THE BELL SYSTEM

Technical Journal

Volume 51 March 1972 Number 3

THE BELL SYSTEM TECHNICAL JOURNAL

THE BELL SYSTEM TECHNICAL JOURNAL is published ten times a year by the American Telephone and Telegraph Company, J. D. deButts, Chairman and Chief Executive Officer, R. D. Lilley, President, J. J. Scanlon, Vice President and Treasurer, R. W. Ehrlich, Secretary. Checks for subscriptions should be made payable to American Telephone and Telegraph Company and should be addressed to the Treasury Department, Room 2312C, 195 Broadway, New York, N. Y. 10007. Subscriptions $10.00 per year; single copies $1.25 each. Foreign postage $1.00 per year; 15 cents per copy. Printed in U.S.A.

THE BELL SYSTEM

TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

Volume 51 March 1972 Number 3

Power Coupling from GaAs Injection Lasers into Optical Fibers

By L. G. COHEN

(Manuscript received October 4, 1971)

Measurements have been made to determine the efficiency of coupling light from GaAs injection lasers (λ = 9000 Å) of stripe geometry into the cores of optical fibers. Laser light was coupled into a fiber across a small air gap which separated the laser from the fiber. The power coupling efficiency was calculated by extrapolating the lasing light power emitted at the end of a short length of fiber back to its input tip and comparing it with the total laser beam power. Experiments were performed with several diffused junction (DJ) lasers and with low-threshold double-heterostructure (DH) lasers having stripes formed by proton bombardment or oxide masking. The beam profile, scanned in its near and far field region, was approximately Gaussian. The beam dimensions at the surface of the laser were estimated from far field measurements and were used to predict the coupling efficiency of single-mode and multimode fibers. Power coupling efficiencies of about 70 percent were measured for DJ lasers feeding 10-μm multimode fibers. A coupling efficiency of 25 percent, almost identical to the theoretical estimate, was achieved with a DH laser having a 1-μm × 13-μm proton-bombarded stripe, feeding the 3.2-μm core of a single-mode fiber. The coupling efficiency was greater than 40 percent when a DJ laser fed the same fiber. A cylindrical lens is proposed to increase the power coupling perpendicular to the junction plane.

Permanent self-supporting couplers were made by applying epoxy between structures which supported the fiber and the laser.

I. INTRODUCTION

This paper describes experiments performed to determine the efficiency of coupling light power from GaAs injection lasers* ($\lambda \approx$ 9000 Å) into the cores of 3.2-μm and 3.7-μm single-mode fibers and several sizes of multimode fibers.† At one end of a fiber the light from a laser was coupled in across a small air gap which separated the laser from the fiber. The power coupling efficiency was computed by extrapolating the lasing light power emitted from a 30-cm length of fiber back to its input tip and comparing it with the total laser power. Permanent coupling of injection lasers and fibers was accomplished by applying epoxy between structures which supported the fiber and the GaAs chip. One other coupling technique was tried. It involved bonding (with epoxy) a fiber onto one of the laser mirrors. However, this technique was abandoned because application of epoxy to the active stripe on a GaAs chip raised the lasing threshold by 30 percent. Only 30 percent of this rise was attributable to the decrease in mirror reflectivity at the epoxy interface. The remaining discrepancy may have been caused by strains introduced to the mirror as the epoxy hardened.

The body of this paper is divided into three sections. In Section 2.1 we describe measurements to determine the properties of light beams emanating from double-heterostructure (DH) lasers and also from diffused junction (DJ) lasers of striped geometry. Guiding stripes on DH lasers were formed by oxide masking or proton bombardment. In proton-bombarded DH lasers the current is confined within a stripe by high-resistivity regions produced by proton bombardment at the edges of the stripe. The details of the fabrication and performance of proton-bombarded lasers are discussed in Ref. 1. Section 2.2 describes power coupling measurements for lasers feeding single-mode and multimode fibers. In Section III we propose a technique to increase the power coupling efficiency by attaching a cylindrical lens onto the fiber tip. The lens was designed to collimate the laser beam perpendicular to the junction plane in order to transform its cross section from a rectangular to a square shape. Finally, in Section IV we indicate how theoretical estimates of power coupling efficiency were made based on the measured properties of the laser beam.

* The lasers were fabricated by J. C. Dyment at Bell Laboratories, Murray Hill, N. J.

† The 3.7-μm fiber was fabricated by Corning Glass Works of Corning, N. Y. All the other fibers were drawn from glass preforms by DeBell and Richardson, Inc., of Hazardville, Conn.

II. MEASUREMENTS

Experiments have been performed to determine the efficiency for coupling light ($\lambda \approx 9000$ Å) into optical fibers from the radiation field of GaAs injection lasers. Laser light emanated from a rectangular aperture on the cleaved surface of the GaAs chip. The beam dimensions were about 3 μm $\times$ 6 μm for DJ lasers and about 1 μm $\times$ 3 μm for DH lasers. The GaAs laser chip was held, by spring contact, with its p-side face down on a copper block and was driven at room temperature at a 100-Hz repetition rate by negative current pulses, 100 nanoseconds wide, which were applied between the spring and the grounded copper block. Peak power measurements were displayed on a storage sampling oscilloscope.

2.1 *Laser Field Measurements*

The properties of the light beam emanating from the lasing aperature were determined by scanning its near and far field radiation patterns.

Far field measurements were made by rotating a laser about one of its cleaved surfaces in the plane parallel to the laser's junction plane and also in the plane perpendicular to it. The power distribution in the beam was measured as a function of angle, 1.5 inches from the laser, through a 10-mil slit mounted on a photomultiplier (refer to Fig. 1). The shape of the profiles was independent of the slit width and the axial separation between laser and slit.

Near field measurements[2] were made by using a lens to magnify the light pattern on the laser mirror. Magnification of X140 was obtained when a X40 microscope objective lens was used to focus the light pattern on a photomultiplier displaced 75 cm from the laser. This blown-up image was scanned through a 0.5-mil slit, attached to

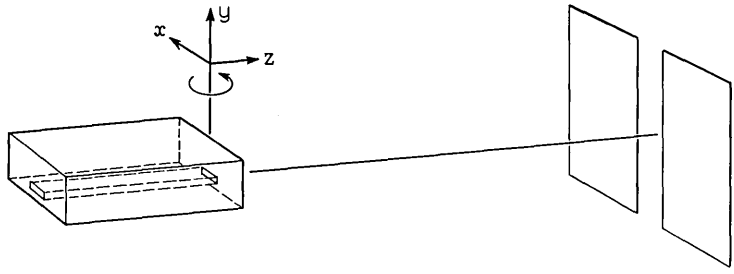


Fig. 1—Experimental arrangement to measure radiation profiles in the far field of a laser beam. The laser was rotated about one of its mirrors and the light power transmitted through a 10-mil slit was measured as a function of rotation angle.

the photomultiplier and aligned parallel or perpendicular to the junction plane.

Sample far field patterns are illustrated in Fig. 2 for a diffused junction laser, and also for two double-heterostructure lasers, one with an oxide-masked stripe and the other with a proton-bombarded stripe. Along the junction plane (Fig. 2a) the pattern of the DJ laser is more Gaussian-like than either of those for the DH lasers. The profile corresponding to the DH laser with the oxide-masked stripe has two

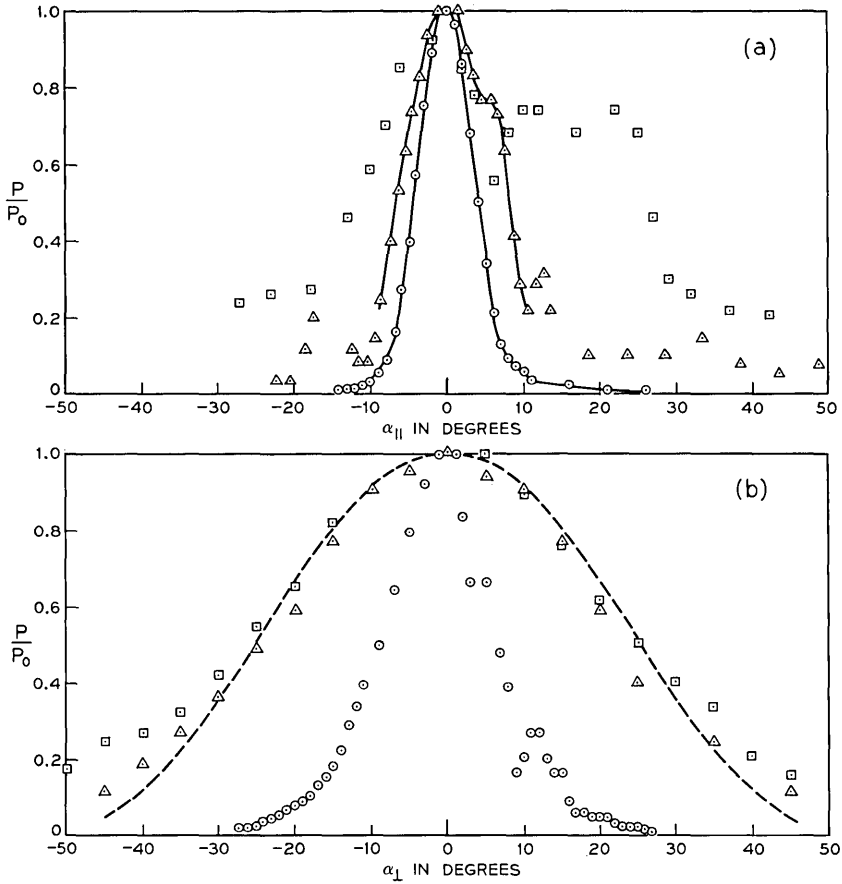


Fig. 2—Far field patterns from junction lasers. Legend: ⊙ L-229 #1, DJ laser with stripe width = 6 μm. ⊡ L-352 #2, DH laser with an oxide-masked stripe, 13 μm wide. △ L-363 #1, DH laser with a proton-bombarded stripe, 13 μm wide. (a) The lasers were rotated through an angle, $\alpha_{\parallel}$, in the plane parallel to the junction. (b) The lasers were rotated through an angle, $\alpha_{\perp}$, in the plane perpendicular to the junction. The dashed curve is a Gaussian distribution fitted to △ data.

peaks implying a multimode field distribution along the laser's junction plane. There is also more fluorescent light within the oxide-masked stripe than within the proton-bombarded stripe as evidenced by the fact that the curve described by ⊡ data points does not fall to zero (20 percent of the peak laser power propagates at large angles from the center of the junction plane).

The far field patterns perpendicular to the junction plane (Fig. 2b) are approximately Gaussian shaped for the oxide-masked and proton-bombarded DH lasers. (The dashed curve in Fig. 2b is a Gaussian distribution fitted to △ data points.) However, on the DJ profile there is evidence of a secondary lobe on the n-side of the junction plane.[2]

The width, between $1/e$ points, of far field profiles like those in Fig. 2 was used to calculate the beam dimensions on the laser mirror in directions parallel and perpendicular to the junction plane through Fig. 9 or equation (8) in Section IV. The broader profiles for DH lasers, in comparison to DJ types, illustrate their larger confinement perpendicular to the junction plane.

Sample near field measurements are contained in Fig. 3 which illustrate the normalized power distribution parallel and perpendicular to the junction plane of a DJ laser and also the profile perpendicular to the junction plane of a DH laser. The larger confinement of the DH

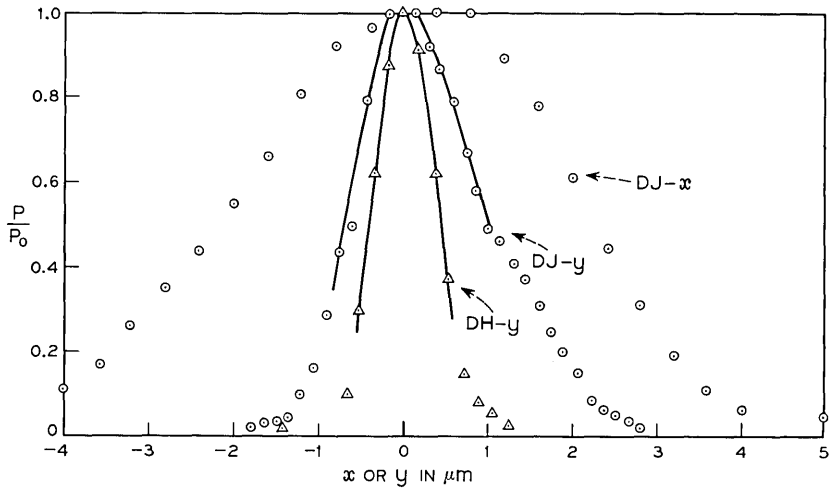


Fig. 3—Near field power distributions on the surface of junction lasers. Legend: ⊙ L-229 #1, DJ laser with stripe width = 6 μm. △ L-363 #1, DH laser with stripe width = 13 μm. Normalized power, P/P_o, is plotted versus distance, x, along the junction and distance, y, perpendicular to the junction.

laser perpendicular to the junction plane is illustrated by the curve following △ data points which is much narrower than the curves following ⊙ data points.

2.2 *Power Coupling Into Fibers*

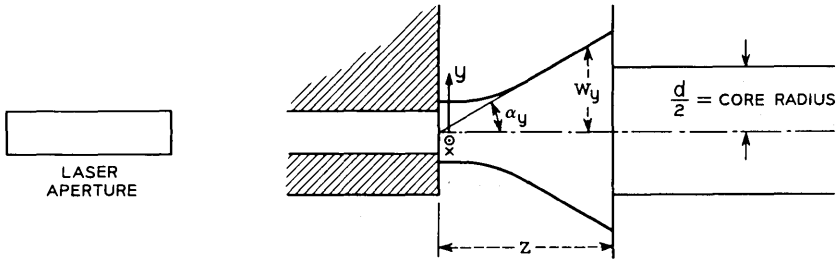
At one end of a fiber the light emanating from the rectangular aperture on a laser was coupled into the circular core (diameter d) of a fiber across a small air gap separating the laser from the fiber. The coupling technique is diagramed in Fig. 4a for a beam which diverges at angle α_y perpendicular to the junction plane.

Power coupling measurements were made for two kinds of single-mode fibers with very different parameters. The 3.7-μm-diameter fiber, fabricated by Corning Glass Works, had a relatively low transmission loss (20 dB/km). Its characteristic number at 9000 Å is $V = 1.55$ based on Corning specifications. This information plus an estimate of the core refractive index enabled us to estimate the fiber's acceptance angle, $\theta = 7$ degrees, as well as the radius, a, to the $1/e$ point of the propagating electric field within the fiber. This is discussed further in Section IV where we find $a = 4.6$ μm. The 3.2-μm-diameter glass fiber was fabricated by DeBell and Richardson. Its parameters[3] are transmission loss $\approx$ 2 dB/m, $n_1 = 1.61$, $V = 2.4$ at 9000 Å, and $\theta = 12$ degrees. In Section IV we estimate $a = 2.6$ μm for this fiber. In this section we will also be describing measurements made with 10-μm and 20-μm multimode fibers ($\theta = 12$ degrees) fabricated by DeBell and Richardson.

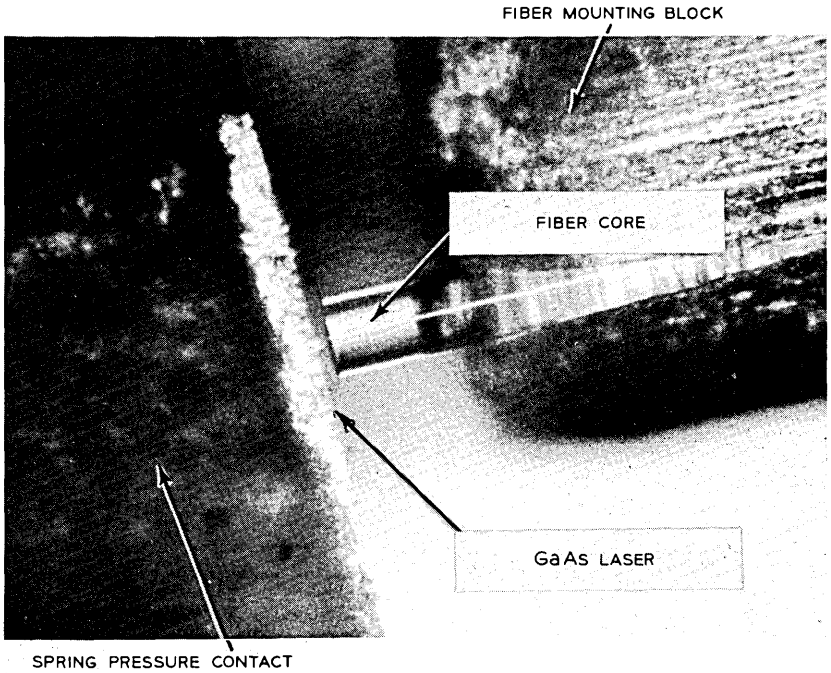
The input tip of a fiber was cut with a razor but was not polished. It was supported on a small copper block and was held in place with epoxy. The input tip extended 20 mils over the edge of the block. The copper blocks supporting laser and fiber were mounted on separate three-dimensional micromanipulators. The output end of the fiber was immersed in oil to eliminate light reflections and light power traveling in the cladding was scattered out by immersing several inches of the fiber in oil matched to the index of the cladding. Light power leaving the output end of a fiber was maximized by adjusting the relative axial and transverse position between the fiber core and the active stripe on the GaAs chip. The micromanipulator holding the fiber had a positioning resolution of 0.125 μm. Provision was also made for reducing the tilt angles between the axes of the laser stripe and the axis of the fiber. Figure 4b contains a photograph of a permanent laser-to-fiber coupler. The picture, taken through a microscope, illustrates the 0.5-mil air gap between the surface of a DJ laser and the input tip of a 10-μm fiber. The power coupling efficiency for this coupler

was 67 percent. It did not change when the coupler was made permanent by applying epoxy between the block supporting the fiber and the block supporting the laser.

Light power from the end of a fiber was measured along with the light power radiating from the exposed laser mirror face. The detectors



(a)



(b)

Fig. 4—(a) The Gaussian light beam emanating from a laser aperture (typical dimensions: 1 μm $\times$ 13 μm) is injected into the core of an optical fiber. In its far field the beam diverges at an angle, α_y, perpendicular to the junction plane. (b) Photograph taken through a microscope shows the 0.5-mil air gap across which light is coupled from a GaAs laser into the core of a 10-micron fiber.

used were Schottky barrier silicon photodiodes, each having an 8-nanosecond rise time, a sensitivity of 4 μA/μW at 9000 Å, and a sensitive area equal to 0.99 cm^2. Their outputs of peak power were simultaneously displayed on a dual-channel sampling oscilloscope.

The power coupling coefficient will be diminished if the axis of the laser beam is displaced from the fiber axis. This effect was measured by offsetting the fiber axis, from its optimum position, parallel and perpendicular to the junction plane. Sample results are illustrated in Fig. 5a where the fiber output power, normalized with respect to its maximum value, is plotted versus displacement along the junction plane, normalized with respect to the fiber radius. Data points ⊙, ⊗, and ◈ apply to DJ lasers. Data points ◬ were obtained with a DH laser (proton-bombarded stripe). The axial offset between laser and fiber was 0.5 mil for the 10-μm fiber and was less than 0.25 mil for the 3.2-μm and 3.7-μm fibers. The profile shapes are similar. For a displacement equal to one fiber radius the 10-μm fiber's output power drops to 50 percent of its optimum value. The profiles for the 3.2-μm and 3.7-μm fibers are wider between $1/e$ points possibly because mode energy extends further into the cladding. Similar measurements were made in the plane perpendicular to the junction plane of the lasers. In that plane the profiles for 3.2 μm and 3.7 μm were broader than their counterparts in Fig. 5a) reflecting the larger divergence angle perpendicular to the junction. They are not included in Fig. 5 since normalized power decreases more rapidly for offsets along the junction plane.

The effect of axially separating fibers from junction lasers is illustrated in Fig. 5b in which P/P_o is plotted versus z. Data points ◈ and ◬ apply to the single-mode fibers. The dashed curves are theoretical estimates based on equation (9) which is discussed in Section IV, and was derived in Ref. 4 for coupled Gaussian beams. The deviation of the ◈ data points, from the theoretical dashed curve, for large z could have resulted because of deviations of the laser beam profile from a Gaussian distribution or because the field energy propagating within the 3.7-μm fiber extended far enough into the cladding to invalidate a Gaussian approximation for the mode profile. The 10-μm and 20-μm multimode fibers (data points ⊗ and ○) begin to lose light when they are axially separated by more than a critical distance, $Z_c \approx (d/2)/\tan\theta$, at which some rays of laser light within the fiber's acceptance cone do not intercept the fiber's core. The critical distance for fibers with a $\theta = 12$ degrees acceptance angle is 24 μm (experimental distance $\approx$ 19 μm) for 10-μm-diameter fibers and 46 μm (experimental

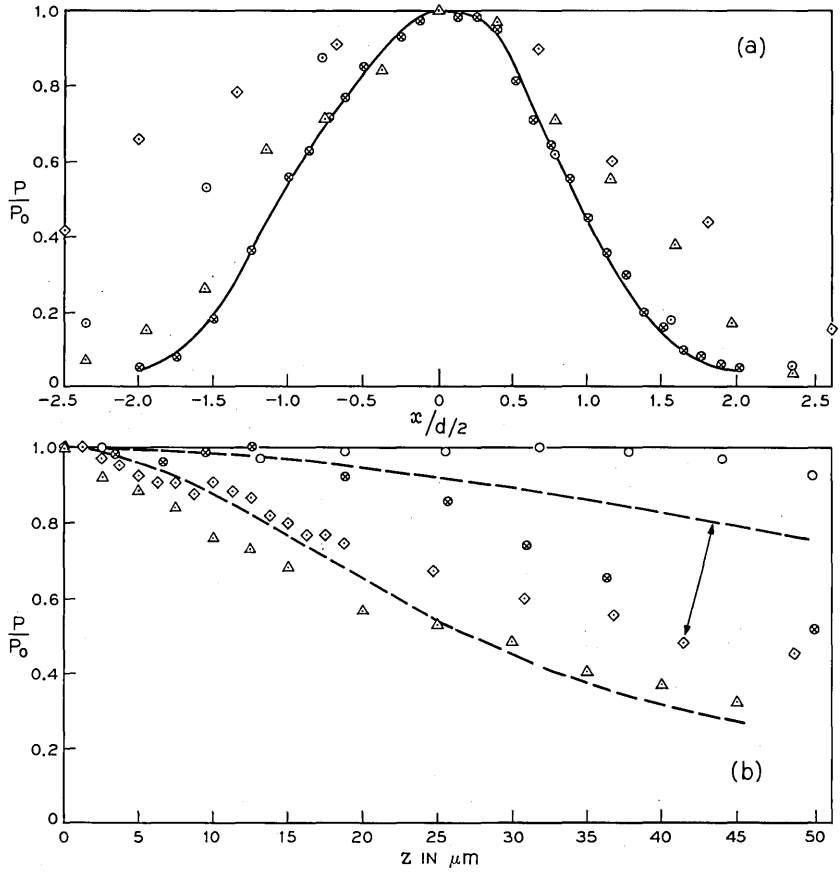


Fig. 5—Profiles of the normalized power emitted from the end of a fiber which was displaced from its optimal position relative to the laser stripe.

Legend: ⊙ 3.2-μm fiber
⊗ 10-μm fiber } L-229 #1 DJ laser
○ 19-μm fiber
◈ 3.7-μm fiber , L-229 #2 DJ laser
△ 3.2-μm fiber , L-363 #1 DH laser.

(a) The fiber core was offset in the plane of the junction. (b) The fiber core was axially separated from the laser surface. The dashed curves were based on equation (9) with $\lambda = 9000$ Å. ($w_{o_x} = 2.7$ μm, $w_{o_y} = 1.3$ μm, $a = 4.6$ μm for ◈ data; $w_{o_x} = 1.5$ μm, $w_{o_y} = 0.36$ μm, $a = 2.6$ for △ data.)

distance $\approx$ 45 μm) for 20-μm fibers. Agreement with experiment is very good for the multimode fibers.

Table I is a summary of experiments performed with two lasers from a diffused junction batch (L-229, stripe width = 6 μm), two lasers from a double-heterostructure batch with proton-bombarded stripes (L-363, stripe width = 13 μm), and two lasers from a DH batch with oxide-masked stripes (L-352, stripe width = 13 μm). Measured power coupling coefficients, κ, are listed for a variety of fibers. The fiber parameters, listed in Table II, are core diameter, d; percent index mismatch between core and cladding, Δ; acceptance angle, θ; characteristic number, V; and number of propagating modes, N. The laser parameters are threshold current, I_{th} ; far field divergence angles, α_x and α_y ; and the beam half-widths, w_{o_x} and w_{o_y} , at the surface of the laser. Parameters w_{o_x} and w_{o_y} were determined in two ways: directly ($w_{o_{x\,\mathrm{exp}}}$, $w_{o_{y\,\mathrm{exp}}}$) from near field measurements or by extrapolating ($w_{o_{x\,\mathrm{theor}}}$, $w_{o_{y\,\mathrm{theor}}}$) from measurements of the far field divergence angles α_x , α_y by using Fig. 9 or equation (8) in Section IV. The mode dimension $w_{o_{x\,\mathrm{theor}}}$ for DH laser L-352 #1 was extrapolated from the width between $1/e$ points of the larger of the two peaks on the multimode far field profile in Fig. 2a. The $1/e$ points were measured relative to the fluorescent light level, P/P_o = 0.2. For every laser $w_{o_{\mathrm{exp}}} > w_{o_{\mathrm{theor}}}$ which implies a greater beam expansion than can be theoretically accounted for by a Gaussian beam having its waist located on the end surface of the laser. The discrepancy implies that the beam waist is located several microns from the end of the lasing stripe. An estimate of the location of the beam waist within the laser cavity is derived from the Gaussian beam expansion curves discussed in Section IV and plotted in Fig. 10a using the beam waist radius, w_o , as the parameter. For DH lasers, the beam waist parallel to the junction plane is located inside the laser around 8 μm from the laser mirror; perpendicular to the junction plane the beam waist is located also inside the laser, about 7 μm from the mirror. For DJ lasers, the beam waist perpendicular to the junction plane is approximately on the laser mirror; parallel to the junction plane, the beam waist is about 8 μm from the mirror.

Theoretical estimates of the power coupling coefficient, κ, based on the laser and fiber parameters, were computed from equation (9) for a Gaussian laser beam. They are listed in Table I for comparison with the measurements. The far field pattern (Fig. 2a) for DH laser L-352 #1 implies a multimode field distribution along its junction plane. The theoretical estimate of κ for DH laser L-352 #1 assumes that 67

TABLE I—SUMMARY OF MEASUREMENTS FOR DIFFUSED JUNCTION LASER (DJ) AND DOUBLE-HETEROSTRUCTURE LASERS (DH) WITH PROTON-BOMBARDED (DH-P) OR OXIDE-MASKED STRIPES (DH-O)

	Laser Parameters							κ							
#	I_{th} (amps)	α_x (deg)	α_y (deg)	w_{0_x} theor. (μm)	w_{0_x} exp. (μm)	w_{0_y} theor. (μm)	w_{0_y} exp. (μm)	Fiber 1 theor. (%)	Fiber 1 exp. (%)	Fiber 2 theor. (%)	Fiber 2 exp. (%)	Fiber 3 theor. (%)	Fiber 3 exp. (%)	Fiber 4 theor. (%)	Fiber 4 exp. (%)
DJ															
L-229 #1	5.6	5	10	2.3	3.6	1.2	1.4		—	72	48	86	73	86	75
#2	7	4.3	8.9	2.7	3.2	1.3	1.5	37	26	77	42	88	64	88	67
DH-P															
L-363 #1	0.59	7.5	29	1.5	2.1	0.36	0.65	8.7	4.6	23	25	38	29		
#2	0.48	12	27	0.95	—	0.40	—		—	18	13	35	24		
DH-O															
L-352 #1	0.73	9.3	26	1.3	1.6	0.43	0.73	8.8	4.3	22	14	40	26		
#2	0.83	13	32	0.88	—	0.33	—	3.5	2.5	9.8	6.5	20	18		

TABLE II—FIBER PARAMETERS (λ = 9000 Å)

#	d (μm)	Δ (%)	θ (deg)	V*	N
1	3.7	0.36	7.1	1.6	1
2	3.2	0.82	12	2.3	1
3	10	↓	↓	7.2	15
4	19	↓	↓	14	98

* $V < 2.41$ for single-mode fibers

percent of the beam power is contained within the dominant Gaussian mode of the laser. The calculation of κ_{theor}, in Table I, assumes that the beam waist is located on the laser mirror. The error caused because the beam waist may actually be located 8 μm from the end of the stripe is derived from Fig. 11 [plotted from equation (9)] which contains curves of κ versus z, with w_o as the parameter, for 3.2-μm and 3.7-μm single-mode fibers. The coefficient, κ, is reduced by 5 percent when a w_{o_y} = 0.5 μm beam is separated by 7 μm from a 3.2-μm fiber. The reduction is 3 percent when a w_{o_x} = 1.5 μm beam is separated by 8 μm from a 3.2-μm fiber. Therefore, the theoretically expected values of κ for 3.2-μm fibers listed in Table I may be too high by as much as 8 percent because the actual beam waist occurs inside the lasing cavity. Similar arguments, applied to 3.7-μm fibers, indicate that K_{theor} may be too high by 3 percent.

The measured values of κ, listed in Table I, were obtained by extrapolating* the lasing light power emitted from a 30-cm length of fiber back to its input tip and comparing it with the total lasing power radiating from the active stripe on the GaAs chip. Power radiating from the stripe is fluorescence until the pump current exceeds the laser's threshold current. Fluorescent light was identified from measurements of total radiated power versus pump current. Figure 6 contains measurements for two DH lasers, one with a proton-bombarded stripe (data points: $\triangle$) and the other with an oxide-masked stripe (data points: $\square$) in which fluorescent light is more prevalent. In the curve for the oxide-masked stripe, fluorescent power is 46 percent of the total laser power when the pump current is 10 percent above threshold.

* The 3.7-μm fiber had a relatively low transmission loss (20 dB/km). Therefore, the power at an input tip is almost equal to the power leaving the output end of a 30-cm length. The transmission loss for DeBell and Richardson fibers is approximately 2 dB/m at 9000 Å. Thus the power at an input tip is about 1.2 times greater than the power emanating from the output end of a 30-cm length.

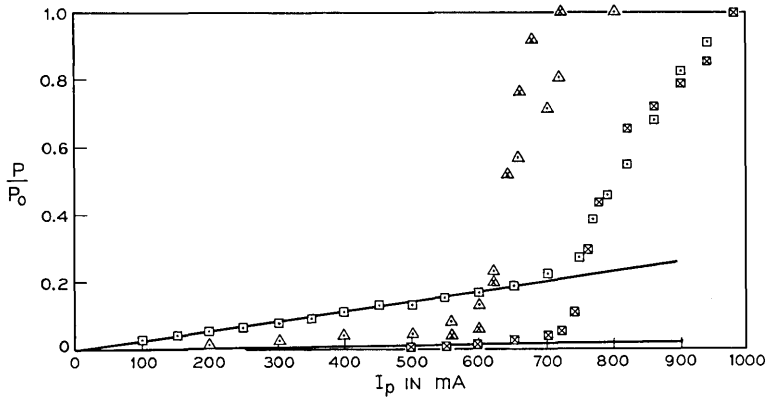


Fig. 6—Normalized power radiated from the laser surface versus pump current: ⊡ L-352 #1, DH laser with oxide-masked stripe; △ L-363 #1, DH laser with proton-bombarded stripe. Normalized power leaving the end of a 3.2-μm fiber versus pump current: ⊠ L-352 #1, DH laser; ▲ L-363 #1, DH laser.

The coupling coefficient with fluorescent light included is 56 percent of its value when the fluorescent component is subtracted away from the total power radiating from the laser.

The coupling coefficients (fluorescent power not included), listed in Table I, were measured for pump currents about 10 percent above threshold within the laser's linear operating region. In general the measured power coupling coefficients in Table I are in good agreement with theory. The cases where the discrepancy is large may have arisen because of deviations of the beam profile from a Gaussian shape or because the field energy propagating within the single-mode fibers penetrated far enough into the cladding to invalidate a Gaussian approximation for the mode profile. The relatively small theoretical coupling coefficients listed in Table I occur because the beam half-width at its waist, $w_{oy\ theor}$, is considerably smaller than the radius, a, of the propagating mode in the 3.7-μm ($a = 4.6\ \mu$m) and 3.2-μm ($a = 2.6\ \mu$m) fibers. In Section III we propose to increase κ by using a cylindrical lens attached to the input tip of a fiber.

III. LENS DESIGN FOR BEAM COLLIMATION

The Gaussian beam emanating from the rectangular aperture on a GaAs injection laser diverges much more rapidly perpendicular to the junction plane than it does parallel to the junction plane. A cylindrical lens can be designed to collimate the laser beam perpendicular to the junction plane and in that way eliminate decoupling due to the

mismatch between the curved wavefront of the propagating mode in a single-mode fiber. Furthermore, the beam cross section can be made square instead of rectangular if the lens is designed to collimate when the beam width perpendicular to the junction is equal to the beam width along the junction. Then, if needed, two spherical lenses can be used to focus the beam into a circular fiber of arbitrary diameter.

Figure 7a illustrates one arrangement with the lens in contact with the fiber core. The parameters to be calculated are the radius of curvature of the lens and its axial offset from the surface of the laser.

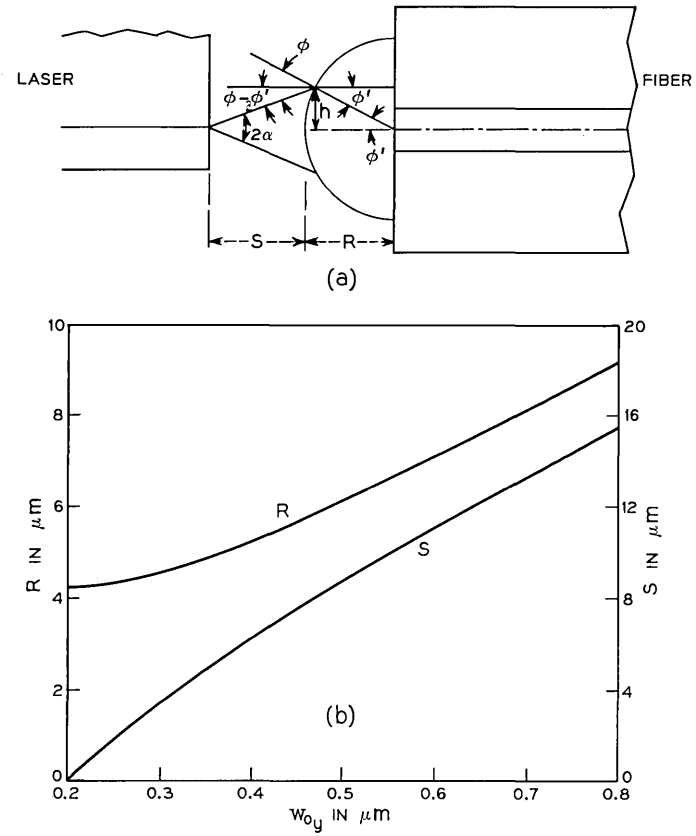


Fig. 7—(a) Cylindrical lens arrangement for collimating the laser beam perpendicular to the junction plane. (b) The radius of curvature, R, of a cylindrical lens and its axial offset, s, from the surface of a laser are plotted versus w_{0_y}, the mode width perpendicular to the junction plane. The lens is designed to collimate the beam (λ = 9000 Å) to a 6-μm half-width perpendicular to the junction plane.

These parameters are functions of the divergence angle, α, of the laser beam; the height at which the beam is to be collimated, h; and the refractive index, n', of the lens glass. Using Fig. 7a as a reference we may write down a system of equations to determine the lens parameters.

The angle of refraction, ϕ', within the lens is defined by:

$$\sin \phi' = \frac{h}{R}. \tag{1}$$

From Snell's Law, ϕ' may be written in terms of the angle of incidence, ϕ,

$$\frac{\sin \phi}{\sin \phi'} \approx \frac{\phi}{\phi'} = n'. \tag{2}$$

The divergence angle, α, of the laser beam may be related to ϕ and ϕ' through

$$\alpha = \phi - \phi'. \tag{3}$$

The offset distance, s, from the surface of the laser may be determined by applying the law of sines:

$$s = R\left(\frac{\sin \phi}{\sin \alpha} - 1\right). \tag{4}$$

In the following we assume that the refractive index of the glass is $n' = 1.5$ and that the Gaussian light mode fills a 12-μm-wide stripe along the junction ($w_{o_x} = 6$ μm, $h = w_{o_x}/\sqrt{2} = 4.2$ μm). The lens parameters, R and s, may easily be recomputed for different mode widths, w_{o_x}, because they are linearly related. Equations (1) through (4) (with $\lambda = 9000$ Å) were used in Fig. 7b to plot R and s as a function of $w_{o_y} = \lambda/(\sqrt{2}\ \pi \tan \alpha)$, the radius of the beam waist perpendicular to the junction. The smallest beam waist (maximum divergence angle) which can be collimated to 12 μm is $w_{o_y} = 0.2$ μm ($\alpha_y = 45$ degrees). As a practical example, consider $w_{o_y} = 0.5$ μm ($\alpha_y = 22$ degrees) for which $R = 6.1$ μm, $s = 8.7$ μm.

One technique for fabricating the cylindrical lens might involve grinding an unclad cylindrical fiber in half.[5] To perform this operation[6] we have constructed a heavy brass lap as illustrated in Fig. 8. The top surface of the lap was surface-ground and 90-degree v-grooves, 2 mils deep, were scribed across it. The procedure involves laying unclad fibers on the bottom of the grooves and holding them in place with wax that can be dissolved by acetone. The grinding is done by turning the brass lap over and polishing its surface down against a wet slurry containing 1-μm carborundum particles.

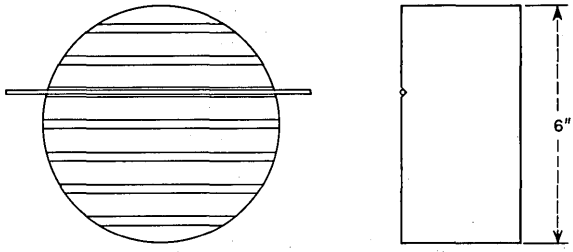


Fig. 8—Brass lap for grinding the surface of an unclad cylindrical fiber.

In a properly fabricated laser-lens-fiber system the sole cause for power decoupling should be due to the mismatch in size between a 6-μm $\times$ 6-μm laser beam and a field of radius, a, propagating within a single-mode fiber. From equation (9) (with $z = 0$) in Section IV we have $\kappa = 2/(6/a + a/6)^2$. If the DeBell and Richardson fiber ($a = 2.6$ μm) is used then $\kappa = 51$ percent* is a lower bound on κ for $1.1 < w_{o_x} < 6$ μm. If the 3.7-μm Corning fiber ($a = 4.6$ μm) is used then $\kappa = 90$ percent* is a lower bound for $3.6 < w_{o_x} < 6$ μm.

IV. COMPUTATION OF POWER COUPLING COEFFICIENTS

The dominant transverse mode of an injection laser has an astigmatic Gaussian spatial distribution. When this mode is injected into an optical fiber a set of modes of the fiber is excited. If the fiber is multimode then it can accept all power within the laser Gaussian distribution that radiates within a cone angle which is less than the fiber's acceptance angle. If the fiber is single-mode then it can only support the HE_{11} mode which we will approximate by a Gaussian distribution. An existing theory[4] for coupling power between two Gaussian beams will be used to compute the power coupling coefficient for injection lasers feeding single-mode fibers.

4.1 *Power Coupling Into a Single-Mode Fiber*

In the following we assume that a Gaussian field $\epsilon(x, y, z)$ emanates from a rectangular aperture on a cleaved GaAs chip (refer to Fig. 4a) and propagates in space along direction z:

$$\epsilon = E(z) \exp\left\{-\left[\left\{\left(\frac{x}{w_x}\right)^2 + \left(\frac{y}{w_y}\right)^2\right\} + j\frac{K}{2}\left\{\frac{x^2}{R_x} + \frac{y^2}{R_y}\right\}\right]\right\}. \quad (5)$$

* These estimates include a 4-percent reflection loss at the input surface of the lens. If the lens surface is made reflectionless then $\kappa = 53$ percent is a lower bound for the 3.2-μm fiber and $\kappa = 94$ percent is a lower bound for the 3.7-μm fiber.

The beam parameters at a convenient reference plane separated by distance, z, from the laser surface are the beam radii w_x , w_y parallel and perpendicular to the laser's junction plane and the wavefront radii of curvature R_x , R_y parallel and perpendicular to the junction plane. For a one-dimensional beam, parameters w and R, at the reference plane, may be written in terms of their values ($w = w_o$, $R_o = \infty$) at the surface of the laser, $z = 0$.[4]

$$w^2 = w_o^2\left(1 + \frac{z^2}{w_o^2}\tan^2\alpha\right), \tag{6}$$

$$\frac{1}{R} = \frac{2\,\frac{z}{w_o^2}\tan^2\alpha}{\left(1 + 2\,\frac{z^2}{w_o^2}\tan^2\alpha\right)}, \tag{7}$$

where

$$\tan\alpha = \frac{\lambda}{\sqrt{2}\,\pi w_o}. \tag{8}$$

In its far field, the width to the $1/e$ point of a Gaussian beam power distribution diverges at angle α. Figure 9, based on equation (8) with $\lambda = 9000$ Å, illustrates a graph of α versus w_o , the half-width of the electric field at the beam waist. Figure 10 contains plots of w and R versus z with w_o , as the parameter, and $\lambda = 9000$ Å. The beam width is collimated in its near field. The wavefront leaves the laser aperture as a plane wave but begins to curve as the beam width begins to enlarge.

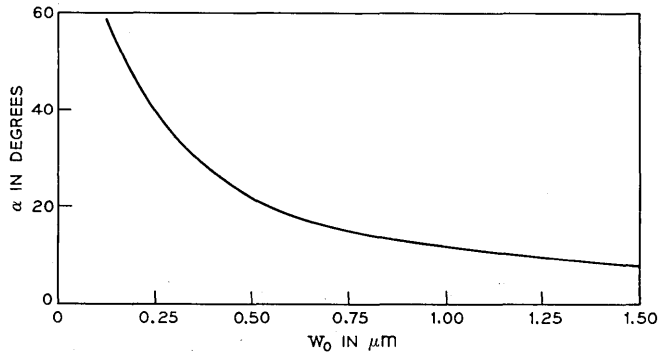


Fig. 9—The far field divergence angle, α, of a Gaussian power distribution is plotted versus the half-width, w_o, of the electric field at the beam waist ($\lambda = 9000$ Å). $\alpha = \tan^{-1}\lambda/\sqrt{2}\,\pi w_0$.

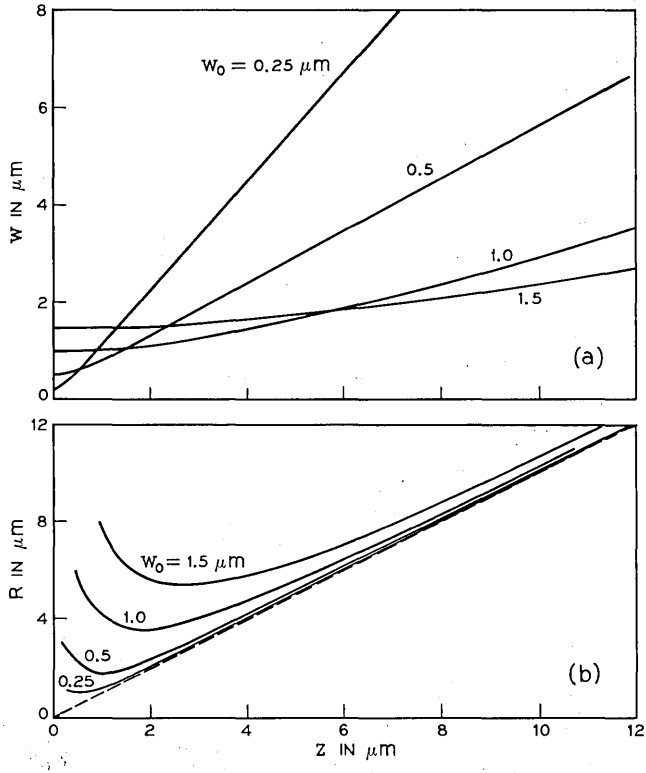


Fig. 10—(a) The half-width, w, of a Gaussian beam is plotted versus z, the separation of the reference plane from the plane of the beam waist (λ = 9000 Å). The half-width, w_o, of the electric field at z = 0 is the parameter. (b) The wavefront radius of curvature, R, is plotted versus z (λ = 9000 Å). w_o is the parameter.

In the far field of the beam both w and R expand linearly with Z. The width of the beam, $2w_o$, at the laser surface may be deduced from equation (8) and far field measurements of α as described in Section 2.1. Sample results are listed under the columns marked α and $w_{o_{theor}}$ in Table I. The near field measurements discussed in Section 2.1 illustrate how to measure w_o directly. Sample measurements are listed in the column marked $w_{o_{exp}}$.

For lasers with rectangular symmetry, the expression for the percent power, κ, coupled from a Gaussian laser beam represented by parameters w_{o_x}, w_{o_y} into another Gaussian distributed field (for the fiber mode) represented by parameters $R = \infty$, $w = a$ is:[4]

$$\kappa = \frac{2}{\left(\frac{w_{ox}}{a} + \frac{a}{w_{ox}}\right)\sqrt{1 + \frac{2z}{K(w_{ox}^2 + a^2)}}} \cdot \frac{2}{\left(\frac{w_{oy}}{a} + \frac{a}{w_{oy}}\right)\sqrt{1 + \frac{2z}{K(w_{oy}^2 + a^2)}}} \tag{9}$$

where $K = (2\pi)/\lambda$ is the propagation constant in free space.

The parameter, a, is the mode radius within a single-mode fiber. It can be estimated from knowledge of the characteristic number, V, of the fiber waveguide which was $V = 2.3$ ($\lambda = 9000$ Å) for the 3.2-μm DeBell and Richardson fiber and was $V = 1.6$ ($\lambda = 9000$ Å) for the 3.7-μm Corning fiber. Since $V < 2.4$ for both fibers, the HE_{11} mode is the only one guided and the percent power in the core may be estimated from the curves of P_{core}/P_{clad} versus V in Ref. 7. For the 3.2-μm fiber 79 percent of the power travels within the core but for the 3.7-μm fiber only 57 percent of the power travels within the core. If the HE_{11} mode is approximated by a Gaussian distribution then the radius, a, of the $1/e$ point of the field intensity can be determined from knowledge of $P_{core}/(P_{core} + P_{cladding})$. We find that $a = 2.6$ μm for the 3.2-μm fiber and $a = 4.6$ μm for the 3.7-μm fiber.

The one-dimensional form of equation (9) was used in Fig. 11a to plot κ (parallel or perpendicular to the junction plane) versus axial separation, z, between either the 3.2-μm fiber ($a = 2.6$ μm, solid curves) or the 3.7-μm fiber ($a = 4.6$ μm, dashed curves) and lasers parametrized by w_o (the radius of the beam waist parallel or perpendicular to the junction plane). The coefficient, κ, sloughs off gradually in the laser's far field due to the increasing mismatch between the beam and mode diameters. The phase mismatch between the curved laser beam wavefront and the planar mode has a marked influence on κ in the near field region where the radius of the wavefront is in the vicinity of its minimum value.

4.2 *Power Coupling Into Multimode Fibers*

The core of a fiber waveguide can accept those rays of laser light which diverge from the lasing slit at angles equal to or less than the acceptance angle, θ, of the fiber where

$$\begin{aligned} \theta &= \sin^{-1}(n_1^2 - n_2^2)^{\frac{1}{2}}, \\ n_1 &= \text{refractive index of core}, \\ n_2 &= \text{refractive index of cladding}. \end{aligned} \tag{10}$$

Assume that the laser beam is Gaussian-distributed along x and y

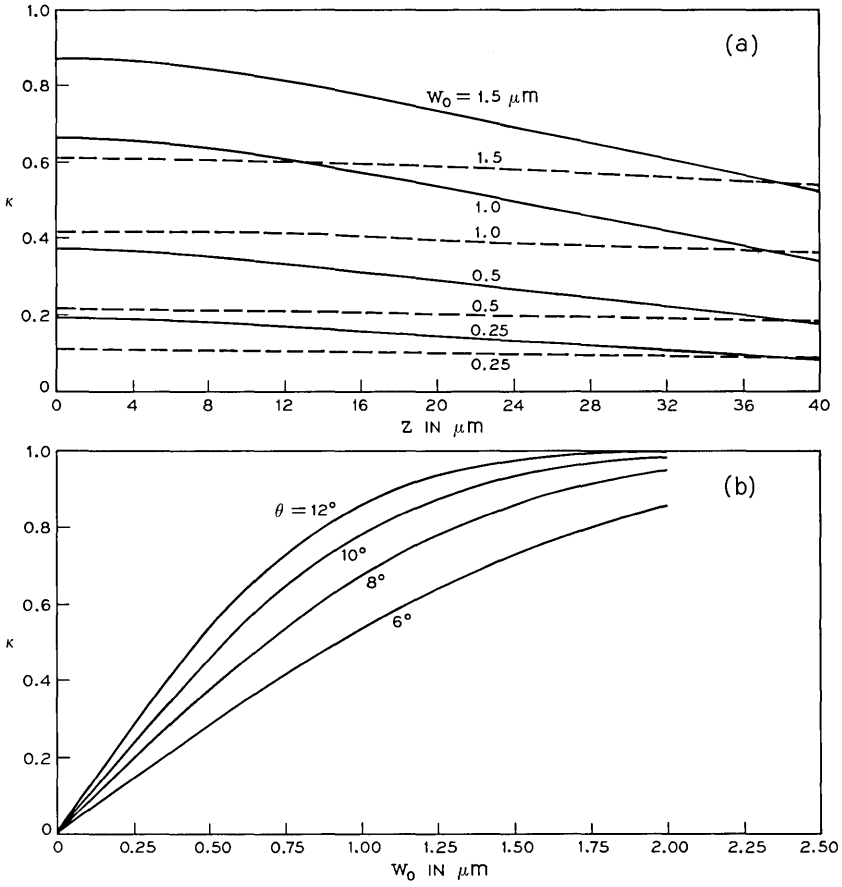


Fig. 11—(a) Power coupling coefficient, κ, is plotted versus axial separation, z, between a laser and a single-mode fiber (λ = 9000 Å). The half-width, w_o, of the laser beam waist is the parameter. Solid curves apply to a 3.2-μm fiber (mode radius = 2.6 μm). Dashed curves apply to a 3.7-μm fiber (mode radius = 4.6 μm). (b) κ is plotted versus w_o (λ = 9000 Å). Fiber acceptance angle, θ, is the parameter.

and that the beam remains collimated along the junction plane. In its far field perpendicular to the junction, the beam half-width increases linearly,

$$w = \sqrt{2}\, z \tan \alpha. \tag{11}$$

The Gaussian power distribution may be expressed as follows:

$$P \propto e^{-2y^2/w^2}. \tag{12}$$

We now compute the fraction of the laser power which travels within

a cone angle equal to θ, the acceptance angle of the fiber. The area subtended by this cone on the fiber end has a half-width, θz. Therefore, the power coupling coefficient, κ, into a multimode fiber with acceptance angle, θ, is:

$$\kappa = \frac{\int_0^{\theta z} e^{-2y^2/w^2}\, dy}{\int_0^{\infty} e^{-2y^2/w^2}\, dy} = \frac{\int_0^{2\pi w_o \theta/\lambda} e^{-t^2/2}\, dt}{\int_0^{\infty} e^{-t^2/2}\, dt} \tag{13}$$

where:

$$t = \frac{2y}{w} = \frac{2\pi w_o y}{\lambda z}. \tag{14}$$

The coefficient, κ, is plotted versus w_o (the half-width of the electric field on the laser mirror) in Fig. 11b. Fiber acceptance angle, θ, is the parameter and λ = 9000 Å.

V. CONCLUSIONS

Experiments have been performed with diffused junction lasers and with double-heterostructure lasers whose stripes were formed by proton bombardment or oxide masking. The far field radiation patterns of DH lasers were considerably broader perpendicular to the junction reflecting their larger confinement in that plane. The patterns along the junctions of DJ lasers were more Gaussian-like than those for DH lasers. Fluorescent light was more prevalent and higher-order modes were more evident with oxide-masked stripes than for those with proton-bombarded stripes.

Theoretical estimates have been calculated for κ, the efficiency for coupling light power from junction lasers into single- and multimode fibers. The highest measured values for κ were:

(*i*) 73 percent (theoretical estimate = 86 percent) for coupling power from DJ lasers into 10-μm fibers.

(*ii*) 48 percent (theoretical estimate = 72 percent) for coupling power from DJ lasers into 3.2-μm single-mode fibers and 26 percent (theoretical estimate = 37 percent) for coupling power from DJ lasers into 3.7-μm single-mode fibers.

(*iii*) 25 percent (theoretical estimate = 23 percent) for coupling power from a DH laser, with a proton-bombarded stripe, into a 3.2-μm single-mode fiber and 5 percent (theoretical estimate = 9 percent) for coupling into a 3.7-μm single-mode fiber.

From theory, the major causes for reduced power coupling between two well-aligned Gaussian beams is a mismatch between the beam diameters at their waists or else a mismatch between the shapes of their wavefronts if the beam waists are axially separated from one another. Our measurements in the near field of DH and DJ laser beams[2] imply a larger beam expansion than can be theoretically accounted for by a Gaussian beam having its waist located on the end surface of the laser. In Section 2.2 we showed that the reduction in the expected value of κ is small (less than 8 percent) if the laser beam waist is located, within the lasing cavity, approximately 8 μm from the end of the stripe. The primary reason for relatively small coupling into the single-mode fibers we measured is the mismatch in diameter of the laser beam waist perpendicular to the junction from the larger diameter field propagating within the fiber. A cylindrical lens has been designed to increase κ by:

(*i*) Allowing the beam radius to increase perpendicular to the junction and then collimating it when it more closely matches the diameter of the propagating mode within the fiber (the beam is essentially collimated along the junction).

(*ii*) Transforming the beam's rectangular cross section into a square so that a spherical lens can be used to further match the beam diameter to the field diameter within the fiber.

The lens design curves indicate the need for a 6-μm radius of curvature. We are attempting to fabricate lenses by grinding cylindrical fibers in half.

VI. ACKNOWLEDGMENTS

The author gratefully acknowledges the helpful suggestions of T. P. Lee and J. C. Dyment.

REFERENCES

1. Dyment, J. C., D'Asaro, L. A., North, J. C., Miller, B. I., and Ripper, J. E., "Proton Bombardment Formation of Stripe Geometry Heterostructure Lasers for 300°K C. W. Operation," to be submitted for publication in Proc. IEEE.
2. Zachos, T. H., and Dyment, J. C., "Resonant Modes of GaAs Junction Lasers—III: Propagation Characteristics of Laser Beams with Rectangular Symmetry," IEEE J. Quant. Elect., *QE-6*, (June 1970), pp. 317–324.
3. Tynes, A. R., Pearson, A. D., and Bisbee, D. L., "Loss Mechanisms and Measurements in Clad Glass Fibers and Bulk Glass," J. Opt. Soc. Amer., *61*, No. 2 (February 1971), pp. 143–153.
4. Kogelnik, H., "Coupling and Conversion Coefficients for Optical Modes," in *Proceedings of the Symposium on Quasi-Optics*, edited by J. Fox (Polytechnic Press, Brooklyn, 1964).
5. Marcatili, E. A. J., private communication.
6. Glynn, P., private communication.
7. Gloge, D., unpublished work.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 51, No. 3, March, 1972
Printed in U.S.A.

The Identification of Modal Resonances in Ferrite Loaded Waveguide Y-Junctions and Their Adjustment for Circulation

By B. OWEN

(Manuscript received September 24, 1971)

This paper reports on extensive eigenvalue measurements made on an X-band waveguide Y-junction containing different ferrite geometries. The frequency dependence of the eigenvalues is used to identify the principal field modes involved, to examine their sensitivities to various junction parameters, and to arrange their correct displacement for circulation. The knowledge gained from these measurements is used to explain the mode of operation of the partial height ferrite Y-junction circulator, and to introduce other novel configurations.

I. INTRODUCTION

The waveguide Y-junction circulator was first introduced by H. N. Chait and T. R. Curry in 1959.[1] During the past ten years it has become a widely used device. Surprisingly, its exact mode of operation has never been clearly understood. Numerous analyses and theories have been proposed for the device, but none has proven entirely satisfactory.

A fundamental theory for the circulator in terms of its external properties was first presented by B. A. Auld in 1959.[2] He showed that the junction scattering matrix was characterized by three eigenvalues which had to be phase displaced by 120 degrees for circulation. The theory, however, did not show how this displacement was to be achieved since this required a more detailed field theory analysis of the junction involved.

Such an analysis for the stripline circulator was first carried out by H. Bosma in 1962.[3] Based on Bosma's results, C. E. Fay and R. L. Comstock in 1965 presented a simplified "two resonant mode" theory for the stripline device together with experimental evidence of the validity of their approximations.[4] This theory was later improved by W. H. von Aulock and Fay in 1968.[5] For lack of a suitable alternative,

the "two resonant mode" theory also became the accepted explanation for the waveguide circulator. The theory stated that circulation occurred in the region between two diametrically determined resonances which were split apart by the applied field. The ferrite saturation magnetization had to be large enough to provide adequate splitting and low enough to avoid low field loss; and the ferrite dimensions had to be suitable for resonance at the desired frequency. A quarterwave transformer also had to be included for broadband matching.

While this explanation seemed satisfactory for the stripline device, it was inadequate for the waveguide case. Attempts to broadband models designed on this basis were rarely successful, and the development was always completed in a cut-and-try manner. An important outcome of these empirical adjustments, however, was the discovery of the partial height ferrite circulator. This seemed capable of larger bandwidths than its full height ferrite counterpart, and it rapidly became established as a standard design. In fact, most broadband devices in use today are of this nature. The discrepancies between the "two resonant mode" theory and the mode of operation of the partial height ferrite device were never clearly explained.

Field theory analyses for the waveguide circulator itself were carried out by J. B. Davies in 1962 and H. J. Butterweck in 1963.[6,7] Davies' analysis provided a solution for the Y-junction boundary value problem by matching the dominant mode fields in the connecting waveguides to a summation of fields due to modes within the junction. The mathematical complexity of the analysis, however, obscured any simplified explanation for circulation. Nevertheless, it did provide a numerical means of circulator synthesis. The only restriction imposed was that the field variation in the direction parallel to the symmetry axis be zero. This limited the analysis to junctions with full height components only. It excluded the partial height ferrite circulators designed since 1959, and narrow bandwidths were at first anticipated. However, in 1965 Davies extended the technique to junctions with more complex full height elements, and predicted much larger bandwidths.[8] The ferrite had a small conducting pin along its axis, and was surrounded by a dielectric sleeve. These predictions were confirmed by C. G. Parsonson, et al., in 1968, and since then near full waveguide bandwidth devices of this nature have been reported by J. B. Castillo and L. E. Davis.[9,10]

In this paper, a useful technique for determining the mode of operation of waveguide circulators is described. It involves a measuring set, operating at X-band, that is capable of displaying the junction eigen-

values as functions of frequency. The phase-frequency responses of the eigenvalues are used to identify the principal field modes involved in each case, and to examine their sensitivities to various junction parameters. The experimental results taken provide an explanation for various types of circulators, and also introduce other novel configurations.

II. THEORY

2.1 *Eigenvalue Analysis*

The scattering matrix $[S]$ of a 3-port junction is given by:

$$[S] = \begin{bmatrix} S_{11} & S_{12} & S_{13} \\ S_{21} & S_{22} & S_{23} \\ S_{31} & S_{32} & S_{33} \end{bmatrix}, \tag{1}$$

where S_{pp} is the reflection coefficient at port p, and S_{pq} is the transmission coefficient from port q to port p. If the junction is symmetrical then:

$$[S] = \begin{bmatrix} S_{11} & S_{12} & S_{13} \\ S_{13} & S_{11} & S_{12} \\ S_{12} & S_{13} & S_{11} \end{bmatrix}, \tag{2}$$

where

$$\begin{aligned} S_{11} &= S_{22} = S_{33} , \\ S_{12} &= S_{23} = S_{31} , \\ S_{13} &= S_{21} = S_{32} . \end{aligned} \tag{3}$$

The scattering matrix of the symmetrical junction is characterized by three eigensolutions:

$$[S][x]_i = \phi_i[x]_i ; \qquad i = 1, 2, 3 \tag{4}$$

where $[x]_i$ are the eigenvectors and ϕ_i are the eigenvalues. A vector in this case represents three signals applied simultaneously at the three ports of the junction. Equation (4) states that for three such excitations $[x]_1$, $[x]_2$, and $[x]_3$,* the reflected signals at each port are equal to the incident signals times a constant ϕ_1 , ϕ_2 , and ϕ_3 respectively. Since the ratio of reflected to incident signals at every port is ϕ_i , it is clear that

* Hence called eigen-excitations.

the eigenvalues are simply the reflection coefficients for the excitations. If the junction is lossless, the eigenvalues have a magnitude of unity $|\phi_i| = 1$, and are distinguishable from one another in phase only.

It has been shown by several authors that the eigen-excitations and eigenvalues are given by

$$[x]_1 = K\begin{bmatrix}1\\1\\1\end{bmatrix}; \qquad [x]_2 = K\begin{bmatrix}1\\1\ \exp(+j120^\circ)\\1\ \exp(-j120^\circ)\end{bmatrix};$$

$$[x]_3 = K\begin{bmatrix}1\\1\ \exp(-j120^\circ)\\1\ \exp(+j120^\circ)\end{bmatrix}; \quad (5)$$

$$\begin{aligned}\phi_1 &= S_{11} + S_{12} + S_{13}\\ \phi_2 &= S_{11} + S_{12}\exp(+j120^\circ) + S_{13}\exp(-j120^\circ)\\ \phi_3 &= S_{11} + S_{12}\exp(-j120^\circ) + S_{13}\exp(+j120^\circ)\end{aligned} \quad (6)$$

where K is a normalizing constant.[2,11-13] $[x]_1$ is the in-phase excitation with the signals in each port having equal magnitude and phase. $[x]_2$ and $[x]_3$ are the clockwise and anticlockwise rotating excitations respectively with the signals in each port having equal magnitudes but differing in phase by 120 degrees.

The simultaneous application of all three eigen-excitations to the junction is given by

$$K\begin{bmatrix}1\\1\\1\end{bmatrix} + K\begin{bmatrix}1\\1\ \exp(+j120^\circ)\\1\ \exp(-j120^\circ)\end{bmatrix} + K\begin{bmatrix}1\\1\ \exp(-j120^\circ)\\1\ \exp(+j120^\circ)\end{bmatrix} = 3K\begin{bmatrix}1\\0\\0\end{bmatrix}. \quad (7)$$

This is equivalent to exciting one port only. If unity power is applied at this port, the normalizing constant K is 1/3. An input on one port is the most familiar form of excitation. As it is the sum of the three eigen-excitations, and as the S parameters are functionally related to the eigenvalues ϕ_i through equation (6), it is evident that the output in the three ports can be determined directly from the eigenvalues.

2.2 *Transmission and Return Losses*

The reflection and transmission coefficients, S_{pp} and S_{pq}, are given from equation (6) by

$$
\begin{aligned}
S_{11} &= 1/3\{\phi_1 + \phi_2 + \phi_3\} \\
S_{31} = S_{12} &= 1/3\{\phi_1 + \phi_2 \exp(-j120^\circ) + \phi_3 \exp(+j120^\circ)\} \\
S_{21} = S_{13} &= 1/3\{\phi_1 + \phi_2 \exp(+j120^\circ) + \phi_3 \exp(-j120^\circ)\}.
\end{aligned} \tag{8}
$$

If the junction eigenvalues are phase displaced from one another such that

$$
\begin{aligned}
\phi_1 &= 1 \\
\phi_2 &= 1 \exp j\theta_2 \\
\phi_3 &= 1 \exp j\theta_3 ,
\end{aligned} \tag{9}
$$

the return loss at port 1 and the transmission losses from ports 1 to 2 and 1 to 3 are given respectively by

$$
\begin{aligned}
&\text{Return Loss } (1 \rightarrow 1) = -20 \log |\, S_{11} \,| \\
&= -20 \log \{1/3 \,|\, 1 + \exp j\theta_2 + \exp j\theta_3 \,|\} \\
&= -20 \log \{1/3[3 + 2\,\mathrm{Cos}\,\theta_2 + 2\,\mathrm{Cos}\,\theta_3 + 2\,\mathrm{Cos}\,(\theta_2 - \theta_3)]^{\frac{1}{2}}\}
\end{aligned} \tag{10}
$$

$$
\begin{aligned}
&\text{Transmission Loss } (1 \rightarrow 2) = -20 \log |\, S_{21} \,| = -20 \log |\, S_{13} \,| \\
&= -20 \log \{1/3 \,|\, 1 + \exp j(\theta_2 + 120^\circ) + \exp j(\theta_3 - 120^\circ) \,|\} \\
&= -20 \log \{1/3[3 + 2\,\mathrm{Cos}\,(\theta_2 + 120^\circ) + 2\,\mathrm{Cos}\,(\theta_3 - 120^\circ) \\
&\quad + 2\,\mathrm{Cos}\,(\theta_2 - \theta_3 + 240^\circ)]^{\frac{1}{2}}\}
\end{aligned} \tag{11}
$$

$$
\begin{aligned}
&\text{Transmission Loss } (1 \rightarrow 3) = -20 \log |\, S_{31} \,| = -20 \log |\, S_{12} \,| \\
&= -20 \log \{1/3 \,|\, 1 + \exp j(\theta_2 - 120^\circ) + \exp j(\theta_3 + 120^\circ) \,|\} \\
&= -20 \log \{1/3[3 + 2\,\mathrm{Cos}\,(\theta_2 - 120^\circ) + 2\,\mathrm{Cos}\,(\theta_3 + 120^\circ) \\
&\quad + 2\,\mathrm{Cos}\,(\theta_2 - \theta_3 - 240^\circ)]^{\frac{1}{2}}\}.
\end{aligned} \tag{12}
$$

These losses are shown plotted for θ_2 and θ_3 in the range zero to 360 degrees in Figs. 1, 2, and 3. As was shown by Auld, circulation occurs when the eigenvalues are mutually displaced by 120 degrees.[2] This condition is satisfied both at X and Y in Figs. 1, 2, and 3, and these points represent the opposite senses of circulation. At X, anticlockwise circulation takes place with

$$
\begin{aligned}
\phi_1 &= 1 \\
\phi_2 &= 1 \exp(+j120^\circ) \\
\phi_3 &= 1 \exp(-j120^\circ),
\end{aligned} \tag{13}
$$

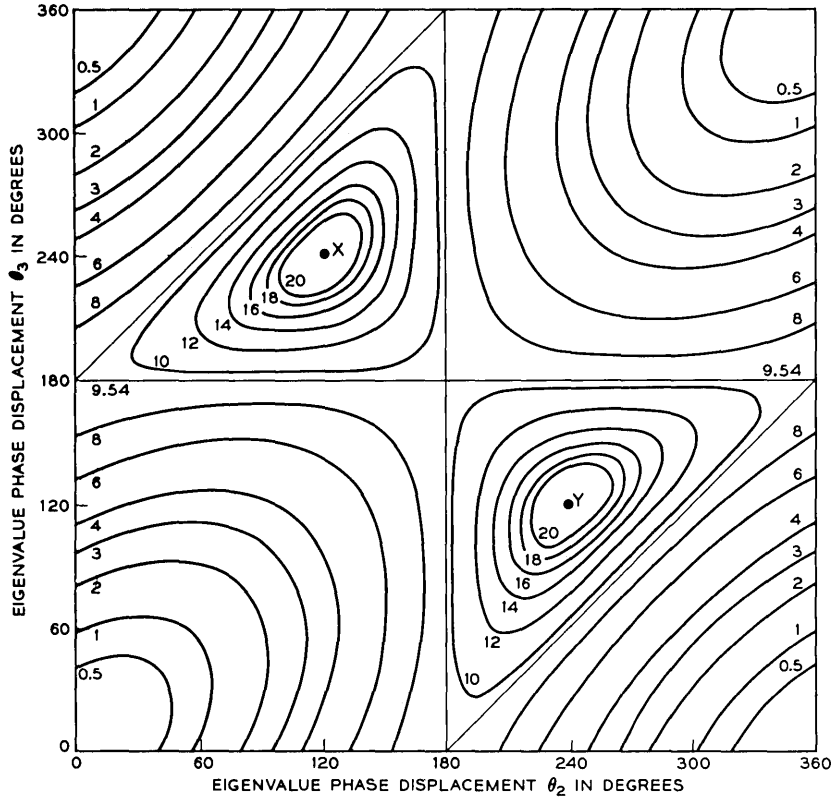


Fig. 1—The return loss in decibels on port 1 is shown as a function of the phase angles between the eigenvalues. The curves are the locii of constant loss.

and at Y, clockwise circulation occurs with

$$\begin{aligned} \phi_1 &= 1 \\ \phi_2 &= 1 \exp(-j120^\circ) \\ \phi_3 &= 1 \exp(+j120^\circ). \end{aligned} \tag{14}$$

Figure 4 schematically displays how circulation is achieved in terms of the eigen-excitations and eigenvalues. The inner and outer circles on each port represent the ingoing and outgoing eigen-excitation components respectively. The ingoing components cancel in ports 2 and 3 and sum in port 1. Only port 1 is then excited. The eigenvalues are the ratios of the outgoing to ingoing wave components for each

eigen-excitation. If the eigenvalue arguments are $\angle\phi_1 = 0°$, $\angle\phi_2 = -120°$, and $\angle\phi_3 = +120°$ as in Figure 4a, the outgoing components cancel in ports 1 and 3 and sum in port 2. The junction then circulates in the clockwise direction. If the eigenvalue arguments are $\angle\phi_1 = 0°$, $\angle\phi_2 = +120°$, and $\angle\phi_3 = -120°$ as in Figure 4b, the outgoing components cancel in ports 1 and 2 and sum in port 3. The junction now circulates in the anticlockwise direction.

Referring again to Figs. 1, 2, and 3, it is interesting to note that the return loss on the input port and the transmission loss to the isolated port are identical only when two of the three eigenvalues are mutually displaced by 120 degrees. Figure 5 shows the losses when θ_2 is fixed at 240 degrees and θ_3 is allowed to vary from 80 degrees to 160 degrees.

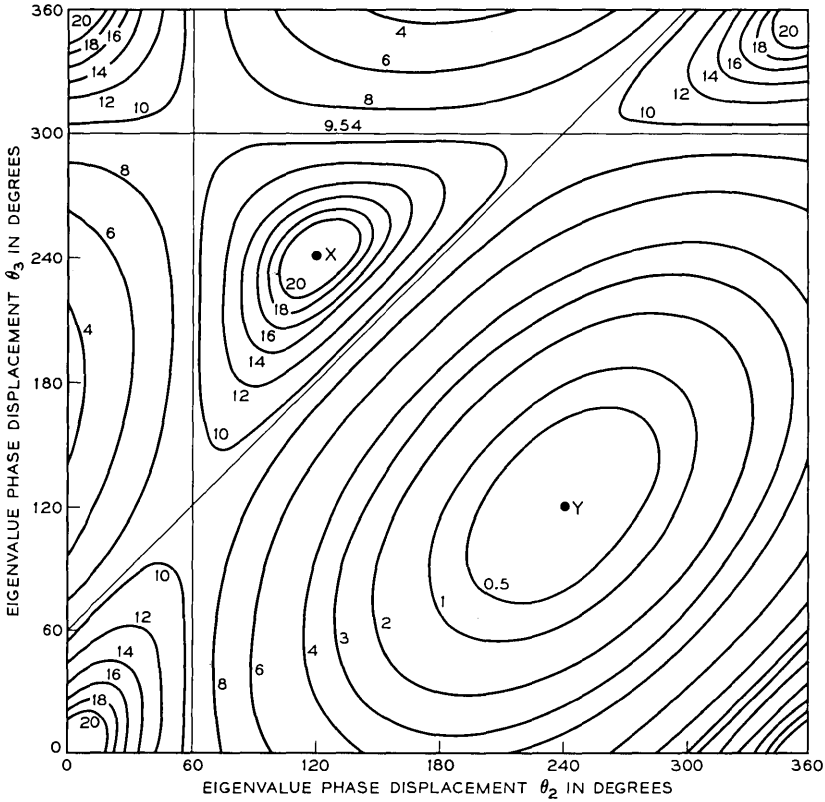


Fig. 2—The transmission loss in decibels from port 1 to port 2 is shown as a function of the phase angles between the eigenvalues. The curves are the locii of constant loss.

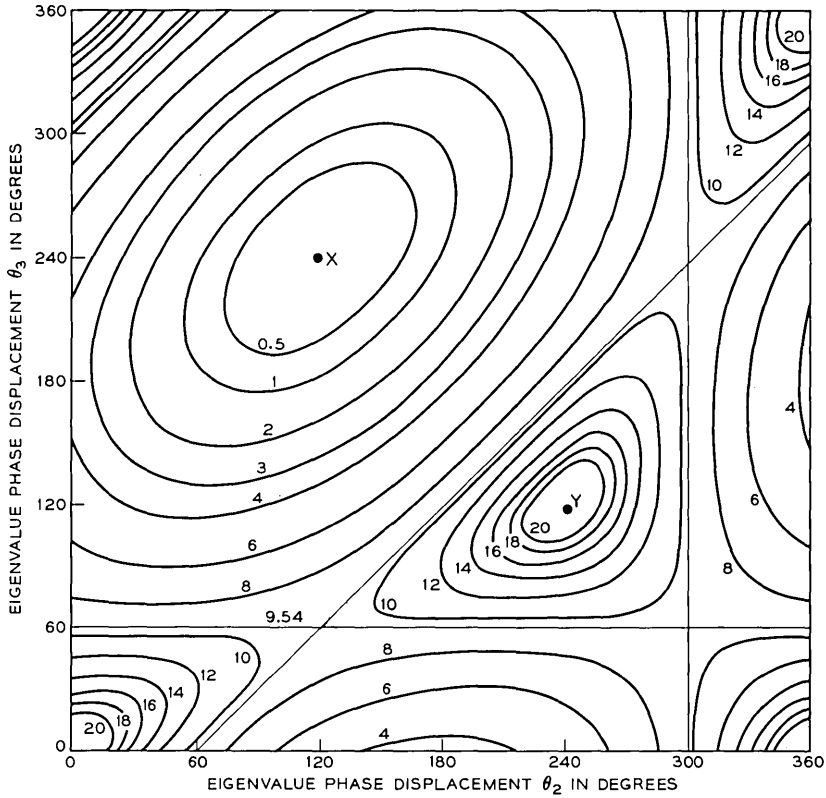


Fig. 3—The transmission loss in decibels from ports 1 to 3 is shown as a function of the phase angles between the eigenvalues. The curves are the locii of constant loss.

The isolation and return loss are both reduced to 30 dB and 20 dB when θ_3 deviates by ± 5.5 degrees and ± 17 degrees respectively from its 120 degrees relative position. This indicates the approximate limits on the phase of the eigenvalues for a satisfactory circulator characteristic.

For broadband circulation, the symmetrically disposed junction components must be arranged so that the eigenvalues are correctly displaced over a wide frequency range. The phase-frequency responses of the three eigenvalues and their sensitivities to various parameters are then important characteristics in broadband circulator design. A circuit capable of examining the eigenvalues individually is described in the following section.

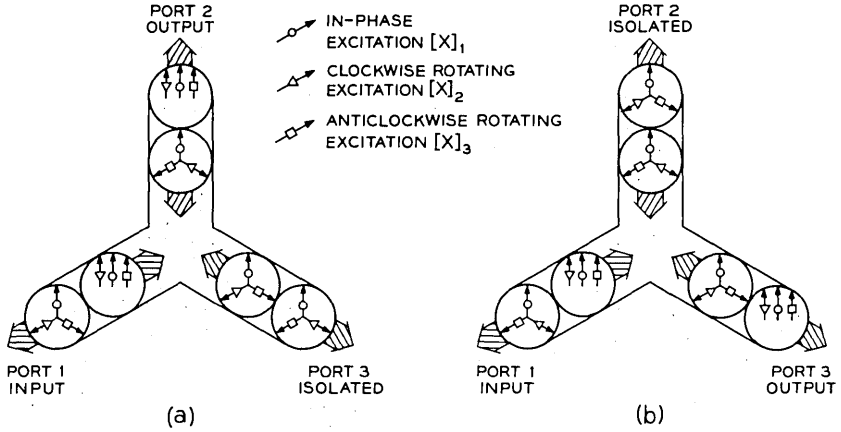


Fig. 4—A phase displacement of 120 degrees between the eigenvalues results in circulation. In (a), ϕ_2 is phase retarded and ϕ_3 phase advanced for clockwise circulation. In (b), ϕ_2 is phase advanced and ϕ_3 phase retarded for anticlockwise circulation.

III. EIGENVALUE MEASUREMENTS

For maximum utilization a circuit is required that is capable of displaying the magnitudes and phases of the three eigenvalues over a full waveguide band. Several eigenvalue measurement techniques are available in the literature, but only a few seem capable of operating on a swept frequency basis.[11] One method is to measure the reflection and transmission coefficients S_{11} , S_{12} , and S_{13} , and to determine the eigenvalues directly from equation (6). The computation involved,

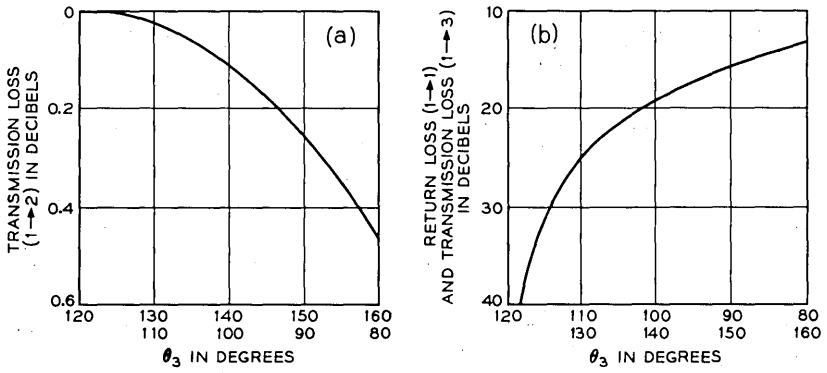


Fig. 5—(a) Transmission loss from ports 1 to 2 and (b) transmission loss from ports 1 to 3 and return loss on port 1 as a function of θ_3 with θ_2 fixed at 240 degrees.

however, is rather extensive, and it does not provide a real time display. The technique to be reported here is similar to a method proposed by F. M. Magalhaes for synthesizing lumped element circulators.[14] The three eigen-excitations are individually applied to the junction, and the eigenvalues are determined from the reflection coefficients on any one port.

A circuit capable of such measurements at X-band is shown in Fig. 6. The 3-way power divider couples signals of equal magnitude to the three arms connected to the Y-junction. These arms are of the same mechanical length, and are made up of identical components which track in phase and loss as a function of frequency. The 20-dB attenuators prevent errors due to circulating energy in the waveguide circuit, and the variable attenuators equalize any loss differences in the three arms. The phase shifters are adjustable for the in-phase, and rotating eigen-excitation conditions.

To avoid the use of two couplers on each arm for reflection coefficient measurements, the eigenvalues are determined by sampling the incident signal on one port and the reflected signal on the other. The coupler on arm 3 is terminated, and is included only for phase equalization. The other two couple the incident and reflected signals on arms 1 and 2 respectively to a network analyzer and XY recorder. The components then measured for the three eigen-excitations are ϕ_1 , ϕ_2 exp $(+j120°)$, and ϕ_3 exp $(-j120°)$ respectively. A zero-phase adjustment of 0°, $-120°$, and $+120°$ on the analyzer enables ϕ_1 , ϕ_2 , and ϕ_3 , respectively, to be displayed on the XY recorder. An alternative method which avoids this zero-phase adjustment is discussed in the Appendix.

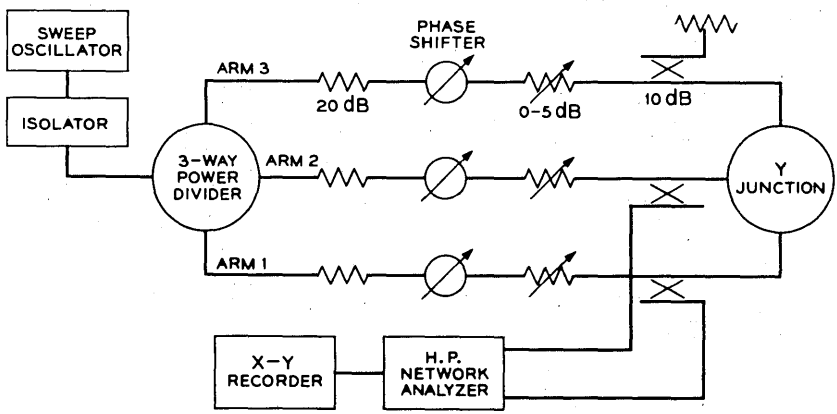


Fig. 6—The eigenvalue measuring set.

IV. RESULTS

The Y-junction used in the experimentation was made from 0.900-inch wide and 0.450-inch high X-band waveguide. Most commonly used waveguides have this 2 to 1 aspect ratio. Its selection here was motivated by the desire to scale synthesized circulators to other frequency bands. For compatibility with the measuring set, broadband transitions to standard 0.400-inch high guide were used on each port. The junction was milled in brass, and threaded plugs at the center of the lid facilitated the interchange of ferrite posts. Some plugs had a finely threaded hole at the center so that metallic pins of various diameters could be introduced along the junction symmetry axis. The reflection coefficient measurements were referenced to planes on the three ports 0.437 inch from the junction center. The polarity of the biasing field was kept constant, and its magnitude was determined using a Hall effect probe with the ferrite post removed.

Eigenvalue measurements were made on the junction with a number of different ferrite geometries enclosed. Several different ferrites were also used. Their saturation magnetizations and loss tangents were small enough to avoid magnetic low field losses and to minimize dielectric losses. The junction was then essentially free of dissipation, and the eigenvalue magnitudes were very close to unity. Only the phases of the eigenvalues need then be considered. Most of the results were taken on Trans-Tech magnesium based TT1-1500 ferrite. This was available in large quantities at the time, and was selected for this reason only. It had the following properties:

Saturation magnetization, $4\pi M_s = 1500$ Gauss
Dielectric constant, $\epsilon = 12$
Dielectric loss tangent, Tan $\delta = 0.00025$
Line width, $\Delta H = 180$ Oersted.

Normally a ferrite with up to twice this saturation magnetization would be selected for use in broadband X-band circulators. Care was taken to demagnetize each ferrite sample before it was used in the junction.

4.1 *Junction with a Full Height Ferrite*

Consider the junction with a full height ferrite post, and excited from the connecting waveguides by the dominant TE_{10} mode. The electric field is everywhere parallel to the symmetry axis, and no variations occur in this direction. The junction modes are then of the $TM^{*}_{\pm m,n,0}$ type. Integer m indicates the number of full period

* Transverse to the symmetry axis.

azimuthal variations with the field pattern rotating in the clockwise (+) and anticlockwise (−) directions. Integer n indicates the number of half-period radial variations. The $TM_{\pm m,n,0}$ ($m > 0$) rotating modes are excited by the $[x]_2$ and $[x]_3$ eigen-excitations. Their magnetic field components are circularly polarized in opposite directions at the junction center. The $TM_{0,n,0}$ in-phase modes are excited by the $[x]_1$ eigen-excitation. Their magnetic field components are zero at the junction center.

The eigenvalue phase-frequency response of such a junction with a 0.260-inch diameter ferrite is shown in Fig. 7. Only the rotating $TM_{\pm 1,1,0}$ modes are resonant within band. The $[x]_2$ eigen-excitation couples to the $TM_{+1,1,0}$ mode, and the $[x]_3$ eigen-excitation couples to the $TM_{-1,1,0}$ mode. With no field applied, $\angle\phi_2$ and $\angle\phi_3$ are identical, and both resonances occur at 9.5 GHz. As field is applied, the degeneracy is removed and the resonances split apart. This occurs because the magnetized ferrite presents different permeabilities to the two circularly

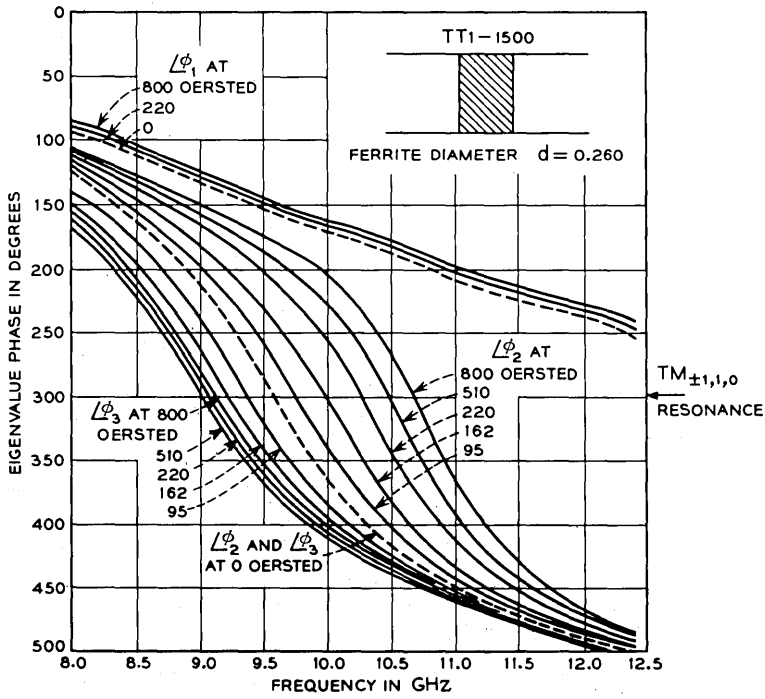


Fig. 7—The eigenvalue phase-frequency response of a junction with a full height ferrite post and various biasing fields.

polarized modes. The resonant frequency of the $TM_{+1,1,0}$ mode increases with field while the resonant frequency of the $TM_{-1,1,0}$ mode decreases until the ferrite is saturated, and then increases as shown in Fig. 8. An identical phenomenon was observed for the stripline circulator by Fay and Comstock in 1965.[4] The phases of eigenvalues ϕ_2 and ϕ_3 are significantly displaced only in the immediate vicinity of the resonances where the fields in the ferrite are at a maximum. None of the in-phase modes are resonant within band, and $\angle\phi_1$ is only slightly modified by the applied field. The transmission losses from ports 1 to 2 and ports 1 to 3 are determined from the measured eigenvalues using equations (11) and (12). They are shown in Fig. 9. The eigenvalues are correctly displaced for circulation at 10.3 GHz when the applied field is 220 Oersted. The 20-dB isolation bandwidth is 350 MHz (3.4 percent).

The circulating frequency may be adjusted by varying the $TM_{\pm1,1,0}$ resonances. This may be accomplished over a limited range by using ferrite with different dielectric constants, as shown in Fig. 10. The resonant frequency of the zero field $TM_{\pm1,1,0}$ mode varies inversely as the square root of the dielectric constant, and decreases from 11.2 GHz to 10 GHz as the dielectric constant is increased from 11.3 to 14.5. Over a wider band, changing the ferrite post diameter is much more effective. The zero field resonance varies inversely with the post diameter, and sweeps from 12.4 GHz to 8.0 GHz in Fig. 11 as the diameter is increased from 0.200 inch to 0.300 inch. The slope of $\angle\phi_1$ is essentially unchanged over the range in both cases, and the circulator bandwidth remains nearly constant.

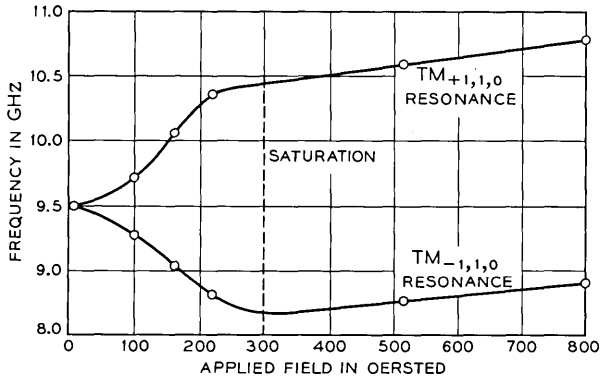


Fig. 8—The splitting of the $TM_{+1,1,0}$ and $TM_{-1,1,0}$ resonances in a junction with a full height ferrite post as a function of the biasing field. The results are obtained from Fig. 7.

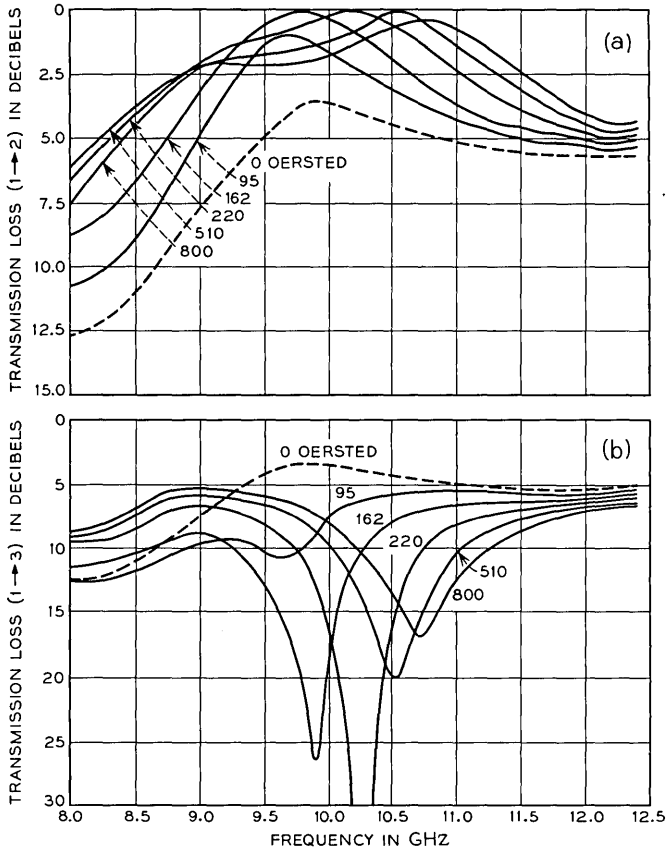


Fig. 9—Transmission losses from (a) ports 1 to 2 and (b) ports 1 to 3 in a junction with a full height post for various biasing fields. The results are obtained from Fig. 7 and equations (11) and (12).

Circulation is also possible in the vicinity of the higher order $TM_{\pm 2,1,0}$ mode which comes within band when the ferrite diameter is larger than 0.320 inch. It is closely accompanied by the in-phase $TM_{0,2,0}$ mode. The $[x]_2$ eigen-excitation now couples to the $TM_{-2,1,0}$ mode, and the $[x]_3$ eigen-excitation to the $TM_{+2,1,0}$ mode. For the same polarity of biasing field, the splitting occurs in opposite directions from the splitting in the vicinity of the $TM_{\pm 1,1,0}$ modes. The $TM_{+2,1,0}$ resonant frequency decreases with field while the $TM_{-2,1,0}$ resonant frequency increases as shown in Fig. 12. The effect was also predicted by Fay and Comstock in 1965.[4]

4.2 *Junction with a Centrally Pinned, Full Height Ferrite*

It was shown by K. Kurokawa that the electric field components of the in-phase excitation sum at the junction center while those of the rotating excitations mutually cancel.[13] This feature can be used to provide an independent control for the phase of eigenvalue ϕ_1 .

A thin metallic pin inserted along the symmetry axis does not affect the rotating modes, and leaves $\angle\phi_2$ and $\angle\phi_3$ unchanged. It significantly affects the phase of ϕ_1 by introducing $TM_{0,n,z}$ type modes. These have components of electric field perpendicular to the symmetry axis, and resonate between the discontinuity at the open end of the pin and the broadwall of the waveguide junction. They are demonstrated in Fig. 13. As the pin is inserted into the junction, two resonances sweep across the band from the high end. They occur when the pin is approximately one quarter and three quarter wavelengths long in the ferrite medium. They are the $TM_{0,1,\delta}$ and $TM_{0,1,1+\delta}$ ($\delta \approx 0.5$) resonances respectively. When the pin extends fully across the junction, continuity in the direction parallel to the symmetry axis is restored and the resonances disappear. The phases of ϕ_2 and ϕ_3 are essentially unaffected by the pin position.

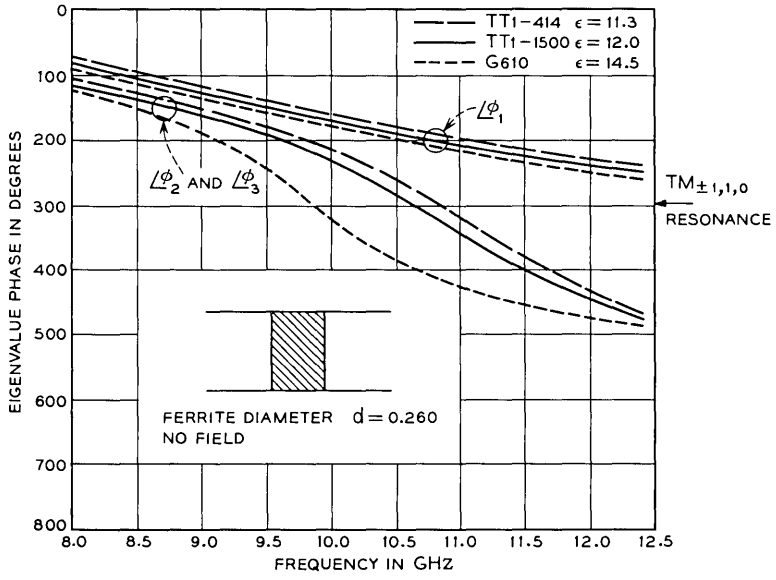


Fig. 10—The eigenvalue phase-frequency response of a junction with a full height ferrite post with various dielectric constants. No biasing field is applied.

The $TM_{0,n,z}$ resonances are not only a function of the ferrite diameter, ferrite dielectric constant, and pin height but also of the ferrite hole and pin diameters. They move to higher frequencies as these diameters are made larger as shown in Fig. 14. This is expected since both the

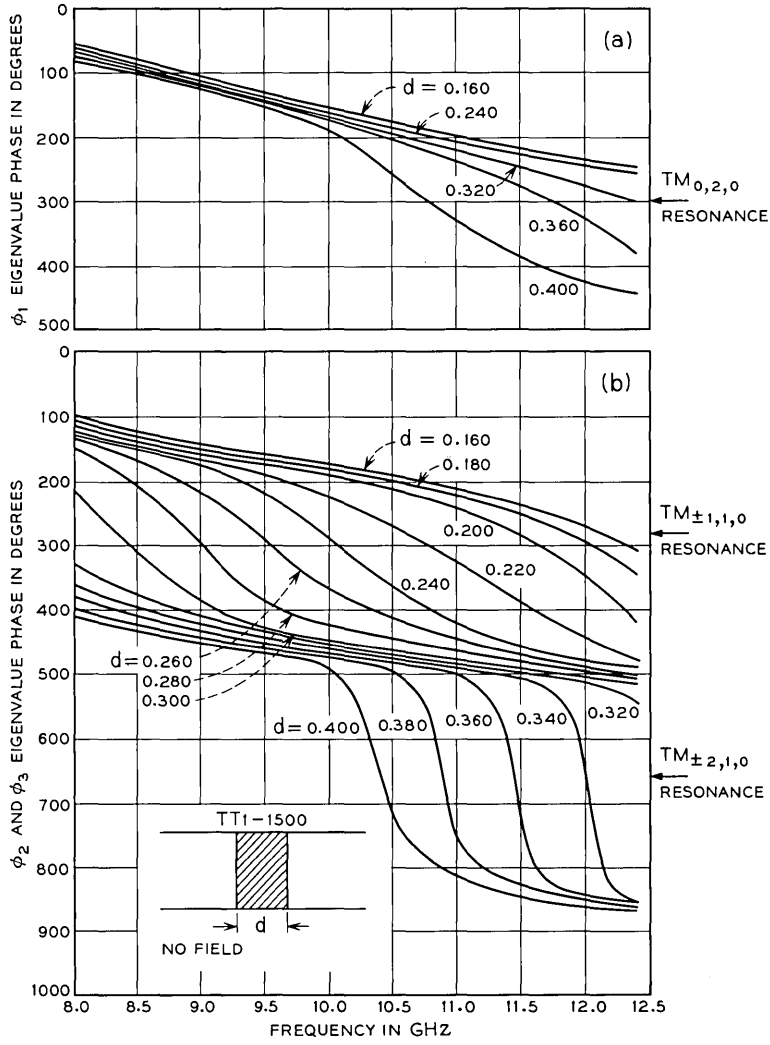


Fig. 11—The eigenvalue phase-frequency response of a junction with a full height ferrite post of various diameters. The results display (a) the lower-order $TM_{0,n,0}$ modes in $\angle\phi_1$, (b) the lower-order $TM_{\pm m,n,0}$ modes in $\angle\phi_2$ and $\angle\phi_3$.

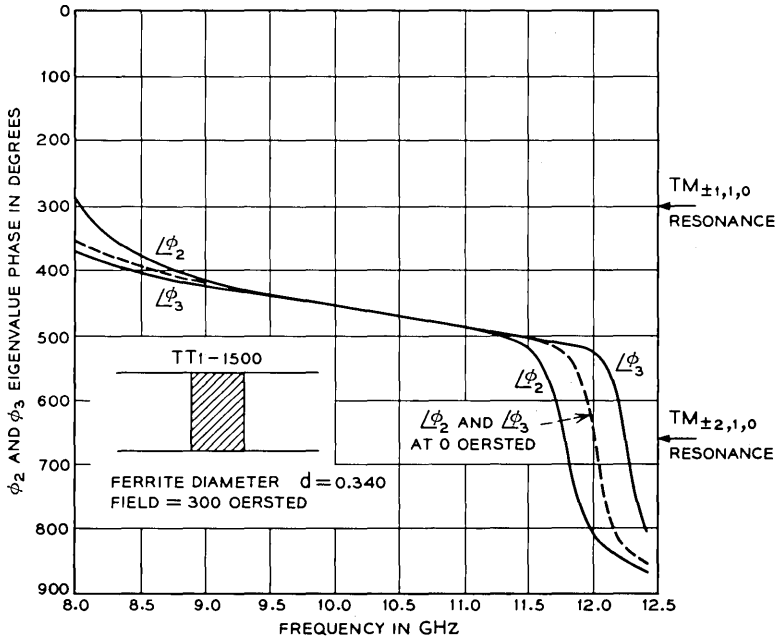


Fig. 12—The eigenvalue phase-frequency response of a junction with a full height ferrite post showing the $TM_{\pm1,1,0}$ and $TM_{\pm2,1,0}$ resonant modes splitting in opposite directions with field.

larger hole and pin tend to exclude fields from the ferrite, and thereby increase the frequency at which the pin resonates. The change in the phases of ϕ_2 and ϕ_3 is initially very small. However, when the pin diameter reaches 0.098 inch, a resonance occurs at 10.8 GHz. The discontinuity at the open end of the pin is now sufficiently large to excite rotating modes that resonate along the ferrite axis. These modes are of the $HE_{\pm m,n,z}$ type. They are also excited by a partial height ferrite, and are discussed in detail in Section 4.4.

The hole in the ferrite has other effects that have to be considered. The permeability difference for the rotating modes is a maximum at the junction center where the magnetic fields are circularly polarized. At the ferrite edge where the polarization is elliptical, the permeability difference is much less. Removing ferrite from the center then reduces the phase displacement of ϕ_2 and ϕ_3 with field as shown in Fig. 15. The demagnetized curves also indicate that the resonances move to higher frequencies as the hole diameter is made larger. This is due to the reduction in the effective dielectric constant of the ferrite cylinder.

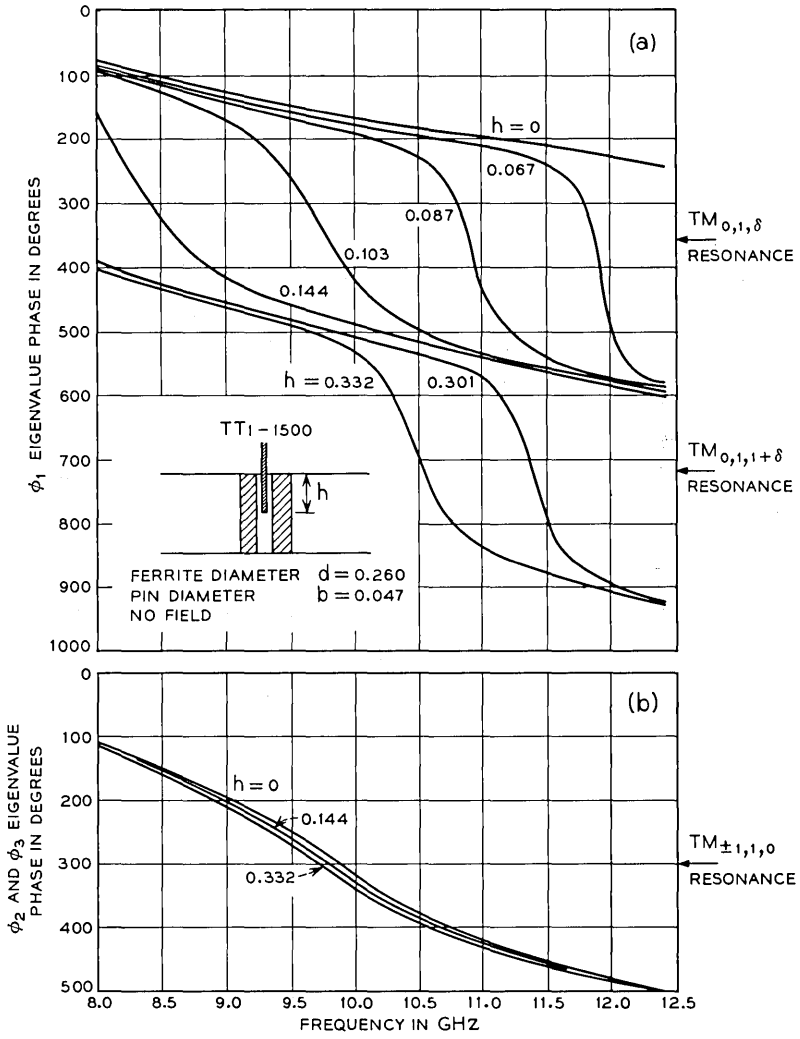


Fig. 13—The eigenvalue phase-frequency response of a junction with a centrally pinned ferrite post. The results display (a) the $TM_{0,1,\delta}$ and $TM_{0,1,1+\delta}$ resonances that are induced in $\angle\phi_1$ by the variable height pin and (b) how $\angle\phi_2$ and $\angle\phi_3$ remain unchanged.

The independent control on $\angle\phi_1$ provided by the thin pin is very useful for circulator synthesis as is demonstrated in Fig. 16. The ferrite is magnetized to separate $\angle\phi_2$ and $\angle\phi_3$ by 120 degrees at 9.4 GHz. A pin-induced $TM_{0,1,\delta}$ resonance is then adjusted to position $\angle\phi_1$ 120 degrees away from both $\angle\phi_2$ and $\angle\phi_3$ at the same frequency. The resultant circulator has a 20-dB isolation bandwidth of 600 MHz

(6.4 percent). The bandwidth improvement from the non-pinned junction is due to the $TM_{0,1,\delta}$ resonance. The slope of $\angle\phi_1$ in its vicinity more nearly matches the slopes of $\angle\phi_2$ and $\angle\phi_3$. If all three resonances had lower and more comparable eigenvalue phase-frequency slopes,

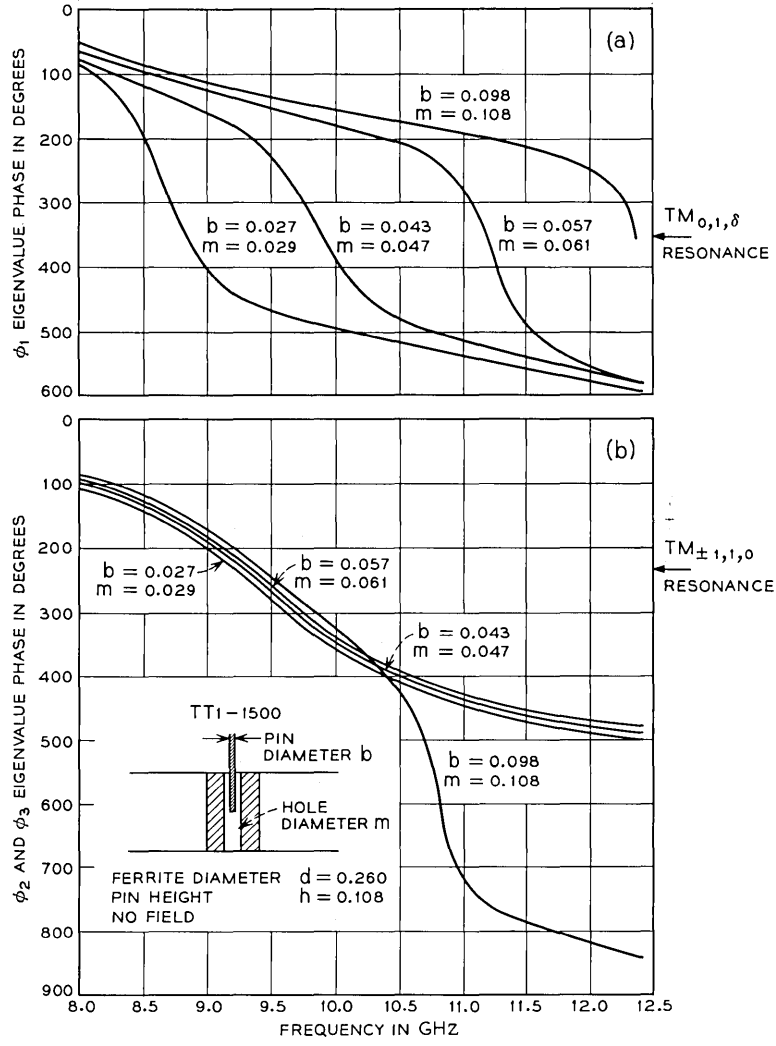


Fig. 14—The eigenvalue phase-frequency response of a junction with a full height ferrite post and containing various diameter holes and various diameter axial pins. In (a), the $TM_{0,1,\delta}$ resonance in $\angle\phi_1$ increases in frequency as the pin diameter is made larger. In (b), $\angle\phi_2$ and $\angle\phi_3$ are unaffected until the pin is sufficiently large to excite axially resonating $HE_{\pm m,n,z}$ modes.

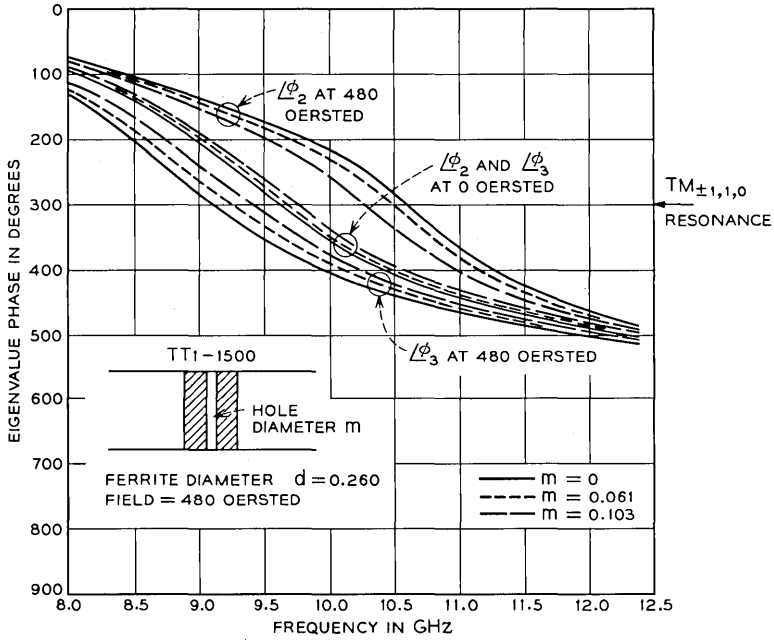


Fig. 15—The eigenvalue phase-frequency response of a junction with a full height ferrite post and with different diameter holes along its axis. The results are displayed both with and without field.

the bandwidth improvement would have been even more substantial.

4.3 *Junction with a Partial Height Ferrite*

It was recently proposed that some circulators with a partial height ferrite operate in a turnstile fashion with rotating modes propagating along the ferrite axis.[15] These modes are excited by the dielectric discontinuity at the open end of the ferrite, and are characterized by electric field components perpendicular to the symmetry axis. When the rotating eigen-excitations are applied, the signals from the three ports induce three such transverse components on the ferrite, each space displaced from the other by 120 degrees. Together they couple to the $HE_{\pm m,n,z}$ series of modes that resonate along the ferrite axis. For the in-phase excitation the three transverse components are in phase with one another, but the 120 degrees spatial separation results in cancellation at the junction center. They couple to the $TM_{0,n,z}$ series of modes as before.

Figure 17 demonstrates these effects on a 0.300-inch diameter ferrite.

The full height results are shown dotted. The $TM_{\pm 1,1,0}$ diametric resonance is centered at 8.2 GHz, and no $TM_{0,n,0}$ mode is resonant within band. A small reduction in the ferrite height produces rapid 360-degree phase transitions in $\angle\phi_2$ and $\angle\phi_3$ at 8.9 GHz and 11.2 GHz. These are the $HE_{\pm 1,1,1+\delta}$ and $HE_{\pm 1,1,2+\delta}$ axial resonances, respectively. They resonate between the open end of the ferrite on the one side, and the broadwall of the waveguide junction on the other. As the ferrite height is reduced, the resonances move to higher frequencies. Below 0.250 inch, the lowest-order $HE_{\pm 1,1,\delta}$ resonance comes within band. The ferrite is approximately an odd number of quarter wavelengths long ($\delta \approx 0.5$) at these HE resonate frequencies. The physical dimensions are in good agreement with those predicted for the resonances when the ferrite is assumed to be a dielectric rod waveguide with dielectric constant $\epsilon = 12$ and relative permeability $\mu \approx 1$.

Recalling from Section 4.2 that a resonance is necessary to separate the phases of ϕ_2 and ϕ_3, circulation now becomes possible in the vicinity of both the $HE_{\pm m,n,z}$ and $TM_{\pm m,n,0}$ resonances. The bandwidths are comparable in both cases. Further, as the resonances are differentially affected by the ferrite height, they are easily "staggered" for a larger

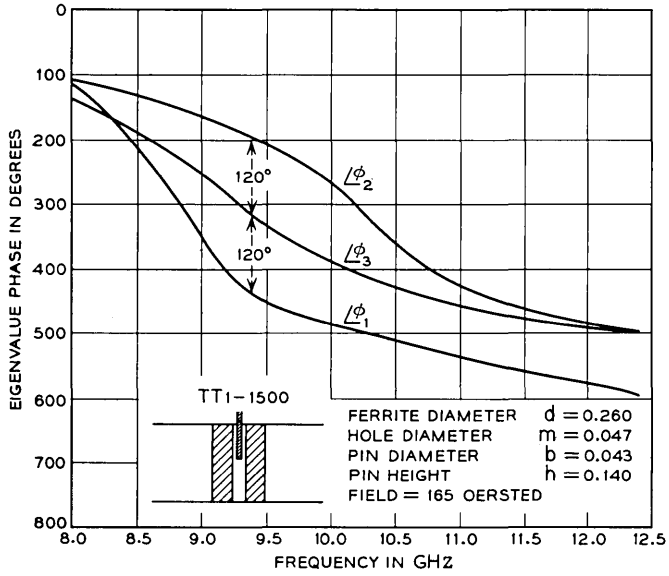


Fig. 16—The eigenvalue phase-frequency response of a junction with an axially pinned ferrite post. The biasing field and pin have been adjusted for circulation at 9.4 GHz.

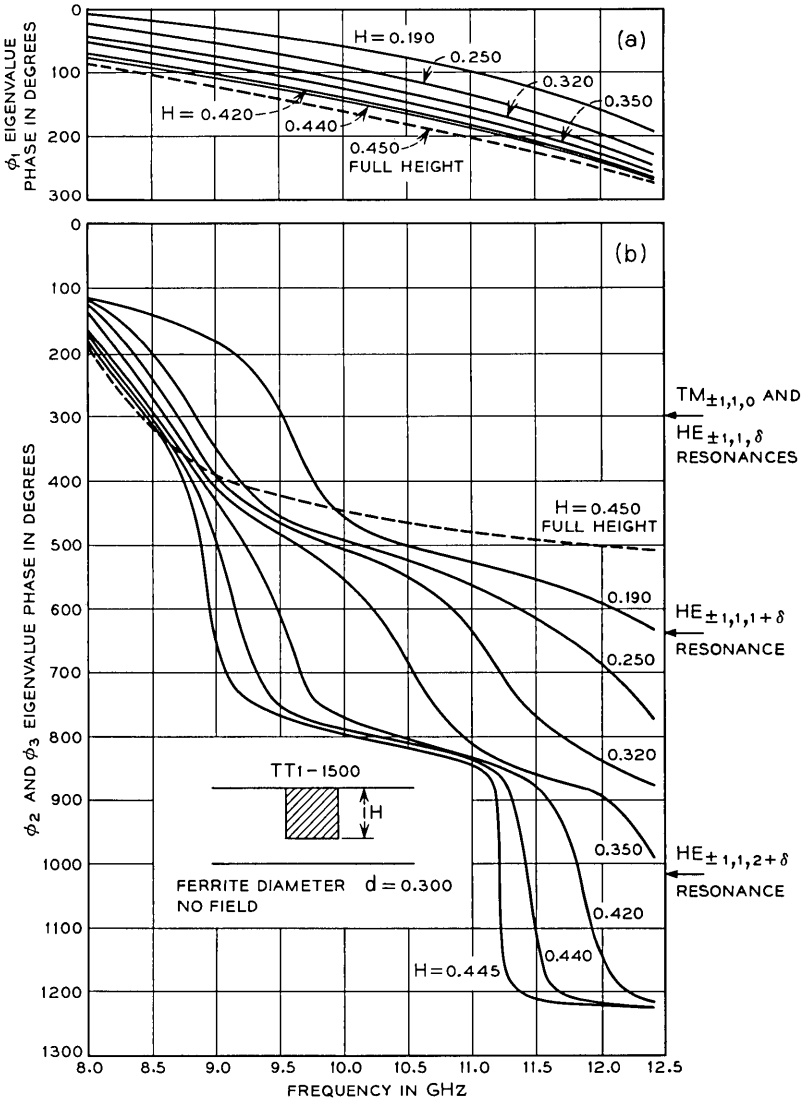


Fig. 17—The eigenvalue phase-frequency response of a junction with a partial height ferrite post. The ferrite diameter is selected to position the $TM_{\pm 1,1,0}$ resonance at the low end of the band. The results show (a) the phase of ϕ_1 is only slightly changed as the ferrite height is reduced, (b) the axially resonating $HE_{\pm m,n,z}$ modes in $\angle\phi_2$ and $\angle\phi_3$ increase in frequency.

bandwidth. Clearly the eigenvalue displacement with field must be in the same direction in both cases. This is automatically satisfied if the modes have the same azimuthal, m, variation. The results for one such "stagger-tuned" circulator are shown in Fig. 18. The $TM_{\pm 1,1,0}$ resonance is centered at 8.4 GHz, and the $HE_{\pm 1,1,1+\delta}$ resonance at 9.9 GHz. With biasing field applied, $\angle\phi_2$ and $\angle\phi_3$ are now displaced by 120 degrees over a wider band. A $TM_{0,1,\delta}$ pin-induced resonance provides the correct displacement and slope for $\angle\phi_1$. The resultant circulator has a 20-dB isolation bandwidth of 700 MHz (7.6 percent).

The coupling to the HE modes may be from one side of the ferrite

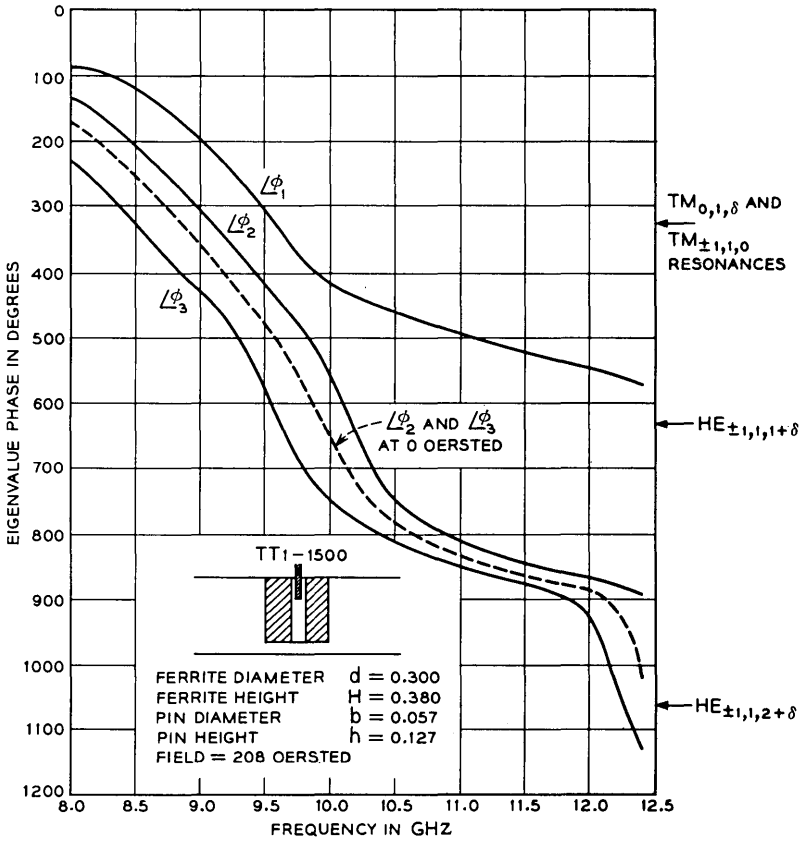


Fig. 18—The eigenvalue phase-frequency response of a junction with a partial height and pinned ferrite. The ferrite height and diameter have been selected to "stagger" the $TM_{\pm 1,1,0}$ and $HE_{\pm 1,1,1+\delta}$ modes, and the $TM_{0,1,\delta}$ pin-induced resonance has been adjusted for circulation at 9.3 GHz.

only or from both sides simultaneously. This provides for an additional variation. In the double-sided case, low dielectric constant spacers support the ferrite on either side, and provide the discontinuity necessary to launch the axially propagating modes. The resonances now occur when the ferrite is approximately an integral number of half wavelengths long ($\delta \approx 1$), as shown in Fig. 19. The Q-factor is also reduced because of the larger coupling.

4.4 *Junction with a Transformed Ferrite*

The bandwidths of the circulators synthesized so far are less than 10 percent. The limitation is imposed by the junction resonances. The band over which the required displacement is achieved is confined

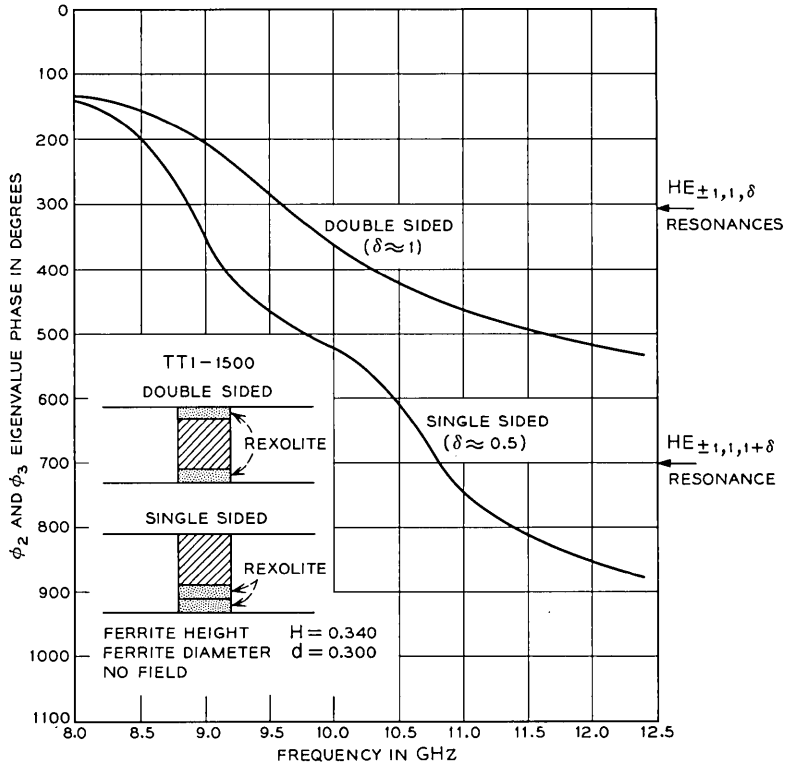


Fig. 19—The phase-frequency response of ϕ_2 and ϕ_3 for a junction with a partial height ferrite of both the single- and double-sided type. In the single-sided case, the HE resonances occur when the ferrite is approximately an integral number of quarter wavelengths long ($\delta \approx 0.5$). In the double-sided case, they occur when the ferrite is approximately an integral number of half wavelengths long ($\delta \approx 1$).

to the immediate vicinity of the rotating mode resonances. Staggering $HE_{\pm m,n,z}$ and $TM_{\pm m,n,0}$ resonances as in Fig. 18 provides little relief. The resonance of the in-phase eigen-excitation then limits the bandwidth.

Several methods exist for overcoming this restriction. The most commonly used technique is to introduce a quarter-wave transformer into the junction. This may be formed from a dielectric cylinder surrounding the ferrite, or from a metallic pedestal placed between the ferrite and the broadwall of the waveguide junction. Only the metal transformer type is considered here. Its effect on the eigenvalues is shown in Fig. 20. The radial difference, w, between the ferrite and the transformer is one quarter wavelength in free space at 9.2 GHz. The transformer step height is t. With no transformer present ($t = 0$), the $TM_{\pm 1,1,0}$ resonance is centered at 9.5 GHz. As t is increased, the phase-frequency responses of ϕ_2 and ϕ_3 become more linear. An optimum is reached when t is approximately 0.130 inch. The linearity permits a more constant displacement of $\angle\phi_2$ and $\angle\phi_3$ with field as shown in Fig. 21. Referring again to Fig. 20, a spurious resonance moves into the band when $t > 0.130$ inch. This is an evanescent mode associated with the transformer. Little energy is coupled to the ferrite in its vicinity, and no displacement occurs when field is applied. Figure 20 also displays the effect of the transformer on the phase of ϕ_1 . For $t = 0$, $\angle\phi_1$ is essentially linear with frequency. As t is increased, $\angle\phi_1$ takes on a curvature suggesting a low-Q resonance near or just below the lower end of the band.

The transformer and full height ferrite geometry of Fig. 20 does not directly lend itself to broadband circulation. The transformer cannot easily equalize the phase-frequency slopes of the eigenvalues, and simultaneously arrange for their correct displacement. Another degree of freedom is required for the adjustment of the eigenvalues, and this accounts for the use of partial height ferrites in many existing designs. From Section 4.3, it is apparent that a central conducting pin would also provide the additional degree of freedom required.

An example in the use of the pin is shown in Fig. 22. Except for the hole in the ferrite center and the pin, the junction is identical to the one shown in Fig. 21. The transformer linearizes $\angle\phi_2$ and $\angle\phi_3$ in the vicinity of the $TM_{\pm 1,1,0}$ resonance at 9.5 GHz, and a biasing field is applied to displace these eigenvalues by 120 degrees over a wide band. The pin-induced $TM_{0,1,\delta}$ resonance in $\angle\phi_1$ is likewise linearized by the transformer, and its position adjusted for optimum circulation. The 20-dB isolation bandwidth is 1.35 GHz (15.2 percent).

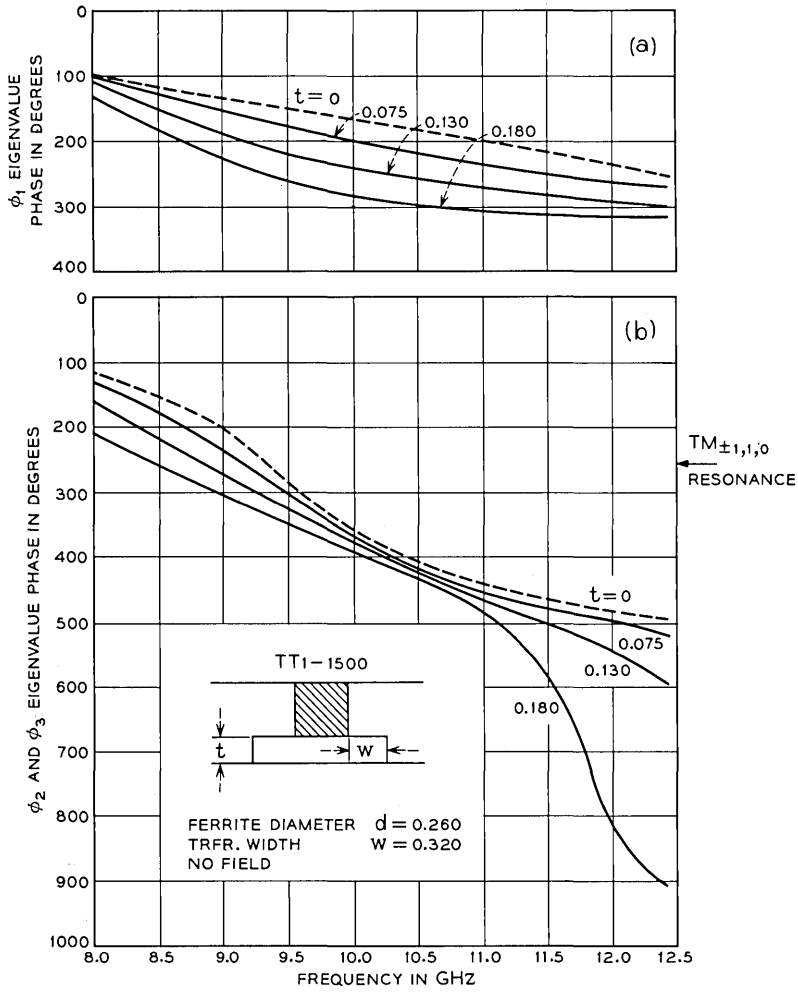


Fig. 20—The eigenvalue-phase frequency response of a junction with a full height ferrite post and different height transformers. The results show (a) the curvature in $\angle\phi_1$ as the transformer height is increased and (b) the linearization of $\angle\phi_2$ and $\angle\phi_3$ in the vicinity of the $TM_{\pm 1,1,0}$ resonance.

The effect of the transformer is essentially the same for both the $TM_{\pm m,n,0}$ and $HE_{\pm m,n,z}$ modes. Recalling that these modes are differentially affected by the ferrite height, a very wideband circulator of this nature then becomes possible by having both a $TM_{\pm m,n,0}$ and a $HE_{\pm m,n,z}$ resonance within band. With a suitable transformer the

eigenvalue linearization can extend into both resonance regions. The proper combination of pin height, transformer, and applied field can also provide linearization for a pin-induced resonance in $\angle\phi_1$, and the correct mutual eigenvalue displacement over this extended range. Such a circulator synthesized in standard 0.900-inch wide and 0.400-inch high X-band waveguide is shown in Fig. 23. A slight recess in the transformer directly beneath the ferrite provides the desired coupling for the $HE_{\pm 1,1,1+\delta}$ mode, and a sleeve around the transformer provides minor adjustments for an optimum 20-dB bandwidth. The losses calculated from the eigenvalue phases using equations (11) and (12) are compared with those measured on a transmission loss test set in Fig. 24. The 20-dB isolation bandwidth is 3.1 GHz (31 percent) centered at 10 GHz. It nearly extends over the full waveguide band.

V. CONCLUSIONS

Examining a symmetrical nonreciprocal 3-port junction in terms of its eigenvalues is an extremely useful method for determining the mode of operation of a Y-circulator. It provides a direct relation between the scattering matrix theory, the inner junction modes, and the resultant device characteristics. For many circulator geometries, the exact nature

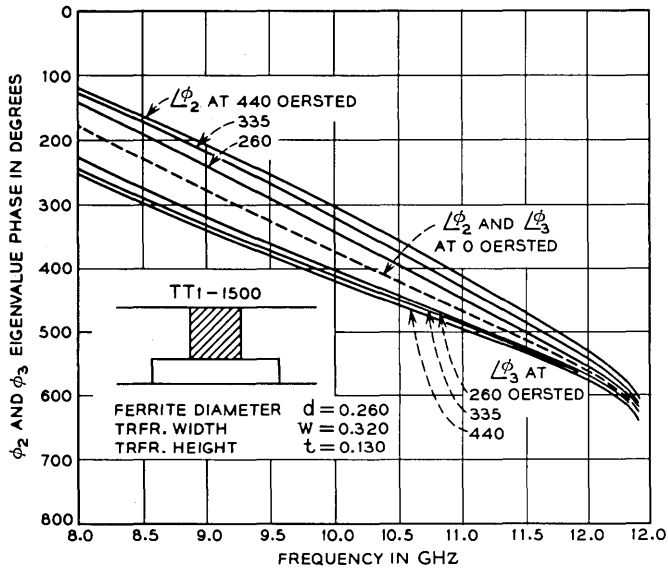


Fig. 21—The phase-frequency response of ϕ_2 and ϕ_3 for a junction with a transformer and full height ferrite post. The linearization due to the transformer permits a more even displacement of the eigenvalues with applied field.

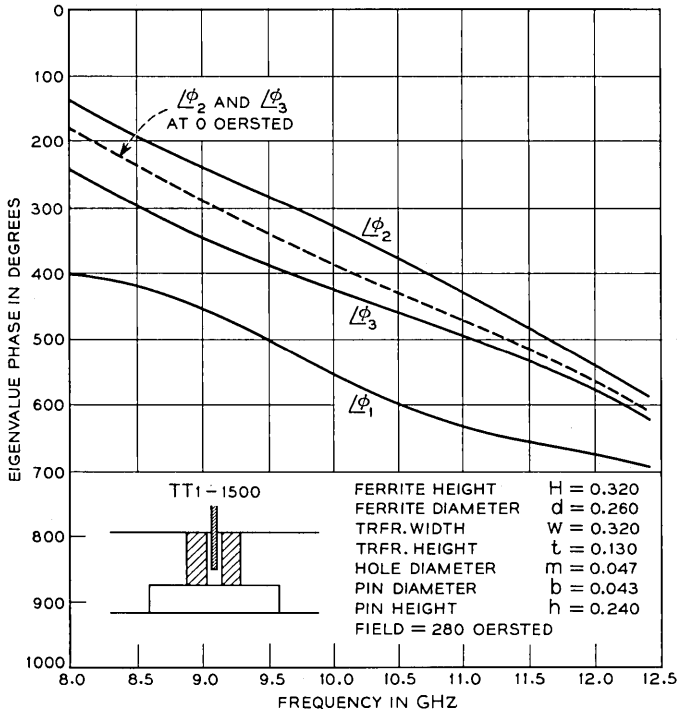


Fig. 22—The eigenvalue phase-frequency response of a junction with a transformer and pinned, full height ferrite. The eigenvalues have been linearized by the transformer in the vicinity of the $TM_{\pm 1,1,0}$ and $TM_{0,1,\delta}$ resonances. The field and pin have been adjusted for circulation.

of this relationship was not clearly understood before, and little was known about their modes of operation. The principal geometries considered here together with their modal responses are summarized in Table I.

The eigenvalue measuring set easily lends itself to direct circulator synthesis. It is much more powerful than the impedance plotting techniques previously used, and provides a quick-and-easy way of "arranging" for circulation. Only the phases of the eigenvalues need be considered when the losses are small. If losses are excessive or if a higher accuracy is required, then the eigenvalue magnitudes must also be taken into account. Since network analyzers are designed to measure both phase and amplitude, this does not present any measurement problem. The analysis of the results, however, is appreciably more difficult.

VI. ACKNOWLEDGMENTS

The author wishes to thank T. W. Mohr and J. J. Kostelnick for useful discussion, and K. P. Steinmetz for mechanical contributions. The advice, guidance, and contributions of C. E. Barnes throughout this work was invaluable and is greatly appreciated.

APPENDIX

From equation (8) and (3) respectively, the junction S parameters are given by,

$$\begin{aligned} S_{11} &= 1/3\{\phi_1 + \phi_2 + \phi_3\} \\ S_{31} &= 1/3\{\phi_1 + \phi_2 \exp(-j120^\circ) + \phi_3 \exp(+j120^\circ)\} \\ S_{21} &= 1/3\{\phi_1 + \phi_2 \exp(+j120^\circ) + \phi_3 \exp(-j120^\circ)\} \end{aligned} \tag{15}$$

where

$$\begin{aligned} S_{11} &= S_{22} = S_{33} \\ S_{31} &= S_{12} = S_{23} \\ S_{21} &= S_{13} = S_{32} \,. \end{aligned} \tag{16}$$

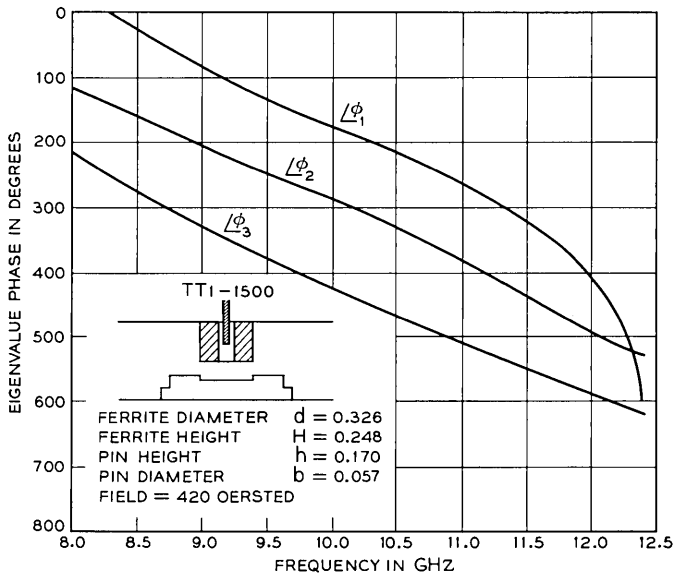


Fig. 23—The eigenvalue phase-frequency response of a very wideband circulator synthesized in a junction made from standard 0.900-inch wide and 0.400-inch high X-band waveguide.

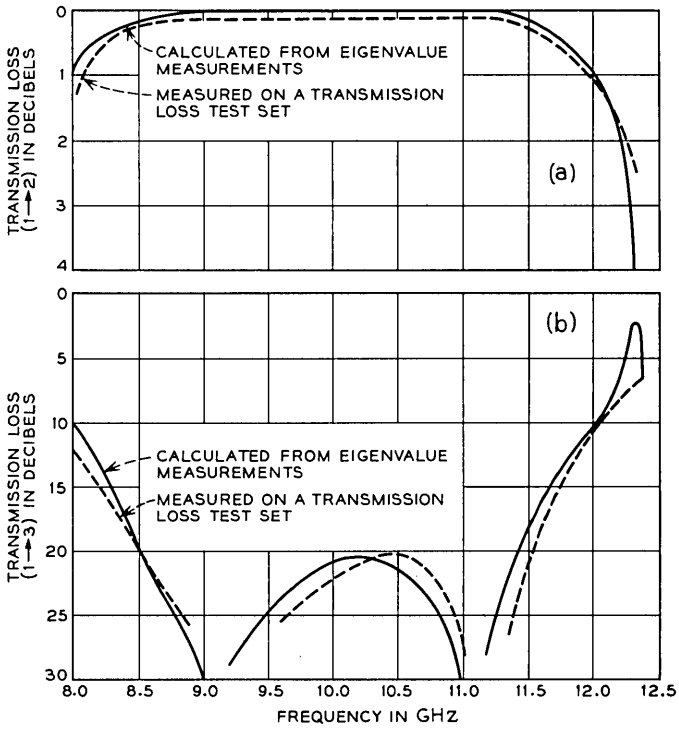


Fig. 24—Transmission loss from (a) ports 1 to 2 and (b) ports 1 to 3 for the junction in Fig. 23. The losses calculated from the eigenvalue phase are compared with those measured on a loss test set.

In more general terms, equation (1) is given by

$$S_{pq} = 1/3 \sum_{i=1}^{i=3} \alpha_i^{(pq)} \tag{17}$$

where S_{pq} is the ratio of the signal out of port p to the signal in at port q when only port q is excited; and $\alpha_i^{(pq)}$ is the ratio of the signal out of port p to the signal in at port q when the junction is excited by the ith eigen-excitation. When

(*i*) $p = q = 1$, then

$$S_{pq} = S_{11} \quad \text{and} \quad \alpha_i^{(11)} = \phi_i \ ,$$

(*ii*) $p = 3$ and $q = 1$, then

$$S_{pq} = S_{31} \quad \text{and} \quad \alpha_i^{(31)} = \phi_i \exp\{-j(i-1)120^\circ\}, \tag{18}$$

(*iii*) $p = 2$ and $q = 1$, then

$$S_{pq} = S_{21} \quad \text{and} \quad \alpha_i^{(21)} = \phi_i \exp\{+j(i-1)120^\circ\}.$$

We are interested in conditions in the junction when the S parameters are either zero or unity. Three possibilities exist, namely,

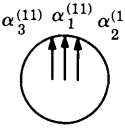
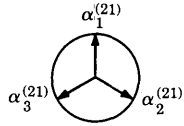
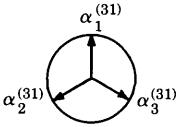
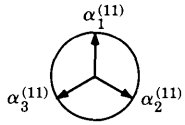
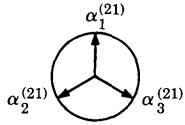
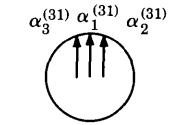
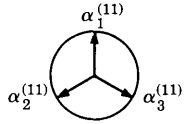
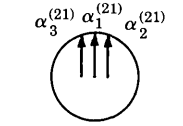
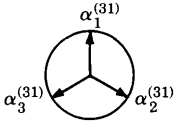
(*i*) $S_{11} = 1$ $S_{31} = 0$ $S_{21} = 0$ Total Reflection,
(*ii*) $S_{11} = 0$ $S_{31} = 1$ $S_{21} = 0$ Anticlockwise Circulation,
and (*iii*) $S_{11} = 0$ $S_{31} = 0$ $S_{21} = 1$ Clockwise Circulation.

S_{pq} is zero when the three components $\alpha_i^{(pq)}$ ($i = 1, 2, 3$) are phase displaced by 120 degrees from one another; and S_{pq} is unity when the same three components are in phase with each other. Suppose $\alpha_i^{(pq)}$ is measured for the three eigen-excitations by sampling the output signal on port p and the input signal on port q. Table II shows how these components must be displaced from one another for total reflection and for circulation. When the components measured are in phase with each other and $p \neq q$, the junction circulates. This provides a superior alternative to the 120 degrees displacement requirement normally aimed for in synthesis, since deviations from the ideal are so much more obvious. When the components measured are separated by 120 degrees from one another, the junction may be circulating or may be totally reflecting. One can determine which condition exists by noting the sequency of the measured phases. If circulation is indicated, a reversal of the biasing field will cause the measured components to coincide once more for the easy detection of variations.

TABLE I—A SUMMARY OF THE DIFFERENT RESONANT MODES EXCITED BY THE IN-PHASE AND ROTATING EIGEN-EXCITATIONS FOR DIFFERENT FERRITE GEOMETRIES

Ferrite Configuration	In-Phase Modes		Rotating Modes	
	Radial $TM_{0,n,0}$	Axial $TM_{0,n,z}$	Radial $TM_{\pm m,n,0}$	Axial $HE_{\pm m,n,z}$
Full Height	present		present	
Pinned and full height	present	present	present	present with large dia. pin only
Partial height	present		present	present
Pinned and partial height	present	present	present	present

TABLE II—PHASE REQUIREMENTS FOR CIRCULATION AND TOTAL REFLECTION WHEN THE INCIDENT AND REFLECTED WAVES ARE SAMPLED AT PORTS q AND p RESPECTIVELY

	Outgoing Signal Sampled at Port p Where		
	p=1	p=2	p=3
Ingoing Signal Sampled At Port q Where q=1.	Total Reflection When 	Total Reflection When 	Total Reflection When 
	Anticlockwise Circulation When 	Anticlockwise Circulation When 	Anticlockwise Circulation When 
	Clockwise Circulation When 	Clockwise Circulation When 	Clockwise Circulation When 

REFERENCES

1. Chait, H. N., and Curry, T. R., "Y Circulator," J. Appl. Phys., *30*, April 1959, p. 152.
2. Auld, B. A., "The Synthesis of Symmetrical Waveguide Circulators," IRE Trans. Microwave Theory and Techniques, *MTT-7*, No. 4 (April 1959, pp. 238–246.
3. Bosma, H., "On the Principle of Stripline Circulators," Proc. IEE, *109*, Part B Suppl., No. 21 (January 1962), pp. 137–146.
4. Fay, C. E., and Comstock, R. L., "Operation of the Ferrite Junction Circulator," IEEE Trans. Microwave Theory and Techniques, *MTT-13*, No .1 (January 1965), pp. 15–27.
5. Von Aulock, W. H., and Fay, C. E., *Linear Ferrite Devices for Microwave Application*, New York: Academic Press, 1968, p. 116.
6. Davies, J. B., "An Analysis of the m-Port Symmetrical H-Plane Waveguide Junction with Central Ferrite Post," IRE Trans. Microwave Theory and Techniques, *MTT-10*, No. 11 (November 1962), pp. 596–604.
7. Butterweck, H. J., "The Y Circulator," Arch. Elek. Ubertragung, *17*, No. 4 (December 1963), pp. 163–176.
8. Davies, J. B., "Theoretical Design of Wideband Waveguide Circulators," Electronics Letters, *1*, No. 3 (May 1965), pp. 60–61.
9. Parsonson, C. G., Longley, S. R., and Davies, J. B., "The Theoretical Design

of Broadband 3-Port Waveguide Circulators," IEEE Trans. Microwave Theory and Techniques, *MTT-16*, No. 4 (April 1968), pp. 256–258.
10. Castillo, J. B., Jr., and Davis, L. E., "Computer Aided Design of 3-Port Waveguide Junction Circulators," IEEE Trans. Microwave Theory and Techniques, *MTT-18*, No. 1 (January 1970), pp. 25–34.
11. Montgomery, C. G., Dickie, R. H., and Purcell, E. M., *Principles of Microwave Circuits*, New York: McGraw-Hill Book Co., Inc., 1948, p. 420.
12. Altman, J., *Microwave Circuits*, New York: D. Van Nostrand Co. Inc., 1964, p. 101.
13. Kurokawa, K., *An Introduction to the Theory of Microwave Circuits*, New York: Academic Press, 1969, p. 234.
14. Magalhaes, F. M., private communication.
15. Owen, B., and Barnes, C. E., "The Compact Turnstile Circulator," IEEE Trans. Microwave Theory and Techniques, *MTT-18*, No. 12 (December 1970), pp. 1096–1100.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 51, No. 3, March, 1972
Printed in U.S.A.

On Maxentropic Discrete Stationary Processes

By D. SLEPIAN

(Manuscript received September 24, 1971)

This paper is concerned with the following mathematical problem. Let **X** *denote a stationary time-discrete random process whose variables,* $\cdots, X_{-1}, X_0, X_1, \cdots$, *take values from the finite set of real numbers* $\{x_1, x_2, \cdots, x_K\}$. *Let* **X** *have mean zero and a given covariance sequence* $\rho_k = EX_iX_{i+k}$, $j, k = 0, \pm 1, \pm 2, \cdots$. *What is the largest entropy that* **X** *can have and what is the probability structure of this most random process of given second moments?*

I. INTRODUCTION

Let **X** denote a stationary time-discrete random process whose variables, $\cdots, X_{-1}, X_0, X_1, \cdots$, take values from the finite set of real numbers $\{x_1, x_2, \cdots, x_K\}$. Let **X** have mean zero and a given covariance sequence $\rho_k = EX_iX_{i+k}$, $j, k = 0, \pm 1, \pm 2, \cdots$. What is the largest entropy that **X** can have and what is the probability structure of this most random process of given second moments?

Our interest in this question arose from the consideration of certain pulse-type communication systems used for the transmission of digital data. In such systems, a customer provides data in the form of an infinite sequence of binary digits that can be represented by a stationary process **Y** whose variables, $\cdots, Y_{-1}, Y_0, Y_1, \cdots$, are independent random variables each taking values zero and one with equal probabilities. An encoder transforms **Y** into a K-level process **X** of the sort described above, whose random variables are then used as amplitudes for successive pulses of a train. The transmitted signal is thus of the form

$$s(t) = \sum_{n=-\infty}^{\infty} X_n g(t - nT + \theta) \tag{1}$$

where $g(t)$ is the pulse shape and $T > 0$ is the pulse repetition period

of the system. It is easy to compute that the power density spectrum of the stochastic process (1) is given by

$$\Phi_s(f) = \frac{|\, G(f)\,|^2}{T}\,\Phi_{\mathbf{x}}(fT) \tag{2}$$

where $G(f)$ is the Fourier transform of $g(t)$ and

$$\Phi_{\mathbf{x}}(f) = \sum_{-\infty}^{\infty} \rho_n e^{2\pi i n f} \tag{3}$$

is the spectrum of the discrete-amplitude process **X**. Here it has been assumed that θ is uniformly distributed in $(0, T)$.

Many different encoding schemes for mapping the customer's data stream **Y** onto the pulse amplitude stream **X** have been proposed in the past. Typical are dicode, partial response, pseudo-ternary, run-length-limited codes, etc.. Entry to the literature on this subject can be made through Refs. 1–4. In general, these encoding schemes are employed to give $\Phi_{\mathbf{x}}(f)$, and hence $\Phi_s(f)$, some desirable shape that will be particularly well-suited to the transmission medium, the noise, and the demodulation process. However, such deviations of $\Phi_{\mathbf{x}}(f)$ from a flat shape (Φ = constant) are bought at the price of a decreased information rate for the system as will be seen in an example below. Solution to the problem posed in the opening paragraph would yield the maximum information rate possible with given amplitudes $x_1, x_2, \cdots, x_K$ and given spectrum $\Phi_{\mathbf{x}}(f)$.

To illustrate these matters, consider the simple case of dicode for which the encoding is

$$X_n = Y_n - Y_{n-1}, \quad n = 0, \pm 1, \pm 2, \cdots .$$

Here $K = 3$ and the allowed pulse amplitudes are $x_1 = 1$, $x_2 = 0$, $x_3 = -1$. It is readily computed that for this amplitude-process $\rho_0 = \frac{1}{2}$, $\rho_1 = \rho_{-1} = -\frac{1}{4}$, and $\rho_n = 0$ for $n = \pm 2, \pm 3, \cdots$ and so $\Phi_{\text{dicode}}(f) = \sin^2 \pi f$. This spectrum vanishes like f^2 at zero frequency, a frequently desirable property. But, this 3-level scheme signals at a rate of only one bit of information per pulse whereas a rate of $\log_2 3 = 1.58$ bits per pulse could be had by appropriate mapping of the customer's binary digits onto independent random variables taking the same amplitude values, -1, 0, and 1 each with probability 1/3. This latter encoding would, of course, yield a flat spectrum. Thus dicode achieves a desired spectrum at the cost of about a 1/3 decrease in rate. Can any scheme with the same values and spectrum as dicode attain a rate

greater than one bit per pulse? What is the highest rate so achievable?

We have been unable to answer even these seemingly simple specific questions. Quite apart from applications to pulse-amplitude data transmission systems, the general question of finding a maxentropic finite state discrete process of given second moments is of interest in its own right. As we shall see, such a process is a natural finite state analog of the Gaussian process and could serve as a convenient model in many contexts. We have been able to make only slight progress in solving this more general problem.

It is the purpose of this paper to record the progress we have made and the approaches we have followed in pursuing these goals, and to exhibit the difficulties encountered as well. It is hoped that others who may become interested in this problem can thereby avoid some pitfalls and be guided to more successful approaches.

II. REDUCTION TO THE MARKOV CASE

Let $\mathbf{X}$ be a stationary process $\cdots, X_{-1}, X_0, X_1, \cdots$ where each X takes values from the set of K real numbers $\{x_1, x_2, \cdots, x_K\}$. We denote the probability distribution of n successive Xs by

$$p_n(\epsilon_1, \cdots, \epsilon_n) = \Pr\{X_{i+1} = x_{\epsilon_1}, \cdots, X_{i+n} = x_{\epsilon_n}\}. \tag{4}$$

Here each index $\epsilon_1, \epsilon_2, \cdots, \epsilon_n$ takes values from the set $\{1, 2, \cdots, K\}$. We have, of course,

$$\sum_{\epsilon} p_n(\epsilon_1, \epsilon_2, \cdots, \epsilon_n) = 1 \tag{5}$$

and

$$p_n(\epsilon_1, \cdots, \epsilon_n) \geqq 0, \quad \epsilon_1, \epsilon_2, \cdots, \epsilon_n = 1, 2, \cdots, K. \tag{6}$$

The stationarity of $\mathbf{X}$ implies that the left of (4) is independent of i, and furthermore that

$$\sum_{\alpha=1}^{K} p_n(\epsilon_1, \epsilon_2, \cdots, \epsilon_{n-1}, \alpha) = \sum_{\alpha=1}^{K} p_n(\alpha, \epsilon_1, \epsilon_2, \cdots, \epsilon_{n-1})$$

$$= p_{n-1}(\epsilon_1, \cdots, \epsilon_{n-1}), \qquad \epsilon_1, \epsilon_2, \cdots, \epsilon_{n-1} = 1, 2, \cdots, K. \tag{7}$$

The statements (4) through (6) are to hold for $n = 1, 2, \cdots$ and (7) for $n = 2, 3, \cdots$. Note that if (7) holds for $n = n_o$, this implies the validity of (7) for $n = 2, 3, \cdots, n_o$.

The entropy of $\mathbf{X}$ is defined by

$$H(\mathbf{X}) = \lim_{n\to\infty} \frac{1}{n} H_n(\mathbf{X}), \tag{8}$$

$$H_n(\mathbf{X}) = -\sum p_n(\epsilon_1, \cdots, \epsilon_n) \log p_n(\epsilon_1, \cdots, \epsilon_n), \tag{9}$$

where the sum is over the K^n allowed values of the ϵ's. We seek to maximize (8) by suitable choice of a hierarchy of distributions $p_n(\epsilon_1, \cdots, \epsilon_n)$, $n = 1, 2, \cdots$ that satisfy (5), (6), and (7) and the constraints

$$EX_i = \sum_{\epsilon_1=1}^{K} x_{\epsilon_1} p_1(\epsilon_1) = 0, \tag{10}$$

$$EX_i X_{i+k} = \sum_{\epsilon} x_{\epsilon_1} x_{\epsilon_{k+1}} p_{k+1}(\epsilon_1, \cdots, \epsilon_{k+1}) = \rho_k,$$

$$k = 0, 1, 2, \cdots. \tag{11}$$

Here the ρ_k are given and the sum is over all allowable values of $\epsilon_1, \epsilon_2, \cdots, \epsilon_{k+1}$.

We do not know how to proceed directly with this problem. One approach is to attempt to solve the problem when the constraint (11) is imposed only for $k = 0, 1, 2, \cdots, L$. That is, we seek the process of maximum entropy whose first $L + 1$ covariance elements are prescribed. Let $H^{(L)}(\mathbf{X})$ denote this maximum entropy and let $p_n^{(L)}(\epsilon_1, \cdots, \epsilon_n)$, $n = 1, 2, \cdots$, be the corresponding distribution. We would then investigate the behavior of these quantities as $L \to \infty$. We have, of course, $H^{(L)}(\mathbf{X}) \geqq H(\mathbf{X})$.

In Appendix A we establish

Theorem 1: *The K-valued stationary discrete process of largest entropy with mean zero, given values* $x_1, \cdots, x_K$, *and given values of* $\rho_0, \rho_1, \cdots, \rho_L$ *is an Lth-order Markov process.*

An Lth-order Markov process is characterized by the fact that

$$\Pr\{X_n = x_{\epsilon_n} \mid X_{n-1} = x_{\epsilon_{n-1}}, \cdots, X_{n-L} = x_{\epsilon_{n-L}}, X_{n-L-1} = x_{\epsilon_{n-L-1}}, \cdots\}$$
$$= \Pr\{X_n = x_{\epsilon_n} \mid X_{n-1} = x_{\epsilon_{n-1}}, \cdots, X_{n-L} = x_{\epsilon_{n-L}}\}$$

for all n and all allowable values of the ϵ's. A stationary Lth-order Markov process can be specified by K^{L+1} transition probabilities

$$q_L(\epsilon_{L+1} \mid \epsilon_1, \cdots, \epsilon_L) \equiv \Pr\{X_{L+1} = x_{\epsilon_{L+1}} \mid X_1 = x_{\epsilon_1}, \cdots, X_L = x_{\epsilon_L}\}$$
$$\epsilon_1, \cdots, \epsilon_{L+1} = 1, 2, \cdots, K$$

and a corresponding Lth-order distribution $p_L(\epsilon_1 , \cdots , \epsilon_L)$ that satisfies

$$\sum_{\epsilon_1=1}^{K} q_L(\epsilon_{L+1} \mid \epsilon_1 , \cdots , \epsilon_L)p_L(\epsilon_1 , \cdots , \epsilon_L) = p_L(\epsilon_2 , \cdots , \epsilon_{L+1})$$

$$\epsilon_2 , \cdots , \epsilon_{L+1} = 1, 2, \cdots , K. \tag{12}$$

We have, of course

$$q_L(\epsilon_{L+1} \mid \epsilon_1 , \cdots , \epsilon_L) \geqq 0, \tag{13}$$

$$\sum_{\alpha=1}^{K} q_L(\alpha \mid \epsilon_1 , \cdots , \epsilon_L) = 1, \qquad \epsilon_1 , \cdots , \epsilon_{L+1} = 1, 2, \cdots , K. \tag{14}$$

Equations (12) and (14) guarantee that the normalized solutions p_L of (12) have property (7) (with $n = L$). The general term p_n of the probability distribution for such a process is given in terms of p_L by the product rule

$$p_n(\epsilon_1 , \cdots , \epsilon_n) = p_L(\epsilon_1 , \cdots , \epsilon_L)q_L(\epsilon_{L+1} \mid \epsilon_1 , \cdots , \epsilon_L)$$
$$q_L(\epsilon_{L+2} \mid \epsilon_2 , \cdots , \epsilon_{L+1}) \cdots q_L(\epsilon_n \mid \epsilon_{n-L} , \epsilon_{n-L+1} , \cdots , \epsilon_{n-1}) \tag{15}$$

for $n > L$. For $n < L$,

$$p_n(\epsilon_1 , \cdots , \epsilon_n) = \sum_{\alpha_1=1}^{K} \cdots \sum_{\alpha_{L-n}=1}^{K} p_L(\epsilon_1 , \cdots , \epsilon_n , \alpha_1 , \cdots , \alpha_{L-n}). \tag{16}$$

It is easy to show that for a stationary Lth-order Markov process the entropy (8) through (9) is given by

$$\begin{aligned} H &= -\sum_{\epsilon} p_L(\epsilon_1 , \cdots , \epsilon_L) \sum_{\alpha} q_L(\alpha \mid \epsilon_1 , \cdots , \epsilon_L) \log q_L(\alpha \mid \epsilon_1 , \cdots , \epsilon_L) \\ &= -\sum p_{L+1}(\epsilon_1 , \cdots , \epsilon_{L+1}) \log p_{L+1}(\epsilon_1 , \cdots , \epsilon_{L+1}) \\ &\quad + \sum p_L(\epsilon_1 , \cdots , \epsilon_L) \log p_L(\epsilon_1 , \cdots , \epsilon_L) \\ &= H_{L+1} - H_L . \end{aligned} \tag{17}$$

III. THE DETAILED DISTRIBUTION

Now to find the most random stationary Lth-order Markov process with given $\rho_0 , \rho_1 , \cdots , \rho_L$, we must maximize $H_{L+1} - H_L$ by proper choice of the K^{L+1} quantities $\rho_{L+1}(\epsilon_1 , \cdots , \epsilon_{L+1})$ subject to certain linear constraints of the form

$$\sum_{\epsilon} a_i(\epsilon_1 , \cdots , \epsilon_{L+1})p_{L+1}(\epsilon_1 , \cdots , \epsilon_{L+1}) = b_i \qquad i = 1, 2, \cdots , M. \tag{18}$$

We assume this system is of rank $M' \leqq M$.

There are two ways to proceed: (*i*) by the method of Lagrange multipliers which treats the unknown p_{L+1}'s symmetrically; (*ii*) by expressing $H_{L+1} - H_L$ in terms of $K^{L+1} - M'$ independent p_{L+1}'s obtained by solving (18). Both methods lead to unwieldy higher-order algebraic equations with which we have been able to do little in the general case. The form of the solutions is not without some interest, however, and we record it here.

To avoid unnecessary superficial complications, we shall henceforth assume that if x is one of the allowed values $x_1, x_2, \cdots, x_K$, then $-x$ is also in the set of allowed values. This condition will assure that

$$\Pr(X_1 = x_{\epsilon_1}, \cdots, X_n = x_{\epsilon_n}) = \Pr(X_1 = -x_{\epsilon_1}, \cdots, X_n = -x_{\epsilon_n})$$

in the optimal process and that $EX_j = 0$, $j = 0, \pm 1, \cdots$.

3.1 *Lagrange Multipliers*

Let us define the sample lag sums

$$\begin{aligned}
l_n^{(0)}(\epsilon_1, \cdots, \epsilon_n) &\equiv x_{\epsilon_1}^2 + x_{\epsilon_2}^2 + \cdots + x_{\epsilon_n}^2 \\
l_n^{(1)}(\epsilon_1, \cdots, \epsilon_n) &\equiv x_{\epsilon_1}x_{\epsilon_2} + x_{\epsilon_2}x_{\epsilon_3} + \cdots + x_{\epsilon_{n-1}}x_{\epsilon_n} \\
&\vdots \\
l_n^{(j)}(\epsilon_1, \cdots, \epsilon_n) &\equiv x_{\epsilon_1}x_{\epsilon_{1+j}} + x_{\epsilon_2}x_{\epsilon_{2+j}} + \cdots + x_{\epsilon_{n-j}}x_{\epsilon_n} \\
&\vdots \\
l_n^{(n-1)}(\epsilon_1, \cdots, \epsilon_n) &\equiv x_{\epsilon_1}x_{\epsilon_n}
\end{aligned} \tag{19}$$

and the function

$$h_n(\epsilon_1, \cdots, \epsilon_n; \lambda_0, \lambda_1, \cdots, \lambda_{n-1}) \equiv \exp \sum_{j=0}^{n-1} \lambda_j l_n^{(j)}(\epsilon_1, \cdots, \epsilon_n). \tag{20}$$

Then the Lagrange solution can be stated as follows. Solve the homogeneous system of equations

$$\sum_{\epsilon_{L+1}} h_{L+1}(\epsilon_1, \cdots, \epsilon_{L+1}; \lambda_0, \cdots, \lambda_L) f(\epsilon_2, \cdots, \epsilon_{L+1}) = \frac{1}{c} f(\epsilon_1, \cdots, \epsilon_L) \qquad \epsilon_1, \epsilon_2, \cdots, \epsilon_L = 1, 2, \cdots, K \tag{21}$$

for the K^L f's and c. Then the transition probabilities and initial stationary distribution of the maxentropic process are given by

$$q_L(\epsilon_{L+1} \mid \epsilon_1, \cdots, \epsilon_L) = c h_{L+1}(\epsilon_1, \cdots, \epsilon_{L+1}; \lambda_0, \cdots, \lambda_L) \frac{f(\epsilon_2, \cdots, \epsilon_{L+1})}{f(\epsilon_1, \cdots, \epsilon_L)}, \tag{22}$$

$$p_L(\epsilon_1 , \cdots , \epsilon_L) = kf(\epsilon_1 , \cdots , \epsilon_L)f(\epsilon_L , \cdots , \epsilon_1),$$

$$\epsilon_1 , \cdots , \epsilon_{L+1} = 1, 2, \cdots , K. \qquad (23)$$

In (23), $k > 0$ must be chosen so that the p_L sum to unity. A derivation of these equations is given in Appendix B.

While equations (21), (22), and (23) are a formal solution to our problem, in practice they are of little value. The solutions p_L and q_L contain the Lagrange multipliers $\lambda_0 , \lambda_1 , \cdots , \lambda_L$ in a complicated way, and these must be determined to give the prescribed covariance elements $\rho_0 , \rho_1 , \cdots , \rho_L$. Presumably that eigenvalue c of (21) should be taken which gives maximum entropy and yields $q_L \geqq 0$ in (22). In the small examples we have carried out, p_L and q_L turned out to be independent of the eigenvalue chosen in (21), but we have been unable to prove anything in general about this situation. For particular processes, say the symmetric binary process with $K = 2$, $x_1 = 1$, $x_2 = -1$, for example, equations (21) take a special simple seductive form that suggests the possibility of explicit closed-form solution. We have been unable to find one.

Perhaps the best that can be said for this curious Lagrange solution is that (23) shows clearly that in the maxentropic process $p_L(\epsilon_1 , \cdots , \epsilon_L) = p_L(\epsilon_L , \cdots , \epsilon_1)$. It is not hard to see that the product rule (15) and the form of (22) and (23) propagate this property so that for arbitrary n, $p_n(\epsilon_1 , \cdots , \epsilon_n) = p_n(\epsilon_n , \cdots , \epsilon_1)$. The maxentropic process treats past and future in a symmetric manner.

3.2 *The Independent Variable Approach*

We seek to maximize

$$J'' = -\sum_{\epsilon} p_{L+1}(\epsilon_1 , \cdots , \epsilon_{L+1}) \log p_{L+1}(\epsilon_1 , \cdots , \epsilon_{L+1}) + \sum_{\epsilon} p_L(\epsilon_1 , \cdots , \epsilon_L) \log p_L(\epsilon_1 , \cdots , \epsilon_L) \qquad (24)$$

by choice of the K^{L+1} quantities $p_{L+1}(\epsilon_1 , \cdots , \epsilon_{L+1})$. Here we define

$$p_L(\epsilon_1 , \cdots , \epsilon_L) \equiv \sum_{\alpha} p_{L+1}(\epsilon_1 , \cdots , \epsilon_L , \alpha),$$

$$\epsilon_1 , \cdots , \epsilon_L = 1, 2, \cdots , K. \qquad (25)$$

The p_{L+1}'s must satisfy

$$\sum_{\epsilon} p_{L+1}(\epsilon_1 , \cdots , \epsilon_{L+1}) = 1, \qquad (26)$$

$$\sum_{\alpha} p_{L+1}(\epsilon_1 , \cdots , \epsilon_L , \alpha) - \sum_{\alpha} p_{L+1}(\alpha, \epsilon_1 , \cdots , \epsilon_L) = 0,$$

$$\epsilon_1 , \cdots , \epsilon_L = 1, 2, \cdots , K, \tag{27}$$

$$\sum_{\epsilon} x_{\epsilon_1} x_{\epsilon_{k+1}} p_{L+1}(\epsilon_1 , \cdots , \epsilon_{L+1}) = \rho_k , \qquad k = 0, 1, \cdots , L. \tag{28}$$

These $K^L + L + 2 \equiv M$ equations are of the form (18). We suppose that they can be solved for $M' \leqq M$ of the p_{L+1}'s in terms of the remaining $K^{L+1} - M' \equiv M''$ ones. We denote these M'' independent p_{L+1}'s by the variables $\xi_1 , \xi_2 , \cdots , \xi_{M''}$ and denote the M' dependent p_{L+1}'s by $\eta_1 , \cdots , \eta_{M'}$. Thus we write equations (26), (27), and (28) in the form

$$\eta_i = \alpha_i + \sum_{j=1}^{M''} \beta_{ij} \xi_j , \qquad i = 1, 2, \cdots , M'. \tag{29}$$

It is convenient also to adopt a single index notation for the p_L of (25) which we now denote by $\zeta_1 , \zeta_2 , \cdots , \zeta_N$, where $N = K^L$. By means of (29) the right of (25) can be expressed in terms of the ξ's. We write

$$\zeta_i = \delta_i + \sum_{j=1}^{M''} \gamma_{ij} \xi_j , \qquad i = 1, 2, \cdots , K^L. \tag{30}$$

We further note that (26) and (25) imply that

$$\sum_{1}^{M'} \xi_i + \sum_{1}^{M''} \eta_i = 1, \qquad \sum_{1}^{N} \zeta_i = 1.$$

Since these are to hold as identities in the ξ's we must have

$$\sum_{1}^{M'} \alpha_i = \sum_{1}^{N} \delta_i = 1, \qquad \sum_{i=1}^{N} \gamma_{ij} = 1 + \sum_{i=1}^{M'} \beta_{ij} = 0,$$

$$j = 1, 2, \cdots , M''. \tag{31}$$

In this new notation (24) becomes

$$J'' = -\sum_{1}^{M''} \xi_i \log \xi_i - \sum_{1}^{M'} \eta_i \log \eta_i + \sum_{1}^{N} \zeta_i \log \zeta_i$$

where the η's and ζ's are explicit linear functions of the ξ's given by (29) and (30). Maximizing with respect to the ξ's gives

$$\frac{\partial J''}{\partial \xi_j} = 0 = -1 - \log \xi_j - \sum_{i=1}^{M'} (1 + \log \eta_i)\beta_{ij} + \sum_{1}^{N} (1 + \log \zeta_i)\gamma_{ij} .$$

On taking account of (31), we find finally

$$\xi_j = \frac{\prod_{i}^{N} \xi_i^{\gamma_{ij}}}{\prod_{i}^{M'} \eta_i^{\beta_{ij}}}, \qquad j = 1, \cdots, M''. \tag{32}$$

These are M'' equations for the M'' ξ's. There seems to be little that can be said in general about them, and so the trail ends here. (We note only that the form of these equations is appropriate for an iterative numerical solution: a trial set of ξ's used to evaluate the products yields directly a new set of ξ's.)

IV. SOME SIMPLE EXAMPLES

We consider first the binary case and set

$$x_1 = 1, \quad x_2 = -1, \tag{33}$$

In this case we must take $\rho_0 = 1$.

When $L = 1$, we find

$$p(1, 1) = p(2, 2) = \tfrac{1}{4}(1 + \rho_1)$$

$$p(1, 2) = p(2, 1) = \tfrac{1}{4}(1 - \rho_1)$$

$$H = -\tfrac{1}{2}(1 + \rho_1) \log \tfrac{1}{2}(1 + \rho_1) - \tfrac{1}{2}(1 - \rho_1) \log \tfrac{1}{2}(1 - \rho_1)$$

$$\rho_n = \rho_1^n, \qquad n = 0, 1, 2, \cdots,$$

$$\Phi_x(f) = \frac{1 - \rho_1^2}{1 + \rho_1^2 - 2\rho_1 \cos 2\pi f}.$$

When $L = 2$,

$$p(1, 1, 1) = p(2, 2, 2) = \tfrac{1}{8}(1 + 2\rho_1 + \rho_2)$$

$$p(1, 1, 2) = p(2, 2, 1) = \tfrac{1}{8}(1 - \rho_2)$$

$$p(1, 2, 2) = p(2, 1, 1) = \tfrac{1}{8}(1 - \rho_2)$$

$$p(1, 2, 1) = p(2, 1, 2) = \tfrac{1}{8}(1 - 2\rho_1 + \rho_2).$$

Let

$$\alpha_\pm = \frac{1}{2(1 - \rho_1^2)} [\rho_1(1 - \rho_2) \pm \sqrt{4\rho_1^4 + \rho_1^2(\rho_2^2 - 6\rho_2 - 3) + 4\rho_2}].$$

Then

$$\rho_n = \frac{1}{(1 + \alpha_+\alpha_-)(\alpha_+ - \alpha_-)} [(1 - \alpha_-^2)\alpha_+^{n+1} - (1 - \alpha_+^2)\alpha_-^{n+1}]$$

$$\Phi_x(f) = \frac{(1 - \rho_1^2)(1 - \rho_2)(1 - 2\rho_1^2 + \rho_2)}{1 - \rho_1^2 + 2\rho_1^4 - 4\rho_1^2\rho_2 + \rho_2^2 + \rho_1^2\rho_2^2 - 2\rho_1(1 - \rho_2)^2 \cos 2\pi f + 2(\rho_1^2 - \rho_2)(1 - \rho_1^2) \cos 4\pi f}.$$

When $L = 3$, we are already in algebraic difficulties. Equations (26), (27), and (28) in the present case permit us to solve for all the p's in terms of $\xi \equiv p(1, 1, 1, 1)$. We find

$$\begin{aligned}
p(1, 1, 1, 2) = p(1, 2, 2, 2) &= p(2, 1, 1, 1) = p(2, 2, 2, 1) \\
&= \tfrac{1}{8}(1 + 2\rho_1 + \rho_2) - \xi \\
p(1, 1, 2, 1) = p(1, 2, 1, 1) &= p(2, 1, 2, 2) = p(2, 2, 1, 2) \\
&= \tfrac{1}{8}(1 + \rho_1 + \rho_2 + \rho_3) - \xi \\
p(1, 1, 2, 2) = p(2, 2, 1, 1) &= \tfrac{1}{8}(1 - \rho_1 - 2\rho_2 - \rho_3) + \xi \\
p(1, 2, 1, 2) = p(2, 1, 2, 1) &= \tfrac{1}{8}(-3\rho_1 - \rho_3) + \xi \\
p(1, 2, 2, 1) = p(2, 1, 1, 2) &= \tfrac{1}{8}(-2\rho_1 - 2\rho_2) + \xi \\
p(2, 2, 2, 2) &= \xi.
\end{aligned}$$

On setting $Z = 8\xi$, equation (32) becomes

$$Z = \frac{[1 + 2\rho_1 + \rho_2 - Z]^4[1 + \rho_1 + \rho_2 + \rho_3 - Z]^4}{Z[-3\rho_1 - \rho_3 + Z]^2[-\rho_1 - 2\rho_2 - \rho_3 + Z]^2[-2\rho_1 - 2\rho_2 + Z]^2}.$$

One can take the square root of both sides of this equation, clear fractions, and expand to obtain a cubic equation in Z. It is not hard to show that there are no roots rational in ρ_1 , ρ_2 , and ρ_3 , so that the simple dependence of p on the ρ's exhibited for the cases $L = 1$ and 2 fails here.

We next consider the case $K = 3$ and choose

$$x_1 = 1, \quad x_2 = 0, \quad x_3 = -1.$$

Here with $L = 1$ we already meet with higher-degree algebraic equations. The constraints permit solution of all p's in terms of $p(1, 1) = \xi$. We find

$$\begin{aligned}
p(1, 2) = p(2, 1) = p(2, 3) = p(3, 2) &= \tfrac{1}{2}(\rho_0 + \rho_1) - 2\xi \\
p(3, 1) = p(1, 3) &= -\tfrac{1}{2}\rho_1 + \xi \\
p(2, 2) &= 1 - 2\rho_0 - \rho_1 + 4\xi \\
p(3, 3) &= \xi.
\end{aligned}$$

On setting $Z = 4\xi$, equation (32) becomes

$$Z = \pm\frac{(\rho_0 + \rho_1 - Z)^4}{(-2\rho_1 + Z)(1 - 2\rho_0 - \rho_1 + Z)^2} \tag{34}$$

which is a cubic in Z. One finds

$$\rho_n = \rho_0\left(\frac{\rho_1}{\rho_0}\right)^n, \qquad n = 0, 1, 2, \cdots \tag{35}$$

quite independent of the value chosen for ξ. The spectrum is given by

$$\Phi_x(f) = \frac{\rho_0(\rho_0^2 - \rho_1^2)}{\rho_0^2 + \rho_1^2 - 2\rho_0\rho_1 \cos 2\pi f}. \tag{36}$$

Using the dicode values $\rho_0 = \frac{1}{2}$, $\rho_1 = -\frac{1}{4}$, one finds from (34) that $\xi = 0.0103$. The entropy of the resulting simple Markov process is found to be 1.299 bits, which is greater than the one-bit rate of dicode. While the first two terms of the covariance sequence agree with the dicode values, the higher terms are given by (35) and the spectrum, as given by (36), does not vanish for $f = 0$.

The case $K = 3$, $L = 2$ begins to reveal the complexity of the general case. We denote each of the 27 quantities $p(i, j, k)$ by x with a subscript ranging from 1 to 27. The association is made by listing the p's in order, interpreting (ijk) as a three-digit number. Thus $x_1 = p(1, 1, 1)$, $x_2 = p(1, 1, 2)$, $x_3 = p(1, 1, 3)$, $\cdots$, $x_{27} = p(3, 3, 3)$. Equations (26), (27), and (28) can be solved to express all the x's in terms of five of them. Equations (29) are

$$\begin{aligned}
\eta_1 &= -\tfrac{1}{4} + \rho_1 - \tfrac{1}{4}\rho_2 - \xi_1 + 3\xi_2 + 4\xi_3 + \tfrac{3}{2}\xi_4 + \tfrac{1}{4}\xi_5 \\
\eta_2 &= \tfrac{1}{4} - \tfrac{1}{2}\rho_1 + \tfrac{1}{4}\rho_2 + \xi_1 - 2\xi_2 - 3\xi_3 - \tfrac{3}{2}\xi_4 - \tfrac{1}{4}\xi_5 \\
\eta_3 &= \tfrac{1}{8} + \tfrac{1}{4}\rho_0 - \tfrac{1}{8}\rho_2 - \tfrac{1}{2}\xi_1 + \tfrac{1}{4}\xi_4 + \tfrac{1}{8}\xi_5 \\
\eta_4 &= \tfrac{1}{2}\rho_0 - \rho_1 + \tfrac{1}{2}\rho_2 + \xi_1 - 4\xi_2 - 4\xi_3 - \xi_4 \\
\eta_5 &= \tfrac{1}{4} - \tfrac{1}{2}\rho_0 + \tfrac{1}{2}\rho_1 - \tfrac{1}{4}\rho_2 - \xi_1 + 2\xi_2 + 2\xi_3 + \tfrac{1}{2}\xi_4 - \tfrac{1}{4}\xi_5 \\
\eta_1 &= x_1 = x_{27} \\
\eta_2 &= x_2 = x_{10} = x_{18} = x_{26} \\
\eta_3 &= x_3 = x_9 = x_{19} = x_{25} \\
\eta_4 &= x_4 = x_{24} \\
\eta_5 &= x_5 = x_{13} = x_{15} = x_{23} \\
\xi_1 &= x_6 = x_{22} \\
\xi_2 &= x_7 = x_{21} \\
\xi_3 &= x_8 = x_{12} = x_{16} = x_{20} \\
\xi_4 &= x_{11} = x_{17} \\
\xi_5 &= x_{14}\ .
\end{aligned} \tag{37}$$

Equations (30) become

$$\begin{aligned}
\zeta_1 &= -\tfrac{1}{8} + \tfrac{1}{4}\rho_0 + \tfrac{1}{2}\rho_1 - \tfrac{1}{8}\rho_2 - \tfrac{1}{2}\xi_1 + \xi_2 + \xi_3 + \tfrac{1}{4}\xi_4 + \tfrac{1}{8}\xi_5 \\
\zeta_2 &= \tfrac{1}{4} - \tfrac{1}{2}\rho_1 + \tfrac{1}{4}\rho_2 + \xi_1 - 2\xi_2 - 2\xi_3 - \tfrac{1}{2}\xi_4 - \tfrac{1}{4}\xi_5 \\
\zeta_3 &= -\tfrac{1}{8} + \tfrac{1}{4}\rho_0 - \tfrac{1}{8}\rho_2 - \tfrac{1}{2}\xi_1 + \xi_2 + \xi_3 + \tfrac{1}{4}\xi_4 + \tfrac{1}{8}\xi_5 \\
\zeta_4 &= \tfrac{1}{2} - \rho_0 + \rho_1 - \tfrac{1}{2}\rho_2 - 2\xi_1 + 4\xi_2 + 4\xi_3 + \xi_4 + \tfrac{1}{2}\xi_5
\end{aligned} \tag{38}$$

where

$$\begin{aligned}
\zeta_1 &= p(1, 1) = p(3, 3) \\
\zeta_2 &= p(1, 2) = p(3, 2) = p(2, 1) = p(2, 3) \\
\zeta_3 &= p(1, 3) = p(3, 1) \\
\zeta_4 &= p(2, 2).
\end{aligned}$$

For the case at hand, equations (32) become

$$\begin{aligned}
\xi_1^2 &= \frac{\zeta_1^{-1}\zeta_2^{4}\zeta_3^{-1}\zeta_4^{-2}}{\eta_1^{-2}\eta_2^{4}\eta_3^{-2}\eta_4^{2}\eta_5^{-4}}, \qquad \xi_2^2 = \frac{\zeta_1^{2}\zeta_2^{-8}\zeta_3^{2}\zeta_4^{4}}{\eta_1^{6}\eta_2^{-8}\eta_4^{-8}\eta_5^{8}} \\
\xi_3^4 &= \frac{\zeta_1^{2}\zeta_2^{-8}\zeta_3^{2}\zeta_4^{4}}{\eta_1^{8}\eta_2^{-12}\eta_4^{-8}\eta_5^{8}}, \qquad \xi_4^2 = \frac{\zeta_1^{\frac{1}{2}}\zeta_2^{-2}\zeta_3^{\frac{1}{2}}\zeta_4^{1}}{\eta_1^{3}\eta_2^{-6}\eta_3\eta_4^{-2}\eta_5^{2}} \\
\xi_5 &= \frac{\zeta_1^{\frac{1}{4}}\zeta_2^{-1}\zeta_3^{\frac{1}{4}}\zeta_4^{\frac{1}{2}}}{\eta_1^{\frac{1}{2}}\eta_2^{-1}\eta_3^{\frac{1}{2}}\eta_5^{-1}}.
\end{aligned} \tag{39}$$

The right members of these equations can be written in terms of the ξ's by using (37) and (38). Equations (39) can then be written as five multinomial equations in the five ξ's. In principle, by using Sylvester's method,[5] the ξ's could be systematically eliminated to yield a single high-order polynomial equation for ξ_1 . The other ξ's can be similarly determined. To carry this out in practice would be a formidable task.

V. THE COVARIANCE PROBLEM

We have seen that the maxentropic discrete stationary process with given values x_1 , x_2 , $\cdots$, x_K and given truncated covariance sequence ρ_0 , ρ_1 , $\cdots$, ρ_L is an Lth order Markov process. In Section III a formal solution was given to the problem of determining the complete probability structure of this process. This structure in turn determines the remaining elements ρ_{L+1} , ρ_{L+2} , $\cdots$ of the covariance sequence. It is shown in Appendix C that for a K-valued Lth order Markov process the covariance sequence can always be written in the form

$$\rho_n = \sum_{j=1}^{K^L} A_j \theta_j^n , \qquad n = 0, 1, 2, \cdots . \tag{40}$$

Thus, only a restricted class of covariance sequences, those expressible as a finite sum of exponentials, can be obtained by our procedure. The dicode covariance, $\rho_0 = \frac{1}{2}$, $\rho_1 = -\frac{1}{4}$, $\rho_n = 0$, $n = 2, 3, \cdots$, is excluded, for example.

This raises an important pertinent question that we have sidestepped thus far: what are the possible covariance sequences for a discrete stationary process taking values $x_1 , x_2 , \cdots , x_K$? When the restriction on allowed values of the process is removed, one has the elegant Bochner theorem[6] that characterizes the covariance sequences as Fourier cosine-series coefficients of non-negative finite measures, that is, as non-negative definite sequences. No comparable description is available for the proper subset of these non-negative definite sequences that comprises the covariance sequences of processes restricted to the values $x_1 , x_2 , \cdots , x_K$. The situation has been discussed by B. McMillan[7] and L. A. Shepp.[8]

More germaine to our discussion, and less ambitious, is the question: "What sequences of $L + 1$ numbers, $\rho_0 , \rho_1 , \cdots , \rho_L$, can be the first $L + 1$ terms of the covariance sequence of a discrete stationary process taking values $x_1 , \cdots , x_K$?" If we consider such a sequence as a point in $\mathcal{E}_{L+1}$, Euclidean space of $L + 1$ dimensions, then the region $\mathcal{R}$ of admissible points is a convex one bounded by fewer than $2K^L$ hyperplanes. This is shown in Appendix D. Such a region can be characterized as the convex hull of its extreme points, or vertices (finite in number), and a convenient description of the region is a list of these vertices. An alternate economical description is a list of the hyperplane boundaries of $\mathcal{R}$. We have been unable to sort out the combinatorics involved, even in the simple case $K = 2$, $x_1 = 1$, $x_2 = -1$, to provide such lists for arbitrary values of L.

It is to be expected that the formalisms of Section III will have solutions $p_{L+1}(\epsilon_1 , \cdots , \epsilon_{L+1})$ that are probability distributions if and only if $\rho_0 , \rho_1 , \cdots , \rho_L$ is contained in $\mathcal{R}$. In the few cases that we have been able to carry out in detail, this is indeed the case. For example, from Section IV, we see that the solution presented is valid only if

$$1 + 2\rho_1 + \rho_2 \geqq 0,$$

$$1 - 2\rho_1 + \rho_2 \geqq 0,$$

$$1 - \rho_2 \geqq 0.$$

These inequalities do indeed describe the region of admissible values of ρ_1 and ρ_2 for a process having values $+1$ and -1.

VI. THE UNIFILAR MARKOV MESSAGE SOURCE

The dicode process is not an Lth order Markov process for any L. It can, however, be described simply in terms of a two-state Markov chain. Consider the chain with states S_1 and S_2 and transition probabilities $\frac{1}{2}$ as shown in Fig. 1. Along each transition path in the figure is an associated number enclosed in a box. When the chain passes along a path from one state to another, the associated number is "emitted." The sequence of emitted numbers is the dicode process.

The foregoing is an example of a class of discrete processes which we call unifilar Markov message sources. An ergodic Markov chain with states S_1 , S_2 , $\cdots$, S_N is given along with the transition probabilities $p_{ij} = \Pr\{\text{next state is } S_j \mid \text{last state is } S_i\}$. Associated with each pair of states S_i , S_j for which $p_{ij} > 0$ is a number $X(S_i , S_j)$ that is emitted when the chain passes from S_i to S_j . The word unifilar refers to the fact that we demand that whenever $S_j \neq S_k$, then $X(S_i , S_j) \neq X(S_i , S_k)$, $i = 1, 2, \cdots , N$. If this condition is met and the initial state of the chain is known, the sequence of emitted letters determines the sequence of states followed by the chain and a simple formula is available for the entropy of the emitted X process, namely

$$H = -\sum_{i,j} p(S_i)p_{ij} \log p_{ij} . \tag{41}$$

(See Ref. 9, p. 68.) Here $p(S_i)$, the probability that the chain be in state S_i , is the stationary measure satisfying

$$\sum_i p(S_i)p_{ij} = p(S_j), \qquad j = 1, 2, \cdots , N.$$

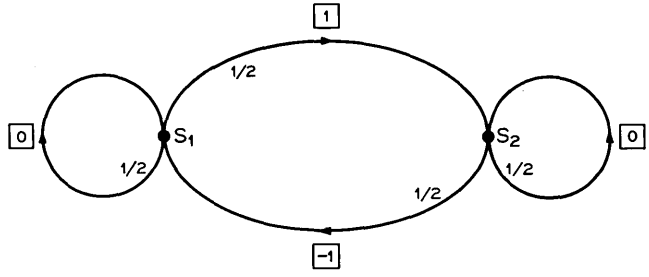


Fig. 1—Diagram of a two-state Markov chain with states S_1 and S_2 and transition probabilities 1/2.

It is easy to write an expression for the covariance sequence of a unifilar Markov message source. When $n \geqq 2$, we have

$$\rho_n = \sum_{\substack{i,j \\ k,l}} X(S_i , S_j)p(S_i)p_{ij}p_{jk}^{(n-1)}p_{kl}X(S_k , S_l) \tag{42}$$

where $p_{jk}^{(n)}$ is the probability that the chain be in state S_k after n transitions given that it started from S_j . We have also

$$\rho_1 = \sum_{i,j,k} X(S_i , S_j)p(S_i)p_{ij}p_{jk}X(S_j , S_k), \tag{43}$$

$$\rho_0 = \sum_{i,j} X(S_i , S_j)^2 p(S_i)p_{ij} . \tag{44}$$

Since $p_{ij}^{(n)}$ has an expression analogous to (75), in Appendix C, equation (42) can be written

$$\rho_n = \sum_{1}^{N} B_i\varphi_i^n , \qquad n = 2, 3, \cdots . \tag{45}$$

Comparison with (40) shows that the covariance sequences achievable are of almost the same form as for the Lth-order Markov processes. For the unifilar Markov message source, deviation from the sum of exponential form may occur for ρ_0 and ρ_1 .

To find the unifilar Markov message source of largest entropy with N states and given truncated covariance sequence appears to be a most difficult problem. We have not found a unifilar Markov source with values 0, ± 1, with the dicode covariance sequence and an entropy greater than unity.

VII. THE n-VARIATE GAUSSIAN ANALOGUE

Closely related to the problem we have been discussing is the following question. Let $X_1 , X_2 , \cdots , X_n$ be n random variables each taking values from the set $x_1 , x_2 , \cdots , x_K$. What distribution for the Xs, $p_n(\epsilon_1 , \cdots , \epsilon_n)$ say, has maximum entropy and given second moments $EX_iX_j = \rho_{ij}$? Using Lagrange multipliers, one finds at once that

$$p_n(\epsilon_1 , \cdots , \epsilon_n) = c \exp\left\{-\frac{1}{2}\sum_{1}^{n} \sigma_{ij}x_{\epsilon_i}x_{\epsilon_j}\right\}. \tag{46}$$

Here

$$\frac{1}{c} \equiv S = \sum_{\epsilon} \exp\left\{-\frac{1}{2}\sum_{i,j=1}^{n} \sigma_{ij}x_{\epsilon_i}x_{\epsilon_j}\right\} \tag{47}$$

and the σ's must be determined so that

$$\rho_{ij} = -\frac{2}{S}\frac{\partial S}{\partial \sigma_{ij}}$$

$$= c \sum_{\epsilon} x_{\epsilon_i} x_{\epsilon_j} \exp\left\{-\frac{1}{2}\sum_{i,j=1}^{n} \sigma_{ij} x_{\epsilon_i} x_{\epsilon_j}\right\}, \qquad i, j = 1, \cdots, n. \quad (48)$$

The entropy of (46) can be written

$$H = -\sum_{\epsilon} p_n(\epsilon_1, \cdots, \epsilon_n) \log p_n(\epsilon_1, \cdots, \epsilon_n)$$

$$= \log S + \tfrac{1}{2} \sum \rho_{ij}\sigma_{ij} . \quad (49)$$

The analogy of (46) with the n-variate Gaussian density is striking. Let $Y_1, Y_2, \cdots, Y_n$ be n real-valued random variables having probability density $\hat{p}_n(y_1, y_2, \cdots, y_n)$. Let $EY_iY_j = \rho_{ij}$. Under these constraints, the density having largest entropy is the Gaussian density

$$\hat{p}_n(y_1, \cdots, y_n) = \hat{c} \exp\{-\tfrac{1}{2} \sum \hat{\sigma}_{ij} y_i y_j\} \quad (50)$$

Here

$$\frac{1}{\hat{c}} \equiv \hat{S} = \int_{-\infty}^{\infty} dy_1 \cdots \int_{-\infty}^{\infty} dy_n \exp\{-\tfrac{1}{2} \sum \hat{\sigma}_{ij} y_i y_j\} \quad (51)$$

and the $\hat{\sigma}$'s are related to the ρ's by

$$\rho_{ij} = -\frac{2}{\hat{S}}\frac{\partial \hat{S}}{\partial \hat{\sigma}_{ij}}, \qquad i, j = 1, 2, \cdots, n. \quad (52)$$

The entropy of (50) can be written

$$\hat{H} = -\int_{-\infty}^{\infty} dy_1 \cdots \int_{-\infty}^{\infty} dy_n \hat{p}_n(y_1, \cdots, y_n) \log \hat{p}_n(y_1, \cdots, y_n)$$

$$= \log \hat{S} + \tfrac{1}{2} \sum \rho_{ij}\hat{\sigma}_{ij} . \quad (53)$$

Note the complete parallel between (46) through (49) and (50) through (53).

In the case of the Gaussian density, the integral (51) can be performed to yield the simple expression

$$\hat{S} = \frac{(2\pi)^{n/2}}{|\hat{\sigma}|^{\frac{1}{2}}} \quad (54)$$

where $|\hat{\sigma}|$ is the determinant of the matrix with elements $\hat{\sigma}_{ij}$. Equation (52) then gives at once the well-known results $\rho_{ij} = \hat{\sigma}^{-1}_{ij}$ or $\hat{\sigma} = \rho^{-1}$, where we use obvious matrix notation. Surprisingly, the σ_{ij} are rational in the ρ_{ij} in spite of the more complicated nature of the dependence of $\hat{S}$ on the $\hat{\sigma}_{ij}$, as given by (54).

In the discrete case, matters are not so simple. For example, when $n = 3$, $K = 2$ and $x_1 = 1$, $x_2 = -1$,

$$S = 2e^{-\frac{1}{2}(\sigma_{11}+\sigma_{22}+\sigma_{33})}[e^{-\frac{1}{2}(\sigma_{12}+\sigma_{13}+\sigma_{23})} + e^{+\frac{1}{2}(\sigma_{12}-\sigma_{13}-\sigma_{23})} + e^{-\frac{1}{2}(-\sigma_{12}+\sigma_{13}-\sigma_{23})} + e^{-\frac{1}{2}(-\sigma_{12}-\sigma_{13}+\sigma_{23})}].$$

One finds

$$\sigma_{12} = -\tfrac{1}{8}\log\frac{\beta_1\beta_2}{\beta_3\beta_4}$$

$$\sigma_{13} = -\tfrac{1}{8}\log\frac{\beta_1\beta_3}{\beta_2\beta_4}$$

$$\sigma_{23} = -\tfrac{1}{8}\log\frac{\beta_1\beta_4}{\beta_2\beta_3}$$

$$c = \tfrac{1}{8}[\beta_1\beta_2\beta_3\beta_4]^{\frac{1}{4}}$$

where

$$\beta_1 = 1 + \rho_{12} + \rho_{13} + \rho_{23}$$

$$\beta_2 = 1 + \rho_{12} - \rho_{13} - \rho_{23}$$

$$\beta_3 = 1 - \rho_{12} + \rho_{13} - \rho_{23}$$

$$\beta_4 = 1 - \rho_{12} - \rho_{13} + \rho_{23}\ .$$

Thus σ_{12} , σ_{13} , and σ_{23} are not rational in the ρ's. (The quantities σ_{11} , σ_{22} , and σ_{33} can be chosen to be zero in this binary case.) The probabilities themselves, however, turn out to be linear in the ρ's. One has $p_3(1, 1, 1) = p_3(2, 2, 2) = \frac{1}{8}\beta_1$, $p(1, 1, 2) = p(2, 2, 1) = \frac{1}{8}\beta_2$, $p(1, 2, 1) = p(2, 1, 2) = \frac{1}{8}\beta_3$, $p(1, 2, 2) = p(2, 1, 1) = \frac{1}{8}\beta_4$. When $n > 2$, the p's become algebraic in the ρ's.

VIII. CONCLUDING REMARKS

In addition to the methods discussed here, I have pursued several other attacks on the problem at hand. All approaches seem to end in unmanageable algebraic complexities. Perhaps it is the nature of the problem; perhaps the answer can't be stated in simple terms. The mathematician, whose pleasure it is to make order out of chaos, will likely disagree. He will feel that surely so basic a construct as we consider here must be simple at heart and that we have only failed to find the appropriate language to make it so appear. In analogy, to

the uninitiated, the relationship found in the last section between ρ and $\hat{\sigma}$, namely $\hat{\sigma} = p^{-1}$, must surely at first have appeared formidable. The matrix language of Cayley brings us apparent order here. Can we find the right point of view in which to describe the discrete maxentropic process?

IX. ACKNOWLEDGMENT

I am happy to acknowledge many most pleasant and profitable talks with L. A. Shepp on all phases of the work reported here.

APPENDIX A

We are concerned with maximizing (8) subject to (5), (6), (7), (10), and the $L + 1$ constraints

$$\sum_{\epsilon_1=1}^{K} \cdots \sum_{\epsilon_{k+1}=1}^{K} x_{\epsilon_1} x_{\epsilon_{k+1}} p_{k+1}(\epsilon_1, \cdots, \epsilon_{k+1}) = \rho_k, \qquad k = 0, 1, \cdots, L. \tag{55}$$

We proceed by maximizing H_n of (9), subject to these same constraints, for each $n > L + 1$.

Observe now that (55) and (10) can be stated solely in terms of p_{L+1}:

$$\sum_{\epsilon_1=1}^{K} \cdots \sum_{\epsilon_{L+1}=1}^{K} x_{\epsilon_1} x_{\epsilon_{k+1}} p_{L+1}(\epsilon_1, \cdots, \epsilon_{L+1}) = \rho_k, \qquad k = 0, 1, \cdots, L \tag{56}$$

$$\sum_{\epsilon_1=1}^{K} \cdots \sum_{\epsilon_{L+1}=1}^{K} x_{\epsilon_1} p_{L+1}(\epsilon_1, \cdots, \epsilon_{L+1}) = 0. \tag{57}$$

Thus the maximization can be carried out by first maximizing H_n subject to (5), (6), and (7) *given* the K^{L+1} quantities $p_{L+1}(\epsilon_1, \cdots, \epsilon_{L+1})$, $\epsilon_1, \cdots, \epsilon_{L+1} = 1, 2, \cdots, K$, then maximizing further over these quantities subject to the additional constraints (56) and (57). For this first maximization problem, the constraints are (5), (6), (7), and

$$\begin{aligned} p_{L+1}(\epsilon_1, \cdots, \epsilon_{L+1}) &= \sum_{\alpha} p_n(\epsilon_1, \cdots, \epsilon_{L+1}, \alpha_1, \cdots, \alpha_{n-L-1}) \\ &= \sum_{\alpha} p_n(\alpha_1, \epsilon_1, \cdots, \epsilon_{L+1}, \alpha_2, \cdots, \alpha_{n-L-1}) \\ &= \sum_{\alpha} p_n(\alpha_1, \alpha_2, \epsilon_1, \cdots, \epsilon_{L+1}, \alpha_3, \cdots, \alpha_{n-L-1}) \\ &\vdots \\ &= \sum_{\alpha} p_n(\alpha_1, \cdots, \alpha_{n-L-1}, \epsilon_1, \cdots, \epsilon_{L+1}) \end{aligned}$$

$$\epsilon_1, \cdots, \epsilon_{L+1} = 1, 2, \cdots, K \tag{58}$$

where we regard the p_{L+1} as given. These quantities, of course, must themselves satisfy (5), (6), and (7) with $n = L + 1$.

Introducing Lagrange multipliers we seek the maximum of

$$J = -\sum_{\epsilon} p_n(\epsilon_1, \cdots, \epsilon_n) \log p_n(\epsilon_1, \cdots, \epsilon_n) + \lambda \sum_{\epsilon} p_n(\epsilon_1, \cdots, \epsilon_n)$$
$$+ \sum_{j=0}^{n-L-1} \sum_{\alpha, \epsilon} \nu^{(j)}_{\epsilon_1 \cdots \epsilon_{L+1}}$$
$$p_n(\alpha_1, \cdots, \alpha_j, \epsilon_1, \cdots, \epsilon_{L+1}, \alpha_{j+1}, \cdots, \alpha_{n-L-1}) \tag{59}$$

where in the last sum for $j = 0$ the first argument of p_n is ϵ_1 and for $j = n - L - 1$ the last argument of p_n is ϵ_{L+1}. In (59) we have omitted terms corresponding to the constraint (7). It turns out that this constraint will be met automatically. Differentiation of (59) with respect to $p_n(\epsilon_1, \cdots, \epsilon_n)$ yields

$$-1 - \log p_n(\epsilon_1, \cdots, \epsilon_n)$$
$$+ \lambda + \nu^{(0)}_{\epsilon_1 \cdots \epsilon_{L+1}} + \nu^{(1)}_{\epsilon_2 \cdots \epsilon_{L+2}} + \cdots + \nu^{(n-L-1)}_{\epsilon_{n-L} \cdots \epsilon_n} = 0$$

or

$$p_n(\epsilon_1, \cdots, \epsilon_n) = c \exp \{\nu^{(0)}_{\epsilon_1 \cdots \epsilon_{L+1}} + \nu^{(1)}_{\epsilon_2 \cdots \epsilon_{L+2}} + \cdots + \nu^{(n-L-1)}_{\epsilon_{n-L} \cdots \epsilon_n}\} \tag{60}$$

where c is independent of the ϵ's. Equation (58) and (5) serve in principle to determine c and the $K^{L+1}(n - L)$ Lagrange multipliers $\nu^{(j)}_{\epsilon_1 \cdots \epsilon_{L+1}}$.

Note now that from (60) we find that

$$\Pr\{X_n = x_{\epsilon_n} \mid X_1 = x_{\epsilon_1}, \cdots, X_{n-1} = x_{\epsilon_{n-1}}\} = \frac{p_n(\epsilon_1, \cdots, \epsilon_n)}{\sum_{\alpha=1}^{K} p_n(\epsilon_1, \cdots, \epsilon_{n-1}, \alpha)}$$
$$= \exp(\nu^{(n-L-1)}_{\epsilon_{n-L} \cdots \epsilon_n}) \Big/ \sum_{\alpha=1}^{K} \exp(\nu^{(n-L-1)}_{\epsilon_{n-L} \cdots \epsilon_{n-1} \alpha}) \equiv f_n(\epsilon_{n-L}, \epsilon_{n-L+1}, \cdots, \epsilon_n),$$

since of the ν's in (60) only $\nu^{(n-L-1)}$ involves ϵ_n. Similarly, for each m satisfying $L + 1 \leqq m \leqq n$ we find from (60) that

$$\Pr\{X_m = x_{\epsilon_m} \mid X_1 = x_{\epsilon_1}, \cdots, X_{m-1} = x_{\epsilon_{m-1}}\}$$
$$= \frac{\Pr\{X_1 = x_{\epsilon_1}, \cdots, X_m = x_{\epsilon_m}\}}{\Pr\{X_1 = x_{\epsilon_1}, \cdots, X_{m-1} = x_{\epsilon_{m-1}}\}}$$
$$= f_m(\epsilon_{m-L}, \epsilon_{m-L+1}, \cdots, \epsilon_m), \qquad L + 1 \leqq m \leqq n, \tag{61}$$

that is, this conditional probability depends only on $\epsilon_{m-L}, \epsilon_{m-L+1}, \cdots, \epsilon_m$. Writing (61) in another form,

$$\Pr\{X_1 = x_{\epsilon_1}, \cdots, X_m = x_{\epsilon_m}\} = f_m(\epsilon_{m-L}, \cdots, \epsilon_m) \Pr\{X_1 = x_{\epsilon_1}, \cdots, X_{m-1} = x_{\epsilon_{m-1}}\}$$

and then summing on $\epsilon_1, \epsilon_2, \cdots, \epsilon_{m-L-1}$, we find that

$$\Pr\{X_{m-L} = x_{\epsilon_{m-L}}, \cdots, X_m = x_{\epsilon_m}\} = f_m(\epsilon_{m-L}, \cdots, \epsilon_m) \cdot \Pr\{X_{m-L} = x_{\epsilon_{m-L}}, \cdots, X_{m-1} = x_{\epsilon_{m-1}}\}. \quad (62)$$

We have then, substituting the value of f_m from (62) into (61),

$$\begin{aligned}\Pr\{X_m = x_{\epsilon_m} \mid X_1 = x_{\epsilon_1}, \cdots, X_{m-1} = x_{\epsilon_{m-1}}\} &= \frac{\Pr\{X_{m-L} = x_{\epsilon_{m-L}}, \cdots, X_m = x_{\epsilon_m}\}}{\Pr\{X_{m-L} = x_{\epsilon_{m-L}}, \cdots, X_{m-1} = x_{\epsilon_{m-1}}\}} \\ &= \Pr\{X_m = x_{\epsilon_m} \mid X_{m-1} = x_{\epsilon_{m-1}}, \cdots, X_{m-L} = x_{\epsilon_{m-L}}\}, \\ & \qquad L + 1 \leqq m \leqq n. \end{aligned} \quad (63)$$

Let us now define

$$p_L(\epsilon_1, \cdots, \epsilon_L) \equiv \sum_{\alpha=1}^{K} p_{L+1}(\epsilon_1, \cdots, \epsilon_L, \alpha) \quad (64)$$

$$q_L(\epsilon_{L+1} \mid \epsilon_1, \cdots, \epsilon_L) \equiv \frac{p_{L+1}(\epsilon_1, \cdots, \epsilon_{L+1})}{p_L(\epsilon_1, \cdots, \epsilon_L)}. \quad (65)$$

Repeated application of (63) then shows that

$$\begin{aligned}\Pr\{X_1 = x_{\epsilon_1}, \cdots, X_m = x_{\epsilon_m}\} &= p_L(\epsilon_1, \cdots, \epsilon_L) q(\epsilon_{L+1} \mid \epsilon_1, \cdots, \epsilon_L) \\ &\cdot q(\epsilon_{L+2} \mid \epsilon_2, \cdots, \epsilon_{L+1}) \cdots q(\epsilon_m \mid \epsilon_{m-L}, \epsilon_{m-L+1}, \cdots, \epsilon_{m-1}), \\ & \qquad m = L + 1, L + 2, \cdots, n. \end{aligned}$$

This expression is independent of n. Thus, among stationary processes with a given $(L + 1)$st-order distribution $p_{L+1}(\epsilon_1, \cdots, \epsilon_n)$, the Lth-order Markov process generated by the initial distribution (64) and the transition mechanism (65) has maximum entropy H_n for every $n \geqq L + 1$. Q.E.D.

APPENDIX B

We seek to maximize $H_{L+1} - H_L$ by choice of the $p_{L+1}(\epsilon_1, \cdots, \epsilon_{L+1})$ subject to the stationarity constraints

$$\sum_{\alpha} p_{L+1}(\epsilon_1 , \cdots , \epsilon_L , \alpha)$$

$$= \sum_{\alpha} p_{L+1}(\alpha, \epsilon_1 , \cdots , \epsilon_L) \qquad \epsilon_1 , \cdots , \epsilon_L = 1, \cdots , K, \tag{66}$$

the distribution constraint

$$\sum_{\epsilon} p_{L+1}(\epsilon_1 , \cdots , \epsilon_{L+1}) = 1, \tag{67}$$

and the covariance restrictions

$$\sum_{\epsilon} l_{L+1}^{(j)}(\epsilon_1 , \cdots , \epsilon_{L+1}) p_{L+1}(\epsilon_1 , \cdots , \epsilon_{L+1}) = (L + 1 - j)\rho_j$$

$$j = 0, 1, \cdots , L \tag{68}$$

where the $l_{L+1}^{(j)}$ are defined in (19). The constraints (68) treat the variables in a more symmetric manner than do the constraints

$$\sum_{\epsilon} x_{\epsilon_1} x_{\epsilon_{k+1}} p_{L+1}(\epsilon_1 , \cdots , \epsilon_{L+1}) = \rho_k , \qquad k = 0, 1, \cdots , L. \tag{69}$$

It is easy to show that (66), (67), and (68) are equivalent to (66), (67), and (69).

Introducing Lagrange multipliers, we must maximize

$$\begin{aligned} J' = & - \sum_{\epsilon} p_{L+1}(\epsilon_1 , \cdots , \epsilon_{L+1}) \log p_{L+1}(\epsilon_1 , \cdots , \epsilon_{L+1}) \\ & + \sum_{\epsilon} p_{L+1}(\epsilon_1 , \cdots , \epsilon_{L+1}) \log [\sum_{\alpha} p_{L+1}(\epsilon_1 , \cdots , \epsilon_L , \alpha)] \\ & + \sum_{\epsilon} \nu_{\epsilon_1 \cdots \epsilon_L} [\sum_{\alpha} p_{L+1}(\epsilon_1 , \cdots , \epsilon_L , \alpha) - \sum_{\alpha} p_{L+1}(\alpha, \epsilon_1 , \cdots , \epsilon_L)] \\ & + \mu \sum_{\epsilon} p_{L+1}(\epsilon_1 , \cdots , \epsilon_{L+1}) \\ & + \sum_{j=0}^{L} \lambda_j \sum_{\epsilon} l_{L+1}^{(j)}(\epsilon_1 , \cdots , \epsilon_{L+1}) p_{L+1}(\epsilon_1 , \cdots , \epsilon_{L+1}). \end{aligned}$$

Differentiation with respect to $p_{L+1}(\epsilon_1 , \cdots , \epsilon_{L+1})$ gives the necessary condition

$$-1 - \log p_{L+1}(\epsilon_1 , \cdots , \epsilon_{L+1}) + 1 + \log \sum_{\alpha} p_{L+1}(\epsilon_1 , \cdots , \epsilon_L , \alpha)$$

$$+ \nu_{\epsilon_1 \cdots \epsilon_L} - \nu_{\epsilon_2 \cdots \epsilon_{L+1}} + \mu + \sum_{j=0}^{L} \lambda_j l_{L+1}^{(j)}(\epsilon_1 , \cdots , \epsilon_{L+1}) = 0. \tag{70}$$

With the notation $c = e^{\mu}$, $f(\epsilon_1 , \cdots , \epsilon_L) = e^{-\nu_{\epsilon_1 \cdots \epsilon_L}}$ and the definition (20), (70) becomes (22).

Inserting (22) in (14) yields (21). Again using (22) for q_L in (12) gives

$$\sum_{\epsilon_1} h_{L+1}(\epsilon_1 , \cdots , \epsilon_{L+1}) \frac{p_L(\epsilon_1 , \cdots , \epsilon_L)}{f(\epsilon_1 , \cdots , \epsilon_L)} = \frac{1}{c} \frac{p_L(\epsilon_2 , \cdots , \epsilon_{L+1})}{f(\epsilon_2 , \cdots , \epsilon_{L+1})}$$

$$\epsilon_2 , \cdots , \epsilon_{L+1} = 1, 2, \cdots , K,$$

where for simplicity we suppress the λ dependence of h. Relabeling variables, this can be written

$$\sum_{\epsilon_{L+1}} h_{L+1}(\epsilon_{L+1} , \cdots , \epsilon_1) \frac{p_L(\epsilon_{L+1} , \cdots , \epsilon_2)}{f(\epsilon_{L+1} , \cdots , \epsilon_2)} = \frac{1}{c} \frac{p_L(\epsilon_L , \cdots , \epsilon_1)}{f(\epsilon_L , \cdots , \epsilon_1)}$$

$$\epsilon_1 , \cdots , \epsilon_L = 1, 2, \cdots , K. \tag{71}$$

But from the definition (20) and (19), $h_{L+1}(\epsilon_{L+1} , \cdots , \epsilon_1) = h_{L+1}(\epsilon_1 , \cdots , \epsilon_{L+1})$. Equation (71) is then

$$\sum_{\epsilon_{L+1}} h_{L+1}(\epsilon_1 , \cdots , \epsilon_{L+1} ; \lambda_0 , \cdots , \lambda_L) \cdot \frac{p_L(\epsilon_{L+1} , \cdots , \epsilon_2)}{f(\epsilon_{L+1} , \cdots , \epsilon_2)} = \frac{1}{c} \frac{p_L(\epsilon_L , \cdots , \epsilon_1)}{f(\epsilon_L , \cdots , \epsilon_1)}.$$

Comparison with (21) now shows that we must have

$$\frac{p_L(\epsilon_L , \cdots , \epsilon_1)}{f(\epsilon_L , \cdots , \epsilon_1)} = kf(\epsilon_1 , \cdots , \epsilon_L) \tag{72}$$

if the eigenvalue $1/c$ is not degenerate, which is the general case. A change of notation reduces (72) to (23).

APPENDIX C

We have considered stationary Lth-order Markov processes $\cdots X_{-1} , X_0 , X_1 , \cdots$ whose variables take values $x_1 , x_2 , \cdots , x_k$. The probability structure of such a process can be generated from transition probabilities $q_L(\epsilon_{L+1} \mid \epsilon_1 , \cdots , \epsilon_L)$ via the mechanism of equations (12) through (16). Such a process can also be regarded as a function on the states of an ordinary Markov chain. The chain has K^L states, each one labeled by an Ltuple of integers $\alpha = (\alpha_1 , \alpha_2 , \cdots , \alpha_L)$. The conditional probability that the chain pass in one move from state α to state β is given by

$$p(\beta \mid \alpha) = q_L(\beta_L \mid \alpha_1 , \alpha_2 , \cdots , \alpha_L)\delta_{\alpha_2\beta_1}\delta_{\alpha_3\beta_2} \cdots \delta_{\alpha_L\beta_{L-1}} \tag{73}$$

where δ_{ij} is the Kronecker symbol. The chain thus has a very special

nature. One can pass only to a state whose initial $L-1$ labels agree with the last $L-1$ labels of the state just left. The states correspond to successive Ltuples, X_i , X_{i+1} , $\cdots$, X_{i+L-1} , in the Lth-order Markov process. We regain that process from the chain by defining on the states of the chain the numerical valued function

$$X(\alpha) = x_{\alpha_1} . \tag{74}$$

From well-known results in the theory of finite Markov chains (see for example Ref. 10, Chapt. 16, Sec. 1), we see that the probability of going from state α to state β in exactly n moves can be written in the form

$$p^{(n)}(\beta \mid \alpha) = \sum_{j=1}^{K^L} u_\alpha^{(j)} v_\beta^{(j)} \lambda_j^n \qquad n = 1, 2, \cdots$$

$$\alpha_1 , \alpha_2 , \cdots , \alpha_L , \beta_1 , \cdots , \beta_L = 1, 2, \cdots , K. \tag{75}$$

Here the u's and v's are left and right eigenvectors of $p(\beta \mid \alpha)$,

$$\sum_\beta u_\beta^{(j)} p(\beta \mid \alpha) = \theta_j u_\alpha^{(j)} \qquad \sum_\beta p(\alpha \mid \beta) v_\beta^{(j)} = \theta_j v_\alpha^{(j)}$$

$$j = 1, 2, \cdots , K^L, \quad \alpha_1 , \alpha_2 , \cdots , \alpha_L = 1, 2, \cdots , K,$$

normalized so that

$$\sum_{j=1}^{K^L} u_\alpha^{(j)} v_\beta^{(j)} = \delta_{\alpha\beta} .$$

Note that this gives (75) the special value

$$p^{(0)}(\beta \mid \alpha) = \delta_{\alpha\beta} .$$

In terms of this Markov chain, the covariance of the X process can be written

$$\rho_n = EX_i X_{i+n} = \sum_\alpha \sum_\beta x_{\alpha_1} x_{\beta_1} p(\alpha) p^{(n)}(\beta \mid \alpha) \qquad n = 0, 1, 2, \cdots \tag{76}$$

where $p(\alpha)$ is the stationary distribution for the chain, i.e.,

$$\sum_\alpha p(\alpha) p(\beta \mid \alpha) = p(\beta).$$

Using (75) in (76) we have the desired result

$$\rho_n = \sum_{j=1}^{K^L} A_j \theta^n, \qquad n = 0, 1, 2, \cdots$$

where

$$A_j = \sum_{\alpha} x_{\alpha_1} p(\alpha) u_{\alpha}^{(j)} \sum_{\beta} x_{\beta_1} v_{\beta}^{(j)}.$$

APPENDIX D

For any discrete stationary process taking values $x_1, x_2, \cdots, x_k$, the truncated covariance sequence can be formed from the $(L+1)$st-order distribution $p_{L+1}(\epsilon_1, \cdots, \epsilon_{L+1})$. Thus

$$\rho_j = \sum_{\epsilon} x_{\epsilon_1} x_{\epsilon_{j+1}} p_{L+1}(\epsilon_1, \cdots, \epsilon_{L+1}) \qquad j = 0, 1, \cdots, L. \tag{77}$$

Such a distribution satisfies the stationarity conditions

$$\sum_{\alpha} p_{L+1}(\alpha, \epsilon_2, \cdots, \epsilon_{L+1}) = \sum_{\alpha} p_{L+1}(\epsilon_2, \cdots, \epsilon_{L+1}, \alpha)$$

$$\epsilon_2, \epsilon_3, \cdots, \epsilon_{L+1} = 1, 2, \cdots, K, \tag{78}$$

the constraint

$$\sum_{\epsilon} p_{L+1}(\epsilon_1, \cdots, \epsilon_{L+1}) = 1, \tag{79}$$

and the inequalities

$$p_{L+1}(\epsilon_1, \cdots, \epsilon_{L+1}) \geqq 0 \qquad \epsilon_1, \epsilon_2, \cdots, \epsilon_{L+1} = 1, 2, \cdots, K. \tag{80}$$

Conversely, from K^{L+1} quantities $p_{L+1}(\epsilon_1, \cdots, \epsilon_{L+1})$ satisfying (78) through (80) we can construct a discrete stationary (Lth-order Markov) process having values $x_1, \cdots, x_K$ and truncated covariance given by (77). To do so, define

$$p_L(\epsilon_1, \cdots, \epsilon_L) \equiv \sum_{\epsilon_{L+1}} p_{L+1}(\epsilon_1, \cdots, \epsilon_{L+1})$$

$$\epsilon_1, \epsilon_2, \cdots, \epsilon_L = 1, 2, \cdots, K.$$

Let

$$q_L(\epsilon_{L+1} \mid \epsilon_1, \cdots, \epsilon_L) = \frac{p_{L+1}(\epsilon_1, \cdots, \epsilon_{L+1})}{p_L(\epsilon_1, \cdots, \epsilon_L)}.$$

It is easy to verify that (12), (13), and (14) are satisfied, so that the measure described by (15) and (16) defines the desired process.

Thus equations (77) through (80) serve to define parametrically the region $\mathfrak{R}$ of admissible truncated covariance sequences. Consider an $(L+1)$st-order density $p_{L+1}(\epsilon_1, \cdots, \epsilon_{L+1})$ as a point in a Euclidean space $\mathcal{E}_{K^{L+1}}$ of dimension K^{L+1}. Equations (78), (79), and (80) define a convex region $\mathcal{U}$ in this space that is bounded by no more than $K^L + 1 + K^{L+1} \leqq 2K^{L+1}$ hyperplanes ($K \geqq 2$). Equations (77) provide

a linear mapping of $\mathcal{E}_{KL+1}$ into $\mathcal{E}_{L+1}$ and in particular the image of $\mathfrak{V}$ is $\mathfrak{R}$. The hyperplane boundaries of $\mathfrak{V}$ map into hyperplanes in $\mathcal{E}_{L+1}$ that include all the hyperplane boundaries of $\mathfrak{R}$. Q.E.D.

REFERENCES

1. Kobayashi, H., "Coding Schemes for Reduction of Intersymbol Interference in Data Transmission Systems," IBM J. Res. & Dev., *14*, No. 4 (July 1970), pp. 343–353.
2. Croisier, A., "Introduction to Pseudoternary Transmission Codes," IBM J. Res. & Dev., *14*, No. 4 (July 1970), pp. 354–367.
3. Kobayashi, H., and Tang, D. T., "Application of Partial-Response Channel Coding to Magnetic Recording Systems," IBM J. Res. & Dev., *14*, No. 4 (July 1970), pp. 368–375.
4. Franaszek, P. A., "Sequence-State Methods for Run-Length-Limited Coding," IBM J. Res. & Dev., *14*, No. 4 (July 1970), pp. 376–383.
5. van de Waerden, B. L., *Modern Algebra*, New York: Frederick Ungar Pub. Co., 1953, Vol. 1, pp. 84–85.
6. Loève, M., *Probability Theory*, New York: D. Van Nostrand Co., 1960, p. 207.
7. McMillan, B., "History of a Problem," J. Soc. Ind. and Appl. Math., *3*, (1955), pp. 114–128.
8. Shepp, L. A., "Covariances of Unit Processes," Unpublished Conference Notes, Working Conference on Stochastic Processes, IDA, Santa Barbara, Calif., 1969.
9. Gallager, R., G. *Information Theory and Reliable Communication*, New York: John Wiley & Son, 1968.
10. Feller, W., *An Introduction to Probability Theory and Its Applications*, Vol. I, New York: John Wiley & Son, 1957.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 51, No. 3, March, 1972
Printed in U.S.A.

Fabrication and Performance Considerations of Charge-Transfer Dynamic Shift Registers

By C. N. BERGLUND and R. J. STRAIN

(Manuscript received September 30, 1971)

Two similar dynamic shift register schemes have recently been proposed, the insulated-gate-field-effect transistor (IGFET) version of the bucket-brigade register and the charge-coupled device (CCD). These charge-transfer dynamic shift registers show great promise for many digital and analog applications because of their small size and simplicity. In this paper the fabrication, performance, and drive characteristics of the registers are considered in detail to bring out common and comparative capabilities, the discussion being supplemented with presently available experimental data. In order to be specific in comparing the two devices, we assume 10-micrometer metallization tolerances and emphasize digital rather than analog operation of the registers. In addition, a refractory gate technology with two levels of metallization has been assumed so that the devices can be compared using similar technologies. With respect to the common capabilities, it is found that the registers have several significant advantages over other existing shift register schemes—a minimum of processing steps leading to areas of under 3 mil^2 per bit, with the possibility of areas down to 1.1 mil^2 per bit using a refractory gate technology; operation up to frequencies of 10 MHz p-channel or 50 MHz n-channel; and power requirements under 5 μW per bit at a clock frequency of 1 MHz, power varying approximately linearly with clock frequency. From a comparative point of view, it is found that the charge-coupled device and the IGFET bucket brigade are so similar that area limitations, voltage, current, and power requirements, and high- and low-frequency operating limitations arise from the same mechanisms and hence are essentially the same within less than a factor of two. There appear to be only two major differences. First, the fabrication requirements are somewhat different. The CCD requires no diffusions in its active region and may be less sensitive to mask realignment when two levels of metallization are available. However,

under the restrictions of a single level of metallization and 10 μm tolerances, only bucket-brigade registers have presently been successfully fabricated. Second, at intermediate frequencies of operation both registers are capable of transfer efficiencies in excess of 99.9 percent. The residual inefficiency in the bucket-brigade register is due to the nonzero IGFET dynamic drain conductance; the CCD has no analogous limitation so its intermediate frequency performance, limited only by interface state trapping of mobile carriers, has the potential for more efficient charge transfer.

I. INTRODUCTION

There has been a renewed interest recently in the use of charge storage on p-n junctions or capacitors for memory and shift register applications.[1-3] Because leakage currents gradually degrade the stored information, the memories require periodic refreshing, but the simplicity, speed, small size, and modest power requirements of many schemes and their compatibility with silicon technology can often offset the disadvantage of the regeneration requirement. This class of memories can be divided into two general groups—those which are designed for random access, and those which are inherently dynamic shift registers and can only be accessed serially. This article is particularly concerned with the dynamic shift register.

Several dynamic shift registers have been proposed which use charge storage on metal-oxide semiconductor (MOS) capacitors[3-6] and at least one of these is presently available commerically.[6] Two of the most promising are the CCD first described by Boyle and Smith,[3] and the IGFET bucket-brigade shift register reported by Sangster and Teer.[4,5] These two register schemes are similar in many respects, and we shall refer to both of them as charge-transfer shift registers. The purpose of this paper is to describe in detail the operation, fabrication, and performance of these two MOS charge-transfer shift registers, with particular emphasis on digital applications, and to bring out both their similarities and differences. In order to provide a basis for comparison, we assume 10 μm metallization photolithographic tolerances. It should be recognized that such an assumption, while it simplifies the comparison, greatly restricts the possible fabrication schemes. Hence our results and conclusions may change if a different set of ground rules is defined, and care must be taken in extrapolating the results to improved or different fabrication capabilities.

In the next section, a qualitative description of charge-transfer shift register operation is given, and the points where the two registers differ are identified; in the third section, the fabrication requirements

and techniques are outlined and the ultimate size limitations described; in the fourth section, some experimental and theoretical results on performance limitations are discussed; and in the final section the common and comparative capabilities of charge-transfer shift registers are summarized.

II. DESCRIPTION OF CHARGE-TRANSFER SHIFT REGISTERS

This description of the two dynamic charge-transfer shift registers the IGFET bucket brigade and the CCD, will proceed for convenience on the assumption that both devices are fabricated on n-type silicon substrates. Both operate on the basic principle of moving holes from position to position along the silicon surface by making the potential of the forward position more negative (hence attractive to holes) while the trailing position is made less negative. The devices differ in the mechanisms used to establish unidirectional information transfer, and this difference is reflected in the device construction.

The bucket-brigade concept has been described by Janssen[7] and by Hannan, et al.,[8] and the operation of the IGFET version has been described by Sangster[5] and by Berglund and Boll.[9] By its nature, the bucket brigade is a two-phase device although extension to multiphase operation is straightforward. This means that each element or half-bit has an asymmetry which insures that propagation will occur only in the forward direction. In the MOS bucket brigade, the half-bit consists of a p-region covered with SiO_2 and coupled capacitatively to a metal clock electrode which is arranged to induce a field-effect transistor (FET) channel between the p-region and that associated with the preceding half-bit. The structure which will accomplish this is diagrammed in Fig. 1 which shows a four-bit MOS bucket-brigade shift register. The two-phase clocking is accomplished by driving all the odd gates with a periodic signal $\varphi_1(t)$ and the evens with $\varphi_2(t)$; these signals are ordinarily the same but displaced in time by one-half period.

The unloaded state of the bucket brigade is that state where all the p-islands are strongly reverse biased with respect to the substrate; this means they have a deficiency of holes. A signal would be introduced by adding some holes to the first p-island, and this reduces its negative potential. Information transfer is effected by driving φ_2 to a negative value and at the same time driving φ_1 toward a more positive potential. The p-island coupled to φ_2 now becomes the drain of an IGFET which has its channel coupled to the φ_2 gate, and the p-island coupled to φ_1 is the source. If the source is carrying some signal charge, a channel will be induced between the two p-islands, and the charge will flow

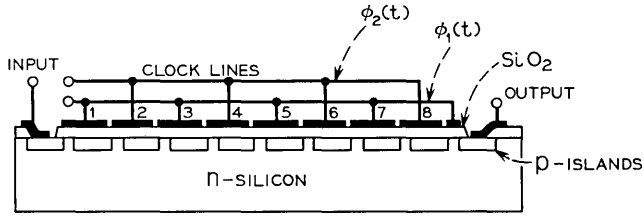


Fig. 1—Simple four-bit IGFET bucket-brigade shift register.

forward into the drain until the potential on the source p-island is reduced to the point where the difference between the source voltage and the gate voltage equals the threshold voltage; then there is no more channel. As clock signals φ_1 and φ_2 reverse their phases, the *evens* become sources and the *odds* drains, and the entire cycle is repeated. By this means, information is advanced two p-islands (one bit) for each cycle of the clock. Transfer in the reverse direction does not occur because the voltage on the gate over the channel is always more positive than the voltage on the p-island capacitatively coupled to that gate, so that the reverse channel never turns on.

The CCD concept was first described by Boyle and Smith.[3] The device consists of a series of MOS capacitors placed very close together and driven by two, three, or more clock signals. In its unloaded state, all the capacitors are driven into deep depletion (a transient state) and no minority carriers are present at the surface. Information would be represented as a controlled number of minority carriers trapped under one or more of the plates. In its simplest realization, diagrammed in Fig. 2, three electrodes are used per bit, and the device is driven by three sequential periodic signals $\varphi_1(t)$, $\varphi_2(t)$, and $\varphi_3(t)$. In this case the holes representing a signal would be injected into the device when the first phase is most negative, and the holes are trapped under the first electrode. Transfer is effected when φ_2 becomes more negative than φ_1 ; then the holes move from the first electrode to the second by a combination of drift and diffusion. The sequential application of the clock potentials moves the information through the device, and reverse transfer does not occur because φ_1 , φ_2 , and φ_3 occur in a sequence which assures that when the electrode ahead of the charge is becoming more negative, the one behind is becoming more positive. Thus, the electrostatic attraction is always in the forward direction.

If two-phase drive is desired for CCD operation, the simple structure shown in Fig. 2 will not suffice. The charge transfer mechanism has no inherent directionality, so each half-bit must be designed to inhibit

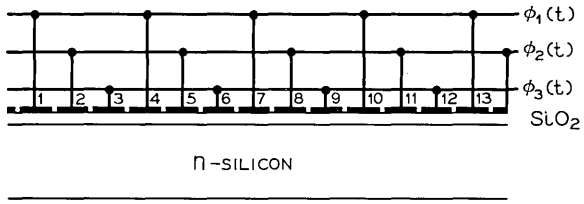


Fig. 2—Simple three-phase charge-coupled device.

propagation in the reverse direction. This can be achieved, for example, by arranging a periodic alteration in the channel threshold voltage—such as by using two different oxide thicknesses, or two different doping levels at the surface.[10] Such devices will be described in the next section.

III. FABRICATION CONSIDERATIONS

3.1 *IGFET Bucket Brigade*

3.1.1 *Description*

The basic steps required for fabrication of an IGFET bucket-brigade shift register (two-phase) using conventional diffusion and metallization technology are listed in Table I and illustrated by the cutaway view

TABLE I—TWO TYPES OF BUCKET-BRIGADE FABRICATION

Bucket-Brigade Elements (Regenerator Elements)	Conventional Processing	Refractory Gate Processing
1. Field Oxide (~1 μm)	Oxidation Photomasking Etching	Oxidation Photomasking Etching
2. Diffused Islands (Sources and Drains)	Photomasking Etching Diffusion	
3. Gate Insulation	Oxidation CVD Insulator	Oxidation CVD Insulator (Optional)
4. Refractory Metallization	No	Deposition from Chemical Vapor Deposition of Insulator Photomasking
5. Refractory Gate Definition (Gates)	No	Etching Diffusion (Ion Implantation Optional)
6. Sealing Insulation	No	Sputtered or CVD Insulator
7. Contact Holes	Photomasking Etching	Photomasking Etching
8. Metallization	Evaporation Photomasking Plating (Optional) Etching	Evaporation Photomasking Plating (Optional) Etching

in Fig. 3. An experimental register chip fabricated this way is shown in Fig. 4.[9] There are no critical steps or photolithography tolerances beyond those normally encountered in IGFET processing. Apart from input and output, the operation of the register is self-compensating for variations in IGFET threshold,[9] and since somewhat heavier channel doping than for most IGFET applications will normally be used, the register should have less stringent requirements on stability, interface state density, and reproducibility than other IGFET circuits.

For comparison to CCD registers, it is also of interest to examine the fabrication of an IGFET bucket-brigade shift register assuming that a refractory gate (either silicon or refractory metal) technology is available. The basic fabrication steps in this case are also listed in Table I assuming silicon gates as a specific example, and they can be followed by referring to the cutaway view in Fig. 5. Even though bucket-brigade registers can be fabricated without a two-level metallization, the fabricational advantages of such a technology should be emphasized at this point. First, a self-aligned gate is achieved so that undesirable capacitance from one gate to the preceding p-island is minimized without a critical alignment step; second, the p-islands are also self-aligned with respect to the thin oxide; third, a second level of metallization is available for associated circuitry on the chip; and finally, these advantages are all achieved with, at most, an increase of only one mask, depending on the particular fabrication scheme.

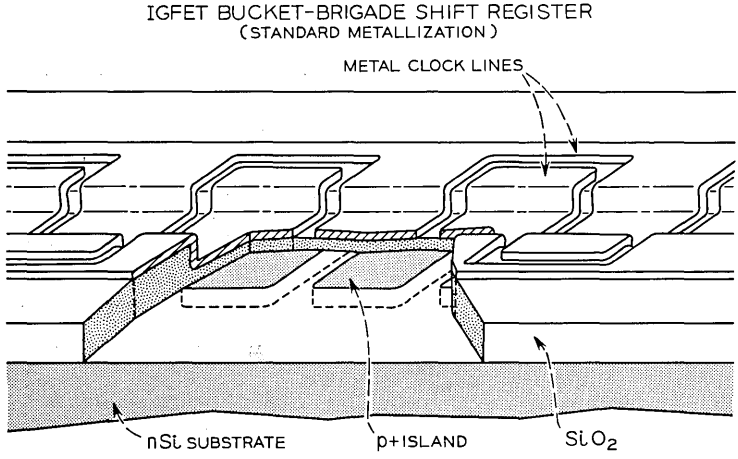


Fig. 3—Cutaway view of integrated bucket-brigade register—standard metallization.

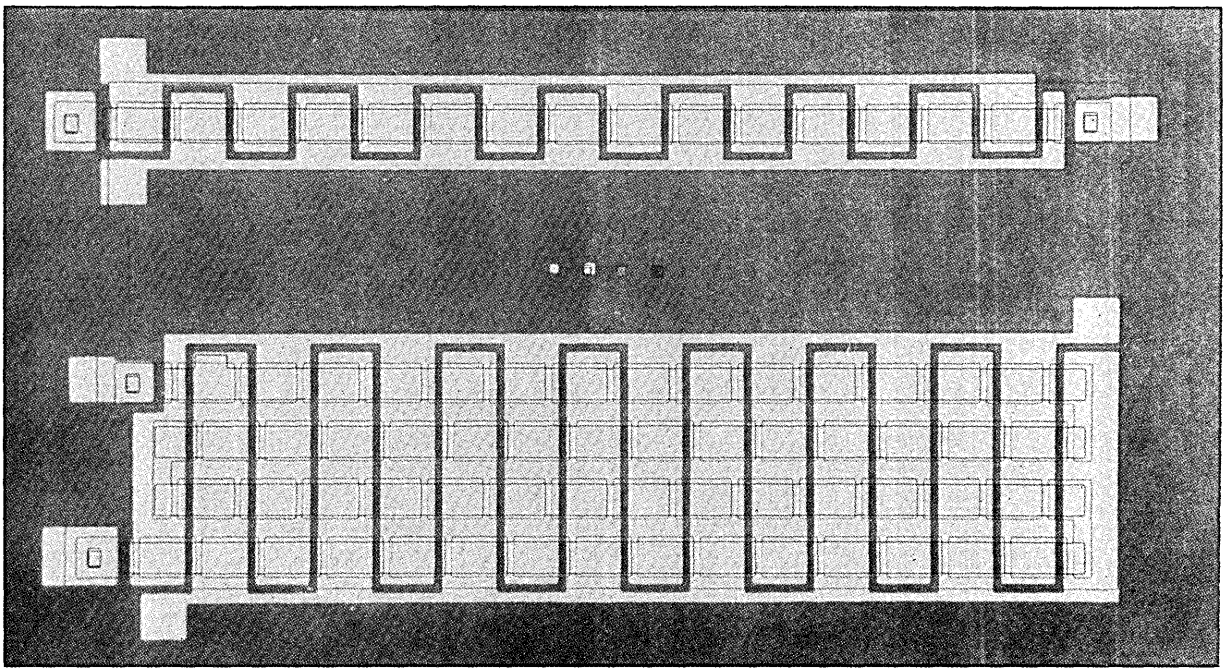


Fig. 4—Photograph of experimental bucket-brigade register chip showing an 8-bit string and a 31-bit string. Metal clock lines are approximately 3 mils wide.

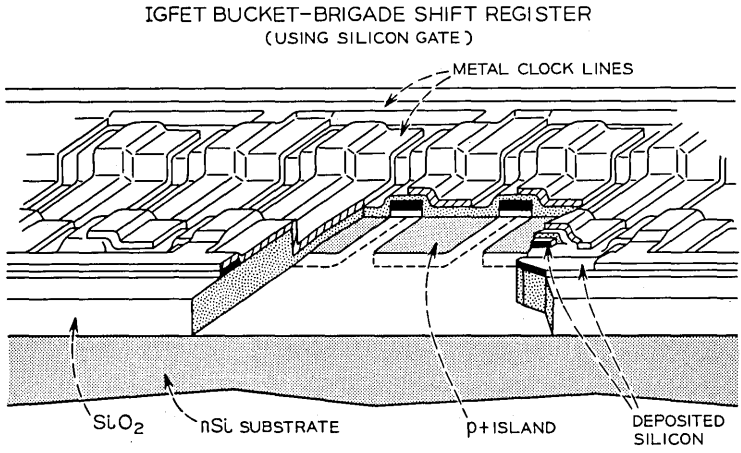


Fig. 5—Cutaway view of integrated bucket-brigade register—silicon gate technology.

3.1.2 *Shift Register Size and Tolerance Limitations*

In the shift register fabrication which we have described, we have restricted ourselves to the basic processing steps and ignored the extra steps associated with channel-stop diffusions, second insulator deposition, and beam-lead processing which may be necessary for any particular application. These have not been included because of the large number of alternatives available to the designer. Similarly, we have not discussed input, regeneration, and output circuitry and their compatibility with the register, including any restriction they may place on the geometry, operating voltages, and minimum charge that must be transferred (i.e., minimum size of the shift register). For these reasons, an estimate of the minimum area per bit in the register obtained only from the basic processing steps restricted by photolithographic tolerances will be somewhat optimistic. Nevertheless, such an estimate does provide a basis for comparison of similar shift register schemes, and it is some measure of processing simplicity.

In this paper we shall estimate the minimum shift register area by arbitrarily assuming the following tolerances in the device fabrication: minimum metallization etching tolerance, 10 μm; mask realignment tolerance, 4 μm; and minimum oxide etching tolerance, 7.5 μm. Under these restrictions, an IGFET bucket-brigade register fabricated using standard metallization according to Table I must have a length per bit (along the direction of charge transfer) which corresponds to two p-islands and two IGFET gates. The gates can be 7.5 μm long, but

the p-islands must be large enough to allow a 4 μm metal overlap with the preceding metal gate, 10 μm metallization spacing, and at least 18 μm for the intentional overlap capacitance. This last figure is derived from the requirement that the desired overlap must exceed the unavoidable overlap with the previous p-island even under worst case misalignment. Hence, the minimum bit length for such a shift register is 79 μm. The width of each bit must include p-regions of at least 15.5 μm, 8 μm due to mask realignment, and 7.5 μm for the oxide etch tolerance, plus an additional 7.5 μm for spacing between adjacent shift register strings. Hence, the minimum bit width is 23 μm and the minimum bit area is 1.8×10^3 square micrometers (2.8 mils^2).

If we now assume the register is to be fabricated using refractory gate technology according to Table I and Fig. 5, the minimum length per bit will include two 10 μm self-aligned gates, and two p-islands, each of which must include 4 μm for spacing between the metallization and the silicon gate and a minimum of 10 μm for p-island overlap capacitance. This gives a total bit length of 48 μm. The minimum width per bit, however, is not as easily calculated when using refractory gates since it is determined by the nature of the signal flow in a shift register array. Figure 6 illustrates two principal types of signal flow—parallel and serpentine. In parallel flow, the signal in every string moves in the same direction as might be useful, for example, in a multichannel or multiplexed register or in an imaging device. The serpentine signal flow is that type of flow which will be most prevalent in digital circulating memory applications. Here, the signal reverses itself from left to right as it goes through the array (Fig. 4). For parallel flow, the minimum bit width will be simply two oxide etch tolerances 15 μm,

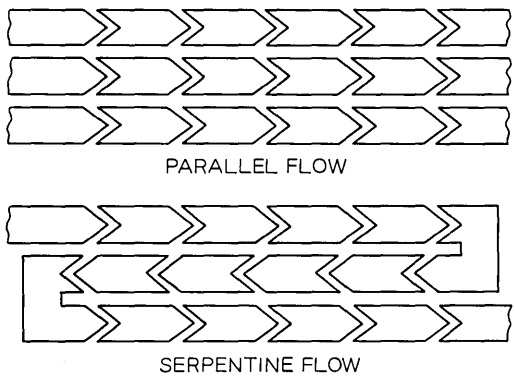


Fig. 6—Illustration of shift register layout for serpentine or parallel signal flow.

one for the p-island and one for the spacing to the next string. Hence, the minimum bit area will be 0.72×10^3 square micrometers (1.1 mils^2). However, for serpentine flow the refractory gate lines and metal lines must crisscross from string to string and additional area will be required. For example, even if the gate length is increased from 10 μm to 14 μm, photolithographic tolerances can only be maintained using a p-island width of 7.5 μm as before but increasing the spacing between strings to 10 μm. Under these conditions the minimum bit length will be 56 μm and minimum bit width will be 17.5 μm for a minimum bit area of 0.98×10^3 square microns (1.5 mils^2). An illustration of the layouts for two-phase arrays for both serpentine and parallel flow using refractory gates is shown in Fig. 7.

3.2 *Charge-Coupled Device*

3.2.1 *Description*

While the conceptual charge-coupled device, a series of MOS capacitors, would appear from its elegant simplicity to be easy to fabricate, there are a number of complicating factors which alter this conclusion. First, for digital applications, it is virtually a necessity to include diffused regions and associated contact holes either for the input, output, and regeneration, or because of the polyphase clock requirement. The eight-bit register of Tompsett, Amelio, and Smith[11] is representative of this simplest class of device, fabricated using conventional metallization techniques. Because it is a three-phase register, it has been necessary to provide diffused crossunders at the expense of area and complexity. This fabrication approach is also limited because the thin channel oxide is exposed, thus making the device performance prone to gross alterations of its operating characteristics when ionic contamination reaches this sensitive surface. The simple expedient of subsequent passivating insulator deposition over the device may reduce the device performance instabilities, but in many cases this will not be a satisfactory solution. Even if the instability were eliminated, experimental and theoretical studies[12] indicate that while CCD operation is possible with wide gaps between metal plates, best operation, particularly with p-channel, is normally achieved when the gaps are smaller than approximately three micrometers, a spacing smaller than the allowable photolithographic tolerances that we have assumed.

In the long run the best fabrication procedure will probably involve the use of two levels of metallization such as that available from re-

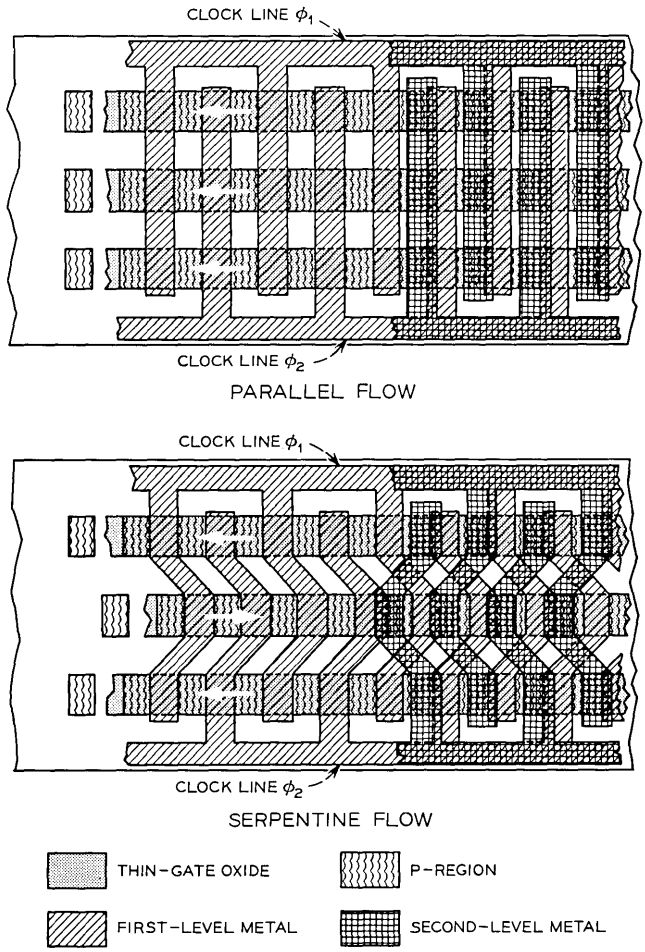


Fig. 7—Layout for serpentine and parallel flow arrays assuming a two-level metal technology for a bucket brigade.

fractory gate technology since the problem of the spacing between adjacent plates can be solved without requiring more stringent photolithographic tolerances. Within this context, it is natural to make two-phase or four-phase[13] rather than three-phase devices primarily because of the difficulty in laying out serpentine signal flow in three-phase. Four-phase CCD's may have some performance advantages and be more versatile, while two-phase registers may be smaller and easier to lay out in arrays. In this article we restrict ourselves to CCD's

made using two-level metallization and we ignore schemes for fabricating CCD's using conventional metallization. Figure 8 is a sectional drawing of a multilevel CCD. It is clear that the problems associated with the gaps between adjacent capacitor plates has been eliminated because the gaps are reduced to the oxide thickness; further, the channel is entirely sealed by metallization to prevent ionic contamination of the oxide. For two-phase operation, directionality may be incorporated in the half-bits by providing two levels of oxide thickness, or by differential surface doping.[10]

Conventional CCD operation will be most efficient when the silicon substrate doping is relatively light. This is in contrast to the relatively heavy channel doping required for the IGFET bucket-brigade register. As a result, it may be necessary to provide channel-stop doping to inhibit the inversion of the surface outside the CCD channel.

3.2.2 *Shift Register Size and Tolerance Limitations*

For comparison with Fig. 7, arrays with serpentine and parallel layouts of two- or four-phase CCD's using refractory gate technology according to Table II are shown in Fig. 9. An estimate of minimum bit size will be obtained only for these registers for comparison to the IGFET bucket-brigade registers.

Applying the same fabrication tolerances as previously described, the minimum length per bit must include four metallization tolerances

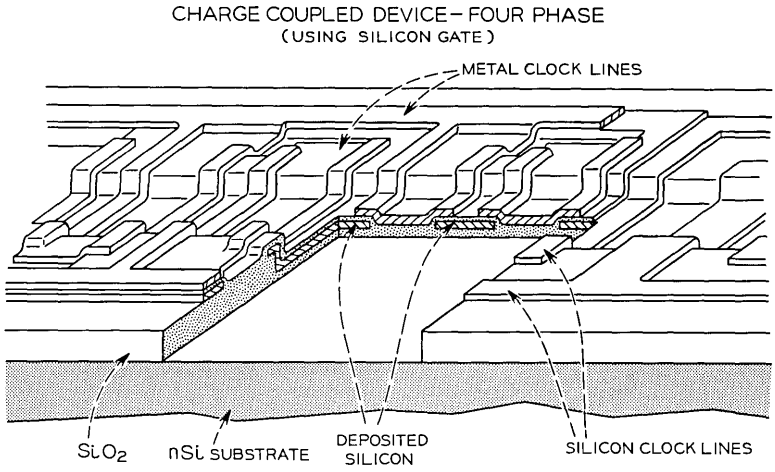


Fig. 8—Sectional drawing of a multilevel charge-coupled device.

TABLE II—REFRACTORY GATE PROCESSING OF CHARGE-COUPLED DEVICES

CCD Elements (Regenerator Elements)	Processing
1. Field Oxide	Oxidation Photomasking Etching
2. Gate Insulation	Oxidation CVD Insulators (Optional)
3. Refractory Metallization	Deposition from Chemical Vapor Deposition of Insulator
4. Input and Output Diffusions (Sources and Drains)	Photomasking Etching Diffusion (Ion Implantation Optional)
5. Refractory CCD Plates Definition (Gates)	Photomasking Etching
6. Sealing Insulation	Oxidation CVD Insulator Optional
7. Contact Holes	Photomasking Etching
8. Metallization	Evaporation Photomasking Plating (Optional) Etching

for a total of 40 μm. However, because of the overlap requirement between the two metallization levels, the refractory metallization must be increased by two realignment tolerances per half-bit, or a total of 16 μm. For two-phase operation, then, the minimum bit length will be 56 μm, and for four-phase operation where the plates would usually be the same size, the minimum bit length will be 72 μm. For parallel strings, the minimum bit width will be the same as for the IGFET bucket-brigade register, 15 μm. Hence, the minimum bit area of a CCD will be 0.84×10^3 square microns (1.3 mils^2) and 1.08×10^3 square microns (1.7 mils^2) for two- and four-phase operation respectively. For serpentine arrays, there again must be an increase in area per bit because of the crisscrossing of the two metallization levels. While there are several compromises available, for comparison with the bucket-brigade register we will increase the minimum length per bit on the CCD to allow the same minimum width of 17.5 μm for the serpentine array. This requires an increase in bit length to 76 μm for the two-phase register and 92 μm for the four-phase register. Hence, for serpentine signal flow the minimum CCD bit area will be 1.33×10^3 square microns (2.0 mils^2) and 1.61×10^3 square microns (2.5 mils^2) for two- and four-phase operation respectively.

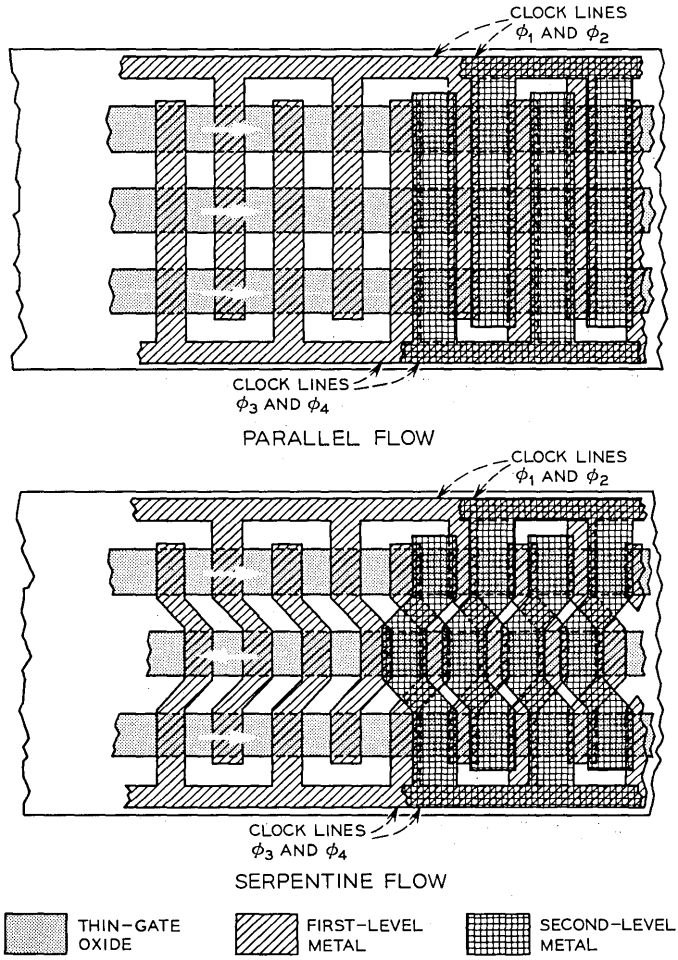


Fig. 9—Layout for serpentine and parallel flow arrays assuming a two-level metal technology for charge-coupled devices.

3.3 *Regeneration and Output Circuitry*

When a shift register is to be used for digital applications, periodic regeneration of the pulses will be required. This will be the case even if the shift register has a perfect transfer characteristic because of generation currents in the reverse-biased p-islands of the IGFET bucket-brigade register or in the depletion layers of the CCD. Regeneration of the pulse train can, in principle, be accomplished off the

chip if a sufficiently low leakage current and high transfer efficiency can be achieved so that hundreds of bits can be allowed before regeneration. However, in general the charge transferred in a register is so small that some amplification and output circuitry is always required on the chip in order to obtain usable impedance and signal levels off the chip. Since the processing steps necessary for regeneration are the same as those required for this amplification and output circuitry, there will often be no fabricational advantage to attempting regeneration off the chip. If this is the case, the extra area required to put regenerators every 10 bits, for example, rather than every 100 bits may be negligible when considered together with the improved margins and less severe performance requirements that accompany more closely spaced regenerators. Depending on the complexity of the regenerator, some compromise spacing will be found, and it will probably be of the order of 5 to 50 bits.

Since for digital applications some output circuitry will always be required on the chip and regenerators may be required every 5 to 50 bits, the compatibility between the extra circuitry and the basic shift register properties and fabrication requirements may become rather important. For example, because of the shift register simplicity, the fabrication steps required for the basic register may be a subset of the steps required for the output and regeneration circuitry. In this case, the fabrication complexity of the output and regeneration circuitry would determine the fabrication complexity of the register chip. Similarly, the speed of the regenerator or output circuit may be the important limitation to the speed of the register operation, and the minimum signal required to drive the regenerator or output circuitry may provide a more important limitation to the size of the register than photolithography tolerances. In this way, it is possible that the output and regeneration circuitry could provide important factors in determining many of the fabricational and operational limitations of charge-transfer shift registers.

IV. SHIFT REGISTER PERFORMANCE

The performance of both the bucket brigade and the charge-coupled device as shift registers must be described in terms of two rather different but interrelated properties, signal transmission and drive requirements. It will be shown that the signal degradation properties of these registers depend very strongly on signal amplitude, and as a consequence, device performance is invariably enhanced if information is represented by rather large quantities of charge and by assuming that all cells are

never totally devoid of charge during operation. The charge-carrying ability of a register is roughly proportional to the drive voltage. On the other hand, the drive power increases with the drive voltage. Hence, there will always be a close link between signal transmission characteristics and drive power.

While the practical limitations to overall performance of a register may often be determined by the input, regeneration, and output circuitry, the inherent limitations of the register itself are of great importance. These must arise from the characteristics of the step-by-step charge transfer operation within the register. In each step the charge is moved forward by one element, but invariably some of the charge lags behind. This effect will always provide a limit to the number of bits in a register before unacceptable signal degradation occurs. The less time available for charge transfer, as at increased clock frequencies, the more charge is left behind in each transfer. Ultimately this will result in a high-frequency limitation to shift register operation for any given number of bits.

At low frequencies, another limitation arises. Generation centers in the depleted regions of the semiconductor and states at the interface give rise to thermal generation currents; and these tend to fill the potential wells which have been intentionally left partially empty. This provides a general shift to higher charge levels in all bits and will eventually lead to unacceptable signal degradation as the register begins to be overloaded. Depending on the temperature of operation, generation currents will thus impose a low frequency limit for operation.

To simplify the discussion which follows, the important charge transfer and drive characteristics of the bucket brigade and charge-coupled device will be discussed separately but in a parallel fashion. Then those general characteristics which are common to the two registers, such as low-frequency limitations and waveform and signal level effects, will be described. The discussion of signal transmission characteristics of charge transfer shift registers, however, is complicated not only by the fact that the signal degradation is extremely nonlinear but also by the fact that the amount of degradation depends on the nature of the signal;[14] i.e., in a digital context it depends on the bit sequence. This means that the exact relationship between the performance of a given shift register and the charge transfer characteristics of each stage is very complex. One approach to simplifying this problem is to linearize the transfer characteristics by making a small-signal approximation.[14] In this way an estimate of the dependence of shift register performance on the nature of the signal can be made in terms

of an incomplete transfer parameter characteristic of each stage of charge transfer.[14,15] Another approach is to assume a worst case situation, a signal which is an alternating series of ONE's and ZERO's, and assume the fractional signal degradation of all stages in the register is the same as that of the first stage.[16] In this way the nonlinearity of the incomplete transfer can be illustrated by varying the magnitudes of the ONE and the ZERO. Both approaches will be used here, but it should be emphasized that both are approximations and are intended only to give an indication of the limitations of charge transfer shift registers.

In the paragraphs that follow, for convenience, the particular devices analyzed theoretically will be those cited as the minimum size devices in the previous section; and they both assume the same design rules and a two-level metallization technology. Because the field of charge-transfer devices is still evolving, some of the theoretical expressions represent approximations which are in the process of being refined. Consequently, the estimated performance represents, in each case, the application of the best theoretical work currently in hand. These best estimates will be subject to further refinement both by improved analysis and subsequent experimental work.

4.1 *Bucket-Brigade Performance*

4.1.1 *Transfer Characteristics*

It is possible to consider the operation of the bucket brigade as a sequence of capacitor discharges, where the discharging current path is an insulated gate field effect transistor with finite nonlinear transconductance, and the current sink is another identical capacitor C. Using this approach, Berglund and Boll have determined the charge Q_r left behind in the course of a single transfer, assuming a square-wave drive.[9] The result is an algebraic function of the signal level and transfer time t:

$$\frac{V_r}{V_i} = \frac{1}{1 + KV_i t}, \tag{1}$$

where V_i is the voltage associated with the initial charge Q_i stored in the capacitor C, V_r is the voltage associated with Q_r , Q_r/C, and K is a normalizing factor which is a function of device parameters and capacitance C. The behavior of V_r for two specific signal levels, V_i of 2 volts and 10 volts, is shown in Fig. 10 where time has been arbitrarily normalized in units of K^{-1}. Note that for times beyond $t' = 1$, the

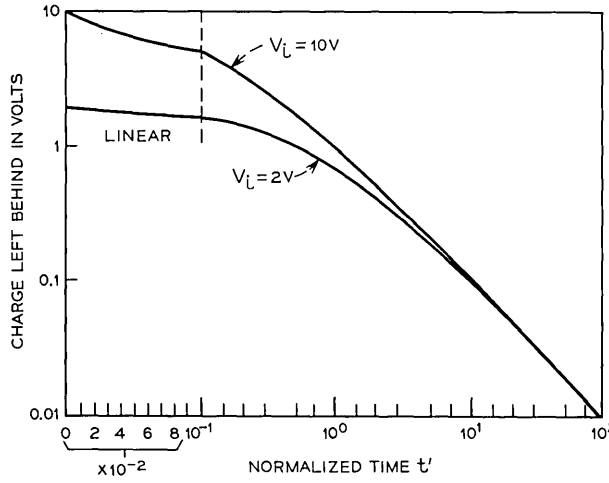


Fig. 10—Charge remaining on a bucket-brigade capacitor versus normalized time for two initial values of charge.

amount of charge remaining, V_r , becomes relatively independent of the initial charge. For this reason signal degradation should be reduced by assuring that relatively large quantities of charge are always transferring through the register.[9]

This effect can be made more evident by defining a large signal digital degradation factor Γ. This is the fractional change in the difference between an adjacent ONE and ZERO in a sequence of alternating ONE's and ZERO's on transferring through the first stage of a shift register.

$$\Gamma = 2p \frac{(Q_{r1} - Q_{r0})}{(Q_{i1} - Q_{i0})}. \tag{2}$$

The factor 2 comes from the fact that the charge lost by a ONE is added to the ZERO, and the p, representing the number of phases in the register, is the number of charge transfers per stage. Figure 11 shows the variation of Γ with the ratio Q_0/Q_1 for a register with the previously described minimum dimensions operating at 10 MHz. The abscissa is the ratio of the amplitude of a ZERO to that of a ONE, keeping the ONE constant at 5 volts, and curves are presented for n-channel (mobility, μ = 500 cm^2/V-sec) and p-channel (μ = 150 cm^2/V-sec) devices. Three points are evident from this plot. First is the advantage gained by having the higher mobility of n-channel devices (at large values of Q_0 , Γ varies approximately as μ^{-2}); second

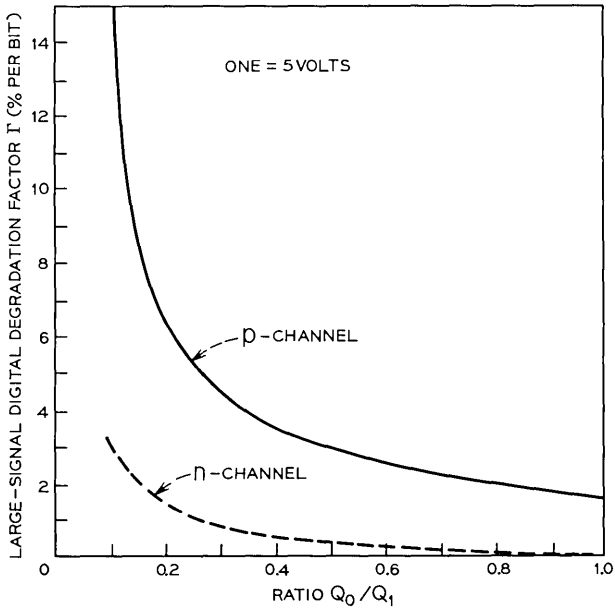


Fig. 11—Variation of signal degradation with size of the ZERO in a bucket brigade.

is the steady diminution of Γ with increasing size of the ZERO; third is the tendency for Γ to decrease less rapidly as the size of the ZERO increases.

Several other features of charge transfer in bucket-brigade shift registers can be conveniently illustrated by using the small-signal incomplete transfer parameter α defined by Berglund.[14]

$$\alpha = \frac{dQ_r}{dQ_i}. \tag{3}$$

This is the differential charge left behind after a single transfer due to a differential change in the charge being transferred. Thornber[17] has proposed that since α is usually quite small, the various contributions to incomplete transfer can be described by writing

$$\alpha = \frac{\partial Q_r}{\partial Q_i} + \sum_i \frac{\partial Q_r}{\partial X_i}\frac{dX_i}{dQ_i}. \tag{4}$$

In this expression, X_i might be any one of the parameters associated with the transfer, like the times of initiation and completion of transfer or the channel length, which had been assumed constant in calculating Q_r but which in fact depends somewhat on Q_i .

The first term in equation (4) is the intrinsic term α_i and has been analyzed by Berglund and Boll[9] for square-wave clock drive and by Thornber[17] for arbitrary clock waveform. Thornber concludes that this term can be made exponentially small with respect to the sum of the terms in equation (4) provided that the quantity (Cf_c/AV_a) is much smaller than unity. In this expression, C is the capacitance in which the signal charge is stored (typically it will consist primarily of the metal gate overlap with the p-islands, but it will also include the effects of gate-to-channel capacitance and the p-island-to-substrate capacitance), f_c is the clock frequency, V_a can be considered to a first approximation to be the voltage associated with the average signal charge $Q_a(Q_a \cong CV_a)$, and A is the parameter which relates drain current I_D to source-gate voltage V_{SG} in an ideal IGFET through

$$I_D = A(V_{SG} - V_T)^2, \tag{5}$$

where V_T is the IGFET threshold voltage. Figure 12 illustrates the clock frequency dependence of the intrinsic contribution to incomplete transfer, α_i , by the dashed lines as calculated by Thornber for the specific case of sine-wave drive. In this calculation the previously described minimum dimensions and a signal voltage V_a of 5 volts have been assumed. Note that this contribution to the total α represented

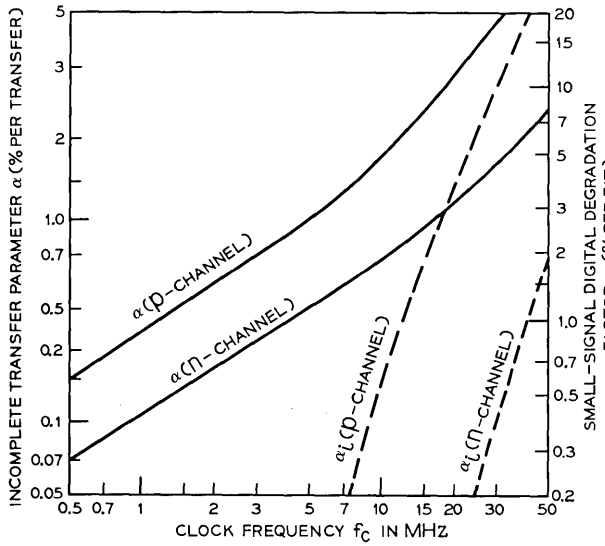


Fig. 12—Theoretical incomplete transfer parameter versus clock frequency for minimum size bucket-brigade shift register.

by the solid lines is modest even to frequencies as high as 50 MHz, and decreases exponentially as the frequency decreases.

In the previous discussion, it was pointed out that α_i varies exponentially with the quantity Cf_c/AV_a . In all other analyses performed to date on incomplete transfer including that leading to Fig. 10, the clock frequency always appears in the product Cf_c/AV where the voltage V is a voltage associated with the signal or drive level, or some combination of the two. Since the signal and the drive voltages are typically within a factor of two, it is useful to define the quantity C/AV_a as the characteristic bucket-brigade time constant τ_B . Also, since AV_a is the IGFET transconductance at a source-gate bias of $V_a/2$, τ_B is the time constant for charging the capacitor C through the IGFET transconductance at source-gate bias $V_a/2$. Alternately, for comparison to charge-coupled devices, τ_B can be written in terms of geometrical and materials properties as

$$\tau_B = \frac{L_G L_O}{\mu V_a} , \tag{6}$$

where μ is the field effect mobility, L_G is the channel length, and L_O is the effective gate overlap length defined as that required to make the overlap oxide capacitance equal to C. Because of the contributions of gate and p-island capacitances, L_O will always be somewhat larger than the actual overlap.

As α_i decreases with decreasing clock frequencies, other contributions to α begin to dominate. One of the most important terms in equation (4) has been found to be that associated with the IGFET dynamic drain conductance.[9,17] Since the transfer characteristic is dependent not only on the source-to-gate voltage but also weakly on the drain voltage, and since the drain voltage depends on the charge transferred, nonzero dynamic drain conductance introduces an additional dependence of Q_r on Q_i . Assuming that this effect is entirely due to channel length modulation, Thornber has calculated this contribution to α, α_D , for sinusoidal clock waveform, and his results can be sumsumarized by the approximate expression

$$\alpha_D \cong \frac{0.3}{L_G} (f_c \tau_B)^{\frac{2}{3}} \left(\frac{V_a}{N_c}\right)^{\frac{1}{2}}, \tag{7}$$

where L_G is the channel length in microns, V_a is the signal voltage in volts, and N_c is the channel doping density in units of 10^{16} cm^{-3}. This contribution has been included in Fig. 12 to yield the solid curves assuming N_c equal to 10^{16} cm^{-3}.

It should be noted that if the actual dependence of τ_B on L_G is included in equation (7), there is an optimum channel-length-to-overlap-length ratio which minimizes α_D . Depending on channel doping, this minimum is rather broad and is centered approximately near L_G equal to the actual overlap length. The minimum value of α_D decreases as L_G decreases.

Interface states in the channel region also make a contribution to α. However, they typically add less than 0.1 percent even when the interface state density is in the 10^{11} states/cm^2-eV range and nonoptimum clock waveforms are assumed. For this reason and because interface state effects will be described in detail in the section on charge-coupled devices, no additional discussion will be included here.

While experimental data on bucket-brigade registers is relatively sparse, most of the qualitative features of incomplete transfer described above have been experimentally observed. The large signal effects illustrated in Fig. 11 have been demonstrated using a register constructed from discrete components, and experimental data from integrated registers in the 1 to 10 MHz range have verified the frequency, clock waveform, and voltage magnitude effects predicted in the discussion of equation (4). Figure 13 shows measured values of α as a function of clock frequency for a register using a distorted trapezoidal waveform. Also included for comparison are theoretical curves for the dimensions and device parameters used assuming sine and ideal trapezoidal clock waveforms. Over the measured frequency range, the channel length modulation effect dominates α and can be seen to explain both the approximate magnitudes and the frequency dependence of the data. Additional measurements on registers in which the channel doping was intentionally increased have verified that α_D is the major contribution to α under most conditions. By such a technique it appears that values of α smaller than 10^{-3} at 1 MHz can be achieved. Figure 14 shows the measured dependence of α at 1 MHz on the amplitude of the clock waveform for a signal level chosen to give optimum results. The main features of the experimental data are quite accurately reproduced; namely, the rapid increase in α at the smaller clock voltages as the intrinsic contribution to α becomes important, and the tendency for the measured α to saturate at higher clock voltages as the dynamic drain conductance term dominates. While the data indicate that shift register operation at frequencies up to and exceeding 10 MHz with α less than 0.01 should be easily attainable, it is possible that other terms in equation (4) not considered here may become important as the frequency increases. For this reason no estimate of the ultimate

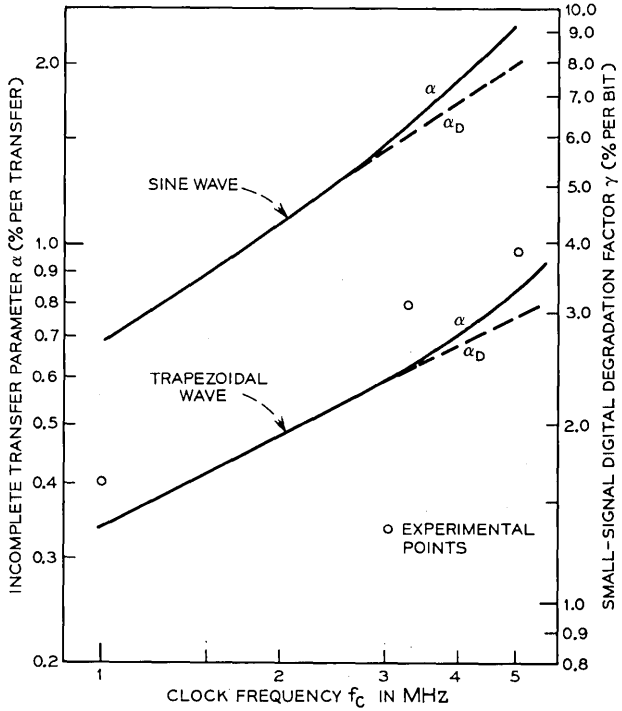


Fig. 13—Comparison of experimental and theoretical incomplete transfer parameter as a function of clock frequency for a bucket brigade.

high-frequency limit for IGFET bucket-brigade operation is available.

4.1.2 *Drive Characteristics*

The drive characteristics of a bucket-brigade register must include the voltage and current requirements on the clock power supply and the power dissipated on the chip. These characteristics are all clock waveform dependent, and hence difficult to describe exactly for the general case. However, clock power supply requirements of major interest are the peak values of current and voltage rather than the detailed time dependences, and these peak values as well as the power dissipated on the chip can be estimated in a relatively simple way. The clock waveform effects will be discussed in a separate section.

The minimum voltage levels required by the IGFET bucket-brigade shift register and the relationship between these voltages and the maximum charge that can be transferred have been briefly discussed by Berglund and Boll.[9] The values are best illustrated by assuming

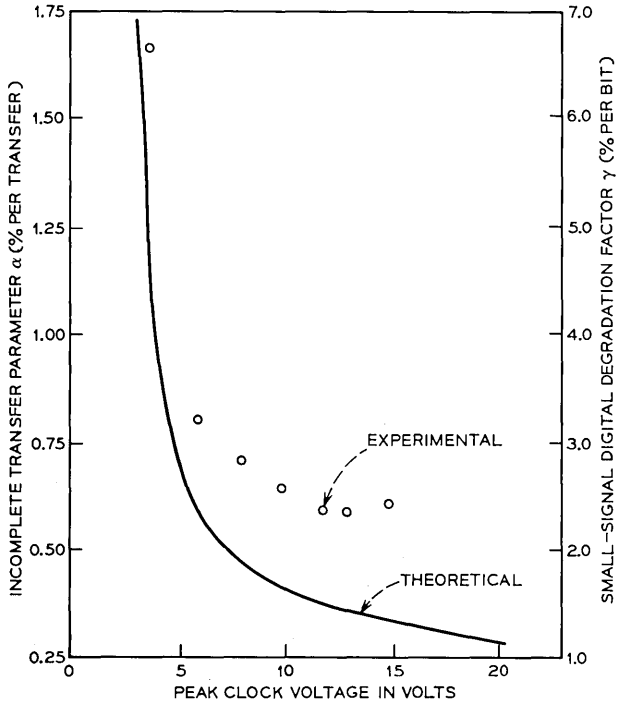


Fig. 14—Comparison of experimental and theoretical incomplete transfer parameter as a function of peak clock voltage for a bucket brigade.

that we are dealing with a p-channel register and that the capacitance C is entirely made up of oxide overlap capacitance. Then if V_n and V_p are the most negative and most positive values of clock voltage respectively, each p-island in the register under the no charge transfer condition will vary in potential from a most positive value of $V_n - V_T$ to a most negative value of $2V_n - V_p - V_T$ where V_T is the IGFET threshold voltage. Since each p-island must always be negatively biased, V_n must at least be sufficiently negative that $V_n - V_T$ is a negative voltage. This condition defines the minimum allowable negative excursion of the clock voltage. Also, since the maximum charge that can be transferred is $Q_M = C(V_p - V_n)$, and since the register performance has been shown to be very sensitive to the magnitude of charge transferred, the minimum difference between V_p and V_n can be defined given the performance requirements of the register. Hence, the most positive and most negative values of the clock voltage for any application are rather simply defined by the IGFET threshold

voltage under operating conditions and the desired register characteristics. Because of back gate bias, it should be noted that the IGFET threshold voltage will be different from that when source is shorted to the substrate.

The peak current that will flow in a clock line will consist of two contributions. One term will be due to the substrate capacitive loading on the clock lines and will include the displacement current to the substrate through this capacitance. Generally, registers will be designed to minimize this capacitive loading such that this term will be negligible. The other term will be the current which flows from one clock line to the other due to the charge transfer itself. Its value will depend on the quantity of charge being transferred in each bit. In any given charge transfer event the peak value of this current will typically occur approximately τ_B after initiation of charge transfer where τ_B is given by equation (6), and at this time the peak current contributed by that bit of the register will be given approximately by

$$I_p = C \frac{dV_{cc}}{dt} , \tag{8}$$

where V_{cc} is the voltage between clock lines. The peak current required from the clock power supply will be given by equation (8) multiplied by the number of bits in the register. Figure 15 illustrates the values of I_p per bit assuming sinusoidal clock voltages of 10 volts peak-to-peak for a register fabricated with the previously defined minimum dimensions.

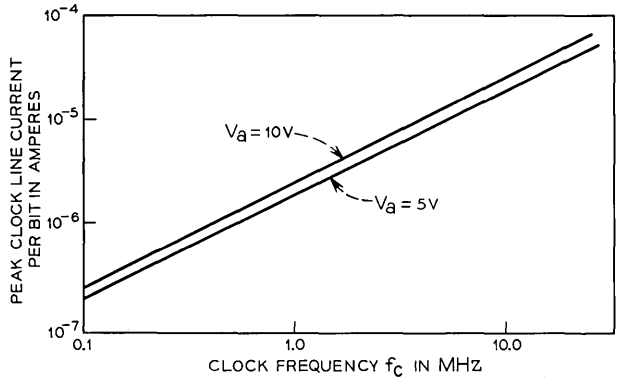


Fig. 15—Theoretical peak clock current required per bit as a function of clock frequency for a bucket brigade.

The average power dissipation on the chip, like the peak current and voltage requirements on the clock supply, is dependent on clock waveform. However, a relatively simple expression can be derived in which the detailed dependence on clock waveform can be lumped into one term. This term will often be negligible under practical operating conditions.

Assuming series resistance of the substrate is negligible, the major power loss on the chip is associated with charge transfer from one capacitor C to the next. Suppose that the capacitance C is entirely due to the oxide overlap capacitance and that a periodic clock voltage $V_c(t)$ with peak-to-peak magnitude V_{pp} is applied to the two clock lines, one delayed one-half a clock period with respect to the other. Then during transfer the source-to-drain voltage magnitude is given by

$$V_{SD} = 2V_c(t) + \frac{2Q}{C} - \frac{Q_a}{C}, \tag{9}$$

where Q is the instantaneous charge remaining on the source capacitance and Q_a is the total charge to be transferred. The power dissipation per bit, P_{DISS}, is given by the energy loss per cycle multiplied by the number of transfers per clock period and by the clock frequency f_c.

$$\begin{aligned} P_{\text{DISS}} &= 2f_c \int_0^{Q_a} \left[2V_c(t) + \frac{2Q}{C} - \frac{Q_a}{C} \right] dQ \\ &= 2f_c \int_0^{Q_a} 2V_c(t)\, dQ. \end{aligned} \tag{10}$$

From equation (5), $V_c(t)$ is related to Q by

$$I = \frac{dQ}{dt} = A\left[2V_c(t) - V_{pp} + \frac{Q}{C} \right]^2. \tag{11}$$

Hence, from equations (10) and (11)

$$P_{\text{DISS}} = 2f_c \left[V_{pp}Q_a - \frac{Q_a^2}{2C} + \int_0^{Q_a} \sqrt{\frac{I}{A}}\, dQ \right]. \tag{12}$$

The last integral in equation (12) is the only term dependent on the details of the clock waveform, and can be recognized as the integral over a clock period of the IGFET source-gate voltage above threshold. Defining this integral as $V_{av}Q_a$ where V_{av} is some average source-gate voltage above threshold during transfer of Q_a, equation (12) becomes

$$P_{\text{DISS}} = 2Q_a f_c \left[V_{pp} + V_{av} - \frac{Q_a}{2C} \right]. \tag{13}$$

Note that V_{av} will tend to zero in the limit of low frequencies and will

probably never be comparable to V_{pp} in normal operation. For this reason, depending on the clock waveform, V_{av} can probably be neglected compared to V_{pp} in order to obtain estimates of P_{DISS} . This means that a first-order estimate of power dissipation in a two-phase bucket-brigade register is independent of the details of the clock waveform but dependent only on Q_a , f_c , and the peak-to-peak clock voltage V_{pp} .

Figure 16 illustrates the 1-MHz power dissipation per bit as a function of charge being transferred, assuming the previously defined minimum dimensions. While the previous discussion has centered on the power dissipated on the substrate of the charge transfer shift register, the driving supply must provide the current (e.g., Fig. 15) necessary to both move the charge, and establish the pattern of stored energy associated with the new charge distribution. Figure 16 has been prepared assuming an ideal sine wave voltage driver, so the power to move the charge is the real power, while the modifications of stored energy can be considered to be reflected in the reactive power. The reactive power is simply the product of the sine wave clock voltage and the quadrature Fourier component of the current. Both power components rise with increasing charge, and the real power starts linearly from zero, while the reactive power is fairly large for zero charge because of the capacitative loading of the substrate (assumed to be doped to 10^{16} carriers/cm^3). This component can be lowered by using nonuniform doping profiles. The total power (since this is a sine wave analysis) is the phasor sum of the real and reactive powers.

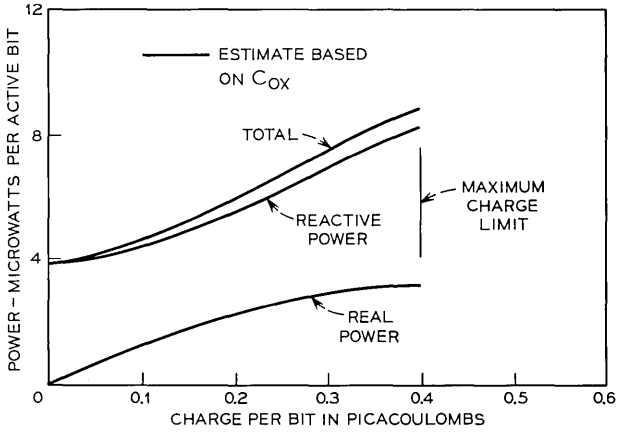


Fig. 16—Theoretical power requirements per bit versus signal charge at 1 MHz for a bucket brigade: peak-to-peak clock voltage of 10 volts.

Figure 17 shows the frequency dependence of the total power dissipation for a bucket brigade loaded to approximately half its absolute maximum value. This figure also illustrates the fact that the drive power varies roughly as the square of the drive voltage. At higher frequencies, the drive power will tend to rise more rapidly than linearly, because the importance of V_{av} increases when more charge is left behind.

4.2 *Charge-Coupled Device Performance*

4.2.1 *Transfer Characteristics*

Difficult as the analysis of a bucket brigade is, the charge-coupled device presents even more formidable difficulties. In the general case, charge motion in CCD's is governed by a nonlinear relative of the diffusion equation, with driving forces coming from external fields, electrostatic repulsion of the mobile charges, and density gradients. Numerical solutions of this equation have been obtained in specific cases, but in the absence of general solutions, many approximations have been employed to reach reasonable estimates of CCD performance.

In charge-coupled devices with relatively large plates, there is negligible penetration of fringing fields under the plates, and the charge transport is governed by this approximate equation:[16]

$$\frac{\partial q}{\partial t} = \frac{\mu}{C}\left(\frac{\partial q}{\partial x}\right)^2 + \mu\left(\frac{q}{C} + \frac{kT}{e}\right)\frac{\partial^2 q}{\partial x^2}, \tag{14}$$

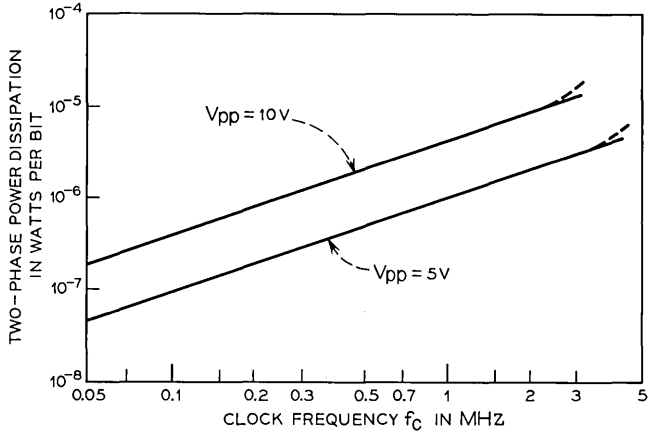


Fig. 17—Theoretical power dissipation per bit versus clock frequency for a bucket brigade.

where q is the mobile surface charge density, n is the surface mobility, C is the capacitance per unit area, approximately equal to the oxide capacitance ϵ_o/δ, but modified by space charge repulsion and depletion layer loading. The other symbols have their usual meanings. The assumption of negligible fringing fields is analogous to the "gradual channel" approximation in IGFET theory.

In solving this problem, it is found that time scales naturally against τ_o defined as

$$\tau_o = \frac{L_p^2}{\mu V_o}, \tag{15}$$

where L_p is the length of a CCD plate and V_o is arbitrarily taken to be one volt. Also, the total charge is most readily described by a voltage $V_a = \int q\, dx/CL$. The results[16] of a numerical solution of equation (14), presented as the amount of charge remaining under the plate as a function of time normalized to τ_o, in analogy to the similar calculation [equation (1) and Fig. 10] for bucket brigade, appear in Fig. 18. Note that for normalized times exceeding one, the charge quickly becomes independent of the initial charge. While time in Fig. 18 is normalized

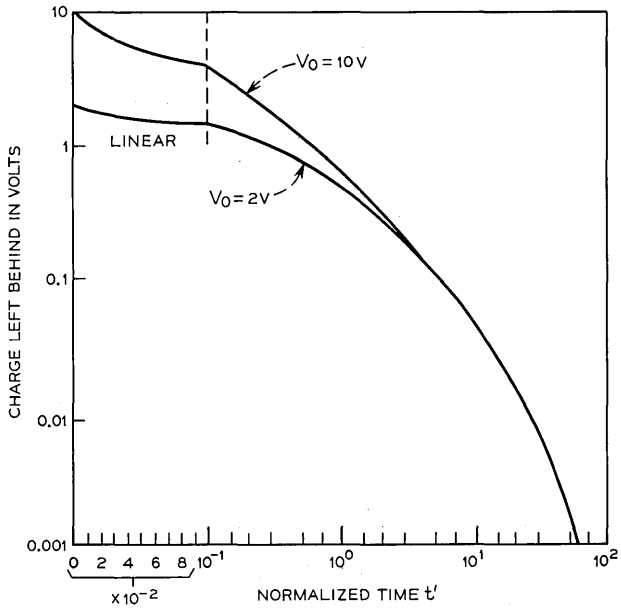


Fig. 18—Charge remaining on a CCD capacitor versus normalized time for two initial values of charge.

to τ_o as defined by equation (15), it is useful to also define a characteristic time for the CCD $\tau_c = L^2/\mu V_a$ for direct comparison to the bucket brigade τ_B in equation (6). As long as V_a is much greater than kT/e, it is found that the transport as determined by solving equation (14) scales with τ_c in much the same way that bucket-brigade transport scales with τ_B .

In Fig. 18 the advantage of a "fat" ZERO in digital charge-coupled devices is strongly suggested,[18] and it becomes clearer in Fig. 19. The large-signal digital degradation factor Γ is shown for a 5-volt ONE at 10 MHz, again in analogy with the bucket-brigade results (Fig. 11), with the ZERO varying from completely empty to nearly equal to the ONE. Equation (2) has been applied to the specific four-phase charge-coupled device described earlier, with four 18-μm plates formed by two levels of metallization. While the loss drops to the order of one percent in n-channel devices, the p-channel devices have a much higher percentage of loss at this operating frequency.

The loss associated with the finite rate of charge transport, illustrated by the calculated curves in Figs. 18 and 19, is similar to the intrinsic contribution to the incomplete transfer parameter, α_i , in bucket brigade. As is the case with the bucket-brigade register, this effect

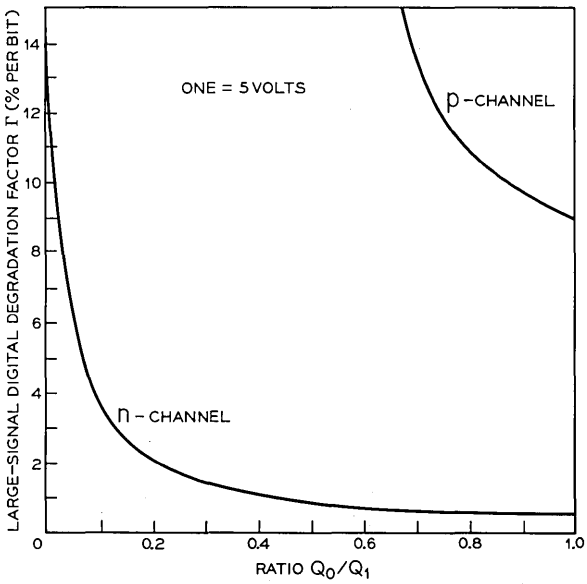


Fig. 19—Variation of signal degradation with size of the ZERO in a charge-coupled device.

becomes dominant at high frequencies and ultimately provides the high-frequency limitation for operation. The dashed curves in Fig. 20 show the frequency dependence of $\partial Q_r/\partial Q_i$ as calculated using equation (14) assuming a four-phase CCD with the previously defined minimum dimensions. To make it easier to include interface-state effects later, we have plotted the small-signal equivalent γ of the digital degradation factor Γ defined in equation (2),

$$\gamma = \lim_{Q_0 \to Q_1} \Gamma, \tag{16}$$

rather than the incomplete transfer parameter α. The two are simply related by

$$\gamma = 2p\alpha \tag{17}$$

so the bucket brigade results have been plotted using both parameters in order to simplify comparisons. Note that this intrinsic contribution to incomplete transfer decreases exponentially as the clock frequency decreases (as is the case with the bucket-brigade register), and that it becomes important at a somewhat lower frequency. The latter is due

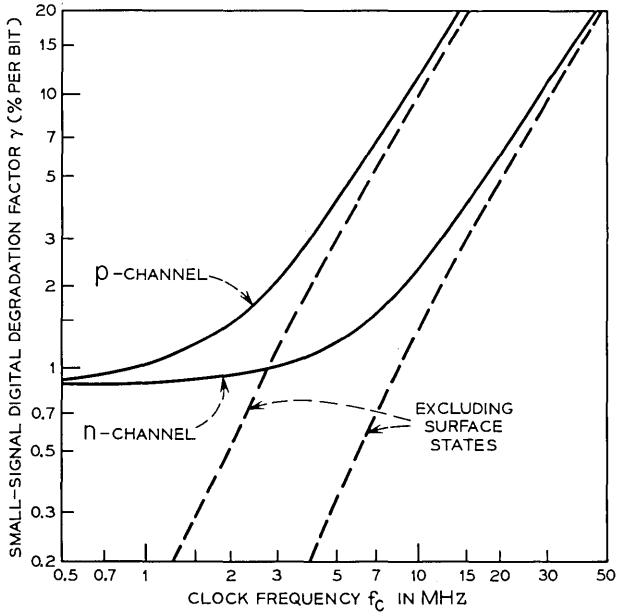


Fig. 20—Theoretical digital degradation factor versus clock frequency for minimum size four-phase CCD.

partly to the fact that the minimum register dimensions over which minority carrier transport occurs are somewhat larger for the CCD and partly to the four-phase mode of operation.

As the frequency of operation decreases, other contributions to incomplete transfer must begin to dominate. However, the CCD has no contribution analogous to the dynamic drain conductance term in the bucket brigade. Since it is believed that interface states will make a significant contribution to incomplete transfer at the lower frequencies, it is this term which will be considered next.

There are three principal ways for the interface states to interact with the information transport. The first interaction occurs through the generation of minority carriers at the surface, the second is by acting as a recombination site, and the third and most serious is their ability to retard mobile carriers, leading both to attenuation and distortion of the information. The generation of minority carriers at the surface adds to the bulk generation in depleted areas, and acts to limit low-frequency operation, as discussed separately in a later section. The recombination process only assumes importance if, at some time during the charge transfer cycle, a portion of the surface is drawn into accumulation. When this happens, majority carriers are trapped in states across a considerable portion of the bandgap, and when information in the form of minority carriers reaches this surface, recombination effects attenuation of the signal. This difficulty is readily eliminated in both charge transfer shift registers by assuring that no surface is ever drawn into accumulation.

Incomplete charge transfer due to interface-state trapping, on the other hand, is a more serious effect, and can be illustrated by using the formulation introduced in equation (4). If we define the charge left behind in interface states as Q_τ' and the interface-state contribution to incomplete transfer as α_{ss}, we can write

$$\alpha_{ss} = \frac{dQ_\tau'}{dQ_i}. \tag{18}$$

Now suppose there is a time τ following transfer of the signal charge beyond which charges still trapped in interface states will be left behind. Then equation (18) can be written

$$\alpha_{ss} = \frac{\partial Q_\tau'}{\partial Q_i} + \frac{\partial Q_\tau'}{\partial \tau}\frac{d\tau}{dQ_i}. \tag{19}$$

Since interface-state emission times vary exponentially with energy, after time τ essentially all the interface states below some energy

ϵ_1 corresponding to an emission time τ will be filled and all those above will be empty. Hence, the first term in equation (19) can be made exponentially small by making Q_i large enough to fill the interface states to a level above ϵ_1 on each cycle; i.e., the use of "fat" ZERO's will make the first term in equation (19) negligibly small. The second term, however, reflects the difference in total emission time available to interface states as a function of signal level, and will be very dependent on clock waveform.

The interface-state effects have been analyzed for CCD operation using an idealized traveling wave model,[18] and the results have been used to yield the solid curves in Fig. 20 assuming a sinusoidal clock voltage with 10 volts peak-to-peak. In this model, the optimum transfer characteristics occur when the signal charge is 5 volts. An interface-state density of 10^{11} states/cm^2/eV was assumed in preparing the data. The interface-state contribution to incomplete transfer leads to the tendency for γ to saturate as the clock frequency decreases as shown in Fig. 20. Further, the value to which the γ tends is independent of the type of carrier. Hence, when interface states dominate as at low frequencies, CCD operation will be relatively independent of whether it is fabricated on n- or p-type substrates.

In the preceding discussion of CCD transfer characteristics we have analyzed, for simplicity, the performance of four-phase registers. Two-phase CCD operation, while probably very similar to that of a four-phase CCD, will depend in detail on the scheme used to achieve directionality, but three-phase operation can be predicted with only a minor change in the four-phase theoretical treatment. This has been done to compare the theory to existing experimental data in Fig. 21. The experimental points were obtained from an 8-bit, three-phase, p-channel register with 50 μm plates operating with a sawtooth clock voltage varying between -1.5 volts and -10 volts.[19] Also shown for comparison are results from a two-phase p-channel register made using silicon-gate technology with 50 μm plates operating with a square-wave clock voltage between -2 volts and -8 volts.[20] Two-phase operation was achieved by using two thicknesses of oxide. The reasonable agreement between theory and experiment and the strong frequency dependence of Γ indicates that the intrinsic transfer rate is limiting the performance of both CCD's, and that the theory of Strain and Schryer[16] represents a reasonable theoretical approximation to this limitation. On the other hand, greatly improved performance over that shown in Fig. 21 will result by going to smaller sizes. In fact, transfer efficiencies in excess of 99.9 percent at 1 MHz have already been reported.[19]

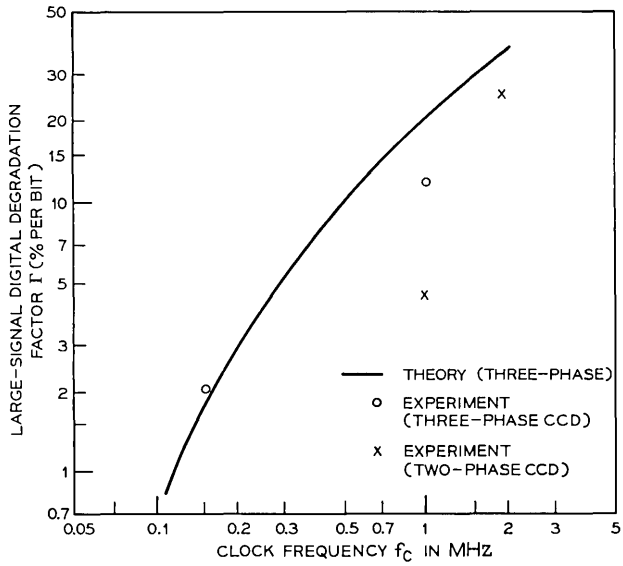


Fig. 21—Comparison of experimental and theoretical digital degradation factor as a function of clock frequency for a charge-coupled device.

4.2.2 *Drive Characteristics*

The operating voltage requirements of the CCD will also vary depending on the clock waveform, but in general they can be determined by considerations similar to those used for the bucket-brigade register. Assuming a p-channel register, the most positive voltage V_p on any clock line must not allow accumulation of majority carriers at the interface because of the loss that would be introduced through interface states. Hence, if the most negative flat-band voltage in the active region of the CCD is V_{FB} , V_p must be more negative than V_{FB} .

The minimum peak-to-peak clock voltage swing, V_{pp} , will be determined, as is the case for the bucket brigade, by the desired operating frequency and required transfer characteristics. Since the signal degradation is reduced as the charge being transferred is increased, and since the maximum charge that can be transferred is directly proportional to V_{pp} , the clock voltage requirements are rather simply defined by the flat-band voltage and the performance requirements of the register.

The peak current required from the clock power supply will be made up of the same two components existing in the bucket-brigade register—the displacement current to the substrate through the capacitive loading and the current between adjacent clock lines associated with

the charge transfer. The substrate current will generally be small, so the current associated with the transfer will usually dominate. Its peak value will occur approximately τ_c after initiation of transfer where τ_c is given by equation (15), and it too will be given approximately by equation (8) with C equal to the CCD plate capacitance and V_{cc} representing the voltage between adjacent clock lines. Thus, as a first approximation, the peak current required from a CCD clock supply is independent of the number of phases, given that the storage capacitance C_p is fixed, and is essentially identical to that required by a bucket-brigade register with the same storage capacitance.

The power dissipated on the chip of a two-phase CCD register can be calculated in the same way as that of the bucket-brigade register. However, in this case the dissipation will be dependent on the height of the potential barrier V_s introduced to achieve the two-phase operation as well as on the charge transferred and clock frequency. An analysis similar to that resulting in equation (12) yields

$$P_{\text{DISS}} = 2Q_a f_c \left[V_s + V_{av} - \frac{Q_a}{2C_p} \right], \tag{20}$$

where Q_a is the signal charge, C_p is defined by the half-bit area less the barrier area, and V_{av} here is the average excess voltage above the potential barrier through which the charge moves during transfer.

The power dissipated for polyphase operation can be considerably less than that given by equation (20) for two-phase operation. The major difference comes from the fact that directionality in the register is achieved by proper phasing of the clock lines rather than by building in a potential step. This means that the charge transfer is achieved by moving the carriers through a smaller potential gradient, resulting in less dissipation. A general analysis of this dissipation has not been carried out because of its mathematical complexity and dependence on clock waveform. However, a lower bound of its magnitude can be obtained from the sinusoidal traveling-wave CCD model described by Strain[18] in which the average power dissipation per bit was derived to be

$$P_{\text{DISS}} = Q_a \frac{16 L_p^2 f_c^2}{\mu}. \tag{21}$$

Figure 22 shows this power as a function of frequency for the particular case of $Q_a = 0.5$ picacoulomb. The dissipation can be seen to be very modest over the entire frequency range of interest. Even though equation (21) may provide a somewhat low estimate, it can be concluded

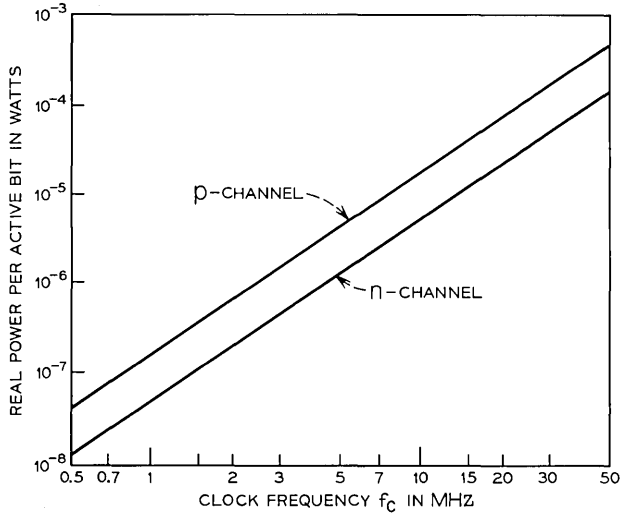


Fig. 22—Theoretical power dissipation per bit versus clock frequency for a polyphase charge-coupled device.

that the power dissipation of a polyphase CCD can be made to be considerably less than that of a two-phase CCD or bucket-brigade register. On the other hand, a polyphase CCD presents a significant reactive load to its driver. Since there is no advantage to using a highly doped substrate for a charge-coupled device, the substrate loading contributes very little to the reactive power required for a CCD, as Fig. 23 indicates. In the traveling-wave model, the reactive power rises to $\frac{1}{2}\omega C_{ox}V^2$ when the device is carrying the maximum possible charge. Optimum operation occurs with approximately one-third that charge, and there the power is just under half that predicted from C_{ox} . So far as the driving source is concerned, the loading is totally reactive; substrate dissipation contributes negligibly.

4.3 *General Considerations*

4.3.1 *Low-Frequency Limitations*

When a charge-transfer shift register is operated at low frequencies, or if the register operation is stopped for some time interval between regenerations, there are two problems which must be considered. First, generation currents due to interface states or bulk generating centers in the space-charge regions of the registers will contribute excess charge to both the ONE's and the ZERO's. Ultimately this

effect will overdrive the register, providing a lower frequency limit for operation and a limit on the maximum number of stages between regenerators. Second, any output and regeneration circuitry must be designed with the realization that only a small, finite amount of charge is available, and any shunt conductance may render the device inoperative at low frequencies. This effect may become very important as the register size is reduced, but since it depends on the detailed design of the regenerator, it will not be considered further here.

Using the usual carrier generation statistics, it can be shown that the generation current density due to interface states J_G is

$$J_G = \frac{\pi}{2} kTn_i v_{th} N_{ss} , \tag{22}$$

where kT is thermal energy, v_{th} is the mean carrier thermal velocity, n_i is the intrinsic carrier density, and N_{ss} is the interface state density near midgap. In the bucket-brigade register, only those interface states in the gate region contribute to J_G and then only for that part of a clock cycle during which the gate region is not transferring charge. In the CCD, the entire surface over which charge is not being transferred

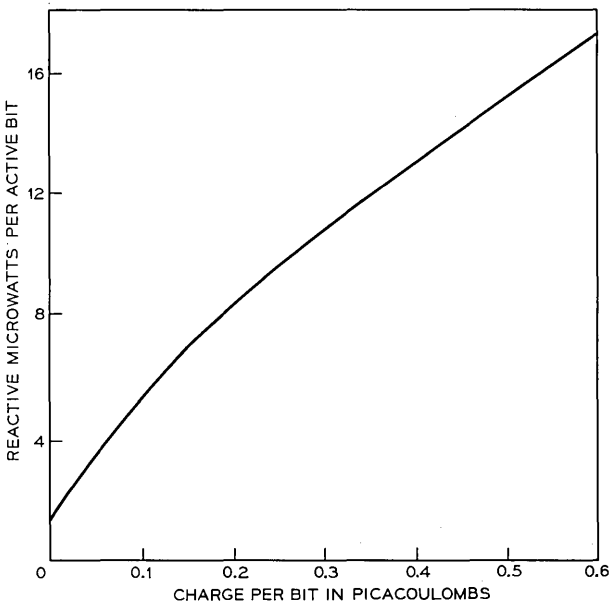


Fig. 23—Theoretical reactive power per bit versus signal charge for four-phase CCD for sinusoidal clock drive.

is active so interface state generation currents will be somewhat more important. Since charge will accumulate in proportion to the time a bit spends in residence in the shift register, Fig. 24 presents an estimate from equation (22) of the ratio of the charge accumulated per microsecond of residence to the nominal charge in a bit, using the minimum shift register sizes that appear throughout the paper and assuming $N_{ss} \cong 5 \times 10^{10}/\text{cm}^2\text{eV}$. The independent variable is temperature and the activation energy is that associated with n_i . Since charging due to bulk centers is likely to be negligible, they have been neglected in preparing Fig. 24.

The compromise between interface state generation currents and charge transfer efficiency can be summarized by relating the various signal degradation parameters to the maximum length of a single-string shift register between regenerators. This has been done in Figs. 25 and 26 which present the maximum length as a function of operating clock frequency assuming the following two limits apply: (*i*) $\Gamma n \leqq 1$ where n is the number of stages in the register, and (*ii*) the fractional increase in the signal charge due to interface state generation current

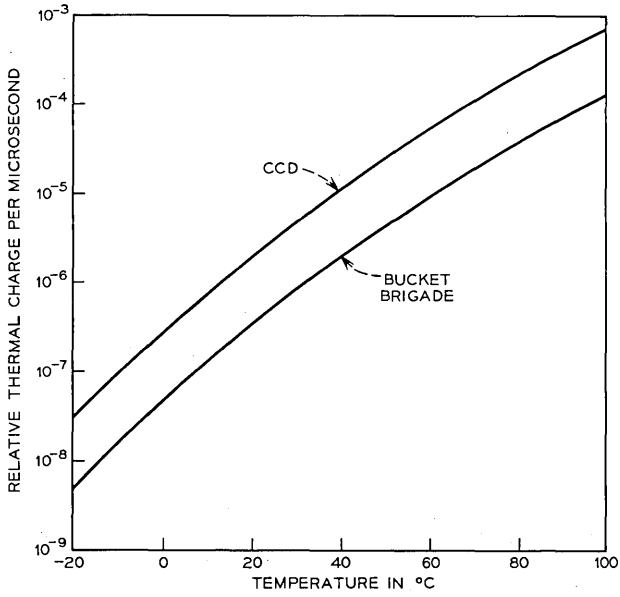


Fig. 24—Ratio of thermal generation charge accumulated per microsecond to signal charge as a function of operating temperature for bucket brigade and charge coupled devices. Only interface states are assumed to contribute to thermal generation.

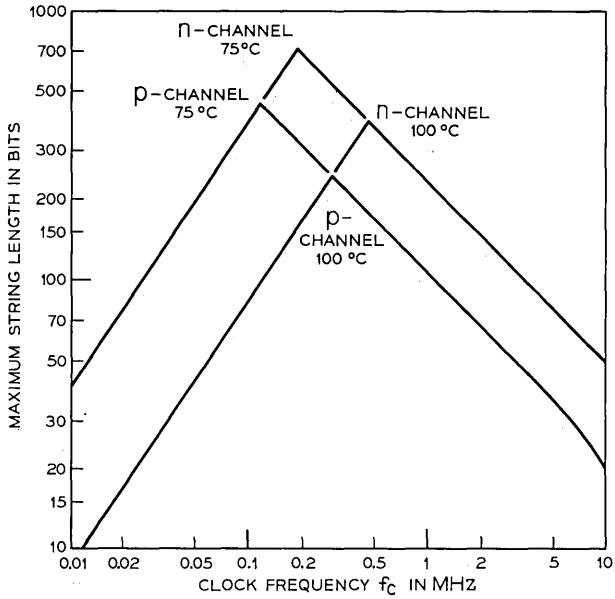


Fig. 25—Maximum length of single shift register string versus clock frequency as limited by thermal generation and incomplete transfer in a bucket brigade.

$\leqq 0.1$. The positive slope portions of the curves represent the limitation due to thermal generation currents; whereas, the negative slope portions represent the limitations of incomplete transfer effects. It is evident that for frequencies above approximately 0.7 MHz, generation currents are not expected to significantly affect charge transfer shift register operation even at 100°C. Further, bit strings up to 50 bits long appear feasible at approximately 2 MHz in p-channel and 10 MHz in n-channel for both CCD's and bucket-brigade registers.

4.3.2 *Clock Waveform and Signal Level Effects*

In the previous sections a dependence of shift register characteristics on both signal level and clock waveform was pointed out and several examples were given. However, because of its importance, it is felt that a separate section describing the effects in more detail is warrented.

The dependence of incomplete transfer effects on the signal charge Q_a reflects the nonlinear transfer characteristics inherent in both register schemes, and without exception all contributions to incomplete transfer which have been considered here decrease as the drive voltage, and consequently the signal charge, is made larger. Assuming this is a

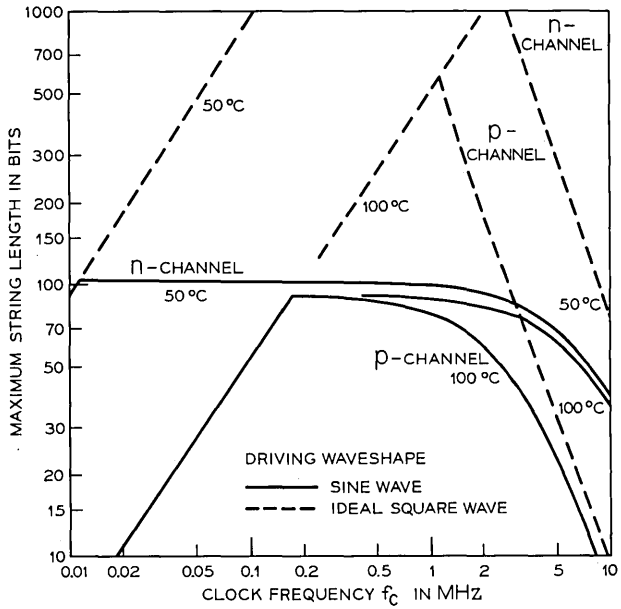


Fig. 26—Maximum length of single shift register string versus clock frequency as limited by thermal generation and incomplete transfer in a charge-coupled device. The dotted lines represent the maximum length if it is assumed interface states do not contribute to incomplete transfer.

general result, it can be concluded that best register operation will occur when the largest possible clock voltages are used and when the optimum fraction of that voltage is used for the signal. For digital applications this means that for fixed clock voltages, for example, a ONE should be represented by something near or greater than one-half the absolute maximum charge that can be transferred and a ZERO by some smaller but still reasonably large quantity of charge; for analog applications this means that the signal should be superimposed on a dc background. In the latter case it is important to recognize the noise disadvantage and the necessity for compromise between transfer efficiency and reduced nonlinear distortion, and signal-to-noise ratio.

The dependence of the various contributions to incomplete transfer on clock waveform differs depending on the particular contribution so that no ideal waveform can be defined. However, all of the contributions for the charge-transfer dynamic registers will either be decreased somewhat or, at worst, remain unchanged by choosing a

clock waveform with both a short rise time and a short fall time. Also, a waveform should be chosen which makes the time interval over which transfer current flows as long as possible. While precautions of this kind can decrease signal degradation, at the same time they place more stringent requirements on the clock power supply. With this in mind it is important to note that either register will operate quite satisfactorily even with sinusoidal clock voltages, and that the frequency characteristics presented in the figures of the previous sections were often calculated assuming sinusoidal clocks and always measured using waveforms with rather poor rise and fall characteristics.

Ignoring for a moment the disadvantages associated with different clock voltage waveforms, it is instructive to illustrate the dependence of the theoretical incomplete transfer effects on the waveform. Figures 27 and 28 show the calculated values for p-channel bucket-brigade and CCD registers respectively, keeping the signal charge and peak-to-peak clock voltage constant but varying the waveform. A channel doping of 10^{16} cm^{-3} has been assumed. Note that bucket-brigade performance is best at high frequencies using trapezoidal waveforms but best at low frequencies using square waves. The crossover comes about as a result of different magnitudes and frequency dependencies

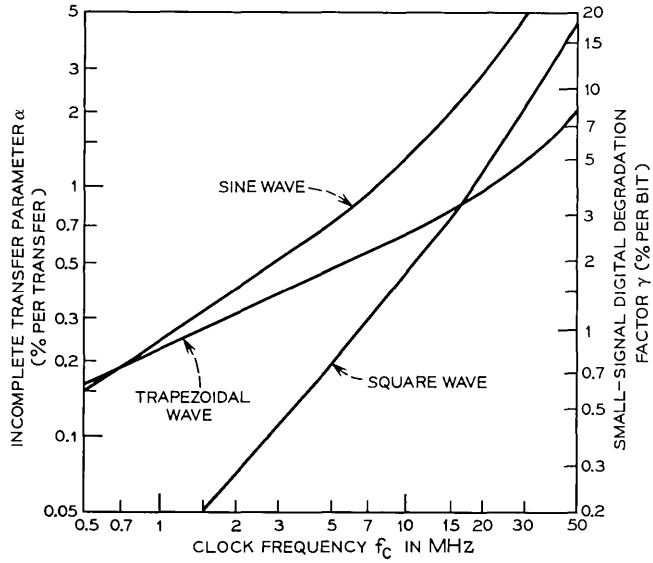


Fig. 27—Illustration of clock waveform dependence of incomplete transfer in a bucket brigade.

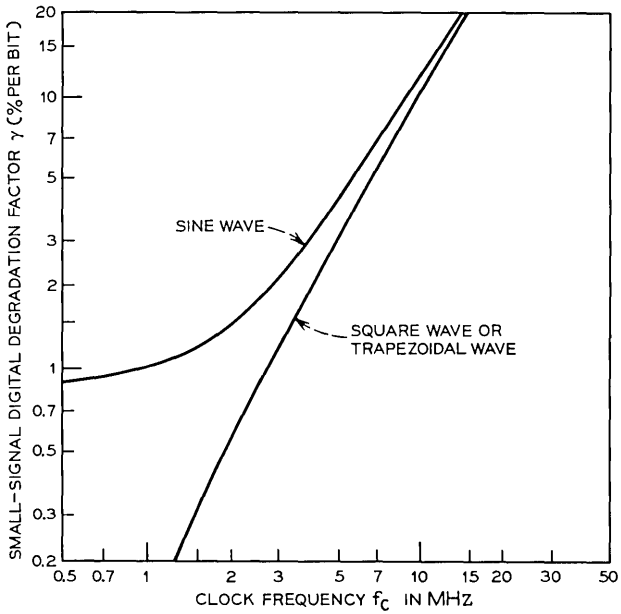


Fig. 28—Illustration of clock waveform dependence of incomplete transfer in a charge-coupled device.

of the intrinsic and drain conductance contributions to α. Note also that degradation due to interface states in the CCD is effectively reduced to zero by using clock waveforms with negligible rise and fall times. This can be explained using equation (19). The first term is α_{ss} is made exponentially small by using a large signal charge, and the second term becomes zero since $d\tau/dQ_i$ is zero.

There is another performance-limiting effect in addition to the increased clock current requirements associated with very short rise and fall times in the clock waveform. To illustrate this effect, suppose that the time required to change a clock line voltage from its most negative value to its most positive value is negligible compared to the time for most of the charge to transfer from one capacitor to the next. Then if a relatively large quantity of charge was present on a capacitor, the surface potential in the case of the CCD or the p-island potential in the case of the IGFET bucket-brigade register could be driven positive with respect to the substrate for a short period of time. If this occurred, some of the holes representing the signal would be injected into the semiconductor bulk where they would be lost to recombination. This injection effect can seriously limit the maximum

charge that can be transferred, thus limiting the transfer efficiency and the maximum frequency of operation. To eliminate this additional signal degradation contribution, one can either include an appropriate dc offset of the clock drive, or limit the clock waveform rise and fall times to at least several τ_B in the case of the bucket brigade and several τ_c in the case of the CCD.

V. SUMMARY

The detailed considerations of the preceding pages can be summarized and the major conclusions stated. Because of the large number of considerations involved, we will subdivide the summary into four sections: size and fabrication considerations, performance limitations and characteristics, drive and power requirements, and regeneration and output requirements.

5.1 *Size and Fabrication Considerations*

If it is assumed that charge-transfer shift register size is limited only by 10 μm photolithographic tolerances, minimum sizes of the two registers, fabricated in serpentine geometry using such tolerances, are indicated below:

	Conventional Metal	Refractory Gate Technology
Bucket Brigade	1.8×10^3 μm^2 (2.8 $mils^2$)	10^3 μm^2 (1.5 $mils^2$)
Two-Phase CCD	not considered	1.3×10^3 μm^2 (2 $mils^2$)
Four-Phase CCD	not considered	1.6×10^3 μm^2 (2.5 $mils^2$)

While factors other than photolithographic tolerances may often be more important in limiting shift register size, these results show that charge-transfer shift registers have the potential for significantly smaller size than other existing shift register schemes.

The fabrication of CCD's using conventional metallization has not been considered because of the difficulty in defining a practical scheme within the restrictions of a 10 μm metallization tolerance. While some novel technology may lead to simplified fabrication schemes for either or both charge transfer registers, the aim here was to consider a practical but relatively simple technology for CCD fabrication within the assumed tolerance restrictions which would allow comparisons with bucket-

brigade registers yet be compatible with additional circuitry on a register chip. This was the reason for choosing a two-level metallization scheme. Assuming such a technology, it can be concluded that the bucket-brigade register can be made slightly smaller than either CCD register. This conclusion comes about because there is no necessity to guarantee metal overlap between adjacent half-bits of the bucket brigade register. On the other hand, the CCD may be less sensitive to the precision of alignment between the two metal levels. With these fabrication differences, it is evident that significant reduction in photolithography tolerances may alter this comparison by making it possible to fabricate practical CCD's using a single level of metallization and by simultaneously emphasizing the realignment problem for bucket brigades.

Approximately the same number of processing steps and masks are required for fabrication of both charge transfer shift registers. However, both registers are so simple and include such a small number of steps than if a two-level metallization is available, the steps will inevitably be included in the processing steps required for the input, output, and regeneration circuitry. The same conclusion applies to the bucket brigade with conventional metallization. Hence, it will be the fabrication complexity of the associated circuitry rather than that of the shift register which will often determine the fabrication complexity of a register chip.

5.2 *Performance Limitations*

Charge-transfer shift register performance becomes limited when the signal transferring through the register is unacceptably degraded. For digital operation this will occur when the difference between ONE's and ZERO's is significantly reduced or becomes noticeably dependent on the bit sequence. Since most signal degradation mechanisms in these shift registers are cumulative, for operation at a given frequency an upper limit to the number of stages will result, while for operation with a given number of stages, a minimum and maximum allowable clock frequency will be defined.

Both charge-transfer shift registers will be limited similarly at low frequencies by generation currents. These currents will come partially from bulk generation in the space-charge regions and partially from interface states in the channel regions of the bucket brigade and over the active surface of the CCD. The difference in area over which interface states can contribute represents the major low-frequency difference between the two registers. If too long a period of time is left between

regeneration of the signal in the register, the generation currents will add to the signal charge and tend to overload the register. With presently available technology it should be possible to operate 50-bit p-channel charge transfer shift registers at frequencies as low as 100 KHz even at temperatures of 100°C; i.e., periods approaching a millisecond between regenerations should be achievable.

At high frequencies, register performance will be limited by the intrinsic rate at which charge can transfer from one capacitor to the next. This leads to a characteristic time constant for transfer which is $L_p^2/\mu V_a$ for the CCD and $L_G L_O/\mu V_a$ for the IGFET bucket brigade, where μ is the field effect mobility, V_a is approximately the voltage associated with the signal charge, L_p is the CCD capacitor plate length, L_G is the IGFET channel length, and L_O is the effective p-island overlap length in the bucket brigade. Because the signal degradation due to this effect varies exponentially with clock frequency, it provides a high-frequency operating limit, but it is inconsequential for operation at lower frequencies. Using the minimum dimensions assumed throughout the paper, it is found that the bucket-brigade register should operate to somewhat higher frequencies than the CCD. However, for both registers it should be possible to operate p-channel up to approximately 10 MHz and n-channel up to approximately 50 MHz even with the assumed 10 μm photolithographic tolerances. With improved tolerances the gradual channel approximation used here for both registers will cease to be valid as fringing fields become important, so that a simple extrapolation of the results using smaller dimensions may be somewhat pessimistic.

At intermediate frequencies both generation current and intrinsic transfer rate effects will be negligible, and other mechanisms will lead to performance limitations of charge transfer shift registers. For the bucket brigade, the most important mechanism in this range is due to the IGFET dynamic drain conductance. Primarily due to channel length modulation, it gives rise to an incomplete charge-transfer contribution which varies slowly with clock frequency. However, the limitation is not serious for most applications since by appropriate doping of the channel region charge transfer efficiencies in excess of 99.9 percent at 1 MHz can be achieved.

For the CCD, there is no contribution to incomplete transfer analogous to the dynamic drain conductance effect in the bucket brigade. However, there is a larger area over which interface states are active. For these reasons interface state effects, which were ignored in the bucket brigade, will probably provide the important limitation to CCD

performance at intermediate frequencies. It is found that their contribution to incomplete transfer is relatively independent of clock frequency and carrier type, but very dependent on the details of the clock voltage waveform. As an example, an interface state density of 10^{11} states per cm^2 per eV will limit transfer efficiency at intermediate frequencies to less than 99.9 percent if the clock waveform is sinusoidal. However, interface states will introduce no limitation at all if the clock waveform is an ideal square wave. Hence, it may be possible to achieve extremely high transfer efficiencies at intermediate frequencies with the CCD by using clock voltages with very short rise and fall times. As will be pointed out later, such operation will be limited by the peak current capabilities of the clock power supply, but in most cases it should be possible to achieve more efficient mid-frequency transfer than with the bucket-brigade register.

All contributions to incomplete transfer in both registers are nonlinear. The nonlinearity is such that best performance is achieved when rather large quantities of charge are transferring through the register at all times. For digital applications, for example, this means that a ONE should be represented by a large quantity of charge and a ZERO should be represented by a slightly smaller, nonzero quantity of charge.

5.3 *Drive Characteristics*

The drive characteristics of charge-transfer shift registers can be described in terms of three related properties: power dissipation on the register chip per bit, clock voltage magnitudes and waveform, and peak current requirements from the clock supply. For two-phase operation, the power dissipation of both charge-transfer shift registers is approximately linear with clock frequency and is a maximum when the largest quantities of charge are being transferred. This maximum power is relatively independent of clock waveform but varies as the square of the peak-to-peak clock voltage. Typical values at 1 MHz for the previously described minimum dimensions will be in the 1-to-5-microwatt-per-bit range assuming a peak-to-peak clock voltage of 10 volts, and will be approximately the same for both the CCD and the bucket brigade. Polyphase CCD's may dissipate considerably less power depending on the details of the clock waveform, perhaps up to one or two orders of magnitude less at 1 MHz. Since directionality is achieved by proper phasing of the clock lines, the charge does not have to transfer through as great a potential difference at each step for polyphase operation.

While the dc level of the clock voltage is an important factor to some of the operating conditions of a register, it is only the peak-to-peak clock voltage magnitude which enters into power dissipation and limits the maximum charge that can be transferred. Because of their nonlinearity, all of the incomplete transfer effects considered here are reduced as the peak-to-peak clock voltage, hence the peak signal charge carried in the register is increased. In this way, the performance requirements of the register place a minimum on the peak-to-peak clock voltage which, for a given storage capacitance, is essentially the same for both charge transfer registers.

There is some performance advantage to shaping the clock voltage waveform, especially by reducing the rise and fall times. However, the peak current from the clock power supply, being essentially displacement current through a capacitance, increases as the rise and fall times are reduced. Since peak current values of the order of one microampere per bit are found to be required for 10-volt operation of minimum size registers at 1 MHz even under sinusoidal operation, the current requirements of a register chip might become completely unrealistic if the rise and fall times are required to be too short. In addition, for some operating conditions, charge injection effects into the semiconductor bulk will limit the maximum charge that can be transferred as the rise and fall times are reduced.

5.4 *Regeneration and Output Circuitry*

In the preceding sections the minimum sizes and optimum performance characteristics of the basic charge-transfer register have been described. Because of the small quantities of charge transferred and small capacitance in which the charge is stored, all register chips will probably have additional circuitry for amplification on the chip. For digital applications, the finite generation currents and incomplete transfer effects will make periodic threshold regeneration necessary in a register string, and this regeneration will most often be performed by circuitry on the register chip itself. This additional output and regeneration circuitry on a register chip may impose serious limitations to both fabrication and performance. The fact that fabrication complexity of the additional circuitry will often determine completely the fabrication complexity of a register chip has already been pointed out. In a similar way the speed of the regeneration or output circuitry may be the important limitation to the speed of the register operation, and the minimum signal required to drive the circuitry may provide a more important size limitation than photolithography tolerances.

In these ways it is possible that output and regeneration circuitry may provide the more important fabrication, size, and performance limitations to charge-transfer shift register chips.

In devices which are designed for production, margins must be taken into considerations; this has not been done in this article, in part because the inputs to a margin analysis are very intimately associated with the input, output, and regeneration circuitry and with the details of device manufacture. Because of the importance of margin considerations, however, some general comments are in order. In the size range discussed in this paper, margin limitations and requirements will probably be very similar for the two registers, particularly since input, output, and regeneration circuitry are likely to be essentially the same. Differences will arise primarily because of the fact that the charge in the bucket brigade register is stored in diffused regions; but they will be modest unless metallization dimensions approach diffusion depths either through improved photolithography or through extended diffusion times.

In the final analysis, one of the most important comparative features of these devices will be fabrication yield. The major fabrication difference is associated with the addition of a diffusion at each storage site in the bucket-brigade register, but no detailed information concerning yield is presently available.

REFERENCES

1. Note, for example, Smits, F. M., "Charge Storage Semiconductor Devices," presented at the European Semiconductor Device Research Conference, Munich, March, 1971.
2. Waaben, S., Digest of Technical Papers, ISSCC, Philadelphia, 1970, p. 46.
3. Boyle, W. S., and Smith, G. E., "Charge Coupled Semiconductor Devices," B.S.T.J. *47*, No. 4 (April 1970), pp. 587–593.
4. Sangster, F. L. J., and Teer, K., IEEE Journal of Solid State Circuits, *SC-1*, p. 131 (1969).
5. Sangster, F. L. J., presented at the International Solid State Circuits Conference, Philadelphia, February 18–20, 1970.
6. Intel, Electronics, *43*, No. 12, p. 56 (June 8, 1970). Motorola, *Microelectronics Data Book* (Supplement 2 to 2nd Ed), (September 1970).
7. Janssen, J. M. L., Nature, *4291*, (149) January 26, 1952.
8. Hannan, W. J., Schanne, J. F., and Woywood, D. J., IEEE Trans. Milt. Elec., p. 246, July–October, 1965.
9. Berglund, C. N. and Boll, H. J., talk presented at the International Electron Device Meeting, Washington, D. C., October 1970.
10. Walden, R. H., Strain, R. J., and Krambeck, R. H., unpublished work.
11. Tompsett, M. F., Amelio, G. F., and Smith, G. E., Appl. Phys. Letters *17*, p. 111 (1970).
12. Krambeck, R. H., unpublished work.
13. Strain, R. J., presented at the International Electron Device Meeting, Washington, D. C., October 1970.
14. Berglund, C. N., IEEE Trans. Solid State Circuits, *SC-6*, p. 391 (1971).

15. Joyce, W. B., and Bertram, W. J., "Linearized Dispersion Relation and Green's Function for Discrete-Charge-Transfer Devices with Incomplete Transfer," B.S.T.J. *50*, No 6 (July–August 1971), pp. 1741–59.
16. Strain, R. J., and Schryer, N. L., "A Nonlinear Diffusion Analysis of Charge-Coupled-Device Transfer," B.S.T.J. *50*, No. 6 (July–August 1971), pp. 1721–40.
17. Thornber, K. K., IEEE Trans. Electron Devices, *ED-18*, pp. 941–50 (1971).
18. Strain, R. J., unpublished work.
19. Boyle, W. S., and Smith, G. E., Spectrum, *7*, (July 1971), p. 18.
20. Powell, R. J., private communication.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 51, No. 3, March, 1972
Printed in U.S.A.

Computer Modeling of Charge-Coupled Device Characteristics

By G. F. AMELIO

(Manuscript received August 16, 1971)

Properties of various charge-coupled device (CCD) configurations are investigated by means of a computer model. The model is based on a numerical solution of the Poisson equation for a unit cell of the CCD structure. The surface potential and the tangential surface electric field are obtained to an estimated accuracy of one percent and used to calculate transfer characteristics. On this basis various devices are compared as a function of both geometrical and electrical parameters. The geometrical parameters include oxide thickness, gap length, and electrode length. The electrical parameters include such things as doping density, fixed interface charge, and applied voltage. The influence of surface states is omitted from the treatment.

The principal results indicate that (i) for dimensions of practical interest electrode lengths of the same order as the interelectrode spacings are desirable, (ii) moderately thick oxides enhance the tangential surface electric fields and increase the effectiveness of the channel-stop diffusion, (iii) lightly doped p-substrates are more resistant to the formation of electrostatic barriers in the gaps and yield faster devices because p-type conductivity silicon has a higher minority carrier mobility, and (iv) fixed charge at the Si-SiO_2 interface can have a significant influence on the device characteristics.

It is concluded that proper choice of both geometrical and electrical parameters is essential in obtaining optimum CCD performance; however, for such an optimized design, the transfer efficiency is for all practical purposes not limited by electrostatic considerations and is probably limited only by surface states. Theoretical limits of transfer efficiency based on these calculations are reported.

I. INTRODUCTION

With the invention[1] and initial investigation[2,3] of charge-coupled devices (CCDs), it became apparent that although conventional one-

dimensional considerations can lead to qualitative and heuristic arguments concerning the device operation, the phenomenon of charge coupling laterally along a semiconductor surface is essentially two-dimensional in nature and any quantitative understanding of the effect must begin with this premise. The purpose, therefore, of the present work is to use the solutions of the two-dimensional Poisson–Boltzmann equation obtained within the bounds of certain reasonable approximations to try to infer the basis of operation for the structures reported in the literature as well as to study new structures and predict their anticipated performance. Such inferences are obviously dependent upon a large number of considerations. For purposes of tractability, however, attention will be restricted to what is believed to be the key attributes of the device. Central to our considerations, therefore, will be the static surface potential and electric field profiles. These will be obtained through numerical iteration of the two-dimensional equations. In addition, the analysis is limited to the simple three-phase devices as originally described. The question of charge motion is then treated for the case of a sufficiently small charge density so that the fields are not appreciably altered by its presence. The equations of motion are then solved by an explicit quasi-static method. The usefulness and validity of such an artifical approach to the dynamic behavior of the device is discussed.

II. THEORETICAL MODEL

2.1 *General*

In Fig. 1, one unit cell of a three-phase CCD is illustrated in cross section (not to scale). For purposes of the model, the structure is assumed infinitely long in the direction normal to the page. Each electrode has a finite thickness t and length l and all are identical. They are spaced from each other by a distance g and from the semiconductor surface by the insulator thickness d. The interelectrode distance is taken as w $(= l + g)$ and the unit cell length as L $(= 3w)$. Although not shown in Fig. 1, in some calculations it will be assumed that the region between adjacent electrodes is occupied by a dielectric as may well be the case in an actual device manifestation. It is hypothesized that there exists an immobile charge of density Q_{ss} in the insulator in a small region adjacent to the insulator-silicon interface. The insulator is assumed to be silicon dioxide. The silicon is postulated as uniformly acceptor doped and sufficiently thick so that punch-through is not a consideration.

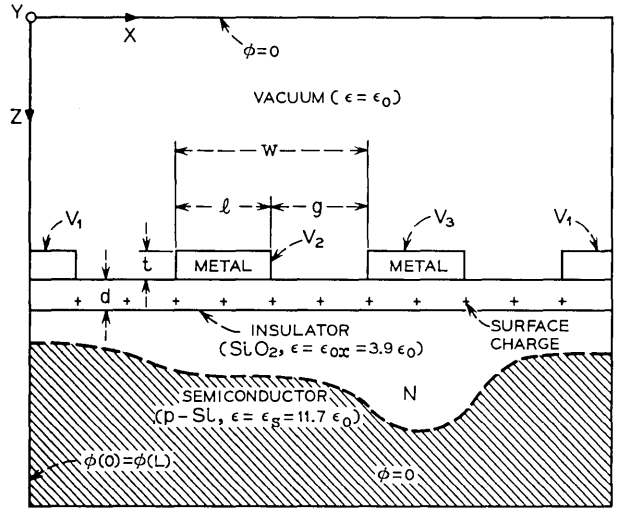


Fig. 1—Schematic cross section of a charge-coupled device unit cell showing the parameters and boundary conditions used in computer analysis.

Transfer of mobile charge at the interface by such a three-phase structure has been described qualitatively in the literature.[1–3] Briefly, electrons residing in an inversion layer are localized at the interface under the most positive electrode. This charge packet is moved to an adjacent area by increasing the potential of the neighboring electrode and decreasing the voltage on the initial electrode. Three electrodes per unit cell are required to insure directionality of charge transfer. Under ideal circumstances, the charge packet should move quickly, and in its entirety, from one area to the next when the proper voltages are applied to the electrodes. Assuming that the possibly important effects of trapping can be ignored, this charge motion is governed by the transport equation which in turn depends on the surface potential and electric field profiles. Our chief concern in the model, then, is to accurately obtain these profiles for specified electrode potentials and boundary conditions.

2.2 *The Dimensionless Poisson Equation*

The fundamental equation governing the electrostatic behavior in a semiconductor is, of course, the Poisson–Boltzmann equation

$$\nabla_x^2 \varphi = -\frac{\rho}{\epsilon} = \frac{q}{\epsilon_s}(N_A - N_D + n - p) \text{ : semiconductor,} \qquad (1)$$

where N_A is the ionized acceptor concentration, N_D is the ionized donor density, and n and p are the electron and hole densities respectively. Similarly, in the oxide the appropriate equation is

$$\nabla_{\mathbf{x}}^2\varphi = -\frac{Q_{ss}}{\epsilon_{ox}}\,\delta(x - x_s) : \text{insulator}, \tag{2}$$

where $\mathbf{x}_s$ is a vector defining the interface, and in the vacuum

$$\nabla_{\mathbf{x}}^2\varphi = 0 : \text{vacuum}. \tag{3}$$

Equations (1) through (3) all possess the same form and, for purposes of generality and computational convenience, can be simultaneously recast into dimensionless form by use of new variables defined as

$$\begin{aligned} v &= \frac{\varphi}{V_o} \\ \boldsymbol{\alpha} &= \frac{\mathbf{x}}{x_o} \end{aligned} \tag{4}$$

where V_o and x_o are dimensioned constants arbitrarily chosen and usually assuming a value suggested by the problem parameters. The basic P–B equation then becomes

$$\nabla_{\boldsymbol{\alpha}}^2 v = -\frac{x_o^2}{V_o}\frac{\rho}{\epsilon} = -r(v, \boldsymbol{\alpha}). \tag{5}$$

In like manner, the dimensionless electric field is given by

$$\boldsymbol{\mathcal{E}} = -\nabla_{\boldsymbol{\alpha}} v \tag{6}$$

and is related to the true electric field by

$$\mathbf{E} = \frac{V_o}{x_o}\,\boldsymbol{\mathcal{E}}. \tag{7}$$

Note that the solution v is unchanged if $r(v, \boldsymbol{\alpha})$ is invariant to changes in x_o , V_o , and ρ. Assuming v to be available, the dimensional results can then be easily extracted by use of the scaling equations (4) and (7). To facilitate description of the physical operation of the device, the results in Section III are presented in dimensional form and the equations above can be easily applied to rescale to other dimensions or densities of interest.

2.3 *Approximations and Boundary Conditions*

In practice it is perhaps probable that, with the statement of boundary conditions, the problem can be solved by some numerical technique

without approximation. In the present problem such an approach is unnecessarily difficult and likely to be uneconomical as well.

In the normal operation of CCDs the potentials on *all* electrodes are kept sufficiently positive at all times to insure that the surface is maintained in a state of depletion (or inversion). Experimentally, it has been determined that this leads to the more efficient charge transfer because majority carriers cannot then reach the surface and be trapped to subsequently recombine with minority carriers in a passing inversion packet. In the analysis, this means that the depletion boundary never terminates at the surface and hence it may be justifiable to treat the majority carrier density in the P–B equation in a less than rigorous way. This observation leads to a significant reduction in computational labor. In the depleted (but not inverted) volume near the surface of a device with a p-type substrate, n, p, and N_D are much less than N_A and hence equation (1) reduces to a simple form of the Poisson equation. Conversely, deep in the bulk the hole density is large, although the electron density is still very small and $N_A - p \approx 0$. Consequently, equation (1) reduces to Laplace's equation if, as is commonly assumed, N_D is small. Within a few Debye lengths of the depletion boundary, however, the hole density is rapidly changing from near zero to N_A and cannot be ignored. However, as stated above, in a CCD the depletion boundary does not come close to the surface in normal operation and hence the precise manner in which the potential changes in the depletion boundary region is relatively uninteresting. Thus we assume that the hole density p is zero up to the depletion boundary and is equal to N_A beyond. This step function treatment of the majority carrier density is the essence of the "depletion edge" approximation and is expected to lead to accurate estimates of the true surface potential.

This method of treating the majority carrier density is the essential approximation of this work and with its statement the boundary conditions can now be specified:

(*i*) At a distance "far above" the device surface, the potential is uniform and equal to zero.

(*ii*) Each electrode is at a specified uniform potential (V_1, V_2, or V_3).

(*iii*) Deep within the bulk the potential is zero. The potential is never allowed to go negative in the semiconductor.

(*iv*) The potential is translationally periodic; i.e., $\varphi(o, z) = \varphi(L, z)$ where L is the unit cell length and z is the direction normal to the surface.

(v) At the vacuum-dielectric boundary and at the dielectric-semiconductor interface, the tangential component of the electric field and the normal component of the displacement field are conserved.

The numerical formulation of the problem and the relevant error considerations are treated in the Appendix.

III. RESULTS

3.1 *Static Surface Potential and Electric Field Profiles*

The surface potential, and consequently the electric field, throughout one bit of a CCD is a function of the semiconductor and oxide parameters, the geometrical configuration of the electrodes, and, of course, the impressed voltages. In this section the influence of each of these variables on the static surface potential and electric field is illustrated by several examples. For convenience these are presented in terms of specific dimensions as opposed to normalized coordinates. It is to be understood, however, that these results can be scaled to other dimensions by means of the equations of the previous section.

3.1.1 *Geometrical Influences*

Variation of the electrode length, the interelectrode spacing, and the oxide thickness has a significant effect on the surface potential profile and the relationship between these parameters must be considered carefully in predicting CCD behavior. The effect of modifying the electrode length is considered first. In Figs. 2a and b the surface potential and electric field are presented for electrode lengths of 3, 6, and 12 μm. For the other variables an oxide thickness of 3000 Å, a gap length of 3 μm, a doping density of 5×10^{14} cm^{-3}, and a positive oxide charge at the interface of 10^{11} cm^{-2} are assumed. At the instant of time in question, the electrode potentials are $V_1 = 0$, $V_2 = 4$ V, and $V_3 = 16$ V and it is assumed no minority carriers are present (the alternative case is considered separately in the next section). In this figure it is clear that the smaller electrode-length-to-gap-length ratio yields a larger tangential electric field under the center of the second electrode. A similar conclusion applies to the electric field under the first electrode. Elsewhere, the fields are very much alike. From the point of view of charge transfer, the tangential electric field under the second electrode is of particular interest. From Fig. 2b the minimum values of this field are about 40 V/cm, 500 V/cm, and 2000 V/cm for the 12-μm, 6-μm, and 3-μm electrode geometries respectively. For electrodes less than

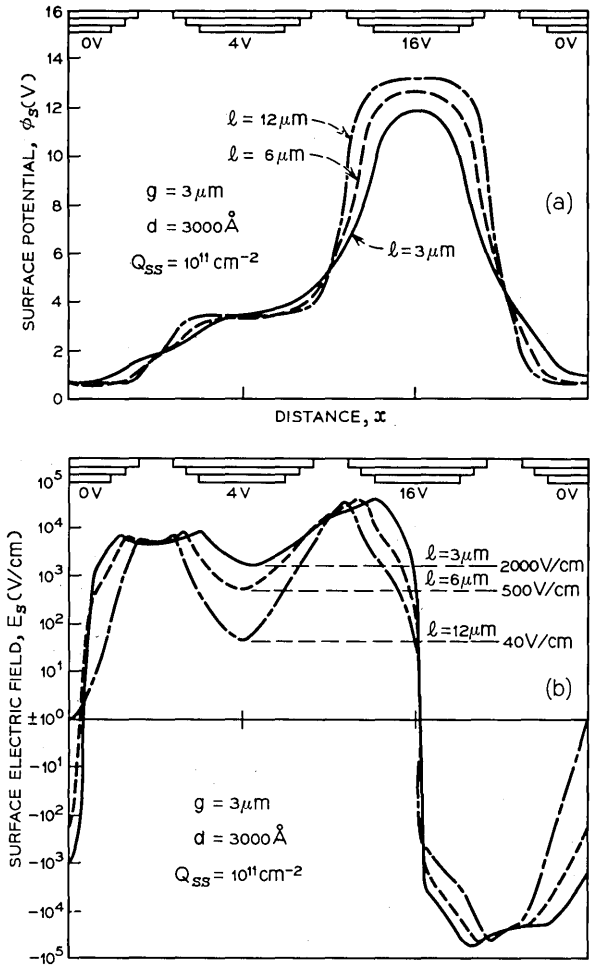


Fig. 2—Surface potential (a) and surface electric field (b) for three charge-coupled devices which differ in the electrode size but have the same gap length of 3 μm.

3 μm, the minimum field slowly continues to increase to about 10^4 V/cm; however, such electrode-length-to-gap-length ratios are not practical in terms of charge storage and efficient usage of silicon area. Moreover, for 2000 V/cm the region of nonlinear mobility is beginning and larger fields are of diminishing benefit.

Some effects of increasing the gaps can be inferred from the results of Fig. 2 and the scaling equations of the previous section without

additional computation. From equation (5), it follows from invariance of $r(v, \alpha)$ that as the dimensions are increased the voltage required to produce similar surface potential profiles increases quadratically. This will lead to similar performance since the surface electric field increases linearly if it is assumed that the mobility remains constant and the carrier drift velocity increases in proportion to the tangential field. Thus, direct scaling of the structures to larger dimensions rapidly results in inordinately large voltages. If the voltage is not increased, charge transfer may still occur but proceeds more slowly due to the decreasing fields and the increasing size of the unit cell. In addition, the surface potential in the large gap region becomes more dependent on the local charge density with the electrode voltages imparting a lesser influence. Thus, as will become clear in the next section, the control of such things as interface charge becomes more critical. For the case of constant voltages and constant unit cell length, the effect on the surface potential as the gaps are increased is shown in Fig. 3. Note the presence of an electrostatic barrier when the gap equals or exceeds 4 μm.

In a similar manner, it is expected that the electric fields will be altered by changing the oxide thickness. Using the same parameter values as in Fig. 2 for the structure with 3-μm electrodes, the surface electric field is shown in Fig. 4 for oxide thicknesses of 1500 Å, 3000 Å, and 5000 Å. For the 1500 Å oxide, the electric field dips to about 10^3 V/cm under the second electrode but peaks to almost 10^5 at the edge

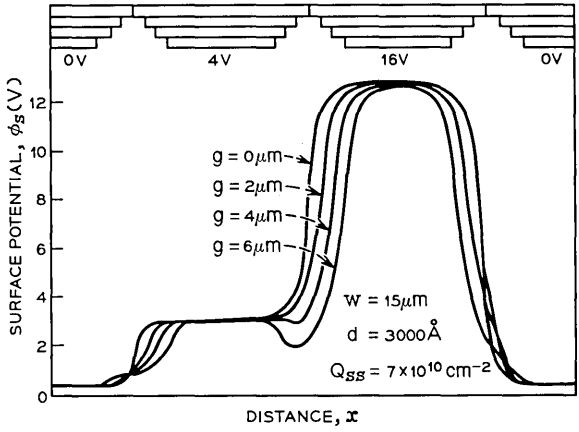


Fig. 3—The surface potential for charge-coupled devices with the same unit cell length but varying gaps and electrode sizes.

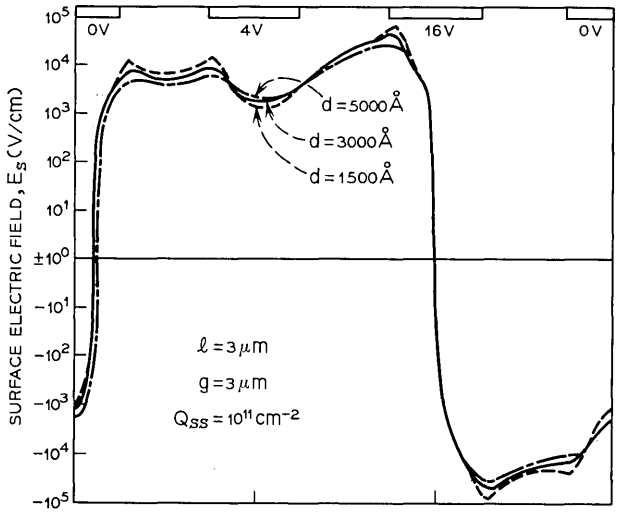


Fig. 4—Surface electric field for three charge-coupled devices which differ only in the oxide thickness.

of the third electrode. For the thicker oxides, the fields are more uniform with a minimum field of about 2×10^3 V/cm and a maximum of about 3×10^4 V/cm. For even thicker oxides, the tangential field begins to fall for a fixed set of gate voltages due to the rapidly decreasing surface potential. The better oxide thickness for this particular structure is approximately 3000 Å (or 0.1 of the electrode length) because it yields the greatest oxide capacitance per unit area (and hence charge handling capability) without sacrificing the magnitude of the electric field parallel to the surface. The above results imply that the moderately thick oxide enhances the ability of adjacent electrodes to couple effectively. In this context, two electrodes are said to be coupled when the tangential electric field at the surface is positive in the region between their centers. The larger the minimum field in this region, the more strongly the electrodes are said to be coupled.

The geometrical influences as presented in Figs. 2 and 4 indicate that in the range of practical interest (g = 1 to 10 μm), a preferred structure should possess an electrode-length-to-gap-length ratio of about unity and an oxide-thickness-to-electrode-length ratio of about 0.1, assuming all parameters are scaled appropriately. Furthermore, the smaller electrode lengths generally imply larger tangential fields, although the gains are probably marginal below 3 μm. These conclusions are in general consistent with intuitive notions about the effects of

capacitive fringing on which charge coupling is based. In an MOS capacitor, however, the magnitude of the fringing depends not only on the geometry but also on the charge density in the silicon and at the interface.

3.1.2 *Impurity Charge Density Influences*

In addition to the geometry of the device, the surface potential in a CCD is a function of the semiconductor impurity doping density and the (usually positive) ionic charge density at the interface. Ignoring for a moment the interface charge, the influence on the surface potential profile by modifying the substrate doping may be easily inferred from the scaling of equation (5). Thus, if the doping density is halved, the applied voltage may be also halved and a similar surface potential profile results in which the amplitude is reduced by a factor of two. Alternatively, the same voltages may be maintained but the structure enlarged by a factor of $\sqrt{2}$. In the former case the electric field also diminishes by a factor of two but the decrease in performance may be acceptable if the voltage is limited in the expected application. In the latter case the fields are undiminished but the increased bit length will proportionally reduce the charge transfer performance. Again, this may be acceptable if very fine features prove to be a difficulty. Unfortunately, in practice, materials more lightly doped than 5×10^{14} cm^{-3} are frequently nonuniform. Such nonuniformities could conceivably eliminate the anticipated benefits and result in a nonfunctioning device. Using more heavily doped material appears unprofitable unless one is capable of fabricating features smaller than 3 μm.

The influence of a charge density Q_{ss} at the interface is not so easily inferred. In Fig. 5a is plotted the surface potential for a CCD with 12-μm electrodes, 3-μm gaps, and differing Q_{ss} . The other parameters are the same as in Fig. 2. Note that for $Q_{ss} = 0$ there is a barrier-to-electron transfer in the gap between the second and third electrodes. As the magnitude of positive interface charge is increased, the barrier diminishes, eventually disappearing entirely. Thus, in an n-channel device, the normally occurring positive interface charge can be of substantial benefit. This is illustrated somewhat more dramatically in Fig. 5b. The parameters are as before except that the solid curve corresponds to a device with $Q_{ss} = 2 \times 10^{11}$ cm^{-2} and impressed voltages of 0, 4, and 16 volts. The dashed curve corresponds to a device with $Q_{ss} = 0$ and impressed voltages of 2.78, 6.78, and 18.78 volts. (The flatband voltage V_{FB} for 2×10^{11} cm^{-2} is 2.78 volts.) Quite clearly, the lack of Q_{ss} cannot be compensated for by adjusting the level of the

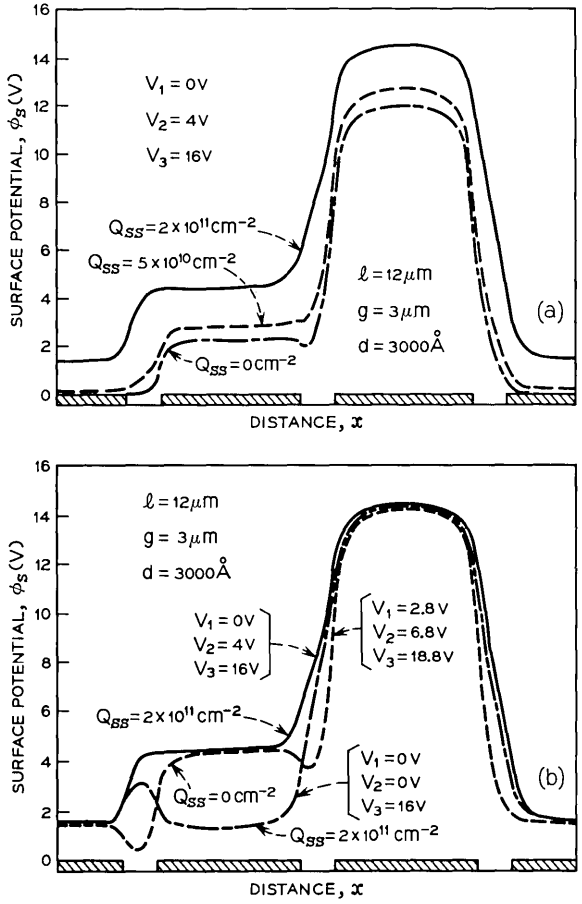


Fig. 5—Surface potential profiles for a 12-μm electrode charge-coupled device for differing values of interface charge and electrode voltage. (a) Electrode voltages held constant ($V_1 = 0$, $V_2 = 4$, $V_3 = 16$) but Q_{ss} varied. Note the "barrier" for $Q_{ss} = 0$. (b) The effect of interface charge on the surface potential. Solid line shows $\varphi_s(x)$ for $Q_{ss} = 2 \times 10^{11}$ cm^{-2}. Dashed curve shows $\varphi_s(x)$ for $Q_{ss} = 0$ and electrode voltages altered to compensate for flatband shift. Note the formation of large barriers in interelectrode gap. Dot-dash curve shows change of solid curve when second electrode is reduced to zero volts. Note the formation of a "pocket" in the gap between electrodes.

applied voltages without encountering large barriers to electron transport. Since the surface potential under the electrodes can be controlled by the applied voltage, the effect of this solid curve can be obtained by placing charge in the gaps only. However, there does not appear to be any advantage to this approach.

The presence of fixed charge at the interface can also cause some difficulties. The dot-dash curve in Fig. 5b illustrates the surface potential for a surface charge Q_{ss} of 2×10^{11} cm^{-2} and impressed voltages of 0, 0, and 16 volts. Note that in the gap between the first and second electrode there is a "pocket" in which electrons may be temporarily stored. Thus, for example, an electron briefly trapped in a surface state at the interface under the second electrode may, when emitted, proceed to the right or left. Those proceeding to the left are acquired by a trailing charge packet resulting in loss of transfer efficiency. This pocket may be eliminated by increasing the voltage on the first and second electrodes. The condition at which there is no pocket or barrier occurs when the gate voltage equals the surface potential. This can be easily calculated from the familiar equation relating the gate voltage to the surface potential

$$V_G - V_{FB} = \varphi_s + \alpha\varphi_s^{\frac{1}{2}} \tag{8}$$

where

$$\alpha = \frac{\sqrt{2\epsilon_s q N_A}}{C_o},$$

C_o = oxide capacitance.

If $V_G = \varphi_s$ it follows that when

$$V_1 = V_2 = V_G = \left(\frac{V_{FB}}{\alpha}\right)^2 \tag{9}$$

the surface potential is "level" and no barrier or pocket exists in the gap between the first and second electrode. Conversely, difficulties with small pockets may be avoided by allowing V_2 to return to its minimum potential slowly enough so that the charge has the maximum opportunity to proceed to the right. For larger pockets such an approach may not be adequate. Both pockets and barriers become larger for greater interelectrode spacing.

In addition to adjusting the doping, the electrode length, and the oxide thickness, the strength of coupling may be influenced by including a dielectric in the interelectrode spacing. If the surface potential and electric field for the 12-μm electrode, 3-μm gap configuration discussed in Fig. 5 is recalculated for the case when Q_{ss} is set equal to zero, the solid curves of Figs. 6a and b result. Adding insulating material in the gaps with the same dielectric constant as SiO_2 but changing nothing else yields the dotted curves. Although the barrier has not

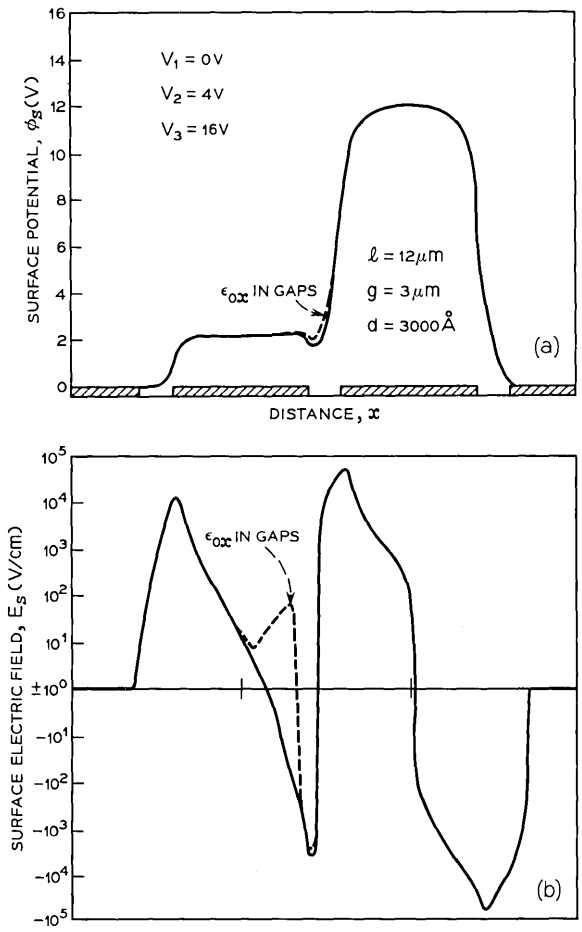


Fig. 6—Surface potential (a) and surface electric field (b) for two charge-coupled devices differing only in the presence of the dielectric in the region between electrodes.

vanished, it has definitely decreased and the fields indicate a more favorable coupling posture. In these calculations the electrode thickness is assumed to be 2000 Å and the observed results are believed attributable to more effective coupling of the electrode edges to the silicon surface. A material in the gaps with a higher dielectric constant would be even more effective in reducing or perhaps eliminating the barrier.

Before proceeding to the discussion of minority carrier transport in these structures, the surface potential in the direction normal to the channel is treated so that the effects of the channel-stop diffusion

may be considered. The appropriate model geometry is shown in Fig. 7a. The p^+ channel-stop diffusion is hypothesized 2 μm deep with an abrupt junction. Assuming a substrate doping of 5×10^{14} cm^{-3}, a channel-stop diffusion of 10^{18} cm^{-3}, and a gate voltage of 25 V, the surface potential and electric field shown in Fig. 7b and 7c result for oxide thicknesses of 2000 Å and 3000 Å. Note that near the channel-stop diffusion the maximum electric field approaches the breakdown field for silicon in both cases. Avalanche breakdown at the channel edges gives rise to unwanted currents thereby degrading device performance. In seeking methods to reduce the peak electric field it was observed that field profiles are relatively insensitive to parameter changes, with the exception of substrate doping and, of course, the applied voltage. Increasing the oxide thickness helps to the extent of reducing the surface potential for a given gate voltage and, therefore, has only a small effect on the maximum field as illustrated in Fig. 7c. The reduced potential, however, keeps the surface in the channel-stop region well below threshold thereby preventing large quantities of spurious dark current from coupling with the active region. Changing the doping of the channel-stop diffusion has only a very small effect on the maximum field, since by its very nature this diffusion must effectively hold the surface potential to small values in this region. It is concluded, then, that avalanching can most easily be avoided by limiting the voltage operation to values less than about 25 V, although an increase in the substrate doping density from 5×10^{14} cm^{-3} to larger values is also a possibility.

3.2 *Charge Transport*

When the surface potential and electric field profiles in a CCD are like those shown, for example, in Fig. 2, any electrons (minority carriers) under the second electrode will be rapidly transferred to the region under the third electrode. In this section we discuss the time behavior of this charge transfer and attempt to infer its dependence on surface potential, surface electric field, and minority carrier charge density.

3.2.1 *Analytic Transport Equation*

Before launching into a presentation of the computer predictions for the time dependence of charge transfer, it is useful to investigate the nature of the transport phenomenon by analytically studying the relevant equations. From such an analysis we can glean basic functional forms which shall prove useful in the interpretation of the computer results.

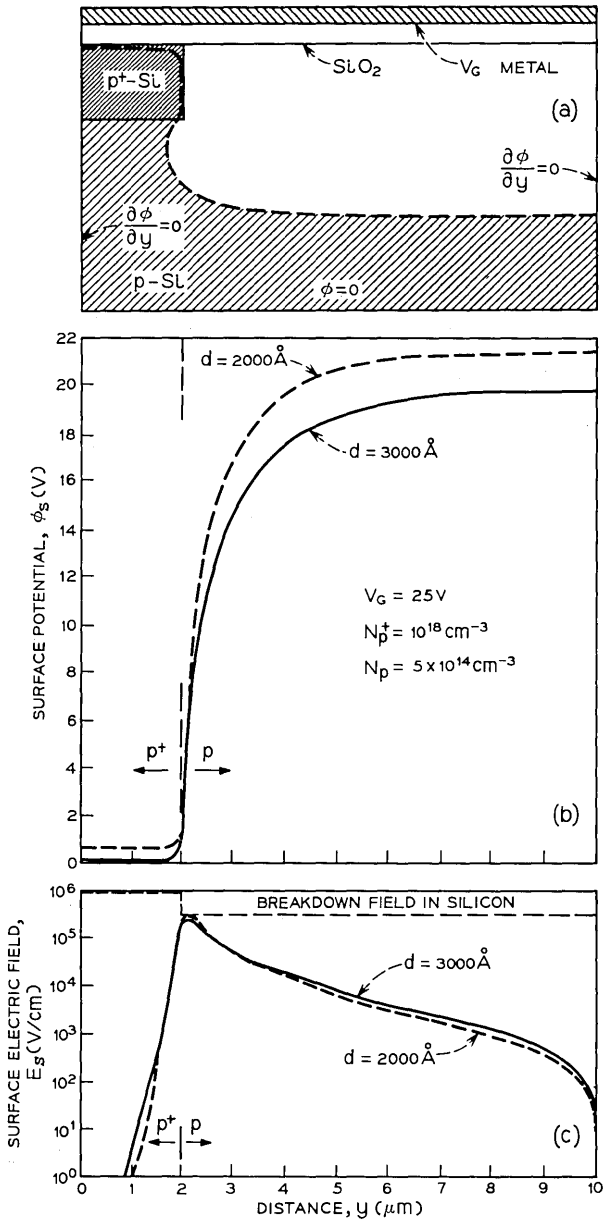


Fig. 7—Cross section (a) of charge-coupled device electrode normal to the channel direction showing channel-stop diffusion as used in computer analysis. Surface potential (b) and surface electric field (c) across channel-stop boundary for fixed electrode voltage and two oxide thicknesses.

In one dimension, the equation for current flow in a semiconductor is well known to be

$$j = \mu(E)\left(\rho E - \frac{kT}{q}\frac{\partial \rho}{\partial x}\right) \tag{10}$$

where $\mu(E)$ is the mobility, E is the surface electric field, and ρ is the minority carrier charge density. It is convenient to divide the surface electric field into two contributions

$$E = E_\rho + E_s$$

where E_s is the electric field arising from the device geometry as described earlier and E_ρ is the field contribution resulting from variations in the charge density of minority carriers. This latter contribution is easy to derive for the case of a large MIS capacitor. From equation (8), it follows almost immediately that the inversion charge density is given by

$$\rho = C_o(V - \varphi_s) - \sqrt{2qN\epsilon_s}\,\varphi_s^{\frac{1}{2}} \tag{11}$$

where $V = V_G - V_{FB}$. Differentiating (11) with respect to x results in

$$\begin{aligned}\frac{\partial \rho}{\partial x} &= -\left(C_o + \sqrt{\frac{qN\epsilon_s}{2\varphi_s}}\right)\frac{\partial \varphi_s}{\partial x} \\ &= -(C_o + C_d)E_\rho\end{aligned}$$

or

$$E_\rho = -\frac{1}{C_o + C_d}\frac{\partial \rho}{\partial x} \tag{12}$$

in which C_d is the depletion capacitance.

The thermal contribution in equation (10) may also be treated in terms of an effective field if it is divided by the charge density.

$$E_{\text{th}} = -\frac{kT}{q}\frac{1}{\rho}\frac{\partial \rho}{\partial x} = -\frac{kT}{q}\frac{\partial(\ln \rho)}{\partial x}. \tag{13}$$

In many instances the "thermal" field is small, initially, in comparison to E_ρ and E_s but can be important in determining the asymptotic behavior of the charge decay.

It is now possible to identify three physical cases as suggested by the form of equation (10). These occur when E_s is small and the electric field E is dominated by E_ρ, when E_s dominates and E_ρ is small, and when E_ρ and E_s are comparable. In each case E_{th} is assumed to be

small initially compared to the dominant field. In any of these cases it is possible to eliminate the current dependence by use of the continuity equation

$$\frac{\partial \rho}{\partial t} + \frac{\partial j}{\partial x} = 0. \tag{14}$$

The analytic transport equation is then

$$\frac{\partial \rho}{\partial t} = -\frac{\partial}{\partial x}\left\{\mu(E)\left[\rho E_s - \left(\frac{\rho}{C_o + C_d} + \frac{kT}{q}\right)\frac{\partial \rho}{\partial x}\right]\right\}. \tag{15}$$

In the case when the tangential fields resulting from geometrical factors are small ($E_\rho + E_{\text{th}} \gg E_s$), equation (15) reduces to a form of the diffusion equation

$$\frac{\partial \rho}{\partial t} = \frac{\partial}{\partial x}\left\{\left[\mu\left(\frac{\rho}{C_o + C_d} + \frac{kT}{q}\right)\right]\frac{\partial \rho}{\partial x}\right\}. \tag{16}$$

If the temperature-dependent term is negligible and μ and C_d are assumed constant, the variables easily separate* and the normalized time-dependent part, which we shall designate $\epsilon'(t)$, has the form

$$\epsilon'(t) = \frac{t_0}{t + t_o} \tag{17}$$

where t_o is a constant of integration determined by the boundary conditions.[4] Note that in this case $\epsilon'(t)$ goes to zero very slowly in time thereby implying that the transfer efficiency will suffer at high frequencies. However, when ρ becomes sufficiently small, the temperature-dependent term can no longer be ignored. In the limit when the thermal term dominates, equation (16) assumes the standard Fick's equation form. Following R. J. Strain and N. L. Schryer,[5] the gradient of the charge density may be assumed to be zero under the left edge of the electrode where the charge is initially located (charge transfers to the right). The solution is then an infinite series of the form

$$\rho = \sum_{n=1}^{\infty} a_n \exp\left\{-\left[\frac{(2n-1)\pi}{2(w + x'_o)}\right]^2 Dt\right\} \cos\left[\frac{(2n-1)\pi}{2}\frac{x'}{w + x'_o}\right] \tag{18}$$

where D is the diffusion coefficient ($= \mu\, kT/q$), x' is a new spatial variable with its origin at the left edge of the electrode, and x'_o is a constant determined by the boundary condition at $x' = L$. When the carriers reach $x' = L$, they move off at a constant velocity v_L deter-

* In this and the ensuing cases it is assumed that separable boundary conditions are applicable.

mined by the electric field E_L present at this edge (and in the inter-electrode gap region) through the relation $v_L = \mu E_L$. Thus the appropriate boundary condition is

$$\frac{j(x' = L)}{\rho(x' = L)} = -D \frac{\partial \ln \rho}{\partial x'}\bigg|_{x'=L} = v_L .$$

For nominal values of v_L , x_o' is much less than w and may be assumed negligible corresponding to the case $v_L = \infty$. The coefficients a_n are determined by the initial condition for the charge density distribution. If the distribution is such that it is rich in harmonics, the time decay of the charge density for small t is a large summation of exponentials not unlike an error function. After a sufficiently long time, however, the leading term will dominate and the decay will be purely exponential with the charge density distribution assuming a cosine form.

In the other extreme, when ρ is large, changes in C_d may not be ignorable because in this case the depletion region is almost totally collapsed. Hence, for a very short time interval near zero, $\epsilon'(t)$ will deviate from (17) but the intermediate and long-time behavior are unaffected. Thus, in the case when $E_\rho + E_{\text{th}} \gg E_s$, the charge density under an electrode decays in approximately a hyperbolic fashion initially and exponentially asymptotically.

In the case when the geometrically induced fields are large compared to the charge induced fields ($E_s \gg E_\rho + E_{\text{th}}$), equation (15) reduces to the field-aided form

$$\frac{\partial \rho}{\partial t} = -\frac{\partial}{\partial x} [\mu \rho E_s]. \tag{19}$$

As above, the variables separate and the normalized time-dependent part has the form

$$\epsilon'(t) = \exp\left\{-b\left(\frac{\mu E_s}{w}\right)t\right\} \tag{20}$$

where b is a constant of integration and it has been assumed that μ and E_s are constant. Note that in this case the form of the charge decay is similar to the asymptotic form of the previous case. It is clear, however, that (20) will approach zero more rapidly than (18) when

$$E_s > \frac{kT}{wq} \tag{21}$$

if we make the reasonable assumption that the constant b is comparable to $(\pi/2)^2$. If the electrode spacing w is taken to be 10 μm, inequality (21) becomes

$$E_s > 26 \text{ V/cm}. \tag{22}$$

This condition can be easily achieved by designing the CCD structure according to the geometrical considerations outlined earlier.

Finally, in the case when E_ρ and E_s are comparable, equation (15) does not simplify and the variables are not separable. Thus we can say nothing quantitative. Qualitatively, however, it is expected that $\epsilon'(t)$ in this case would be an admixture of the above results with an approximate hyperbolic-exponential form.

3.2.2 *Computer Analysis of Charge Transport*

A numerical approach to the question of charge transport in a CCD may proceed in one of two ways. Perhaps the most obvious is to return to equation (15) and solve it numerically using a standard technique for dealing with partial differential equations (such as the Crank–Nicholson scheme) subject to some reasonable boundary conditions. The required electric field profile is either given some approximate mathematical form or taken digitally from the results of Section 3.1. Alternatively, one may return to the current equation (10) and recast it in the form of an effective carrier velocity for each point along the surface. Then proceeding sequentially in time, the trajectory of each carrier is computed for a small time increment Δt during which the velocity is assumed constant. After Δt, a new velocity is computed for each carrier resulting in a new trajectory and so on. At time zero, all carriers are under electrode #2 and the time-dependent transfer inefficiency $\epsilon(t)$ is taken to be the number of carriers *not* under electrode #3 divided by the total number of carriers as a function of time. The transfer inefficiency $\epsilon(t)$ is analogous to the time-dependent solution $\epsilon'(t)$ discussed in the previous section but without the spatial dependence separated out.

Clearly, the latter approach, although conceptually more straightforward, is economical only for a relatively small number of minority carriers. This, however, is just the case we wish to examine. In order to treat the case of large carrier densities, it is necessary to include these carriers explicitly in the solution of the Poisson equation. Whereas this in itself is not difficult, it is necessary to recompute the solution at *every* time step. It is possible that the cost of such an analysis would not be prohibitively large; nonetheless, it is essential to ask what additional knowledge is obtained. Reflection on the discussion in the previous paragraph reveals that the only significant addition is the very short time behavior when electrostatic repulsive forces between

carriers are large. The transfer inefficiency function then falls quickly under a combined form of the exponential and hyperbolic behavior discussed earlier. This mode of behavior, however, lasts for a time very short compared to the asymptotic behavior, which has the principle influence in determining how much charge is left behind in a practical device application.

Based on these considerations much can be learned by investigating only small charge densities. This is done by means of the effective velocity approach and using nearest and next nearest neighbors to estimate local charge density. Thus, from (10), the distance Δx traveled by an electron in time Δt is

$$\Delta x \cong \mu(E)\left(E(x) - \frac{kT}{q}\frac{\partial \ln \rho}{\partial x}\right)\Delta t. \tag{23}$$

For the field-dependent mobility the empirical relation

$$\mu(E) \approx \frac{\mu_0 v_s}{\mu_o\left|E - \frac{kT}{q}\frac{\partial \ln \rho}{\partial x}\right| + v_s} \tag{24}$$

is used where v_s is the scattering limited velocity of the electrons in silicon and μ_o is the low field mobility. Using time increments of 0.002 ns and charge densities of a few times 10^9 cm^{-2} the charge motion is computed for the first nanosecond for each of the structures described by the results of Fig. 2. The time-dependent transfer inefficiency for each of these is plotted in Fig. 8. Note that in each trace the inefficiency initially stays at unity for 0.15 to 0.30 ns, during which time the first carriers move across the interelectrode gap. This is followed by a rapid fall, corresponding to the hyperbolic behavior indicated earlier, after which the decay lapses into an exponential. For each curve an exponential is matched to the appropriate data. The resulting time constants are

$$\tau = \begin{array}{l} 0.08 \text{ ns: } 3\text{-}\mu\text{m electrodes} \\ 0.5 \text{ ns: } 6\text{-}\mu\text{m electrodes} \\ 3.6 \text{ ns: } 12\text{-}\mu\text{m electrodes.} \end{array} \tag{25}$$

Note that if these time constants are used in equation (20) to calculate the electric field modified by the undetermined constant, bE_s, quantities are obtained which are consistent with the average fields in the transfer region as inferred from Fig. 2b if $b \approx 1$.

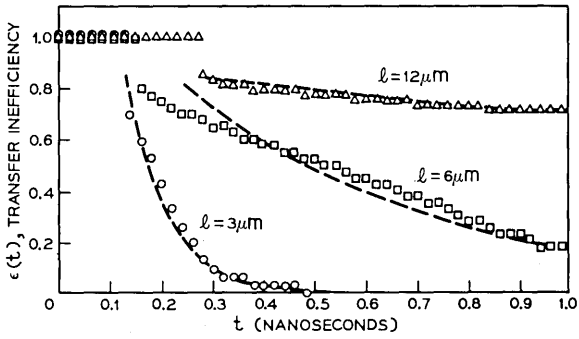


Fig. 8—Time-dependent transfer inefficiency for the same charge-coupled devices as in Fig. 2.

$$bE_s \approx \begin{cases} 10^4 \text{ V/cm: 3-}\mu\text{m electrodes} \\ 3 \times 10^3 \text{ V/cm: 6-}\mu\text{m electrodes} \\ 7 \times 10^2 \text{ V/cm: 12-}\mu\text{m electrodes.} \end{cases}$$

Thus, for times large compared to the characteristic time constant, equation (20) with $b = 1$ is useful in estimating transfer inefficiency if the average field is available. Based on such an exponential form, the transfer efficiency $\eta(= 1 - \epsilon)$ after N unit cells can be expressed

$$\eta(t) = (1 - \epsilon(t))^{3N} \tag{26}$$

where

$$\epsilon(t) = e^{-t/t_o}.$$

If ϵ is assumed small compared to unity, equation (26) becomes

$$\eta(t) \approx e^{-3N\epsilon} \approx 1 - 3N\epsilon \tag{27}$$

where the first approximation is valid if $3N\epsilon^2 \ll 1$ and the second is valid if $3N\epsilon \ll 1$. The appropriate time t used depends on the frequency of operation and t_o depends on the average electric field as described above. It is clear from (27) and previous discussion that an extremely high transfer efficiency is obtainable in principle in a properly designed charge-coupled device.

IV. CONCLUSIONS

The choice of geometrical and electric factors in the construction of a charge-coupled device influence the charge transfer performance of such structures. In particular, the following points are evident.

(*i*) For dimensions of practical interest (1–10 μm) electrode lengths comparable to interelectrode spacings are desirable.

(*ii*) Generally, moderately thick oxides ($\sim$3000 Å) enhance the tangential surface electric fields as well as increase the efficacy of the channel-stop diffusion without a serious loss in charge handling capability.

(*iii*) The presence of a dielectric in the interelectrode gaps enhances the coupling strength of adjacent electrodes. A slightly conductive material (a so-called resistive sea) which allows charge to move between electrodes over the oxide surface in the gap region can be made to accomplish the same objective.

(*iv*) Uniform, lightly doped substrates are generally more resistant to the formation of electrostatic barriers in the gaps than more heavily doped material.

(*v*) A p-type conductivity silicon substrate is preferable to n-type material because of the increased mobility and the favorable influence of the normally occurring positive surface charge, although at the anticipated performance levels suitable p-channel devices can be made if these restrictions are taken into consideration.

(*vi*) There is an optimum surface charge density depending on the oxide thickness and substrate doping density.

(*vii*) For a CCD which is nearly optimized with respect to the considerations discussed in this paper, the transfer efficiency is for all practical purposes not limited by electrostatic considerations.

Remaining to be investigated are the proposed two (or four) phase structures which call for either two levels of metallization or the addition of surface charge near the interface.[6] Furthermore, the effect of surface states on transfer efficiency has been completely ignored and for an appropriately designed CCD will likely represent the limiting factor. With the exposition of the basic electrostatic considerations described in the present work, these important questions can now be investigated on a solid physical foundation and future CCDs can be designed confidently.

APPENDIX

Numerical Formulation

Except for cases of simple geometry, the Poisson equation cannot be solved analytically and the use of numerical methods is necessary. As a result of the great interest in elliptic equations and in the Poisson

equation in particular, a number of satisfactory numerical techniques exist for finding the solution within a bounded region Ω with perimeter Γ on which the solution is specifiable.

The region Ω and the boundary conditions on Γ usually suggest themselves by considerations of symmetry. These were discussed earlier for the problem of a three-phase charge-coupled device and are illustrated in Fig. 1. The geometry of the region is seen to be a simple rectangle characterized by linear interface boundaries. In such cases, an explicit, finite difference method employing successive overrelaxation (SOR) is often satisfactory.[7] While more elegant methods such as the implicit or finite element techniques are equally valid, adequate accuracy and economy is possible with the simple SOR approach. With more exotic geometrical configurations or boundary conditions, the alternative methods may be more advantageous.

The region Ω is segmented into $p \times q$ rectangular cells and each node is labelled (i, j). Each rectangular cell is not necessarily the same size and the length of the sides are given as h_i and k_j . In this notation the potential at any mesh point is estimated by

$$\varphi^e_{i,j} = M^2\left\{\frac{h_{i-1}\varphi_{i+1,j} + h_i\varphi_{i-1,j}}{h_ih_{i-1}(h_i + h_{i-1})} + \frac{k_{j-1}\varphi_{i,j+1} + k_j\varphi_{i,j-1}}{k_jk_{j-1}(k_j + k_{j-1})}\right\} + \frac{\rho_{i,j}}{2\epsilon_{i,j}} \tag{28}$$

where

$$M^2 = \frac{h_ih_{i-1}k_jk_{j-1}}{h_ih_{i-1} + k_jk_{j-1}},$$

$\rho_{i,j}$ is the charge density, and only nearest neighbors have been used in estimating the derivatives (the so-called five point approximation). To compute the solution, the potential is successively calculated using (28) for each node in the field. After completing a step through the field, the process is continued repeatedly until by some means (discussed below) it is determined that the result is within a certain specified precision of the solution and the iteration process is ended. The solution is often approached more rapidly if, instead of (28) above, the potential at each node point is estimated by a linear combination of (28) with the previous estimate. Thus, after n iterations of the region Ω, the potential at any node is given by

$$\varphi^{(n)}_{i,j} = (1 - \alpha)\varphi^{(n-1)}_{i,j} + \alpha\varphi^e_{i,j} \tag{29}$$

where α is the relaxation factor and determines the rate of convergence. For any given problem there exists an α for which the rate of convergence is maximized provided the sequence in which the nodes are calculated possesses a consistent ordering designated "Property A."[4] If α is less than unity, the convergence is "underrelaxed"; and if α is greater than unity, the convergence is "overrelaxed." In practice the maximum rate of stable convergence is usually obtained when $\alpha > 1$ and hence the name successive-overrelaxation.

A detailed description of consistent orderings of the nodes possessing Property A is available in the literature. In the present work the common "checkerboard" array has been used. The nodes of region Ω are divided into two groups, A and B, and each iteration is composed of a step through the A group followed by a step through the B group. All the nodes in Ω are categorized into A or B in a manner exactly analogous to the familiar black and white squares of the chessboard. Note that when computing the potential at a node point in group A, only the potentials at node points of group B are used. All consistently ordered sequences possess this property.

When the region Ω is rectangular and the calculation sequence is consistently ordered, it is possible to analytically determine the optimum relaxation factor. It is given by the expression

$$\alpha_{\text{opt}} = 2\left(1 - \frac{\pi}{pq}\sqrt{\frac{p^2 + q^2}{2}}\right). \tag{30}$$

When Ω is not rectangular or circular, the optimum relaxation factor cannot be computed in closed form and methods exist for estimating it numerically.[6] As a result of the unusual boundary condition for the depletion edge, the region in the present problem is not truly rectangular. Nonetheless, the relaxation factor calculated on the basis of (30) is almost identical to the value estimated numerically, so long as the bottom of the rectangle is within a few mesh points of the lowest part of the depletion edge. This is fortuitous because, in general, the depletion boundary changes with every problem and recomputation of α_{opt} each time would be prohibitive.

Finally, it is possible to estimate the error of the numerical solution. Actually, there are two kinds of error to consider. One arises from the quantization of the space and the replacement of the differential equation by a finite difference equation. The other kind of error arises from the termination of the calculation after a finite number of iterations. The former is of order μ^2, where μ^2 is the dimensionless equivalent of M^2 defined in equation (28) and obtained by scaling the area of Ω to unity.

The quantization error is typically not a problem and can usually be made acceptably small without using a prohibitively large number of mesh sites. It is to be noted that, whereas the quantization error in the potential is $O(\mu^2)$, the error in the electric field is $O(\mu)$.

Truncation error can also be estimated. After many iterations (the exact number depending on the specific problem) the ratio of the errors e^n from one iteration to the next approaches a constant λ at any given mesh point

$$\frac{|e^{n+1}|}{|e^n|} \approx \lambda. \tag{31}$$

The constant λ is in turn related to the "residuals" d^n

$$\lambda^2 \approx \frac{d^{n+1}}{d^n} \tag{32}$$

where a residual is defined

$$d^n = \left[\frac{1}{pq}\sum_{i,j}^{p,q} (v_{i,j}^{(n)} - v_{i,j}^{(n-1)})^2\right]^{\frac{1}{2}}$$

and $v_{i,j}^{(n)}$ is the normalized potential [equation (4)]. Thus, in the iteration region where the convergence is monotonic the mean error after n iterations is given as

$$\langle e^n \rangle = \sqrt{d^n}. \tag{33}$$

In the CCD problem discussed in the text, the region Ω is divided into approximately 50×90 cells. In the direction parallel to the semiconductor surface, the nodes are equally spaced. In the direction normal to surface, the spacing is 500 Å in the vicinity of the interface but increases linearly deep into the bulk and quadratically into the vacuum. In this way the voltage change from one node to the next is always about the same. Such an approach yields the greatest economy with no real loss in accuracy. Based on (33), a mean error in the potential of less than 10^{-2} is typically obtained after less than 10^2 iterations. The constant λ defined in (31) generally is established after about 20–30 iterations. In any practical problem this represents a lower limit to the required number of iterations. Once λ is known, the number of additional iterations needed to reduce the error to an acceptable value can be determined.

REFERENCES

1. Boyle, W. S., and Smith, G. E., "Charge Coupled Semiconductor Devices," B.S.T.J., *49*, No. 4 (April 1970), pp. 587–593.

2. Amelio, G. F., Tompsett, M. F., and Smith, G. E., "Experimental Verification of the Charge Coupled Device Concept," B.S.T.J., *49*, No. 4 (April 1970), pp. 593–600.
3. Tompsett, M. F., Amelio, G. F., and Smith, G. E., "Charge Coupled 8-Bit Shift Register," Appl. Phys. Ltrs., *17*, 1970, pp. 111–115.
4. Engler, W. E., Tiemann, J. S., and Baertsch, R. D., "Surface Charge Transport in Silicon," Appl. Phys. Ltrs., *17*, 1970, p. 469.
5. Strain, R. J., and Schryer, N. L., unpublished work.
6. Krambeck, R. H., Walden, R. H., and Pickar, K. A., "Implanted Barrier Two-Phase Charge-Coupled Device," Appl. Phys. Ltrs. *19*, 1971, pp. 520–522.
7. Smith, G. D., *Numerical Solution of Partial Differential Equations,* New York: Oxford Univ. Press, 1965.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 51, No. 3, March, 1972
Printed in U.S.A.

A Study of Frequency Selective Fading for a Microwave Line-of-Sight Narrowband Radio Channel

By G. M. BABLER

(Manuscript received October 6, 1971)

The spectral and temporal characteristics of a narrowband radio channel subject to multipath fading were estimated from a detailed sampling of channel loss variations. The data base for this characterization was obtained during a 59-day experiment in which the amplitudes of a set of coherent tones spanning a band of 33.55 MHz and centered at 6034.2 MHz were continuously monitored. The more significant observations were:

(*i*) *For fade depths less than 30 dB the frequency selectivity is accurately described by linear and quadratic components (in frequency) of amplitude distortion. The derived statistical distributions of such distortion parameters exhibit slopes of a decade of decrease in probability of occurrence for each 10 dB increase in distortion, for bandwidths greater than 5 MHz.*

(*ii*) *For fade depths greater than 30 dB and bandwidths in excess of 5 MHz the amplitude distortion exceeds second order.*

(*iii*) *Maximum observed rates of change for the linear and quadratic distortion were 90 and 60 dB/second, respectively.*

I. INTRODUCTION

Unusual atmospheric conditions may support microwave propagation over two or more distinct paths between two line-of-sight radio antennas. The various signal paths will typically differ in their propagation delay, thereby permitting constructive and destructive interference at the receiving antenna. When the relative delay is significant with respect to the radio frequency signal period, the interference can be quite selective, with deep nulls in parts of the radio band, and smaller variations at adjacent frequencies. The variation in received power is called fading, and the variation in the amount of fading with radio frequency is known as frequency selective fading.

Experimental data on selective fading are difficult to obtain because of the long time periods (millions of seconds) of continuous measurement required to obtain a sufficient sampling of the fading events for a meaningful characterization. As a result, the experimental literature on the subject generally has been sparse, has been incomplete, or has tended to de-emphasize the magnitude of the propagational effects. W. T. Barnett[1] has explored these frequency effects for discrete microwave signals separated by 20 to 500 MHz, and in an important early study R. L. Kaylor[2] observed maximum amplitude deviations as large as several tens of decibels for the same bandwidths. Because of the ever-increasing emphasis on performance and the need for more efficient use of the microwave frequency spectrum, as well as for other reasons, an extensive experimental program was undertaken in 1970 to more precisely characterize the spectral and temporal effects of frequency selective fading within a narrowband microwave radio channel during a fading season. An additional objective was to obtain information which permits a better understanding of the complex physical processes of multipath fading.

The experiment described in this paper was of 59 days' duration and included the amplitude measurement of 62 uniformly spaced, coherent tones spanning 33.55 MHz at 6 GHz transmitted over a 26.4-mile radio path to Palmetto, Georgia. The tones were sampled five times per second and the results recorded whenever significant activity occurred.

In the following, we present the results of the data analysis and their interpretation. The organization of the report is (*i*) an experimental description, (*ii*) an overview of the fading activity observed, (*iii*) the fading behavior of single tones, (*iv*) a selectivity characterization of multiple tone activity by means of a three-tone amplitude difference technique, (*v*) an error analysis for estimating the higher-order selectivity structure, and (*vi*) observations describing an atypical fading period and comments on the temporal and spatial properties of selective fading.

II. SUMMARY

A listing of the findings follows:

(*i*) The fade depth distributions for individual tones during periods of typical multipath fading were essentially the same and had the expected power law of deep fades with a slope of a decade of probability per 10 dB change in fade depth.[3]

(*ii*) The degree of frequency selectivity was characterized by statistical distributions of linear and quadratic amplitude distortion constructed from the measured amplitudes (in dB) of three uniformly spaced tones spanning different bandwidths in the narrowband channel. The distributions exhibited slopes of a decade of probability per 10 dB change in observed distortion for bandwidths greater than 5 MHz and greater slopes for bandwidths less than 5 MHz. For a bandwidth of 20.35 MHz the linear and quadratic distortions exceeded 15 and 9 dB, respectively, for 10^{-5} of the observation time. As anticipated, the selectivity structure increased with fade depth.

(*iii*) An analysis was made to determine at what fade levels and bandwidths the linear and quadratic amplitude distortion characterization becomes inadequate. The measure of failing was the residual or difference between the observed amplitude-frequency characteristic and a three-term, power series analytic approximation constructed from the linear and quadratic distortions. This study indicated that, for fade depths greater than 30 dB, the amplitude distortion generally exceeded second order.

(*iv*) During deep selective fades the distortion evolves rapidly. The maximum rates of change observed for the linear and quadratic distortion for a bandwidth of 20.35 MHz were 90 and 60 dB/second, respectively.

(*v*) The statistical distributions of amplitude distortion derived from the loss in the narrowband channels as received on two vertically spaced antennas of different beamwidths were found to be nearly identical.

III. EXPERIMENTAL DESCRIPTION

3.1 *The Experiment*

A signal generator located at a microwave station in Atlanta (Fig. 1) generated a field of 62 coherent tones spaced 550 kHz apart and centered on 70.4 MHz. The envelope of the tone field was constant in time and flat to within ± 0.5 dB. The transmitter translated the 33.55-MHz-wide tone field to a 6.0342-GHz center frequency, and radiated the signal via a standard horn reflector antenna located 260 feet above the ground. After propagating 26.4 miles along a line-of-sight microwave path (Fig. 2), the test signal was received both by a horn reflector antenna (1.25 degrees half-power beamwidth) 330 feet above ground and a 6-foot dish (2 degrees) located 19 feet 3 inches below the horn, on a microwave relay tower located outside of Palmetto, Georgia.

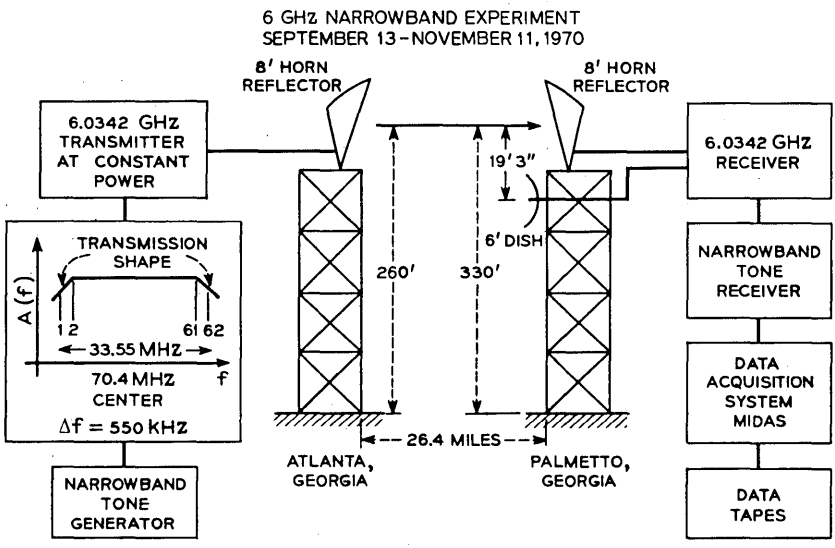


Fig. 1—Experimental layout, Atlanta to Palmetto, Georgia.

Both narrowband tone fields were translated back to 70.4 MHz and a specially designed tone receiver controlled by a multiple input data acquisition system (MIDAS) selected individual tones in a predetermined time sequence for measurement. The subset of tones measured is shown in Fig. 3; each tone was measured every 0.2 second. Using a common detector, the tone amplitudes were converted to dc voltages, quantized into 1-dB steps over a 55-dB range, and recorded on magnetic tape by MIDAS synchronously with timing and tone identification information. The recording rates were 1 sample per 30 seconds, 1 sample per 2 seconds, and 5 samples per second (normal, intermediate, and fast rates) depending on the current fading activity. The higher recording rates were initiated by monitoring the rates of change of tone amplitudes, a necessary step for maintaining a manageable data base for extended surveillance propagation studies.

3.2 *Calibration of Reference Levels and Determination of Transmission Shape*

The first step in analyzing fading data is to determine the nonfaded reference levels for proper calibration of the received tone amplitudes during fading conditions.

During normal transmission, when the lower atmosphere was well-behaved, the received tone field was nearly identical in shape to the

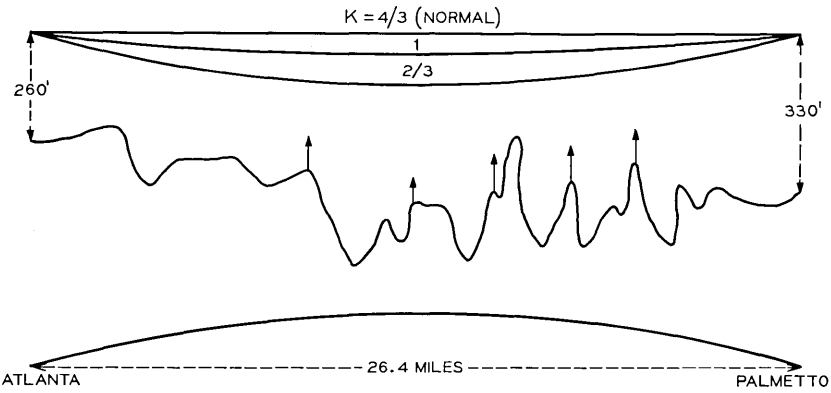


Fig. 2—Atlanta-Palmetto path profile and line-of-sight radio path for the normal (an equivalent earth radius factor $k = 4/3$) and the less frequent ($k = 1, 2/3$) atmospheric configurations. Clearance is adequate even for the extreme case of $k = 2/3$.

transmitted field and was attenuated by the free-space path loss due to the usual spherical radiation spreading. A precise determination of the free-space path loss as well as losses due to radio equipment was done in the usual manner[1] by using midday periods from 12 A.M. to 2 P.M., during which time there was generally no fading. Such quiet periods, selected by visual examination of time plots of received signal levels, were processed by computer to establish the average nonfaded received amplitude levels. These reference levels showed no long-term variations in excess of ± 1 dB throughout the measurement period. By using these results, calibration curves were constructed for the tone receiver detector.

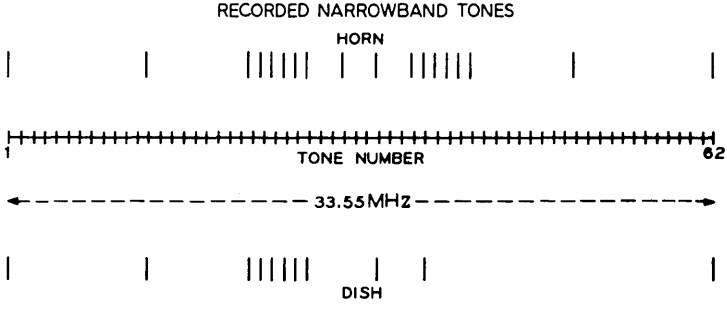


Fig. 3—Subset of 62 tones received on horn reflector and dish antenna measured and recorded serially.

Even during these favorable midday propagation periods, however, the received amplitudes of the tones were undergoing small time-varying random deviations called scintillation. By using specially implemented amplitude measuring capabilities of the tone receiver, amplitude data of all 62 tones were obtained for several 1-hour intervals at a constant rate of 5 measurements per second with an amplitude resolution of 1/8 dB. One such period showing the amplitude statistics of the 2 tones at the extremities of the narrowband is given in Fig. 4. The abscissa is the tone level in dB and the ordinate is the fraction of time that the received tone amplitude has exceeded the indicated level. The horizontal scale is logarithmic and the vertical scale is normal; thus, the scintillation amplitude statistics for both tones are log-normally distributed (normally distributed in dB). The standard deviation of 0.5 dB was found to be independent of tone frequency over the narrowband. The amplitude distribution of the microwave scintillation was found to agree with other experimental studies over the same radio path in previous years.[4]

In addition to the long-term variations in the reference values as well as the short-term scintillation, a third effect was the amplitude deviation across the narrowband arising from the experimental equipment. Figure 5 shows the average received tone level for all 62 tones during the same midday period shown in Fig. 4. As indicated, some tones (e.g., 16, 32) showed spurious amplitude effects. These effects were due to the modulation technique used in the tone generator and such tones were not used in the characterization of the selective fading. In addition, the amplitude depression of the extremities of the tone

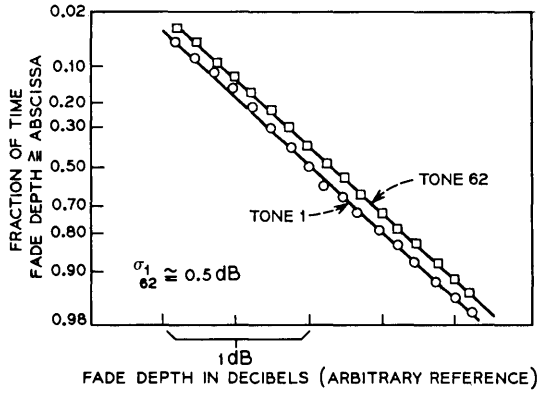


Fig. 4—Amplitude statistics of microwave scintillation for tones 1 and 62 received on horn reflector antenna during a 50-minute midday period on September 21, 1970.

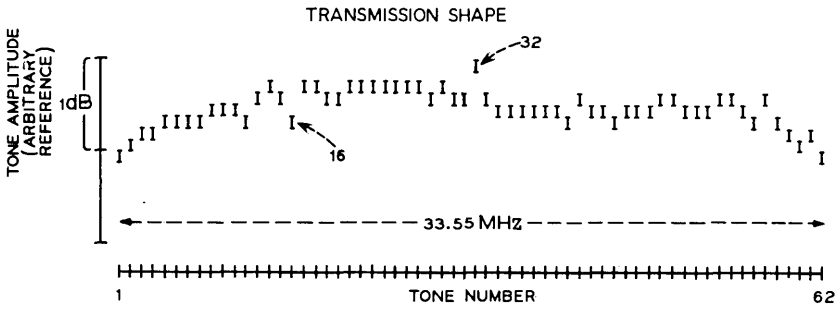


Fig. 5—Transmission shape of narrowband tone field received on horn reflector antenna.

field due to passive filtering were compensated for by suitable modification of the calibration curves for the outer tones.

The reference levels were calibrated in dB relative to midday normal by combining the calibration curves for the detector, the results of the statistical determination of the scintillation effects, and the transmission shape. The rms variation in the reference values for the tones used was estimated to be less than one dB.

IV. OBSERVED FADING ACTIVITY

The fading activity of the narrowband tone fields received by the horn reflector and dish antennas was continuously monitored from September 13 to November 11, 1970, and recorded for almost all of the 59 days (5.1×10^6 seconds). More than 1.8 billion measurements were made and the recorded data base was stored on 20 magnetic tapes. To condense this mass of raw data, a manual preselect was employed which included all periods with any fading in excess of 10 dB. The hour immediately preceding or succeeding each such event was also included. This process condensed the data base to 2 magnetic tapes containing data spanning 1.08×10^6 seconds (21.2 percent) of the total measurement period. This process resulted in a data base with 5 times the fading activity per unit time as compared to that originally measured, and contained all fades in excess of 10 dB. The effects and importance of this apparent increase in normalization of the fading activity will be discussed further in Section V.

An overview of the daily fading activity received on the horn antenna is shown in Fig. 6. The lower half of the figure shows the daily distribution of the fraction of the total time the received amplitude was faded 20 dB for tones 1, 25, 62. The upper half of the figure is for the

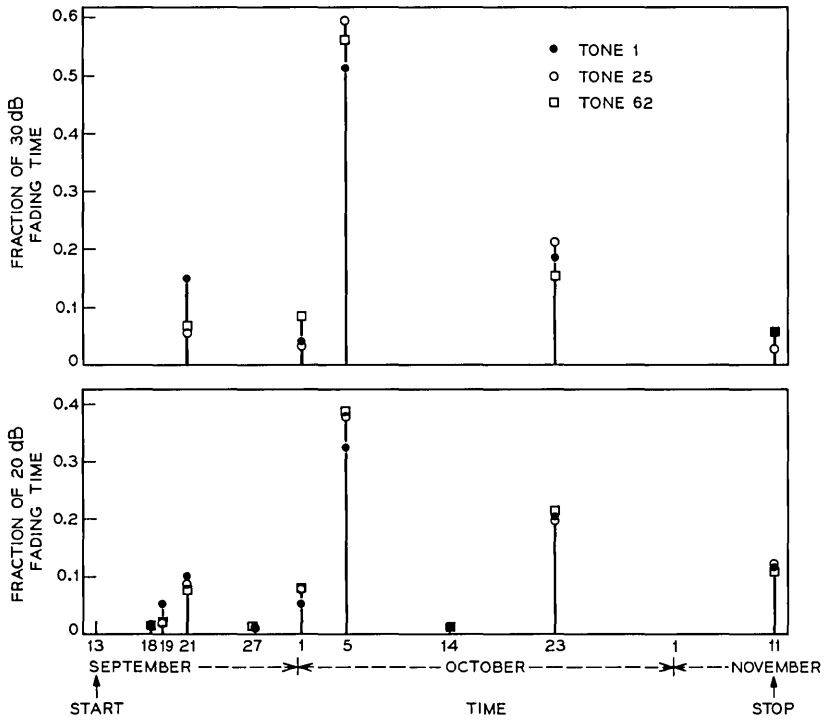


Fig. 6—Daily history of 20 and 30 dB fading activity for tones 1, 25, and 62 during the entire measurement period.

30-dB level. Tone 25 was used instead of tone 32 because of the spurious modulation effects in tone generation mentioned earlier. Several observations can be made from Fig. 6. First, as is typical of line-of-sight multipath fading, the deep fading at these levels occurred in unpredictable bursts; second, higher fading activity occurred in more concentrated bursts and for deeper fades (a rare event phenomena); and third, there was increasingly different fading activity of the closely spaced tones for deeper fades. This last observation is an indication that frequency selective fading is more pronounced for deeper fades as previously observed by Kaylor.[2]

V. FADING BEHAVIOR OF SINGLE TONES

The data were processed by computer to determine the total amount of time during which any tone was faded below a certain level. The resulting fade depth distributions for tones 1, 25, and 62 received on

the horn reflector antenna are given in Fig. 7. The abscissa is the fade depth in dB relative to midday normal and the ordinate is the fraction of the 1.08×10^6 seconds that the tones were faded the indicated amount. It is apparent that these amplitude statistics are not the same below 30 dB with tone 25 in the middle of the narrowband channel having suffered more fading. A preferential amount of fading in the channel, not expected from physical considerations of the overland microwave link, indicates the dominance of one particular fading event and the lack of events at other points in the channel and will now be examined closely.

In a recent study, S. H. Lin[3] has presented a general analysis of the statistical behavior of the envelope of a fading signal. His model indicates that for typical overland line-of-sight microwave links, designed to avoid a single dominant interfering multipath echo, the amplitude V of the received fading signal fades below a specified signal level L according to the following fade depth distribution

$$P(V \leqq L) \alpha L^2 \qquad L \leqq 0.1 \text{ (20 dB)} \qquad (1)$$

where fade depth $= -20 \log L$. For the special atypical case of a dominant (single) echo the model indicates the distribution

$$P(V \leqq L) \alpha L \qquad L \leqq 0.1 \text{ (20 dB)}. \qquad (2)$$

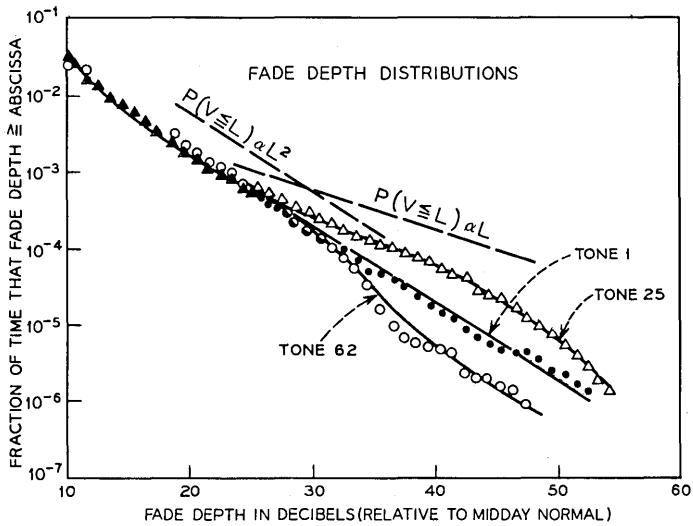


Fig. 7—Fade depth distributions for tones 1, 25, and 62 received on the horn reflector antenna.

Both distributions (1) and (2) are shown on Fig. 7 with slopes of a decade decrease in probability of occurrence per 10 and 20 dB, respectively.

For fade depths less than about 30 dB all three tones exhibit the L^2 dependence, indicating that the majority of the events at these levels were not due to dominant echo interference. For fade depths greater than 30 dB, tone 25 sustained significantly more deep fading and exhibited the L dependency. The indication was that at some time during the 5.1×10^6 seconds of the measurement period, the lower atmospheric structure was sufficiently stable and produced a dominant echo with appropriate amplitude and time delay (phase) resulting in substantial selective fading in the middle of the narrowband radio channel (tone 25). Studies of line-of-sight propagation at the same[3] and other microwave links[5,6] indicate that single dominant echo fading is not the typical multipath mechanism for overland fading. This period during dominant echo interference was assumed as atypical and was extracted from the total data base for special study (Section VIII). It is important to note, as indicated in Fig. 8, that the fade depth distributions for the same tones received on the dish antenna (located 19 feet 3 inches below the horn antenna) all exhibited the L^2 dependency. This reflects the strong spatial sensitivity of multipath fading and will be discussed further in Section VIII.

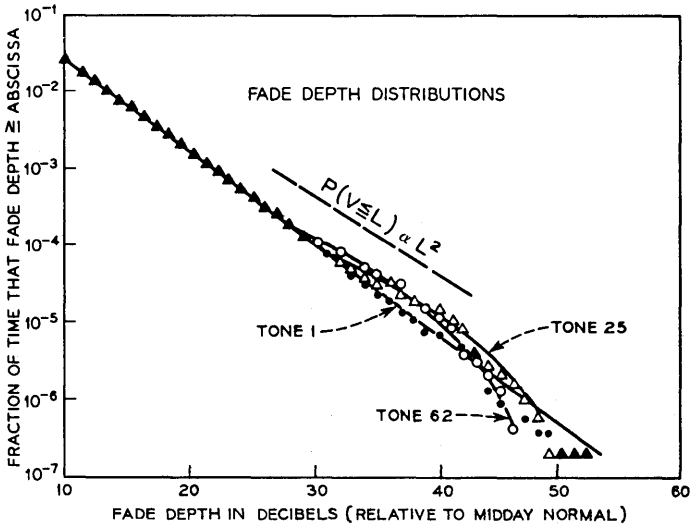


Fig. 8—Fade depth distributions for tones 1, 25, and 62 received on the dish antenna.

A day-by-day examination of the fading activity indicated that the atypical dominant echo event occurred on the morning of October 5 between 4 and 7:30 A.M.. With this time period removed from the data base the fade depth distribution for tone 25, shown in Fig. 9, exhibited more closely the L^2 dependency. The faded depth distribution of tone 62 remained unchanged because of the small sample size at the extremity of the narrowband.

Whereas the slope of the single-tone fade depth distribution curve gives information about the multipath fading process, the ordinate intercept gives a measure of the occurrence of such fading. Previous experience[1] at the same frequency and at approximately the same path length and path roughness indicates that the proportionality constant of equation (1) is about 0.5 which implies that $P(V \leqq 0.032) \cong 5 \times 10^{-4}$ at the 30-dB fade depth [$-20 \log (0.032) = 30$] as plotted in the figures. Figure 9 shows that the fading activity was below expectation at all fade depths; about 1/5 of that expected. This lack of observations was compensated for by using the condensed data base as discussed in Section IV. That is, we had 1/5 the number of fades, but we condensed the time base by five to compensate. This allows us to extrapolate these observations of frequency selective fading derived from a modest but, we believe, not atypical set

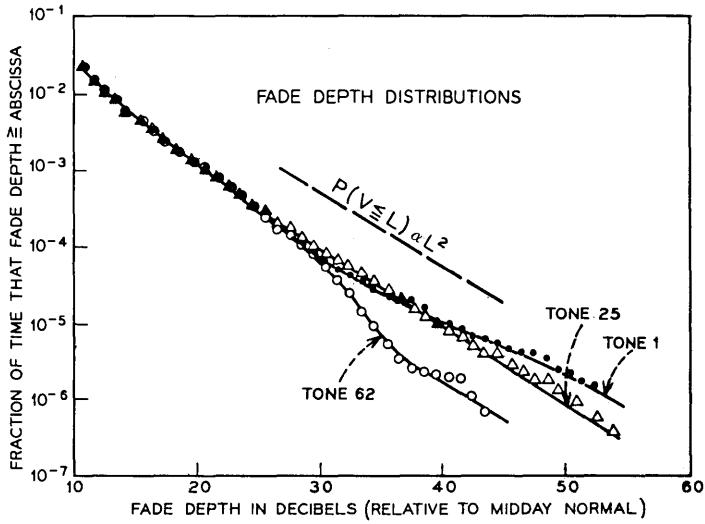


Fig. 9—Fade depth distributions for tones 1, 25, and 62 received on the horn reflector antenna for the entire 1970 measurement period excluding the fading activity of October 5.

of measurements during this late fall fading season to what is believed to be more representative of a typical summer fading month.

VI. SELECTIVITY CHARACTERIZATION OF MULTIPLE TONE ACTIVITY

To give a flavor of the variety and nature of the selective fading observed during the 1970 measurement period and to motivate the characterization of such events, two fading periods are presented in Figs. 10a and 10b. These time sequences of the channel transmission loss (displayed here as a continuous amplitude-frequency characteristic rather than discrete measurements) indicate the dynamics and dispersion of the channel. For figure compactness, the sequential scans of the tone field have arbitrary 0 dB reference; the absolute fade depth of tone 1 is indicated. The time between scans was 0.2 second. Event (a) was extremely rapid, and was preceded by 2 and followed by 6 days of near free-space propagation conditions. During the period shown in event (a), all tones in the narrowband channel were faded 20 dB. Event (b) was preceded by 9 days of free-space conditions and was followed by additional selective fading periods within the hour. For this early morning event the highly selective portion of the fade was superimposed on what appears to be a broader-band selective fade as indicated by the slopes of the amplitude characteristic before and after the events of greatest dispersion.

In event (a) the selectivity swept through the narrowband, suggesting the appearance of a very small echo with rapidly changing delay (phase), which maintained, at least approximately, a constant amplitude. In contrast, in event (b) the selectivity develops and dissipates in-band, suggesting a continual change in the relative amplitude of the echoes present. All of the highly selective fading events observed exhibited activity somewhere between these two extremes and the study of events such as these will undoubtedly sharpen our thinking about the mechanisms of multipath fading.

The choice of form of characterization of such selective fading events is bounded by the two constraints of generality and simplicity. Study of events like those shown in Fig. 10, and others, suggested that the amplitude-frequency characteristic most frequently exhibited either simple slopes or simple curvatures (or combinations of slopes and curvatures) and only for the deeper fades did higher-order structure (for example, cusps) occur. Therefore, the detailed variation across the amplitude-frequency characteristic was parameterized by monitoring the amplitudes of three symmetrically spaced reference tones as shown

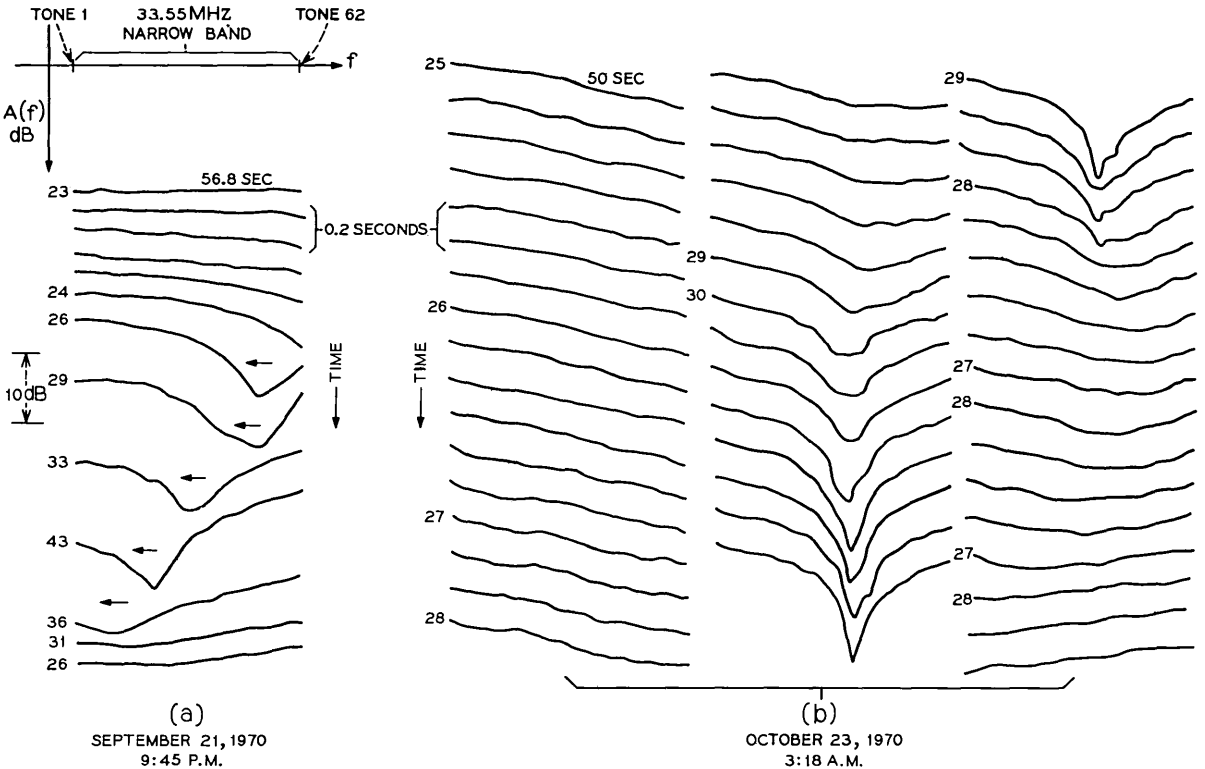


Fig. 10—Time sequential plots of selective faded amplitude-frequency characteristic. The fade depth of the lowest frequency tone (#1) is indicated when changes in amplitude occurred.

in Fig. 11. The three amplitudes, $A(f_1)$, $A(f_2)$, and $A(f_3)$, were used to construct the simple first difference

$$\Delta A = A(f_3) - A(f_1), \tag{3}$$

which is the linear amplitude distortion (slope) in dB across a narrow-band channel of width $f_3 - f_1 = 2\Delta f$, and the second difference

$$\frac{\Delta^2 A}{2} = \frac{A(f_1) + A(f_3)}{2} - A(f_2), \tag{4}$$

which is the quadratic amplitude distortion (curvature) in dB across the $2\Delta f$ channel. ΔA, $\Delta^2 A/2$, and $A(f_2)$ form a complete set; $A(f_2)$ is defined to be the fade depth of the event, $2\Delta f$ is the bandwidth parameter. For the example shown in Fig. 11, the fade depth is $A(f_2) = -34$ dB and for a bandwidth of $2\Delta f = 20.35$ MHz, the linear distortion is $\Delta A = 10$ dB and the quadratic distortion is $\Delta^2 A/2 = 6$ dB.

The 1.05×10^6 seconds of data were processed and the observed experimental distributions for $|\Delta A|$ and $|\Delta^2 A/2|$ were accumulated both for conditional and unconditional fade depths [conditioned on $A(f_2)$] as well as for seven different bandwidths ($2\Delta f = 1.65$ to 33.55 MHz). The results of this selectivity characterization are presented in the following sections in the order:

(*i*) Unconditional linear and quadratic distortion for a bandwidth of 20.35 MHz,

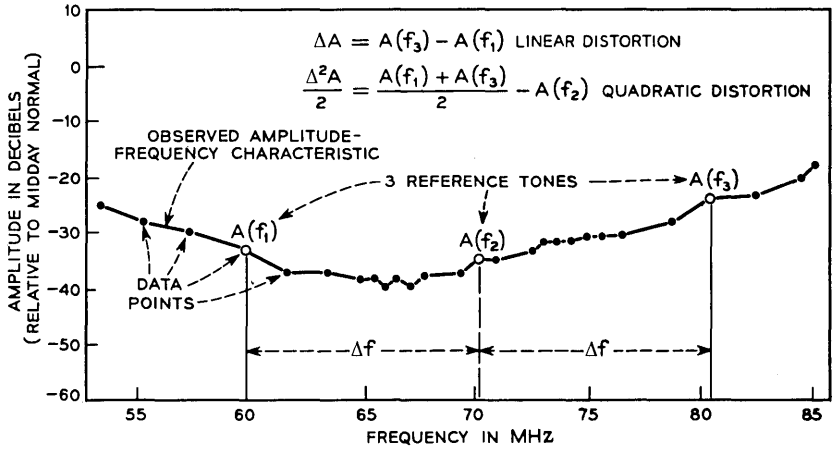


Fig. 11—Selectivity characterization of the observed amplitude-frequency characteristic.

(*ii*) Conditional linear and quadratic distortion for 20.35 MHz,
(*iii*) Unconditional distortion for other bandwidths,
(*iv*) Crossplots of the above results showing growth in selectivity structure with bandwidth and fade depth.

6.1 *20.35-MHz Bandwidth*

The unconditional linear and quadratic distortion distributions for 20.35 MHz are shown in Fig. 12. The abscissa is the amount of distortion in dB and the ordinate is the fraction of time (1.05×10^6 seconds) that the linear or quadratic distortion equaled or exceeded the corresponding abscissa value. For example, $\Delta A \geqq 15$ dB and $\Delta^2 A/2 \geqq 9.5$ dB for 10^{-5} of the time (about 10 seconds). As indicated for smaller fractions of the time even greater distortion was observed. The roll-off at the tails below 10^{-6} is the result of too few samples. The data points below 10^{-4} can be approximated by lines with slopes of a decade of probability of occurrence per 10 dB of distortion and will be discussed in Section 6.2.

It is appropriate to comment here on the effects of quantization. Since the measured amplitude of each tone has a maximum uncertainty of ± 0.5 dB, there results a maximum uncertainty of ± 1 dB in both

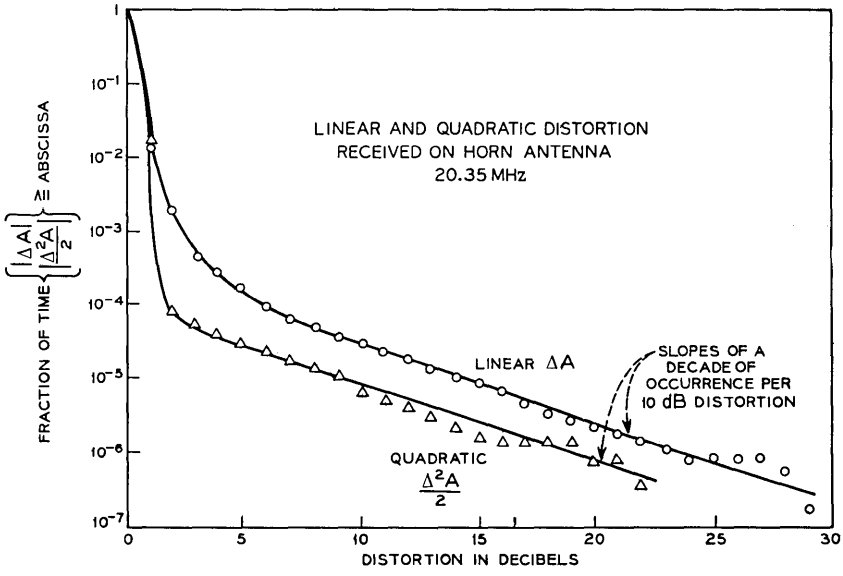


Fig. 12—Unconditional linear and quadratic distortion distributions for a bandwidth of 20.35 MHz.

linear and quadratic distortion parameters. Assuming that the error in each tone measurement is uniformly distributed about 0 dB (a reasonable assumption for a many-measurement statistic) results in an average error of 0 dB for the distortion parameters ΔA, and $\Delta^2 A/2$. Thus we conclude that the quantization effects are secondary and manifest themselves only at the distribution's tails where the samples are few. In addition, it is important to realize that for high-performance microwave systems the relevant fraction of time for propagation considerations of a single radio hop is in the range of 10^{-4} to 10^{-5}. For these fractions of time the data shows consistency.

The linear and quadratic distortion distributions for a 20.35-MHz bandwidth conditioned on the fade depth of the middle tone $[A(f_2)]$ are shown in Figs. 13 and 14, respectively. Note that the time base for each curve is different and is the total time the middle tone was faded the indicated amount. For example, for half of the 30-dB fading time (about 7.4 seconds) the linear distortion $\Delta A \geqq 3.4$ dB and the quadratic distortion $\Delta^2 A/2 \geqq 1.8$ dB. The point scatter of the linear distortion at the 25-dB level is difficult to explain; the irregular position of the 35-dB curve suggests that at (and below) these fade levels a simple linear distortion characterization parameter is an inadequate descriptor of the frequency selectivity in the narrowband channel. As anticipated, the distortion values increase with fade depth, particularly the quadratic distortion below the 30-dB level.

6.2 *Other Bandwidths*

By processing different sets of three reference tones, information regarding the bandwidth of the selective fading process was obtained.

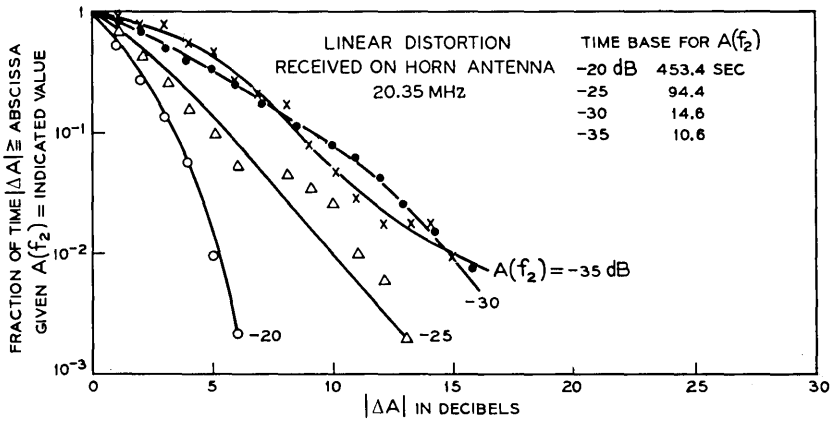


Fig. 13—Conditional linear distortion distributions for 20.35 MHz.

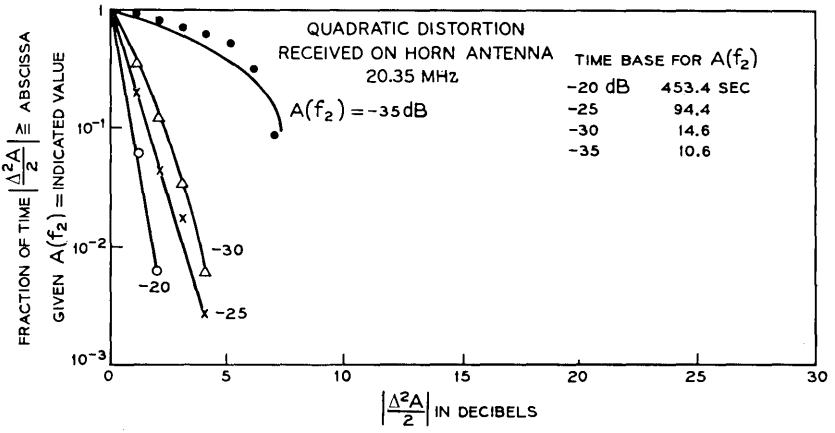


Fig. 14—Conditional quadratic distortion distributions for 20.35 MHz.

Figure 15 shows a summary of the unconditional linear distortion distributions for different bandwidths and Fig. 16 indicates the quadratic distributions. Remembering again that the data below 10^{-6} are less reliable because of fewer samples, the linear distortion distribution shows good consistency for the different bandwidths. The quadratic distributions show more scatter; the unusual behavior of the quadratic distortion for the 33.55-MHz case indicates the inadequacy of characterizing the selective fading amplitude characteristic by monitoring three reference tones spanning a bandwidth as great as 33.55 MHz. For bandwidths less than 5 MHz the tones are highly correlated and the amplitude distortion parameters are smaller.

Both the linear and quadratic distortion curves for the larger bandwidths exhibit slopes of a decade decrease of probability for 10 dB increase in distortion. This is a general result for deeply faded signals and their differences when such quantities are expressed in dB.

To examine more closely the increase in amplitude distortion with increased bandwidth and fade depth, several crossplots of the previously shown results were constructed. Figure 17, a crossplot of Figs. 15 and 16, shows the growth in distortion as a function of channel bandwidth for 10^{-5} fraction of the time. As indicated, the distortion increases rapidly with increasing bandwidth up to about 5 MHz; beyond 5 MHz the increase is less. For the quadratic distortion the transition point appears to occur around 20 MHz.

Crossplots of Figs. 13, 14, and others (not shown) for 10^{-2} fraction of the middle reference tone $[A(f_2)]$ fade time give the results shown in Fig. 18 for bandwidths of 10.45 and 20.35 MHz. The growth in

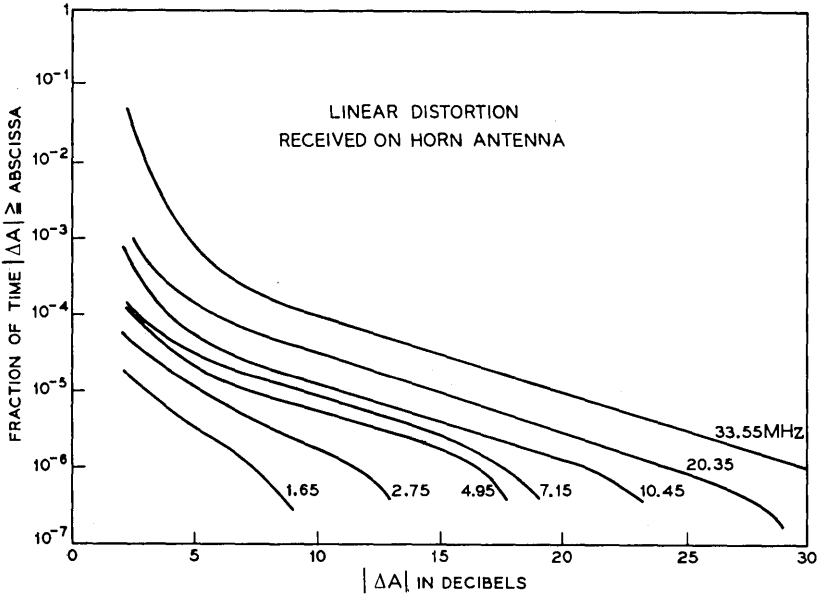


Fig. 15—Summary of unconditional linear distortion distributions for different bandwidths.

distortion is clearly a function of both fade depth and bandwidth. The dashed lines showing the estimated growth in the quadratic distortion below 30 dB were obtained by extrapolation of conditional distortion curves to the appropriate fade depths.

VII. AN ERROR ANALYSIS OF THE HIGHER-ORDER DISTORTION

As previously indicated, the selectivity structure increases with fade depth and monitoring the relative amplitudes of only three tones will not provide a complete description of the frequency selectivity for deep fades. In this section we discuss these higher-order selective effects and describe how a determination was made of the fade depth at which higher-order effects become significant.

This determination was made by making use of the multiple-tone amplitudes measured in the following way: For each amplitude-frequency characteristic measurement the three-term power series approximation was constructed

$$A(f) = A(f_2) + \frac{\Delta A}{2}\left(\frac{f - f_2}{\Delta f}\right) + \frac{\Delta^2 A}{2}\left(\frac{f - f_2}{\Delta f}\right)^2, \qquad (5)$$

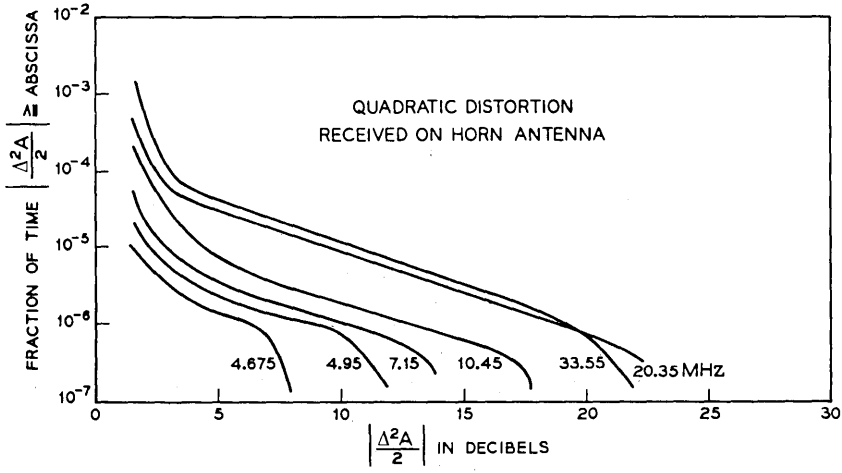


Fig. 16—Summary of unconditional quadratic distortion distributions for different bandwidths.

where ΔA and $\Delta^2 A/2$ are again the distortion parameters as defined in (3) and (4). Then by using the measured amplitudes, A^M, of tones (called here intertones) falling within the $2\Delta f$ bandwidth, the errors (deviations) between the measured values and the values calculated by the power series approximation (5) were found by

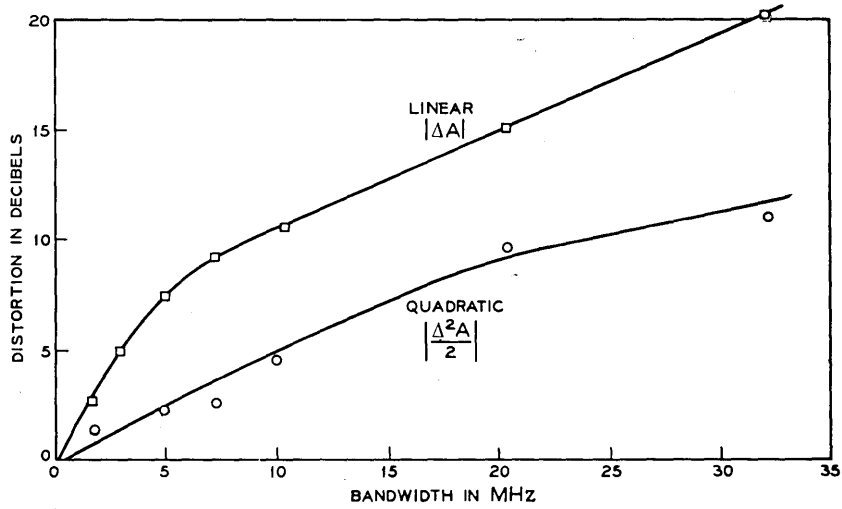


Fig. 17—Observed unconditional linear and quadratic distortions for 10^{-5} fraction of time for different bandwidths.

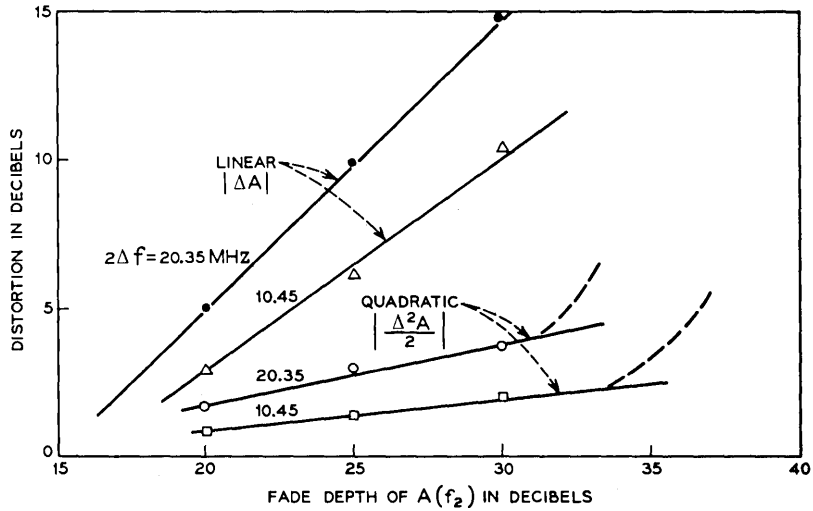


Fig. 18—Observed conditional linear and quadratic distortions for 10^{-2} fraction of the middle reference tone's fade time for 10.45 and 20.35 MHz.

$$E_i = E(f_i) = A^M(f_i) - A(f_i). \tag{6}$$

Figure 19 shows, for an instant in time, the observed selectivity, the power series approximation, and two of the intertone errors. The maximum of the absolute value of the set $\{E_i\}$ called MAX | ERROR | (maximum inband amplitude deviation) was monitored as a function

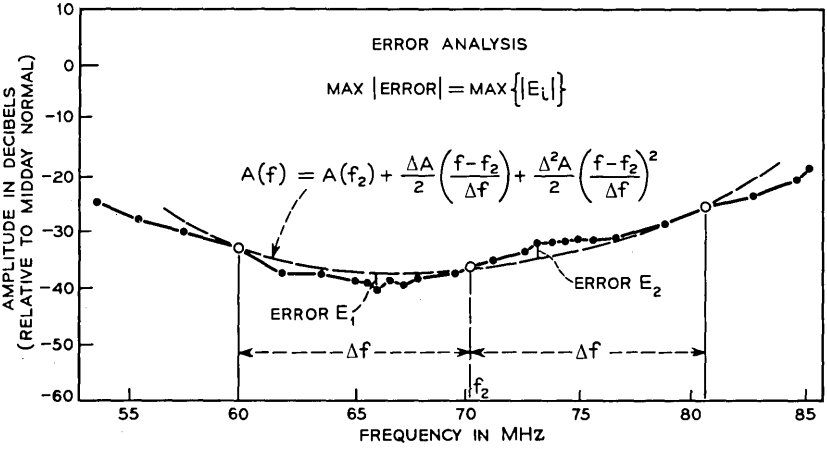


Fig. 19—Definition of errors as differences between the observed selectivity structure (solid line) and the power series approximation (dashed line).

of fade depth and tone spacing Δf. Figure 20 shows the experimental distribution of this MAX | ERROR | for a channel bandwidth of 20.35 MHz. Although the scatter of data is large (undoubtedly due to the higher-order selectivity structure), it is clear that for small fractions of the time the MAX | ERROR | can become significantly large (6 dB at 10^{-5} fraction time). The effect of decreasing the bandwidth is shown in Fig. 21. Figure 22, a crossplot of Fig. 21, indicates that, at 10^{-5} of the time, halving the bandwidth approximately halved the observed MAX | ERROR |. But, as indicated, the bandwidth would have to be limited to values considerably less than 5 MHz to limit MAX | ERROR | to less than 1 dB.

The large values of MAX | ERROR | occurred only during the deep selective fades, and we now explore these effects. The observed distribution of MAX | ERROR | for a bandwidth of 20.35 MHz conditioned on the fade depth of the middle reference tone is given in Fig. 23. Note again that the time base for each curve is different. As indicated, the MAX | ERROR | is a sensitive function of fade depth. Half of the time the middle reference tone $A(f_2)$ was faded 20, 25, or 30 dB, the MAX | ERROR | was less than 1 dB. But for the same

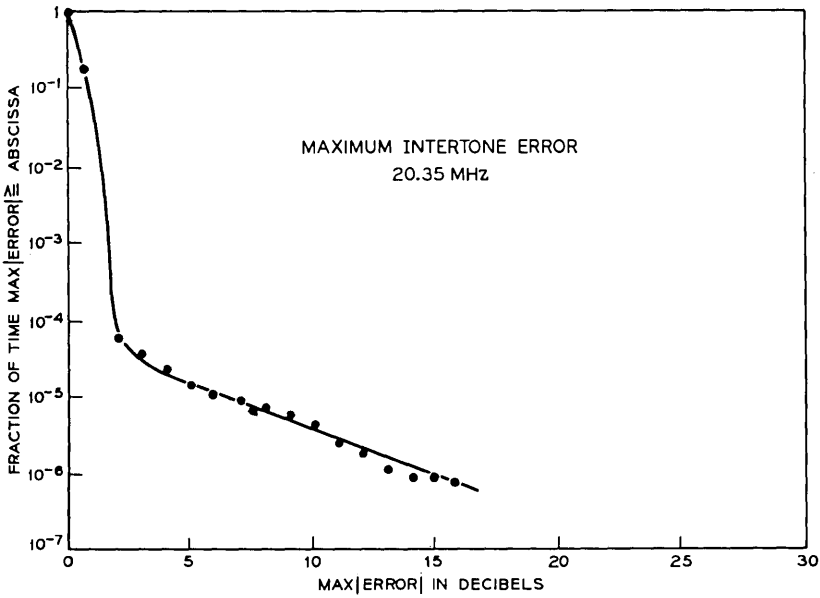


Fig. 20—Unconditional distribution of the MAX |ERROR| for a bandwidth of 20.35 MHz.

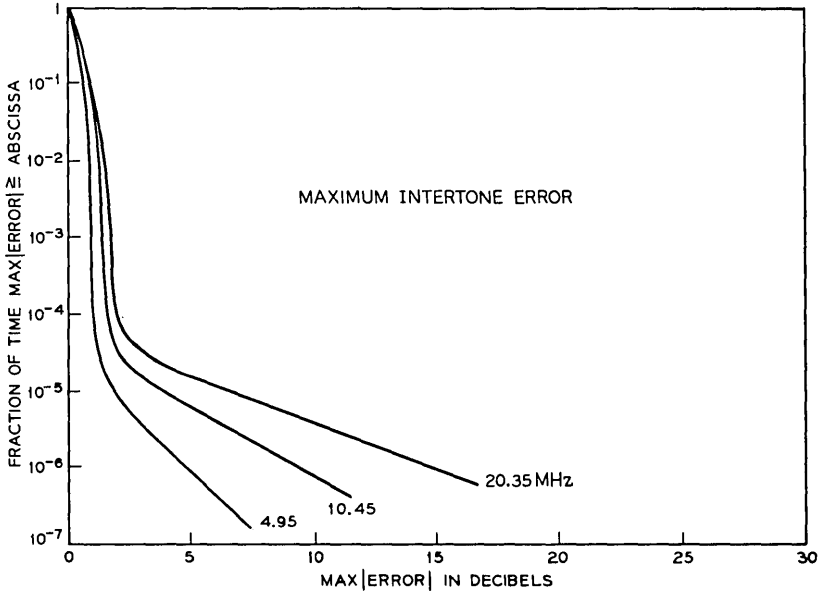


Fig. 21—Summary of unconditional distributions of MAX |ERROR| for additional bandwidths.

fraction of time when the middle tone was faded 35 dB, the MAX | ERROR | was greater than 2.5 dB. The same rapid growth in MAX | ERROR | below 30 dB was also observed for the other bandwidths as indicated in Fig. 24, which is for the 10^{-2} fraction of the middle reference tone's fade time. This figure indicates that for bandwidths from 5 to 20 MHz and fades not in excess of 30 dB, the three-term power series approximation constructed from the linear and quadratic distortion parameters approximates the actual channel to within 2 dB. Below 30 dB the selectivity structure is often greater than second order, and higher-order distortion parameters constructed from more reference tones are required to more precisely quantify the frequency selectivity of channel loss.

VIII. SPECIAL TOPICS

8.1 *Temporal Activity*

Although the physical process responsible for multipath fading may be of relative slow time scale, reflecting the huge inertia of the extended propagating medium, the faded microwave signal, which

is the vector sum of multiple echoes, may itself exhibit much greater rates of change. An example of this was indicated in Fig. 10a.

Figure 25 illustrates a particular fade, in which the dynamics of the linear distortion, quadratic distortion, and MAX | ERROR | for a 20.35-MHz bandwidth are displayed. Note that as the selectivity develops the distortion grows to significant values. The three-tone power series approximation initially matches the channel frequency selectivity character to within 1 dB. During the 58th second, however, the selectivity structure is of such high order (again see Fig. 10a) that the power series approximation fails substantially to match the amplitude structure, resulting in MAX | ERROR |s in excess of 6 dB. The 30-dB fading period for the middle reference tone $A(f_2)$ is shown and further supports the earlier observation of the inadequacy of using only three tones to monitor selective fading over bandwidths of a few tens of MHz for fades below 30 dB. The rates of change $(d/dt)(\Delta A) \sim 90$ dB/second and $(d/dt)(\Delta^2 A/2) \sim 60$ dB/second, which existed for only a fraction of a second, were the maximum observed during the 1970 measurement period.

8.2 *Spatial Activity*

The fading activity experienced by a dish antenna located 19 feet 3 inches below the horn antenna was also monitored throughout the measurement period. The results, less complete than those for the

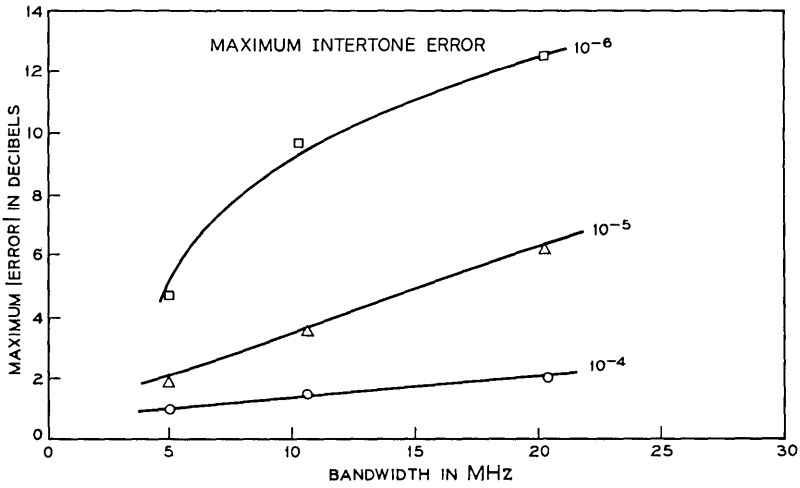


Fig. 22—A crossplot of Fig. 21 showing the observed MAX |ERROR| as a function of bandwidth for 10^{-4}, 10^{-5}, and 10^{-6} fractions of the time.

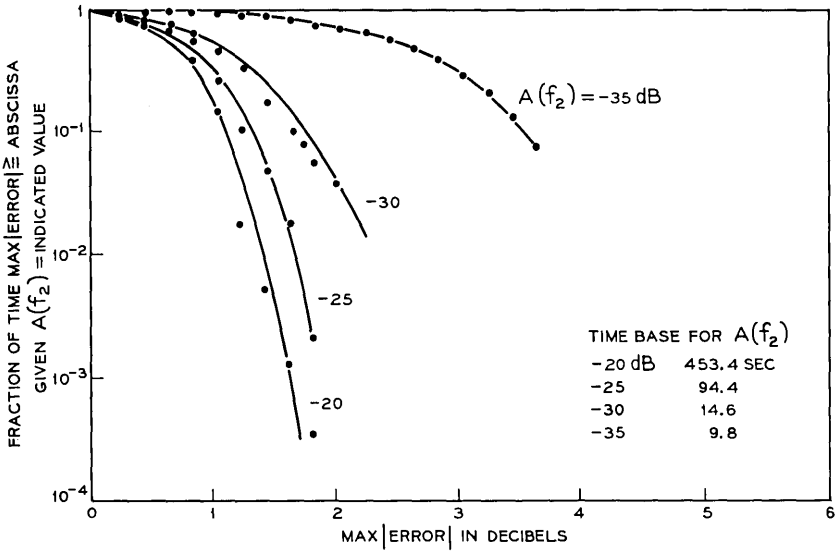


Fig. 23—Conditional distributions of MAX |ERROR| for a bandwidth of 20.35 MHz and different fade depths.

horn reflector because of fewer tones recorded (see Fig. 3), are shown in Fig. 26. These distributions also have slopes of a decade of probability per 10 dB of distortion for the main portion of the curves. The roll-off at the tails is a small samples effect. Comparing the selectivity characterization plots for the horn in Figs. 15 and 16 to the plots for the dish in Fig. 26, we see the same amount of selectivity structure. The

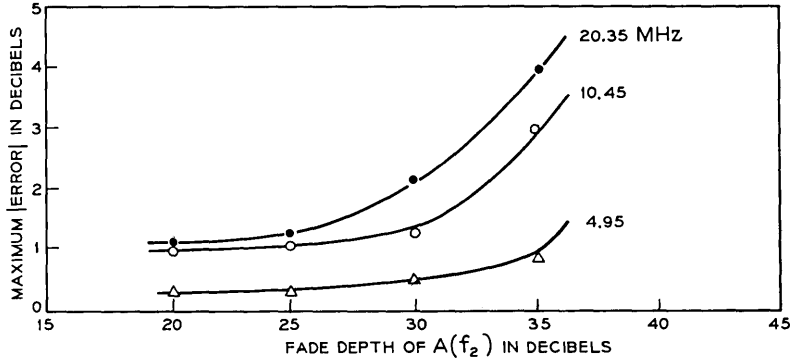


Fig. 24—Observed MAX |ERROR| for 10^{-2} fraction of the middle reference tone's fade time as a function of fade depth and bandwidth.

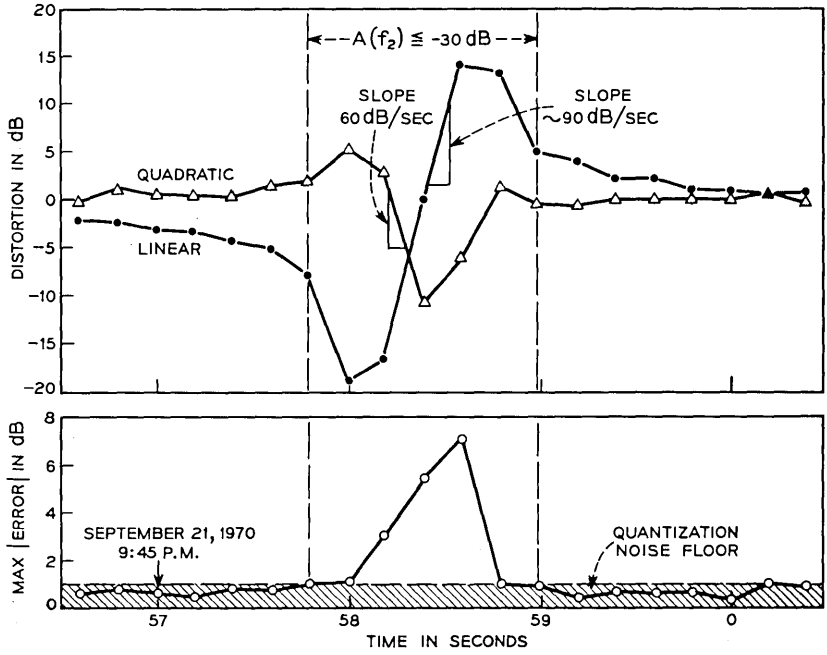


Fig. 25—Dynamics of the linear and quadratic distortion and the MAX |ERROR| for a bandwidth of 20.35 MHz. This is the event shown in Fig. 10a.

conclusion is that the selective fading structure as observed in a narrowband radio channel at these frequencies is insensitive to the beamwidth or to the precise vertical location of the antenna.

Although the joint activity of selective fading in the narrowband channels received on both antennas was not directly processed, examination of all significant fading periods showed very few simultaneous events with appreciable selectivity.

8.3 *The Atypical Event of October 5*

For completeness, the linear and quadratic distortion distributions were accumulated for the entire measurement period including the assumed atypical event of October 5. The distributions for a 20.35-MHz bandwidth are shown in Fig. 27. The distortion distributions for this inclusive period have the same slope as before (see Fig. 12), indicating that the frequency selective structure which occurred during this unusually long dominant echo period was similar to the selective structure occurring during the remainder of the experiment.

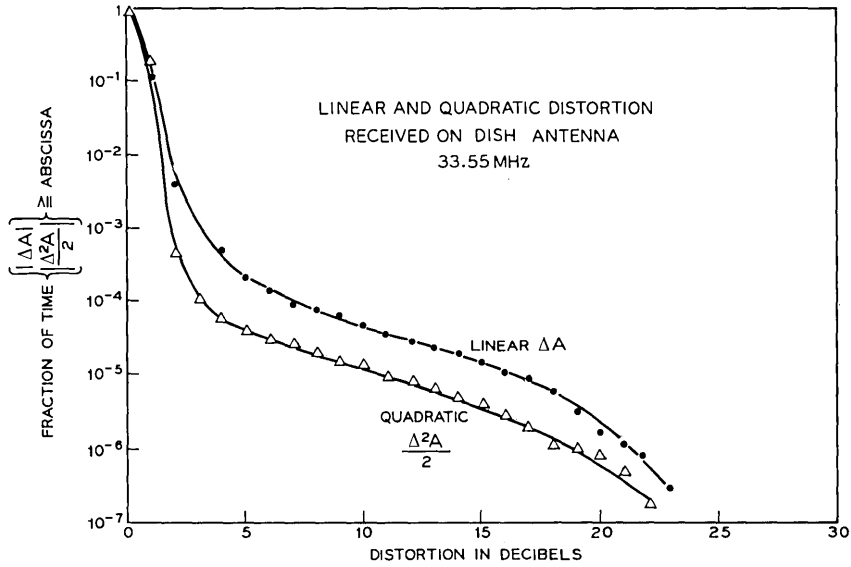


Fig. 26—Unconditional distributions of the linear and quadratic distortion received on the dish for a bandwidth of 33.55 MHz.

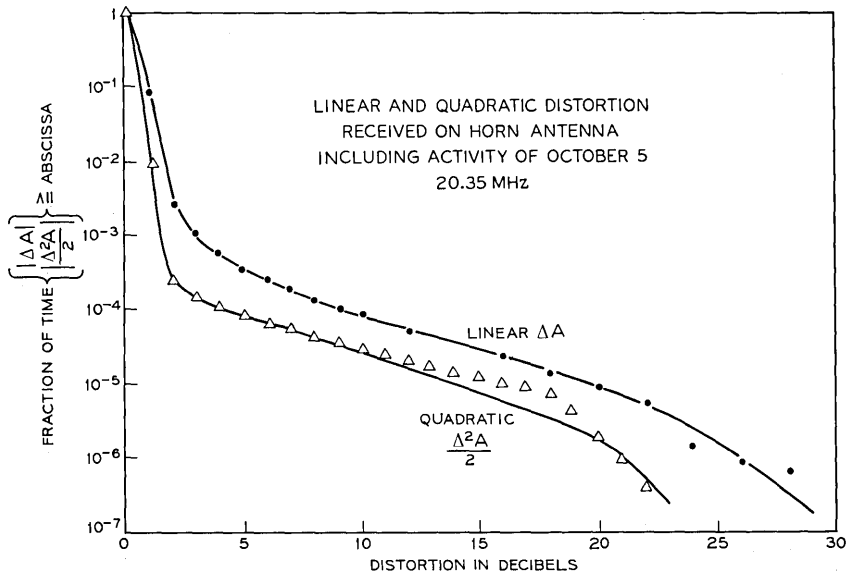


Fig. 27—Unconditional linear and quadratic distortion distributions for a bandwidth of 20.35 MHz and for the entire 1970 measurement period including the activity of October 5.

IX. ACKNOWLEDGMENTS

The narrowband data come from an experiment to which many colleagues at Bell Laboratories have contributed. Indebtedness is extended to W. T. Barnett who was instrumental in the conception and supervision of the experiment; to H. J. Bergmann and L. J. Morris who installed and calibrated the radio link; especially to G. A. Zimmerman who created the multiple-tone generator, multiple-tone receiver, and the MIDAS; and to C. H. Menzel who provided the necessary computer programs required for interfacing the author's analyses programs to the narrowband data.

REFERENCES

1. Barnett, W. T., "Microwave Line-of-Sight Propagation With and Without Frequency Diversity," B.S.T.J., *49*, No. 8 (October 1970), pp. 1827–1871.
2. Kaylor, R. L., "A Statistical Study of Selective Fading of Super High Frequency Radio Signals," B.S.T.J., *32*, No. 5 (September 1953), pp. 1187–1202.
3. Lin, S. H., "Statistical Behavior of a Fading Signal," B.S.T.J., *50*, No. 10 (December 1971), pp. 3211–3270.
4. Babler, G. M., "Scintillation Effects at 4 and 6 GHz on a Line-of-Sight Microwave Link," IEEE Trans. Ant. and Prop., *AP-19*, No. 4 (July 1971), pp. 574–575.
5. Crawford, A. B., and Jakes, W. C., "Selective Fading of Microwaves," B.S.T.J., *31*, No. 1 (January 1952), pp. 68–90.
6. DeLange, O. E., "Propagation Studies at Microwave Frequencies by Means of Very Short Pulses," B.S.T.J., *31*, No. 1 (January 1952), pp. 91–103.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 51, No. 3, March, 1972
Printed in U.S.A.

Capacitances of a Shielded Balanced-Pair Transmission Line

By C. M. MILLER

(Manuscript received October 4, 1971)

The exact formulae (calculable to any accuracy) for mutual and conductor-to-ground capacitance (C_m and C_g) for a shielded balanced pair are expressed as infinite determinants. Convergence of these determinants is rapid except as the conductors of the pair approach each other or the shield. Approximate expressions developed by Philips Research, though not extremely accurate, are simple and in closed form thereby allowing capacitance surfaces to be plotted. These surfaces show qualitatively the variation of capacitance with dimensions.

I. INTRODUCTION

The shielded balanced pair consists of two straight cylindrical conductors immersed in a homogeneous dielectric, surrounded by an electrically thick tubular shield. Figure 1 is a cross section of the shielded-pair structure.

Dimensional restrictions are imposed which serve to keep the three conductors of the structure from touching. These restrictions are $S > d$ and $D > S + d$. In terms of the traditional dimensional ratios u and V, $u < 0.5$ and $V < 1/(1 + 2u)$. Thus, the variables u and V are contained in the shaded area in Fig. 2.

The exact capacitance expression for a pair in free space (unshielded) is derived in most texts.

$$C_m = \pi\epsilon_0/\cosh^{-1}(0.5/u) \tag{1}$$

where $\epsilon_0 = (c^2\mu_0)^{-1} = 8.8541853 \times 10^{-12}$ farads/meter.

As the spacing between the conductors approaches zero, $u \to 0.5$ and $C_m \to \infty$. The effect of placing a shield around the pair at a large relative distance cannot alter the limiting values of C_m as $u \to 0.5$. Also, intuitively, as the inner conductors of the shielded pair structure approach the outer conductor, $V \to 1/(1 + 2u)$ and $C_g \to \infty$. The capacitance values at the limiting dimensions appear in Table I.

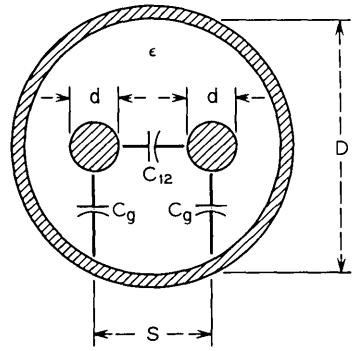


Fig. 1—The shielded-pair transmission line.

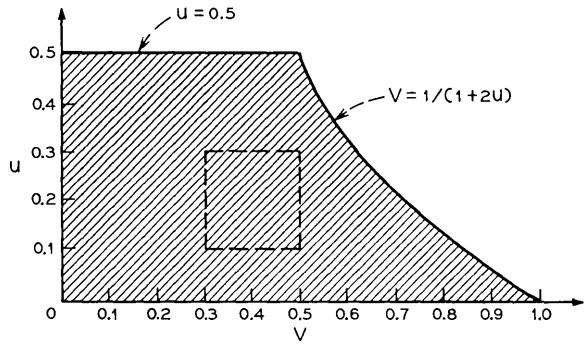


Fig. 2—Dimensional restrictions for the shielded pair. Typical values of u and V for existing shielded balanced-pair cables and equivalent u and V for multipair cables generally lie within the dashed square.

II. EXACT CAPACITANCE EXPRESSIONS

J. W. Craggs and C. J. Tranter[1] derived the exact expression for the mutual capacitance of the shielded balanced pair. The method assumed a Fourier surface charge density on the conductors, the coefficients being determined from the constancy of the potential over the surfaces. The factor δ_{12} is expressed in an infinite determinant set equal to zero.

$$\frac{C_m}{\epsilon} = \frac{0.044766}{\log_e \left(\frac{1}{u} \cdot \frac{1 - V^2}{1 + V^2}\right) - \delta_{12}} \ \mu\text{F/mile} \qquad (2)$$

$$\begin{vmatrix} -\delta_{12} & \alpha_1(2u) & \alpha_2(2u)^2 & \cdots \\ -\frac{\alpha_1}{1}(2u) & 1 + A_{11}(2u)^2 & A_{21}(2u)^3 & \cdots \\ -\frac{\alpha_2}{2}(2u)^2 & A_{12}(2u)^3 & 1 + A_{22}(2u)^4 & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{vmatrix} = 0, \qquad (3)$$

where

$$\alpha_m = \left(-\frac{1}{2}\right)^m + \left(\frac{V^2}{1 - V^2}\right)^m - \left(\frac{-V^2}{1 + V^2}\right)^m$$

and

$$A_{mn} = -{}^mB_n\left(-\frac{1}{2}\right)^{m+n} - 2\sum_q^{\infty} {}^{2q+1}C_n \, {}^mB_{2q+1-m} V^{2(2q+1)},$$

and q equals $(n - 1)/2$ or $(m - 1)/2$, whichever is greater. B represents binomial coefficients and C represents combinatorial coefficients.

TABLE I—LIMITING CAPACITANCE VALUES

Dimensional Condition	Capacitance Value
$D = \infty$, $V = 0$ $d \to S$, $u \to 0.5$ $S + d \to D$, $V \to 1/(1 + 2u)$	$C_m = 0.044765\epsilon_r/\cosh^{-1}(0.5/u)$ uF/mile $C_m \to \infty$ $C_g \to \infty$

This solution lends itself well to computer calculation and for small values of u and V, a 2×2 determinant will give accurate answers. As u and V increase, more terms must be included until at $u = 0.45$, a 10×10 determinant is required to give five-digit accuracy. The convergence of this determinant is very slow as the conductors approach each other or the shield (u and V approach the limiting values).

The exact expression for C_g, capacitance to ground of one conductor, was derived by the author using the method of Craggs and Tranter. This derivation yields another infinite determinant equal to zero.

With polar coordinates (r, θ) let the surface charge density on $r = r_1$ be

$$f(\theta) = \frac{Q}{2\pi r_1}\left(a_0 + \sum_{n=1}^{\infty} a_n \cos n\theta\right).$$

An even function is selected since the conductors will be equipotential circles. The potential due to this charge density is then

$$^{\dagger}\ V = -2Qa_0 \log r_1 + Q \sum_{n=1}^{\infty} \left(\frac{r}{r_1}\right)^n \frac{a_n \cos n\theta}{n} \quad \text{for} \quad r \leqq r_1 \tag{4}$$

or

$$^{\dagger}\ V = -2Qa_0 \log r + Q \sum_{n=1}^{\infty} \left(\frac{r_1}{r}\right)^n \frac{a_n \cos n\theta}{n} \quad \text{for} \quad r \geqq r_1\ . \tag{5}$$

Since three conductors, each with different polar origins, are involved, several coordinate transformations are required. For polar coordinates (ρ, ω) with origin at $r = c$, $\theta = 0$, and $c > r_1$ and $\rho < c - r_1$ use equation (5). Using

$$\left(\frac{1}{1-x}\right)^n = \sum_{m=0}^{\infty} {}^nB_m x^m, \qquad {}^nB_m = \frac{(n+m-1)!}{m!\,(n-1)!}$$

and

$$\log\left(1 + 2\frac{\rho}{c}\log\omega + \frac{\rho^2}{c^2}\right) = -2\sum_{n=1}^{\infty} \frac{\cos n\omega}{n}\left(-\frac{\rho}{c}\right)^n,$$

then

$$\begin{aligned} ^{\dagger}\ V = {} & -2Qa_0 \log c + 2Qa_0 \sum_{n=1}^{\infty} \frac{\cos n\omega}{n}\left(-\frac{\rho}{c}\right)^n \\ & + Q\sum_{n=1}^{\infty} \frac{a_n}{n}\left(\frac{r_1}{c}\right)^n {}^nB_m\left(\frac{-\rho}{c}\right)^m \cos m\omega. \end{aligned} \tag{6}$$

For $c < r_1$ and $\rho < r_1 - c$ and $\rho < c$ use equation (4). Using

$$(1+x)^n = \sum_{m=0}^{n} {}^nC_m x^m, \qquad {}^nC_m = \frac{n!}{m!\,(n-m)!}\ ,$$

then

$$^{\dagger}\ V = -2Qa_0 \log r_1 + Q \sum_{n=1}^{\infty} \frac{a_n}{n}\left(\frac{c}{r_1}\right)^n \sum_{m=0}^{n} {}^nC_m\left(\frac{\rho}{c}\right)^m \cos m\omega. \tag{7}$$

For $c > r_1$ and $\rho > c + r_1$ use equation (5). Using previously stated identities,

$$\begin{aligned} ^{\dagger}\ V = {} & -2Qa_0 \log \rho + 2Qa_0 \sum_{n=1}^{\infty} \frac{\cos n\omega}{n}\left(-\frac{c}{\rho}\right)^n \\ & + Q\sum_{n=1}^{\infty} \left(\frac{r_1}{\rho}\right)^n \frac{a_n}{n} \sum_{m=0}^{\infty} {}^nB_m\left(-\frac{c}{\rho}\right)^m \cos(m+n)\omega. \end{aligned} \tag{8}$$

† This result is stated in Ref. 1.

For polar coordinates (ρ, ω) with origin at $r = c$, $\theta = \pi$, and $c > r_1$ and $\rho > c + r_1$ use equation (5).

$$V = -2Qa_0 \log \rho + 2Qa_0 \sum_{n=1}^{\infty} \frac{\cos n\omega}{n} \left(\frac{c}{\rho}\right)^n + Q \sum_{n=1}^{\infty} \left(\frac{r_1}{\rho}\right)^n \frac{a_n}{n} \sum_{m=0}^{\infty} {}^nB_m \left(\frac{c}{\rho}\right)^m \cos (m + n)\omega. \tag{9}$$

For the shielded balanced pair structure let the inner and outer conductors be numbered as shown in Fig. 3.

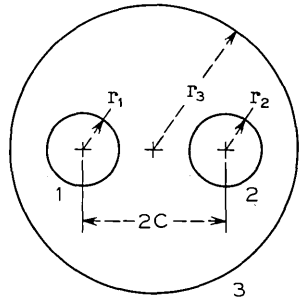


Fig. 3—Polar coordinates for the three conductors of the shielded-pair structure.

Assuming a charge distribution on conductor 1

$$f(\theta) = \frac{Q}{2\pi r_1} \left(1 + \sum_{n=1}^{\infty} a_n \cos n\theta\right),$$

and on conductor 2

$$f(\pi - \theta) = \frac{Q}{2\pi r_1} \left[1 + \sum_{n=1}^{\infty} a_n \cos n(\pi - \theta)\right] = \frac{Q}{2\pi r_1} \left[1 + \sum_{n=1}^{\infty} (-1)^n a_n \cos n\theta\right],$$

and on conductor 3

$$f(\theta) = \frac{Q}{2\pi r_3} \left[b_0 + \sum_{n=1}^{\infty} b_n \cos n\theta\right].$$

Writing the equation for the potential on conductor 3 with both inner conductors at $+V_0$ potential and conductor 3 at zero potential yields

$$0 = V_{31} + V_{32} + V_{33},$$

where V_{ab} is the potential on (a) due to a charge distribution on (b). From (8) with $\rho = r_3$ and $2c$ separation between conductors

$$\frac{V_{31}}{Q} = -2 \log r_3 + 2 \sum_{n=1}^{\infty} \left(-\frac{c}{r_3}\right)^n \frac{\cos n\omega}{n} + \sum_{n=1}^{\infty} \left(\frac{r_1}{r_3}\right)^n \frac{a_n}{n} \sum_{m=0}^{\infty} {}^nB_m \left(-\frac{c}{r_3}\right)^m \cos (m + n)\omega.$$

From (9) with $\rho = r_3$,

$$\frac{V_{32}}{Q} = -2 \log r_3 + 2 \sum_{n=1}^{\infty} \left(\frac{c}{r_3}\right)^n \frac{\cos n\omega}{n} + \sum_{n=1}^{\infty} \left(\frac{r_1}{r_3}\right)^n \frac{a_n}{n} (-1)^n \sum_{m=0}^{\infty} {}^nB_m \left(\frac{c}{r_3}\right)^m \cos (m + n)\omega.$$

From (4) or (5) with $r = r_1 = r_3$,

$$\frac{V_{33}}{Q} = -2b_0 \log r_3 + \sum_{n=1}^{\infty} \frac{b_n \cos n\omega}{n}.$$

Substituting $k = m + n$ yields the total potential on conductor 3,

$$0 = -2(2 + b_0) \log r_3 + 2 \sum_{n=1}^{\infty} \left(\frac{c}{r_3}\right)^n \frac{\cos n\omega}{n} [1 + (-1)^n] + \sum_{n=1}^{\infty} \left(\frac{r_1}{r_3}\right)^n \frac{a_n}{n} \sum_{k=n}^{\infty} {}^nB_{k-n} \left(\frac{c}{r_3}\right)^{k-n} (-1)^n [1 + (-1)^k] \cos k\omega + \sum_{n=1}^{\infty} \frac{b_n \cos n\omega}{n}.$$

From this relationship we select the charge distribution on conductor 3 so as to eliminate the angle (ω) dependence.

$$f(\theta) = \frac{Q}{2\pi r_3} \left(-2 + \sum_{n=1}^{\infty} b_{2n} \cos 2n\theta\right) \tag{10}$$

Using

$${}^nB_{k-n} = 0 \quad \text{for} \quad k < n$$

and

$$\sum_{n=1}^{\infty} \sum_{k=n}^{\infty} {}^nB_{k-n} f(n, k) = \sum_{k=1}^{\infty} \sum_{n=1}^{k} {}^nB_{k-n} f(n, k),$$

then

$$b_{2n} = -4\left(\frac{c}{r_3}\right)^{2n}\left[1 + n\sum_{k=1}^{2n}\left(\frac{r_1}{c}\right)^k \frac{a_k}{k}\,{}^kB_{2n-k}\right]. \tag{11}$$

Equation (11) gives a functional relationship between the Fourier coefficients for the assumed charge distributions. For conductor 2 the total potential $V_0 = V_{21} + V_{22} + V_{23}$. From (6) with $\rho = r_1$ and $2c$ separation between conductors,

$$V_{21} = -2Q \log 2c + 2Q\sum_{n=1}^{\infty}\frac{\cos n\omega}{n}\left(-\frac{r_1}{2c}\right)^n + Q\sum_{n=1}^{\infty}\frac{a_n}{n}\left(\frac{r_1}{2c}\right)^n \sum_{m=0}^{\infty}{}^nB_m\left(-\frac{r_1}{2c}\right)^m \cos m\omega.$$

From (4) or (5) with $r = r_1$,

$$V_{22} = -2Q \log r_1 + \sum_{n=1}^{\infty}\frac{a_n}{n}(-1)^n \cos n\omega.$$

From (7) and (10) with $\rho = r_1$ and $r_1 = r_3$,

$$V_{23} = +4Q \log r_3 + Q\sum_{n=1}^{\infty}\frac{b_{2n}}{2n}\left(\frac{c}{r_3}\right)^{2n}\sum_{m=0}^{2n}{}^{2n}C_m\left(\frac{r_1}{c}\right)^m \cos m\omega.$$

The total potential on conductor 2 is then,

$$\frac{V_0}{Q} = -2 \log 2cr_1 + 2\sum_{n=1}^{\infty}\frac{\cos n\omega}{n}\left(-\frac{r_1}{2c}\right)^n + \sum_{n=1}^{\infty}\frac{a_n}{n}(-1)^n \cos n\omega + \sum_{m=1}^{\infty}\frac{a_m}{m}\left(\frac{r_1}{2c}\right)^m \sum_{n=0}^{\infty}{}^mB_n\left(-\frac{r_1}{2c}\right)^n \cos n\omega + 4 \log r_3 + \sum_{m=1}^{\infty}\frac{b_{2m}}{2m}\left(\frac{c}{r_3}\right)^{2m}\sum_{n=0}^{2m}{}^{2m}C_n\left(\frac{r_1}{c}\right)^n \cos n\omega.$$

For $n = 0$,

$$\frac{V_0}{Q} = 2 \log \frac{(r_3)^2}{2cr_1} + \sum_{m=1}^{\infty}\frac{a_m}{m}\left(\frac{r_1}{2c}\right)^m + \sum_{m=1}^{\infty}\frac{b_{2m}}{2m}\left(\frac{c}{r_3}\right)^{2m}. \tag{12}$$

Using ${}^{2m}C_n = 0$ for $n > 2m$,

$$\sum_{m=0}^{\infty}\sum_{n=0}^{2m}{}^{2m}C_n f(m, n) = \sum_{n=0}^{\infty}\sum_{2m=n}^{\infty}{}^{2m}C_n f(m, n).$$

For $n = 1, 2, 3 \cdots$

$$0 = +\frac{2}{n}\left(-\frac{r_1}{2c}\right)^n + (-1)^n\frac{a_n}{n} + \left(-\frac{r_1}{2c}\right)^n\sum_{m=1}^{\infty}{}^mB_n\left(\frac{r_1}{2c}\right)^m\frac{a_m}{m} + \left(\frac{r_1}{c}\right)^n\sum_{2m=n}^{\infty}{}^{2m}C_n\left(\frac{c}{r_3}\right)^{2m}\frac{b_{2m}}{2m}. \tag{13}$$

To obtain the final solution, substitute (11) into (12) and (13) to obtain a system of $n + 1$ homogeneous equations in $n + 1$ unknowns. The determinant of this system is then equal to zero.

$$\frac{V_0}{Q} + 2 \log \frac{2cr_1}{(r_3)^2} + 4 \sum_{k=1}^{\infty} \left(\frac{c}{r_3}\right)^{2(2k)} \frac{1}{2k} - \sum_{m=1}^{\infty} \frac{a_m}{m} \left(\frac{r_1}{c}\right)^m \alpha_m = 0.$$

$$\frac{2}{n} \left(\frac{r_1}{c}\right)^n \alpha_n + (-1)^n \frac{a_n}{n} + \left(\frac{r_1}{c}\right)^n \sum_{m=1}^{\infty} A_{mn} \left(-\frac{r_1}{c}\right)^m \left(-\frac{a_m}{m}\right) = 0,$$

where

$$\alpha_m = \left(-\frac{1}{2}\right)^m - \left(\frac{c^2}{r_3^2 - c^2}\right)^m - \left(\frac{-c^2}{r_3^2 + c^2}\right)^m,$$

$$A_{mn} = {}^mB_n \left(-\frac{1}{2}\right)^{m+n} - 2 \sum_{q}^{\infty} {}^{2q}C_n \, {}^mB_{2q-m} \left(\frac{c}{r_3}\right)^{2(2q)};$$

and

$$q = m/2 \text{ or } n/2, \quad \text{whichever is greater.}$$

Using

$$\frac{V_0}{Q} = \frac{\epsilon}{C_g},$$

and transferring to the traditional dimensional ratios, $u = r_1/2c$ and $V = c/r_3$, yields $n + 1$ equations in the variable $(-1)^m a_m/m$ so that

$$\frac{C_g}{\epsilon} = \frac{0.089532}{\log \frac{1}{4uV^2} \prod_{k=1}^{\infty} \left[\frac{1 - V^{2^{k+1}}}{1 + V^{2^{k+1}}}\right]^{(\frac{1}{2})^k} - \Delta_g} \ \mu\text{F/mile}, \tag{14}$$

where

$$\begin{vmatrix} -\Delta_g & -\frac{\alpha_1}{2}(2u) & -\frac{\alpha_2}{2}(2u)^2 & \cdots \\ \frac{2\alpha_1}{1}(2u) & 1 + A_{11}(2u)^2 & A_{21}(2u)^3 & \cdots \\ \frac{2\alpha_2}{2}(2u)^2 & A_{12}(2u)^3 & 1 + A_{22}(2u)^4 & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{vmatrix} = 0$$

The convergence characteristics of this solution are similar to the C_m solution with the added effect of the infinite product as $V \to 1$.

With these two exact solutions and the computer programs to calculate them, other closed-form approximate solutions can be evaluated.

III. APPROXIMATE CAPACITANCE EXPRESSIONS

Several approximate expressions (Refs. 2 through 6) using various methods and useful over limited ranges of u and V have been derived.

The Philips[3] approximate equations for C_m and C_g have several useful characteristics.

$$\frac{C_m}{\epsilon} = \frac{0.17906}{4 \cosh^{-1} \rho} \quad \frac{\mu F}{\text{mile}}, \tag{15}$$

where

$$\rho = \frac{1}{2u} \cdot \frac{1 - V^2(1 - 4u^2)}{1 + V^2(1 - 4u^2)}.$$

$$\frac{C_g}{\epsilon} = \frac{0.17906}{2 \cosh^{-1} q} \quad \frac{\mu F}{\text{mile}}, \tag{16}$$

where

$$q = \frac{1}{2u} \cdot \frac{1 - V^4(1 - 4u^2)}{4V^2}.$$

These equations are simple, easy to compute, and in closed form. Additionally, they give the values of capacitance shown in Table I at the dimensional limits. Tables II and III show that the overall accuracy of these relationships in the range of u and V corresponding to practical cables is not high.

IV. NUMERICAL TECHNIQUES

Several numerical techniques have been described to calculate capacitance by mapping equipotential lines. A grid is usually assumed to set values of x and y and known boundary values of potential are used. The value $V(x, y)$ is taken as the average of the four neighboring values, and when this is true for all points, Laplace's Equation, $\partial^2 V/\partial x^2 + \partial^2 V/\partial y^2 = 0$, has been solved. This relaxation method is used in a computer program[7] to calculate capacitance for a system of shielded circular conductors.

It is difficult when using a relaxation method to determine the accuracy of the final solution (due to the finite grid and computer errors). The program[7] was run with the proper parameters to calculate

TABLE II—PERCENT ERROR IN PHILIPS C_m EQUATION
$(C_{PHILIPS} - C_{EXACT})^* 100/C_{EXACT}$

	u									
V	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
0	0	0	0	0	0	0	0	0	0	0
0.05	0.001	0.004	0.012	0.024	0.044	0.073	0.114	0.174	0.267	0
0.1	0.003	0.017	0.047	0.097	0.176	0.292	0.458	0.700	1.07	0
0.15	0.007	0.038	0.104	0.218	0.394	0.657	1.04	1.59	2.47	0
0.2	0.013	0.066	0.182	0.383	0.699	1.17	1.86	2.89	4.52	0
0.25	0.019	0.101	0.279	0.592	1.09	1.84	2.96	4.64	7.39	0
0.3	0.026	0.142	0.395	0.843	1.56	2.67	4.35	6.94	11.3	0
0.35	0.035	0.189	0.528	1.14	2.13	3.68	6.09	9.93	16.7	0
0.4	0.44	0.240	0.677	1.47	2.79	4.90	8.28	13.9	24.4	0
0.45	0.054	0.295	0.843	1.86	3.58	6.41	11.1	19.4	36.5	0
0.5	0.064	0.355	1.03	2.30	4.53	8.33	15.0	28.1	61.7	0
0.55	0.075	0.422	1.24	2.84	5.73	11.0	21.5	51.0		
0.6	0.087	0.496	1.49	3.52	7.44	15.7				
0.65	0.099	0.584	1.82	4.49	10.8					
0.7	0.114	0.695	2.28	6.5						
0.75	0.132	0.854	3.22							
0.8	0.157	1.15								
0.85	0.200									
0.9	2.49									

Note: Typical values of u and V for shielded-pair and multipair cables generally lie within the enclosed area.

C_m and C_g of a shielded pair with $u = V$ between 0.05 and 0.45 in 0.05 increments. The finest grid (33 × 33 in each quadrant) was used and Table IV shows the error in the relaxation solution. The conductor diameter for $u = V = 0.05$ did not include enough of the grid to compute. Unlike the approximate equations, the errors in the relaxation method, for the most part, are not a strong function of u and V. For most values of u and V ($u = V > 0.2$) the relaxation method is more accurate than the Philips equations.

V. CAPACITANCE SURFACES

The complexity of the exact formulae for calculating the capacitance of a shielded balanced pair structure causes difficulty in visualizing how capacitance is affected by changes in dimensions. The Philip's

TABLE III—PERCENT ERROR IN PHILIPS C_g EQUATION $(C_{PHILIPS} - C_{EXACT})^* 100/C_{EXACT}$

	u									
V	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
0.05	−0.033	−0.144	−0.342	−0.630	−1.01	−1.47	−2.01	−2.62	−3.27	0
0.1	−0.040	−0.180	−0.435	−0.811	−1.31	−1.93	−2.66	−3.49	−4.40	0
0.15	−0.046	−0.211	−0.517	−0.974	−1.59	−2.36	−3.29	−4.34	−5.50	0
0.2	−0.052	−0.241	−0.596	−1.14	−1.87	−2.81	−3.94	−5.24	−6.68	0
0.25	−0.057	−0.270	−0.677	−1.31	−2.17	−3.28	−4.64	−6.22	−7.99	0
0.3	−0.062	−0.300	−0.762	−1.49	−2.50	−3.81	−5.42	−7.32	−9.47	0
0.35	−0.067	−0.330	−0.853	−1.68	−2.85	−4.40	−6.31	−8.58	−11.2	0
0.4	−0.073	−0.364	−0.952	−1.9	−3.26	−5.07	−7.35	−10.1	−13.2	0
0.45	−0.079	−0.400	−1.06	−2.15	−3.73	−5.87	−8.59	−11.9	−15.9	0
0.5	−0.085	−0.440	−1.19	−2.44	−4.30	−6.86	−10.2	−14.4	−19.8	0
0.55	−0.092	−0.487	−1.34	−0.280	−5.03	−8.20	−12.6	−19.1		
0.6	−0.099	−0.542	−1.53	−3.27	−6.06	−10.4				
0.65	−0.109	−0.611	−1.77	−3.96	−8.00					
0.7	−0.120	−0.703	−2.15	−5.38						
0.75	−0.135	−0.841	−2.91							
0.8	−0.159	−1.11								
0.85	−0.199									
0.9	−2.08									

Note: Typical values of u and V for shielded-pair and multipair cables generally lie within the enclosed area.

TABLE IV—PERCENT ERROR USING A RELAXATION METHOD

(C_m, μF/mile)

u	V	C_{EXACT}	C_{RELAX}	Percent Error
0.05	0.05	0.01498054	0	−100.
0.1	0.1	0.01969278	0.01985524	0.825
0.15	0.15	0.02442631	0.02445576	0.121
0.2	0.2	0.02987059	0.02984397	−0.089
0.25	0.25	0.03669688	0.03665908	−0.103
0.3	0.3	0.04594544	0.04589750	−0.104
0.4	0.4	0.08314346	0.08311094	−0.039
0.45	0.45	0.13697544	0.13692383	−0.038

(C_g, μF/mile)

u	V	C_{EXACT}	C_{RELAX}	Percent Error
0.05	0.05	0.01178299	0	−100.
0.1	0.1	0.01624484	0.01363114	−16.1
0.15	0.15	0.02090778	0.02052972	−1.81
0.2	0.2	0.02632710	0.02623542	−0.348
0.25	0.25	0.03308222	0.03298556	−0.292
0.3	0.3	0.04210923	0.04203434	−0.178
0.35	0.35	0.05529502	0.05520572	−0.162
0.4	0.4	0.07744843	0.07732726	−0.156
0.45	0.45	0.12737824	0.12717648	−0.158

equations provide simple functions which can be plotted to give a qualitative picture of a capacitance surface. Computer plotting of a function of two variables[8] has been described and will be used to display capacitance surfaces along with scales, coordinate axes and a "cube of reference."

Figures 5, 6, and 7 are plots of capacitance surfaces as functions of d/D and S/D. The dimensional restrictions with these ratios, map into the region shown in Fig. 4.

These plots in the interval $0.05 \leqq d/D \leqq 0.3$ and $0.35 \leqq S/D \leqq 0.65$ can be interpreted as capacitance surfaces normalized to some constant outer conductor diameter (say $D = 1$). Figure 5 shows C_{12}/ϵ to be a monotonically increasing function of d/D and a monotonically decreasing function of S/D with a maximum slope as the inner conductors approach each other. Figure 6 shows C_g/ϵ is an increasing function of both d/D and S/D with a maximum slope as the inner conductors approach the shield. C_m/ϵ, plotted in Fig. 7, is seen to contain the sum of previous effects with approximately equal weighting.

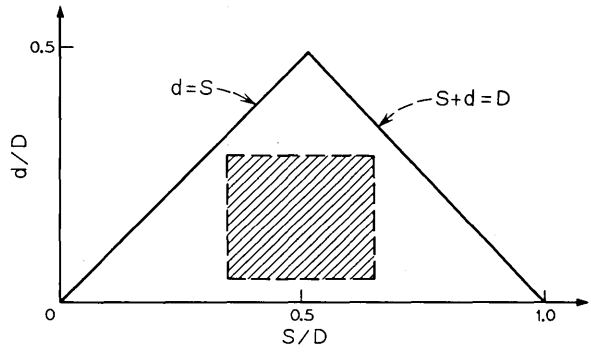


Fig. 4—Dimensional limits as functions of d/D and S/D. The shaded area defines the plotting interval in Figs. 5, 6, and 7.

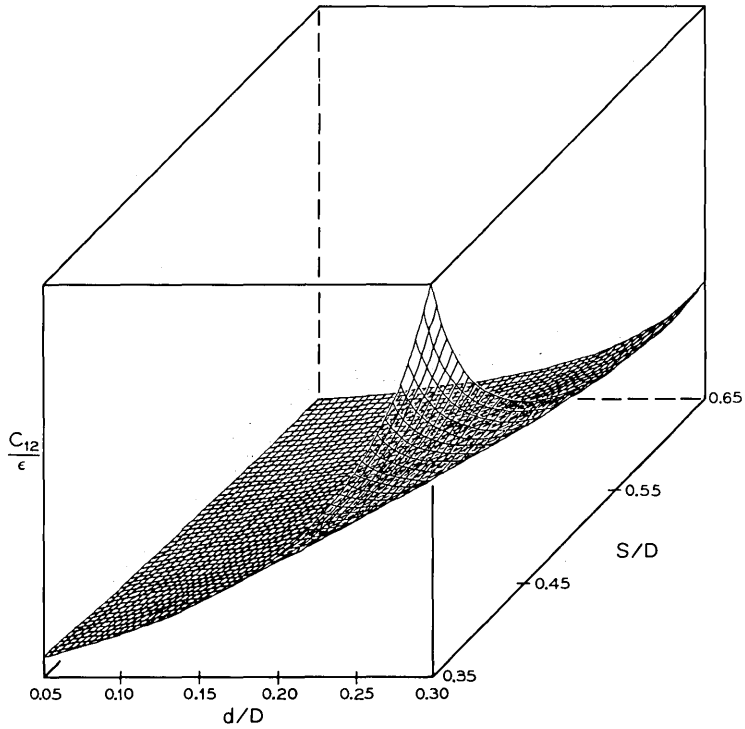


Fig. 5—C_{12}/ϵ as a function of d/D and S/D.

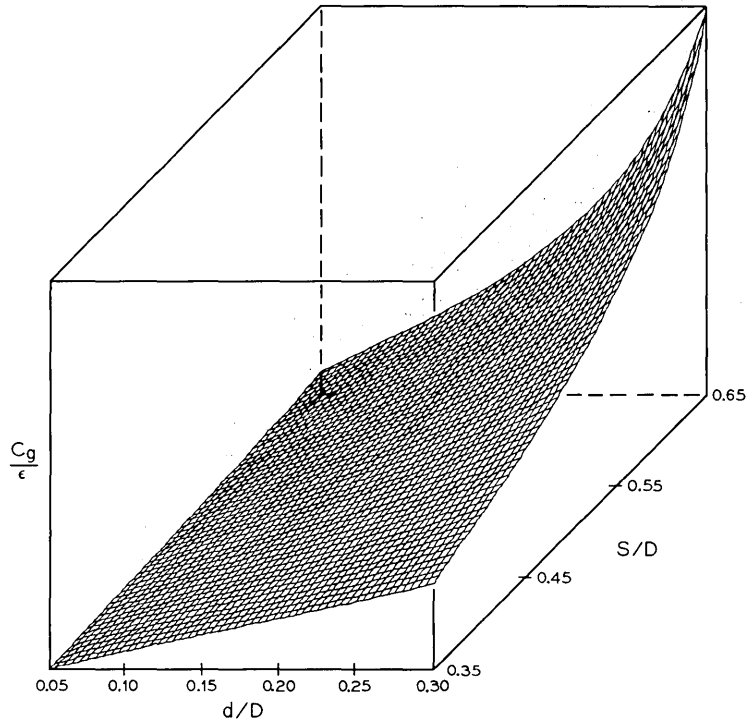


Fig. 6—C_g/ϵ as a function of d/D and S/D.

The previous three plots show qualitatively how C_m and its components vary with dimensions. They also reinforce the previous intuitive statements made regarding behavior at dimensional limits. Thus, for a given d/D, C_g/ϵ is seen to increase as S/D increases (wires approach shield) whereas C_{12}/ϵ is seen to increase as S/D decreases (wires approach each other). C_m/ϵ, the functional sum of C_g/ϵ and C_{12}/ϵ, therefore has a minimum with respect to S/D. The position of this minimum is only a slight function of d/D; for $0.05 \leqq d/D \leqq 0.3$, C_m/ϵ is minimum for $0.48 \leqq S/D \leqq 0.5$.

VI. INVERSE PLOTS OF PHILIPS EQUATION

With the Philips equations, values of u and V can be obtained for a given C_m/ϵ and C_g/ϵ (as could be measured in a cable). Thus from equations (15) and (16),

$$\rho + \sqrt{\rho^2 - 1} = \exp \frac{0.17906\epsilon}{4C_m} \tag{17}$$

$$q + \sqrt{q^2 - 1} = \exp \frac{0.17906\epsilon}{2C_g} \tag{18}$$

using $\cosh^{-1} a = \log_e (a + \sqrt{a^2 - 1})$ for $a \geqq 1$.

Solving (17) and (18) iteratively for ρ and q yields two simultaneous iterative equations in two unknowns.

$$\rho = \frac{1}{2u} \cdot \frac{1 - V^2(1 - 4u^2)}{1 + V^2(1 - 4u^2)}. \tag{19}$$

$$q = \frac{1}{2u} \cdot \frac{1 - V^2(1 - 4u^2)}{4V^2}. \tag{20}$$

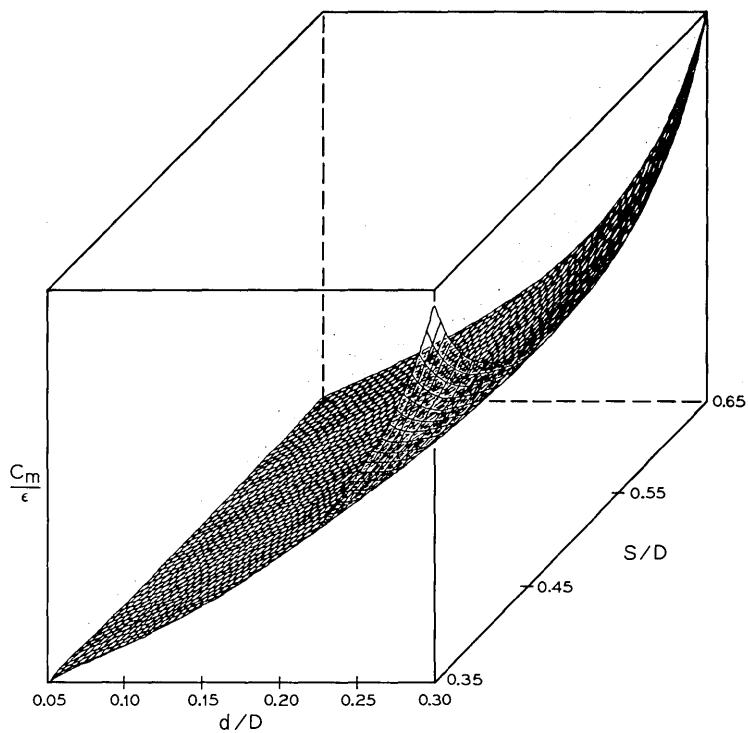


Fig. 7—C_m/ϵ as a function of d/D and S/D.

Solving (19) for V^2 and substituting in (20) yields,

$$V^2 = \frac{1 - 2u\rho}{1 - 4u^2 + 2u\rho - 8u^3\rho} \tag{21}$$

and

$$\frac{8u\rho(1 - 2u\rho)}{1 - 4u^2 + 2u\rho - 8u^3\rho} = 1 - \frac{(1 - 2u\rho)^2(1 - 4u^2)^2}{(1 - 4u^2 + 2u\rho - 8u^3\rho)^2}. \tag{22}$$

Equation (22) can now be iteratively solved for u and substituted in (21) to get V. The variables u and V are again inconvenient from the standpoint of intuitive visualization. In the manufacture of a shielded balanced-pair cable, D and S are difficult to control; however, d can be accurately controlled. Even if d varies down the length of the cable this can be related to die wear or to elongation, which can be modeled. Normalizing S and D with respect to d yields,

$$\frac{S}{d} = \frac{1}{2u} \tag{23}$$

and

$$\frac{D}{d} = \frac{1}{2uV}. \tag{24}$$

Figures 8 and 9 are surfaces of S/d and D/d as functions of C_m/ϵ and C_g/ϵ in the following regions.

$$0.02 \leqq C_m/\epsilon \leqq 0.07$$

$$0.015 \leqq C_g/\epsilon \leqq 0.035.$$

No attempt is made to define the bounding values for C_m/ϵ and C_g/ϵ; however, the above represents one of the largest rectangular regions the author could find by trial and error. These surfaces have the expected general shape ($S/d \to \infty$ as $C_m/\epsilon \to C_g/2\epsilon$ and $D/d \to \infty$ as $C_g/\epsilon \to 0$).

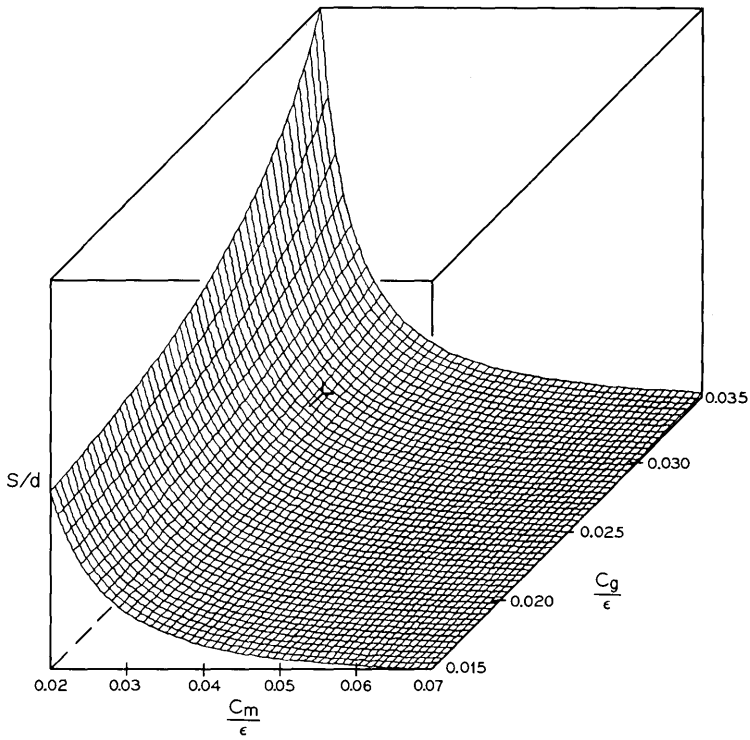


Fig. 8—S/d as a function of C_m/ϵ and C_g/ϵ.

VII. CONCLUSIONS

The exact expressions for C_m and C_g must be used if a high degree of accuracy is desired. As the conductors approach each other or the shield, the convergence of the exact solutions is slower and the accuracy of approximate expressions is less. For large values of u and V, the relaxation method yields a more accurate result than the Philips approximate equations; however, the error is difficult to ascertain. Capacitance surfaces using the Philips relationships show a great deal about the manner in which capacitance varies with dimensions and enables inverse plots of dimensions versus capacitance.

VIII. ACKNOWLEDGMENT

The author wishes to thank H. P-H Yuen for his help in deriving the exact expression for C_g .

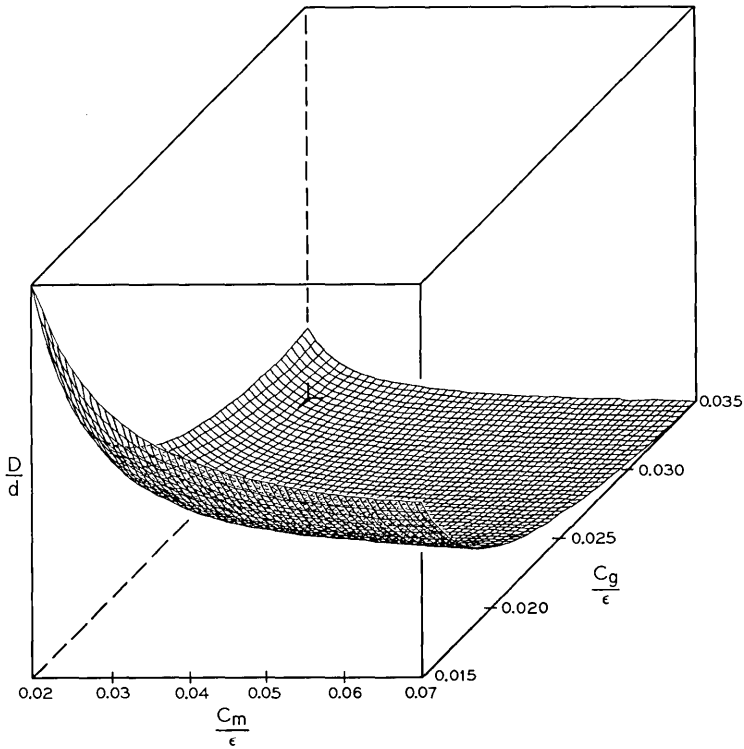


Fig. 9—D/d as a function of C_m/ϵ and C_g/ϵ.

REFERENCES

1. Craggs, J. W., and Tranter, C. J., "The Capacity of Two-Dimensional Systems of Conductors and Dielectrics with Circular Boundaries," Quart. J. of Math. (Oxford) Series 1, *17*, 1946.
2. "Notes on Cable Development and Design," 1952, Section 3, Bell Laboratories (Baltimore, Md.), pp. 26–30.
3. van Hofweegen, J. M., and Knol, K. S., "The Universal Adjustable Transformer for UHF Work," Philips Research Reports, *3*, No. 2 (April 1948), pp. 140–155.
4. Lynch, J. K., "Impedance of Balanced Screened Transmission Lines," Research Laboratory Report, No. 3698, Melbourne, (October 13, 1953).
5. von Burmester, Arthur, "Die Berechnung von Kapazitaten bei Kabeln mit einfachem Querschmitt unter Berucksichtigung der inhomogenen Isolierung," Arch Eleke Vebertr, *24*, (September 1970).
6. Gent, A. W., "Capacitance of Shielded Balanced-Pair Transmission Line," Electrical Communication, *33*, (September 1956).
7. Friesen, H. W., Bell Laboratories (Baltimore, Md.), computer program (undocumented).
8. Miller, C. M., "Computer Plotting of a Function of Two Variables," (December 12, 1969), unpublished work.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 51, No. 3, March, 1972
Printed in U.S.A.

The Equivalent Group Method for Estimating the Capacity of Partial-Access Service Systems Which Carry Overflow Traffic

By SCOTTY NEAL

(Manuscript received August 17, 1971)

We present a technique, called the Equivalent Group method, for estimating the capacities of partial-access service systems which carry overflow traffic. The basic idea is to find a full-access group which has the same capacity as the partial-access system when the arrival process is Poisson. We consider these groups to be "equivalent" and use the capacity of the full-access group when it is offered overflow traffic to estimate the capacity of the partial-access group if it is offered the same overflow traffic.

The Equivalent Group method is used to estimate the capacity of Step-by-Step graded multiples which carry overflow traffic. The resulting procedures can account for day-to-day variations in the offered load, and the computations can be carried out by appropriate use of existing engineering tables.

I. INTRODUCTION

In telephone traffic engineering, it is frequently necessary to understand the behavior of partial-access service systems, i.e., systems in which arriving customers do not have access to all servers. The analysis of such systems is difficult even when the arrivals are adequately approximated by a Poisson process. However, when the arrivals are the overflows from some other service system, there are no simple methods for estimating capacity.

In this note, we present a new procedure, called the Equivalent Group method, for estimating the capacity of partial-access service systems which carry overflow traffic. The basic idea is to find a full-access group of servers which has the same capacity as the partial-access system when the arrival process is Poisson. We consider these groups

to be "equivalent" and use the capacity of the full-access group when it is offered overflow traffic to estimate the capacity of the partial-access group if it is offered the same overflow traffic.

We then consider the Step-by-Step switching system, wherein the trunk groups that interconnect the selectors are sometimes arranged so as to form partial-access systems called graded multiples.[1,2] Although graded multiples have been studied extensively, almost all of the results have had to be based on the assumption that arrivals occur according to a Poisson process. When graded multiples are used as alternate routes, the arrivals are not adequately approximated by a Poisson process.[3] Consequently, the standard approximations (which assume Poisson arrivals) are inadequate. Here we apply the Equivalent Group method to estimate the capacities of such gradings.

Graded multiples which serve overflow traffic have been studied in two other instances. S. Neal[3] and A. Lotze[4] have obtained approximations but their results require extensive calculations. Moreover, their methods are not applicable when grading capacity is significantly affected by the properties of the system in which the grading is imbedded, eg., the Step-by-Step switching system.[1]

II. THE EQUIVALENT GROUP METHOD

A partial-access service system can be viewed as a service device; customers arrive and are either served or blocked.* Hence, to each offered load a for a partial-access group $\mathcal{G}$, there corresponds an "equivalent" full-access system of $s = s(a, \mathcal{G})$ servers† which will experience the same blocking at the load a. That is, if $B_{\mathcal{G}}(a)$ denotes the blocking probability for $\mathcal{G}$ when the offered load is a erlangs, then s is determined by the relation $E_{1,s}(a) = B_{\mathcal{G}}(a)$ where $E_{1,s}(a)$ is the first Erlang loss-function.

If an overflow process having mean a but variance $v > a$ were offered to $\mathcal{G}$, a blocking probability $B_{\mathcal{G}}(a, v) > E_{1,s}(a)$ would result. If the same overflow stream were offered to the equivalent full-access group of $s(a, \mathcal{G})$ servers, the blocking probability would be $B_s(a, v) > E_{1,s}(a)$. Since the blocking probabilities $B_s(a, v)$ have been tabulated,[5] it would be very useful if $B_s(a, v)$ could be used to estimate $B_{\mathcal{G}}(a, v)$.

* All service systems under consideration are in statistical equilibrium and obey a blocked-calls-cleared discipline; a customer arriving to find no server available is blocked, leaves the system, and does not return. Unless specified otherwise, arrivals constitute a Poisson process. Service times are independent and identically distributed according to a negative-exponential distribution.

† There is no apparent relation between $s(a, \mathcal{G})$ and the equivalent number of servers obtained by the extended Equivalent Random method.[3]

The phrase "Equivalent Group method" denotes the entire procedure outlined above. That is, determine $s = s(a, \mathcal{G})$ so that $E_{1,s}(a) = B_{\mathcal{G}}(a)$ and then use $B_s(a, v)$ to estimate $B_{\mathcal{G}}(a, v)$. Of course, we have not described exactly how $B_s(a, v)$ should be used. This point is covered below.

It is not our intention to determine the range of applicability of the Equivalent Group method. We desire only to point out the method and to show how it was used successfully to solve certain engineering problems in the Step-by-Step system.

III. APPLICATIONS

The Equivalent Group method evolved from a search for a simple procedure for engineering Step-by-Step graded multiples which serve overflow traffic. Initially we wanted to make an analytical comparison between $B_{\mathcal{G}}(a, v)$ and $B_s(a, v)$ to determine how $B_s(a, v)$ should be used to estimate $B_{\mathcal{G}}(a, v)$. However, we did not find a feasible analytical method. Having no alternative, we resorted to a numerical study using a simulation.

3.1 *Capacity Estimates for Step-by-Step Gradings*

We chose Step-by-Step gradings of 11, 19, 25, 37, and 45 trunks for experimentation (the gradings used in Ref. 1). The inherent load-balancing present in Step-by-Step systems was represented by the approximate model described in Ref. 1. The arrival processes having specified mean and variance were generated by using the swinging-gate approximation, as analyzed by A. Kuczura,[6] for an overflow process.

For each grading-selector configuration, we used a simulation to generate estimates of the grading load-loss relations for several values of the variance-to-mean ratio $z = v/a$ of the input (overflow) process. Figure 1 contains the results for the 25-trunk grading with 40 selectors. (We will come back to Fig. 1 later on.) The results of the other cases are of a similar nature. The results for $z = 1$ were obtained from Ref. 1.

The first step of the Equivalent Group method is to determine $s = s(a, \mathcal{G})$ so that $E_{1,s}(a) = B_{\mathcal{G}}(a)$ ($\mathcal{G}$ denotes the 25-trunk graded multiple). The results are displayed in Fig. 2. [Notice that $s(a, \mathcal{G})$ is not very sensitive to changes in the load a.] Next, we find $B_s(a, v)$ for each value of $z = v/a$ of interest. This can be done by using Wilkinson's tables[5] together with linear interpolation to obtain values of $B_s(a, v)$ for noninteger s. Figure 1 displays $B_s(a, v)$ as a first approximation to $B_{\mathcal{G}}(a, v)$. Again, the relations for the other gradings were of a similar nature.

For the Step-by-Step applications, we desire results for $1 \leqq z \leqq 3$. Comparing the curves in Fig. 1, we see that $B_s(a, v) > B_g(a, v) > B_g(a)$ for all a and $z > 1$. However, the differences between the load-loss relations obtained by simulation and those obtained from $B_s(a, v)$ appear to be simply related to the size of z. In fact, if we merely decrease z by $\Delta z = 0.2\ (z - 1)$ when carrying out the computations, the modified results are in excellent agreement with the simulation results (see Fig. 3). Moreover, except for the gradings having only two first-choice subgroups, the same adjustment was just as successful for the other cases. The Δz correction was not necessary for the gradings with only two first-choice subgroups.

3.2 *An Engineering Procedure for Step-by-Step Graded Multiples*

Any method used for the engineering of Step-by-Step systems must be able to incorporate into the system model day-to-day variations

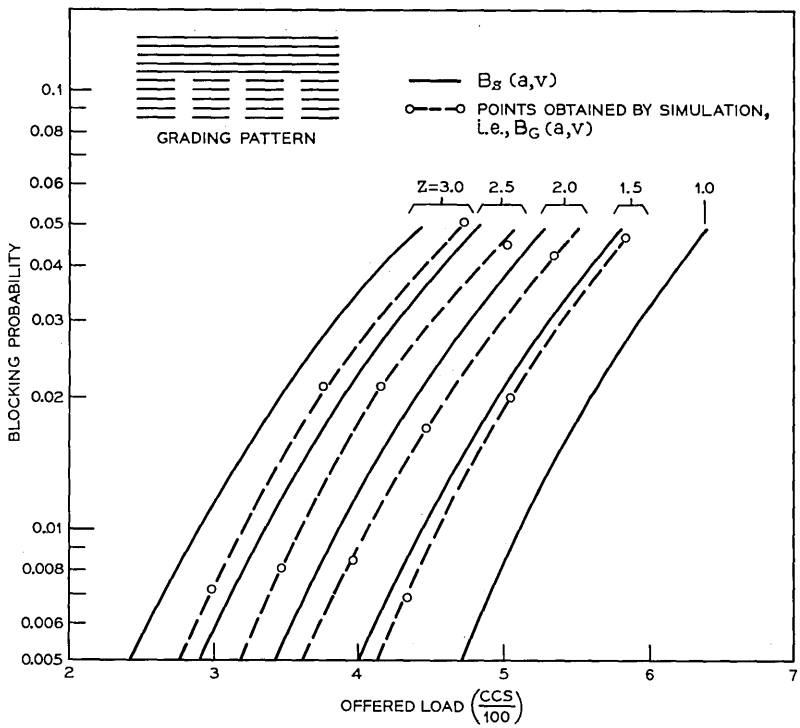


Fig. 1—Approximate load-loss relations for the 25-trunk graded multiple with 40 selectors.

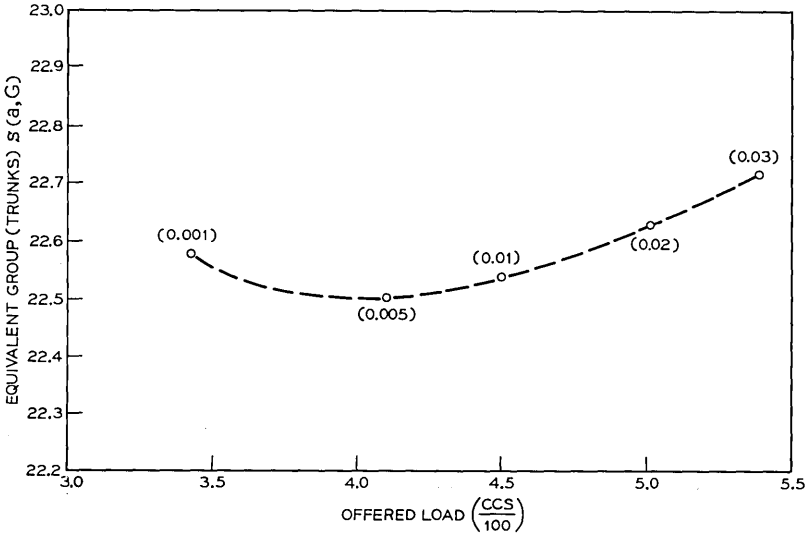


Fig. 2—Equivalent group vs offered load for the 25-trunk graded multiple with 40 selectors. Corresponding blocking enclosed in parentheses.

in the offered load.[7] We account for day-to-day variations by being consistent in the determination of $B_{\mathcal{G}}(a)$, $s(a, \mathcal{G})$, and $B_s(a, v)$. That is, if one is interested in results for day-to-day variations in the offered load, then the same level (low, medium, or high) of day-to-day variations must be assumed at each of the steps* outlined above. An example should clear up this point.

Suppose we are asked to decide which Step-by-Step graded multiple should be connected to 320 selectors in order to carry 13 erlangs (468 CCS) of overflow traffic having $z = 1.7$. We know that high day-to-day variations in the offered load are present and that the average blocking should not exceed 2 percent.

From the preceding, $B_s(a, v) = 0.02$, $a = 13$, and $z = 1.7$. For the first step, decrease z by $\Delta z = 0.2\ (z - 1) = 0.14$.† Using the $\bar{B}$ tables in Ref. 5 we see that 24.4 trunks will serve 13 erlangs under the specified conditions. In the same table, observe that 24.4 trunks will serve 14.5

* In the past, Step-by-Step selector-multiple tables have not explicitly accounted for day-to-day variations in the offered loads. However, revised tables that include the effects of day-to-day variations have been made and will soon be distributed[1].

† We are anticipating that one of the larger gradings will be required. Recall that the Δz correction is not required for the graded-multiples having only two first-choice subgroups.

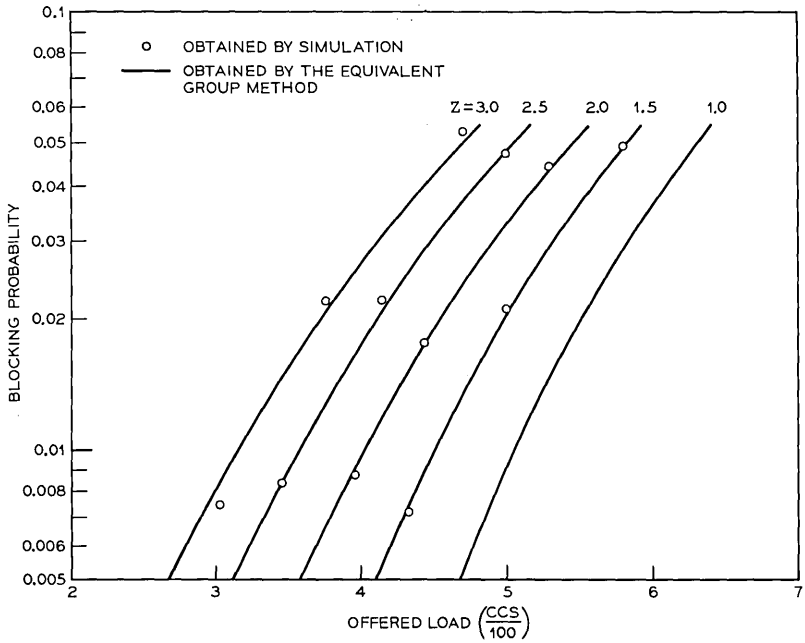


Fig. 3—Load service relations for the 25-trunk graded multiple with 40 selectors.

erlangs of Poisson traffic at $\bar{B}.02$.* Thus, the equivalent group contains 24.4 trunks and the equivalent load is 14.5 erlangs. Consequently, we must find a Step-by-Step grading which, when connected to 320 selectors, will serve an average of 14.5 erlangs of Poisson traffic at $\bar{B}.02$ when high day-to-day variations in the load are present.

Consulting the section for high day-to-day variation in the new Step-by-Step selector-multiple tables, we see that the closest entry is 29 trunks. Thus, the 29-trunk graded multiple connected to 320 selectors will serve an average of 13 erlangs of traffic with $z = 1.7$ at a slightly lower average blocking probability than $\bar{B}.02$, when high day-to-day variations in the offered load are present.

IV. CONCLUSIONS AND SUMMARY

The Equivalent Group method can be used to estimate the capacity of partial-access service systems which serve overflow traffic. In general,

* Since $s(a, \mathcal{G})$ changes very little as a function of a, for computational simplicity we are holding it constant during the computations for this example. The results are not sensitive to this approximation.

we can only say that the Equivalent Group method seems reasonable; an analytical study of the method seems very difficult. We were able to use the method to obtain good estimates of the capacities of Step-by-Step graded multiples for $1 \leqq z \leqq 3$. For the Step-by-Step application, the Equivalent Group method is easy to apply. Moreover, there is no reasonable alternative presently available.

In our applications of the Equivalent Group method, an empirical adjustment (i.e., the Δz correction) was required to maintain accuracy throughout the range of interest, $1 \leqq z \leqq 3$. Hence, one should be very cautious when attempting to use the Equivalent Group method for something other than the Step-by-Step applications discussed above.

REFERENCES

1. Buchner, M. M., Jr., and Neal, S. R., "Inherent Load-Balancing in Step-by-Step Switching Systems," B.S.T.J., *50*, No. 1 (January 1971), pp. 135–165.
2. Lotze, A., "History and Development of Grading Theory," Proceedings of the Fifth International Teletraffic Congress, New York, June 14–20, 1967, pp. 148–161.
3. Neal, S., "Combining Correlated Streams of Nonrandom Traffic," B.S.T.J., *50*, No. 6 (July–August 1971), pp. 2015–2037.
4. Lotze, A., "A Traffic Variance Method for Gradings of Arbitrary Type," Proceedings of the Fourth International Teletraffic Congress, London, July 15–21, 1964, Document No. 80.
5. Wilkinson, R. I., "Nonrandom Traffic Curves and Tables for Engineering and Administrative Purposes," Bell Telephone Laboratories, Traffic Studies Center, 1970.
6. Kuczura, A., "The Interrupted Poisson Process As An Overflow Process," unpublished work, October 4, 1968.
7. Wilkinson, R. I., "A Study of Load and Service Variations in Toll Alternate Route Systems," Proceedings of the Second International Teletraffic Congress, The Hague, July 7–11, 1958, Vol. 3, Paper 29.

Contributors to This Issue

GILBERT F. AMELIO, Ph.D. (Physics), 1968, Georgia Institute of Technology; Bell Laboratories, 1968–1971. Mr. Amelio has been engaged in the development of the silicon diode array camera tube target and in the investigations associated with the study of charge-coupled devices. Member, American Physical Society, IEEE, Sigma Xi.

G. M. BABLER, B.S., 1963, M.S., 1965, and Ph.D. (Physics), 1968, University of Missouri; Bell Laboratories, 1968—. Mr. Babler has done modeling and data analysis work on various aspects of electromagnetic wave propagation in random media. He presently is performing studies to more precisely quantify the atmospheric propagational constraints on line-of-sight microwave radio communication channels.

C. N. BERGLUND, B.Sc. (E.E.), 1960, Queen's University, Kingston, Ont.; M.S.E.E., 1961, Massachusetts Institute of Technology; Ph.D. (E.E.), 1964, Stanford University. Research Assistant, M.I.T., 1960–61; Research Associate, Department of Electrical Engineering, Queen's University, Kingston, 1961–62; Research Assistant, Stanford Electronics Laboratories, 1962–64. Bell Laboratories, 1964—. Mr. Berglund is a supervisor in the Semiconductor Device Laboratory. Member, APS.

LEONARD G. COHEN, B.E.E., 1962, City College of New York; Sc.M., 1964, and Ph.D. (Engineering), 1968, Brown University; Bell Laboratories, 1968—. At Brown University, Mr. Cohen was engaged in research on plasma dynamics. At Bell Laboratories, he has concentrated on the study of optical transmission techniques. Member, IEEE, Sigma Xi, Tau Beta Pi, Eta Kappa Nu.

CALVIN M. MILLER, B.S.E.E., 1963, North Carolina State University; M.S.E., 1966, Akron University; Bell Laboratories, 1967—. Mr. Miller has been engaged in developing equipment and methods for transmission line characterization. His present interest is in the area of exploratory transmission lines.

SCOTTY R. NEAL, B.A. (Mathematics), 1961, M.A. (Mathematics), 1963, and Ph.D. (Mathematics), 1965, University of California, River-

side; Research Mathematician, Naval Weapons Center, China Lake, California, 1964–1967; Bell Laboratories, 1967—. Since coming to Bell Laboratories, Mr. Neal has been primarily concerned with the analysis of various aspects of telephone traffic systems. He has also worked on applications of optimal linear estimation theory and certain aspects of communication theory. Member, American Mathematical Society.

B. Owen, B. Tech. (Eng.), 1963, Welsh College of Advanced Technology, Cardiff, Wales; Ph.D. (E.E.), 1967, Birmingham University, England; Postdoctoral Fellow, Birmingham University, 1967–68; Bell Laboratories, 1968—. Mr. Owen is a member of the Solid State Microwave Device Department, and is presently engaged in the development of millimeter-wave circulators.

David Slepian, University of Michigan, 1941–1943; M.A., 1947, and Ph.D., 1949, Harvard University; Bell Laboratories, 1950—. Mr. Slepian has been engaged in mathematical research in communication theory and noise theory, as well as in a variety of aspects of applied mathematics. During the academic year 1958–59, he was a Visiting Mackay Professor of Electrical Engineering at the University of California at Berkeley and during the Spring semesters of 1967 and 1970 he was a Visiting Professor of Electrical Engineering at the University of Hawaii. He was Editor of the Proceedings of the IEEE during 1969 and 1970. Fellow, IEEE, Institute of Mathematical Statistics. Member, AAAS, American Mathematical Society, SIAM.

R. J. Strain, B.S.E.E., 1958, M.S., 1959, Ph.D., 1963, University of Illinois. After working with Standard Telecommunication Laboratories, England, Mr. Strain joined Bell Laboratories in 1965. His initial activities in the Electron Device Laboratory were concerned with electroluminescent display devices, particularly GaP diodes. In 1968 he joined the Semiconductor Device Laboratory as a supervisor, and he is now in charge of the Charge Transfer Device Group. Member, APS, AAAS, ECS, IEEE, Sigma Xi.


Bell System