RCA Review

June 1976

Volume 37 No. 2

RCARCI 37(2) 157–276 (1976)

RCA Review, published quarterly in March, June, September and December by RCA Research and Engineering, RCA Corporation, Princeton, New Jersey 08540. Entered as second class matter July 3, 1950 under the Act of March 3, 1879 Second-class postage paid at Princeton, New Jersey, and at additional mailing offices. Effective Jan. 1, 197-, subscription rates as follows: United States and Canada: one year $6.00, two years $10.50, three years $13.50; in other countries, one year $6.40, two years $11.30, three years $14.70. Single copies (except for special issues) up to five years old $3.00.

RCA Review

A technical journal published quarterly by RCA Research and Engineering in cooperation with the subsidiaries and divisions of RCA.

Contents

June 1976 Volume 37 Number 2

High Radiance LED for Single-Fiber Optical Links*

James P. Wittke, Michael Ettenberg, and Henry Kressel

RCA Laboratories, Princeton, N. J. 08540

Abstract–The power coupled into a numerical aperture N.A. = 0.14, 90-μm core, multimode fiber from an edge-emitting LED has been increased over previous levels by the development of a narrow-active-region, double-heterojunction type of diode. As much as 0.8 mW has been coupled into a single fiber. The new diode structure results in a narrow (~25° FWHM) emitted beam (in the direction perpendicular to the junction plane), and leads to optical coupling efficiencies 3 to 5 dB better than previously achieved. The radiance of the emitting facet is greatly increased, to over 1000 W/cm^2-sr, a factor of 10 greater than previously reported. The new structure also results in LED's that can be efficiently modulated at high frequencies (>150 MHz). Coupling from the narrow active region of double heterojunction diodes into the fiber is further enhanced (by about 3 dB) by the use of a microlens melted onto the fiber end. Overall coupling efficiencies of −9 to −11 dB are achieved. With the high modulation speed and coupled power of these diodes, material dispersion effects in the fiber, rather than signal power or diode speed, limit the usable system distance–bandwidth product.

1. Introduction

The low-loss multimode fibers needed for long-distance links have numerical apertures (N.A.) on the order of 0.14. Incoherent LED sources, where they can be used, are a desirable alternative to laser diodes for moderate data rates because LED's do not require feedback stabilization

* The research reported in this paper was jointly sponsored by the Department of the Navy, Office of Naval Research, under Contract No. N00014-75-C-0365, and by RCA Laboratories, Princeton, N. J.

of any kind, and their drive circuitry is relatively simple. However, the optical power that can be coupled into a single fiber from an LED is significantly less than from an injection laser. The primary goal of the work reported in this paper was to develop (AlGa)As LED's that would couple the highest possible light power into fibers of N.A. = 0.14, without sacrificing diode speed or operating lifetime.

High coupled power is achieved by combining a source of high radiance (and of appropriate size and shape) with means to render the emitted light beam directive so that it can be coupled into the restricted acceptance aperture of the fiber. Surface-emitting LED's have been reported with radiances R as high as $\sim$100 W/cm^2-sr,[1] emitted from a 50-μm active emitting diameter. For fibers with core diameters exceeding 50 μm, 80 to 140 μm can be coupled from these Lambertian surface-emitting diodes into a single fiber of N.A. = 0.14. Similar results have recently been reported for this type of diode by King, et al.[2]

Edge-emitting diodes of 100 W/cm^2-sr radiance have previously been made with the emission pattern, perpendicular to the junction plane, peaked in the forward direction, permitting more efficient coupling into low-aperture fibers than from Lambertian surface emitters. Moreover, only with the edge-emitting configuration had high-response-frequency diodes been reported.[3,4] For these reasons, edge-emitting diodes hold considerable promise for fiber communications.

2. Diode Design and Experimental Results

2.1 Diode Design

Previous work[3] has shown that the layered structure of double-heterojunction LED's could influence the emitted beam pattern in the direction perpendicular to the plane of the junction. This is illustrated by the far-field emission patterns shown in Fig. 1 for two different LED structures. The introduction of heterojunctions into LED's produces several results. An important one is that the injected carriers can be confined to a recombination region between two heterojunctions. Apart from weak surface-recombination effects at the heterojunctions, the presence of such junctions does not, for low injection rates, alter the recombination rate of the injected minority carriers on which the diode frequency response depends. However, since the bandgaps and, hence, the indices of refraction of the materials on either side of a heterojunction differ, a heterojunction will play a role in altering the edge-emission radiation emission pattern of the LED. The recombination region has a lower bandgap and a higher index than the surrounding regions in a double-heterojunction (DH) diode. Thus, light generated inside the

active region and striking the junction will either be totally internally reflected and confined to the active region or, if the angle of incidence on the junction is large enough, will be transmitted, but refracted, toward the junction plane. This will tend to give a more directed (than Lambertian) beam from an emitting-edge facet and will make the effective source size appear more slit-shaped than would otherwise be the case. This sort of argument can be used to explain the results of Fig. 1.

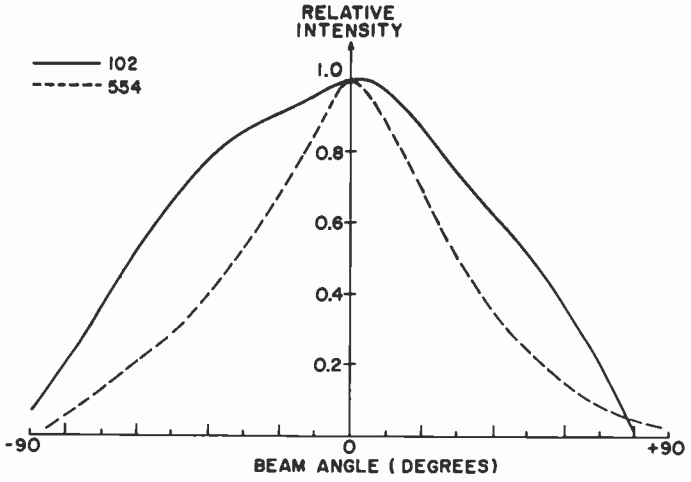

Fig. 1—Far-field radiation patterns perpendicular to the plane of the p-n junction for two "conventional" edge-emitting light-emitting diodes.

For radiation with a wavelength in air of 820 nm, the wavelength within the LED, where $n \approx 3.6$, is only about 230 nm. At a DH spacing in the 500- to 1000-nm range, i.e., several wavelengths, the recombination region between the junction can be treated as an optical resonator that sustains many modes. With such heterojunction spacings, the light-ray picture may be considered to be approximately correct. As the DH spacing is decreased, however, one enters a transition region where a ray picture is no longer truly applicable and where a radiation mode picture must be used.

By going to a DH spacing of under 100 nm, as in our new diodes,[5] the junctions form an optical waveguide in which only a modal description of the fields does justice to the situation. Depending on the index difference across the junction (Δn), the junction spacing (d) can be made such that only one transverse mode can propagate. This requires that,[6] assuming a symmetrical junction,

$$d < \frac{\lambda_0}{2\sqrt{n_1^2 - n_2^2}} \approx \frac{\lambda_0}{2\sqrt{2n\Delta n}}. \quad [1]$$

Δn is roughly proportional to the difference in the aluminum content across the heterojunctions; writing the composition as $(Al_xGa_{1-x})As$,[7]

$$\Delta n \approx 0.6\,\Delta x. \quad [2]$$

Thus, for heterojunctions where $\Delta x = 0.5$, $d < 280$ nm, while for $\Delta x = 0.1$, $d < 620$ nm, for the single-mode case. (Actually, there are two, nearly degenerate modes, one TE and one TM[6]).

In the case of active region thicknesses in this single-mode region, the spontaneous emission that is coupled into this guided mode radiates into a far-field pattern characteristic of the mode. Its properties have been examined in detail in the literature[8,9] for the case of lasers. The radiation pattern perpendicular to the junction plane is the same, whether or not the mode excitation is from spontaneous or stimulated emission. For a thick active region, the lowest order optical wavefront is broad in the transverse direction, and its diffraction pattern is relatively narrow. For $\Delta x > 0.06$, the width of the active region in the single mode regime is less than λ_0, and the beam is still relatively broad, $> 120°$ FWHM. As the thickness d is reduced, the beam widens as the mode width decreases. With still narrower active regions, the optical fields of the mode start to be squeezed out of the active region, and exponentially decaying (evanescent) fields extending past the heterojunctions into the lower-index bounding regions start to increase (in amplitude and in extent), beyond the junctions. The mode width correspondingly begins to broaden, and the radiated beam pattern narrows dramatically. With thicknesses below 100 nm and moderately strong heterojunctions ($\Delta x = 0.5$), beam patterns $< 30°$ FWHM can be achieved.

In the case of injection lasers, where all the stimulated emission can be made to occur in this lowest order transverse mode, it is clear why the observed far-field beam pattern is this narrow. In the case of a spontaneously emitting, incoherent LED, the situation is not so obvious. The radiating electrons can send their radiation into any of the infinite number of modes, nearly all of which are not the guided modes discussed above, but are the usual, quasi-plane wave unguided modes of the ray picture of LED action. One might therefore expect that the radiation into the narrow-beam guided modes would be unobservable in the quasi-Lambertian background of radiation in the unguided modes. This is not the case, as is shown in detail in Ref. [10]. The guided modes, though relatively few in number, have localized field configurations, in which their energy is confined to the region near the radiating (source) electrons. In this case, the spontaneously emitting electrons couple much

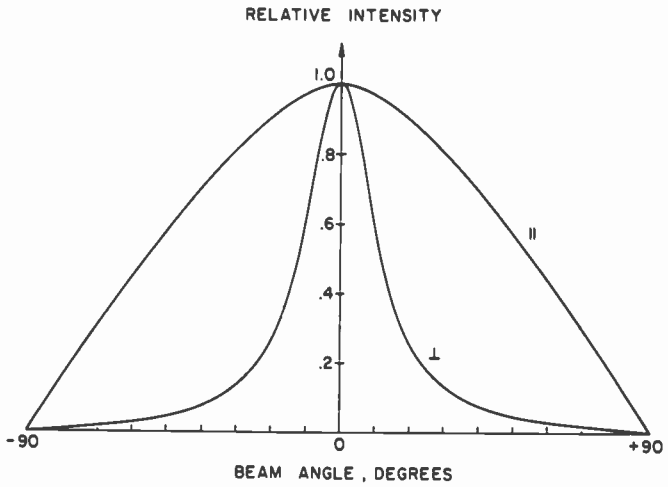

Fig. 2—Far-field radiation pattern for an LED of the new, narrow active region type. Patterns are shown for two directions, in the plane of the junction (‖) and perpendicular to the junction plane (⊥).

more strongly to these guided modes than to the unguided, plane wave modes. The observed result is that a large fraction of the total observed external radiation is confined to the lowest-order mode.

The far-field beam pattern from such a narrow-active-region LED is shown in Fig. 2, where the beam pattern in the plane of the junction is unguided by the heterojunctions and remains Lambertian. Table 1 shows the measured far-field beam widths and theoretical values[5] for a group of LED's covering a range of active-region thicknesses and heterojunction

Table 1—LED Beam Angles

Sample No.	λ (μm)	d (μm)	x_{Al}	θ (degrees)	
				Experimental	Calculated
295	0.88	0.08	0.25	35	26
297	0.88	0.05	0.48	25	31
298	0.82	0.08	0.45	65	50
305	0.81	0.04	0.32	30	18
309	0.81	0.06	0.35	25	17
313	0.83	0.06	0.35	28	29
314	0.81	0.10	0.50	67	70
321	0.81	0.10	0.25	30	35
327	0.80	0.07	0.35	31	35
344	0.87	0.03	0.65	29	25

strengths Δx. The agreement is quite good, considering the inaccuracies attendant on measuring the widths of the active regions, estimated as $\pm 15\%$.

For highest efficiency and longest operating lifetime, it is important to keep the operating temperature in the active region of the diode as low as possible. Since considerable heat is generated in the very small active region, the geometry must be chosen to facilitate heat removal. The most convenient geometry that we have found is an oxide-defined stripe contact similar to that used for our cw laser diodes.[11] Here, the high aluminum-doped p-region adjacent to the active region is kept as thin as possible, consistent with adequate optical confinement (thicker than the optical attenuation distance in this low-index region). Adjacent to this, a thin, highly doped p-GaAs region is used to permit good electrical contacting, and this is then indium-soldered to a copper heat sink. Except for the diode length, which is kept under 200 μm to maximize the edge emission efficiency, these LED's are very similar in their external configuration to cw lasers.

2.2 Output Power and Radiance

The new diodes have high emission efficiency for several reasons. (1) Bimolecular recombination reduces the relative importance of nonradiative recombination processes. (2) The optical structure increases the spontaneous emission rate. (3) The reflectivity at the air–semiconductor interface is lower for the guided mode than for the average ray (mode) striking the interface in a diode with a thick, multimode active region. The result is an external quantum efficiency of $\sim$1% from a single emitting facet.

The light output from one edge of one of our best diodes is shown in Fig. 3. The diode is 150-μm long, has a 65-μm-wide contact, and has reflective and antireflective coatings applied to the back and emitting cleaved facets, respectively. At a drive current of 400 mA, the applied voltage was 1.8 V, giving a power efficiency, under these conditions, of $\sim$1%. The nonlinear behavior of this diode apparent at low drive currents is due to electrical leakage in the (metallic) reflective coating. In general, these diodes exhibit linear output with drive current; this is illustrated, with a pulsed drive, in Fig. 4 for a broad-area diode that does not have a reflective coating. The nonlinear behavior at the higher drive currents is due to the onset of lasing; diode heating would prevent this under cw operating conditions.

We measured the radiance of this diode in two ways—first, by using a previously described[12] lens and pinhole arrangement and, second, by measuring the emission area and the power emitted into a cone of N.A.

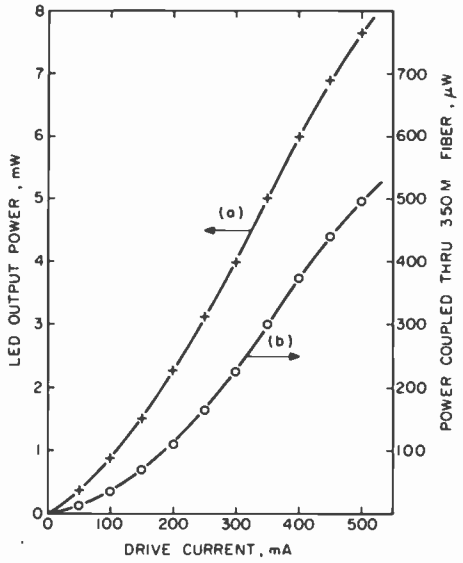

Fig. 3—Total power output from one edge of a stripe-contact LED and power coupled through 350 m of 8 dB/km, N.A. = 0.14, step-index, 90-μm core fiber, as a function of drive current. A melted-lens fiber coupler was used.

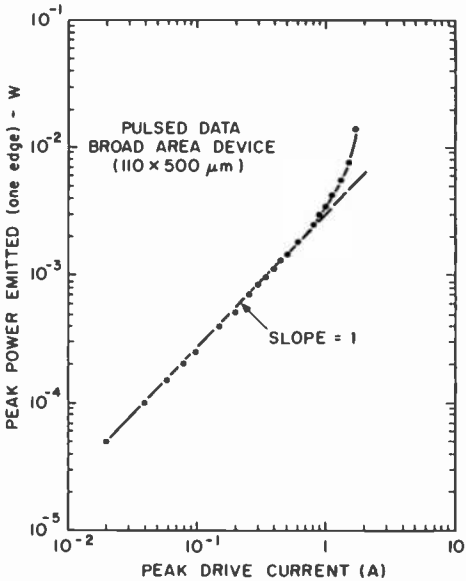

Fig. 4—Peak light output as a function of peak drive current, for a LED from wafer 313 (see Table 1). The pulse length was 200 ns, and the repetition rate was 1 kHz.

= 0.1. For the lens and pinhole measurement, a 16×, N.A. = 0.15 microscope objective lens was used, and the power through a 50-μm-diameter pinhole was 5.7 μW, giving a (peak) source radiance of 1100 W/cm^2-sr. Using the second method, the source size was first determined to be 6 μm × 75 μm (to a −6 dB from peak level) by scanning in the image plane of the lens used in the first method. From this source size, the power collected in a cone of N.A. = 0.1 was 0.25 mW, at a 400-mA dc drive level. This corresponds to a radiance of 1700 W/cm^2-sr. Part of the discrepancy in the two measurements is due to the neglect of losses in the microscope lens. In any case, the source radiance was measured to be over 1000 W/cm^2-sr, about ten times greater than the highest value reported previously.[1,12,13]

Because of the high radiance of this source, large amounts of power can be coupled into small-diameter, low-numerical-aperture fibers with relatively high coupling efficiency. Fig. 3 shows total output power and the power through 0.3 km of a N.A. = 0.14, 90-μm-diameter, low-loss, step-index fiber as a function of drive current. If we assume a 2-dB loss for this fiber length, the maximum power coupled into the fiber is 800 μW, which is nearly an order of magnitude higher than previously reported.[14] The coupling loss is approximately −10 dB, compared with calculated values of −17 dB for Lambertian surface-emitting LED's. A long length of fiber was employed for these measurements, because measurements with relatively short lengths (<10 m) often includes lossy skew modes and power coupled into the cladding.

2.3 LED Coupling to Fiber

An exact calculation of the coupling efficiency of an LED to a fiber is extremely complex, and, in fact, even the concept of "coupling-efficiency" is inexact and generally ill-defined. In the case of a single-mode (cylindrical) fiber, the ratio of optical power flowing in the single (actually, doubly degenerate) propagating mode to total output power from a source can be determined and has practical significance. However, we are here concerned with multimode, step-index fibers that can propagate a large number of modes. For example, for a step-index fiber with a core diameter of 90 μm and N.A. = 0.14, the fiber can sustain about 1165 propagating or bound modes.

In addition to these modes, however, energy can sometimes be transmitted down a fiber in two other types of modes: leaky modes and

cladding modes. In an ideal, lossless fiber, all these modes are not truly bound and suffer attenuation as they propagate. With suitable provisions for mode-stripping, the cladding modes can be attenuated to negligible levels in a very short distance, and hence need not be considered. The leaky modes, on the other hand, which correspond to skew rays propagating in a helical fashion down the fiber, can, under favorable circumstances, travel significant distances without appreciable loss. Because of them, the definition of "coupling efficiency" becomes somewhat arbitrary. Many methods that are often used to measure fiber coupling will contain a significant contribution from such modes, especially if a relatively short length of fiber is used in the measurement, and despite attempts to limit the measured transmission to bound modes by limiting the aperture of the beam from the end of the fiber that is coupled into the detector.

Nevertheless, in many systems, the transmission distance is such that energy from the source reaches the detector via these modes, and, practically, these modes should be treated as truly propagating ones. Therefore, one should use care in using the term "coupling efficiency," to make clear just what is its operational significance.

The simplest coupling scheme is to put the end of the fiber against, or in proximity to, the emitting surface. Colvin[15] and others[16] have made calculations of coupling between surface emitters with various characteristics and step-index fibers. The calculations are based on the assumption that geometrical (ray) optics are valid. Colvin's work, in particular, indicates the importance of skew, leaky rays, and shows "that for small W/S ratios (core-to-source diameter ratio), small fiber–LED separations, and low numerical apertures, ray skewing makes a significant contribution to coupling efficiency."[15]

Considering the complexity of the problem, it becomes clear that various approximate simplified models can be useful in estimating coupled power levels to be expected under various conditions, but that detailed measurements will probably provide the only reliable values. One simple model for step-index fibers assumes that all light rays are propagated that are incident on the exposed end of the core and that make a (meridional) angle with the fiber axis less than the critical angle θ_c,

$$\theta_c = \sin^{-1}(\sqrt{n_1^2 - n_2^2}) = \sin^{-1}(N.A.), \qquad [3]$$

while light rays making an angle $\theta > \theta_c$ are rejected. This gives, for a source smaller than the core and in contact with the fiber end and assuming cylindrical symmetry, a coupling efficiency η of

$$\eta = \frac{\int_0^{\theta_c} I(\theta) \sin\theta \, d\theta,}{\int_0^{\pi/2} I(\theta) \sin\theta \, d\theta}. \qquad [4]$$

For a Lambertian source, $I(\theta) = I_0 \cos\theta$, giving

$$\eta = \sin^2 \theta_c = (\text{N.A.})^2. \qquad [5]$$

For a fiber with N.A. = 0.14, this gives 2.0%, or −17.1 dB, coupling efficiency.

Numerical techniques have been employed to evaluate the integrals in Eq. [4] for edge-emitting diodes. Results are shown in Fig. 5 for a diode of the new structure (LED 297) and for a diode of "conventional" char-

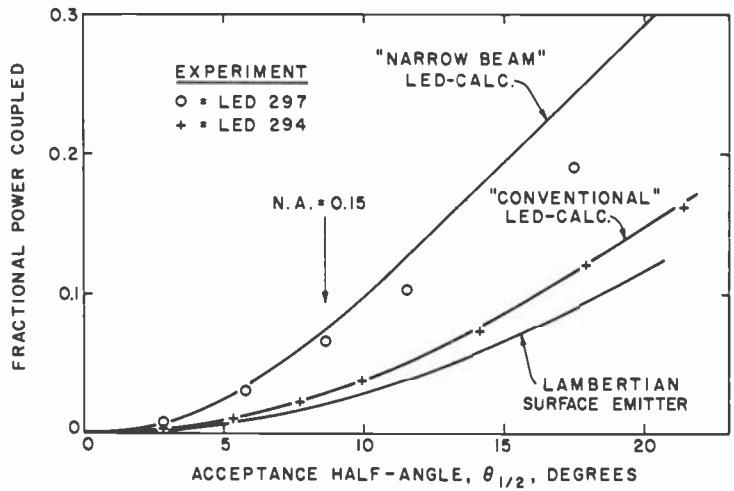

Fig. 5—Calculated and measured coupling efficiencies from three diode types into cones of various numerical apertures.

acteristics (LED 294). Also shown are measurements made on fractional power coupled into a detector subtending various solid cone angles (numerical apertures). For N.A. = 0.14, the predicted best LED coupling is 7.4% (−11.3 dB), over 5 dB better than for a Lambertian source.

There is a small Fresnel reflection at the air–fiber interface when an LED is brought close to the fiber. This can be reduced by index matching, but the gain to be achieved in this way is negligible. The Fresnel reflection at the air–fiber interface for a fiber of N.A. = 0.14 is only 4% (−0.17

dB) for the worst-case ray (θ_c). (We coat the emitting facets of the LED with a $\lambda/4$ antireflecting film, to "match" them to radiate into air. This is necessary to avoid the sizable (>30%, ~1.7 dB) loss that could otherwise occur at the (AlGa)As–air interface.)

Rather than the ~ 0.1-dB gain obtainable by index-matching the LED to the fiber, much larger (>3 dB) gains can be achieved by using a lens to collect and focus the LED radiation into the fiber core. Several forms of lenses have been suggested—spherical,[17,18] cylindrical,[18,19] and hyperbolic.[20] However, as pointed out by Kato,[17] surface tension can be easily used to form a spherical lens onto the (melted) end of a fiber, while much more complex polishing or photoresist techniques are required to form other lens shapes. We, therefore, concentrated our efforts on spherical lenses formed by melting the end of the fiber under controlled conditions. The fibers used are high-silica ones in which the melting temperatures of core and cladding are apparently almost equal. Following Kato,[17] we used an oxy-hydrogen torch to melt the fiber end. The radius of the spherical lens surface could be controlled by adjusting the volume of melted fiber. Lenses formed in this way are shown in Fig. 6.

The effects of one such (optimal) lens are illustrated in Fig. 7. It should be emphasized that the powers (and coupling values) shown here are after transmission down 350 m of fiber. Our best estimate of the fiber losses in this length are 2 dB (~ 6 dB/km) at 820 mm. Thus, the power coupled into the end of the fiber into modes that propagate at least 350 m is about 2 dB (a factor of 1.6) higher than shown; similarly, the "coupling efficiency" into the propagating fiber modes is better than −10 dB, more than a factor of 5 better than available with a Lambertian surface emitter. The illustration clearly shows that the melted fiber lens increases the power coupled by about 4 dB (a factor of 2.5).

2.4 Frequency Response

The frequency response of an LED is determined by two major factors: the diode junction space-charge capacitance[4] and the recombination time of the injected carriers.[21] In a surface-emitting diode, the geometry generally leads to a large junction capacitance[4] and correspondingly low modulation response speeds. In an edge-emitting configuration, on the other hand, the junction capacitance can be greatly reduced, and, as a result, the carrier recombination time determines the diode speed.[21] In the case where the doping level in the active region determines the recombination time, the relative response is given by

$$R(\omega) = \frac{1}{\sqrt{1 + (\omega\tau)^2}}, \qquad [6]$$

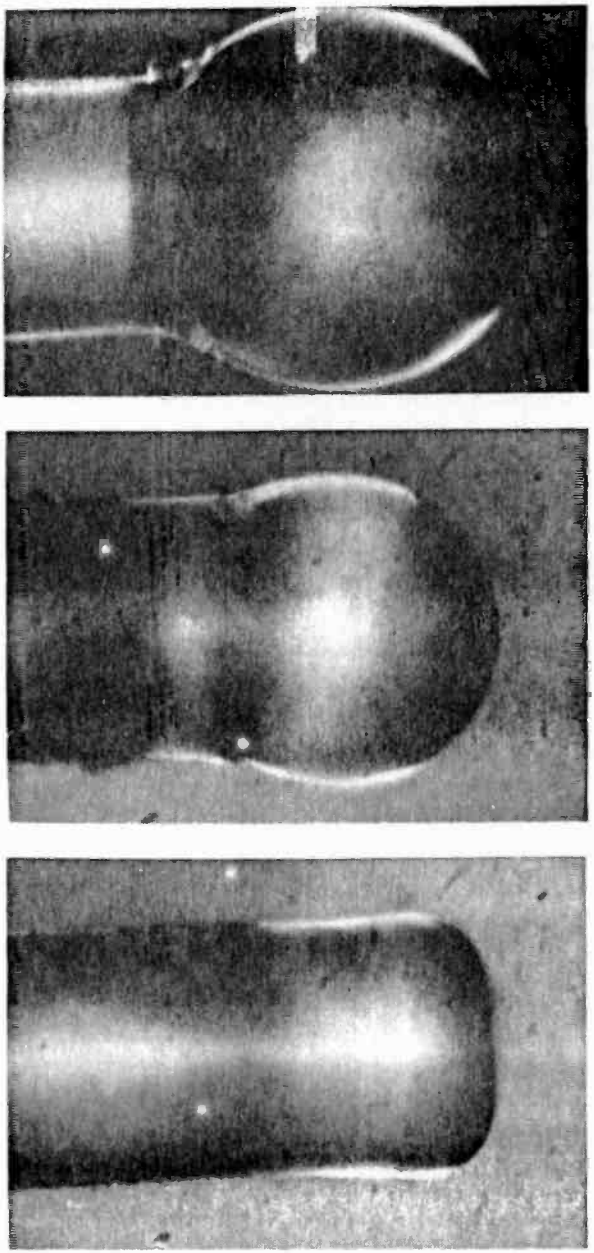

Fig. 6—Photographs of lenses melted onto the ends of high-silica optical fibers. By controlling the melt conditions, different focal length lenses can be made. Fiber diameter (over cladding) is 125 μm.

where ω is the (circular) modulation frequency and τ is the recombination time. This expression is valid for both homojunction[22] and double heterojunction[23] LED's. The recombination time can be shortened by increasing the active region doping. In previous work,[23] we found that

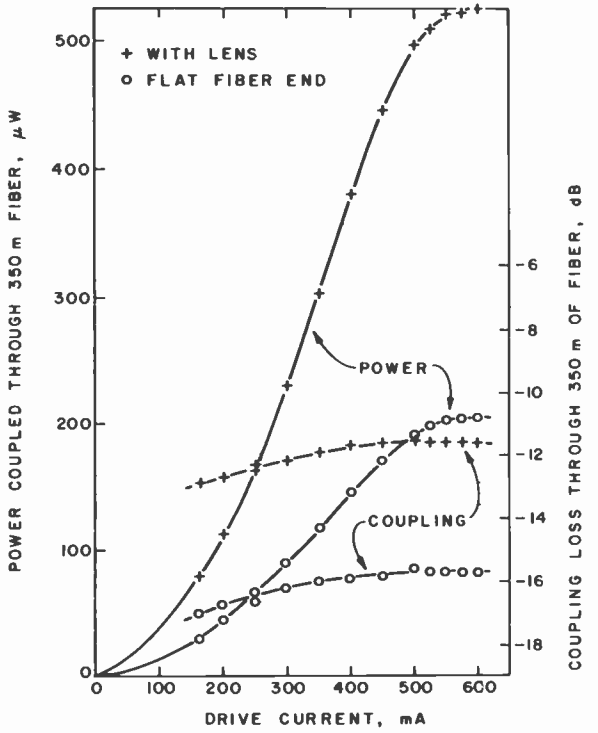

Fig. 7—Effect of melting an optimal lens (see Fig. 6(b)) onto a silica fiber of N.A. = 0.14. The lens increases the power coupled into the fiber by a factor of about 2.5 (+4 dB).

diodes doped with germanium to a density of 2×10^{19} cm^{-3} gave a lifetime of $\tau = 1.0$ ns, corresponding to a 3-dB (electrical signal power) frequency of about 160 MHz. To obtain this high speed, however, one paid the price of lower diode emission efficiency because of the relatively high internal absorption in the heavily doped active regions.

With our new, narrow-active-region structure, the injected carrier density can be high enough that "bimolecular" recombination dominates,

and the radiative lifetime can be very short, even in the absence of heavy doping. This has been discussed by Namizaki et al.[24] The total recombination rate is due to recombination of the electrons injected into the p-type active-region with both the normal majority carriers present in the absence of injection and the injected holes required to maintain charge-neutrality:

$$\frac{1}{\tau} = Bp_0 + B\Delta n. \qquad [7]$$

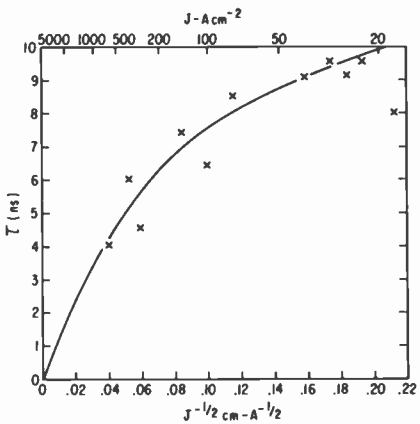

Fig. 8—Lifetime as a function of the inverse square root of the injection current density for a narrow-active-region DH LED with $d = 0.1$ μm and $p_0 = 10^{18}$ cm^{-3}. The active region was GaAs and the surrounding material was $Al_{.5}Ga_{.5}As$. The measured data give the times after the current is cut off for the recombination radiation to decay to I/e of its initial value. They thus represent an average bimolecular recombination effect.

The injected carrier density, $\Delta n = \Delta p$, is given, in a steady-state condition, by

$$\Delta n = \frac{J\tau}{ed}, \qquad [8]$$

where J is the injected current density, e is the electronic charge, and d is the thickness of the active region (between mobile-charge confining heterojunctions). Thus,

$$\frac{1}{\tau} = Bp_0 + \frac{BJ\tau}{ed}, \qquad [9]$$

or

$$\tau = \frac{edp_0}{2J}\left(\sqrt{1 + \frac{4J}{edp_0{}^2B}} - 1\right). \qquad [10]$$

Eq. [10] is plotted in Fig. 8, together with measured lifetimes of one of our new diodes. The doping in the active layer was $p_0 = 10^{18}$ cm^{-3}, the active region thickness was $d = 0.1$ μm, and the Namizaki et al value of the band-to-band recombination parameter, $B = 9 \times 10^{-11}$ cm^3/s, was used. (Interfacial recombination that might be occurring at the heterojunctions is neglected.) The agreement is quite good. The relatively low doping level in the active region leads to higher external radiative efficiency than found for diodes with heavy ($> 10^{19}$ cm^{-3}) doping in the active region.

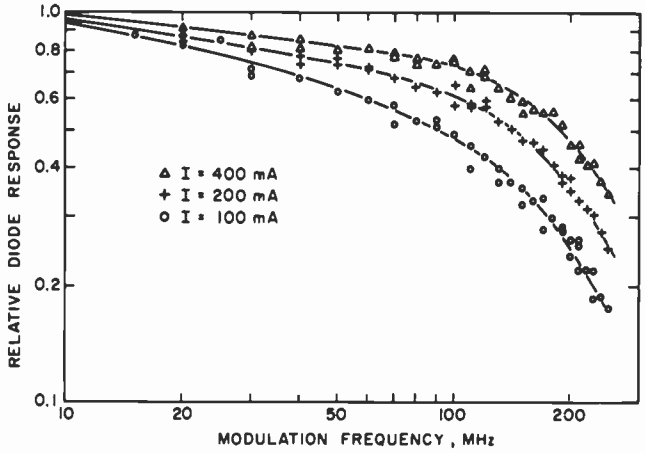

Fig. 9—Frequency response of a narrow active region LED to small modulated drive currents, at different dc bias current levels. An approximate $\tau \propto J^{-1/2}$ dependence of the response times is indicated by the data. The relative electrical signal amplitude is shown; the −3dB signal power point is at a relative response of 0.707.

One important consequence is that the response speed of these diodes is dependent on the drive current density, varying as $J^{-1/2}$ for high drive levels. This is shown in Fig. 9 for a narrow active region LED; the frequency response is seen to vary roughly as $J^{-1/2}$ in the range of current densities shown. However, a detailed examination of Fig. 9 shows that the data are not well fit by the expression above (Eq. [6]), and a simple recombination time does not accurately represent the diode performance.

2.5 Spectral Width and Fiber Dispersion

The doping level in the active region affects not only the recombination rate and, hence, the response time, but also the LED efficiency and the spectrum of the emitted radiation. This is shown in Table 2, which is

taken from previous work.[23] The values shown are representative of trends, rather than limits for any doping type. Silicon-doped diodes are the most efficient, but also have the longest recombination times. This is to be expected, since silicon is amphoteric in (AlGa)As and forms deep centers and bandtail states from which much of the emission originates. Germanium, at high doping levels, gives the shortest recombination times, but diode efficiency suffers somewhat at the required doping

Table 2—Characteristics of Various Double-Heterojunction LED's

Doping Type	Light Output* (μW)	Peak of Emission (nm)	Modulation Capability (MHz)†
Si	800	930	10
Si	600	880	—
Si	250	800	—
Sn	200	900	70
Sn	200	870	—
Ge	300	900	250
Ge	250	850	90
Ge	200	820	70
Unintentional doping	200	800	60
Unintentional doping	350	800	—

*From a narrow facet of a 400-μm × 100-μm LED at 100-mA drive current.
†Frequency at which the depth of optical intensity modulation falls to one-half of that at very low modulation frequencies when the depth of drive current modulation is kept constant (and small). This corresponds to a —6dB signal power level.

levels. With the new narrow-active-region structure, these differences for different dopants become less important, as bimolecular recombination begins to dominate the recombination process, controlling both diode speed and efficiency.

However, the doping type and level will still determine, to a large extent, the spectral width of the emission. This is illustrated in Figs. 10 through 12, where the spectra of three narrow-active-region LED's are shown. Silicon doping in the active regions is seen to broaden the spectrum to over 60 nm, while diodes that have no deliberate doping in their active regions have considerably narrower emission spectra.

These differences in spectral width can have an important effect on system bandwidth, as shown in Fig. 13, where the frequency responses measured with different sources are shown for 1 km of multimode graded-index fiber. The fiber index was graded to reduce pulse broadening. The frequency response of the fiber was measured (by the manufacturer, Corning Glass Co.) using a pulsed laser, under conditions in

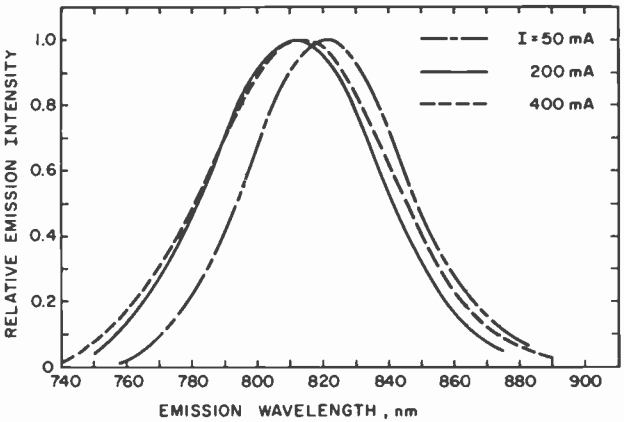

Fig. 10—Emission spectrum of diode 313DH at several drive currents. The active region of the diode is doped with silicon.

which broadening due to material dispersion is negligible. However, with spectrally broad sources, such as LED's, material dispersion can be an important cause of bandwidth limitation, as shown. The broadest LED has a spectral width of 60.5 nm (at 200 mA; see Fig. 14) and despite the (nominal) 180-MHz fiber bandwidth (3-dB electrical signal point), as measured with the broad LED sources the bandwidth is only 45 MHz. With LED's having narrower spectra (Figs. 11 and 12), the measured

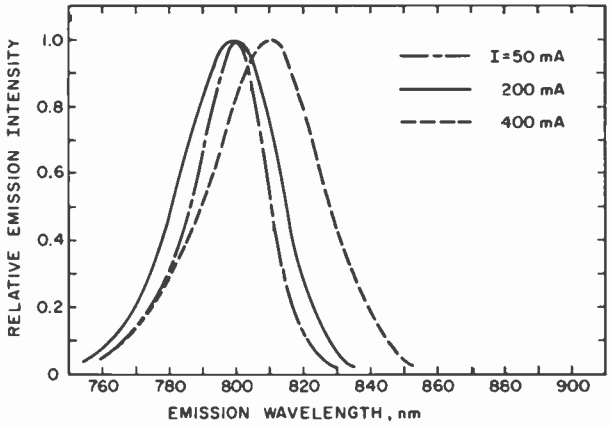

Fig. 11—Emission spectrum of diode 331DH at several drive currents. The active region of the diode is not deliberately doped, although there is evidence of some unintentional doping.

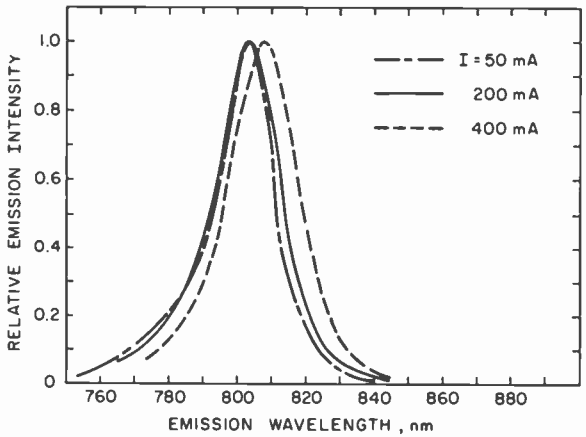

Fig. 12—Emission spectrum of diode 332DH at several drive currents. The active region is not deliberately doped.

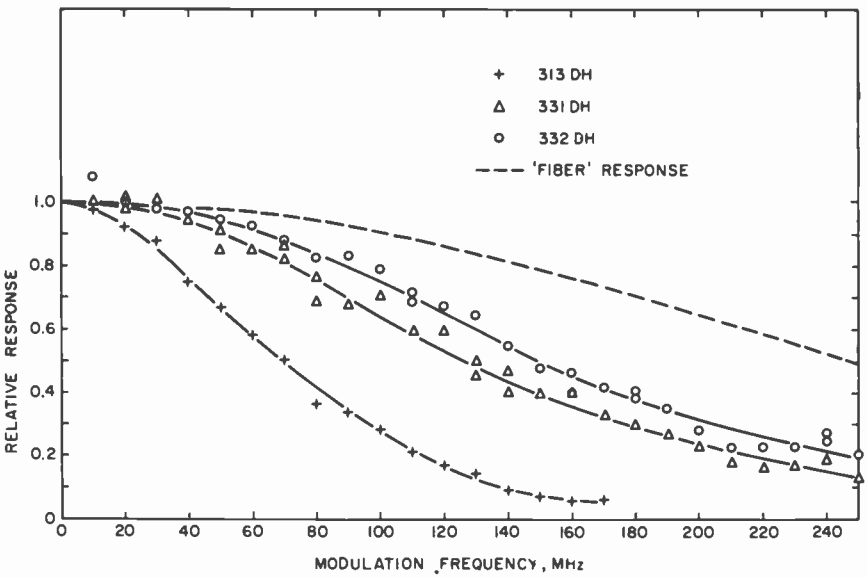

Fig. 13—Frequency response of 1 km of graded-index, multimode fiber using LED sources of various spectral widths. The dashed curve shows the manufacturer's quoted frequency response characteristic. The relative electrical signal amplitude is shown; the −3 dB signal power point is at a relative response of 0.707.

fiber response extends to significantly higher frequencies, as shown in the figure and in Table 3. The pulse broadening due to material broadening has been discussed by Dyott and Stern,[25] and can be written

$$\tau_{\text{Mat'l Disp.}} = \frac{L}{c}\frac{\Delta\lambda}{\lambda}\cdot\lambda^2\frac{d^2n}{d\lambda^2}. \qquad [11]$$

Here L is the fiber length, $\Delta\lambda/\lambda$ is the fractional spectral width of the source, and $\lambda^2 d^2n/d\lambda^2$ is the material dispersion coefficient for the fiber material. For the silica-rich fiber of Fig. 13, the value of the coefficient is 0.028.[26]

Table 3—Modulation Bandwidths of LED–Graded-Index Fiber System (All Frequency Responses Correspond to a −6 dB Signal Power Level)

	LED		
	313DH	331DH	332DH
Spectral Width ($\Delta\lambda$, nm)	60.5	34.0	20.0
Wavelength (λ, nm)	812	800	803
Calculated* Frequency Response due to Mat'l Dispersion (MHz)	64	110	192
Calculated Overall† Frequency Response (MHz)	62	101	152
Measured Frequency Response (MHz)	70	127	151

*Calculated using Eq. [11].
†Calculated assuming Eq. [12].

How to quantitatively include the effects of (residual) multi-mode dispersion in the fiber is not obvious; in Table 3, we make the simplifying assumption that pulse broadening mechanisms can be superposed as

$$\tau^2 = \tau^2_{\text{Mat'l Disp.}} + \tau^2_{\text{Modal Disp.}}, \qquad [12]$$

where τ is an "overall" response time for the fiber-LED system and $\tau_{\text{Mat'l Disp.}}$ and $\tau_{\text{Modal Disp.}}$ are the corresponding times for material and multi-modal dispersions, respectively. As shown in Table 3, the agreement between modulation bandwidths calculated in this way and those measured (Fig. 13) is quite good.

2.6 Diode Packaging

The LED packaging must satisfy three basic requirements. It must (1) provide protection against the environment (mechanical protection,

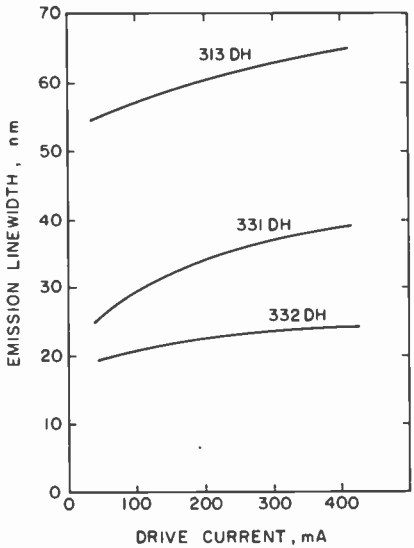

Fig. 14—Broadening of emission spectra as drive current is increased, for three narrow active region LED's.

shielding against humidity and other atmospheric effects), (2) provide for adequate heat removal, and (3) provide and maintain alignment of fiber and LED for optimal coupling. We have developed a prototype LED package that seems to satisfy all three requirements. A cross-sectional view of the package is shown in Fig. 15.

The LED is indium-soldered onto the edge of the hemicylindrical copper heat sink. The fiber, with a focussing lens melted onto its end (see

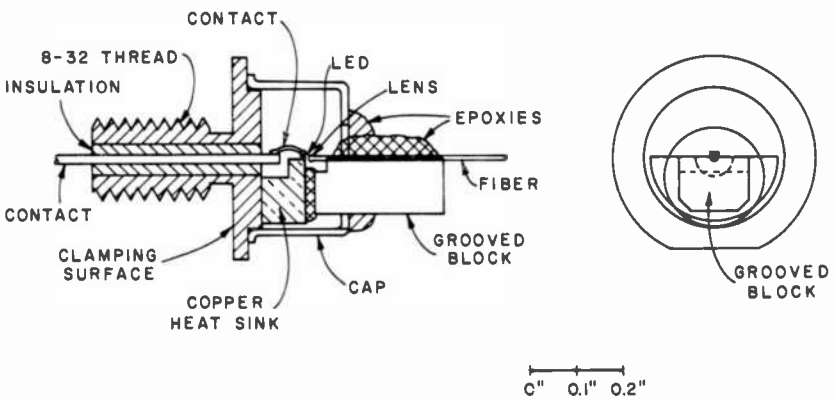

Fig. 15—Cross section of LED–fiber mounting package.

Section 2.3 above) is epoxied onto a small brass mounting block. The block with fiber attached is then positioned to optimize coupling to the (operating) LED and epoxied to the heat sink, forming a mechanically strong connection that maintains the proper alignment between LED and fiber. For protective purposes, a metal cap with an (eccentric) hole in its top is mounted on the basic threaded mount, as shown. This is held in place, and the remaining openings sealed, completing the package. The result is a small, rugged package, with the LED permanently bonded to and aligned with the fiber mount. The entire LED-fiber package can be readily screwed into a system. If necessary, it is an easy matter to replace the LED in the system.

2.7 LED Reliability

Previous work has identified many of the essential diode parameters that affect the reliability at high current densities ($\geq$1000 A/cm^2). These

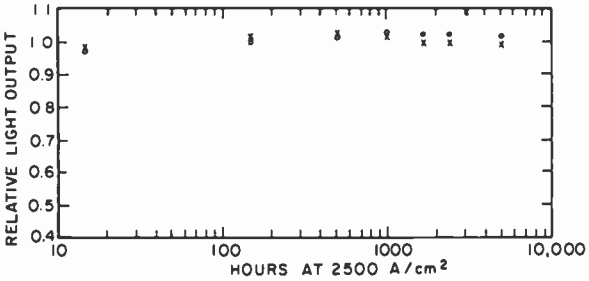

Fig. 16—Light output as a function of time for oxide-defined stripe-contact DH LED with an $Al_{.1}Ga_{.9}As$ active region ~50 nm thick. Diodes were operated continuously at 2500 A/cm^2 in a laboratory ambient.

include proper soldering and assembly techniques to minimize the introduction of defects, low contamination levels, the use of low dislocation substrates, and the removal of work-damaged regions from the recombination region.[27] It is further known that the use of $Al_{.1}Ga_{.9}As$ in the recombination region is beneficial.[28] Double heterojunction diodes with the narrow active region structure have exhibited excellent operating stability in a normal laboratory ambient (not hermetically sealed), as shown in Fig. 16.

Earlier diodes, with wider recombination regions, have been operated for periods of time exceeding 14,000 hours at 1000 A/cm^2, as shown in Fig. 17. The slight increase in the output of the diodes with time is a real

effect, possibly related to the slow annealing of nonradiative centers in the active region of the diode.

3. Conclusions

A new type of edge-emitting LED, with high speed and very high radiance, has been developed. The injected-carrier recombination is forced to occur in a very narrow region, about 50 to 100-nm thick, by means of closely spaced (AlGa)As heterojunctions with a large bandgap differ-

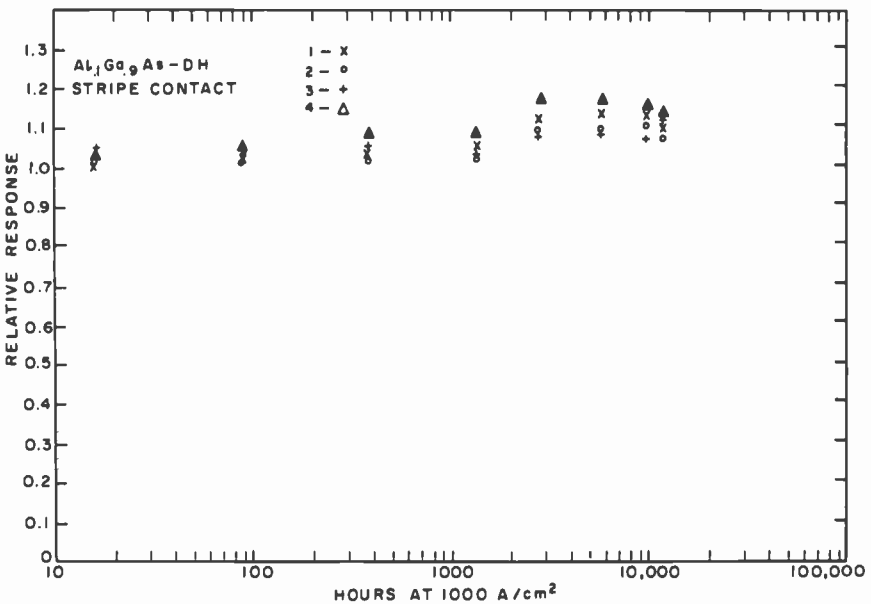

Fig. 17—Light output as a function of time for oxide-defined stripe-contact DH LEDs which have $Al_{.1}Ga_{.9}As$ recombination regions. Diodes were operated continuously at 1000 A/cm² in a laboratory ambient.

ential across them. This structure has several advantages. Because the recombination region is so small, a significant fraction of the recombination is bimolecular even at moderate injection currents. This leads to good efficiency and short recombination time, resulting in diodes with very high-speed modulation characteristics.

In addition to the high-efficiency and fast-response characteristics, the very narrow active region also acts as an optical waveguide, further increasing the efficiency and drastically narrowing the beam angle

(perpendicular to the junction plane) of the emitted radiation. This has resulted in LED radiances more than ten times higher than any previously reported in the literature. Values of over 1000 W/cm^2-sr are readily achieved with this new structure.

Using a simple spherical lens melted onto the end of the fiber to improve coupling efficiency between LED and fiber, we have coupled up to 800 μW into the propagating modes of a single low-loss, step-index, N.A. = 0.14, 90-μm-diameter core fiber.

A method of permanently attaching a short length of such fiber (with spherical-lens end) to the LED in a single sealed package that can readily be inserted into a system has also been developed.

With the new LED structure, high modulation speeds are attainable regardless of the active region dopant. Thus, this can be selected solely to optimize other characteristics. In the past, we have used heavy germanium doping to obtain high speeds, but at some sacrifice in external efficiency. This is no longer required, and lower doping levels (or even no intentional doping) can be used to obtain high output efficiency. Extrapolation from our previous work on LED's and the life of diodes with comparable structures indicate that they should have operating lives considerably in excess of 10,000 hours.

In considering the use of LED's for very-high-bandwidth data transmissions over fibers, the effects of material dispersion due to the broad spectral widths of LED's must be considered. Indeed, this factor can be the limiting one in overall systems performance, as will now be shown.

Consider two variants of the new type of LED— one with an undoped active region (LED 347DH) and one with significant germanium doping in the active region (LED 309DH). Both have a 3-dB (electrical signal power) point of about 150 MHz; the unintentionally doped LED has a spectral width of 34 nm, while the germanium-doped one has a 41-nm width. From Fig. 13, a high-quality, graded-index fiber of a nominal 180-MHz-km pulse dispersion has, because of material dispersion, actual responses of 90 MHz-km and 75 MHz-km for the two diodes. Thus, data occupying 150 MHz of bandwidth can only be transmitted about 0.5 to 0.6 km without suffering significant interference. The power losses in such a short length of low-loss (5 dB/km) fiber are under 3 dB. Under such circumstances, 1 mW of power coupled into the fiber would yield roughly 500 μW at the detector. Even 100 μW of coupled power would give 50 μW at the detector.

The major source of noise at power levels of this sort, assuming that an avalanche-diode detector with reasonable gain (>30) is used, is shot noise in the detected photoelectron current (plus, of course, the noise added in the stochastic avalanche multiplication process).[29] The noise

added by the amplifier stages following the detector can be made unimportant. Under these conditions, 1 mW of power into the fiber gives a +55 dB S/N ratio for a 150-MHz-bandwidth data stream, and even 100 μW into the fiber gives S/N = 45 dB. Since essentially error-free digital data transmission can be obtained with S/N = 25 dB,[30] it is seen that, for such broad-band signals, the material dispersion of the fiber and not the available LED power limits the transmission capabilities of the system.

References:

[1] C. A. Burrus, "Radiance of Small-Area High-Current-Density Electroluminescent Diodes," *Proc. IEEE*, **60**, p. 231 (1972).

[2] E. D. King, A. J. SpringThorpe, and O. I. Szentesi, "High-Power Long-Lived Double Heterostructure LEDs for Optical Communications," Int. Elect. Devices Meeting, 1975, Wash., D. C., Tech. Digest Paper 21.2, p. 480.

[3] M. Ettenberg, H. F. Lockwood, J. P. Wittke, and H. Kressel, "High Radiance, High Speed $Al_xGa_{1-x}As$ Heterojunction Diodes for Optical Communications," Int. Elect. Devices Meeting, 1973, Wash., D. C., Tech. Digest Paper 16.1, p. 317.

[4] C. A. Burrus, T. P. Lee, and W. S. Holden, "Direct-Modulation Efficiency of LEDs for Optical Fiber Transmission Applications," *Proc. IEEE*, **63**, p. 329 (1975).

[5] M. Ettenberg, H. Kressel, and J. P. Wittke, "Very High Radiance Edge-Emitting LED," *IEEE J. Quant. Elect.*, **QE-12**, p. 360, 1976.

[6] R. E. Collin, *Field Theory of Guided Waves*, McGraw-Hill Book Co., New York, (1960), Sect. 11.5.

[7] H. Kressel, H. F. Lockwood, and J. K. Butler, "Measurements of Refractive Index Step and of Carrier Confinement at (AlGa)As-GaAs Heterojunctions," *J. Appl. Phys.*, **44**, p. 4095 (1973).

[8] H. Kressel, J. K. Butler, F. Z. Hawrylo, H. F. Lockwood, and M. Ettenberg, "Mode Guiding in Symmetrical (AlGa)As-GaAs Heterojunction Lasers with Very Narrow Active Regions," *RCA Rev.*, **32**, p. 393 (1971).

[9] W. P. Dumke, "The Angular Beam Divergence in Double-Heterojunction Lasers with Very Thin Active Regions," *IEEE J. Quant. Elect.*, **QE-11**, p. 400 (1975).

[10] J. P. Wittke, "Spontaneous-Emission-Rate Alteration by Dielectric and Other Waveguiding Structures," *RCA Rev.*, **36**, p. 655 (1975).

[11] I. Ladany, J. P. Wittke, and H. Kressel, *Injection Lasers for High Data Rate Optical Communications*, Final Report, Vol. 2, ONR Contract N00014-73-C-0335, October 1975.

[12] M. Ettenberg, K. C. Hudson, and H. F. Lockwood, "High-Radiance Light-Emitting Diodes," *IEEE J. Quant. Elect.*, **QE-9**, p. 987 (1973).

[13] E. D. King and A. J. SpringThorpe, "The Integral Lens Coupled LED," *J. Elect. Mat.*, **4**, p. 243 (1975).

[14] H. Kressel and M. Ettenberg, "A New Edge-Emitting (AlGa)As Heterojunction LED for Fiber-Optic Communications," *Proc. IEEE* ,**63**, p. 1360 (1975).

[15] J. Colvin, "Coupling (Launching) Efficiency for a Light-Emitting Diode, Optical Fibre Termination," *Opto-Elect.*, **6**, p. 387 (1974).

[16] C. C. Timmermann, "Launching Efficiency of Incoherent Light Coupled into Optical Fibres," *Arch. f. Elekt. u. Uebertrag*, **27**, p. 150 (1973); D. A. Barkowitz and E. Tammaru, "Collection Efficiency of Receivers Placed Close to Sources," *J. Opt. Soc. Am.*, **53**, p. 1292 (1963).

[17] D. Kato, "Light Coupling from a Stripe-Geometry GaAs Diode Laser Into an Optical Fiber with Spherical End," *J. Appl. Phys.*, **44**, p. 2756 (1973).

[18] L. G. Cohen and M. V. Schneider, "Microlenses for Coupling Junction Lasers to Optical Fibers," *Appl. Opt.*, **13**, p. 89 (1974).

[19] L. G. Cohen, "Power Coupling from GaAs Injection Lasers into Optical Fibers," *Bell Syst. Tech. J.*, **51**, p. 573 (1972).

[20] K. Kurokawa and E. E. Becker, "Laser Fiber Coupling with a Hyperbolic Lens," *IEEE Trans. Microwave Theory Tech.*, **MTT-23**, p. 309 (1975).

[21] H. F. Lockwood, J. P. Wittke and M. Ettenberg, "LED for High Data Rate, Optical Communications," *Opt. Comm.*, **16**, p. 193 (1976).

[22] Y. S. Liu and D. A. Smith, "The Frequency Response of an Amplitude-Modulated GaAs Luminescence Diode," *Proc. IEEE*, **63**, p. 542 (1975).

[23] M. Ettenberg, J. P. Wittke, and H. Kressel, *High-Speed Light-Emitting Diodes*, Final Report, Vol. 1, ONR Contract N00014-73-C-0335, June 1975.

[24] H. Namizaki, H. Kan, M. Ishii, and A. Ito, "Current Dependence of Spontaneous Carrier Lifetimes in GaAs-$Ga_{1-x}Al_xAs$ Double-Heterostructure Lasers," *Appl. Phys. Letters*, **24**, p. 486 (1974).

[25] R. B. Dyott and J. R. Stern, "Group Delay in Glass-Fibre Waveguide," *Elect Letters*, **7**, p. 82 (1971).

[26] D. Gloge, "Dispersion in Weakly Guiding Fibers," *Appl. Opt.*, **10**, p. 2442 (1971).

[27] I. Ladany and H. Kressel, "Influence of Device Fabrication Parameters on Gradual Degradation of (AlGa)As CW Laser Diodes," *Appl. Phys. Letters*, **25**, p. 708 (1974).

[28] M. Ettenberg, H. Kressel and H. F. Lockwood, "Degradation of $Al_xGa_{1-x}As$ Heterojunction Electroluminescent Devices," *Appl. Phys. Letters*, **25**, p. 82 (1974).

[29] J. P. Wittke, "Optical Fiber Communications Link Design," *Guided Optical Communications, Proc. S.P.I.E.*, **63**, p. 58 (1975).

[30] M. Schwartz, *Information Transmission, Modulation, and Noise*, McGraw-Hill Book Co., New York (1959), 2nd Ed.

Iron-Doped Lithium Niobate as a Read-Write Holographic Storage Medium*

G. A. Alphonse and W. Phillips

RCA Laboratories, Princeton, N.J. 08540

Abstract—The response of iron-doped lithium niobate under conditions corresponding to hologram storage and retrieval is described, and the material characteristics are discussed. The optical sensitivity can be improved by heavy chemical reduction of lightly doped crystals such that most of the iron is in the divalent state, the remaining part being trivalent. The best reduction process found to be reproducible so far is the anneal of the doped crystal in the presence of a salt such as lithium carbonate. It is shown by analysis and simulation that a page-oriented read–write holographic memory with 10^3 bits per page would have a cycle time of about 60 ms and a signal-to-noise ratio of 27 dB. This cycle time, although still too long for a practical memory, represents an improvement of two orders of magnitude over that of previous laboratory prototypes using a thermoplastic storage medium.

1. Introduction

The work at RCA and elsewhere on read–write holographic memory has shown that the storage medium poses the major problem in the realization of a system operating in the submillisecond cycle time. An experimental memory was built by RCA[4] in which the storage medium was thermoplastic. Although the sensitivity of thermoplastic is high, of the order of 0.2 μJ/mm^2 for 1% diffraction efficiency, the memory cycle time is of the order of a few seconds, due to the required processing (charging

* This work was partially supported by the National Aeronautics and Space Administration, George C. Marshall Space Flight Center, Huntsville, Alabama and the RCA Laboratories, Princeton, New Jersey under Contract NAS8-26808.

and thermal fixing). A more suitable storage medium is iron-doped lithium niobate. Its sensitivity has been improved considerably by means of a process that involves light doping of the material followed by heavy chemical reduction.[5] The crystal is easy to grow and to handle, and the chemical treatment process seems to be reproducible.

To determine the usefulness of $LiNbO_3$:Fe as a storage medium, it is necessary to establish its characteristics. To do this, we have developed a preparation process that maximizes the sensitivity, and have refined that process to achieve reproducibility. We have also made a theoretical and experimental study of the time development of holographic storage and erasure, its dependence on beam modulation and recording angle, and its noise characteristics.

The first part of this paper considers the material characterization per se, including a determination of the time response of the material to holographic stimulation and a discussion of the sensitivity and of the methods used to improve it, of the dependence of the holographic efficiency on recording angle, and of the noise characteristics. In the second part, we discuss the systems requirements and limitations and use the result of the time-response analysis and the noise measurements to predict the performance of the read–write holographic storage system using iron-doped $LiNbO_3$. The predicted performance is verified experimentally by means of measurements in a simulated read–write memory environment.

This work demonstrates that the sensitivity of lithium niobate doped with iron has been increased reproducibly by a factor of about 40, and that the noise characteristics have also been improved. It also indicates that a holographic read–write memory with a cycle time of about 60 ms and a high signal-to-noise ratio is now feasible with this material. Although this cycle time is still too low for a practical system, it does represent a two-order-of-magnitude improvement over earlier prototypes.[4]

2. Materials Characterization

2.1 Time Dependence of Holographic Storage and Erasure

To characterize the iron-doped lithium niobate ($LiNbO_3$:Fe) samples, some important questions must be answered. For a given holographic illumination, how much time does it take for the diffraction efficiency to reach a certain value? How does that time depend on the illumination level? What is the time constant for erasure? What determines the material optical sensitivity? What parameters must be manipulated chemically in material preparation to improve that sensitivity? The answers to these questions lie in the rigorous analysis of the complete

space–time development of holograms under writing and erasing illumination.

When exposed to light, a number of electro-optic, ferroelectric, or photorefractive materials exhibit refractive index changes.[6] These changes are attributed to electro-optic effects due to the storage of electric fields[7] in the material as a result of the optical exposure. For very intense light (of the order of several W/mm^2) the stored electric field is the combined result of migration and macroscopic polarization changes in the material due to nonlinear absorption.[8] For light of moderate intensities (less than a few hundred mW/mm^2) the mechanism of electric field storage is that of space-charge migration.[7,9] It is this space-charge model that applies to holographic storage in read–write memory systems of the type under study in this work. In this model electrons are excited from impurity traps by the incident radiation and, upon migration, are retrapped at other locations, giving rise to a frozen-in electric field. The latter induces refractive index changes via the electro-optic effect. In undoped crystals, the traps are provided by small traces of impurities. In doped crystals, the dopants act as donor–acceptor traps via intervalence exchanges such as $Fe^{2+} \rightleftharpoons Fe^{3+}$ in $LiNbO_3$:Fe (Refs. [5], [10]). In general, the migration of charges occurs under the combined influence of diffusion and an electric field. The electric field may be internally generated or externally applied. Although the space-charge field can have arbitrary directions, the component along the optic or c-axis is of most interest to holographers because of the relative size of the r_{33} electro-optic coefficient compared with the other coefficients.[11] This field causes a change to occur mainly in the extraordinary refractive index without affecting the other optical properties of the material. This refractive index change Δn_e is given by[12]

$$\Delta n_e = -\frac{1}{2} n_e{}^3 r_{33} E_s, \qquad [1]$$

where n_e is the extraordinary refractive index, r_{33} is the third diagonal component of the electro-optic tensor, and E_s the stored electric field along the optic axis. When the illumination is spatially periodic, E_s is also periodic and the refractive index pattern is the hologram.

The stored electric field under spatially periodic illumination, as in hologram formation, has been calculated from the space-charge model.[13] The results from Ref. [13] that apply to the subject of this paper are summarized below. When the medium, oriented as shown in Fig. 1, is illuminated by beams I_1 (object) and I_2 (reference) at the incidence angles $\pm\,\theta$, respectively, the total illumination I, as a function of the z-coordinate, is given by

$$I(z) = I_0 (1 + m \cos Kz), \qquad [2]$$

where I_0 is the sum of the intensities I_1 and I_2, $m = 2\sqrt{I_1 I_2}/I_0$ is the modulation index, and K is the angular frequency, given by

$$K = \frac{4\pi}{\lambda} \sin \theta. \qquad [3]$$

λ is the wavelength of the incident light. It is assumed that the lifetime τ of the photoexcited electrons is much shorter (of the order of 10^{-11} s)

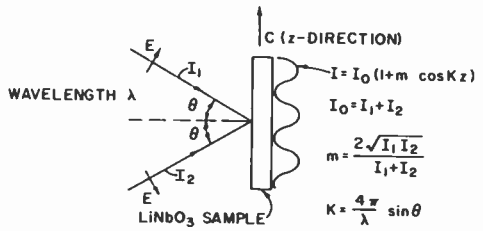

Fig. 1—Setup for recording holograms in $LiNbO_3$ crystal.

than the dielectric relaxation time. The lifetime is given by[14,15]

$$\tau = \frac{1}{N_e V S}, \qquad [4]$$

where N_e is the density of empty (or acceptor) traps in the material, V the thermal velocity of the electrons, and S their capture cross section. The dielectric relaxation time T_0 is given by

$$T_0 = \frac{\epsilon}{n_0 q \mu}, \qquad [5]$$

where ϵ is the dielectric constant, n_0 the average concentration of the photoexcited electrons, q the electronic charge, and μ the mobility. It is also assumed that the migration length L (average distance traveled during a lifetime) is such that $KL \ll 1$; from Ref. [13] this is true for θ less than about 23°. Under those conditions the electron concentration n is the product of the lifetime and the generation rate g, i.e.,

$$n = \tau g, \qquad [6]$$

where[9]

$$g = \frac{\sigma N_f}{h\nu} I(z). \qquad [7]$$

Here σ is the absorption cross section, N_f is the density of filled (or donor) traps, h is Planck's constant, and ν is the optical frequency. It is convenient to write Eqs. [6] and [7] as

$$n = n_0(1 + m \cos Kz) = n_0 n(z),$$

where

$$n_0 = \frac{\sigma N_f}{h\nu} \tau I_0 \qquad [9]$$

is the average concentration discussed earlier and

$$n(z) = 1 + m \cos Kz. \qquad [10]$$

The current density is

$$\mathbf{J} = qDn_0 \nabla n(z) + qn_0 n(z)\mathbf{E}, \qquad [11]$$

where D is the electron diffusion constant, and $\mathbf{E}$ is the total electric field, equal to the sum of the stored field $\mathbf{E}_s$ and of any dc field $\mathbf{E}_a$ that may be applied or that may exist within the material. In Ref. [13] it is shown that when all this information is substituted into the continuity equation

$$\nabla \cdot \mathbf{J} + \epsilon \left(\frac{\partial \mathbf{E}}{\partial t}\right) = 0, \qquad [12]$$

the magnitude of stored electric field (in the z-direction) is

$$E_s(z,t) = \frac{m}{n(z)} E_{max} \{1 - \exp [-n(z)t/T_0]\} \sin (Kz - \phi), \qquad [13]$$

where

$$E_{max} = \left[\left(K \frac{kT}{q}\right)^2 + {E_a}^2 \right]^{1/2} \qquad [14]$$

$$\phi = \tan^{-1} \left(\frac{qE_a}{KkT}\right), \qquad [15]$$

and where k is Boltzman's constant, T is the absolute temperature, and E_a is assumed to be along the z-axis. Due to the factor $n(z)$ in the denominator and in the exponential, E_s contains several harmonics. However, when the object consists of a large number of light spots at discrete locations, the light from each spot or bit creates an illumination in the form of Eq. [2], with $m \ll 1$, and Eq. [13] reduces to the simple sinusoid

$$E_s(z,t) = mE_{max}[1 - \exp(-t/T_0)] \sin(Kz - \phi). \quad [16]$$

This field can be permanently "fixed" in the material.[16] If no fixing is done, then upon constant illumination the stored electric field is erased. If the initial value is $E_0(z)$, the field decays as

$$E_s(z,t) = E_0(z) \exp(-t/T_0). \quad [17]$$

As mentioned earlier, the stored field gives rise to a refractive index change (Eq. [1]), which constitutes a phase hologram or grating. The hologram of an arbitrary object or of a distribution of point sources is a superposition of the individual gratings made by the interference of the reference beam and the point sources constituting the object.

For readout, the hologram is exposed to the reference beam alone, polarized extraordinarily, in order that its propagation be affected by Δn_e. Since the hologram thickness is of the order of 1 mm (equivalent to about 2000 optical wavelengths) and since the reference beam is at the Bragg angle, the diffracted light is a reconstruction of the object beam. The diffraction efficiency η, for negligible absorption, is given by[17]

$$\eta = \sin^2\left(\frac{\pi l|\Delta n_e|}{\lambda \cos\theta}\right), \quad [18]$$

where l is the hologram thickness. In practice the absorption constant α is nonnegligible, and if it is considered uniform, the efficiency is reduced by a factor of $\exp(-2\alpha l)$. Moreover, the quantity in parentheses in Eq. [18] is often much less than $\pi/2$. This gives

$$\eta \cong \left[\frac{\pi l|\Delta n|}{\lambda \cos\theta}\right]^2 \exp(-2\alpha l). \quad [19]$$

The effect of the absorption can severely limit the efficiency. For example, if α is large, differentiation of Eq. [19] with respect to l shows that the optimum thickness is $l = 1/\alpha$ and that the maximum efficiency is 14%. It must be noted that if the hologram is not fixed, reading it out will also erase it, a desirable feature for short cycle time in read–write applications.

2.2 Optical Sensitivity

It follows from Eq. [16] that for $m \ll 1$ the relaxation time T_0, which is the time constant for the decay (Eq. [17]) of the stored field of the hologram, is also the time constant for its build-up. It is inversely proportional to the illumination. By combining Eqs. [4], [5], and [9] we find

$$T_0 = \frac{C}{I_0} , \qquad [20]$$

where

$$C = \frac{\epsilon h \nu V S}{q \mu \sigma} \frac{N_e}{N_f} , \qquad [21]$$

and where N_e/N_f is the ratio of the empty traps to the filled traps. In $LiNbO_3$:Fe this ratio is that of the Fe^{3+} to the Fe^{2+} concentrations. We note that for a given sample, C is a constant that is proportional to the ratio of empty-to-filled trap concentrations. The smaller C is, the smaller T_0 for a given illumination, and the faster the build-up or erasure of a hologram. It is thus appropriate to designate C as the parameter that represents the sensitivity of a sample. Its units are those of energy density (if I_0 is in mW/mm^2, then C is in mJ/mm^2), and it can be obtained experimentally by measuring the decay time constant of a hologram using a known uniform illumination. Indeed, the stored field decays as e^{-t/T_0} or $e^{-I_0 t/C}$, and the holographic efficiency η decays from an initial value η_0 as the square of the field, i.e.,

$$\eta/\eta_0 = e^{-2I_0 t/C}. \qquad [22]$$

A comparison of the value of C for different samples is in fact a comparison of their respective sensitivities. This is the procedure that has been used, together with the measurement of the peak efficiency and the dark storage time constant, to evaluate the suitability of iron-doped lithium niobate samples for read–write applications. The parameter C has also been used to obtain expressions for the write and erase times (cycle time) of holographic memories.

2.3 Improving the Sensitivity of $LiNbO_3$:Fe

An important aspect of our work was the development of material-preparation procedures that will minimize C and of techniques to make the procedure reproducible. The highest sensitivity is obtained when the crystal has a low overall iron concentration and is treated so that all but about 10% of the iron is reduced to the divalent (Fe^{2+}) state, the rest of the iron remaining trivalent (Fe^{3+}). That this criterion is correct can be verified from the expression for the sensitivity parameter C (Eq. [21]), and from the expression for the diffraction efficiency (Eq. [19]) in the presence of absorption. Indeed, because C is proportional to the ratio of Fe^{3+} to Fe^{2+} concentrations, the smaller the ratio, the smaller C, i.e., the higher the sensitivity. However, since the absorption of the material is determined by the Fe^{2+} concentration, excessive absorption would

seriously reduce the diffraction efficiency; hence, the need for the overall absorption to be low.

Two procedures have been employed to reduce the Fe^{3+} concentration to low levels. In the first, crystals are annealed in argon at a temperature near 1100°C, then are cooled rapidly to room temperature. In the second, they are annealed at 550 to 600°C while packed in powdered Li_2CO_3 in an air or oxygen ambient.

When iron-doped crystals are annealed in argon, the Fe^{2+} concentration reaches a high value. However, the Fe^{2+} reoxidizes as the temperature is lowered, making it difficult to retain more than roughly 50% of the iron in the divalent state. If the crystals are cooled rapidly from the anneal temperature, more Fe^{2+} is retained, and this "quenched-in" concentration is stable at room temperature. The amount of Fe^{2+} retained depends on the cooling rate, which in turn depends on the size of the crystal, its placement in the furnace tube, and other variables that are difficult to control.

Annealing of iron-doped crystals in powdered Li_2CO_3 produces reduction of the iron by virtue of lithium ions diffusing into the crystal. This occurs at a comparatively low temperature (500° to 600°C), and the Fe^{2+} concentration thus produced is independent of the cooling rate. However, relatively little is understood about this process, and two questions in particular had to be investigated. First, are crystals produced in this manner equivalent to those produced in argon in terms of storage characteristics? Second, how does the Fe^{2+} concentration thus produced depend on the temperature and time of anneal? In addition, a third problem arises due to increased thermal erasure of Li_2CO_3-reduced crystals. The work described was performed on slices of a $LiNbO_3$ crystal doped with 0.002% Fe. The slices were oriented to lie in the x-z plane. The doping was chosen so that heavily reduced crystals of a convenient thickness (~2.1 mm) would absorb ½ to ¾ of the incident light at 4880 Å.

(*a*) *Sample Preparation by Argon Annealing*—It is estimated that the cooling rate necessary to "quench-in" about 90% of the Fe^{2+} population in argon-reduced crystals would be on the order of 10°C per second, limited by heat loss of the samples and of their support structure. Two experiments were performed to define the degree of difficulty to be encountered in this approach. In both cases, crystals were placed in a 1.25-inch ID quartz tube furnace and annealed at 1100°C for 16 hours. In the first case, samples were placed in a platinum-foil-lined alumina boat. At the end of the anneal, the entire quartz tube assembly was pushed out of the furnace and allowed to cool in the air. In the second experiment, a crystal was supported only by platinum foil. When the

quartz tube was pushed out of the furance after the anneal, it was cooled very rapidly with water.

The optical absorption of the two crystals is shown in Fig. 2. The air-quenched crystal has only about 65% of the absorption of the water-quenched crystal. The diffraction efficiency of holograms written in the water-quenched crystal is extremely small, of the order of 0.01%. This indicates that there is essentially no Fe^{3+} in the crystal. The optical absorption of the air-quenched crystal, on the other hand, indicates that only 65% of the Fe is divalent; the rest is presumed to be trivalent. Since this relative concentration of Fe^{3+} to Fe^{2+} is not optimum, it would be necessary to build an apparatus that would give a cooling rate between

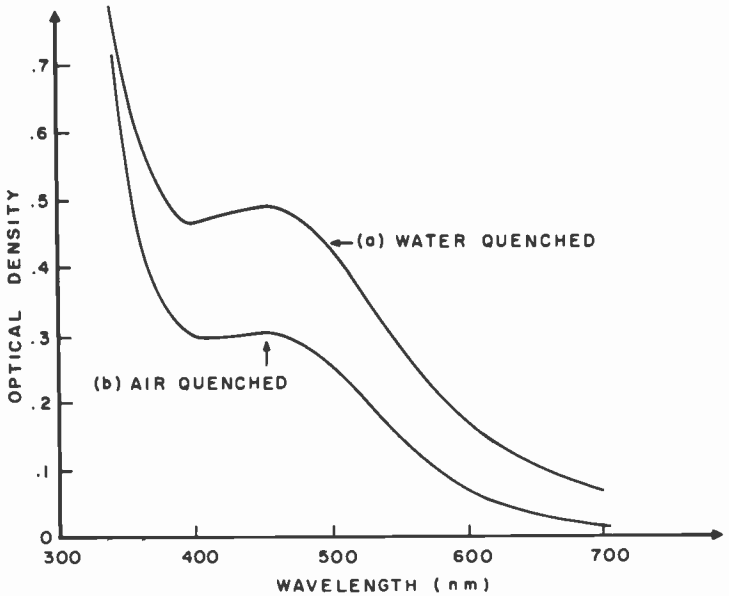

Fig. 2—Optical absorption of $LiNbO_3$:Fe (0.005% Fe) annealed in argon and quenched (a) in water and (b) in air. The water-quenched crystal is fully reduced; the air-quenched crystal has about 65% of Fe as Fe^{2+}.

that obtained with water quenching and that with air quenching. This approach was not pursued in view of the excellent results obtained by the Li_2CO_3 annealing process described below.

(*b*) *Sample Preparation by Li_2CO_3 Treatment*—The procedure for Li_2CO_3 sample preparation is to bury one or more unpolished crystals in powdered Li_2CO_3 contained in a small platinum crucible. The crucible is then heated in an oxygen atmosphere. The temperature is monitored

independently of the control thermocouple by means of a second thermocouple placed at the top of the crucible. At the end of the anneal, the Li_2CO_3 powder is discarded and the crystals are polished.

It is generally known that heavy coloration of the crystals would take place in the 500° to 550°C temperature range. However, insufficient information was available as to whether the reduction of crystals treated at the low end of this temperature range would reach equilibrium at some predetermined level below 100%, i.e., whether reduction was influenced by the temperature. To provide an answer to this question, crystals were annealed in Li_2CO_3 for varying times and temperatures. A sample annealed at 554°C for 60 hours was found to have absorption quite comparable to the water-quenched argon-annealed sample described earlier. This sample is assumed to be close to 100% reduced. However, other samples prepared at lower temperatures (486° to 500°C) have less ab-

Table 1—Summary of $LiNbO_3$:Fe Preparation and Characteristics.*

Preparation*						
Anneal Temp.	Time (hr)	Method	Percent Fe^{2+}	Optical Density at 450 nm	C** (mJ/mm^2)	Dark-Storage Time Constant
1100°C-Ar	60	Water quench	100	0.48	—	—
1100°C-Ar	23	Air quench	65	0.3	2.2	100 days
554°C	60	Li_2CO_3	100	0.44	—	—
485°C	82	Li_2CO_3	88	0.39	2.64	>100 hr
485°C	23	Li_2CO_3	77	0.34	2.1	>400 hr
485°C	48	Li_2CO_3	86	0.38	1.75	>100 hr
500°C	90	Li_2CO_3	90	0.40	—	—

*Argon anneal samples, 0.005% Fe; Li_2CO_3 annealed samples, 0.002% Fe.
**By comparison, a sample with 0.1% iron having 15% of Fe^{2+} has C = 65 mJ/mm^2.

sorption (corresponding to 85 to 90% reduction) even after extended anneal times. This indicates that the Li_2CO_3 process can be used to control the Fe^{2+} to Fe^{3+} ratio in the 85 to 95% region on an equilibrium basis; this is an important point, since it signifies that the process is amenable to close control and high reproducibility. The absorption data for a representative selection of crystals are shown in Fig. 3.

Table 1 is a summary of the characteristics of several samples prepared according to the two procedures. The argon-annealed water-quenched sample and the 554°C Li_2CO_3 samples, which are presumed to be fully reduced (100% Fe^{2+}), did not store holograms with measurable diffraction efficiency. The air-quenched argon-reduced sample shows good

sensitivity with $C = 2.2$ mJ/mm^2 and extremely long dark-storage time constant (100 days). By comparison, a typical sample with 1% iron and 15% reduced, made prior to the development of either reduction technique, had $C = 65$ mJ/mm^2. The most sensitive sample is the 48-hour, 485°C Li_2CO_3-reduced sample. It has a sensitivity of 1.76 mJ/mm^2, an improvement by a factor of 38 over the earlier 0.1% Fe sample. The Li_2CO_3-reduced samples have a shorter dark storage time constant than the argon-reduced samples. The reason for this is not known at the present time. We also do not know why the 500°C sample annealed for 90 hours would not store holograms, in spite of the fact that it does not appear to be 100% reduced. One can postulate, however, that the conductivity of the samples continues to build up during Li_2CO_3 reduction

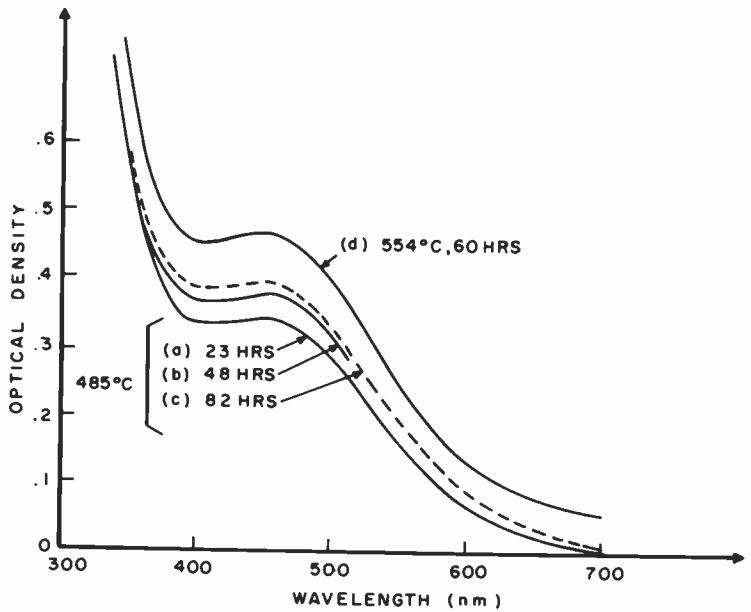

Fig. 3—Optical absorption of $LiNbO_3$:Fe (0.002%) annealed in Li_2CO_3. Samples (a), (b), and (c) are annealed at 485°C for the times shown. Sample (d) is fully reduced.

even after the Fe^{2+} concentration has reached equilibrium. Thus, the crystal annealed for 90 hours had too much conductivity for a significant field to build up during hologram formation.

Fig. 4 shows a comparison of the write–erase characteristics of holograms made in the 0.1% Fe-doped lightly reduced (15%) crystal and in one of the Li_2CO_3-reduced samples. The 0.1% Fe-doped sample, which has a value of $C = 65$ mJ/mm^2, is slow to erase but is capable of high

diffraction efficiency (dashed curve). On the other hand, the 0.002% doped crystal (solid curve), which is 86% reduced with $C = 1.7$ mJ/mm^2, has a much lower saturation efficiency but requires less than 2 mJ/mm^2 to erase significantly. It was also found that the optically induced scattering noise present in the 0.1% doped crystal does not exist in the lightly doped crystal reduced by either process. This characteristic has been observed in all the samples that exhibit low diffraction efficiency.

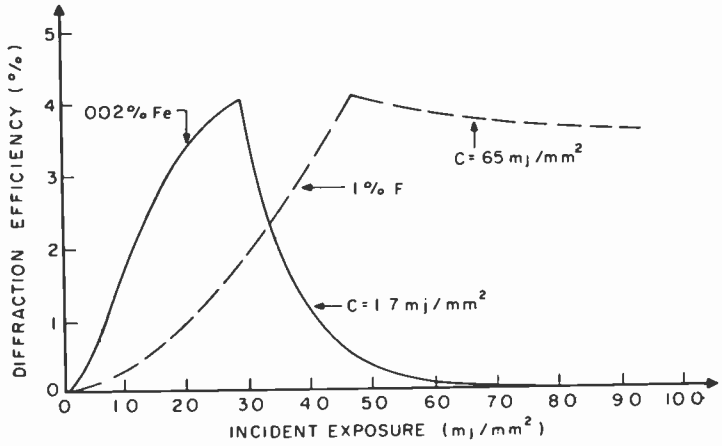

Fig. 4—Comparison of write–erase characteristics of different $LiNbO_3$:Fe crystals. Dashed curve indicates heavily doped, lightly reduced and solid curve lightly doped, heavily reduced in Li_2CO_3.

(*c*) *Experiments with Reduction in Other Powdered Salts*—We also explored the possibility of using powdered salts other than Li_2CO_3 to perform the heavy but incomplete reductions required. Two salts in particular were tried, Li_2SiO_3 and Na_2CO_3. The latter was of particular interest because reduction via sodium ions might be expected to produce somewhat different properties than lithium reduction, and perhaps lead to longer dark storage times. The absorption curves produced by the longest anneals in each of these salts are shown in Fig. 5. For comparison we also show a 100% argon-reduced crystal. Although reduction in these salts yields the expected Fe^{2+} absorption spectrum, the fractional conversion to Fe^{2+} is relatively small. The hologram storage sensitivity of these crystals was very low, as would be expected for crystals with large Fe^{3+} concentrations.

Although the optical absorption produced by annealing in Li_2SiO_3 and Na_2CO_3 continued to increase, the time that would have been required to approach 90% reduction was prohibitively long, and this line of investigation was stopped.

2.4 Spatial Frequency Response

The spatial frequency response is a plot of the holographic efficiency versus recording angle expressed in spatial frequency, at constant recording energy and modulation index. This response determines the allowable range of spatial frequencies in a storage system and the uniformity of the output signal. In the absence of the applied field E_a (a practical consideration in the design of a system), Eqs. [13] and [14] predict that the stored field is proportional to K which, according to Eq.

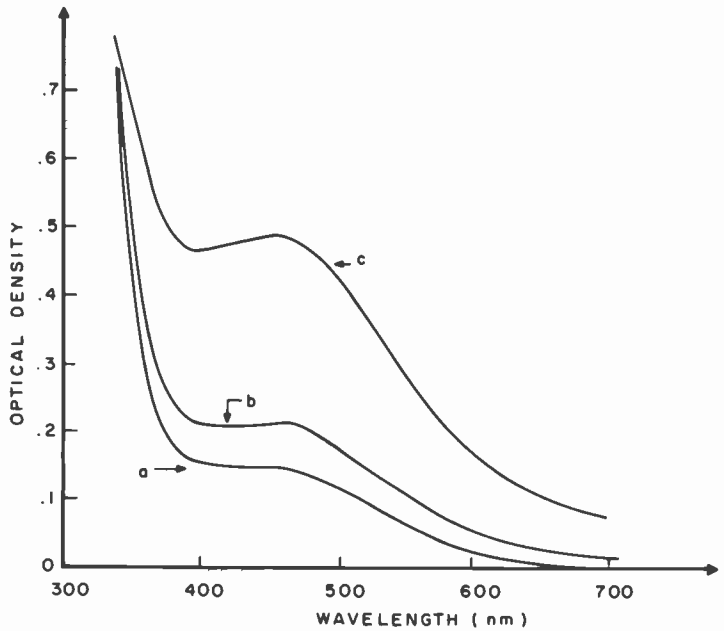

Fig. 5—Optical absorption of $LiNbO_3$:Fe (0.002% Fe) annealed in salts other than Li_2CO_3: (a) in Na_2CO_3 at 570°C for 90 hours, (b) in Li_2SiO_3 at 570°C for 80 hours, and (c) reference curve (fully reduced in Li_2CO_3).

[3], is proportional to the sine of half the recording angle, i.e., $K = (4\pi/\lambda) \sin \theta$. The efficiency, according to our theory, is thus proportional to K^2. However, the theory is valid only as long as the electron pattern remains a replica of the illumination, without smearing out. If $L = (D\tau)^{1/2} = (kT\mu\tau/q)^{1/2}$ is the diffusion length of the electrons in the material, the K^2 dependence will be true as long as $KL \ll 1$. For larger values of K the holographic efficiency will fall off as a result of the smearing of the electron pattern. The spatial frequency response is shown in Fig. 6 as

a plot of holographic efficiency versus the spatial frequency $f = K/2\pi$ lines/mm. For each data point the recording is 2.3 mJ/mm^3, and the modulation index is unity. The value of C for the sample is 2.2 mJ/mm^2 and the thickness is 1.7 mm. The expected K^2 dependence is shown to exist up to a spatial frequency of $f \cong 1600$ lines/mm, corresponding to $\theta \cong 23°$ or $K = 10^4$ per mm, followed by a roll-off approximately as $(1/K)^2$. A calculation[15] using $\mu \cong 16$ cm^2/V-s and $\tau \cong 4 \times 10^{-11}$ s gives $L \cong 4 \times 10^{-5}$ mm and $KL = 0.4$ at the "turning point." The roll-off above the turning point is consistent with a theory by Young and co-workers[18] for the initial time development, in which the restriction on the migration length is removed.

From a systems viewpoint the turning point indicates the optimum spatial carrier frequency or location of the data mask in a storage system, for only around that point is the storage-medium response essentially flat over some spatial frequency bandwidth. For the sample under discussion, the half-power response is from 1350 lines/mm to about 1900 lines/mm centered at 1600 lines/mm.

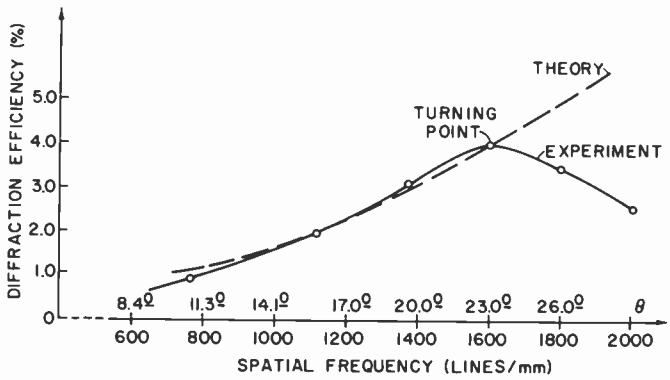

Fig. 6—Spatial frequency response of $LiNbO_3$. In an optical memory, the "turning point" is the optimum location of the data mask.

2.5 Noise Characteristics

In the readout of an optical system, any undesired light falling upon the photodetector array tends to reduce the image contrast and can be considered as noise. In lithium niobate, in addition to the usual noise due to scattering from imperfections in the crystal, there is often an optically induced scattering[19,20] that builds up negligibly initially as a function of exposure, then rises rapidly as a fourth power law of the exposure. This behavior is illustrated in Fig. 7. In the so-called "static" region, the noise contribution is mainly from the usual scattering process.

The induced noise appears only at high exposure. It is accompanied by distortion in the transmitted beam and exhibits the same properties as the holograms made in the sample—growth with exposure, angular sensitivity, polarization dependence, etc. The experimental evidence indicates that the induced noise is the result first of a gradual bending of part of the reference beam, followed by the interference between the bent and unbent portions of the reference beam. The ray bending is due

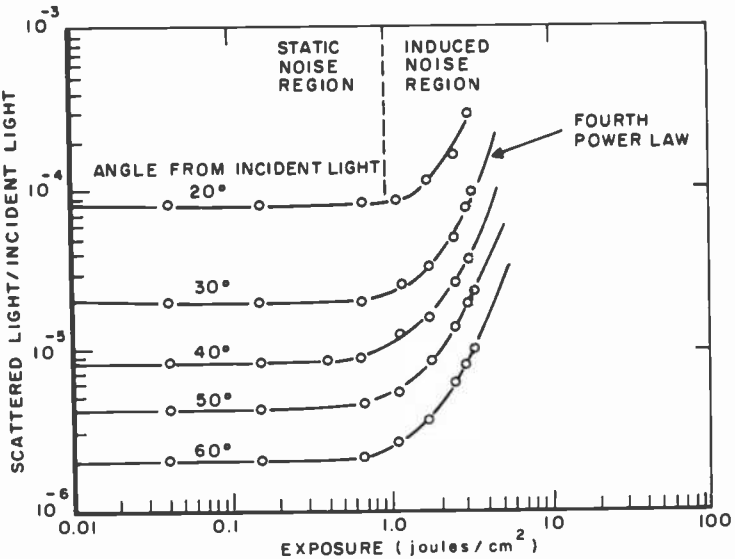

Fig. 7—Plot of the ratio of scattered light to incident light intensities as a function of exposure for 0.1% Fe-doped $LiNbO_3$ crystal. After an exposure of about 1 J/cm^2, the induced noise builds up as the fourth power of the exposure.

to optically induced local index inhomogeneities due to the possibly nonuniform cross section of the beam. The interference pattern is then recorded as a hologram. Experimentally, this scattering is found to occur only in samples capable of high diffraction efficiency. This optically induced noise does not seem to appear in the materials having improved sensitivity. It must be noted that the saturation efficiency of those materials is usually low, of the order of a few percent, possibly due to the effect of absorption. Thus, for all practical purposes, only the static noise is present in those samples.

The Wiener noise spectrum for a typical Li_2CO_3-reduced sample is shown in Fig. 8. A comparison of the Wiener noise at 1000 lines/mm for

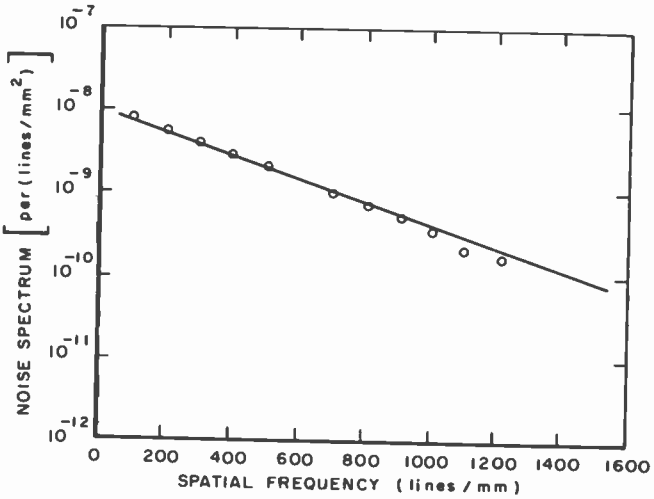

Fig. 8—Wiener noise spectrum of a typical $LiNbO_3$:Fe crystal reduced in Li_2CO_3.

several materials,[21] including Li_2CO_3-reduced iron-doped lithium niobate, is shown in Table 2. It is seen that $LiNbO_3$ compares well with dichromated gelatin (DCG) (used as a reference) and may have a slightly better noise characteristic.

3. Applications

A study was conducted to determine the characteristics of $LiNbO_3$:Fe as the storage medium in a holographic read–write memory. The basic configuration of a holographic memory is shown in Fig. 9.[5] It consists of a laser source, a deflection system, a hololens, a page composer, the storage medium, and a detector array. The deflection system consists of two acousto-optic deflectors at right angle to each other with a resolution of 100 to 1000 positions each (total capacity of 10^4 to 10^6 bits) and with access time of 2 to 10 μs. The "hololens" is a holographic optical element whose input is the light from the deflector and whose output

Table 2—Comparison of the Wiener Noise of $LiNbO_3$:Fe and of Other Storage Media at 1000 lines/mm.

Material	Wiener Noise at 1000 lines/mm	Relative to DCG
DCG (Dichromated gelatin)	8×10^{-10}	1.0
Kodak 649F at 50% Transmission	8×10^{-8}	100
Afga 10E70 at 50% Transmission	2.5×10^{-8}	31
Bleached Kodak 649F	9×10^{-8}	112
Li_2CO_3—Reduced $LiNbO_3$:Fe	5×10^{-10}	0.625

consists of the reference beam, which is directed toward the storage medium, and a cluster of 10^3 to 10^4 low-intensity beams directed toward the page composer. This page composer[22] is an array of electronically controlled light valves that modulate the cluster of beams in a binary fashion (ON = "1", OFF = "0"). Its output and the reference beam are focused to a 1-mm spot at the addressed location on the storage medium. With a 10^6-position deflector and a 10^3-bit page composer, the capacity of such a page-oriented memory is 10^9 bits. The hologram is written-in by the application of the reference beam and the object beams from the page composer, and it is read out by the application of the reference beam alone. The output light signals are converted to electrical signals by the

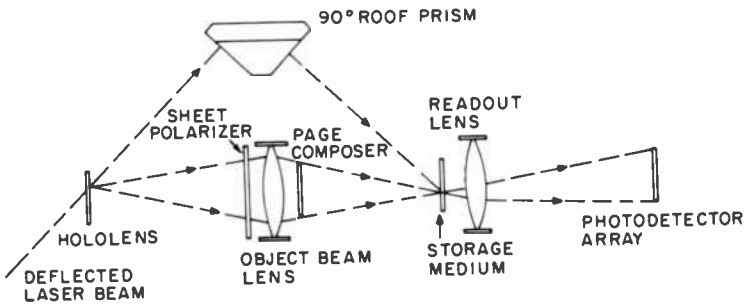

Fig. 9—Schematic of a read–write holographic memory (from Ref. [23]).

photodetector array. In a read–write system the readout is destructive, and the hologram must be rewritten before a new address is given to the deflectors.

3.1 Storage-Medium Requirements

The storage medium is the most important component of the system. It must have the following characteristics:

High-Resolution Capability—Since holographic recording involves high spatial frequencies, the resolution capability of the storage medium should be at least 1000 lines/mm.

Acceptable Dark Storage Time—Present-day computer systems use a hierarchy of memories ranging from small, fast, and volatile semiconductor devices to very large, slow, and permanent tape machines. Data or programs are transferred in blocks from the slow, high-capacity units to fast, smaller-capacity units for processing, and are later returned to the former devices. Thus, the storage ability of an active memory need not be permanent. In optical read–write memories the actual storage time would be determined by the data's ultimate place in the hierarchy.

Depending upon usage, an acceptable storage time may range from a week to several months.

Acceptably Short and Equal Write and Erase Times—The access time of core memories is of the order of 0.4 μs and that of discs is about 3 ms. At present, optical memories tend to be slower, but since the information is stored in the form of blocks or pages containing 10^3 to 10^4 bits, some speed compromise can be tolerated. An acceptable speed should be of the order of a few milliseconds per page and should be approximately the same for both writing and erasing.

Reversibility—The material should be reversible in the sense that a hologram should be erasable upon illumination. This implies that the read-out with the reference beam must be destructive, just as in core memories. However, retaining the information requires that a rewrite step be included in the memory cycle. This situation is acceptable if the writing speed is high enough, i.e., if the storage medium has enough sensitivity not to increase the cycle time beyond practical limits.

Low Noise—Due to transmission losses through the optical components, the amount of light reaching the storage medium is of the order of 5% of the laser output. The storage medium may contain light-scattering structures that may broadcast light over the photodetector array and reduce the signal-to-noise ratio of the system. This "noise", introduced by the storage medium, limits the system capacity.

Although several materials have been studied as potential storage media for holographic memories, iron-doped $LiNbO_3$ is particularly suitable, because it has the above characteristics to a large extent. Its major drawback is that its sensitivity is still insufficient for practical memory. However, it is very promising because of ease of preparation and because further improvements in its sensitivity are possible.

3.2 Component Limitations and Projected Performance

In a typical optical memory, the efficiency of the deflection system (about 20 to 25%) and the transmission and reflection losses through the various optical components are such that only about 5% of the total light reaches the storage medium. From a 1-W laser, only about 50 mW of light is available at the storage medium for writing and reading-erasing. About 10% of this amount (the hololens output) is split into the individual beams to be coded by the page composers and to be stored holographically in the storage medium. For a page composer of 10^3 bits, the reference-to-object beam ratio for each bit is 10^4, taking into account the effect of the hololens, and the modulation index is $m = 0.02$. The important question is the write–erase cycle time that can be expected with only 50 mW of total light available at the storage medium. Also, since

the modulation index m is small and the holographic efficiency is proportional to its square, the efficiency per bit is expected to be small. Then, with a writing time short and equal to the erase time, and with a low light level and low modulation index, will the signal at the detectors be high enough to have reasonable error rates?

To answer these questions from the available theory and verify them experimentally, let us define a suitable erase time T_E as the time required with the available light to erase a hologram to one-tenth of its initial value. From Eq. [22] the erase time thus defined is obtained by writing

$$0.1 = \exp(-2I_0T_E/C) = \exp(-2T_E/T_0). \quad [23]$$

This gives

$$T_E = 1.15T_0 = 1.15C/I_0. \quad [24]$$

To estimate the amount of light per bit at the photodetectors we note that, if the writing time T_W is made equal to T_E, the hologram efficiency reaches about four-tenths of its saturation value in the time T_W. Due to the absorption of the heavily reduced samples and their required small thickness, the measured saturation efficiency in our samples for $m = 1$ is of the order of 0.03. With $m = 0.02$, the efficiency per bit is $\sim 10^{-6}$. Assuming that a total 50 mW of light is available to make the holograms and that the hologram diameter is 1 mm, we obtain $I_0 = 64$ mW/mm^2. With a typical $LiNbO_3$:Fe sample having $C = 1.7$ mJ/mm^2 we obtain 30 ms for T_W and T_E, and 50 nW per bit at the photodetectors. This amount should be easily detected with a well-designed photodiode and sense amplifier system.

3.3 Memory Simulation

The setup of Fig. 10 simulates a memory with the parameters that have just been described. The hololens with 10% efficiency is represented by a 10:1 beam splitter; a 10^3-bit page composer is simulated by the neutral density filter having a transmission of 10^{-3}. This gives the desired 10^4 beam ratio between the reference and object beams for each bit in a page. The light source is the 488-nm line of an argon ion laser. We adjust the laser output and use optics to obtain a total light of 50 mW on a 1-mm-diameter spot on the storage medium. To improve the detector signal-to-noise ratio, we used differential detection with two p-i-n diodes and a band-limited high-gain differential amplifier. The detection sensitivity of the system was about 10^{-11} W.

With the setup described above, holograms were written in and read out in 30 ms each. The resulting sense signal upon readout is shown in

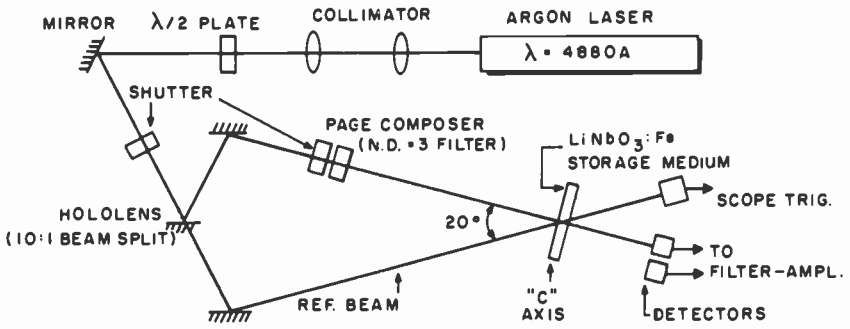

Fig. 10—Simulation of the light levels in a page-oriented read–write holographic memory with 10^3 bits per page.

Fig. 11. The signal corresponds to about 9 nW of light per bit, or a diffraction efficiency of about 10^{-7}. Although the detected light is small, the signal-to-noise ratio is very high, of the order of 27 dB, which is more than adequate from a systems point of view. Table 3 contains a complete set of experimental data for the simulated memory.

4. Conclusions

Holographic memories can be viable if their speed can be made competitive with other systems of comparable capacity. The major component affecting the speed is the storage medium. Photorefractive materials are potential candidates for the role of storage media, but their sensitivity is generally low. However, the sensitivity of lithium niobate doped with iron can be controlled by heavy chemical reduction of lightly doped samples. The most practical method of chemical reduction involves annealing the crystal in the presence of a powdered salt such as lithium

Table 3—Optical Memory Simulation Data

Wavelength	4880 Å	Reference Beam	47 mW (60 mW/mm²)
Storage Medium	Fe-$LiNbO_3$ (C = 1.7 mJ/mm²)	Material Relaxation Time	26 ms
Hologram Diameter	1 mm	Write or Erase Time	30 ms
Hololens Efficiency	10%	Systems Sensitivity	1.3×10^{-11} W
Page Composer	10^3 bits	Sense Signal	9.25×10^{-9} W
Beam Angle	20° (800 lines/mm)	SNR (with differential detection)	>27 dB

carbonate. This has yielded an improvement in optical sensitivity by a factor of about 38. The improvements have made it possible to evaluate the possible performance of a read–write memory using this material. Results on a simulated system indicate a cycle time of 60 ms with a high signal-to-noise ratio. Although still too slow for a practical system, this

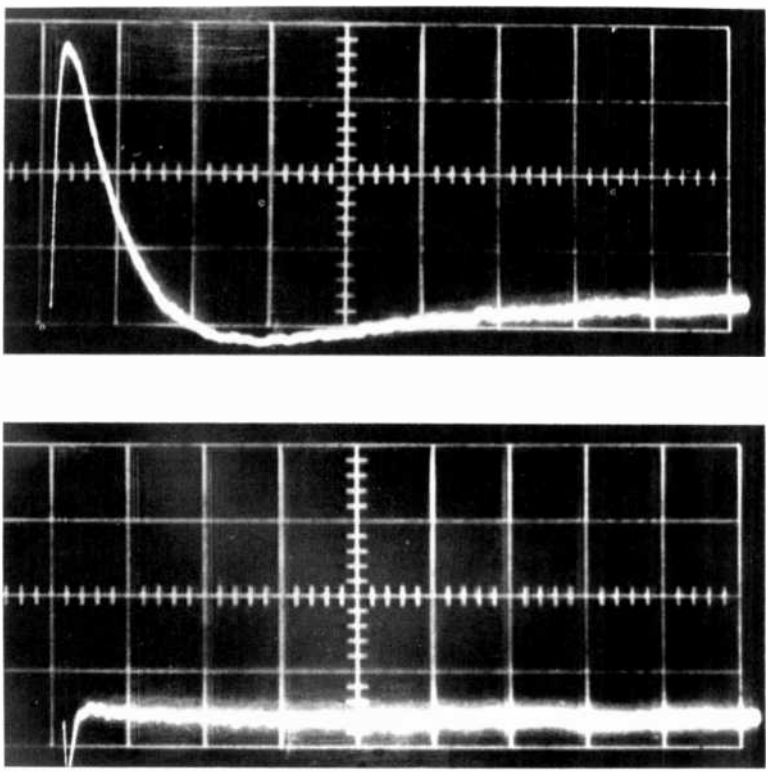

Fig. 11—Readout signals of simulated memory (horizontal = 20 ms/div; vertical = 2.5 nW/div): (a) binary "1" and (b) "0".

represents an improvement of two orders of magnitude over the performance of earlier experimental prototypes.[5]

Acknowledgments

The authors sincerely acknowledge technical discussions with D. L. Staebler, and technical assistance from W. C. Saunders.

References:

[1] L. S. Cosentino et al., Optical Memory Development, Vol. I—Prototype Memory System, Interim Report, Contract NAS8-26808, Nov. 1972.

[2] R. A. Gange and R. S. Mezrich, Optical Memory Development, Vol. II—Gain-Assisted Holographic Storage Media, Interim Report, Contract NAS8-26808, Nov. 1972.

[3] L. S. Cosentino, E. M. Nagle, and W. C. Stewart, Optical Memory Development, Vol. III—The Membrane Light Valve Page Composer, Interim Report, Contract NAS8-26808, Nov. 1972.

[4] W. C. Stewart, R. S. Mezrich, L. S. Cosentino, E. M. Nagle, F. S. Wendt, and R. D. Lohman, "An Experimental Read-Write Holographic Memory," *RCA Review*, **34**, (1973).

[5] W. Phillips and D. L. Staebler, "Control of the Fe^{2+} Concentration in Iron-doped Lithium Niobate," *J. Electronic Materials*, **3**, p. 601 (1974).

[6] A. Ashkin, G. D. Boyd, J. M. Dziedzie, R. G. Smith, A. A. Ballman, J. J. Levinstein, and K. Nassau, "Optically-Induced Refractive Index Inhomogeneities in $LiNbO_3$ and $LiTaO_3$," *Appl. Phys. Lett.*, **9**, p. 72 (1966).

[7] F. S. Chen, "A Laser-Induced Inhomogeneity of Refractive Indices in KTN," *J. Appl. Phys.*, **38**, p. 3418 (1967).

[8] D. von der Linde, A. M. Glass, and K. F. Rodgers, "Multiphoton Photorefractive Processes for Optical Storage in $LiNbO_3$," *Appl. Phys. Lett.*, **25**, p. 155 (1974); also, "High-Sensitivity Optical Recording in KTN by Two-Photon Absorption," *Appl Phys. Lett.*, **26**, p. 22 (1975).

[9] J. J. Amodei, "Electron Diffusion Effects During Hologram Recordings in Crystals," *Appl. Phys. Lett.*, **18**, p. 22 (1971).

[10] G. E. Peterson, A. M. Glass, and T. J. Negran, "Control of the Susceptibility of Lithium Niobate to Laser-Induced Refractive Index Changes," *Appl. Phys. Lett.*, **19**, p. 130 (1971).

[11] P. V. Lenzo, E. G. Spencer, and K. Nassau, "Electro-Optic Coefficients in Single-Domain Ferroelectric Lithium Niobate," *J. Opt. Soc. Amer.*, **56**, p. 633 (1966).

[12] J. F. Nye, *Physical Properties of Crystals*, Oxford Press, p. 252 (1969).

[13] G. A. Alphonse, R. C. Alig, D. L. Staebler, and W. Phillips, "Time-Dependent Characteristics of Photo-Induced Space Charge Fields and Phase Holograms in Lithium Niobate and Other Photorefractive Media," *RCA Review*, **36**, p. 213 (1975).

[14] A. Rose, *Concept of Photoconductivity and Applied Problems*, Wiley-Interscience, New York (1963).

[15] D. L. Staebler and W. Phillips, "Fe-Doped $LiNbO_3$ for Read-Write Applications," *Appl. Optics*, **13**, p. 788 (1974).

[16] J. J. Amodei, W. Phillips, and D. L. Staebler, "Improved Electrooptic Materials and Fixing Techniques for Holographic Recording," *Appl. Optics*, **11**, p. 390 (1972).

[17] H. Kogelnik, "Coupled Wave Theory for Thick Hologram Gratings," *Bell Systems Tech. J.*, **48**, p. 2909 (1969).

[18] L. Young, W. K. Y. Wong, M. L. Theriault, and W. D. Cornish, "Theory of Formation of Phase Holograms in Lithium Niobate," *Appl. Phys. Lett.*, **24**, p. 264 (1974).

[19] R. Magnusson and T. K. Gaylord, "Laser Scattering Induced Holograms in Lithium Niobate," *Appl. Optics*, **13**, p. 1545 (1974).

[20] W. Phillips, J. J. Amodei, and D. L. Staebler, "Optical and Holographic Storage Properties of Transition Metal Doped Lithium Niobate," *RCA Review*, **33**, p. 94 (1972).

[21] D. H. R. Vilkomerson, "Measurements of the Noise Spectral Power Density of Photosensitive Materials at High Spatial Frequencies," *Appl. Optics*, **9**, p. 2080 (1970).

[22] L. S. Cosentino and W. C. Stewart, "A Membrane Page Composer," *RCA Review*, **34**, p. 45 (1973).

Optical Properties of a Dense Two-Dimensional Electron Gas

R. Casanova Alig

RCA Laboratories, Princeton, N. J. 08540

Abstract—Electronic surface charge densities on the silicon side of a silicon–silica interface, which are nearly 100 times larger than the densities that can be generated with MOS structures, may be generated with mobile positive ions. Some optical properties of these dense two-dimensional electron gases, e.g., the reflectivity for light incident on the gas and the dispersion of light coupled to the plasma modes of the gas, are described. For a surface charge density of 10^{15} electrons/cm^2, the characteristic frequency of the electron gas is nearly 3000 cm^{-1}; hence these properties ought to be most interesting in the infrared.

1. Introduction

Whenever the electrostatic potential at an interface becomes sufficiently steep, the motion of bound electrons in the direction normal to the interface becomes quantized. When the separation of these quantized levels exceeds the thermal energy, the electrons bound at the interface behave like a two-dimensional electron gas. Examples of this behavior include electrons bound by the image potential to the surface[1] of liquid He and electrons bound within the inversion[2] and accumulation[3] layers of a semiconductor upon the application of a large electric field across a metal-oxide-semiconductor (MOS) structure.[4] The transport properties of conduction electrons in n-type inversion layers on p-type silicon have been studied extensively.[2]

In an MOS structure, the maximum surface charge density is fixed by the maximum electric field E_M the oxide layer can sustain, which is

about 10^7 volt/cm. Thus the largest surface charge density attainable is $\epsilon\,\epsilon_0 E_M \approx 2 \times 10^{13}$ electrons/cm^2, since the dielectric constant ϵ of silica is about 3.8. It is possible, however, to introduce into the oxide mobile positive ions, e.g. sodium, that can be made to drift toward the oxide–silicon interface. These ions remain bound near the interface due to the existence of negative image charges in the silicon. Using radiotracer techniques, sodium densities as large as 1.3×10^{15} cm^{-2} have been established, and internal photoemission and thermionic emission measurements show that much of this sodium is ionized.[5] Hence it appears that through the use of mobile positive ions, electronic surface-charge densities can be obtained that are nearly two orders of magnitude in excess of those obtainable with MOS structures.

Williams and Woods[6] have pointed out that the dielectric properties of the silicon near the interface are altered when the induced electronic charge density becomes comparable to the density of mobile electrons in the bulk. This can be easily understood by recalling that the complex refractive index of a solid is[7] $(\epsilon + 4\pi i\sigma/\omega)^{1/2}$ where σ is the conductivity, ω is the angular frequency of the wave, and ϵ is the dielectric constant when $\omega \gg \sigma$. If we take the conductivity σ_s of the bulk silicon to be 1 $(\Omega$ cm$)^{-1}$, or 9×10^{11} sec^{-1}, then $\hbar\, 4\pi\sigma_s = 60$ cm^{-1}. Thus for frequencies far in excess of 60 cm^{-1}, silicon behaves as a dielectric with $\epsilon = 11.7$ (excluding lattice and impurity modes); this is a characteristic frequency of the bulk silicon. As we shall see, a characteristic frequency $\omega_0 = 4\pi Ne^2/m^*c$, where N, e, and m^* are the electronic surface density, charge, and effective mass, can be associated with the induced charge density. Since the contributions to the conductivity are additive, when $\omega_0 \lesssim 4\pi\sigma_s$ the induced electronic charge density will not alter the dielectric behavior of the silicon. Since $m^* = 0.2m_0$,[8] $\hbar\omega_0 = 60$ cm^{-1} when $N = 2 \times 10^{-13}$ electrons/cm^2. Hence, unless the silicon is very pure so that σ_s is near the value 10^{-2} $(\Omega$ cm$)^{-1}$ characteristic of intrinsic silicon at room temperature, the silicon dielectric behavior is unaltered for all surface charge densities obtainable with MOS structures. However, when $N = 10^{15}$ electrons/cm^2, $\hbar\omega_0 = 2800$ cm^{-1} so $\omega_0 \gg 4\pi\sigma_s$ and the dielectric properties of the surface layer will differ from those of the bulk silicon; we expect these properties will be those of a two-dimensional metal. In the following paragraphs the optical properties of oxide–silicon interfaces containing high densities of sodium ions, i.e., the effect of this metal-like layer on the optical properties of the interface, will be explored.

These optical properties can be calculated from the dielectric response of the two-dimensional electron gas in the self-consistent-field approximation.[9] However, the two-dimensional electron gas can also be modeled as a thin metal sheet, the thickness of which is allowed to go to zero, to calculate these properties. Indeed the only difference between

a two-dimensional electron gas and a monolayer of metal lies in the mechanisms by which electrons are scattered. That is, in a thin metal sheet, boundary scattering becomes the dominant scattering mechanism as the thickness is diminished whereas in a two-dimensional electron gas, the quantization of the electronic motion normal to the surface inhibits this scattering mode. However, this difference will not be significant in the analysis as the scattering time is introduced phenomenologically. The advantage of this procedure lies, of course, in its conceptual simplicity.

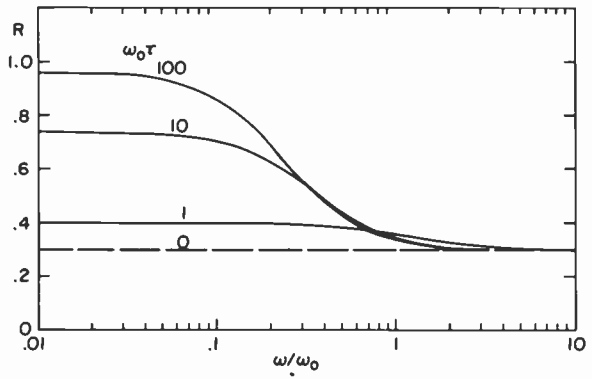

Fig. 1—The optical reflectivity R of a silicon–vacuum interface containing a two-dimensional electron gas with characteristic frequency ω_0, where $\omega_0 \gg \sigma_s$, and relaxation time τ is plotted versus angular frequency ω for selected values of $\omega_0\tau$. The silicon refractive index was taken as 3.8.

The optical reflectivity of a silicon-silica interface containing a large surface concentration ($N \gtrsim 10^{14}$ cm^{-2}) of sodium ions was calculated using this procedure. Electromagnetic radiation normally incident on the interface was considered and the silica layer was assumed thin compared to the wavelength of the radiation. The reflectivity R for selected values of the relaxation time τ is plotted as a function of frequency ω in Fig. 1. Clearly when $N \approx 10^{15}$ cm^{-2}, significant changes in the infrared reflectivity may appear due to the image charge induced by the sodium ions at the interface. Measurements of the infrared reflectivity and/or transmissivity will yield values of N and τ or, equivalently, of the complex dielectric constant of the two-dimensional electron gas.

Coupled waves composed of an electromagnetic wave and plasma oscillations of the two-dimensional electron gas can propagate along the interface; these waves will be called interface polaritons. When the media bounding the interface are identical, the dispersion curve of the interface polariton must lie below that of an electromagnetic wave in the medium

because the polariton has some plasmon character. In fact, in the frequency region about ω_0, the polariton dispersion $\omega(k)$ is found to vary as the square root of the wave vector k. When the media are different, however, generalizations become more difficult. Hence the interface-polariton dispersion for a silicon–vacuum interface containing a large concentration of sodium ions were found numerically (see Fig. 2). Electromagnetic power is coupled into this mode using techniques similar to those used in integrated optics to couple power into a wave-

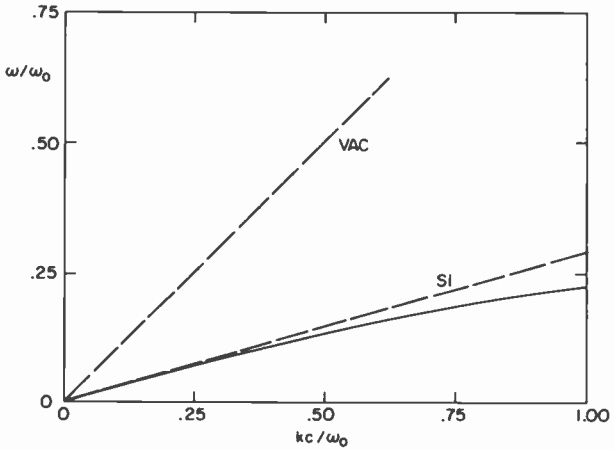

Fig. 2—The dispersion $\omega(k)$ of the polariton associated with a two-dimensional electron gas of characteristic frequency ω_0 at a silicon–vacuum interface is given by the solid curve. The dispersion of electromagnetic waves in vacuum and silicon are shown by the dashed curves. The angular frequency is expressed in units of ω_0 and the wave number k is given in units of ω_0/c.

guide, e.g., grating and prism couplers. Measurements of the dispersion and propagation loss of the interface polariton again yield the dielectric constant of the two-dimensional electron gas.

2. Optical Reflectivity

When an electromagnetic wave of frequency $\omega/2\pi$ impinges on a metal surface, it is reflected after penetrating a characteristic depth δ, called the skin depth, where

$$\delta = c/\sqrt{2\pi\omega\sigma}.$$

The dielectric behavior of the metal is expressed by the complex conductivity σ; this, in turn, can be expressed in terms of a characteristic

frequency ω_p and a characteristic time τ according to the relation

$$4\pi\sigma = \omega_p{}^2\tau/(1 - i\omega\tau).$$

Here $\omega_p{}^2 = 4\pi ne^2/m^*$ defines the plasma frequency ω_p where n, e, and m^* are the electronic density, charge, and effective mass; the relaxation time τ is a measure of the time between electronic collisions.

When $\omega\tau \gg 1$, $\sigma \approx i\omega_p{}^2/4\pi\omega$ so $\delta \approx c/\omega_p \equiv \lambda_p$, where λ_p is an electromagnetic wavelength corresponding to the plasma frequency. Thus, if the metal is replaced by a sheet of thickness d, we expect the sheet to reflect light as if it were infinitely thick whenever $d \gg \lambda_p$. When $d < \lambda_p$ however, the number of electrons that can respond to an electromagnetic wave is reduced, and so we expect the characteristic frequency to fall. If we let ω_0 denote this new characteristic frequency, we expect it to be

$$\omega_0 = \frac{d}{\lambda_p}\,\omega_p = \frac{d}{c}\,\omega_p{}^2 = \frac{4\pi nde^2}{m^*c} = \frac{4\pi Ne^2}{m^*c},$$

where $N = nd$ is the electronic surface density. For $N = 10^{15}$ electrons/cm^2 and $m^* = 0.2m_0$ (corresponding to the electronic effective mass in silicon[8]), $\hbar\omega_0 = 2800$ cm^{-1}.

The optical reflectivity was calculated for the geometry shown in Fig. 3. The region $z < 0$ is a vacuum, the region $z > d$ is filled with a dielectric of refractive index n, i.e., $\omega \gg \sigma_s$ is assumed, and the region $0 \le z \le d$ is filled with a metal characterized by ω_p and τ. A plane electromagnetic wave is incident on the metal from the vacuum. Some of this wave is reflected, some is transmitted through the metal into the dielectric, and some is dissipated in the metal (unless, of course, $\omega_p\tau = \infty$). The amplitudes of these waves may be found by solving Maxwell's wave equation in each medium and requiring the electric and magnetic fields be continuous at $z = 0$ and $z = d$. The solution to this problem is outlined by Stratton[10] and the specific solution for a thin metal sheet is developed in Appendix 1. There it is shown that when $d \ll \lambda_p$, the reflectivity R is

$$R = \left|\frac{1 - n - \omega_p{}^2 d\tau/[c(1 - i\omega\tau)]}{1 + n + \omega_p{}^2 d\tau/[c(1 - i\omega\tau)]}\right|^2.$$

In obtaining this result, it was assumed that $\omega_p \gg \omega$ and $\omega_p\tau \gg 1$; these assumptions should be very good at infrared frequencies.

In terms of the characteristic frequency ω_0 of the two-dimensional electron gas,

$$R = \left|\frac{1 - n - \omega_0\tau/(1 - i\omega\tau)}{1 + n + \omega_0\tau/(1 - i\omega\tau)}\right|^2.$$

Clearly if $\omega_0\tau \ll 1$, the sodium ions at the silica–silicon interface will not affect the reflectivity of the silicon (the silica is assumed thin compared to c/ω). Hence for a reflectivity change to occur we must have

$$\tau \gtrsim \omega_0^{-1} \approx 2\,N^{-1}\ \text{sec/cm}^2.$$

We might expect τ to be of the same order of magnitude as that found in bulk metals. e.g., about 10^{-14} sec at room temperature, so for electronic surface densities N above 10^{14} cm^{-2} this condition should be fulfilled.

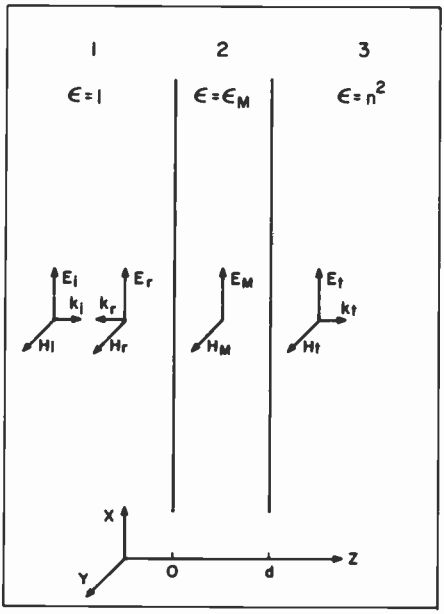

Fig. 3—A diagram of the electric, magnetic, and wave vectors of the system used to derive the optical reflectivity of the two-dimensional electron gas. These vectors are denoted by E, H, and k, and they are subscripted by i, r, t, and M to denote the waves incident, reflected, transmitted, and in the metal sheet. This sheet lies in the region $0 \le z \le d$; it is numbered 2 and described by dielectric constant ϵ_M. The regions $z < 0$ and $z > d$ are numbered 1 and 3 and are described by the dielectric constants 1 and n^2. The problem is one-dimensional along z.

If $\omega\tau \ll 1$, the sodium ions at the interface cause the apparent refractive index of the silicon to change n to $n + \omega_0\tau$. If $\omega\tau \gg 1$, then

$$R = \frac{(n-1)^2 + (\omega_0/\omega)^2}{(n+1)^2 + (\omega_0/\omega)^2}$$

i.e., the interface is totally reflecting when $\omega \ll \omega_0$ and the electron gas

is transparent when $\omega \gg \omega_0$. The reflectivity R is plotted versus ω for selected values of $\omega_0 \tau$ in Fig. 1.

The silica layer, ignored in the preceding analysis, will affect the reflectivity in proportion to its thickness Δ. From Stratton[10] the first-order correction to R due to the silica layer is $-30(\Delta/\lambda)^2$ where λ is the wavelength of the radiation in free space; the coefficient is obtained from the refractive indices of silicon and silica. Since Δ is typically 100 to 1000 Å while λ is near 100 μm, this correction is well below 1%.

It is a well-known[11] experimental observation that a layer of Al about 100 Å thick is required to obtain a reflectance exceeding 50% in the visible. Is this observation consistent with the prediction made here that a monolayer of electrons at metallic density can be nearly perfectly reflecting? In fact, real metal films of thicknesses less than 100 Å resemble disconnected crystallites more than a continuous metal film; hence they are more properly described by a Maxwell–Garnet theory.[11] However, it is interesting to speculate on what optical properties a continuous thin metallic film would have if it could be made. Like the two-dimensional electron gas the film would be less than one-skin-depth thick, and its characteristic frequency ought to be less than ω_p. But, unlike the two-dimensional electron gas, the film would be less than one electronic mean free path in thickness, and the relaxation time τ would be decreased due to boundary scattering.

3. Interface Polariton

In the preceding section we examined the interaction of the two-dimensional electron gas with an electromagnetic wave incident normally upon it. The extension of this analysis to oblique angles of incidence is straightforward. However, when a means to conserve momentum is available, another phenomenon is possible, the coupling of the electromagnetic wave to plasma oscillations in the two-dimensional electron gas. Momentum conservation must, of course, be present because the speed of light in the medium must exceed the speed of any matter wave. Common means of achieving momentum conservation include grating and prism couplers.

When electromagnetic and plasma waves are coupled, measurements of the dispersion and attenuation of this coupled excitation, or interface polariton, yield information about the dielectric properties of the two-dimensional electron gas.[9] While it is true, of course, that this same information can be obtained from the reflection (and transmission) measurements discussed above, the polariton may be easier to observe. For example, a similar mode has recently been observed on liquid helium.[12]

The dispersion relation $\omega(k)$ for the interface polariton of frequency ω and wavevector k associated with a two-dimensional electron gas bounded by a medium of dielectric constant ϵ is[9]

$$k^2 = \epsilon(\omega/c)^2 [1 + \epsilon(2\omega/\omega_0)^2].$$

Clearly when $\omega \gtrsim \omega_0$, we have $\omega \sim k^{1/2}$; in contrast to this, the dispersion relation for plasma oscillations in a three-dimensional electron gas is $\omega = (\omega_p + \alpha k^2)^{1/2}$. In Appendix 2 we derive a similar relation for the case where the two-dimensional electron gas is at the interface separating two different dielectric media.* As shown in Fig. 4, media with dielectric

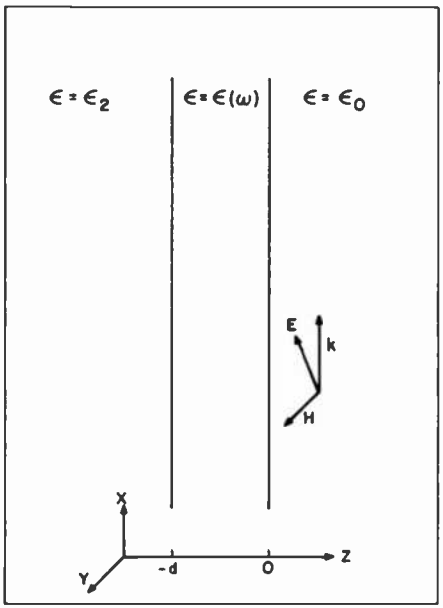

Fig. 4—A diagram of the electric, magnetic, and wave vectors, denoted by E, H, and k, of the system used to derive the dispersion of the interface polariton. Media of dielectric constants ϵ_0 and ϵ_2 occupy the regions $z > 0$ and $z < -d$, respectively. The region $-d \leq z \leq 0$ is occupied by a metal of dielectric constant $\epsilon(\omega)$. The problem is two-dimensional in the xz-plane.

constants ϵ_0 and ϵ_2 occupy the regions $z > 0$ and $z < -d$, respectively. The region $-d \leq z \leq 0$ is occupied by a metal of electronic density n. The wave propagates in the x-direction. Since only TM modes propagate (because a plasma can support only longitudinal electric fields), the magnetic field H lies along the y-direction and the electric field E lies

* This derivation is also described in Ref. [13].

in the xz-plane. The electromagnetic wave equations were solved in each region and the dispersion relation was found from the boundary conditions at $x = -d$ and 0. The dispersion relation for the interface polariton was found by letting d go to zero subject to the restriction that $nd = N$, the surface electron density. This dispersion relation is shown in Fig. 2 for a silicon–vacuum interface.

4. Summary

When the surface concentration of electrons at a silicon–silica interface is large enough that $\omega_0 \gg \sigma_s$, optical phenomena due to these electrons should be observable. More generally these optical phenomena should be observable for any electron layer for which $d \ll \lambda_p$, $\omega_0 \gg \sigma_s$, and $\omega_0 \tau \gtrsim 1$. Measurements* of these optical properties should provide data on the dielectric constant of the electron gas and data about the electron concentration to complement that obtainable from capacitance measurements.[14]

Acknowledgments

Discussions with H. Weakliem, R. Williams, D. Hoffman, R. Crandall, R. Braunstein, E. Priestley, and P. Zanzucchi have been very helpful in the development of these ideas.

Appendix 1—The Infrared Reflectivity

As described in the text, the infrared reflectivity will be calculated from the reflectance of a metal sheet separating two dielectric media. The analysis is elementary and it is modeled after one given by Stratton[10] for a dielectric sheet between two dielectric media. The diagram in Fig. 3 shows the electric, magnetic, and wave vectors of the waves in the vacuum, metal, and silicon. The vectors **E**, **H**, and **k** lie along the x, y, and z axes, respectively; they are subscripted with i, r, t, and M to indicate the waves incident, reflected, transmitted, and in the metal. The vacuum, metal, and silicon have dielectric constants, 1, ϵ_M, and n^2 respectively; these media have been labeled 1, 2, and 3.

For plane waves in a medium where $\mu = 1$, Faraday's law yields

$$\omega \mathbf{H} = c\,\mathbf{k} \times \mathbf{E} \qquad [1]$$

where the plane wave form

* Lattice modes, which frequently extend to several hundred cm^{-1}, however, may complicate these observations when $\hbar\,\omega_0$ also lies in this energy range.

$$\mathbf{E} = \mathbf{E}_0 \exp\{i(\mathbf{k}\cdot\mathbf{r} - \omega t)\} \quad [2]$$

has been assumed. In a homogeneous isotropic medium with waves propagating in the z-direction, the x-component of $\mathbf{E}$ and the y-component of $\mathbf{H}$, i.e., E and H, are related by $E = \pm ZH$ where Z is a characteristic impedance of the medium and the upper and lower signs correspond to waves propagating in the positive and negative z-directions. For these media, $Z = \epsilon^{-1/2}$. Hence in region 1, where $z < 0$, $\epsilon = 1$ so $Z_1 = 1$; in region 3, where $z > d$, $\epsilon = n^2$ so $Z_3 = n^{-1}$; in region 2, where $0 \leq z \leq d$, $\epsilon = \epsilon_M$ so $Z_2 = {\epsilon_M}^{-1/2}$. The metallic dielectric constant is

$$\epsilon_M = 1 + 4\pi i\sigma/\omega \quad [3]$$

and the metallic conductivity is

$$\sigma = \sigma_0/(1 - i\omega\tau) \quad [4]$$

where $\sigma_0 = {\omega_p}^2\tau/4\pi$ and ω_p and τ are the plasma frequency and relaxation time.

If we specify the x-component of the constant field $\mathbf{E}_0$ in Eq. [2] by $E_j^{\pm}$ where j specifies the region and the superscript specifies the direction of propagation along z, the spatial part of the x-components of the electric field are

$$E_i = E_1^{+} \exp\{i\omega z/c\}$$

$$E_r = E_1^{-} \exp\{-i\omega z/c\}$$

$$E_M = E_2^{+} \exp\{ik_M z\} + E_2^{-} \exp\{-ik_M z\}, \text{ and}$$

$$E_t = E_3^{+} \exp\{i\omega nz/c\} \quad [5]$$

where $k_M = \omega\sqrt{\epsilon_M}/c$. Then, from Eq. [1], the spatial part of the y-components of the magnetic field are

$$H_i = E_i/Z_1,$$

$$H_r = -E_r/Z_1,$$

$$H_M = (E_2^{+} \exp\{ik_M z\} - E_2^{-} \exp\{-ik_M z\})/Z_2, \text{ and}$$

$$H_t = H_t/Z_3. \quad [6]$$

Since the electric and magnetic fields must be continuous across the boundaries at $z = 0$ and $z = d$, we obtain the relations

$$E_1^{+} + E_1^{-} = E_2^{+} + E_2^{-},$$

$$E_1^{+} - E_1^{-} = Z_{12}(E_2^{+} - E_2^{-}),$$

$$E_2^{+} \exp\{ik_M d\} + E_2^{-} \exp\{-ik_M d\} = E_3^{+} \exp\{i\omega nd/c\}, \text{ and}$$

$$E_2^{+} \exp\{ik_M d\} - E_2^{-} \exp\{-ik_M d\} = Z_{23}E_3^{+}\exp\{i\omega nd/c\}. \quad [7]$$

where

$$Z_{12} = Z_1/Z_2 = \sqrt{\epsilon_M} \text{ and}$$

$$Z_{23} = Z_2/Z_3 = n/\sqrt{\epsilon_M}. \quad [8]$$

The reflectivity and transmittivity coefficients, R and T, are defined by

$$R = |E_1^-/E_1^+|^2 \text{ and}$$

$$T = |E_3^+/E_1^+|^2; \quad [9]$$

the fractional energy loss in the sheet is $1 - R - nT$. Since the reflectivity is expected to be more convenient to measure, we shall restrict the discussion to this quantity. From Eqs. [7] and [9],

$$R = \left|\frac{(1 - Z_{12})(1 + Z_{23}) + (1 + Z_{12})(1 - Z_{23})e^{2ik_M d}}{(1 + Z_{12})(1 + Z_{23}) + (1 - Z_{12})(1 - Z_{23})e^{2ik_M d}}\right|^2. \quad [10]$$

If we put the thickness of the metal sheet toward zero, $|k_M d| \ll 1$; then

$$R = \left|\frac{1 - n + ik_M d(1 - n + Z_{12} - Z_{23})}{1 + n + ik_M d(1 + n - Z_{12} - Z_{23})}\right|^2. \quad [11]$$

Since we are considering a metal in the presence of infrared radiation, $\sigma_0 \gg \omega$ and $\omega_p \gg \omega$ so $\epsilon_M \approx 4\pi i\sigma/\omega$ and $|\epsilon_M| \gg 1$ for all $\omega\tau$. Hence $|Z_{12}| \gg 1$ and $|Z_{23}| \ll 1$ so

$$R \approx \left|\frac{1 - n + ik_M d Z_{12}}{1 + n - ik_M d Z_{12}}\right|^2. \quad [12]$$

Since

$$k_M d Z_{12} = \omega d \epsilon_M / c = i\omega_0\tau/(1 - i\omega\tau) \quad [13]$$

where $\omega_0 \equiv \omega_p{}^2 d/c$,

$$R = \left|\frac{(1 - n)(1 - i\omega\tau) - \omega_0\tau}{(1 + n)(1 - i\omega\tau) + \omega_0\tau}\right|^2. \quad [14]$$

as given in the main text.

Appendix 2—The Interface Polariton Dispersion

The plasmon dispersion relation for a two-dimensional electron gas bounded by media of differing dielectric constants will be derived in the following paragraphs. The model consists of a plane sheet of metal of thickness d bounded by two semi-infinite, homogeneous, isotropic dielectric media of dielectric constants ϵ_0 and ϵ_2 as shown in Fig. 4. The dielectric constant $\epsilon(\omega)$ of the metal is

$$\epsilon(\omega) = 1 - \omega_p{}^2/\omega^2 \approx -\omega_p{}^2/\omega^2 \qquad [15]$$

where $\omega_p{}^2 = 4\pi ne^2/m^*$ and n, e, and m^* are the electronic density, charge, and effective mass. The approximation $\omega \ll \omega_p$ is possible because ω_p typically lies in the ultraviolet while the interesting optical frequencies ω lie in the infrared. As we shall see, plane waves of the form

$$\mathbf{E} = \mathbf{E}(z)\exp\{i(kx - \omega t)\} \qquad [16]$$

will propagate along the x-direction in this structure. However, this propagation is limited to TM waves, i.e., waves where the y-component of $\mathbf{E}$ is zero, because plasmas can support only longitudinal electric fields. Maxwell's equations provide three homogeneous equations in $E_x(z)$, $E_z(z)$, and $H_y(z)$ which can be solved in each of the three regions. Then continuity of the tangential components of these fields at the boundaries $z = 0$ and $z = -d$ leads to a set of homogeneous equations that has a nontrivial solution only if[15]

$$\left(1 + \frac{\alpha_0}{\alpha_1}\frac{\epsilon(\omega)}{\epsilon_0}\right)\left(1 + \frac{\alpha_2}{\alpha_1}\frac{\epsilon(\omega)}{\epsilon_2}\right) - e^{-2\alpha_1 d}$$
$$\times \left(1 - \frac{\alpha_0}{\alpha_1}\frac{\epsilon(\omega)}{\epsilon_0}\right)\left(1 - \frac{\alpha_2}{\alpha_1}\frac{\epsilon(\omega)}{\epsilon_2}\right) = 0 \qquad [17]$$

where

$$\alpha_0^2 = k^2 - \omega^2\epsilon_0/c^2,$$
$$\alpha_1^2 = k^2 - \omega^2\,\epsilon(\omega)/c^2 \approx k^2 + \omega_p{}^2/c^2, \text{ and}$$
$$\alpha_2^2 = k^2 - \omega^2\epsilon_2/c^2. \qquad [18]$$

This equation can be simplified somewhat if we define

$$x_0 = \alpha_0\epsilon(\omega)/\epsilon_0\alpha_1 \text{ and } x_2 = \alpha_2\epsilon(\omega)/\epsilon_2\alpha_1; \qquad [19]$$

then Eq. [17] becomes

$$(1 + x_0)(1 + x_2) - \exp\{-2\alpha_1 d\}(1 - x_0)(1 - x_2) = 0. \qquad [20]$$

When $\omega_p d/c \ll 1$, the exponential term can be expanded so that

$$(x_0 + x_2) + \alpha_1 d\,[1 - x_0 - x_2 + x_0 x_2] = 0. \qquad [21]$$

Since $\alpha_1 d \ll 1$, this equation has solutions only when $x_0, x_2 \gg 1$ or $x_0, x_2 \ll 1$.

The latter case, which will prove uninteresting, is best illustrated by setting $x_0 = x_2$ so

$$2x_0 = -\alpha_1 d \qquad [22]$$

or

$$-2\alpha_0 \omega_p{}^2/\omega^2 = -\epsilon_0 d\alpha_1{}^2 = -\epsilon_0 d(k^2 + \omega_p{}^2/c^2) \quad [23]$$

Since $k \ll \omega_p/c$ in the infrared, we have from Eqs. [18] and [23]

$$\alpha_0{}^2 \equiv k^2 - \omega^2 \epsilon_0/c^2 = \omega^4 \epsilon_0{}^2 d^2/4c^4, \quad [24]$$

so

$$k^2 = \omega^2 \epsilon_0 (1 + \omega^2 d^2 \epsilon_0/4c^2)/c^2, \quad [25]$$

that is, as $\omega d/c$ goes to zero, the wave approaches a plane wave propagating in the dielectric medium.

The solution of Eq. [21] when $x_0, x_2 \gg 1$ is found from the equation

$$x_0{}^{-1} + x_2{}^{-1} = -\alpha_1 d$$

or

$$\epsilon_0 \alpha_0{}^{-1} + \epsilon_2 \alpha_2{}^{-1} = -\epsilon d = \omega_p{}^2 d/\omega^2 \quad [26]$$

If we define the characteristic frequency

$$\omega_0 = \omega_p{}^2 d/c$$

and

$$x = \omega(\epsilon_0 \epsilon_2)^{1/4}/\omega_0, \; y = kc/\omega_0, \text{ and } \gamma = \sqrt{\epsilon_2/\epsilon_0};$$

then

$$\alpha_0 = \omega_0 x((y/z)^2 - \gamma^{-1})^{1/2}/c \quad [27]$$

and

$$\alpha_2 = \omega_0 x((y/x)^2 - \gamma)^{1/2}/c$$

The dispersion relation $\omega(k)$ is then given by the solution $x(y)$ of the equation

$$x^{-1} = \gamma^{-1}(q^2 - \gamma^{-1})^{-1/2} + \gamma (q^2 - \gamma)^{-1/2} \quad [28]$$

where $q = y/x$. When the dielectric media are identical so $\gamma = 1$, Eq. [28] becomes

$$(y/x)^2 - 1 = 4x^2 \quad [29]$$

or

$$k^2 = (\omega^2 \epsilon_0/c^2) + (4\, \omega^4\, \epsilon_0{}^2/\omega_0{}^2 c^2)$$

in agreement with the result obtained in Ref. [9].

Selected solutions of the dispersion equation for differing media were found numerically; they are displayed in Fig. 2 and discussed in the text.

References

[1] M. W. Cole and M. H. Cohen, "Image-Potential-Induced Surface Bands in Insulators," *Phys. Rev. Lett.*, **23**, p. 1238 (1969); R. S. Crandall and R. Williams, "Properties of Electron Surface States on Liquid Helium," *Phys. Rev.*, **A5**, p. 2183 (1972).

[2] K. F. Komatsubara, K. Narita, Y. Katayama, N. Kotera, and M. Kobayashi, "Transport Properties of Conduction Electrons in n-Type Inversion Layers in (100) Surfaces of Silicon," *J. Phys. Chem. Solids*, **35**, p. 723 (1974).

[3] D. C. Tsui, "Electron-Tunneling Studies of a Quantized Surface Accumulation Layer," *Phys. Rev.*, **B4**, p. 4438 (1971).

[4] F. Stern, "Surface Quantization and Surface Transport in Semiconductor Inversion and Accumulation Layers," *J. Vac. Sci. Technol.*, **9**, p. 752 (1972).

[5] T. H. DiStefano and J. E. Lewis, "The Influence of Sodium on the Si-SiO_2 Interface," *J. Vac. Sci. Technol.*, **11**, p. 1020 (1974).

[6] R. Williams and M. H. Woods, "Image Forces and the Behavior of Mobile Positive Ions in Silicon Dioxide," *Appl. Phys. Lett.*, **22**, p. 458 (1973).

[7] J. M. Ziman, *Principles of the Theory of Solids*, p. 220, Cambridge Univ. Press (1964).

[8] G. Abstreiter, P. Kneschaurek, J. P. Kotthaus, and J. F. Koch, "Cyclotron Resonance of Electrons in an Inversion Layer on Si," *Phys. Rev. Lett.*, **32**, p. 104 (1974); see also S. J. Allen, Jr., D. C. Tsui, and J. V. Dalton, "Far-Infrared Cyclotron Resonance in the Inversion Layer of Silicon," *Phys. Rev. Lett.*, **32**, p. 107 (1974).

[9] F. Stern, "Polarizability of a Two-Dimensional Electron Gas," *Phys. Rev. Lett.*, **18**, p. 546 (1967); A. L. Fetter, "Electrodynamics and Thermodynamics of a Classical Electron Surface Layer," *Phys. Rev.*, **B10**, p. 3739 (1974); A. Eguiluz, T. K. Lee, J. J. Quinn, and K. W. Chiu, "Interface Excitations in Metal-Insulator-Semiconductor Structures," *Phys. Rev.*, **B11**, p. 4989 (1975).

[10] J. A. Stratton, *Electromagnetic Theory*, p. 511, McGraw-Hill Book Co., N.Y. (1941).

[11] O. S. Heavens, *Optical Properties of Thin Solid Films*, Dover Press, N.Y. (1965).

[12] C. C. Grimes and G. Adams, "Observation of Two-Dimensional Plasmons and Electron-Ripplon Scattering in a Sheet of Electrons on Liquid Helium," *Phys. Rev. Lett.*, **36**, p. 145 (1976).

[13] M. Nakayama, "Theory of Surface Waves Coupled to Surface Carriers," *J. Phys. Soc. Japan*, **36**, p. 393 (1974).

[14] K. H. Zaininger, "Automatic Display of MIS Capacitance Versus Bias Characterisitcs," *RCA Review*, **27**, p. 341 (1966); M. H. Woods and R. Williams, "Injection and Removal of Ionic Charge at Room Temperature Through the Interface of Air with SiO_2," *J. Appl. Phys.*, **44**, p. 5506 (1973).

[15] D. L. Mills and A. A. Maradudin, "Properties of Surface Polaritons in Layered Structures," *Phys. Rev. Lett.*, **31**, p. 372 (1973).

Model of Target Cross-Contamination During Co-Sputtering

J. J. Hanak and R. W. Klopfenstein

RCA Laboratories, Princeton, N.J. 08540

Abstract—A simplified model has been developed for cross-contamination of sputtering targets occurring during co-sputtering. Approximate position-dependent expressions have been derived for the relative target-surface composition, target etch rates, and the time dependence of surface composition. Comparison of experimental data with these expressions has been made and qualitative agreement has been obtained. Taking into account the target cross-contamination, modifications have been suggested for existing methods of the compositional calculations of co-sputtered films.

1. Introduction

Co-sputtering has evolved as a practical method of materials synthesis which can greatly facilitate the study of the dependence of materials properties on composition.[1] Methods of compositional determination for co-sputtered multicomponent systems have been developed[2,3] that are based on calculated deposition profiles and easily determined sputtering rates of target materials. In these methods the well-known re-emission[4] of the sputtered species from the substrate and subsequent target cross-contamination[5] have been neglected, because in most instances they cause only small errors in composition. Recent observation of the formation of "cones"[6] on target surfaces as well as drastic changes in the sputtering rates of contaminated targets[3,6,7] may raise questions about the effect of such phenomena on the accuracy of the methods of compositional determination, if not about the ability to co-sputter cer-

tain compositions.[7] Hence, a study of cross-contamination was made which consisted of evolving a model and deriving equations for target surface compositions, for target etch rates, and for film thickness and composition. Experimental verification for the conclusions based on this model was also obtained.

2. Derivation of Basic Equations

2.1 The Model of Target Cross-Contamination

Consider a sputtering arrangement, such as that in Fig. 1, consisting of two half-disk targets a and b made of materials A and B, respectively, and a planar substrate placed parallel above the target plane. Material

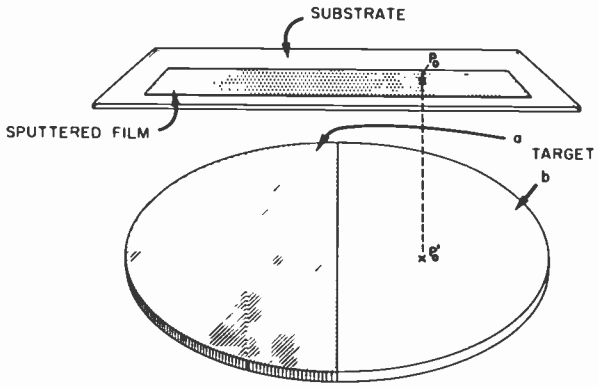

Fig. 1—Schematic arrangement for co-sputtering of two-component systems.

sputtered from any target i should be deposited on the substrate at a location P_0 according to geometric deposition profiles G_i. However, a significant fraction of the sputtered material may fail to deposit on the substrate and may return to the target for several reasons. These include the ionization of neutral species, back-scattering from the sputtering gas, sticking coefficient less than unity, and re-sputtering of the deposited material. The ionized species would tend to return at right angles to the target; the neutral, nonsticking species would return with a random distribution. To simplify the mathematical treatment, it will be assumed that all of the sputtered material arriving at a point P_0 on the substrate and failing to stick there, will return to a point P_0' perpendicularly below on the target. The consequence of this assumption is that the calculated cross-contamination will be greater than actual nearer the contaminating

target and less than actual far away. In this process a fraction of A will land on target b and vice versa, so that target cross-contamination takes place. Because of variable surface topography of a contaminated target, the definition of surface concentrations poses some problems. For example, a flat surface covered uniformly with a monolayer of a contaminant may be said to be 100% contaminated, whereas a surface covered with clusters of a contaminant may be only "slightly" contaminated, although on an average the contaminant may be present in amounts of several monolayers. Clearly, only those atoms that can be sputtered by an impact of an ion should be included in the surface concentration, which includes atoms on or near the surface. Let N_{11} and N_{12} be the number of such "active" surface atoms per unit area of target a and let the total number of the surface atoms be invariant, namely,

$$N_{11} + N_{12} = C.$$

The corresponding relative surface concentrations, ρ_{ij}, defined as

$$\rho_{ij} = \frac{N_{ij}}{C}$$

are related by the expression

$$\rho_{i1} + \rho_{i2} = 1, \qquad i = 1,2. \tag{1}$$

The model of cross-contamination proceeds from differential equations describing the rate of change of surface concentration, $d\rho_{ij}/dt$, which can be expressed in terms of the difference of the number of atoms departing from and returning to the target. Accordingly, the rate of change of concentration of B on target a can be expressed as

$$\frac{d\rho_{12}}{dt} = \frac{1}{C}(\epsilon_{11} - \epsilon_{12} - T_1 + T_2) \tag{2a}$$

where ϵ_{11} and ϵ_{12} are the total emission rates of A and B from target a in units of atoms/area/time and T_1 and T_2 are the rates of return of A and B in the same units. For the sake of simplicity we will assume that when atom A departs from or returns to target a, it will land onto or be emitted from another A atom. This assumption is partially justified because target a is an infinite reservoir of A. It becomes more valid at low concentrations of B. By similar argument we will assume that when an atom of B departs from or returns to target a, it will land onto or be emitted from an atom of A below it. Thus a change of concentration on target a can occur only when an atom B is removed from or arrives at the target. The rate of change of concentration can be expressed as the difference between the total emission rate ϵ_{12} of material B from target a and the rate of arrival T_2 of B from P_0 to P_0':

$$\frac{d\rho_{12}}{dt} = \frac{1}{C}(T_2 - \epsilon_{12}), \qquad [2b]$$

and from Eq. [1] it follows that

$$\frac{d\rho_{11}}{dt} = \frac{1}{C}(\epsilon_{12} - T_2). \qquad [2c]$$

Similarly

$$\frac{d\rho_{22}}{dt} = \frac{1}{C}(\epsilon_{21} - T_1). \qquad [2d]$$

The total emission rates ϵ_{ij} are expressed as

$$\epsilon_{ij} = Ce_{ij}\rho_{ij}, \qquad [3]$$

where e_{ij} are the relative emission rates in units of $(\text{time})^{-1}$. Both e_{ij} and ϵ_{ij} are dependent on concentration and ρ_{ij}, in turn, on position P_0'.

The rate of return T_j to point P_0' on the target is equal to the rate arriving at the point P_0 on the substrate, multiplied by its reemission coefficient k_j. For non-contaminated targets it has been demonstrated[8] that the rate of arrival to point P_0 can be expressed as the product of the emission rate ϵ_j and the deposition profile G_j. The latter is dependent on the target shape and size as well as its position relative to the point P_0 on the substrate. If similar expressions are used for contaminated targets the rate of return of A and B to P_0' from both targets will be

$$T_1 = k_1(\bar{\epsilon}_{11}G_1 + \bar{\epsilon}_{21}G_2) \qquad [4a]$$

and

$$T_2 = k_2(\bar{\epsilon}_{12}G_1 + \bar{\epsilon}_{22}G_2) \qquad [4b]$$

where $\bar{\epsilon}_{ij}$ are the emission rates averaged over each target taken to be $\bar{\epsilon}_{ij} = C\bar{e}_{ij}\bar{\rho}_{ij}$, similar to Eq. [3]. Similar averaging of ϵ_{ij} is implied in the methods of compositional analysis[1,2,3] and is justified, since this method has been shown to be relatively accurate. One can now write Eqs. [2] at steady state as

$$\frac{d\rho_{11}}{dt} = e_{12}(1 - \rho_{11}) - k_2[\bar{e}_{12}(1 - \bar{\rho}_{11})G_1 + \bar{e}_{22}\bar{\rho}_{22}G_2] = 0 \qquad [5a]$$

and, similarly,

$$\frac{d\rho_{22}}{dt} = e_{21}(1 - \rho_{22}) - k_1[\bar{e}_{21}(1 - \bar{\rho}_{22})G_2 + \bar{e}_{11}\bar{\rho}_{11}G_1] = 0. \qquad [5b]$$

Eqs. [5a] and [5b] specify ρ_{11} and ρ_{22} as functions of position in terms of the averages of several known quantities as well as their own

average values. The latter may be determined by taking the "average" of each of Eqs. [5], obtaining

$$\bar{e}_{12}(1 - \bar{\rho}_{11}) - k_2[\bar{e}_{12}(1 - \bar{\rho}_{11})\bar{G}_1 + \bar{e}_{22}\bar{\rho}_{22}\bar{G}_2] = 0, \quad [6a]$$

$$\bar{e}_{21}(1 - \bar{\rho}_{22}) - k_1[\bar{e}_{21}(1 - \bar{\rho}_{22})\bar{G}_2 + \bar{e}_{11}\bar{\rho}_{11}\bar{G}_1] = 0. \quad [6b]$$

Solving Eqs. [6a] and [6b] for $\bar{\rho}_{11}$ and $\bar{\rho}_{22}$ and neglecting all terms containing $k_1k_2\bar{G}_1\bar{G}_2$ we obtain

$$\bar{\rho}_{11} = 1 - \frac{k_2\bar{e}_{22}\bar{G}_2}{\bar{e}_{12}(1 - k_1\bar{G}_2 - k_2\bar{G}_1)}, \quad [7a]$$

$$\bar{\rho}_{22} = 1 - \frac{k_1\bar{e}_{11}\bar{G}_1}{\bar{e}_{21}(1 - k_1\bar{G}_2 - k_2\bar{G}_1)}. \quad [7b]$$

2.2 Target Surface Concentration

The preceding equations can be used in expressions for the position dependent concentrations ρ_{11} and ρ_{22} obtained from Eqs. [5], namely,

$$\rho_{11} = 1 - \frac{k_2}{e_{12}}[\bar{e}_{12}(1 - \bar{\rho}_{11})G_1 + \bar{e}_{22}\bar{\rho}_{22}G_2], \quad [8a]$$

$$\rho_{22} = 1 - \frac{k_1}{e_{21}}[\bar{e}_{21}(1 - \bar{\rho}_{22})G_2 + \bar{e}_{11}\bar{\rho}_{11}G_1]. \quad [8b]$$

Again, to first order in k_1, k_2

$$\rho_{11} \approx 1 - k_2\frac{\bar{e}_{22}}{e_{12}}G_2, \quad [9a]$$

$$\rho_{22} \approx 1 - k_1\frac{\bar{e}_{11}}{e_{21}}G_1. \quad [9b]$$

Recalling Eq. [1], we can also write

$$\rho_{12} = k_2\frac{\bar{e}_{22}}{e_{12}}G_2, \quad [9c]$$

$$\rho_{21} = k_1\frac{\bar{e}_{11}}{e_{21}}G_1. \quad [9d]$$

2.3 Net Target Etch Rates

Target cross-contamination will have an effect on the net etch rate E_{ii} of a given target. At steady state the net target etch rate is equal to the difference in the rates of removal and return of the target material; the

corresponding rates for the contaminant can be neglected as they are equal. Thus, we have

$$E_{11} = \epsilon_{11} - T_1, \tag{10a}$$

$$E_{22} = \epsilon_{22} - T_2. \tag{10b}$$

Substituting appropriate expressions for E_{ii} and T_i and again keeping only terms to first order in k_1, k_2 we obtain

$$E_{11} = Ce_{11}\left(1 - k_2\frac{\bar{e}_{22}}{e_{12}}G_2 - k_1\bar{e}_{11}G_1\right), \tag{11a}$$

$$E_{22} = C\left[e_{22}\left(1 - k_1\frac{\bar{e}_{11}}{e_{21}}G_1\right) - k_2\bar{e}_{22}G_2\right]. \tag{11b}$$

For a pure target sputtered by itself $\bar{e}_{ii} = e_{ii} = e$ and the net etch rate E becomes

$$E = Ce(1 - kG). \tag{12}$$

2.4 Compositional Determination of Films Co-sputtered from Cross-Contaminated Targets

(a) *Review of the Method for "Uncontaminated" Targets*

Co-sputtering of predictable composition as well as the methods of compositional determination are based on an early observation[8] that at reasonably low gas pressures sputtered species from a target are deposited approximately according to the Knudsen Cosine Law distribution. Expressions for calculating such normalized distribution functions, better known as deposition profiles $G_J(P_0)$ dependent only on the geometry of the sputtering arrangement, have been derived[2] for most practical target shapes. They are useful in expressing the film thickness $T_J(P_0)$ of component J, or of the total thickness $T(P_0)$ at any point P_0 on the substrate by the relation

$$T(P_0) = \sum_{J=1}^{n} T_J(P_0) = t\sum_{J=1}^{n} G_J(P_0)R_J(0'), \tag{13}$$

where t is the sputtering time and $R_J(0')$ is the deposition rate at some reference point on the substrate. Eq. [13] is the basis for determining the sputtering rates and eventually composition from measured film thicknesses:[1,2]

$$Vol.\ \%\ J = \frac{T_J(P_0)}{T(P_0)} \times 100. \tag{14}$$

Sputtering rates can also be determined from target etch rates and sticking coefficients as well as from actual chemical analysis of the deposited film.[3]

(b) Non-Uniform Target Etch Rate

The methods of compositional calculation[1,2,3] assume uniform density of emitted atoms over the target surface. As shown above, this assumption does not hold for a non-uniformly contaminated target surface. A method has been worked out[3] for correcting for such deviations. It consists of dividing the contaminated target into two or more areas and treating these areas as separate targets, for each of which a separate average deposition rate is calculated.

(c) Emission of Material B from Target a

The methods of compositional calculation thus far have neglected the emission of the contaminant from the cross-contaminated target. As noted above, at steady state, the amounts of contaminant arriving at and leaving the contaminated target are equivalent. The redeposition of B from target a to a point P_0 on the substrate can be also assumed to take place according to the Knudsen Cosine Law. The contribution to the thickness $T_B(P_0)$ will be

$$\delta T_B(P_0) = tG_a(P_0)R_{aB}(0), \qquad [15]$$

assuming an average value of the concentration of the contaminant as well as an average value of the deposition rate. Thickness of B from all targets n will be

$$T_B(P_0) = t \sum_{j=1}^{n} G_j R_{jB}. \qquad [16]$$

Eq. [13] can now be rewritten as

$$T(P_0) = \sum_{J=1}^{N} T_J = \sum_{J=1}^{N} \sum_{j=1}^{n} G_j R_{jJ}. \qquad [17]$$

Eq. 17 implies that the number of unknown rates is equal to nN, the product of the number of targets and materials. This means that if the rates were to be determined from thicknesses,[2] the minimum number of thickness points would be nN. If the rates were to be determined from chemical analyses,[3] the number of points on the substrates to be analyzed would be n.

Thus far, this modification of the method has not been tested. It is felt that it would be applicable to the determination of rates from chemical

analyses. It would probably fail if measured thicknesses were to be used because of insufficient accuracy of thickness measurements.

2.5 Time Dependence of Target Surface Concentration

The time dependence of target surface concentration $\rho_{ij}(t)$ might be of interest as, for instance, in determining the time necessary to reach a steady state. Appropriate expressions can be obtained by integrating Eqs. [5] with respect to time. In principle, this is straightforward; however, with the exception of k_j and G_i, all parameters are also unknown functions of time. One useful result that can be obtained is the time dependence of "sputter cleaning" of a contaminated target with all other targets removed. In this case, the second term ($k_j[\ ]$) on the right side of Eq. [5a] can be neglected, ρ_{11} changed to ρ_{12} via Eq.[1], and the expression for the concentration of the contaminants becomes

$$\rho_{12}(t) = \rho_{12}(0) \exp\left\{-\int_0^t e_{12}(t)dt\right\}, \qquad [18]$$

where $\rho_{12}(0)$ is the starting concentration and $\rho_{12}(t)$ and $e_{12}(t)$ are time dependent quantities. Although the time dependence of $e_{12}(t)$ is not known *a priori*, Eq. [18] is nevertheless useful, because it predicts an exponential decrease of the contaminant, in agreement with Wehner's observation.[6] From such data (ρ_{12} vs time) the emission rate, $e_{12}(t)$, can be determined from

$$e_{12}(t) = -(\Delta\rho_{12}/\Delta t)/\rho_{12}. \qquad [19]$$

Wehner's data[6] indicate that once the initial presputtering has reduced the impurity level substantially below a monolayer, $e_{12}(t)$ attains a constant value. Nevertheless, Eq. [18] predicts long sputter-cleaning times thereafter.

3. Comparison of Experimental Data With the Model

Experimental verification of this model was carried out using two different target configurations. One consisted of two half-disk targets of Ni and Cu 14.6 cm in diameter and the other of a 2.5-cm diameter Ni target centered on a 14.6-cm diameter Cu target. The targets were presputtered for several hours to reach steady state, as verified by constant target surface contamination, determined by x-ray fluorescence spectroscopy. Etch rates versus position on the target radius were measured for a pure Cu target and for the Cu target with a small, centered Ni target.

In the case of the Ni half disk, there was no detectable Cu contamination over the starting value; for the small Ni disk, minute Cu contamination was barely detectable indicating possibly up to an average of two monolayers of Cu. The nickel surface remained grainy but bright. Of the copper targets, the half disk surface underwent macroscopic contamination of about 200 monolayers of Ni at the Ni-Cu target interface down to about 70 monolayers near the edge away from the Ni target. The large copper disk had about 14 monolayers of Ni in the vicinity of the small Ni disk target and about 3 monolayers near the edge away from the Ni target. The copper targets were covered with cones, with cone density decreasing with decreasing contamination as indicated by the relative darkening of the targets.

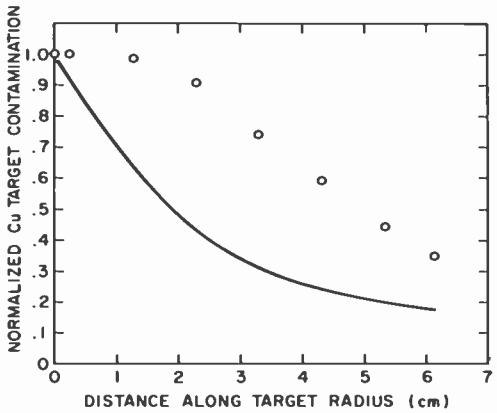

Fig. 2—Observed and calculated cross-contamination of copper half-disk target by a nickel half disk.

Comparison of observed and calculated relative contamination for the two cases is shown in Fig. 2 and 3. In the computations a value of 0.4 was assumed for k_1 and k_2. The ratio of etch rates $\overline{e_{22}}/e_{12}$ calculated at "intermediate" target contamination was 32 and was assumed to be constant. An excellent agreement was obtained for the small disk of Ni and the large disk of Cu (Fig. 3). For the two half disks (Ni and Cu, Fig. 2) the trend is predicted correctly; however, the massive Ni contamination near the Ni target apparently decreases the value of $\overline{e_{21}}/e_{11}$ and contributes to increased Ni concentration everywhere. The relative contamination of the two Cu targets was in good agreement with regard to the relative target areas and deposition profiles $G(P_0)$. It is significant that in spite of the massive contamination of copper surface by nickel and the cone formation, copper continues to be sputtered and the

sputter-etch-rate decrease reaches a limiting value, as has been demonstrated by Hanak et al.[2,3] Furthermore, the methods of compositional determination reviewed above are applicable without the need to take into account the target contamination, although methods have been established for massive target contamination.[3] The ranges of film composition from thus-contaminated targets are normally considerably different from those obtained from clean targets.

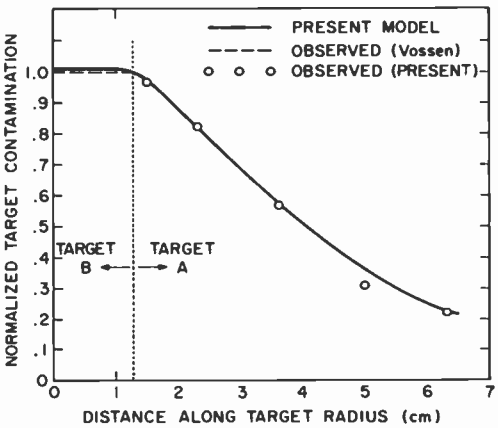

Fig. 3—Cross-contamination of a copper disk target (14.6-cm dia.) by a centered nickel disk target (2.5-cm dia.).

Comparison of calculated and measured target etch rates has been done for copper targets with and without the small centered nickel target. First the dependence of relative etch rates has been calculated for a copper target alone from Eq. [4] for various values of k_1 and plotted along the target radius, as shown in Fig. 4 by the solid curves. These values can be qualitatively compared to observed etch rate data in Fig. 5 obtained at 3 different argon pressures. The decrease in etch rates at higher pressures signifying an increase in k_1, observed by Vossen[5], is thus in agreement with Fig. 4. The calculated etch rates are minimum at the center, whereas the observed rates are maximum. This disagreement may be due to uneven ion density distribution (the target was recessed below the ground shield) and/or sputtering by Cu^+ ions which would be a greater concentration near the center of the target. The etch data for the copper target with the small centered nickel target are shown in Fig. 6, and are in qualitative agreement with the calculated etch rate dependence shown in Fig. 4 (dashed line).

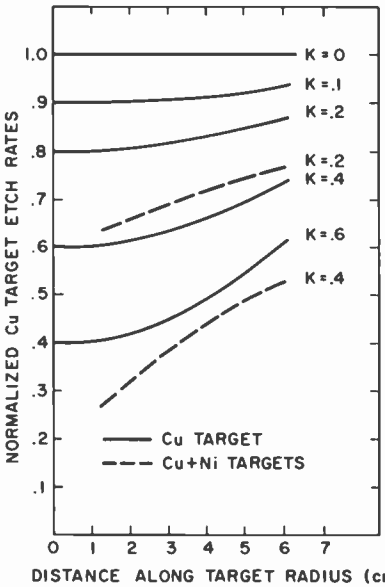

Fig. 4—Calculated sputter etch rates at various values of k_1 for a pure copper disk target and for one contaminated with a small centered nickel disk target.

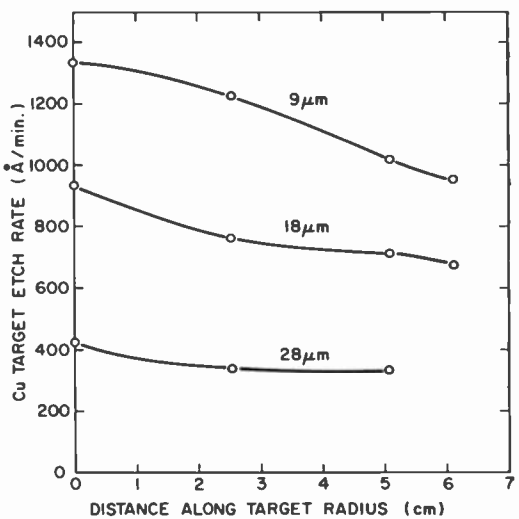

Fig. 5—Net target etch rate for pure copper disk target at different argon pressures.

4. Parameters Affecting Cross-Contamination

(*a*) *The Reemission Coefficient*

The coefficients $k_{1,2}$ can have values between 0 and 1. Experimental techniques for determining them have been established by Maissel et al.[4] Eqs. [9] indicate a direct relationship between target contamination and the value of the reemission coefficient of the contaminating species. Eqs. [11] indicate that net target etch rates are depressed with increasing $k_{1,2}$. High values of $k_{1,2}$ cause redistribution of the sputtered species and,

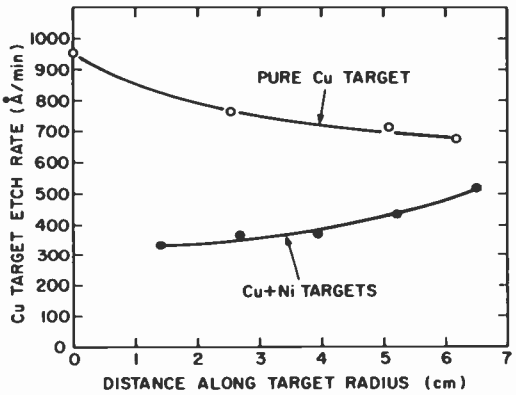

Fig. 6—Experimental net etch rates for a pure copper disk target and for one contaminated by a small centered nickel target.

consequently, a flattening of the concentration gradient in co-sputtered films. Thereby they make the use of modifications (Eqs. [17] and [18]) in the composition determination mandatory. Therefore, in co-sputtering it seems wise to avoid conditions leading to high values of $k_{1,2}$.

(*b*) *The Deposition Profiles* $G_{1,2}$

The calculated deposition profile[2] can also be normalized to range between 0 and 1. To a first approximation $G_{1,2}$ are proportional to the relative target area. Thus, Eqs. [9] indicate that, in general, large targets should suffer relatively low contamination in the presence of small contaminating targets, unless the rates of $\bar{e}_{ii}/e_{ij}$ is large.

The expressions [11a] and [11b] can be simplified for small targets in the presence of large ones by neglecting the last term. However, for large targets the equations cannot be further simplified unless the ratio of the

rates $\bar{e}_{ii}/e_{ij}$ is $\lesssim 1$. For finite values of k of a large target, the net etch rate of a small target will always be lowered substantially.

(c) *Ratio of Emission Rate* $\bar{e}_{ii}/e_{ij}$

Perhaps the most unpredictable yet important variable in Eqs. [9] and [11] is the ratio of the emission rates $\bar{e}_{ii}/e_{ij}$, $i \neq j$. As already noted, when this ratio becomes large, target contamination can become large and the target etch rate can be depressed. This condition can occur when cone formation takes place or when some very stable alloy or compound having low sputtering rate is formed on the target surface. In dc sputtering, the formation of an insulating surface layer may decrease the rate e_{ij} to zero and increase $\bar{e}_{ii}/e_{ij}$ to infinity. Wehner[6] found drastically different emission rates of molybdenum from the surfaces of various materials and observed variation of emission rate with surface concentration. Large values of this ratio of rates (in the absence of cones) would be expected for the case of one or less monolayers of material B on target a. Buildup of material B beyond a monolayer would tend to reduce this ratio to unity, thereby also limiting further buildup. A decrease in the ratio $\bar{e}_{ii}/e_{ij}$ is also possible, as, e.g., in cases when a volatile compound is formed. Moreover, this would exclude large contamination, as implied in Eqs. [9].

(d) *Surface Concentration*

It would be useful to give the units of surface concentrations ρ_{ij} definitively, but this is impossible because of factors such as surface topography and bulk diffusion. To illustrate this point, in the case of cone formation, crystallites of contaminant have been observed on the tips of the cones[7] implying that, on an average, many monolayers of the contaminant may be present on the surface, yet the surface is not fully contaminated. Possibly an unambiguous definition would be the ratio of the probability of sputtering an atom of A or B, divided by the total probability.

Acknowledgments

The authors are grateful to J. P. Pellicane for performing most of the sputtering experiments, to R. J. Paff for doing the x-ray fluorescence spectroscopy analyses, and to J. L. Vossen for constructive discussions.

References:

[1] J. J. Hanak, "The 'Multiple-Sample Concept' in Materials Research: Synthesis, Compositional Analysis and Testing of Entire Multicomponent Systems," *J. Materials Sci.*, **5**, p. 964 (1970).

[2] J. J. Hanak, H. W. Lehmann, and R. K. Wehner, "Calculation of Deposition Profiles and Compositional Analysis of Cosputtered Films," *J. Appl., Phys.*, **43**, p. 1666 (1972).

[3] J. J. Hanak and B. F. T. Bolker, "Calculation of Composition of Dilute Cosputtered Multicomponent Films," *J. Appl. Phys.*, **44**, p. 5142 (1973).

[4] R. E. Jones, C. L. Standley, and L. I. Maissel, "Re-emission Coefficients of Si and SiO_2 Films Deposited Through RF and DC Sputtering," *J. Appl. Phys.*, **38**, p. 4656 (1967).

[5] J. L. Vossen, J. J. O'Neill, Jr., K. M. Finlayson, and L. J. Roger, "Back Scattering of Material Emitted from RF-Sputtering Targets," *RCA Review*, **31**, p. 293 (1970).

[6] G. K. Wehner and D. J. Hajicek, "Cone Formation on Metal Targets During Sputtering," *J. Appl. Phys.*, **42**, p. 1145 (1971).

[7] C. Sella, L. Tertian, P. Regnier, and A. Le Bian, "Problemes Poses par la Preparation de Couches Homogenes d'Alliages Fer-Argent par Copulverisation," *Proc. 1er Colloque International de Pulvérisation Cathodique et Ses Applications, Monpellier, France, Oct. 1973*, p. 199 (Suppl. to No. 165 of the *Review Le Vide, Les Couches Minces*).

[8] A. von Hippel, "Disintegration of the Cathode," *Ann. Physik*, **81**, p. 1043 (1926).

Third-Order Distortion in Amplifiers and Mixers

S. M. Perlow

RCA Laboratories, Princeton, N. J. 08540

Abstract—The transfer characteristic, both amplitude and phase, of a "linear" amplifier or mixer is sufficient information to determine all the spurious responses related to third-order distortion. The linearity of an amplifier is not determined by the linearity of the device alone. The circuit environment in which the device is embedded must also be considered. Equations relating the various forms of third-order distortion have been developed that agree very well with experimental data. These equations coupled with the concept of linear efficiency can be used to arrive at an estimate of the various third-order distortion components in the initial-design stage of an amplifier or mixer.

1. Introduction

The dynamic range of an amplifier or mixer represents the range of signal levels over which it will exhibit its intended signal-processing properties. The lower limit of dynamic range is determined by the noise figure or sensitivity, while the upper limit is determined by the allowable level of signal distortion.

A great deal has been written on both these limitations. Many claims have been made for the distortion properties of certain devices. Masers, FET's, and abrupt junction varactors have all been called the most linear devices which, when incorporated in amplifiers and mixers, will result in a signal processor with the best linearity. Despite all these claims, a fundamental limitation on linearity exists, namely, the gain–distortion (amplitude and phase) of the complete amplifier or mixer. The question as to which device will produce a signal processor with the greatest range

of linearity cannot be answered by considering the linearity of each device by itself. In most cases, the circuit environment should provide the criteria for choosing the type of device.

2. Gain Saturation

Mixers and amplifiers are special forms of power converters. They all transform power at one frequency to power at another. For example, a transistor mixer will convert both dc and local oscillator power to power at the i-f frequency. As power converters, these circuits must follow the fundamental rules of thermodynamics. Consider the two-port amplifier

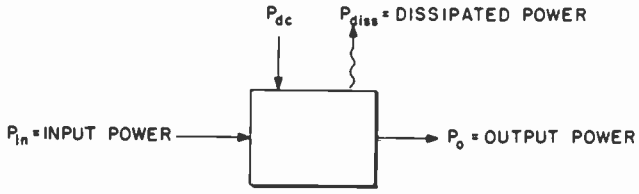

Fig. 1—Two-port amplifier.

of Fig. 1. The total input power must equal the total output power:

$$P_{in} + P_{dc} = P_o + P_{diss}. \quad [1]$$

Defining the power gain G as

$$G = \frac{P_o}{P_{in}} \quad [2]$$

$$P_{diss} = P_{dc} - (G - 1)P_{in}. \quad [3]$$

If the power gain were to remain constant and greater than unity, then at some level of input power the dissipated power would become negative. This is clearly impossible and, therefore, the power gain cannot remain constant. It is important to note that no assumptions about the linearity of the device used in the two-port amplifier were made. Even if the amplifying device were ideally linear, gain saturation would still occur as long as the dc power level remained unchanged. The difference between an ideally linear amplifier and an amplifier utilizing an ideally linear device is now easily recognized. The gain of an ideally linear amplifier, by definition, never changes, but the gain of an amplifier utilizing an ideally linear device does saturate. For the amplifier to be ideally

linear, it would have to have, in addition to an ideally linear device, an infinite source of dc power.

Device nonlinearities reduce the gain saturation level. This is equivalent to stating that the nonlinearities of the device reduce the conversion efficiency. This is seen in the following discussion.

Consider a transfer current gain device that is linear in the first quadrant of the V-I plane. Its output characteristic is shown in Fig. 2.

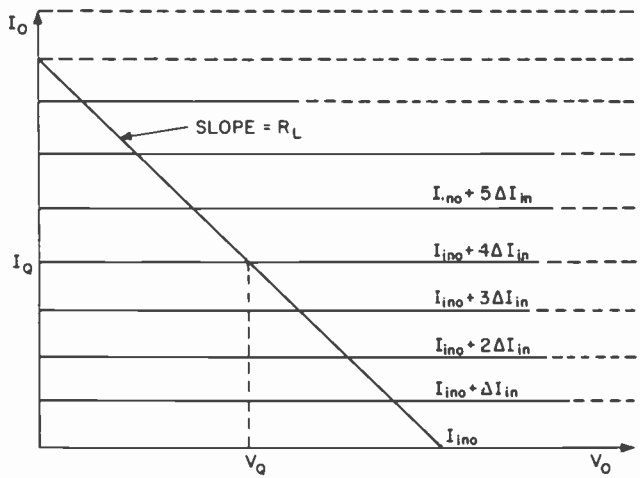

Fig. 2—Linear current-amplifying device.

Note that the idealized characteristics extend toward infinity on both positive voltage and current axes. A dc operating point is supplied to the device so that the bias point (V_Q,I_Q) is centered on the load line established by the load resistor, R_L. Under these conditions the peak to peak signal voltage and signal current swings become $2V_Q$ and $2I_Q$, respectively. The maximum power developed in R_L is, therefore,

$$P_{om} = \frac{V_Q I_Q}{2}. \qquad [4]$$

Since the dc power supplied is $V_Q I_Q$ the maximum output efficiency η_{om}, is

$$\eta_{om} = \frac{P_{om}}{P_{dc}} = \frac{1}{2} = 50\% \qquad [5]$$

The maximum output power can be increased by shifting the operating point away from the origin at a 45° angle, but the maximum efficiency remains constant.

The value of R_L for both maximum efficiency and maximum output power for any operating point is always

$$R_{Lopt} = \frac{V_Q}{I_Q} \quad [6]$$

If R_L is lower than R_{Lopt}, the maximum current swing remains at $2I_Q$ but the voltage swing is reduced below $2V_Q$. The operation of the device can be said to be current limited, and the output power is

$$P_o = \frac{I_Q{}^2 R_L}{2} \; (R_L < R_{Lopt}) \quad [7]$$

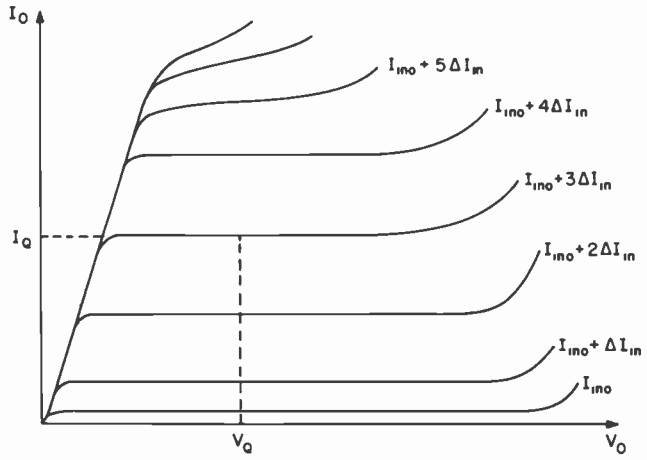

Fig. 3—Current-amplifying device.

Similarly if R_L is greater than R_{Lopt}, the device becomes voltage limited and

$$P_o = \frac{V_Q{}^2}{2R_L} \; (R_L > R_{Lopt}) \quad [8]$$

In both cases since $P_o < P_{om}$ the efficiency and saturation level will decrease.

A real device used for transfer amplification is not ideally linear. Higher order terms in the transfer characteristic of the device result in an output characteristic similar to that shown in Fig. 3.

The device may be considered to be linear if it is operated in a very small region on the output characteristic around the bias point. As the signal amplitude is increased, the output will become increasingly dis-

torted. Under very large signal conditions the transfer characteristic would have to be represented by an infinite series

$$i_o = \sum_{n=1}^{\infty} k_n (i_{in})^n. \qquad [9]$$

In general the coefficients k_n will be complex quantities whose values can be determined by an analysis of the output waveforms. Small-signal operating conditions, i.e., operation in the small region about the bias point, are represented by Eq. [9] when all the coefficients of the expansion are zero except k_1. Then

$$i_{o_{ss}} = k_1 i_{in} \qquad [10]$$

Eq. [10] is used as a linear approximation for small-signal circuit analysis and design.

An important region of operation is the one in which small-signal operation is assumed even though the signal levels have increased enough to produce slight distortion. An amplifier of this type is usually referred to as "linear". To analyze this mode of operation, three terms of Eq. [9] are used

$$i_o = k_1 i_{in} + k_2 i_{in}^2 + k_3 i_{in}^3. \qquad [11]$$

This is the first-order approximation needed to extract information about the intermodulation distortion and is assumed to be valid for small deviations from linear operation.

The gain equation for transfer voltage-amplifying devices is the dual of Eq. [11] with the currents replaced by voltages.

The gain of negative-resistance amplifiers in which the output power is converted from a dc supply, e.g., tunnel diodes, can also be represented by Eq. [10]. In this case, the gain is linear as long as the negative resistance remains constant for the signal voltage or current swing. If the voltage or current swing exceeds the constant-negative-resistance region, the waveform will distort. Fig. 4 shows a typical negative-resistance characteristic. If the negative resistance was constant between V_{min} and V_{max}, then

$$P_{om} = \frac{(V_{max} - V_{min})(I_{max} - I_{min})}{8} = \frac{R_L}{8}(I_{max} - I_{min})^2 = \frac{(V_{max} - V_{min})^2}{8R_L} \qquad [12]$$

and

$$V_Q = \frac{V_{max} + V_{min}}{2}$$

$$I_Q = \frac{I_{max} + I_{min}}{2}$$

The efficiency is

$$\eta_{om} = \frac{P_{om}}{P_{dc}} = \frac{(V_Q - V_{min})(I_Q - I_{min})}{2V_Q I_Q} \qquad [13]$$

Note that as V_{min} and I_{min} approach zero, the maximum efficiency approaches 50%.

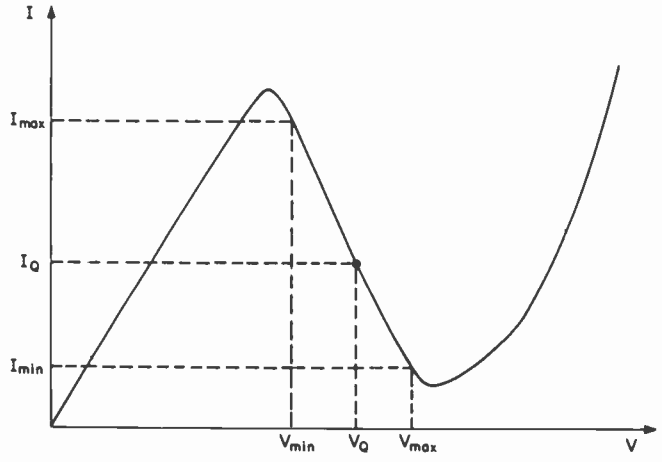

Fig. 4—Negative-resistance device.

3. Gain Saturation for Multiple Signals

As the amplifier deviates slightly from linearity, the gain can be represented in the general form

$$y = k_1 f(x) + k_2 f(x)^2 + k_3 f(x)^3 \qquad [14]$$

where y represents the output and $f(x)$ the input. In general, the transform of each coefficient K_n will be complex quantities whose values can be determined by an analysis of the output waveforms or by using a Voltera series approach.[2]

Simons[1] has investigated this equation for three sinusoidal inputs and real coefficients. Here, we will assume complex coefficients. The input will consist of m sinusoidal signals,

$$f(X) = \sum_{n=1}^{m} A_n \cos X_n = \sum_{n=1}^{m} \mathrm{Re}(A_n e^{jX_n}) \qquad [15]$$

where A_n is the signal amplitude and

$$X_n = 2\pi f_n t + \phi_n$$

Substitution of Eq. [15] into [14] results in a series of sinusoidal terms that contribute to an output at the m fundamental frequencies, their second and third harmonics, and to a variety of mixing frequencies between the fundamentals and harmonics.

The fundamental output consists of

$$y_{fund} = \sum_{n=1}^{m} \mathrm{Re}\left\{\left[K_{1,f_n} + \frac{3}{4} K_{3,f_n}\left(A_n{}^2 + 2\sum_{p=1,p\neq n}^{m} A_p{}^2\right)\right] \times A_n e^{jX_n}\right\}. \quad [16]$$

Note that the terms K_{1,f_n}, K_{3,f_n} indicate that the complex coefficients K_1, K_3 are to be evaluated at frequency f_n; they do *not* indicate that K_1, K_3 are functions of f_n exclusively. The contributions of the first and third terms of Eq. [14] are represented by K_{1,f_n} and K_{3,f_n}, respectively. The third-term contribution consists of two parts. The first, $(\frac{3}{4})K_{3,f_n}A_n{}^2$, represents self-saturation. The second part, $(\frac{2}{3})\ K_{3,f_n}A_p{}^2$, represents "cross-saturation".

At any one of the frequencies f_n, the ratio of the fundamental output amplitude at that frequency to the small signal fundamental output amplitude is the gain saturation, g_c,

$$g_{c,f_n} = 1 + \frac{3}{4}\frac{K_{3,f_n}}{K_{1,f_n}}\left(A_n{}^2 + 2\sum_{p=1,p\neq n}^{m} A_p{}^2\right) \quad [17]$$

Since K_3 and K_1 are complex quantities, gain compression will occur only if the relative phase of K_3 and K_1 is 180°.

The gain saturation of any fundamental output at frequency f_q can be related to the gain saturation of the largest amplitude signal. If the signal with the largest amplitude A_1 is at frequency f_1 and all the signal amplitudes are normalized so that

$$A_p = a_p A_1, \qquad (p \neq 1) \quad [18]$$

then,

$$g_{c,f_q} = (g_{c,f_1} - 1)\frac{K_{1,f_1}}{K_{1,f_q}}\frac{K_{3,f_q}}{K_{3,f_1}}\left(\frac{2 + a_q{}^2 + 2\sum_{p=3}^{m} a_p{}^2}{1 + 2a_q{}^2 + 2\sum_{p=3}^{m} a_p{}^2}\right) + 1. \quad [19]$$

The gain saturation of each signal is different and is a function of their relative levels.

If only one high level signal is present among the m signals, then

$$g_{c,f_q} \approx (g_{c,f_1} - 1)\frac{K_{1,f_1}}{K_{1,f_q}}\frac{K_{3,f_q}}{K_{3,f_1}}2 + 1, \qquad [19a]$$

where

$$g_{c,f_1} = 1 + \frac{3}{4}\frac{K_{3,f_1}}{K_{1,f_1}}A_1^2, \qquad [19b]$$

which is the gain saturation if the large signal is the only one present. The gain saturation for all the small signals is approximately the same but differs from the saturation of the large signal by the relationship shown in Eq. [19a].

. If all the signal amplitudes are the same then

$$g_{c,f_q} = (g_{c,f_1} - 1)\frac{K_{1,f_1}}{K_{1,f_q}}\frac{K_{3,f_q}}{K_{3,f_1}} + 1 \qquad [20]$$

where

$$g_{c,f_1} = 1 + \frac{3}{4}\frac{K_{3,f_1}}{K_{1,f_1}}A_1^2(2m - 1). \qquad [21]$$

Note that g_c is a complex quantity, i.e., it has both magnitude and phase.

4. Intermodulation Distortion

The third term of the gain equation, Eq. [14], produces outputs at frequencies $2f_n \pm f_p$. These outputs are commonly referred to as intermodulation distortion. They appear to be caused by the mixing of the second harmonic of one signal and the fundamental of another. This observation may result in the erroneous conclusion that if the second harmonic is eliminated, intermodulation distortion will be eliminated. Second-harmonic reduction may be accomplished by decreasing the magnitude of K_2, the coefficient of the second term of Eq. [14]. In a practical situation, this is achieved by using a balanced configuration. If this decrease in the magnitude of K_2 does not affect the magnitude of K_3, then the outputs at the intermodulation distortion frequencies are not affected.

The intermodulation distortion outputs are

$$y_{imd} = \frac{3}{4}\sum_{n=1,p=1,p\neq n}^{m} \mathrm{Re}[K_{3,2f_n\pm f_p}A_n^2A_p \exp\{j(2X_n \pm X_p)\}] \qquad [22]$$

If the input consists of m signals, the output contains $2m(m-1)$ frequencies at which intermodulation distortion occurs.

The intermodulation distortion ratio, imr, is defined as the ratio of the amplitude of the intermodulation distortion, to the amplitude of its respective fundamental or desired output. Therefore, the imr at frequency $2f_n - f_p$ with respect to the output at frequency f_n is

$$\text{imr}_{2f-f_p} = \frac{3}{4}\frac{K_{3,2f_n-f_p}}{K_{1,f_n}} A_n A_p. \qquad [23]$$

Similarly, the imr with respect to output at frequency f_p is

$$\text{imr}_{2f_p-f_n} = \frac{3}{4}\frac{K_{3,2f_p-f_n}}{K_{1,f_p}} A_n A_p. \qquad [24]$$

If

$$K_{3,2f_p-f_n} = K_{3,2f_n-f_p} \text{ and } K_{1,f_p} = K_{1,f_n},$$

then $\text{imr}_{2f_p-f_n} = \text{imr}_{2f_n-f_p}$ independent of the relative amplitudes of A_n and A_p. Under these conditions, the intermodulation distortion ratio between a signal and its associated distortion due to a second signal is identical to the intermodulation distortion ratio between the second signal and its associated distortion due to the first signal.

If the signal amplitudes are normalized to the largest signal amplitude, A_1, then

$$\text{imr}_{2f_n-f_p} = \frac{3}{4}\frac{K_{3,2f_n-f_p}}{K_{1,f_n}} a_n a_p A_1^2, \qquad [25]$$

where

$$A_n = a_n A_1,\ A_p = a_p A_1$$

Eq. [25] reduces to

$$\text{imr}_{2f_n-f_p} = \frac{3}{4}\frac{K_{3,2f_n-f_p}}{K_{1,f_n}} A_1^2 \qquad [26]$$

if all the signals are equal level. The ratio of the imr to the equal level imr is

$$\left.\frac{\text{imr}}{\text{imr}_{\text{eq.}}}\right|_{2f_n-f_p} = a_n a_p \qquad [27]$$

Therefore the improvement in imr as the signal level is decreased is proportional to the product of the reductions of the signal levels. In dB, the intermodulation distortion ratio improvement is equal to the sum of the reductions of the signals. For example, if two equal tone signals produce an intermodulation distortion ratio of −40 dB, then a reduction

of one signal by 10 dB will reduce the intermodulation distortion ratio to −50 dB.

Eqs. [27] and [24] provide the intermodulation distortion ratio for any signal level; the intermodulation distortion ratio for equal signal levels is known.

The relationship between gain saturation and intermodulation distortion can be seen by examining Eqs. [17] and [23]. The gain saturation for the largest of m signals is

$$g_{c,f_1} = 1 + \frac{3}{4}\frac{K_{3,f_1}}{K_{1,f_1}} A_1^2 \left(1 + 2\sum_{p=2}^{m} a_p^2\right), \qquad [28]$$

and the intermodulation distortion ratio with respect to this and any other signal at f_q is,

$$\text{imr}_{2f_1-f_q} = \frac{3}{4}\frac{K_{3,2f_1-f_q}}{K_{1,f_1}} A_1^2 a_q. \qquad [29]$$

Eliminating A_1 from both equations results in

$$\text{imr}_{2f_1-f_q} = \frac{a_q}{1 + 2\sum_{p=2}^{m} a_p^2}\frac{K_{3,2f_1-f_q}}{K_{3,f_1}}(g_{c,f_1} - 1). \qquad [30]$$

Similarly the imr produced by any other two signals can be related to the gain saturation of the largest signal as

$$\text{imr}_{2f_r-f_s} = \frac{a_r a_s}{\left(1 + 2\sum_{p=2}^{m} a_p^2\right)}\frac{K_{3,2f_r-f_s}}{K_{1,f_1}}(g_{c,f_1} - 1) \qquad [31]$$

or to the gain saturation of the fundamental output which produced the imr using Eqs. [31] and [19].

A very special case, one which is normally used for test purposes, is the two equal tones case. If just two signals of equal amplitude are used, then

$$\text{imr}_{2f_1-f_2} = \frac{1}{3}\frac{K_{3,2f_1-f_2}}{K_{3,f_1}}(g_{c,f_1} - 1) \qquad [32a]$$

$$\text{imr}_{2f_2-f_1} = \frac{1}{3}\frac{K_{3,2f_2-f_1}}{K_{3,f_2}}(g_{c,f_2} - 1) \qquad [32b]$$

These equations describe the imr as a function of the saturation of one of two signals. If the K_1 and K_3 coefficients are the same at frequencies f_1 and f_2, then $g_{c,f_1} = g_{c,f_2}$ and both signals will have the same degree of saturation. However, if both signals are replaced by a single signal of the

same total power, the degree of saturation will not be the same. The level of a single signal, A_1', that causes the same degree of gain saturation as the level of either of the two signals, A_1, is found by equating the gain saturation for a single signal, g_{c,f_1}', to the gain saturation when two signals are present g_{c,f_1}. Both g_{c,f_1}' and g_{c,f_1} are obtained from Eq. [28]. If

$$g_{c,f_1}' = g_{c,f_1},$$

then

$$(A_1')^2 = 3(A_1)^2 \qquad [33]$$

Therefore a single signal requires three times more power, or 4.77 dB, than the power in either of the two signals to reach the same degree of gain saturation. In terms of total power, the power of a single signal is 3/2 times, or 1.76 dB, greater than the total power of the two signals for the same degree of gain saturation.

The intermodulation distortion ratio can be related to the single signal gain saturation, g_c', by noting that the total power of two signals must equal the power of the single signal,

$$(A_1')^2 = 2A_1^2. \qquad [34]$$

Since

$$g_c' = 1 + \frac{3}{4}\frac{K_{3,f_1}}{K_{1,f_1}}(A_1')^2 = 1 + \frac{3}{4}\frac{K_{3,f_1}}{K_{1,f_1}}2A_1^2,$$

solving for A_1^2 and substituting this value in Eq. [26] results in

$$\text{imr}_{2f_n-f_p} = \frac{1}{2}\frac{K_{3,2f_n-f_p}}{K_{3,f_n}}(g_c' - 1). \qquad [35]$$

Eq. [35] provides the intermodulation distortion ratio for two equal-level signals whose total power is the same as a single signal that produces a gain saturation of g_c'.

The intermodulation distortion ratio for two equal level signals, each of which has a power that is equal to a single signal that produces a gain saturation of g_c' is

$$\text{imr}_{2f_n-f_p} = \frac{K_{3,2f_n-f_p}}{K_{3,f_n}}(g_c' - 1). \qquad [35a]$$

5. Triple-Beat Distortion

One of the most common methods of measuring distortion is the two-tone intermodulation test. Two equal-level signals at frequencies f_1 and f_2 are applied to the input of an amplifier or mixer and the intermodulation distortion at $2f_2 - f_1$ and $2f_1 - f_2$ is measured.

Another test, which seems to have gained acceptance in the CATV industry, is the triple-beat test. This test is performed by applying three signals at frequencies f_1, f_2, and f_3 at the input of an amplifier. f_1 and f_2 are usually quite close together, while f_3 is spaced far away from both of them. The measured distortion component is the output at $f_3 + (f_2 - f_1)$. This output is commonly referred to as the third-order distortion. Since this nomenclature is too general for this discussion, it will be referred to here as the triple-beat.[1]

If there are m input signals present, the output will contain

$$4\binom{m}{3} = 4\,\frac{m!}{(m-3)!3!}$$

triple-beat frequencies or $\binom{m}{3}$ triple-beat outputs of the form,

$$y_{tb} = \mathrm{Re}\left[\frac{3}{2}K_{3,n\pm p\pm q}A_nA_pA_q\exp\{j(X_n \pm X_p \pm X_q)\}\right] \quad [36]$$

This output is also derived from the third term of Eq. [14]. The amplitude of the output at any one of the triple-beat frequencies is

$$\mathrm{tb}_{f_n+f_p-f_q} = \frac{3}{2}K_{3,f_n+f_p-f_q}A_nA_pA_q. \quad [37]$$

Normalizing the signal amplitudes to the largest at frequency f_1, results in

$$\mathrm{tb}_{f_n+f_p-f_q} = \frac{3}{2}K_{3,f_n+f_p-f_q}a_na_pa_qA_1{}^3. \quad [38]$$

The triple-beat ratio with respect to the output at f_n is

$$\mathrm{tbr}_{f_n+f_p-f_q} = \frac{3}{2}\frac{K_{3,f_n+f_p-f_q}}{K_{1,f_n}}a_pa_qA_1{}^2. \quad [39]$$

If just two signals at f_n and f_p are used, the intermodulation distortion ratio with respect to the output at f_n is (from Eq. [25])

$$\mathrm{imr}_{2f_n-f_p} = \frac{3}{4}\frac{K_{3,2f_n-f_p}}{K_{1,f_n}}a_na_pA_1{}^2. \quad [40]$$

Therefore the relationship between the intermodulation distortion and the triple-beat distortion is

$$\mathrm{tbr}_{f_n+f_p-f_q} = 2\,\frac{K_{3,f_n+f_p-f_q}}{K_{3,2f_n-f_p}}\frac{a_q}{a_n}\,\mathrm{imr}_{2f_n-f_p} \quad [41]$$

If the amplitudes of the signals at f_q and f_n are equal or if equal-level

signals are used for both tests and $K_{3,f_n+f_p-f_q} = K_{3,2f_n-f_p}$, then the triple-beat ratio is exactly twice the intermodulation distortion ratio.

$$\text{tbr}_{f_n+f_p-f_q} = 2\ \text{imr}_{2f_n-f_p}. \qquad [42]$$

6. Cross-Modulation Distortion

Cross-modulation distortion is the transfer of modulation from one signal to another. It is also derived from the third term of Eq. [14] because it is a special case of triple-beat distortion.

Let the input to the amplifier consist of a single tone amplitude modulated signal and an unmodulated carrier, i.e.,

$$f(x) = A_c\left(1 + \frac{A_m}{A_c}\cos X_m\right)\cos X_c + A_n \cos X_n$$

$$= A_c \cos X_c + \frac{A_m}{2}\cos (X_c + X_m) + \frac{A_m}{2}\cos (X_c - X_m)$$

$$+ A_n \cos X_n. \qquad [43]$$

The modulated and unmodulated signals may be considered four individual signals of the form

$$f(x) = A_1 \cos X_1 + A_2 \cos X_2 + A_3 \cos X_3 + A_4 \cos X_4$$

where

$$A_1 = A_c,\ A_2 = A_3 = \frac{A_m}{2},\ A_4 = A_n$$

and

$$X_1 = X_c,\ X_2 = X_c + X_m,\ X_3 = X_c - X_m,\ X_4 = X_n$$

At the output, the cross-modulation distortion appears at $X_n + X_m$ and at $X_n - X_m$. Since

$$X_n + X_m = X_4 + X_2 - X_1 = X_4 + X_1 - X_3$$
$$X_n - X_m = X_4 - (X_2 - X_1) = X_4 - (X_1 - X_3),$$

it is easily seen that the cross-modulation sidebands are simply the triple-beat distortion of the four signals. Since the signals at X_2 and X_3 are equally spaced from X_1, the triple-beat at $X_n + X_m$ or $X_n - X_m$ is composed of two parts as shown in Fig. 5.

The triple-beat distortion is

$$\text{tb}_{f_n+f_m} = \text{tb}_{f_4+f_2-f_1} + \text{tb}_{f_4+f_1-f_3}$$
$$\text{tb}_{f_n-f_m} = \text{tb}_{f_4-f_2+f_1} + \text{tb}_{f_4-f_1+f_3}. \qquad [44]$$

The outputs at $f_n + f_m$ and $f_n - f_m$ are, from Eq. [36],

$$y_{tb,f_n+f_m} = \mathrm{Re}\left[\frac{3}{2} A_n A_c A_n K_{3,f_n+f_m} \exp\{j(X_n + X_m)\}\right] \quad [45a]$$

$$y_{tb,f_n-f_m} = \mathrm{Re}\left[\frac{3}{2} A_n A_c A_m K_{3,f_n-f_m} \exp\{j(X_n - X_m)\}\right] \quad [45b]$$

The output at f_n with its newly created sidebands is

$$y = y_{f_n} + y_{tb,f_n+f_m} + y_{tb,f_n-f_m}$$

$$= \mathrm{Re}\left(K_{1,f_n} A_n \left[1 + \mathrm{Re}\left\{\frac{3A_c A_m}{2} \times \left(\frac{K_{3,f_n+f_m} + K_{3,f_n-f_m}}{K_{1,f_n}}\right) e^{jX_m}\right\}\right] e^{jX_n}\right)$$

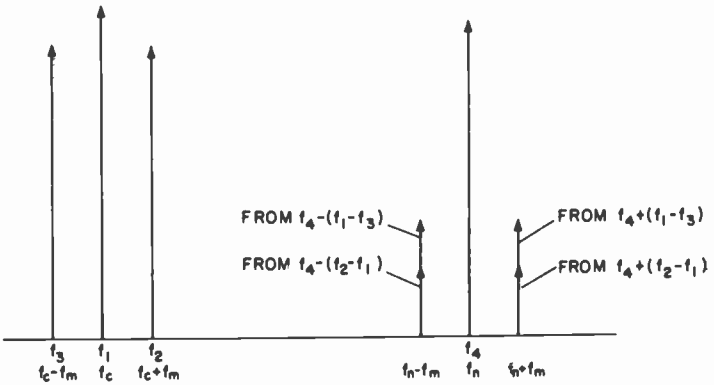

Fig. 5—Cross-modulation distortion.

The degree of modulation, M, on the original modulated signal at f_c was

$$M \equiv \frac{A_m}{A_c} \quad [46]$$

The degree of cross modulation, M_c, on the originally unmodulated carrier at f_n is

$$M_c = \frac{3}{2} A_c A_m \frac{K_{3,f_n+f_m} + K_{3,f_n-f_m}}{K_{1,f_n}}$$

$$= \frac{3}{2} M A_c{}^2 \frac{K_{3,f_n+f_m} + K_{3,f_n-f_m}}{K_{1,f_n}}. \quad [47]$$

The intermodulation distortion for two unmodulated carriers whose level is equal to those in the cross-modulation analysis, i.e., A_c and A_n is

$$\text{imr}_{2f_n-f_c} = \frac{3}{4}\frac{K_{3,2f_n-f_c}}{K_{1,f_n}} a_n A_c{}^2.$$

The cross modulation in terms of intermodulation distortion at $2f_n - f_c$ is

$$M_c = 2M\frac{K_{3,f_n+f_m} + K_{3,f_n-f_m}}{K_{3,2f_n-f_c}}\frac{1}{a_n}\,\text{imr}_{2f_n-f_c}. \qquad [48]$$

If an equal-tone imr test is performed with levels equal to the level of the modulated carrier and $K_{3,f_n+f_m} = K_{3,f_n-f_m} = K_{3,2f_n-f_c}$, then

$$M_c = 2M\,\text{imr}_{2f_n-f_c}|_{equal\ level} \qquad [49]$$

Where more than one interfering modulated signal is present, the cross modulation is the vector sum of each modulating component. This result was also arrived at using a Voltera series approach.[2,3] If m equal-level interfering signals are present and the modulation on all m signals are in phase, then the cross modulation due to these m signals, $M_{c,m}$, is

$$M_{c,m} = mM_c \qquad [50]$$

The combined voltage or current of the m signals must not violate the conditions for Eq. [14].

In some cases, the measurement of cross-modulation distortion is made by making the amplitude of each interfering signal at the peak of the modulation cycle equal to the amplitude of the unmodulated test signal. Therefore

$$A_n = A_c(1 + M).$$

Under these conditions, the intermodulation distortion ratio becomes

$$\text{imr}_{2f_n-f_c} = \frac{3}{4}\frac{K_{3,2f_n-f_c}}{K_{1,f_n}}(1 + M)^2 A_c{}^2$$

and the cross-modulation distortion becomes

$$M_c = 2\frac{M}{(1 + M)^2}\frac{K_{3,f_n+f_m} + K_{3,f_n-f_m}}{K_{3,2f_n-f_c}}\,\text{imr}_{2f_n-f_c}. \qquad [51]$$

The cross-modulation distortion obtained in this method of measurement is lower than the distortion obtained when the amplitudes of

the two signals for the intermodulation distortion test are made equal to the interfering signal's carrier amplitude.

It should also be noted that the National Community Television Association (NCTA) definition of the degree of cross-modulation distortion, X_m, is defined as the peak-to-peak variation of the test signal, as a result of cross modulation, to its amplitude with the interfering signal removed. Therefore

$$X_m = 2M_c. \qquad [52]$$

7. Modulation Distortion

The previous discussions of distortion all entailed multiple carrier signals. If a single modulated signal is applied to the input of an amplifier, the modulation will distort due to gain saturation. Since the modulated signal itself is made up of several frequency components, new frequency components will be generated.

Modulation distortion is a special case of intermodulation distortion. If a single tone modulated carrier is used, three frequency components are present. A resulting intermodulation distortion component will appear at $2(f_c + f_m) - f_c = f_c + 2f_m$, i.e., this component appears as the second harmonic of the modulation frequency.

Let the input to the amplifier be

$$f(x) = A_c(1 + \frac{A_m}{A_c} \cos X_m) \cos X_c, \qquad [53]$$

or

$$f(x) = A_1 \cos X_1 + A_2 \cos X_2 + A_3 \cos X_3$$

where

$$A_1 = A_c, A_2 = A_3 = \frac{A_m}{2},$$

and

$$X_1 = X_c, X_2 = X_c + X_m, X_3 = X_c - X_m.$$

Since three input frequencies are present there will be $2m(m-1) = 12$ intermodulation distortion frequencies. The six frequencies of importance in this analysis are:

$$\left.\begin{aligned} 2f_1 - f_2 &= f_c - f_m \\ 2f_2 - f_3 &= f_c + f_m \end{aligned}\right\} \text{ appear at the fundamental frequency of the modulation}$$

$$\left.\begin{aligned} 2f_3 - f_1 &= f_c - 2f_m \\ 2f_2 - f_1 &= f_c + 2f_m \end{aligned}\right\} \text{ appear at the second harmonic of the modulation}$$

$$\left.\begin{aligned} 2f_3 - f_2 &= f_c - 3f_m \\ 2f_2 - f_3 &= f_c + 3f_m \end{aligned}\right\} \text{ appear at the third harmonic of the modulation}$$

The other six frequencies appear at the sum frequencies.

The intermodulation distortion for each of these cases, from Eq. [25] is,

$$\text{imr}_{f_c \pm f_m} = \frac{3}{4}\frac{K_{3,f_c \pm f_m}}{K_{1,f_c}} M A_c{}^2 \qquad [54a]$$

$$\text{imr}_{f_c \pm 2f_m} = \frac{3}{4}\frac{K_{3,f_c \pm 2f_m}}{K_{1,f_c \pm f_m}} M A_c{}^2 \qquad [54b]$$

$$\text{imr}_{f_c \pm 3f_m} = \frac{3}{4}\frac{K_{3,f_c \pm 3f_m}}{K_{1,f_c \pm f_m}} M^2 A_c{}^2 \qquad [54c]$$

The intermodulation distortion ratio for the distortion appearing at the modulation frequency is taken with reference to the carrier level, while those at the harmonics of the modulation frequency are taken with respect to the modulation amplitude.

If the modulation distortion, M_d, is defined as the ratio of the second harmonic of the modulation to the linearized output of the modulation, then

$$M_d \equiv \text{imr}_{f_c \pm 2f_m}. \qquad [55]$$

When this is related to an equal-level intermodulation distortion ratio (Eq. [27]), the result for the modulation distortion is

$$M_d = M \, \text{imr}_{f_c \pm 2f_m}\big|_{equal\ level} \qquad [56]$$

8. AM to PM Conversion

AM to PM conversion is a direct result of gain saturation. A single unmodulated signal applied to the input of an amplifier produces a gain saturation of

$$g_{c,f_n} = 1 + \frac{3}{4}\frac{K_{3,f_n}}{K_{1,f_n}} A_n{}^2.$$

If the phase of K_{3,f_n} is not equal to or 180° with respect to the phase of K_{1,f_n}, then the phase of the output signal (or g_{c,f_n}) will vary with input signal level. It is important to remember that the amplifier is operating under the conditions that make Eq. [14] valid, i.e., small deviations from linearity, so that the magnitude and phase of K_{3,f_n} and K_{1,f_n} are constant.

Large-signal operation implies that three terms in the expansion (Eq. [14]) are insufficient. The practical aspect of this assumption is that a small degree of distortion, such as clipping, appears on the output waveform accompanied by only a very small change in bias conditions. If this condition is violated, it becomes questionable as to whether the amplifier is still to be considered linear.

An amplitude-modulated signal will vary the phase of the output signal in accordance with its change in amplitude. Eq. [16] will provide the fundamental output of such a signal. In general, the output will be of the form

$$y_{fund} = f_c \cos (X_c + \phi_c) + f_{1s} \cos (X_c - X_m + \phi_{1s}) + f_{us} \times \cos (X_c + X_m + \phi_{us}),$$

where $f_c, f_{1s}, f_{us}, \phi_c, \phi_{1s}, \phi_{us}$ are functions of K_1, K_3, A_m and A_c. Since this signal may contain both in-phase and phase-quadrature components with respect to the carrier, it is both amplitude and narrow-band frequency (or, alternatively small-angle phase) modulated.

The AM to PM conversion is the ratio of the change in output phase angle to the change in input signal level, in dB, for a single unmodulated signal. The change in output phase angle is the phase of the gain saturation shifted 180°. Letting the phase of $K_{1,f_n} = 0°$, the change in phase is

$$\Delta\phi = -\tan^{-1}\left(\frac{(3/4)|K_{3,f_n}|A_n^2 \sin \phi_3}{|K_{1,f_n}| + (3/4)|K_{3,f_n}|A_n^2 \cos \phi_3}\right) \qquad [57]$$

where ϕ_3 is the phase angle of K_{3,f_n}.

9. Even-Symmetry Clipping

The previous analysis allowed K_1, K_2, and K_3 to be complex quantities. A single unmodulated carrier input to the amplifier would result in an output that is represented by a Fourier series. If the clipping of the waveform occurs so that the output waveform has even symmetry, as shown in Fig. 6a, then the Fourier series consists of cosine terms only.* Under these conditions K_1, K_2, and K_3, must be negative real coefficients† and gain saturation becomes pure gain compression.

If the amplifier has sufficient bandwidth to allow all of the distortion products to appear at the output terminals unperturbed, the K_1 coeffi-

* If the clipping is also symmetrical about the horizontal axis, i.e., the x axis, then the output waveform has half-wave symmetry and, as a result, does not contain even harmonics. This property is typical of balanced amplifiers.

† Positive real coefficients would produce output waveforms with accentuated peaks rather than clipped peaks.

cients are all equal and the K_3 coefficients are all equal. With these two conditions, the relationships derived in the preceding sections can be summarized as follows.

9.1 Gain Compression

(1) The relationship between the gain compression of any signal at frequency f_q is related to the gain compression of the largest signal at

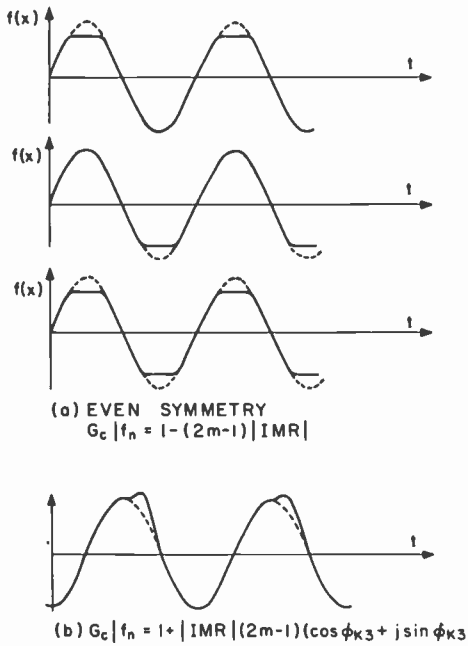

Fig. 6—Output waveforms.

frequency f_1 by

$$g_{c,f_q} = (g_{c,f_1} - 1)\left(\frac{2 + \left(\frac{A_q}{A_1}\right)^2 + 2\sum_{p=3}^{m}\left(\frac{A_p}{A_1}\right)^2}{1 + 2\left(\frac{A_q}{A_1}\right)^2 + 2\sum_{p=3}^{m}\left(\frac{A_p}{A_1}\right)^2}\right) + 1. \qquad [58]$$

If only one large level signal is present then

$$g_{c,f_q} = 2g_{c,f_1} - 1. \qquad [59]$$

(2) The gain compression, g_c, for multiple signals all of which are at the same level as a single signal which provides a compression of g_c' is

$$g_{c,m} = (2m - 1)g_c' - 2(m - 1) \quad [60]$$

where m is the number of signals present.
(3) The gain compression, g_c, for multiple signals the total power of which is equal to the power of a single signal that provides a compression of g_c' is

$$g_{c,m} = \left(\frac{2m - 1}{m}\right) g_c' - \frac{(m - 1)}{m} \quad [61]$$

9.2 Intermodulation Distortion*

(1) The intermodulation distortion ratio between any two signals is the same:

$$\text{imr}_{2f_n - f_p} = \text{imr}_{2f_p - f_n} \text{ (dB)}. \quad [62]$$

(2) The intermodulation distortion ratio for unequal-level signals may be obtained from the intermodulation distortion ratio of equal-level signals by multiplying the latter by the ratio of small signal to large signal amplitudes

$$\text{imr}_{2f_n - f_p}|_{A_n = A_p} = \text{imr}_{2f_n - f_p} - \frac{A_p}{A_n} \text{ (dB)}. \quad [63]$$

The amplitude of the intermodulation distortion referenced to the large signal, $\text{imr}_{2f_n - f_p}$, will be reduced by A_p/A_n (dB) and $\text{imr}_{2f_n - f_p}$ will be reduced by $2(A_p/A_n$ (dB)).
(3) The intermodulation distortion between any two signals is related to the gain compression of the largest signal at the frequency f_1 by

$$\text{imr}_{2f_r - f_s} = \frac{A_r}{A_1} + \frac{A_s}{A_1} + 20 \log |g_c(f_1) - 1|$$
$$- 20 \log \left(1 + 2 \sum_{p=2}^{m} \frac{A_p}{A_1}\right) \text{ (dB)}. \quad [64]$$

The intermodulation distortion ratio for two equal-level signals is related to the gain compression of each signal by

$$\text{imr}_{2f_n - f_p}|_{A_n = A_p} = 20 \log |g_{c,f_n} - 1| - 9.54 \text{ (dB)}. \quad [65]$$

* The intermodulation distortion ratio is referenced to the amplitude of the signal at the frequency of the first subscript; e.g., $\text{imr}_{2f_n - f_p}$ is referenced to A_n, $\text{imr}_{2f_p - f_n}$ is referenced to A_p.

When each signal compresses 1 dB the intermodulation distortion ratio is −28.8 dB.

If a plot of output power versus input power for a two-tone intermod test is used to predict intermodulation distortion at other power levels, it is important to remember that the intermodulation distortion power level is 28.8 dB below the output level of just one of the signals. A superposition of intermodulation distortion on the power transfer characteristic of total input power and total output power requires a scale shift of 3 dB. This is depicted in Fig. 7. Note that if these two plots are intermixed, the intermodulation distortion will be in error by 9 dB.

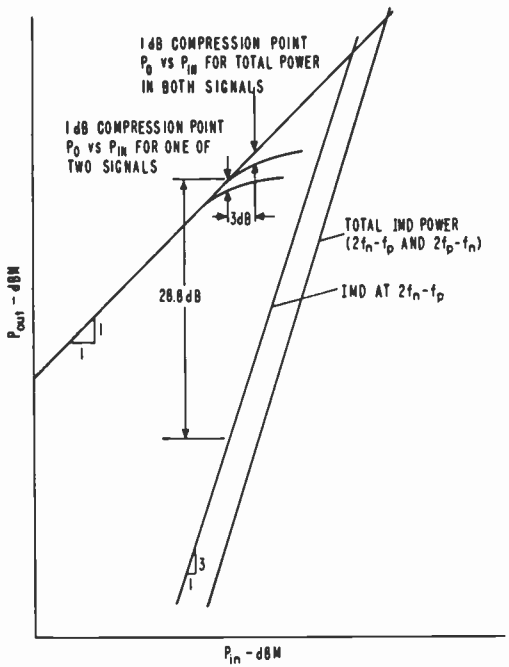

Fig. 7—Two-signal transfer characteristic.

The single-signal gain compression may also be used as a reference. In this case

$$\text{imr}_{2f_n-f_p}|_{A_n=A_p} = 20 \log |g_{c,f_n}' - 1| - 6 \text{ dB}. \quad [66]$$

Two equal-level signals whose total power is equal to the power level of a single signal that produces 1 dB of gain compression, will produce an intermodulation distortion ratio of −25.3 dB. Note that the actual gain

compression of either of these two signals is 1.55 dB. If each of the two signal levels is equal to the level of a single signal at a specified compression point, the intermodulation distortion ratio is

$$\text{imr}_{2f_n-f_p}\big|_{A_n=A_p} = 20 \log |g_{c,f_n}{}' - 1| \text{ dB}. \quad [67]$$

At the 1-dB compression point, the intermodulation distortion is −19.2 dB. The gain compression of the two signals is −3.43 dB.

(4) The analytic expression relating the intermodulation distortion ratio at any signal level to the 1-dB compression point is

$$\text{imr} = 2P_0 - P_{1dB} - 28.8 \text{ dB}, \quad [68]$$

where P_o is the power in one of the two equal level signals and P_{1dB} is the power in one of the two signals at the 1-dB compression point. Where P_{ot} is the total power of both the two equal-level signals and P_{om} is the single-signal 1-dB compression point

$$\text{imr} = 2P_{ot} - 2P_{om} - 25.3 \text{ dB} \quad [69]$$

Where P_o is the power in one of the two equal-level signals and P_{om} is the single-signal 1-dB compression point

$$\text{imr} = 2P_o - 2P_{om} - 19.3 \text{ dB} \quad [70]$$

The previously developed relationships can be used in conjunction with these equations to predict the intermodulation distortion for multiple unequal-level signals.

(5) The intercept point,[4] the point P_{int} at which the straight lines representing the fundamental output and the intermodulation distortion on a log-log plot intersect, is the linearized output power P_o when the imr is zero. Therefore

$$\text{imr} = 2P_o - 2P_{int}$$

where

$$\left.\begin{aligned} P_{int} &= P_{1dB} - 14.4 \\ P_{int} &= P_{om} - 9.65 \end{aligned}\right\} \text{ if } P_o \text{ is the power in each signal}$$

$$P_{int} = P_{om} - 12.65 \text{ if } P_o \text{ is the total power}$$

9.3 Triple-Beat Distortion*

(1) The triple-beat distortion is related to the intermodulation distortion by

* The triple-beat ratio is referenced to the amplitude of the signal at the frequency of the first subscript, e.g., $\text{tbr}_{f_n+f_p-f_q}$ is referenced to f_n.

$$\text{tbr}_{f_n+f_p-f_q} = \text{imr}_{2f_n-f_p} + 6 - \frac{A_q}{A_n}\,\text{dB}. \qquad [71]$$

9.4 Cross-Modulation Distortion

(1) The cross-modulation distortion between a modulated carrier of amplitude A_c and degree of modulation M and an unmodulated carrier of amplitude A_n is related to the intermodulation distortion of two unmodulated signals of amplitudes A_c and A_n by

$$M_c = \text{imr}_{2f_n-f_c} + M + 6 - \frac{A_n}{A_c}\,\text{dB}. \qquad [72a]$$

If $A_c = A_n$,

$$M_c|_{A_n=A_c} = \text{imr}_{2f_n-f_c}|_{A_n=A_c} + M + 6\,\text{dB}. \qquad [72b]$$

(2) The cross-modulation distortion due to m interfering signals, all of the same amplitude synchronously modulated, is

$$M_{c,mA_n} = M_c|_{A_n=A_c} + 20 \log m\ \text{dB}. \qquad [73]$$

(3) If the cross-modulation distortion test is performed by making the amplitude of each interfering signal at the peak of the modulation cycle equal to the amplitude of the unmodulated test signal, then

$$M_c|_{A_n=A_c+A_m} = M_c|_{A_n=A_c} - 40 \log (1 + M)\,\text{dB} \qquad [74]$$

Note that the difference between the two test results would be 7 dB for 50% modulation and 12 dB for 100% modulation.

(4) The National Community Television Association (NCTA) definition of cross modulation distortion increases the above values by 6 dB.

$$M_c|_{NCTA} = M_c|_{A_n=A_c+A_m}\,\text{dB} \qquad [75]$$

9.5 Modulation Distortion

(1) The modulation distortion is the intermodulation distortion ratio between the carrier and its sideband:

$$M_d = \text{imr}_{f_c \pm 2f_m}\,\text{dB}. \qquad [76]$$

In terms of an equal-level intermodulation distortion test in which both signal levels are the same as the carrier level,

$$M_d = \text{imr} + M\,\text{dB} \qquad [77]$$

10. Nonsymmetrical Clipping—AM to PM Conversion

The coefficients of Eq. [14] must be complex quantities if the output waveform does not have even symmetry. The gain characteristic can compress or expand depending upon the relative phase of K_1 and K_3. The object of this discussion is to determine the effect of phase on various distortion components.

Eq. [35], repeated here for convenience, provides the relationship between the intermodulation distortion ratio for two equal-level signals and the single-signal compression;

$$\text{imr}_{2f_1-f_2} = \frac{K_{3,2f_1-f_2}}{K_{3,f_1}} (g_c' - 1). \qquad [35]$$

Since g_c' is complex in this case, both its magnitude and phase are required to determine the intermodulation distortion.

If $|K_{3,2f_1-f_2}| = |K_{3,f_1}|$, Eq. [35] reduces to

$$\text{imr}_{2f_1-f_2} = g_c' - 1 \qquad [78]$$

The relationship between the intermodulation distortion and the magnitude and phase of the gain saturation may be easily recognized by

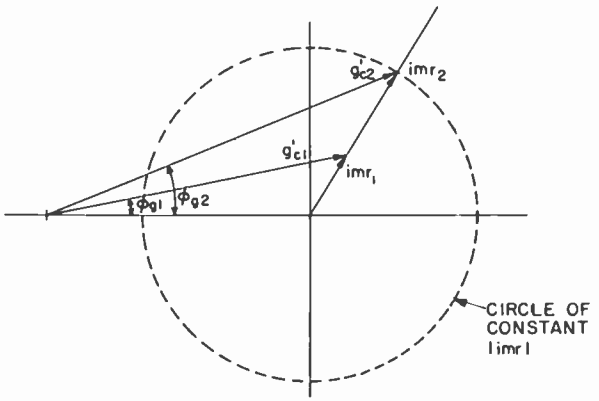

Fig. 8—Complex-plane representation.

graphically representing Eq. [78] in the complex plane as shown in Fig. 8. The locus of points representing a constant intermodulation distortion ratio is a circle centered at the origin with radius equal to the magnitude of the intermodulation distortion ratio. Eq. [26] indicates that the phase angle of $\text{imr}_{2f_1-f_2}$ must remain constant. Therefore, the locus of points

corresponding to $g_c' - 1$ is a straight line extending radially from the origin at an angle

$$\theta = \arctan\left(\frac{|g_c'| \sin \phi_g}{|g_c'| \cos \phi_g - 1}\right) \qquad [79]$$

where $|g_c'|$ and ϕ_g are the magnitude and phase of the gain saturation at any point.

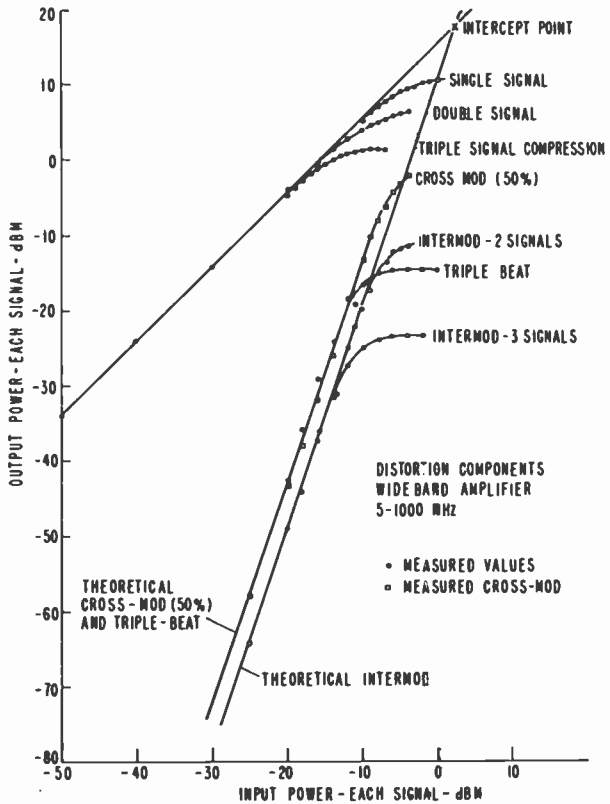

Fig. 9—Distortion components.

Since

$$\text{imr} = 2P_o - 2P_{1dB} - \text{imr}_{P_{1dB}}\,\text{dB}, \qquad [80]$$

the intermodulation ratio for any power level is known if it is known at one power level.

The result of Eqs. [79] and [80] is that if the value of $|g_c'|$ and its corresponding ϕ_g are known for only one power level, the values of $|g_c'|$ and ϕ_g and $|\text{imr}|$ can be determined for any other power level.

The intermodulation distortion ratio in terms of $|g_c'|$ and ϕ_g is

$$|\text{imr}_{2f_1-f_2}| = \sqrt{(|g_c'| \cos \phi_g - 1)^2 + (|g_c'| \sin \phi_g)^2} \qquad [81]$$

The gain compression in terms of the intermodulation distortion ratio is

$$|g_c'| = \sqrt{(|\text{imr}_{2f_1-f_2}| \cos \theta + 1)^2 + (|\text{imr}_{2f_1-f_2}| \sin \theta)^2}$$

$$\phi_g = \arctan \frac{|\text{imr}_{2f_1-f_2}| \sin \theta}{|\text{imr}_{2f_1-f_2}| \cos \theta + 1}. \qquad [82]$$

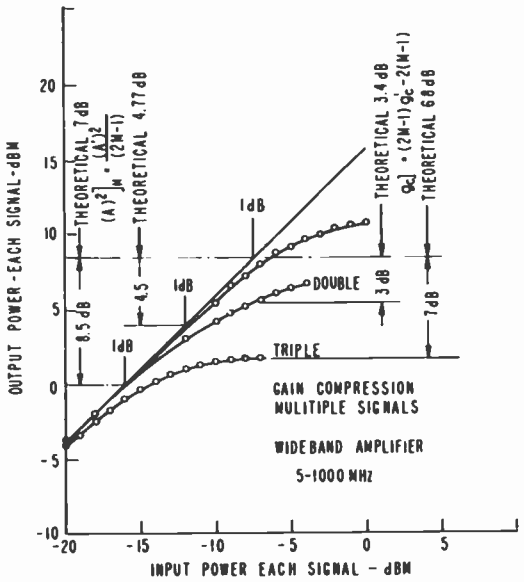

Fig. 10—Multiple signal compression.

If $|K_{3,2f_1-f_2}| \neq |K_{3,f_1}|$ then the ratio of the magnitudes of these two quantities must be determined. This is done by also obtaining the intermodulation distortion ratio for the lower level at which the $|g_c'|$ and ϕ_g are measured. The inequality of the K_3 coefficients does not affect the prediction of AM-PM conversion, i.e., ϕ_g, but will increase or decrease the intermodulation distortion ratio at all power levels by $|K_{3,2f_1-f_2}/K_{3,f_1}|$.

Eq. [80] is identical to Eq. [70] in the case of gain compression if the value of g_c' corresponds to the power level at which the $|\text{im}_1|$ is -19.3 dB. g_c' at this power level may be considered the equivalent of the 1-dB compression point.

11. Efficiency

The dynamic range of an amplifier or mixer represents the range of signal levels over which it will exhibit its intended signal-processing properties. If the upper level of this range is the 1-dB compression point (or its

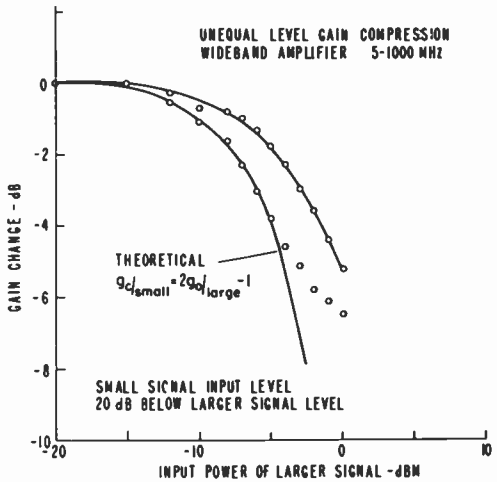

Fig. 11—Unequal-level signal compression.

equivalent in the case of AM to PM conversion), the maximum linear conversion efficiency may be defined as

$$(\eta_L)_{max} = \frac{P_{1dB}}{P_{dc}}, \qquad [83]$$

where P_{dc} is the power delivered by the fundamental source of power.

A device that has the ideally linear characteristic of Fig. 2 will provide a maximum linear efficiency of 50% if the optimum load is used. The degradation of the maximum linear efficiency for other loads is

$$\eta_L = (\eta_L)_{max} \frac{R_L}{R_{Lopt}} \quad \text{for} \frac{R_L}{R_{Lopt}} \leq 1$$

$$= (\eta_L)_{max} \frac{R_{Lopt}}{R_L} \quad \text{for} \frac{R_{Lopt}}{R_L} \leq 1 \qquad [84]$$

A real device that has nonlinearities will decrease the linear efficiency even further. An example is the bipolar transistor, which has saturation, high-current-compression, low-current-cutoff, and breakdown-voltage regions.

The most linear amplifier, i.e., the one that produces the least distortion at a specified power level is the amplifier that has the highest 1-dB compression point. If several devices can handle the same fundamental power, P_{dc}, then the device that has the highest linear efficiency

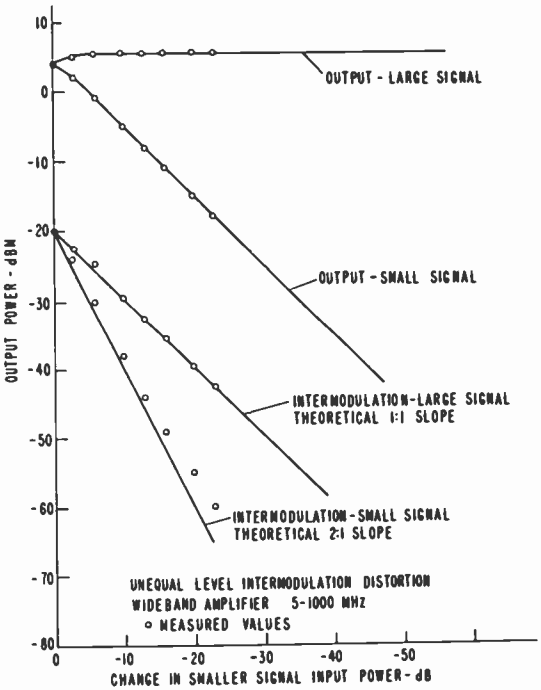

Fig. 12—Unequal-level signal intermodulation.

will provide the least distortion. The linear efficiency is a joint property of the device and its circuit environment, since it is a function of the load, R_L. Unfortunately, many applications dictate the value of R_L based upon requirements other than linearity. For example:

High-frequency gain; A device operated near the limit of its frequency-handling capability would most likely be matched for optimum gain.

Output VSWR; For various reasons an output isolator or balanced type (i.e., hybrid coupled) amplifier cannot be used. If output VSWR is re-

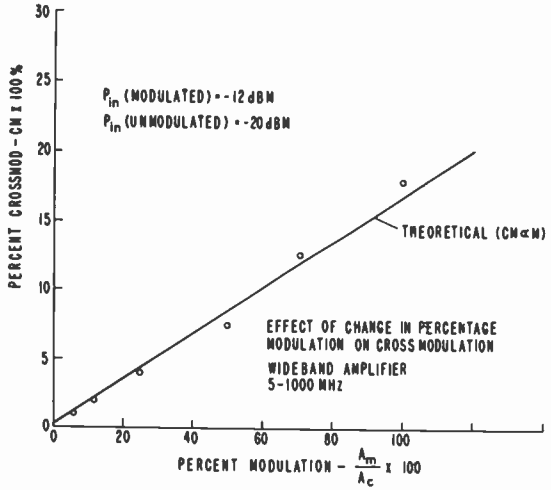

Fig. 13—Cross-modulation distortion.

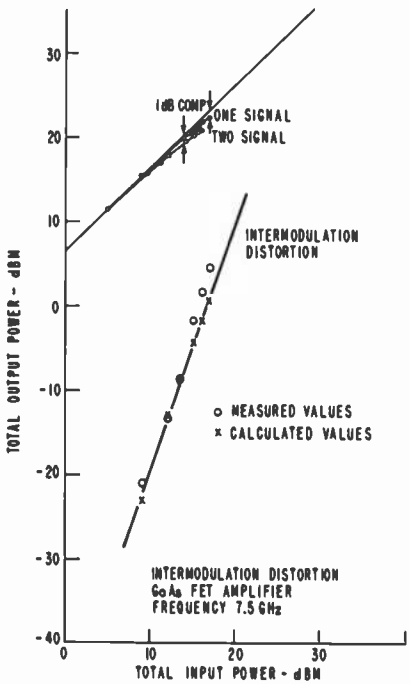

Fig. 14—GaAs FET amplifier.

quired under these circumstances, the output matching has VSWR as a criteria.

Low noise; In many applications the output matching circuit is used in conjunction with feedback to obtain the optimum impedance for noise figure at the input.

In all of these cases, a device that provides maximum linear efficiency for the load dictated by these other requirements is highly desirable. The best device need not and usually is not the same in all cases.

The efficiency may also be used to give an estimate of the various third-order-distortion components in the initial-design stage. If the maximum linear efficiency is assumed to be 50% and the required value of R_L is estimated, the 1-dB compression point can be calculated and used for evaluation of the distortion components. Conversely the optimum load and bias conditions can be estimated for some specified distortion and power level.

12. Mixers

Mixers, as a class of signal processors, may be treated in a similar development to that given above for amplifiers. However, a detailed analysis based upon actual device characteristics has already been performed on varactor up-converters[5] and resistive[6] mixers. It resulted in the following equation for the intermodulation distortion ratio.

$$\text{imr} = 2P_o - 2P_{om} - 19.5 \qquad [85]$$

This is almost identical to Eq. [70], which was developed independent of a particular device characteristic.

13. Experimental Results

A wideband amplifier was used to experimentally verify the relationship between the various third-order distortion components. Fig. 9 shows experimental data for gain compression, intermodulation, cross modulation, and triple beat distortion components. Fig. 10 provides a magnified view of the gain compression region indicating the theoretical relationship between the multiple signal 1-dB-gain compression points. Fig. 11 shows the compression of a small signal in the presence of a large one.

A very interesting experimental result, seen in Fig. 9, is the saturation of the intermodulation, triple-beat, and cross-modulation distortion. This is predictable if the distortion products are considered small signals in the presence of larger ones. For example, consider the intermodulation distortion, which is proportional to the third power of the input signal

level. On a dB–dB plot it has a slope of three. As the input signal level is increased, the output signals start to compress. The intermodulation distortion also starts to compress as if it were a small signal in the presence of two larger ones; however, the compression is relative to the 3:1 slope of the intermodulation rather than the unity slope of a real small signal present at the input of the amplifier. The gain compression of a small signal in the presence of m equal large signals is

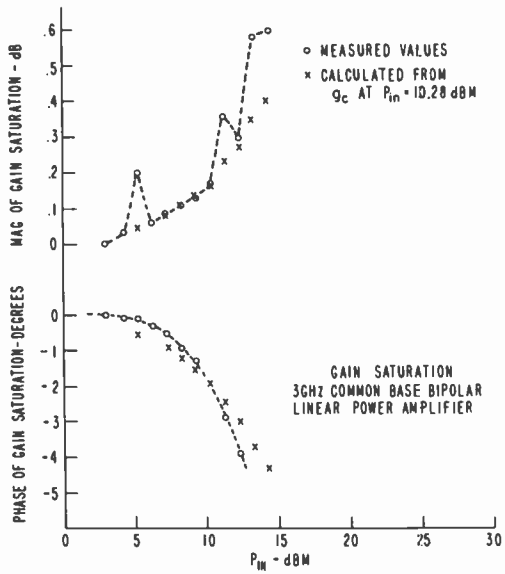

Fig. 15—Power amplifier.

$$g_{c.fq} = 2mg_c' - (2m - 1), \qquad [86]$$

where g_c' is the gain compression of a single signal whose amplitude is the same as each of the large ones. The gain compression of a small signal in the presence of a large signal is −4.96 dB at the small input level that causes a single signal to compress 1 dB. The experimental data shows that the intermodulation distortion is 4.0 dB below the theoretical 3:1 slope. As the number of signals increases[3], the saturation level of the intermod actually decreases as shown by comparing the intermod with two and with three signals present.

Fig. 12 shows how the intermodulation distortion ratio is equal for

unequal signal levels and how it decreases with changes in the level of one signal.

The effect of changing the percentage modulation on the cross modulation distortion is shown in Fig. 13. To check the validity of the results at microwave frequencies, data was obtained for a 7.5 GHz GaAs FET amplifier. These results are shown in Fig. 14.

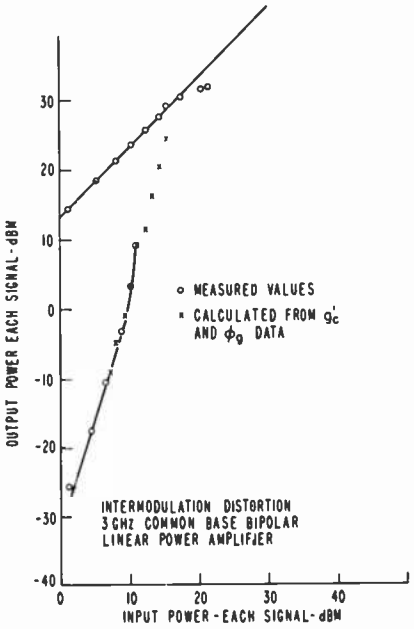

Fig. 16—Power amplifier intermodulation.

An excellent example of gain saturation and AM-PM conversion was observed in a 3-GHz common-base power amplifier. Using the data for $|g_c|$ and ϕ_g at one data point, the magnitude and angle of the intermodulation distortion ratio were calculated. The $|g_c|$ and ϕ_g were then calculated for other power levels. The resulting values are compared to the measured values in Fig. 15. Intermodulation distortion for this amplifier is shown in Fig. 16. Note the gain expansion.

Acknowledgments

The author would like to thank E. F. Belohoubek, F. N. Sechi, and G. E. Theriault for their many helpful discussions and R. L. Camisa for the experimental data on microwave amplifiers.

References:

[1] K. A. Simons, "The Decibel Relationships Between Amplifier Distortion Products," *Proc IEEE*, **58**, No. 7, p. 1071, July 1970.

[2] S. Narayan, "Application of Voltera Series to Intermodulation Distortion Analysis of Transistor Feedback Amplifiers," *IEEE Trans. Circuit Theory*, **CT-17**, p. 518, Nov. 1970.

[3] H. C. DeGraff and T. E. Winkle, "Relationship Between Cross Modulation and Intermodulation," *Electronics Lett.*, **8**, p. 33, Jan. 1972.

[4] F. C. McVay, "Don't Guess the Spurious Level," *Electronic Design*, p. 70, Feb. 1967.

[5] S. M. Perlow and B. S. Perlman, "A Large Signal Analysis Leading to Intermodulation Distortion Prediction in Abrupt Junction Varactor Up-Converters," *IEEE Trans. Microwave Theory and Tech.*, **MTT-13**, No. 6, p. 820, Nov. 1965.

[6] S. M. Perlow, "Intermodulation Distortion in Resistive Mixers," *RCA Review*, **35**, p. 25, March 1974.

Recent Papers by RCA Authors

Listing is alphabetical by name of primary author. For copies of reprints, the reader should contact the publication directly.

M. S. Abrahams, C. J. Buiocchi, J. F. Corboy, Jr., G. W. Cullen, "Misfit Dislocations in Heteroepitaxial Si on Sapphire," **Appl. Phys. Lett.**, Vol. 28, No. 5, p. 275, March 1, 1976.

R. C. Alig and S. Bloom, "Electron-Hole-Pair Creation Energies in Semiconductors," **Phys. Review Lett.**, Vol. 35, No. 22, p. 1522, Dec. 1, 1975.

K. Ametani, "Composition of Single Crystals of Chromium Chalcogenide Spinels Grown by Chemical Vapor Transport," **Bulletin of the Chem. Soc. of Japan**, Vol. 29, No. 2, p. 450, 1976.

V. S. Ban and S. L. Gilbert, "The Chemistry and Transport Phenomena of Chemical Vapor Deposition of Silicon from $CiCl_4$," **J. Crystal Growth**, p. 284, 1975.

J. Blanc, "Thermodynamics of 'Glide' and 'Shuffle' Dislocations in the Diamond Lattice," **The Philosophical Magazine**, Vol. 32, No. 5, p. 1023, Nov. 1975.

R. L. Camisa, J. Goel, and I. Drukier, "GaAs MESFET Linear Power-Amplifier Stage Giving 1W," **Electronics Lett.**, Vol. 11, No. 24, Nov. 27, 1975.

D. J. Channin and D. E. Carlson, "New Results on Triode Optical Gate Light Valves," **Tech. Digest of the 1975 Intern'l Electron Devices Meeting**, p. 398.

R. W. Cohen, G. D. Cody, M. D. Coutts, and B. Abeles, "Optical Properties of Granular Silver and Gold Films," **Phys. Rev. B**, Vol. 8, No. 8, p. 3689, Oct. 15, 1975.

R. S. Crandall and B. W. Faughnan, "The Dynamics of Coloration of Amorphous Electrochromic Films of WO_3 at Low Voltages," **Appl. Phys. Lett.**, Vol. 28, No. 2, p. 96, Jan. 15, 1975.

G. W. Cullen, J. F. Corboy, and R. T. Smith, "The Effect of Rapid Early Growth on the Physical and Electrical Properties of Heteroepitaxial Silicon," **J. of Crystal Growth**, Vol. 31, p. 274, 1975.

W. R. Curtice and P. T. Ho, "Analytical Analysis of Experimental TRAPATT Amplifiers and Oscillators," **Proc. of the IEEE**, p. 1731, Dec. 1975.

W. R. Curtice, "A Transferred-Electron Frequency Memorizer," **IEEE Trans. Microwave Theory & Techniques**, p. 1074, Dec. 1975.

W. Czaja, "Bose-Einstein Condensation and High Density Exciton Behaviour in AgBr," **Molecular Spectroscopy of Dense Phases—Proc. 12th European Congress on Molecular Spectroscopy**, p. 155, July 1–4, 1975.

R. V. D'Aiello, P. H. Robinson, and H. Kressel, "Epitaxial Silicon Solar Cells," **Appl. Phys. Lett.**, Vol. 28, No. 4, p. 231, Feb. 15, 1976.

E. Denlinger, J. Rosen, E. Mykietyn, and E. McDermott, "Microwave Varactor-Tuned Millimeter Impatt Diode Oscillators," **IEEE Trans. Microwave Theory and Techniques**, Vol. MTT-23, No. 12, p. 953, Dec. 1975.

M. Ettenberg, "Very Low Threshold Double Heterojunction $Al_xGa_{1-x}AS$ Injection Lasers," **Appl. Phys. Lett.**, Vol. 27, No. 12, p. 652, Dec. 15, 1976.

J. M. Hammer, "Metal Diffused Stripe Waveguides: Approximate Closed Form Solution for Lower Order Modes," **Optical Soc. of America**, Vol. 15, No. 2, p. 319, Feb. 1976.

J. M. Hammer, "Optical Grating Coupling Between Low Index Fibers and High Index Film Waveguides," **Appl. Phys. Lett.**, Vol. 28, No. 4, p. 192, Feb. 15, 1976.

G. Harbeke and E. Tosati, "Comment on 'The Band Gap Excitons in PbI_2'," **J. Phys. Chem. Solids**, Vol. 37, p. 126, 1976.

W. L. Harrington, R. E. Honig, A. M. Goodman, and R. Williams, "Low-Energy Ion-Scattering Spectrometry (ISS) of the SiO_2/Si Interface," **Appl. Phys. Lett.**, Vol. 27, No. 12, p. 644, Dec. 15, 1975.

T. T. Hitch, E. J. Conlon, and W. I. Rogers, "A Parametric Survey of the Adhesion Strength of Gold Thick Films on Alumina," **1975 Intern'l Microelectronics Symp. Proc.**, sponsored by ISHM, p. 287.

P. T. Ho, A. Rosen, and J. Klatskin, "Power Combination of Broadband Trapatt Amplifiers," **Electronics Lett.**, Vol. 12, No. 1, Jan. 8, 1976.

R. E. Honig, "Surface and Thin Film Analysis of Semiconductor Materials," **Thin Solids Films**, Vol. 31, p. 89, 1976.

H. C. Huang, I. Drukier, S. Y. Narayan, and R. L. Camisa, "GaAs MESFET Performance," **Tech. Digest 1975 Interna'l Electron Devices Meeting**, p. 235.

J. Kane, H. P. Schweizer, and W. Kern, "Chemical Vapor Deposition of Antimony-Doped Tin Oxide Films Formed from Dibutyl Tin Diacetate," **J. Electrochem. Soc.**, Vol. 123, No. 2, p. 270, Feb. 1976.

H. Kiess, "The Physics of Electrical Charging and Discharging of Semiconductors," **RCA Review**, Vol. 36, No. 4, p. 667, Dec. 1975.

H. Kressel, R. V. D'Aiello, and P. H. Robinson, "Epitaxial Solar Cells on Silicon EFG Ribbon Substrates," **Appl. Phys. Lett.**, Vol. 28, No. 3, p. 157, Feb. 1, 1976.

H. P. Lockwood, J. P. Wittke, and M. Ettenberg, "LED for High Data Rate, Optical Communications," **Optics Communications**, Vol. 16, No. 1, p. 193, Jan. 1976.

D. Meyerhofer, "Elastic and Dielectric Constants in Mixtures of Nematic Liquid Crystals," **J. Appl. Phys.**, Vol. 46, No. 12, p. 5084, Dec. 1975.
M. D. Miller, H. Schade, and C. J. Nuese, "Use of Platinum for Lifetime Control in Power Devices," **Tech. Digest 1975 Internat'l Electron Devices Meeting**, p. 180.
G. H. Olsen, "Interfacial Lattice Mismatch Effects in III–V Compounds," **J. Crystal Growth**, Vol. 31, p. 22, 1975.
G. H. Olsen and R. T. Smith, "Misorientation and Tetragonal Distortion in Heteroepitaxial Vapor-Grown III–V Structures," **Phys. Stat. Sol. (A)**, Vol. 31, p. 739, 1975.
R. J. Powell, "Radiation Induced Hole Transport and Electron Tunnel Injection in SiO_2 Films," **IEEE Trans. on Nuclear Science**, Vol. NS-22, No. 6, p. 2241, Dec. 1975.
W. Rehwald and G. K. Lang, "Ultrasonic Studies of Phase Transitions in the Tin Telluride Germanium Telluride System $Sn_xGe_{1-x}Te$," **J. Phys. C:Solid State Phys.**, Vol. 8, 1975.
E. S. Sabisky, P. J. Call, C. H. Anderson, "Tunable Low-Temperature Optical —EPR Gaussmeter," **Rev. Sci. Instrum.**, Vol. 46, No. 12, p. 1632, Dec. 1975.
J. Shefer, R. J. Klensch, H. C. Johnson, and G. S. Kaplan, "A New Kind of Radar for Collision Avoidance," **1974 SAE Transactions**, Sec. 1, Vol. 83, p. 481.
I. Shidlovsky, "The Mechanism of Color Center Production in Iron Doped Photochromic Sodalities," **Solid State Communications**, Vol. 18, p. 155, 1976.
P. D. Southgate, "Room-Temperature Poling and Morphology Changes in Pyroelectric Polyvinylidene Fluoride," **Appl. Phys. Lett.**, Vol. 28, No. 5, p. 250, March 1, 1976.
D. L. Staebler, "Binocularly Induced Motion of Flicker Patterns," **J. Opt. Soc. Am.**, Vol. 66, No. 2, p. 156, Feb. 1976.
E. F. Steigmeier, H. Auderset, and G. Harbeke, "The Central Peak in SbSi," **Phys. Stat. Sol. (b)**, Vol. 70, No. 2, p. 705, 1975.
E. F. Steigmeier and H. Auderset, "Evidence for a Low Frequency Defect Mode in Flux Grown $SrTiO_3$," **Physics Lett.**, Vol. 51A, No. 5, p. 285, March 24, 1975.
E. F. Steigmeier, R. Loudon, G. Harbeke, H. Auderset, and G. Scheiber, "Raman Scattering in $K_2Pt(CN)_4Br_{0.3}.3H_2O$," **Solid State Communications**, Vol. 17, p. 1447, 1975.
A. W. Stephens, J. L. Vossen, and W. Kern, "The Effect of Substrate Bias on the Properties of Reactively Sputtered Silicon Nitride," **J. Electrochem. Soc.**, Vol. 123, No. 2, p. 304, Feb. 1976.
C. W. Struck and W. H. Fonger, "Quantum-Mechanical Treatment of $Eu^{3+}4f \rightarrow 4f$ and $4f \rightleftarrows$ Charge-Transfer-State Transitions in Y_2O_2S and La_2O_2S," **J. Chemical Phys.**, Vol. 64, No. 4, p. 1784, Feb. 15, 1976.
T. Takahasi and O. Yamada, "Mechanism of Voltage Generation During Phase Change of Fe3 Single Crystals," **J. Phys. Chem. Solids**, Vol. 37, p. 161, 1976.
L. J. Vieland, "On the Cubic Rhombohedral Transformation in Magnetite," **Acta Crystallographica**, Vol. A31, part 6, p. 754, Nov. 1975.
Z. Turski, L. A. Kraus, and Ho-Chung Huang, "Properties of an (X)-Cut Single Crystal Lithium Niobate Transducer," **IEEE Trans. Sonics & Ultrasonics**, Vol. Su-22, No. 6, Nov. 1975.
H. S. Veloric, A. Presser, and F. J. Wozniak, "Ultra-Thin RF Silicon Transistors with a Copper-Plated Heat Sink," **RCA Review**, Vol. 36, No. 4, p. 731, Dec. 1975.
J. L. Vossen, "Inhibition of Chemical Sputtering of Organics and C by Trace Amounts of Cu Surface Contamination," **J. Appl. Phys.**, Vol. 47, No. 2, p. 544, Feb. 1976.
C. C. Wang and S. H. McFarlane III, "Crystal Growth and Defect Characterization of Heteroepitaxial III–V Semiconductor Films," **Thin Solid-Films**, Vol. 31, p. 3, 1976.
B. F. Williams, W. J. Burke, and D. L. Staebler, "Mobile Si Ions in Fe-Doped $LiNbO_3$ Crystal," **Appl. Phys. Lett.**, Vol. 28, No. 4, p. 224, Feb. 15, 1976.
R. O. Winder, "Microprocessors in Consumer Markets," **Computer Magazine**, p. 39, Jan. 1976.
J. P. Wittke, "Optical Fiber Communications Link Design," **Proc. of the SPIE**, 19th Annual Meeting, Vol. 63, p. 58, 1975.
J. P. Wittke, "Spontaneous-Emission-Rate Alteration by Dielectric and Other Waveguiding Structures," **RCA Review**, Vol. 36, No. 4, p. 711, Dec. 1975.

Patents Issued to RCA Inventors First Quarter 1976

January

Oded Ben-Dov Short End-Fire Circularly Polarized Antenna (3,932,876)
F. J. Campbell Electron Gun with a Multi-Element Electron Lens (3,932,786)
D. J. Channin Method of Testing an Electrical Circuit (3,934,199)
G. F. Dewdney Method and Apparatus for Securing Wires to Connection Terminals (3,930,606)
W. F. Dietz Start-Up Circuit for a Deflection System (3,936,115)
A. G. Dingwall Semiconductor Device Including a Conductor Surrounded by an Insulator (3,936,859)
D. P. Dorsey and W. E. Rodda Video Frequency Amplifier Operable in Either of Two Bias Conditions (3,936,597)
P. Foldes Multimode Coupling System Including a Funnel-Shaped Multimode Coupler (3,936,838)
A. Garcia Shortened Aperture Dipole Antenna (3,932,873)
W. F. Gehweler Voltage Controlled Oscillator Utilizing Field Effect Transistors (3,931,588).
M. Glogolja Protection Circuit (3,931,547)
L. D. Huff Lid Latch Mechanism for a Disc Record Player (3,930,653)
G. L. Hopkins Audio-Visual Apparatus with Control Signal Operated Gating Means (3,935,591)
H. C. Johnson Doppler Radar Sensor Calibrator (3,935,573)
S. T. Jolly Accurate Control During Vapor Phase Epitaxy (3,930,908)
H. G. Kiess and A. Rose Method of Reproducing an Electrostatic Charge Pattern in Intensified Form (3,934,180)
J. B. Klatskin and A. Rosen Method of Electrically Interconnecting Semiconductor Elements (3,932,226)
M. B. Knight Junction-Isolated Monolithic Integrated Circuit Device with Means for Preventing Parasitic Transistor Action (3,931,634)
W. F. Kosonocky Interlaced Readout of Charge Stored in a Charge Coupled Image Sensing Array (3,932,775)
P. A. Levine Scene Brightness Compensation System with Charge Transfer Imager (3,931,463)
P. A. Levine Blooming Control for Charge Coupled Imager (3,931,465)
A. Macovski Noninteracting Lens System for a Color Encoding Camera (RE 28702)
J. S. McNamee Connector for a Corrugated Conduit (3,934,902)
C. W. Mueller, E. C. Douglas, and C. P. Wu Method of Radiation Hardening and Gettering Semiconductor Devices (3,933,530)
J. A. Olmstead Improved Fluid Flow Sensor Configuration (3,931,736)
J. A. Olmstead Integrated Power Transistor with Ballasting Resistance and Breakdown Protection (3,936,863)
R. C. Palmer Video Disc Recording Apparatus and Methods (3,934,263)
D. R. Patterson Optical Switching System (3,931,514)
W. Phillips Method for the Reduction of Iron in Iron-Doped Lithium Niboate Crystals (3,932,299)
W. Phillips Photochromic Lithium Niobate and Method for Preparing Same (3,933,504)
J. J. Piascinski Conditioning Partially-Completed CRT Bulb Assembly for Storage and/or Transit (3,932,011)
E. S. Poliniak, R. J. Himics, and H. G. Scheible Preparation of Olefin SO2 Copolymer Electron Beam Resist Films and Use of Same for Recording (3,935,331)
E. S. Poliniak and R. J. Mimics Development of Poly(1-Methyl-1-Cyclopentene-SO2) Electron Beam Resist (3,935,332)
M. A. Polinsky Method of Vapor Deposition (3,934,059)

M. Rayl and H. D. Hanson Aerosol Detector (3,932,851)
C. F. Rose Variable Amplitude Timed Alarm System (3,931,621)
A. C. Sheng Proximity Switch Circuit (3,936,755)
T. D. Smith Ultrasonic Wave Transmitter Mechanism (3,934,544)
L. J. Thorpe Television Studio Control Apparatus (3,936,868)
M. Weiss Optical Reflector for Lasers (3,932,029)
C. F. Wheatley Amplifier with Fast Recovery After Input Signal Overswing (3,936,731)
O. M. Woodward Broadband Turnstile Antenna (3,932,874)

February

A. A. Ahmed Threshold Detector (3,937,987)
J. G. Amery Comb Filter for Video Processing (3,938,179)
J. Avins Automatic Luminance Channel Frequency Response Control Apparatus (3,938,181)
O. Ben-Dov Circularly Polarized, Broadside Firing Tetrahelical Antenna (3,940,772)
C. D. Boltz Timing Error Detecting and Speed Control System (3,940,556)
R. S. Crandall and B. W. Faughnan Electrochromic Device Having an Indium Electrode (3,940,205)
R. J. D'Amato Spring Mount for a Cathode Ray Tube Yoke (3,939,447)
S. B. Deal and D. W. Bartch Method for Preparing Haze-Resistant Lithium-Silicate Glare-Reducing Coating (3,940,511)
J. Fech, Jr. and E. S. Poliniak Electron Beam Recording Media and Method of Recording (3,940,507)
J. E. Keigler and L. Muhlfelder Re-Orientation of a Spacecraft Relative to its Angular Momentum (3,940,096)
R. D. Larrabee Suspension for Depositing a Pyroelectric Material (3,939,098)
A. J. Leidich Transistor Amplifier (3,938,054)
H. F. Lockwood Semiconductor Injection Laser (3,938,172)
D. W. Luz Deflection System (3,938,004)
H. R. Mathwich Transmitter for Frequency Shift Keyed Modulation (3,938,045)
J. C. Peer Class D Amplifier (3,939,380)
J. C. Peer Deflection System with Overscan Protection (3,940,661)
M. Pradervand Center Hole Formation in an Information Storing Disc (3,938,810)
S. M. Sherman and L. Goldman, Jr. Binaural Auditory Presentation Apparatus (3,940,769)
L. Sickles, II Doppler Correction of Transmission Frequencies (3,940,695)
H. A. Stern Method of Making Electrical Connections for Liquid Crystal Cells (3,936,930)
A. Sussman Fabrication of Liquid Crystal Devices (3,938,242)
L. A. Torrington and R. R. Oberle Lid Interlock Apparatus for Disc Record Player (3,940,148)

March

A. A. Ahmed Transient Suppressor (3,941,940)
O. Ben-Dov Circularly Polarized Antenna System Using a Combination of Turnstile and Vertical Dipole Radiators (3,943,522)
K. R. Bube Laser-Trimmed Resistor (3,947,801)
C. J. Busanovich, J. T. Fischer, R. M. Moore, and J. A. van Raalte Photoconductor of Cadmium Selenide and Aluminum Oxide (3,947,717)
J. Castleberry, Jr., E. P. Cecelski, and R. H. Storz Ultrasonic Digital Receiver (3,943,491)
D. E. Christensen Channel Number Memory for Television Tuners (3,947,773)
D. J. Dempsey Sidelobe Lock-On Discriminating Method for Search-Track Monopulse (3,943,512)
W. F. Dietz Side Pincushion Correction Circuit (3,946,274)

K. Feher Timing Technique for NRZ Data Signals (3,944,926)
N. Feldstein Method of Electroless Nickel Plating (3,946,126)
M. T. Gale and J. Kane Generation of Permanent Phase Holograms and Relief Patterns in Durable Media by Chemical Etching (3,944,420)
M. T. Gale and A. H. Firester Method for Producing Width-Modulated Surface Relief Patterns (3,945,825)
R. J. Giger and J. M. Lawrence Start-Up Power Supply for a Television Receiver (3,947,632)
N. Goldsmith and P. H. Robinson Method of Growing Thick Epitaxial Layers of Silicon (3,945,864)
J. R. Hall Controlled Gain Amplifier (3,942,129)
R. E. Hanson and T. E. Nolan, Jr. Power Test Means and Methods for Internal Combustion Engines (3,942,365)
J. C. Hartmann Noncontacting Marker (3,943,527)
F. Z. Hawrylo and H. Kressel Metallized Device and Method of Fabrication (3,945,902)
M. L. Henley and L. E. Smith Retrace Pulse Generator Having Improved Noise Immunity (3,944,883)
K. G. Hernqvist Metal Vapor Laser Discharge Device (3,942,062)
K. G. Hernqvist Laser Device (3,947,781)
M. V. Hoover Bridge-Balance Detection Circuit (3,944,859)
S. T. Hsu Amplifier Employing Complementary Field-Effect Transistors (3,946,327)
R. W. Jebens and W. H. Morewood Precision Turntable Rotation in a Vacuum Atmosphere (3,943,275)
L. B. Johnston, Jr. Electron Beam Recording in Thick Materials (3,943,302)
J. Kane and H. Schweizer Process for Depositing Transparent, Electrically Conductive Tin Containing Oxide Coatings on a Substrate (3,944,684)
H. Kawamoto and E. L. Allen, Jr. Mechanical Structure for Mounting Microwave Diode Packages (3,943,469)
S. W. Kessler, Jr. Self-Fusing Transcalent Electrical Device (3,946,429)
R. S. Kutay and K. W. Laughman Gas Laser Tube with Stray Light Restriction (3,942,133)
R. D. Larrabee Method of Fabricating a Charged Couple Radiation Sensing Device (3,941,630)
A. J. Leidich Transistor Amplifier (3,946,325)
A. L. Limberg Low Output Impedance Voltage Divider Network (3,942,046)
M. D. Lippman Serial to Parallel Converter for Data Transmission (3,946,379)
J. J. Lyon Altitude Coding for Collision Avoidance System (3,947,845)
L. W. Martinson Data Processor Reorder Random Access Memory (3,943,347)
W. G. McGuffin Symmetrical Odd Modulus Frequency Divider (3,943,379)
D. K. Morgan and R. C. Heuner Keyed Comparator (3,943,380)
C. W. Mueller and E. C. Douglas SOS Bipolar Transistor (3,943,555)
J. A. Olmstead Fluid Flow Measuring System (3,942,378)
R. L. Pryor Zero Crossover Detector (3,944,936)
O. H. Schade, Jr. Astable Multivibrator (3,942,134)
I. Shidlovsky Substituted Zirconium Pyrophosphate Phosphors (3,941,715)
G. E. Skorup LSI Array Using Field Effect Transistors of Different Conductivity Type (3,943,551)
R. G. Stewart Protection Circuit for Insulated-Gate Field-Effect Transistors (3,947,727)
I. F. Thompson Multi-Layer Toroidal Deflection Yoke (3,947,793)
H. Urkowitz, R. P. Perry, and L. Weinberg Digital Interpolator for Reducing Time Quantization Errors (3,943,346)

AUTHORS

Gerald A. Alphonse attended New York University where he obtained a B.S.E.E. in 1958 and an M.S.E.E. in 1959, after doing research in the field of superconductivity. He joined RCA Laboratories in June 1959 and worked toward the development of superconductive memories. In 1963, he attended the Polytechnic Institute of Brooklyn under an award from RCA Laboratories, obtaining his Ph.D. in Electrophysics in 1967. He has done research on photochromic memories, acousto-optics, holographic storage in photorefractive media, and more recently has been working on ultrasonic imaging for medical diagnostics. Dr. Alphonse received an Outstanding Research Award from RCA in 1960 for his work in superconductivity. Besides his work at RCA, he is a physics instructor in the Evening Division of La Salle College in Philadelphia.

He is a senior member of the IEEE, a member of Eta Kappa Nu, Tau Betta Pi, the American Institute of Physics, and the Optical Society of America. He currently serves as a consultant to the National Science Foundation.

R. Casanova Alig received the B.A. degree from Wabash College in 1963, and the M.S. and Ph.D. degrees in physics from Purdue University in 1965 and 1967. He did theoretical studies of acoustic transport phenomena in metals for his dissertation. In 1967 Dr. Alig joined RCA Laboratories as a member of the Physics and Chemistry of Solids Group in the Physical Electronics Laboratory. He has recently done theoretical studies of the optical spectra of impurities and color centers in crystals. During the academic year 1970–71 he was a visiting professor at the Instituto de Fisica e Quimica, Sao Carlos, S.P., Brasil; in addition, since 1967 he has given courses at Mercer County Community College, LaSalle College, and Princeton University. He is a member of Phi Beta Kappa, Sigma Xi, and the American Physical Society.

Michael Ettenberg received his B.S. degree in Metallurgy from Polytechnic Institute of Brooklyn in 1964, and received an M.S. in 1967 and a Ph.D. in 1969 from New York University. His dissertation was concerned with thermodynamic studies of the interaction of point defects in intermetallic compounds. Dr. Ettenberg joined RCA Laboratories in 1969 and has made major contributions in the area of III–V compound synthesis and devices. His work on reliability of laser diodes has included reduction of catastrophic degradation by facet coating, the study of the effect of growth conditions on long-term operating reliability, and the first studies of reliability in InGaAs alloy diodes for 1.06 μm emission at room temperature. This research has been instrumental in the development of very long-lived electroluminescent diodes and cw laser diodes for optical communications. In 1975 he won an RCA Outstanding Achievement Award for the development of liquid-phase epitaxial processes.

He is a member of the American Institute of Metallurgical Engineers, the American Physical Society, and Alpha Sigma Mu Metallurgical Honor Society.

Joseph J. Hanak received his B.S. in Chemistry at Manhattan College in 1953, his M.S. in Physical Chemistry at the University of Detroit in 1956, and his Ph.D. in Physical Chemistry at Iowa State University in 1959. He worked for the Ethyl Corporation in 1955, and was a Research Assistant at the AEC Institute for Atomic Research, Ames, Iowa, from 1955 to 1959. He has been a member of the technical staff at RCA Laboratories since 1959 and was appointed a Fellow of the Laboratories in 1971. Dr. Hanak's field of interest is the synthesis and characterization of electronically active materials. His materials studies have included such areas as rare earth elements, superconductors, ferrites, magnetic recording heads, dielectrics, ultrasonic delay lines, electrophotography, electroluminescence, displays, and metals and dielectrics for video recording. He received three RCA Laboratories Achievement Awards for his work on the vapor deposition process for the superconductor Nb_3Sn, single-crystal ferrite recording heads, and ultra-high-frequency microwave ultrasonic transducers. He also received the David Sarnoff Award and the John Roebling ASM Award for his work on superconductivity. More recently, he has introduced novel methods of co-sputtering, compositional analysis, and materials testing that greatly increase productivity in materials research.

Dr. Hanak is a member of the American Vacuum Society, SID, Phi Lambda Upsilon, and Sigma Xi.

Ralph. W. Klopfenstein received his B.S. in Electrical Engineering from the University of Washington (Seattle) in 1944. He subsequently studied Applied Mathematics at Iowa State University where he received the M.S. and Ph.D. degrees in 1951 and 1954, respectively. From 1943-46 he served in the United States Navy, completed the United States Navy Radio Material Program, and served as Radio Materiel Officer in the Pacific Fleet. He was employed by the RCA Victor Division, Camden, New Jersey, as an Electrical Engineer in 1948 where he worked on the development of high power filters and transmitting antennas for UHF and VHF television applications. Since 1953, Dr. Klopfenstein has been engaged in research at RCA Laboratories, Princeton, N.J., where he is a Fellow of the Technical Staff. From 1953-57 he was active in research in microwave and UHF filters and antennas. In 1957 he was named Head, Mathematical Services, with principal responsibilities in numerical analysis research and the development of computer programs for applications in scientific problems. From 1962-63 he was on leave of absence and served as Director of the Computation Center and Professor of Mathematics at Iowa State University. Since 1963 he has been engaged in research related to the application of mathematics with a principal interest in the development of algorithms for the numerical solution of differential equations.

Dr. Klopfenstein is a member of the American Mathematical Society, Association for Computing Machinery, Institute of Electrical and Electronic Engineers, Mathematical Association of America, Society for Industrial and Applied Mathematics, and the Society of Sigma Xi.

Henry Kressel received the B.A. degree in 1955 from Yeshiva University; the M.S. in 1956 and the M.B.A. in 1959 from Harvard University; and the Ph.D. in 1965 from the University of Pennsylvania. He has been with RCA since 1959.

As a member of the RCA Solid State Division he contributed to the development of high frequency planar silicon transistors and led a group responsible for the development of high power varactor diodes subsequently used for the Lunar Excursion Module Communication System. He joined the RCA Laboratories, Princeton, N.J., in 1966 and became Head of the Semiconductor Optical De-

vices Research Group in 1969. He pioneered in the field of (AlGa)As-GaAs heterojunction devices, in particular laser diodes, and has been actively engaged in the study of devices and luminescent processes in various III-V compound materials. He is the recipient of three RCA Research Achievement Awards as well as a 1974 David Sarnoff Medal for Outstanding Technical Achievement, RCA's highest award. He is presently Head, Semiconductor Device Research, with responsibility in the area of silicon power and III-V compound devices. Dr. Kressel is a Fellow of the IEEE, and a Fellow of the American Physical Society.

William Phillips received an A.B. degree in physics from Columbia University in 1958. He did his graduate work at the Carnegie-Mellon University, receiving an M.S. degree in physics in 1961 and a Ph.D. in electrical engineering in 1964. His thesis research was concerned with the crystal growth and optical properties of potential laser materials. Dr. Phillips joined the RCA Laboratories in 1964, and from 1964 to 1966 was engaged in research on the synthesis and properties of crystalline materials for laser applications. In 1966 he became involved in work on the synthesis and evaluation of improved photochromic materials, particularly rare-earth-doped calcium fluoride, and sodalite, and in 1970 he became involved in research on improved materials for the storage of volume phase holograms, notably iron-doped lithium niobate. He is currently working on diffused lithium niobate-tantalate films for integrated optical waveguide applications.

Stewart Perlow received his B.E.E. degree from the College of the City of New York in 1963, and his M.S.E.E. degree from the Polytechnic Institute of Brooklyn in 1970. From 1963 to 1967, he was employed by RCA's Advanced Communications Laboratory in New York. In 1967 he left RCA to become a co-founder of National Electronics Laboratories. Upon acquisition of National Electronics Laboratories by Harvard Industries, Mr. Perlow became a Project Manager and subsequently Chief Engineer at KMC Semiconductor Corporation, a division of Harvard Industries. His work has been in the field of microwave solid state devices and ultra low noise UHF amplifiers. In January of 1973 Mr. Perlow joined the RCA Laboratories as a member of the Technical Staff. He is presently involved in UHF and microwave work on ceramic substrates.

Mr. Perlow is a member of Eta Kappa Nu and the IEEE.

James P. Wittke received the M.E. degree from Stevens Institute of Technology, Hoboken, New Jersey in 1949, and the M.A. and Ph.D. degrees in physics from Princeton University, Princeton, N. J., in 1952 and 1955, respectively. His doctoral research was a precision microwave determination of the hyperfine splitting in atomic hydrogen. He was an instructor in physics at Princeton for one year before joining the technical staff of RCA laboratories, Princeton. At RCA he has engaged in paramagnetic resonance and microwave maser studies, in research on optically excited crystal lasers, and in investigations of the motions and properties of impurities in rutile. He has also worked on various classified projects. In 1967 Dr. Wittke was awarded an RCA Fellowship for a year of study abroad, which he spent at the Clarendon Laboratory at Oxford, England. Since then, he has studied the luminescence of heavily doped GaAs, helped develop new infrared-stimulated light emitting phosphors, and worked on the generation of ultra-short laser pulses, using mode-locking techniques.

He is currently studying optical fiber communications, including measurements of the characteristics of fast-light-emitting diodes and of various low-loss fibers.

Dr. Wittke is a member of the Optical Society of America, IEEE, Tau Beta Pi, and Sigma Xi and a Fellow of the American Physical Society.