1}{2}}. \tag{82}$$

We have

$$1 < s'/s < \frac{R}{R-1}. \tag{83}$$

For synchronous orbit

$$1 < s'/s < 1.178. \tag{84}$$

Thus as stated in Section I, for synchronous orbit the approximation of this paper which determines separation between satellites as seen from the center of the earth (rather than from ground stations) ranges from correct to 18 per cent too conservative for the critical cases of closest approach.

REFERENCE

1. Tillotson, L. C., "A Model of a Domestic Satellite Communication System," B.S.T.J., this issue, pp. 2111–2137.

* Practically the minimum elevation must be greater than zero, as in Appendix B. A minimum elevation of 6.4° was assumed in Section VI.

Contributors to This Issue

Bruce E. Briley, B.S., 1958, M.S., 1959, Ph.D., 1963, University of Illinois; Bell Telephone Laboratories, 1966—. Mr. Briley has worked on the design of an exploratory small electronic telephone office, and is now doing exploratory work in telephone switching systems processors. He is past chairman of the Chicago chapter of the IEEE Computer Group, member of the Logical Design Subcommittee of the Computer Group, and Graduate Lecturer at the Illinois Institute of Technology. Member, IEEE, ACM, Sigma Xi, Eta Kappa Nu, AAAS.

Detlef Gloge, Dipl. Ing., 1961, D.E.E., 1964, Braunschweig Technische Hochschule (Germany); research staff, Braunschweig Technische Hochschule 1961–1965; Bell Telephone Laboratories, 1965—. In Braunschweig, Mr. Gloge was engaged in research on lasers and optical components. At Bell Telephone Laboratories, he has concentrated in the study of optical transmission techniques. Member, VDE, IEEE.

Stephen C. Johnson, A.B., 1963, Haverford College; M.A. and Ph.D., 1968, Columbia University; Bell Telephone Laboratories, 1967—. Mr. Johnson has been doing research in the applications of computers to acoustical sound generation, symbolic computation, and psychometrics. Member, Sigma Xi, American Mathematical Society, AAAS, Phi Beta Kappa.

Bela Julesz, Dipl. in E.E., 1950, Budapest (Hungary) Technical University; Kandidat in Technical Sciences, 1956, Hungarian Academy of Sciences; Telecommunication Research Institute (Budapest) 1950–56; Bell Telephone Laboratories, 1956—. He first taught and did research in communication systems and his thesis work reflected his later interest in analyzing and processing pictorial information. At Bell Laboratories he was first engaged in studies of systems for reducing television bandwidth. Since 1959 he has devoted full time to visual research, particularly in depth perception and pattern recognition, about which he has written extensively. Since 1964 Dr. Julesz has been Head of the Sensory and Perceptual Processes Department, responsible for research in visual psychology and neurophysiology. Member IEEE, AAAS, Psychonomic Society, Optical Society of America.

W. A. KESTER, B.S.E.E., 1964, North Carolina State University; M.S.E.E., 1966, Duke University; Bell Telephone Laboratories, 1964—. Mr. Kester initially was involved in the design of high-speed circuits associated with analog-to-digital converters for the Nike-X program. He later participated in the development of the Bell System Reference Frequency Standard and is now engaged in work associated with an exerciser for the Sentinel system. Member, IEEE, Eta Kappa Nu, Phi Kappa Phi, Tau Beta Pi.

S. C. LIU, B.S. in C.E., 1960, National Taiwan University; M.S., 1964, and Ph.D., 1967, University of California at Berkeley; Bell Telephone Laboratories 1967—. Mr. Liu has been doing research is applied mechanics, structural dynamics, random vibrations and earthquake engineering. Member: American Society of Civil Engineers, Seismological Society of America.

ANTHONY G. LUBOWE, A.B., 1957; B.S. in M.E., 1958; M.S., 1959; Eng. Sc.D. in Engineering Mechanics, 1961, all from Columbia University; Bell Telephone Laboratories, 1961—. He has worked on methods of orbit determination and prediction used for the *Telstar*® communications satellite experiment, the Apollo project, and for NASA and Department of Defense studies. Recently he has been concerned with problems of satellite dynamics arising in the Sentinel project. Associate Fellow, A.I.A.A.; member, A.S.M.E., Tau Beta Pi, Phi Beta Kappa.

ELLIOTT R. NAGELBERG, B.E.E., 1959, City College of New York; M.E.E., 1961, New York University; Ph.D., 1964, California Institute of Technology; Bell Telephone Laboratories, 1964—. Mr. Nagelberg is Supervisor of the Electromagnetics Group and is responsible for research on the transmission and radiation aspects of microwave communication systems. Member, IEEE, American Physical Society, Eta Kappa Nu, Sigma Xi, AAAS.

G. PASTERNACK, B.S.E.E., 1966 and M.S., 1967, both from Polytechnic Institute of Brooklyn; Bell Telephone Laboratories, 1961—. Mr. Pasternack was initially engaged in the development of a *Touch-Tone*® receiver for use with key telephone systems. He later became concerned with the exploratory study of digital methods for data communication. Most recently he did development work on a low cost,

line powered data set, and is now doing exploratory work on an all digital multiple data set. Associate Member, Sigma Xi; Member IEEE.

ARNO A. PENZIAS, B.S. 1954, City College of New York; M.A. 1958, Ph.D. 1962, Columbia University; Bell Telephone Laboratories 1961—. A radio astronomer, Mr. Penzias has been concerned mainly with problems relating to cosmology. He has also done work on antenna pointing and calibration, microwave noise measurement, and precipitation effects on atmospheric transmission. He is now working on satellite antenna problems. Member, American Physical Society, American Astronomical Society, International Scientific Radio Union, Phi Beta Kappa.

DANIEL L. POPE, B.C.E., 1953; Ph.D. (Mechanics), 1961, Cornell University; Bell Telephone Laboratories, 1960—. Mr. Pope has been concerned with various aspects of defensive system analysis. He has studied the role of non-nuclear warheads and done structural analysis of various components of high performance missiles. He participated in the early phases of orbital mechanics studies for communication satellites. He has been involved in the development of advanced methods for structural analysis and worked on applying such techniques in antenna design and other fields. He supervises a continuum mechanics group in the Analytical Mechanics Department. Member, Chi Epsilon, Society of Industrial and Applied Mathematics, Tau Beta Pi.

ATTILIO J. RAINAL, University of Alaska, University of Dayton, 1950–52; B.S.E.Sc., 1956, Pennsylvania State University; M.S.E.E., 1959, Drexel Institute of Technology; Dr. Eng., 1963, Johns Hopkins University; Bell Telephone Laboratories, 1964—. He has been engaged in research on noise theory with application to detection, estimation, and radar theory. Member, Tau Beta Pi, Eta Kappa Nu, Sigma Tau, Pi Mu Epsilon, Sigma Xi, IEEE.

HARRISON E. ROWE, B.S., 1948, M.S., 1950, Sc. D., 1952, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1952—. His fields of interest have included parametric amplifier theory, noise and communication theory, propagation in random media, and related problems in waveguide, radio, and optical systems. Member, IEEE, Sigma Xi, Tau Beta Pi, Eta Kappa Nu.

MICHAEL G. TAYLOR, B.A.Sc., 1961, and M.A.Sc., 1962, University of British Columbia, Vancouver, Canada; Ph.D., 1966, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1967—. Mr. Taylor has worked in the areas of digital data processing and communication theory. He is concerned with designing adaptive digital filters for use in digital data transmission systems. Member, IEEE, Tau Beta Pi, Sigma Xi.

LEROY C. TILLOTSON, B.S.E.E., 1938, University of Idaho; M.S. in EE, 1940, University of Missouri; D.Sc. (Hon.), 1966, University of Idaho; Bell Telephone Laboratories, 1941—. Mr. Tillotson has worked on the design of filters and radio systems. He is Director of Radio Research and has been engaged in research on radio and optical wave propagation and systems for terrestrial and satellite applications. He spent more than a year on a leave of absence with the Institute for Defense Analysis, Washington, D. C., assigned to the Advanced Research Projects Agency, where he was concerned with communication satellites and other space-related actives.

DANIEL WEINER, M.S. (physics), 1957, Ph.D. (physics), 1961, University of Chicago; Bell Telephone Laboratories 1966—. Mr. Weiner has done work on high resolution transmission scanning electron microscopy and long range laser communication systems. Member, American Physical Society, Sigma Xi, Phi Beta Kappa.

RONALD L. WHALIN, B.S.M.E., 1959, New Mexico State University; M.S.E.M., 1961, New York University; M.Sc. (E.E.), 1968, Rutgers University; Bell Telephone Laboratories, 1959—. Mr. Whalin has been responsible for circuit and physical design of low-speed serial FM data sets. He is supervisor of the DDD Data Sets and Multiline Systems physical development group. Member IEEE, Sigma Tau, Pi Tau Sigma.