THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

Volume 58 November 1979 Number 9

A Gallium-Arsenide Laser Facsimile Printer

By R. C. MILLER, R. H. WILLENS, H. A. WATSON, L. A. D'ASARO, and M. FELDMAN

(Manuscript received May 23, 1979)

We summarize the development of a high-resolution facsimile printing system in which miniature archival images are machined in metal film by a pulsed, focused, scanning laser beam. The film is designed to machine at optical energies within the pulse-power capabilities of GaAs semiconductor lasers. The images are instantly available, as machined, for rear-projection viewing in a compact desk-top printer. Frames of 8 × 10 mm area containing 3.3 × 10^6 pixels and demonstrating up to seven distinguishable gray levels have been written in 12 seconds. Extensive data on hole machining in partially oxidized, low-melting-temperature, metal films are marshalled to show that single-layer films are inhospitable media for facsimile machining with GaAs lasers. We describe a two-layer film consisting of about 600Å of bismuth followed by approximately 650Å of selenium evaporated onto a low-surface-energy substrate. This selenium thickness is antireflective at the laser wavelength, and the composition is nearly stoichiometric for Bi_2Se_3. A strongly exothermic reaction occurs when molten bismuth and selenium mix in the irradiated area. We present data illustrating the machining performance of this film with a thick-cavity, GaAlAs, double-heterostructure laser providing 300 mW of peak power at 3-percent duty factor and 0.1-μs pulse duration and give examples of the image quality obtained in raster-scanned, hole-array machining. Studies of film adhesion, scratch resistance, and shelf life are summarized. Several diffraction-limited optical systems for acquiring, deflecting, and focusing the GaAs beam onto the film, cylindrically corrected for astigmatism and *ellipticity, are described, and several laser properties which affect image machining adversely are chardcterized.*

I. INTRODUCTION

This paper describes the printing of miniature, high-resolution, pictorial images[1] by machining with a gallium-arsenide laser. The focused beam from a pulse-modulated laser is raster-scanned over a thin metal film deposited on a transparent plastic substrate. The energy in each burst of laser power is controlled in correspondence with a transmitted signal so as to machine an array of holes in the metal film.[2] This signal may correspond to the optical reflectivity of a remotely located, synchronously scanned, original document, or it may represent the machining instructions of a graphics-generating computer.

A capability for high resolution is intrinsic to this printing method. The machined image can readily contain 1600 hole sites per line and 2000 lines per frame, and is similar in appearance to a frame of microfilm or microfiche. The pulse repetition rate, raster scanning speeds, and maximum pulse energy are chosen such that the holes machined in "all-white" regions of the image are nearly contiguous. These regions are about 50 percent transparent to visible light. Unmachined regions are <1 percent transparent to visible light. With suitable projection optics, the received portion of a frame can be viewed while the remainder is still being transmitted. Such a projection receiver need be little larger than a conventional microfilm viewer of comparable screen dimensions. The machined image is available without further processing as an archival record.

Hole machining begins at a threshold laser intensity whose value depends on various optical, thermodynamic, mechanical, and—in the case of certain metal composites—chemical parameters of the film. It also depends on the heat loss mechanism, which is usually dominated[3] by thermal conduction to the plastic substrate. Such conduction losses are greatly decreased if energy is supplied to the film in short-duration, high-peak-power bursts,[2] rather than in a low-average-power, continuous-duty stream.[4] Thus, with careful attention given to the output beam optical quality, cavity-dumped He-Ne, ArII, and Nd:YAG lasers, supplying 1 to 2 W power in 30-ns-duration bursts, have overcome the machining threshold of thin bismuth films sufficiently to produce satisfactory images.

Present GaAs lasers[5,6] cannot machine useful images in pure bismuth films. Pulse energies comparable to those from cavity-dumped lasers can be attained from GaAs only by roughly trebling the pulse duration with consequently increased heat conduction loss in the film and decreased machining efficiency. Further, a significant portion of the

GaAs power is dispersed into secondary nonmachining modes, making its beam inferior for machining to the nearly Gaussian beam profiles produced by gas and YAG lasers. Thus, the development of a film with machining threshold several times less than that of bismuth is the key to realizing GaAs as a practical laser machining source.

The advantages of GaAs are large. The low cost and small size of GaAs lasers, together with the inherent mechanical stability of their cavities and the simplicity with which they can be modulated by low voltage pulses, suggest the possibility of a moderately priced, compact, high-speed, facsimile receiver with superior resolution, a useful gray scale range, and instantly accessible images. The miniature size of these images and the capability of laser-machining additional information onto unwritten areas of the film can be advantageous in many information storage applications.

II. GENERAL DESCRIPTION OF FRAME MACHINING

The film transparent substrate is a 100-μm-thick, polyester base overcoated on one side with a low-surface-energy acrylic polymer. A 600 Å-thick layer of bismuth is evaporated onto the polymer, and about 650Å of selenium is evaporated onto the bismuth. The chosen selenium thickness is antireflective for air-incident radiation at 0.885-μm wavelength, permitting about 85 percent of the incident laser power to be absorbed in the bismuth layer. This type of composite film presents interesting points of comparison with the amorphous arsenic-telluride film recently proposed[7] for obtaining cleanly machined holes in optically recorded video disks.

The energy delivered by the pulsed beam melts a disk of Bi-Se whose radius depends on the intensity profile of the focused beam, on the pulse duration, and on the heat loss rate to the substrate. Surface tension pulls this liquid towards the solid periphery where it piles up and freezes into a narrow rim. An exothermic reaction[8] occurs when the molten bismuth and selenium mix; the released energy, in combination with the increased absorption due to the selenium antireflective property, more than offsets the thermal mass added by the selenium. Assisted by the low surface energy at the polymer interface, the Se/Bi bilayer can achieve machining threshold at laser beam intensities four to six times lower than those required for bismuth alone.

Metal "tektites" are strewn beyond the hole boundary at all machining energies, and some metal is ejected from the film at high machining energies. However, the primary function of the absorbed optical energy is to unlock surface forces which displace metal into the peripheral rim surrounding the hole. Image white areas are local regions of nearly contiguous holes, each bordered by piled-up metal; such regions are about 50 percent transparent to visible light. Image black areas contain no machined holes, and their transparency is less than 1 percent. Gray-scale effects can be achieved by varying the laser pulse duration or amplitude in correspondence with a received analog signal, while maintaining constant center-to-center hole spacing. Figures 1a and 1b are photographs of an individual machined hole and an assemblage of variable size holes defining an image of a human eye. (The pulse duty factor is typically a few percent, so that the beam is essentially stationary during the machining of each hole.) When the

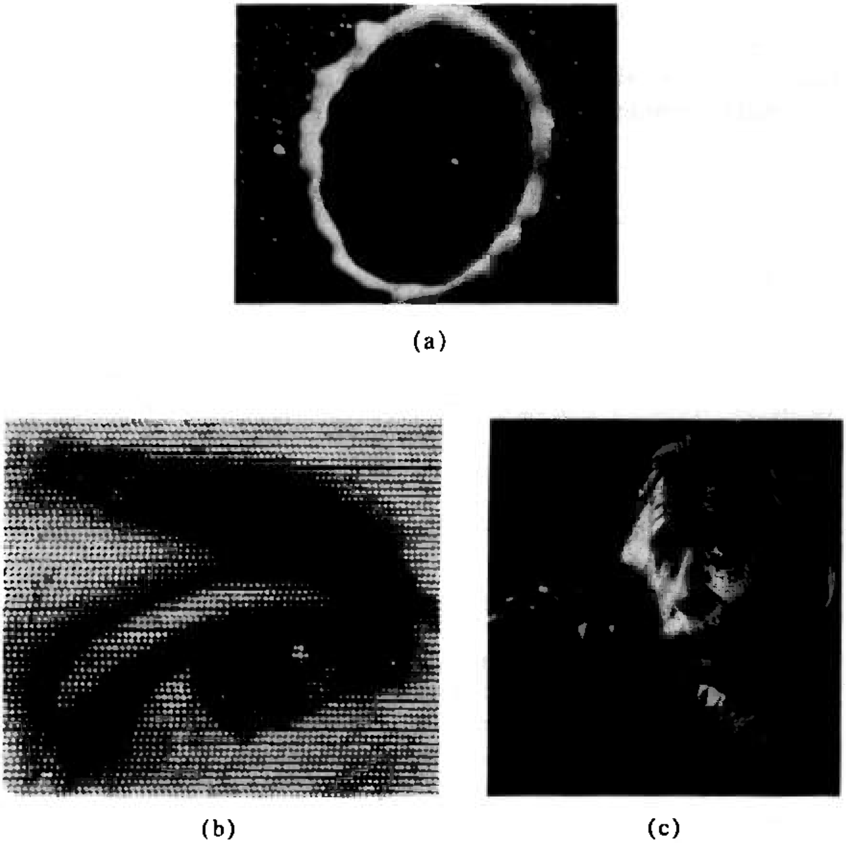


Fig. 1—Formation of microimages by hole machining with a cavity-dumped YAG laser. (a) An isolated 5-μm-diameter machined hole in Se/Bi film deposited on a plastic substrate. Most of the displaced metal is found in a thickened rim around the hole periphery. A few metal droplets remain in the hole, and some appear beyond it. (b) Microimage of a girl's eye. The center-to-center hole spacing is maintained constant. The hole sizes are varied to attain a gray-scale effect. This figure is an enlargement of a portion of an IEEE Test Chart. (c) The capabilities of the cavity-dumped YAG-laser machining printer for rendering detail and gray-scale tonality are illustrated by this microimage of Albert Einstein. The light vertical scratches were caused by a film manufacturing error.

image is projected onto a screen, the eye does not resolve individual holes at normal (25 cm) reading distance, so local regions machined at constant pulse energy appear to be uniformly bright, as indicated in Fig. 1c. (The images of Fig. 1 were made on Se/Bi film with a cavity-dumped YAG laser.)

The rectangular frame area, usually of 17/22 aspect ratio, has been defined by a mechanical raster scan in all machining printers built to date. The line scan (fast scan) is obtained by collimating the laser beam and deflecting it with a galvanometer mirror driven in sawtooth, oscillatory motion about the galvanometer axis. The mirror is centered in the entrance pupil of a writing lens which converts the beam angular deflections into displacements of the focused beam waist over a line in the flat focal plane. Hole machining occurs only during the line scan forward ramp. The galvanometer loaded resonant frequency, usually dominated by the deflecting mirror moment of inertia, dictates a minimum value for the sawtooth flyback time: it becomes difficult or impossible to provide acceptable ramp linearity if the scan rate exceeds about 40 percent of the loaded resonant frequency.[9] Frame widths up to 9 mm have been machined. (An exception to this flat-field method is described in Section 6.1.1.) The frame scan (slow scan) has been obtained either by translating the film past the fast-scanned line focus or by translating this line focus down the length of the stationary frame. Hesitation or jerkiness in the frame scan affects entire lines or groups of lines and produces prominent defects in the printed image. The smoothest scans are obtained by translating strongly tensioned film with a heavily damped capstan.

Because the film has a finite machining threshold and because lateral thermal conduction is negligible[3] in thin films, long, low-amplitude tails on the beam do not contribute to machining—i.e., the optical beams which machine adjacent holes may overlap greatly, but the machined holes need not. Discrete, contiguous machining sites are thereby defined, leading naturally to a capability for high resolution. Resolutions corresponding to 1600 holes per line are easily obtained and are typical of our printers. We find, however, that the optical numerical apertures needed to focus the GaAs machining beam exceed those which would be required if the beam obeyed the propagation law for Gaussian beams.[10] These large aperture requirements present an obstacle to the achievement of high printing speeds. Line scan rates of 166 Hz and pulse repetition rates of 330 kHz, corresponding to 12-s printing time for a 2000-line frame, have been obtained from GaAs laser printers without much difficulty, but appreciably higher scan rates require certain optical modifications, described in a later section. The fastest line scan rate attempted during our printer development was 500 Hz, corresponding to 4-s printing time for a 2000-line frame.

The scan linearity was barely acceptable even when flyback occupied approximately 30 percent of the line period, and successful hole machining required the use of a cavity-dumped gas laser at a pulse repetition rate of 1.0 MHz, producing only 1400 holes during the forward ramp. Attempts to print with GaAs lasers at 1.0-MHz pulse rate have so far been unrewarding; modifications in the laser structure, combined with thermoelectric cooling of the laser mounting stud, may be needed.

The beam emerging from the present GaAs laser is elliptical in cross section. It is also astigmatic: that is, when viewed in a plane normal to the p-n junction, the beam appears to diverge from the output mirror surface, but when viewed in the junction plane it appears to diverge from a point at least 25 μm deep within the laser.[11] All beam-acquisition lens systems, whether fiber-optical[12] or conventional,[13] must include one or more cylindrical elements which accurately remove this astigmatism. Unless this is done, machining, if it occurs at all, will take place with greatly reduced efficiency and very small depth of focus. By proper choice and placement,[12,13] these cylindrical elements can also correct beam ellipticity. This latter correction need not be particularly accurate since nearly round holes are machined at beam ellipticities as large as 2:1, provided the beam energy is not too far above machining threshold.

Astigmatism and ellipticity are easily remedied, but unfortunately the laser cavity design which produces these effects exhibits a more serious and intransigent beam defect. A low-amplitude, rather complicated, mode structure appears parallel to the junction plane at currents slightly above lasing threshold. The amplitudes of these modes usually increase, relative to the fundamental mode, as the current increases. At high currents as much as ⅓ to ½ of the total optical power is diverted into this secondary structure. This power contributes very little to raising the film temperature within the periphery of the main (machining) mode, and its presence represents a serious loss of machining efficiency. The problem is tolerable only because the Se/Bi film machining threshold is exceptionally low. The GaAs beam also evinces several lesser defects which, luckily, can be disposed of without seriously limiting the printer performance. The longevity of present lasers is probably adequate for commercial printing usage provided the average pulse repetition rate is restricted to keep the laser junction near room temperature.

Readers for whom the above general description of frame machining has provided sufficient detail may wish to turn to the final sections of this paper for examples of the image quality which has been attained in GaAs laser printers. The intervening sections describe the methods, problems, and design compromises used in producing those images.

III. THE GALLIUM-ARSENIDE MACHINING LASER

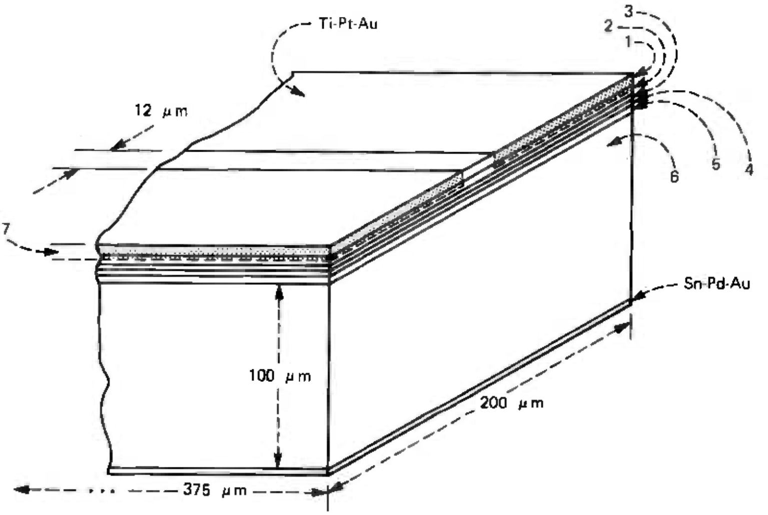
3.1 Machining requirements; epitaxial structure

The GaAlAs double heterostructure laser described next was developed around the needs of machined-image printing. The requirements on the laser evolved during this development, mainly in response to advances in the machining film technology. The present nominal requirement is a single-mirror output of 0.3-W peak pulse power in the fundamental cavity modes parallel and perpendicular to the junction plane, with 3-percent duty-factor capability at room temperature and median life corresponding to the machining of 10^4 to 10^5 frames. Junction-plane filamentation (strongly erratic beam-intensity variations along the junction plane) cannot be tolerated.

Figure 2 depicts the typical epitaxial structure[14] used in machining lasers and also illustrates the confinement of current to a 12-μm-wide stripe in the outermost epitaxial layers. The stripe is defined by masking the 12-μm width from 300-KeV proton bombardment which reduces electrical conductivity in the adjacent, penetrated regions.[15] Proper confinement decreases the total current which must be supplied at the laser threshold current density, and it also leads to gain profile variations which provide mode guidance parallel to the junction plane. The epitaxial layers are grown on a 100-μm-thick, n-type, GaAs substrate. The topmost layer is usually a GaAs, p-type, capping layer included primarily to assist contact formation and to reduce the propagation of contact strains into the recombination volume. This cap is sequentially metalized with layers of titanium, platinium, and gold, and the substrate with layers of tin, palladium, and gold. Individual laser chips are obtained from a wafer by cleaving into 375-μm lengths and sawing into 200-μm widths the cleaved surfaces defining the laser mirrors. The epitaxial side is indium-bonded to a copper stud, and the substrate is contacted by a thermal-compression-bonded gold wire.

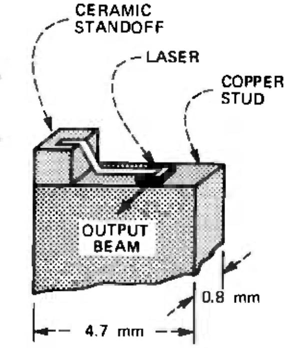
3.2 Mode confinement perpendicular to junction plane

Double heterostructures fabricated by liquid phase epitaxy are among the most efficient room-temperature junction lasers. The energy bandgap and optical refractive index of $Ga_{1-x}Al_xAs$ depend on the aluminum fraction x. Epitaxial layer sequences with different x-values can be arranged to create optical potential wells which define radiation field confinement perpendicular to the junction plane. Simultaneously, this sequencing creates diffusion barriers at some layer interfaces which confine minority carriers to the optical well where most of them recombine radiatively. The optical cavity of a machining laser must be comparatively thick lest the output mirror peak loading approach the catastrophic damage limit,[16] but thick cavities tend to



(a)

LAYER	DESCRIPTION	PRINCIPAL DOPING ELEMENT	THICKNESS
1	p-CAP, x=0	Ge	2 μm
2	P-TERNARY, x=0.24	Ge	2 μm
3	p-ACTIVE LAYER, x=0	Ge	1.3-1.6 μm
4	n-LOSS LAYER, x≃0.02	Te	0.8 μm
5	N-TERNARY, x=0.24	Te	2-3 μm
6	n-SUBSTRATE, x=0	Si	100 μm
7	REGION PENETRATED BY 300 keV PROTONS	—	~2.3 μm



(b)

Fig. 2—Micrographics-laser structure and mounting. (a) Structure. The label x refers to the aluminum fraction in $Ga_{1-x}Al_xAs$. Five epitaxial layers are indicated. The topmost layer (p-cap) is included to facilitate bonding; however, it also increases the thermal resistance and has been omitted from some micrographics wafers. Proton bombardment penetrates only into the top of the $P_{0.24}$ region, as indicated by the dashed line, defining a 12-μm-wide x 375-μm-long conductive cross section. (b) Mounting. The laser is indium-bonded to a copper stud with the output mirror overhanging the edge of the stud by ~ 10 μm. The laser n-contact is wire-bonded to a gold-plated ceramic standoff. The epitaxial layers are overplated with a 10-μm-thick gold heat sink before indium bonding. The illustrated arrangement produces a steady-state temperature rise of 34°C per watt of electrical input power.

favor high-order modes over the fundamental mode. A distinctive feature of the present laser design is the selective introduction of optically lossy dopants[17] into the epitaxial growth sequence to suppress high-order modes.

Figure 3 illustrates the mode and carrier-confining mechanisms for typical refractive index and band-gap energy variations. Band-edge positions for the forward-biased junction are shown in Fig. 3b. Electrons are injected across the p-$n_{0.02}$ junction* by application of forward bias, and their diffusion in the p-region is limited by the large, nearly lossless, potential barrier at the p-$P_{0.24}$ boundary. Thus, carrier radiative recombination—the lasing power source—occurs almost exclusively within the p-region. The optical fields are confined within the $n_{0.02}$-layer and the p-layer by the large refractive-index discontinuities at the $N_{0.24}$-$n_{0.02}$ and $P_{0.24}$-p boundaries shown in Fig. 3c.

The technique[17] for maintaining fundamental mode operation with this thick optical cavity is indicated by Fig. 4, in which 0- and 1-order mode profiles have been superimposed on a cavity refractive-index plot. The small index discontinuity at the p-$n_{0.02}$ interface shifts the 0-order mode center preferentially toward the p-region where most of the gain-producing radiative recombination occurs. The same shift occurs for the 1-order mode, but since a larger volume is needed to accommodate this mode, most of one of its lobes remains in the tellurium-doped $n_{0.02}$-layer. Calculations[18] indicate that the high optical losses created by tellurium doping reduce the 1-order mode net gain below that of the 0-order mode. Hence this design should produce a large 0-order cavity volume compatibly with low threshold current. Both factors are important in achieving long laser life at high pulse powers and at room temperature.

3.3 Junction plane mode guidance

The previous section dealt with the use of real-refractive-index discontinuities to achieve mode confinement perpendicular to the junction plane. This method can provide fundamental mode operation so long as the confining layer thickness is less than about 1.6 μm. The high peak powers which must be produced by a machining laser require cavity cross sections of 15 to 20 μm^2 to keep the mirror loading[16] and lateral current crowding[19,20] within conservative limits. Therefore, if the cavity can be only 1.6 μm thick, it must be made at least 10 μm wide. Fundamental mode operation in this wide a cavity is not easily achieved with structurally defined, real-index discontinuities, although recent advances in real-index mode guidance have been encouraging (see Section 4.2.4). Lateral mode confinement in the present laser has been based on gradual variations in the imaginary part of the dielectric constant. A guidance mechanism of this type is illustrated in Fig. 5.

In Fig. 5a, the junction plane is the $x - z$ plane. A thick, high-

* Ternary layers are distinguished with capital letters, and an optional subscript denotes the aluminum fraction, x.

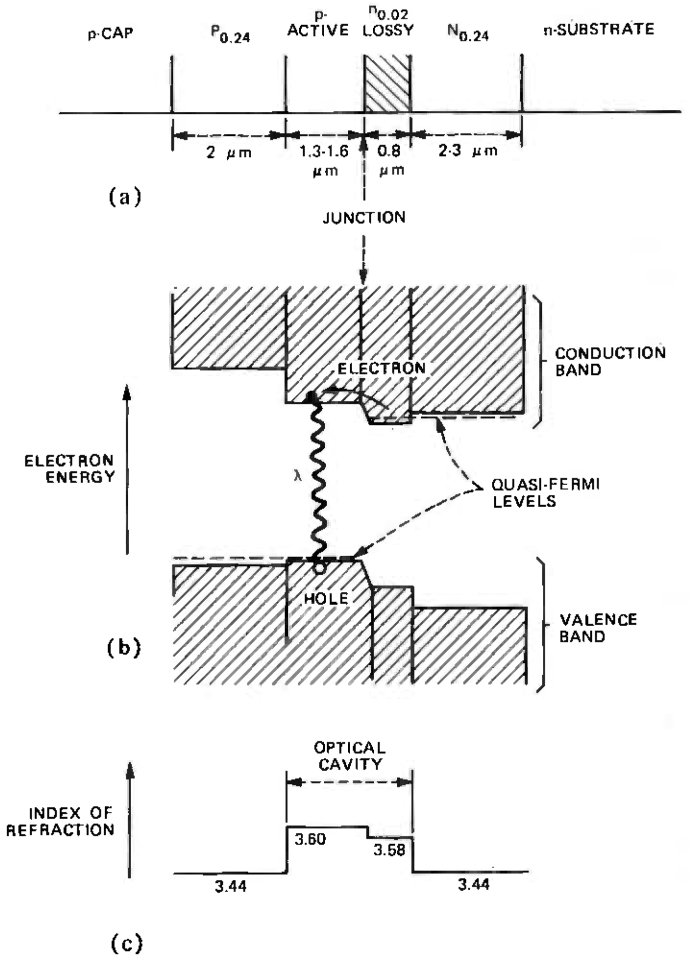


Fig. 3—Optical- and carrier-confinement mechanisms. (a) Epitaxial structure. Forward bias causes electrons to be injected across the $n_{0.02}$-p junction. (b) Carrier well. Electrons injected from the $n_{0.02}$-region into the p-region are transported toward the $P_{0.24}$-region by diffusion. At the p-$P_{0.24}$ boundary, they encounter a potential barrier whose height is large compared to the electron thermal energy, limiting radiative recombination between minority carriers (electrons) and majority carriers (holes) almost entirely to the p-region. (c) Optical well. Optical fields tend to remain confined in the regions of highest (real) refractive index (i.e., the p-active region) because of critical-angle reflections at boundaries with lower index layers. Thus, the optical and carrier wells overlap substantially.

conductivity, n-GaAs region is assumed to occupy the space $y \leq 0$. The lower conductivity, p-GaAs, active region extends from $y = 0$ to $y = d$, and a high conductivity P-$Ga_{1-x}Al_xAs$ region extends from $y = d$ to the external contact at $y = a$. Proton bombardment, penetrating partway into the P-layer, defines a contact width $2w = 12$ μm through which current flows. We assume that the proton-penetrated volume has zero

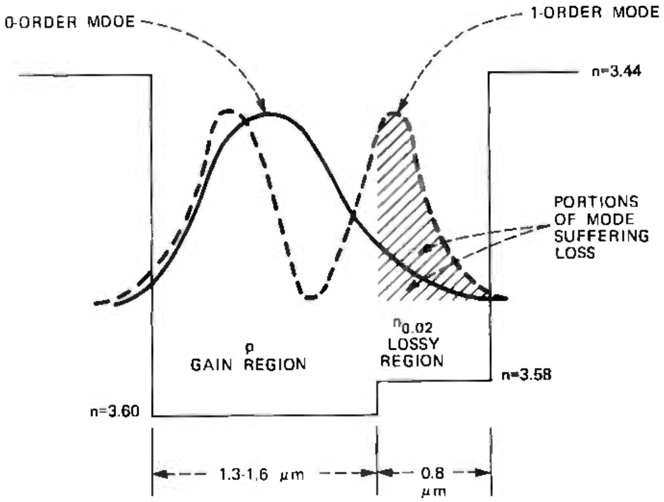


Fig. 4—Mode control perpendicular to junction plane by loss selectivity of optical modes. The $n_{0.02}$-layer is made optically lossy by heavy tellurium doping. A larger fraction of the undesired 1-order mode exists within this lossy region than is the case for the desired 0-order mode. By properly controlling the loss and the p- and $n_{0.02}$-layer thicknesses, the pulse-current threshold for 1-order lasing can be kept above the currents required for micrographics use.

conductivity. When forward bias is applied, electrons are injected across the junction into the p-layer where they move by ambipolar diffusion. Because of the 168-mV-high diffusion barrier at the p-P interface, the electrons are effectively trapped in the p-layer, and nearly all of them recombine there radiatively. Their diffusion is driven by the profile $J_y(x)$ of the junction current density. This current density is concentrated in an area beneath and slightly wider than the stripe, rather than being uniformly distributed over the junction, as would occur if the p-region conductivity were infinite. That is, the finite conductivity and lateral current density $J_x(x)$ within the p-region cause the junction forward bias to depend on x. The bias is largest at $x = 0$ and decreases with increasing $|x|$, effectively shutting off the junction at large $|x|$.

This combination of diffusion and junction-bias variation produces an electron density profile, and hence a laser gain profile, which is peaked at $x = 0$, as shown in Fig. 5b. The ability of gain peaking to guide modes is readily shown[21–23] by treating gain transparency by inclusion of a positive imaginary term in the complex dielectric constant. The modes predicted when the gain profile is approximated by a parabola are constant-curvature cylindrical wavefronts with Hermite-Gaussian intensity distributions centered about $x = 0$. The refraction of these wavefronts at the output mirror, illustrated in Fig. 5c, is responsible for the astigmatism in the beams emitted by stripe-

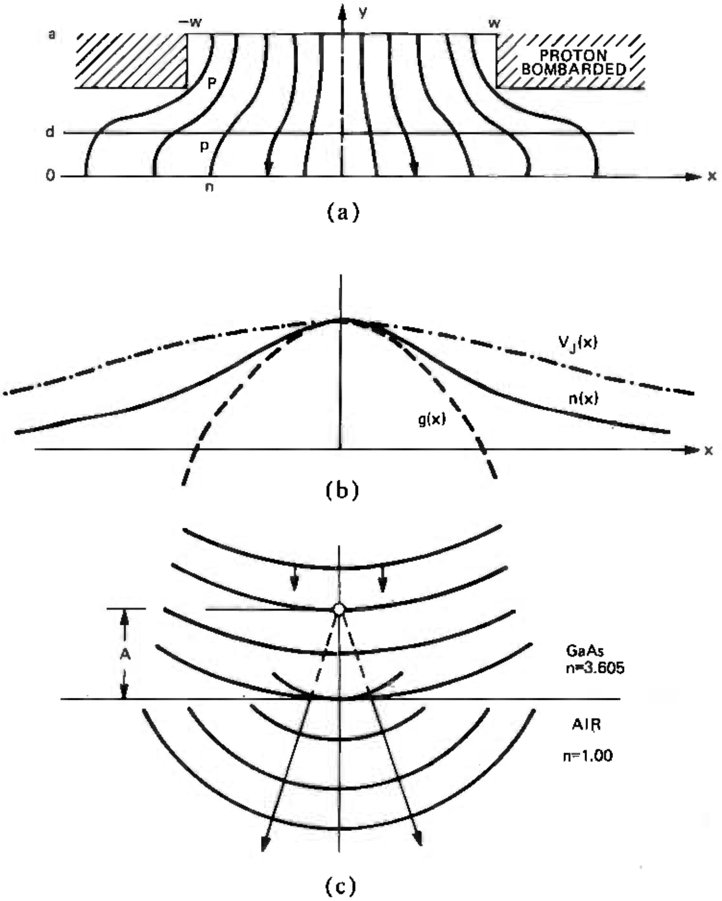


Fig. 5—Mode control parallel to junction plane. (a) Expected current-flow pattern in micrographics laser. (b) Proton bombardment and layer resistivity inhibit the lateral (i.e., x-directed) flow of current and cause the junction forward bias $V_J(x)$ to decrease with increasing $|x|$, confining junction current to the region $|x| \approx 1.5\text{w} - 2\text{w}$. The result is a bell-shaped minority-carrier profile $n(x)$ in the active region, which produces an approximately parabolic variation in optical gain $g(x)$. Ideally, this gain profile is centered symmetrically underneath the stripe. (c) The gain profile supports Hermite-Gaussian modes with constant-curvature wavefronts. This curvature produces astigmatism, since the cylindrical wavefront which is refracted into air at the GaAs-air interface appears to originate from a line focus located a distance A inside the laser.

geometry junction lasers. However, the electron density also changes the real part of the dielectric constant via several effects.[11] The sum of these effects is perhaps slightly anti-guiding,[11,24] but is too small to overcome[25] gain guidance. Some of these real-part effects could be nonsymmetric in x and could displace the mode center from $x = 0$. The retention of cubic (and higher odd-order) terms in approximating the gain profile also predicts a mode whose center is displaced from

$x = 0$. Whatever its physical origin, mode displacement does occur,[26-28] and it leads to degradations in the beam intensity profile which have serious consequences for machining.

The junction-plane behavior is sometimes unstable and erratic. This tendency appears to be correlated with high aluminum content in the ternary P-layer. Wafers with $x = 0.36$ have yielded very few stable machining lasers, whereas $x = 0.24$ wafers have produced many usable lasers. The difference may be related to the ternary electrical conductivity, which is about 10 times higher for $x = 0.24$ than for $x = 0.36$.

IV. DESCRIPTION OF LASER NEAR AND FAR FIELDS

4.1 The major beam defects

The electric-field-amplitude profile of the cylindrical beam waist at the output mirror determines[29] the divergence of rays perpendicular to the junction plane whereas, in the plane of the junction, the beam appears to diverge from a cylindrical waist located a short distance A behind the output mirror because of refraction of the constant-curvature wavefronts.[11] A simplified statement[13] of the printer optical function is that a cylindrical optical element(s) superimposes these two waists onto one another, and spherical elements image this superposition onto the metal film. A characterization of these waists prior to their superposition provides valuable information on the laser beam properties that create difficulties in machining images.

The most important beam defects are:

(*i*) Beam steering in the junction plane. The beam, viewed in the junction plane, emerges at an angle to the mirror normal, and this angle is current dependent.

(*ii*) A sidewise displacement of the junction-plane virtual waist toward one edge of the 12-μm-wide conductive stripe, accompanied by the appearance of side-lobe mode structure. At high currents, the side lobes contain an appreciable fraction of the total power, but contribute very little to hole machining.

(*iii*) Destabilization of the mode by optical feedback. Reflections from beam waists formed by optics external to the laser, e.g., from the metal film, are focused back into the laser cavity. This feedback is superimposed onto the internal feedback from the output mirror, and its phase fluctuates because of unavoidable microphonic, thermal, and convection effects, producing erratic power variations.

(*iv*) Current dependence of the astigmatism distance, A. The value of A increases with increasing current (i.e., the wavefront curvature decreases), and the various junction-plane mode components have differing A values.

The deleterious effects of items (*i*), (*iii*), and (*iv*) on this list can be rather easily remedied or reduced as described in the remainder of

Section IV. However, the avoidance of power loss into junction plane side lobes [item (*ii*)] is somewhat less certain, at least within the context of a fundamental-mode, high-duty-factor, long-lived laser.

4.2 Experimental evidence of beam defects; possible remedies

4.2.1 Beam steering

Beam steering is measured by mechanically scanning the laser emission profile parallel to the junction plane with a slit/photomultiplier arrangement. The plane of the slit can be spaced 2 to 3 cm from the laser mirror, and no optics are needed. The direction of the mirror normal can be determined accurately by reflecting a He-Ne laser beam from the cleaved facet. A family of such profiles is shown in Fig. 6 for an ascending sequence of currents in a selected laser. The structure which appears in the wings of these profiles at high currents indicates that junction-plane side lobes can propagate into the far field at angles different from the principal-ray angle of the main lobe. Below lasing threshold, the main-lobe profile peak is accurately located along the mirror normal but, as the current increases above threshold, the peak shifts monotonically away from the normal, the shift exceeding 8 degrees at high currents. The angular shifts θ_s and the profile angular full-widths at half-amplitude $(2\Delta\theta)_{1/2}$ are plotted in Fig. 7 as functions of current.

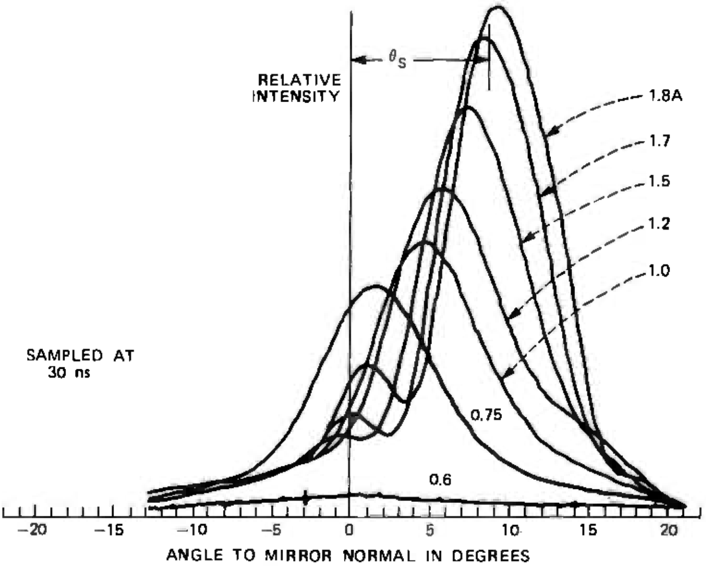


Fig. 6—Intensity angular distribution parallel to junction plane. The angular position of the peak of the far-field intensity distribution parallel to the junction plane defines the beam-steering angle, θ_s. Side-lobe structure in the intensity profiles becomes increasingly pronounced as the current is raised. The data were measured 30 ns after the beginning of a 100-ns-duration pulse.

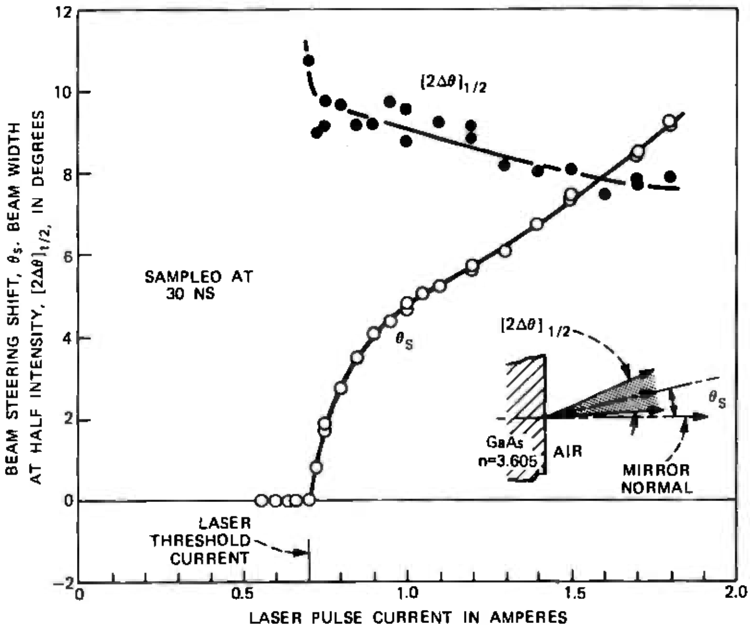


Fig. 7—Dependence on current of beam-steering angle and profile width parallel to junction plane. The far-field intensity full-width at half-maximum amplitude, $[2\Delta\theta]_{1/2}$, depends very little on current, whereas the beam-steering angle, θ_s, increases rapidly with current and actually exceeds $[2\Delta\theta]_{1/2}$ at high currents.

The immediate effect of beam steering is to increase the printer optical apertures needed to accommodate the junction-plane rays at all currents. This works no particular hardship until one reaches the galvanometer mirror. There the consequence of beam steering can be a sizable displacement of the collimated beam parallel to the laser junction plane. This displacement, for a collimated beam full-width at half-amplitude of $(2R)_{1/2}$, equals $[\theta_s/(2\Delta\theta)_{1/2}]\ (2R)_{1/2}$. As seen from Figs. 6 and 7, this quantity can actually exceed the collimated beam diameter. The net effect, if a large range of laser currents must be accommodated, is considerably to increase the galvanometer-mirror moment of inertia and the writing lens entrance-pupil diameter. The moment-of-inertia increase is smallest if the laser junction plane is oriented parallel to the galvanometer axis.

The increases in aperture size demanded by beam steering disappear if only one pulse-current level is used to print images. Time-sampled studies of junction-plane behavior show that the beam-steering angles in Figs. 6 and 7 are attained within 30 ns of the current pulse turn-on instant. Therefore, if the pulse machining energy is controlled by pulse-duration modulation (30 ns $\leqslant \Delta t \leqslant$ 100 ns) at constant pulse-current amplitude, rather than by current modulation at constant pulse duration ($\Delta t = 100$ ns), then beam-steering displacements will be constant and can easily be eliminated by mechanically biasing the laser.

Beam steering is small or absent from profiles scanned perpendicular to the junction plane. The normalized far-field profiles shown in Fig. 8 indicate that a slight (≤2 degrees) shift may have occurred when the current in this laser was raised above threshold. (The illustrated data are chart recorder tracings, uncorrected for aperture inclinations.) The smooth, monotonically decreasing sides and tails of these profiles invite comparison with the junction-plane data of Fig. 6.

4.2.2 Mode displacement

Displacement of the junction-plane principal lobe towards the stripe edge is illustrated in Fig. 9. These profiles were obtained by scanning a ×43 magnified image of the laser output mirror, formed by a 0.68-NA microscope objective, parallel to the junction plane. The shift begins abruptly near laser threshold and proceeds rapidly with increasing current towards an asymptotic value equal to about 0.8 of the stripe half-width. In Fig. 9, one or more unresolved peaks is visible on the right of the higher current profiles, filling in the space left vacant by the shifted main-lobe peak. Mode displacement in itself does not cause important printing defects. But as a manifestation of secondary junction-plane mode structure, it is symptomatic of the rather serious defect which is elaborated on in Section 4.2.4. It has been established, by sampling the profiles at various instants within the 100-ns-long

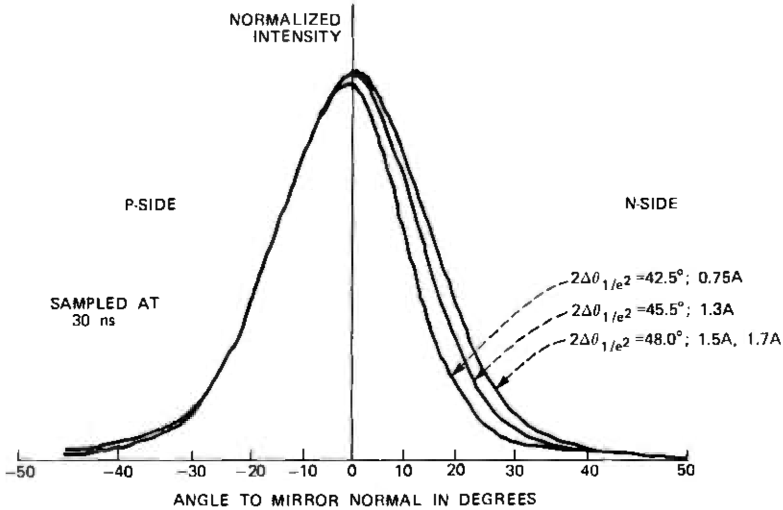


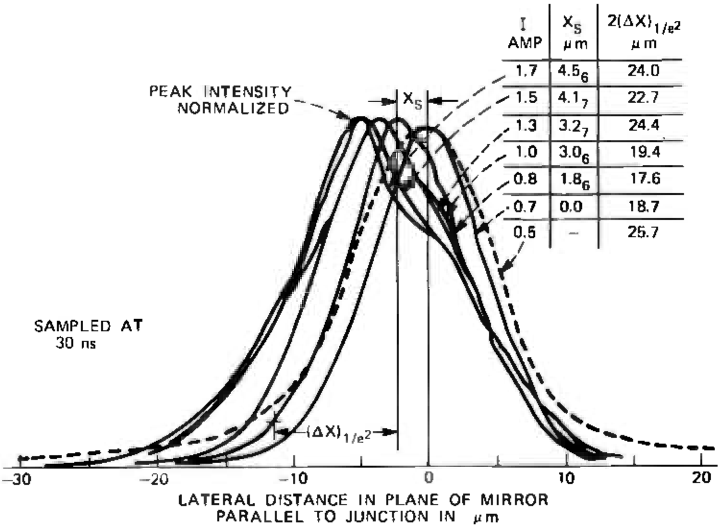
Fig. 8—Intensity angular distribution perpendicular to junction plane. The far-field intensity distribution normal to the junction plane shows evidence only of the 0-order mode. Its width increases slightly with increasing pulse current.

pulse, that mode displacement attains its final value somewhere within 10 to 30 ns of current-pulse turn-on.[28]

The width of the main peak can be roughly estimated by measuring the width of the left-hand half of the Fig. 9 profiles and assuming symmetry. This full width is current-independent at high currents, and its $1/e^2$-intensity value equals twice the stripe width. The same width is found for the spontaneous emission profiles well below threshold. The profiles are significantly narrower than this value only in the neighborhood of threshold.

4.2.3 Mode destabilization

The brass scanning slit used for the data of Fig. 9 was lined on both sides with diffusely reflecting black tape. When the tape was removed, erratic scans such as those illustrated in Fig. 10 were obtained. The shape of these profiles depended on the instant within the 100-ns pulse at which they were sampled, and their reproducibility was poor. If the slit is replaced by a piece of Se/Bi film and an attempt made to machine images with this beam, uncontrollable gradations in the machined hole sizes will appear on many portions of the film, and in some areas as many as half of the intended holes may be missing. Examples of such erratic machining behavior are presented in Section 8.3.



I AMP	X_S μm	$2(\Delta X)_{1/e^2}$ μm
1.7	4.5_6	24.0
1.5	4.1_7	22.7
1.3	3.2_7	24.4
1.0	3.0_6	19.4
0.8	1.8_6	17.6
0.7	0.0	18.7
0.5	–	25.7

Fig. 9—Output-mirror intensity distribution parallel to junction plane. These mirror scans parallel to the junction plane show that the intensity peak shifts to the left by an amount x_s with increasing current; simultaneously, a shoulder appears on the right, indicative of multimode output. The values of $2(\Delta x)_{1/e^2}$ are estimates of the main lobe width.

The type of destabilization shown in Fig. 10 appears only on profiles taken at currents above laser threshold and is correlated with the laser output mirror reflectivity. The erratic machining which results is most serious with mirrors that have been antireflection-coated to forward-angle reflectivities smaller than 1 percent. The observations suggest that the laser cavity is being upset by time-dependent optical feedback. Light reflected from the neighborhood of a beam waist is refocused back onto the laser by the printer optics. This light re-enters the cavity and combines with the internal feedback from the mirror. The internal feedback is small if the mirror reflectivity is small, and the external feedback, whose phase and amplitude can vary slightly over the raster-scanned film, may be comparable to it in magnitude. Hence the observed destabilization consists of power overshoots and undershoots caused by pulse-to-pulse variations in the effective net feedback.

The likelihood of destabilization can be reduced by controlling the output-mirror reflectivity to guarantee the dominance of internal feedback in normal printing situations. The mirror coating designs used to accomplish this are summarized in Section 4.3.

4.2.4 Astigmatism and near-field side lobes

A virtual waist is produced inside the laser when the constant-curvature, junction-plane wavefronts are refracted[11] at the output mirror. Suppose that a 4-mm-focal-length microscope objective is positioned to form an image of the mirror at an appropriate conjugate distance, say, 180 mm. Then, as shown by the composite in Fig. 11 of

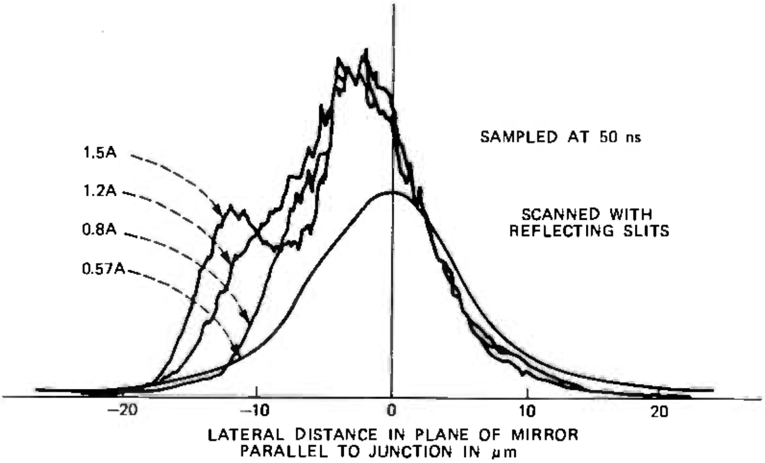


Fig. 10—Output-mirror scans parallel to junction plane, illustrating mode destabilization. Light reflected from the scanning slits and focussed back onto the laser output mirror destabilizes the laser cavity at currents above lasing threshold. There is no destabilizing effect below threshold.

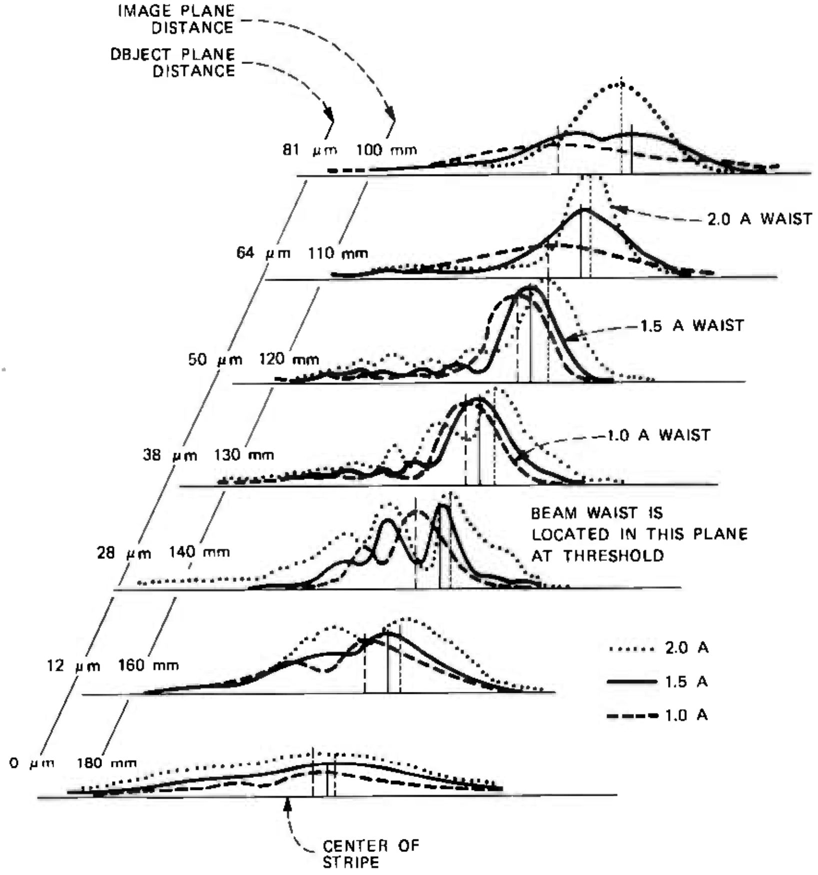


Fig. 11—Junction-plane profiles produced at various conjugate positions by beam-acquisition lens. A 4-mm-focal-length, 0.68-NA, microscope objective was used to focus an image of the laser output mirror at 180-mm distance from the mirror, and the intensity distributions parallel to the junction plane were measured at various intermediate planes between the lens and the 180-mm image position for several currents. The air-equivalent object distances within the laser, conjugate to these intermediate planes, are indicated. The plane at which the beam profile is narrowest, i.e., the astigmatism distance, depends on beam current. The astigmatism distance within the laser increases with current, corresponding to decreasing wavefront curvature of the junction-plane mode. Complicated multimode structure is evident throughout the focal region at machining current levels.

profile scans taken parallel to the junction plane, a real image of the junction-plane virtual waist will occur at some lesser, current-dependent, conjugate distance. At and below threshold this distance is, typically, about 140 mm; it decreases steadily as the current increases, approaching 110 mm at high machining currents. This change corresponds to an increase in the virtual-waist distance behind the mirror from 28 to 64 μm. The apparent width of this virtual waist has been approximated by applying the cylindrical-beam expansion law[30] to the lower current mirror scans of Fig. 9, equating the appropriate astigmatism distances to the wavefront radii of curvature. Values of approximately 3.0 μm for the virtual-waist $1/e^2$-intensity full-width at high currents are obtained by this procedure.

Junction-plane profiles scanned at an axial distance of 125 mm for the laser of Fig. 11 and the mirror-emission profile scanned perpendicular to the junction plane at 180-mm axial distance are shown in Figs. 12a and 12b. The 125-mm plane was chosen to emphasize the strong side-lobe structure which accompanies the main lobe. If the image axial position is kept fixed and the current increased, the position of the main peak shifts laterally, and the side lobes build up and intrude into the main lobe. The loci of individual side lobes can be traced as functions of axial position in the image field with the aid of Fig. 11. One sees that the narrowest width of a given side lobe occurs in an image plane different from that of the main lobe and of other side lobes, and that the steering angle of its principal ray is similarly unique, a fact already known from Fig. 6.

Care must be taken that the above description does not mislead. A real "image" in Figs. 11 or 12a derives its identity from being the position in the image field at which the main junction-plane lobe appears to be narrowest. The virtual waist within the laser is the conjugate of this image as demagnified by the microscope objective. However, it must be kept in mind that the illustrated lobe structure in no sense depicts the shape of cavity traveling-wave modes. (The intensity distributions of these modes should be correctly illustrated by the mirror scans of Fig. 9.) The scans of Figs. 11 and 12 represent the response of the microscope objective (and the detecting photomultiplier) to both phase and amplitude at the mirror, in accordance with the Kirchoff-Huygens formulation.[10] The actual phase variation at the mirror parallel to the junction plane becomes more complicated than square-law at high currents and produces the secondary-lobe intensities found in the lens image plane.

As mentioned in Section 4.1, the final beam waist at the machining film is a cylindrically corrected superposition of a junction-plane profile (such as those in Fig. 12a) and of the mirror-emission profile viewed perpendicular to the junction plane (as in Fig. 12b). The dependence of the main-lobe astigmatism distance on current[11,30] causes the machining-waist diameter also to depend on current. This can complicate the rendition of gray scales when images are printed by pulse-amplitude modulation. However, since time-sampled profile measurements show that the mode structure is established within 30 ns of the pulse turn-on time, the use of pulse-duration modulation, as discussed in Section 4.2.1, should eliminate astigmatism current-dependence as a potential problem.

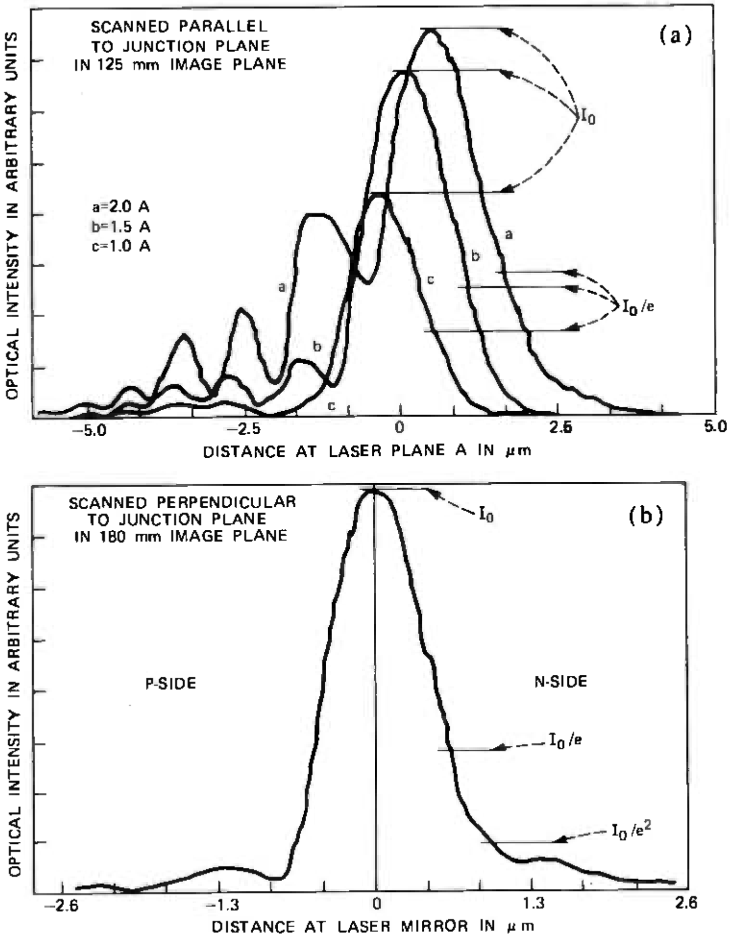


Fig. 12—Intensity profiles in image planes of beam-acquisition lens. (a) At 1.5 A, the profile scanned parallel to the junction is narrowest in the 125-mm image plane. As the current is varied above and below this value, the secondary-lobe structure at 125 mm varies rapidly. All the secondary-lobe peak intensities are below machining threshold, except for the largest secondary lobe at 2.0 A. (b) The 4-mm-focal-length, 0.68-NA, microscope objective focuses the laser output mirror at 180-mm image distance. Profile scans perpendicular to the junction plane at 180 mm are essentially independent of current; the profile full width at $1/e^2$ of peak intensity is 1.50 μm, comparable to the p-region width.

The presence of secondary junction-plane modes* in the machining waist is a different matter. At machining current levels, these lobes

* The permanence of the lobe structure, the regularity of the lobe spacings, and the apparent interrelatedness of their amplitudes are evidence that the structure represents higher-order, guided, junction-plane modes rather than independent lasing filaments. Filamenting, when it occurs, usually involves violent and erratic beam displacements and beam steering within the 100-ns-duration pulse, which bar use of the laser for image printing.

can carry an appreciable fraction of the beam power, yet their individual amplitudes are insufficient to machine (nor would it be desirable to have them machine). Typically, they are strung out asymmetrically on one side of the main (machining) lobe, and because lateral thermal conduction is very small in the film being machined, they contribute little to the attainment of machining threshold within the main lobe. The loss of machining efficiency implied by their existence is a serious defect of the present laser design, which relies on gain guidance to control the junction plane modes. The side-lobe amplitudes could be reduced optically only by non-square-law, cylindrical, phase corrections which, even if physically realizable, would be *ad hoc* with each laser. Practical control of the mode displacement that permits the secondary lobes to arise requires further laser design effort. (The GaAs machining lasers used in video disk recording also utilize gain guidance but operate at much lower powers than in the present application; acceptable junction plane mode control is obtained by narrowing the proton bombardment stripe to 5 μm.[31])

Real-refractive index guidance in the junction plane can be provided by burying the laser active p-region in ternary layers of lower refractive index. By controlling the index differences and the active region width in these buried-stripe heterostructure devices, lowest-order transverse-mode operation parallel to the junction plane has recently been obtained[32] at pulse currents up to nine times lasing threshold. At this level, had the laser been antireflection-coated, the pulsed output power would have been enough to machine Se/Bi film. Astigmatism is absent with real index guidance, and it appears that buried-stripe heterostructures can be designed to yield near-field ellipticities smaller than 3:1. Hence, cylindrical optical corrections may not be needed. Whether the buried-stripe geometry would present more severe thermal limitations at printing duty cycles than the proton-delineated stripe geometry is unknown, but this approach to junction plane mode control certainly merits consideration in future studies of GaAs machining lasers.

4.3 Laser antireflectance coatings

Antireflection coatings are needed on the laser output mirror for a variety of practical reasons, the most important from an image-printing standpoint being laser longevity. Such coatings can double the output-mirror differential quantum efficiency, furnishing machining-level powers at significantly reduced pulse currents; they can suppress higher-order modes and improve power stability, increasing the device yield of thick-cavity wafers; and they can protect the output-mirror facet against harsh chemicals and abrasion. Although refractory coatings are available which can reduce facet reflectivity to values smaller than 0.1 percent, the resultant laser would be highly sensitive to the mode-destabilization effect explained in Section 4.2.3. In practice, the coating small-angle reflectivity must be controlled carefully.

Antireflectance coatings usually broaden the laser longitudinal mode spectrum and increase the optical bandwidth that must be chromatically corrected. Hence, their use in high-optical-resolution applications can have the economic drawback of requiring expensive glasses in lens fabrication. The objection is not trivial since the cost of fabricating diffraction-limited printer optics could become prohibitive unless design simplifications of this type were possible.

The antireflection coatings presently used on printing lasers are: (*i*) single-layer films of ZrO_2 deposited on the output mirror in optical thickness of approximately one-quarter or three-quarter wavelength and (*ii*) bilayer films consisting of ZrO_2 overcoated with Al_2O_3. These refractory oxides are deposited by electron-gun evaporation at $\sim 10^{-5}$ Torr of O_2 partial pressure. The measured ZrO_2-film refractive index is 1.92, and the Al_2O_3-film index is 1.63, at 0.885-μm wavelength.

The refractive index of ZrO_2 is a close antireflection match to the value 3.605 of the refractive index of binary GaAs, and ZrO_2 films can reduce the forward angle reflectivity to values smaller than 0.1 percent when deposited in exact odd-quarter-wavelength optical thicknesses. Unfortunately, its index is not sufficiently larger than the Al_2O_3 index to permit much exploitation of bilayer design possibilities. Higher index materials such as ZnS and TiO_2 are well known. However, the ZnS/Al_2O_3 bilayers are found to suffer from limited room-ambient shelf life, and the high substrate temperatures needed to make thin TiO_2 films in rutile form are a deterrent.

Properly fabricated coatings can greatly assist in achieving exclusive fundamental-mode output from thick cavities; namely, they can provide an assigned small-angle reflectivity together with deep reflection minima in both polarizations[33] at a specified large angle. Both polarizations must be accounted for, since the laser output is usually of mixed polarization at machining power levels. In Fig. 13, we have superposed the angular distributions of 0- and 1-order modes* measured for a particular laser onto the Fresnel-reflectivity angular dependences of a ZrO_2/Al_2O_3-bilayer design to illustrate that reflection losses can be made significantly larger for the 1-order than for the 0-order mode. Discrimination between the 0- and 1-order modes would clearly be largest if coatings whose reflection minima occur near the 1-order mode maximum could fabricated. Unfortunately, with our lim-

* Reflection-loss angular dependence discriminates only between modes perpendicular to the junction plane. The junction-plane modes have comparatively narrow angular distributions which are sensitive only to the reflectance values at forward angles. Those values are asymptotically independent of angle.

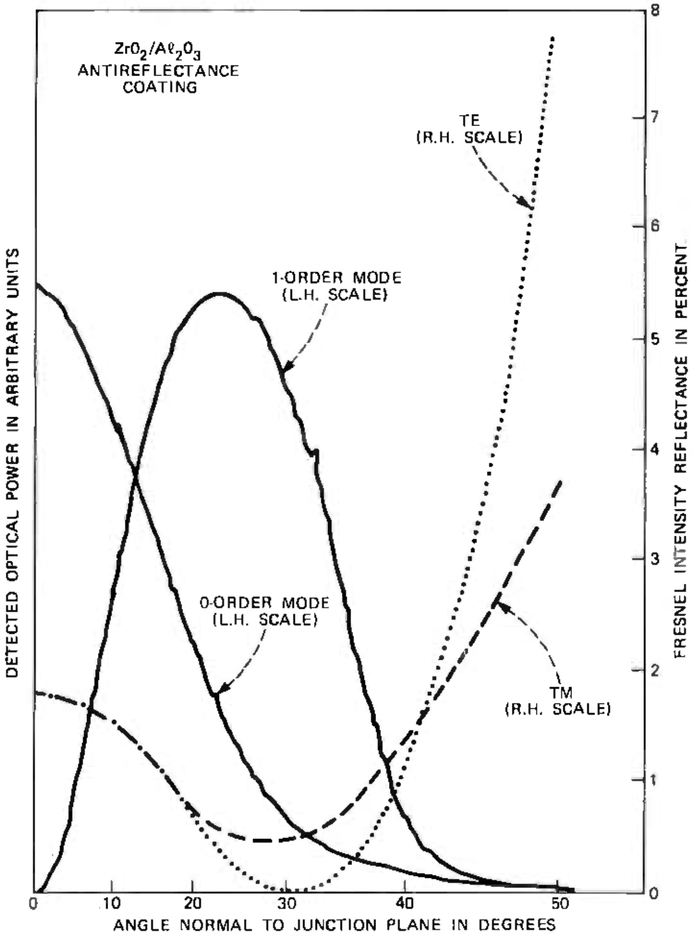


Fig. 13—Mode discrimination perpendicular to junction plane by output-mirror antireflection coating. The antireflectance design is chosen to produce minima in the TE and TM Fresnel reflectivities at angles as close as feasible to the maximum in the 1-order mode intensity. The necessity of maintaining at least a 2-percent small-angle reflectivity value (to inhibit mode destabilization by reflections from the micrographics film) makes it difficult to secure exact overlap between the Fresnel-reflectivity minima and the mode-intensity maximum, but rough approximations to this condition are sufficient to produce significantly higher mirror losses for the 1-order mode than for the 0-order mode.

ited selection of refractory coating materials, reflectivity minima at angles below 25 degrees are predicted only for designs whose 0-degree reflectivity is smaller than 0.7 percent, a situation very susceptible to external-reflection-induced destabilization.

Nevertheless, valuable amounts of mode discrimination can be obtained with the present coatings. Fresnel-reflectance measurements on a selection of coated output mirrors show that our single-layer and bilayer films produce reflection minima for both polarizations in the range 25 to 40 degrees, with 0-degree reflectivities ranging from 2 to 6 percent. These measured Fresnel-reflectivity angular distributions have been multiplied by the mode angular distributions of Fig. 13 and integrated over angle to obtain *mode reflectivities*,[34-36] and hence comparative mirror losses, for the 0- and 1-order modes. The loss discriminations so computed significantly favor 0-order over 1-order oscillation and are compatible in detail with the switching of numerous lasers from oscillation in the 1-order mode (or in a mixture of 0- and 1-order modes) prior to coating to the 0-order mode exclusively after coating. (The 0-order device yield increased from 50 to 83 percent on one thick-cavity wafer.)

An antireflection coating applied to the front mirror increases its differential quantum efficiency and reduces that of the rear mirror, as shown by the laser output characteristics in Fig. 14. These coatings increase the laser threshold current slightly, but allow machining powers to be reached at significantly reduced currents and, more importantly, at lower peak values of laser internal circulating power.[37,38]

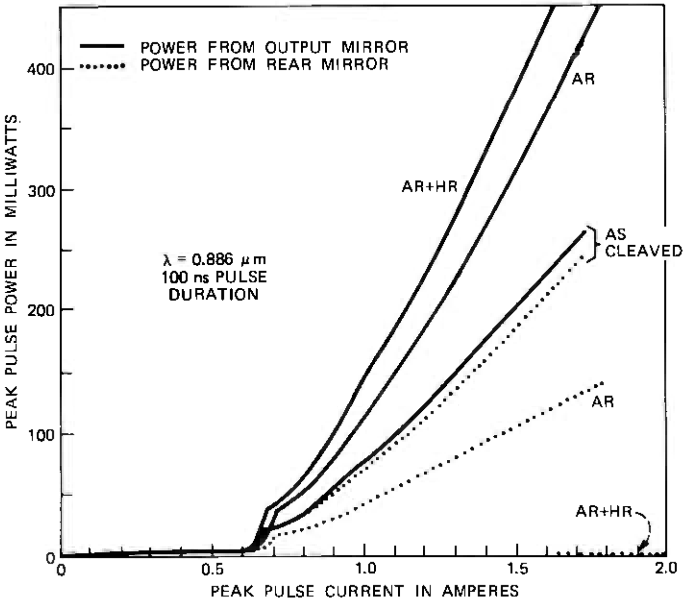


Fig. 14—Effect of mirror reflectance coatings on laser output power. The output- and rear-mirror powers in milliwatts are shown for: (a) The as-cleaved, uncoated mirrors. (b) Antireflection-coated output mirror and uncoated rear mirror, *AR*. (c) Antireflection-coated output mirror and high-reflectance-coated rear mirror, *AR* + *HR*. The increase in threshold current caused by *AR*-coating the output mirror is reduced when an *HR* coating is applied to the rear mirror.

If the rear mirror is now coated with a high reflector (a one-half wavelength optical thickness of ZrO_2, overcoated with aluminum), the threshold current approaches the as-cleaved mirror value, and the front-mirror output power at constant current increases further. This combination of coatings increased the front-mirror differential efficiency by a factor of 2.3 relative to its uncoated value, as determined by averaging over many lasers.

4.4 Laser longevity

End-of-life in thick-cavity, printing lasers has occurred from adverse changes in output mode structure and from a gradual loss of mode stability against external-reflection-feedback, as well as from diminished output power. Only the first and last aging symptoms have been investigated quantitatively, and these only on devices from one wafer. Gradual cavity destabilization has been seen on ZnS/Al_2O_3-coated lasers operated in room ambient; but it is absent, so far, from ZrO_2- and ZrO_2/Al_2O_3-coated devices.

4.4.1 Median life and failure symptoms

Antireflection-coated lasers from one wafer were pulse-current aged at 5 or 10-percent duty factors and at 20°C (time-averaged) junction temperature. The output mirror coatings represented various ZrO_2- and ZrO_2/Al_2O_3-evaporation runs. The rear mirrors of some lasers were high-reflectance coated, producing 20 to 30 percent higher internal optical powers at the output mirrors of these lasers than in the remaining aged devices. The median life against total power failure was 3×10^{11} pulses ($\sigma = 3.45$) independent of duty cycle, with the failure rate of high-reflectance-coated lasers being significantly greater than the average failure rate. Aging produced both favorable and adverse changes in machining performance of the survivors. Specifically, the difference between machining and lasing threshold currents, which is affected by details of the focussed beam profile and hence is sensitive to mode degradation, was—on the average—unchanged among survivors.

Failure was uncorrelated with the post-aging appearance of the mirrors, and mirror erosion was detectable on only two of the aged devices. However, the ratio of output-mirror to rear-mirror power was decreased among survivors, indicating that reduced net transmission of aged, coated mirrors (relative to as-cleaved mirrors subjected to somewhat lower optical power densities) had occurred despite the absence of visible damage. Scanning-electron-microscope examination of several failed lasers, using the electron-bombardment-induced recombination current at various primary energies, suggested that the devices had been damaged over as much as 75 percent of the device length, with the damage originating at or near the output mirror. (The power loading of these mirrors is less than ⅛ the catastrophic damage limit reported[37] for similarly constructed lasers.) An increase of at least 40 mV in junction voltage and a softening of the laser V-I characteristic were observed in nearly all failed or severely degraded lasers, but were absent in healthy survivors.

4.4.2 Duty-factor limitation

Although a commercially feasible life* against power decrease, mode degradation, and mode destabilization seems established (at 20°C junction temperature) for the above devices, the data do not establish that satisfactory images can be machined at the 5 and 10-percent duty factors used in the aging tests. Attempts to machine images at 10-percent duty factor produced unintended hole size decreases from beginning to end of long machined lines, even though the junction was held at a time-averaged temperature of 20°C. The thermal relaxation time of 150 ns estimated for the laser active region is apparently not short enough to prevent a gradual build-up of junction temperature during the machining of many contiguous holes at 1-MHz repetition rate.[39] The build-up raises the laser threshold current and therefore lowers the machining power available for a given pulsed-current amplitude and duration. It is not known whether a significant reduction in the relaxation time could be obtained by redesigning p-cap and $P_{0.24}$-epitaxial thicknesses without degrading the already marginal junction-plane mode stability.

V. THE FILM SYSTEM

The laser machining performance of many *single-element* metal films, as fabricated by evaporation, sputtering, or electrodeposition onto a transparent substrate, has been investigated with high-pulse-power, cavity-dumped lasers. However, single-element metal films are found to be fundamentally limited in their ability to absorb light. In thicknesses which produce acceptable "black-white" contrast ratios at visible wavelengths, none achieves low enough machining threshold to serve satisfactorily with the present GaAs laser. Also, the long-term chemical stability—and hence the machining reproducibility—is poor for low-melting-temperature metals in oxidizing environments, limiting the room-ambient storage life of unused film.

Because of this unsatisfactory single-metal performance, a search for lower machining energies and improved chemical stability has been made among multiconstituent films. Over 70 film systems, including

* The median life of 3×10^{11} pulses is sufficient to machine about 90,000 all-white frames, and probably more than 500,000 negative-image frames.

eutectics, alloys, and metal-dielectric combinations, have been investigated. These were chosen primarily, but not exclusively, from low-melting-point elements. Antireflective combinations which might improve the film chemical stability while increasing its optical absorptance have been of special interest. The compositions were studied for machining performance as functions of laser power, for storage life (i.e., oxidation resistance) as determined by changes in optical transmission and machined hole size, for scratch resistance, and for adhesion. The next sections identify some of the chemical and physical attributes believed to affect hole formation in unoxidized and partially oxidized thin films, and attempt to justify the selection of a Se/Bi composition as a practical metal-dielectric film for facsimile image machining with pulsed GaAs lasers.

5.1 Hole machining in thin metal films

5.1.1 General description; hole-nucleation hypotheses

A machined hole represents the permanent rearrangement of film material in response to a localized temperature pulse exceeding a threshold temperature.* The final state consists of a debris pattern—typically, thickened peripheral rims (viz., Fig. 1a) accompanied by congealed metal droplets—whose surface energy is smaller than that of the unruptured film but whose optical transmission is much higher. Heat is absorbed from the focused laser beam in unruptured film areas and is radially redistributed via the debris pattern. Very nearly all the absorbed heat is eventually lost to the substrate by laminar conduction.

A free liquid surface is required for hole enlargement. The kinetic energy of the molten constituents as they recede towards the solid boundary or are flung slightly beyond it is usually attributed[40] to the surface energy decrease. Debris is ejected from the film surface at high power densities, possibly with the assistance of pressure from entrapped gases. During raster-scanned laser machining of large-area background "whites," these ejecta are evolved from the film in numbers sufficient to form a highly visible plume whose density is a convenient indicator for optimizing the machining beam focus. Redeposition of these "tektites" onto the film can degrade the image (viz., Sec. 8.3 and Fig. 46b) and must be kept under control. The avoidance of machining "splatter" and the production of clean holes with uniform rims are central problems[7,41] in attaining broadcast-quality signal-to-noise ratios with reflection-readout video disks.

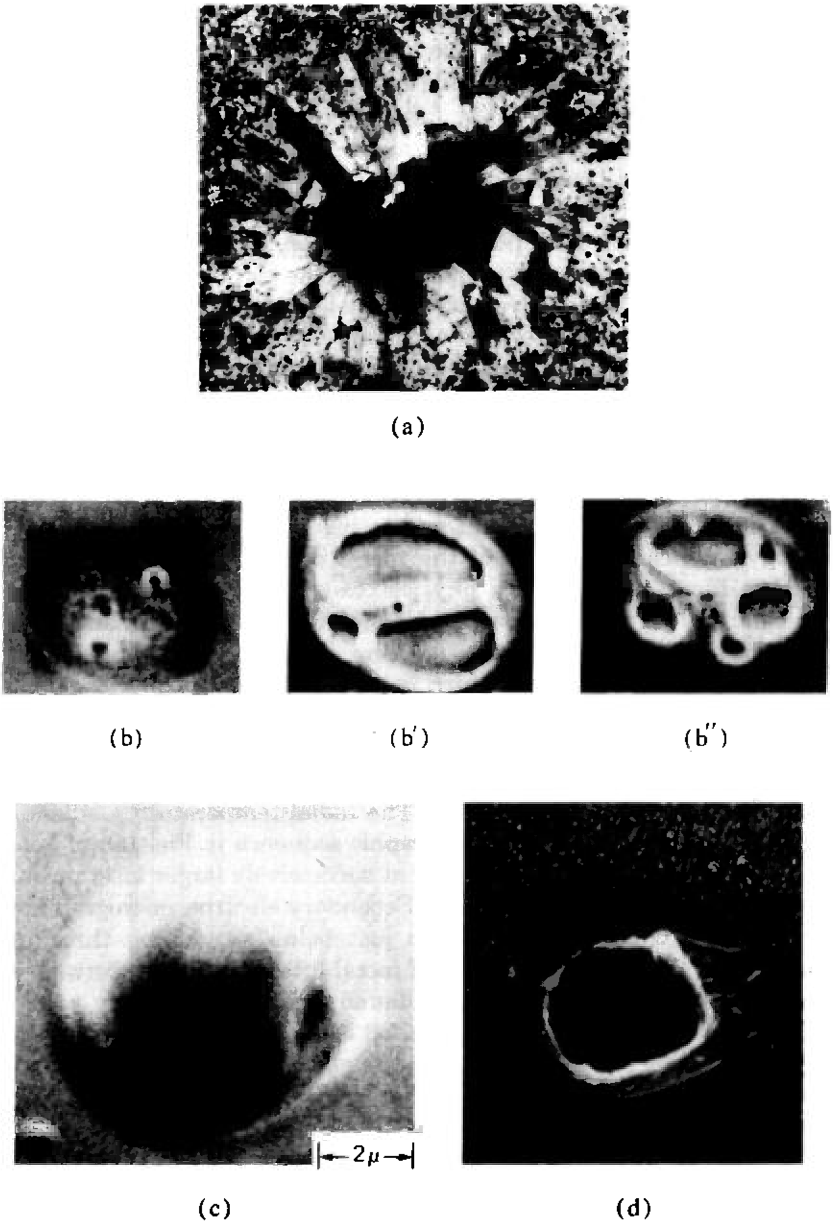
The thickness of a film and the severity of its oxidation strongly

* Holes can also be opened by the repeated application of subthreshold pulses. At a low duty factor, approximately 16 pulses, each containing about 20 percent less energy than the threshold energy for one pulse, were needed to open a hole in 540Å-thick bismuth film.

affect the laser pulse energy needed to nucleate and enlarge a hole. The film-substrate interface also affects hole-opening dynamics via the interfacial surface energy and thermal impedance. Studies of hole-opening kinematics by a finely focused, very-low-power, laser probe beam, incident onto a surface-oxidized, 350 Å-thick bismuth film simultaneously and collinearly with the machining beam, showed that at powers just above machining threshold a hole tends to nucleate after the end of the laser pulse and, at powers five or six times larger than threshold, holes start to open before the end of the pulse, but at powers between these extremes, most holes nucleate near the end of the laser machining pulse. The occurrence of hole nucleation slightly before, or at, the end of the machining pulse is also compatible with an independent experiment on a 1200Å-thick composite film in which close agreement of machined hole areas with the predictions of laminar heat flow theory was obtained (viz., Sec. 5.3).

Hole nucleation in very thin films (≤250Å of bismuth) is generally agreed to originate at pinholes and to involve a reduction in surface energy by local balling up of the metal into numerous droplets. In thicker films, however, the physical causes of hole nucleation are speculative. Figure 15 presents evidence for two hypotheses. Figure 15a shows the onset of hole opening in 540Å-thick, pinhole-free, bismuth film, apparently occurring at a multiplicity of sites in the neighborhood of, but not immediately beneath, beam center. This observation supports a hypothesis of hole nucleation in thick films involving changes in surface-tension gradients[42,43] during the transition from heating to cooling the film. (The radial temperature gradient is zero at beam center.) The photographic sequence in Fig. 15b of holes opened in an unoxidized Ni-P film at successively larger laser powers further illustrates this hypothesis. Secondary electron micrographs of thick films irradiated at energies just below machining threshold sometimes show circular blisters of metal lifted from the substrate, as in Fig. 15c. Blistering has been attributed to pressure from gas evolved from the heated substrate and trapped between it and the metal film, or alternatively to buckling of the film due to thermal expansion of the heated area. In either case, it is hypothesized that nucleation proceeds by rupture of the blister. Further evidence for buckling or loosening of film from the substrate is seen in the wrinkling of the irradiated area in Fig. 15d.

Freezeback of molten material begins when the pulse ends, and a solidification front propagates inwardly[44] towards beam center, terminating the outward displacement of molten material and defining the final hole diameter. The relevant freezing temperature may be the melting temperature of the metal or of a mechanically stiff oxide; in some volume-oxidized films, it may approach the metal boiling tem-



(a)

(b)

(b′)

(b″)

(c)

(d)

Fig. 15—Evidence for hole-nucleation hypotheses. (a) Transmission electron micrograph of 540Å-thick, free-standing, Bi film irradiated just at machining threshold. Three very small holes near, but not beneath, beam center are indicated by arrows. (b) to (b″) Three areas in NiP film irradiated at successively higher powers. (c) A blistered area in NiP film irradiated below machining threshold. (d) Hole in 400Å-thick Bi film. Wrinkled annulus surrounding lip of the hole suggests that film may have lifted from substrate prior to hole nucleation.

perature. Thus, the hole is circumscribed by an annulus of resolidified film material whose outer diameter may be only slightly larger than the machined hole diameter if the film oxide structure is fragile, but which can be considerably larger than the hole diameter if the oxide structure is strong (particularly for holes machined close to threshold).

5.1.2 Experimental description of laser-machined holes; hole-enlargement hypothesis

The sizes of holes machined by pulsed laser beams in thin metal films, and the patterns of the resulting debris, depend on the film thickness, the machining pulse intensity and duration, and the extent of film oxidation (film age). Figure 16 is a montage of scanning electron

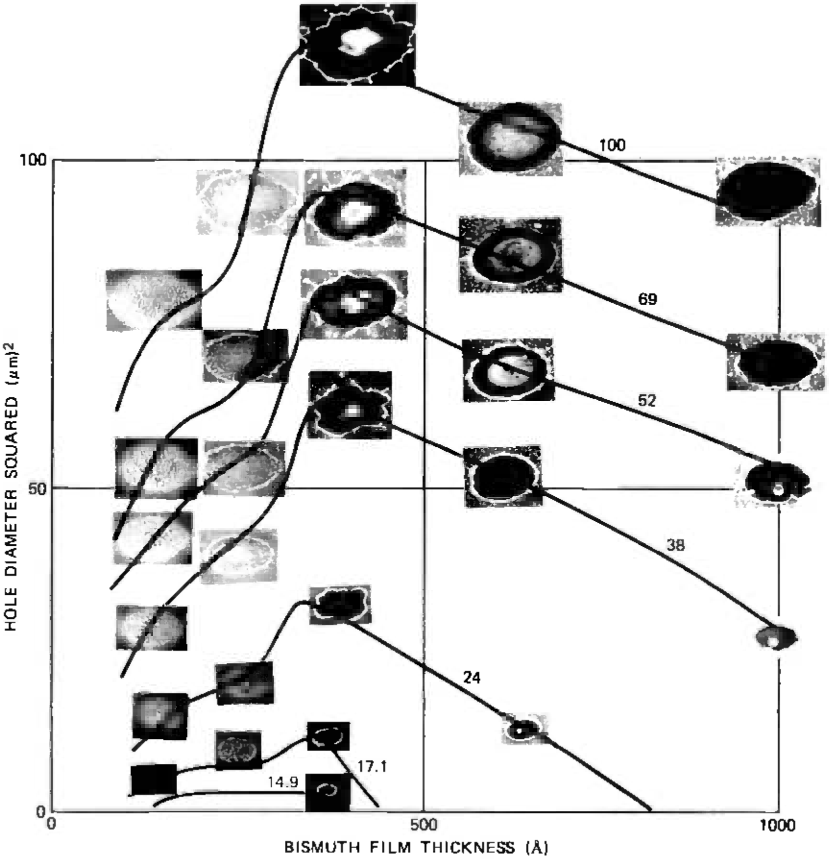


Fig. 16—Size and appearance of holes machined in bismuth films of various thickness, with laser pulse energy as a parameter. The curves connect data obtained at the same laser pulse energy and are labeled by the energy value as a percentage of 120×10^{-9} joules (corresponding to 4.0-W peak power at the measured 30-ns duration of the cavity-dumped YAG-laser pulse). See text for discussion.

micrographs showing the holes machined in evaporated bismuth films of various thicknesses by 30-ns-duration laser pulses of variable intensity. The focused beam diameter was 6.8 μm at 1/e intensity. These data reveal the following:

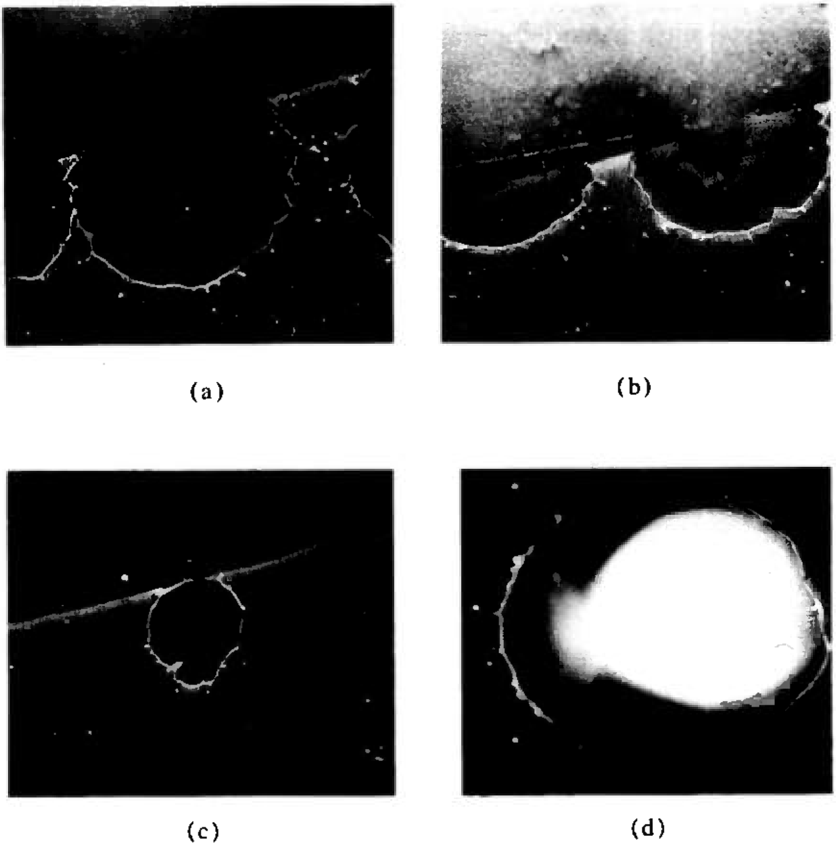
(*i*) The hole areas are maximized in the neighborhood of 350 to 400 Å film thickness. The debris in 400 Å-thick film machined at low intensities congeals into a raised lip surrounding the hole, and at high intensities into a characteristic pattern of balled-up metal droplets flung out beyond the lip. This type of splatter has been attributed[7] to the relatively low viscosity of molten bismuth. The photographs in Fig. 16 support the generalization that droplet size is an increasing function of film thickness.

(*ii*) Machining efficiencies diminish rapidly as the film thickness is decreased below 350Å. This thickness is coincident with the value below which the continuity of bismuth film is poor,[45] resulting in a proliferation of pinholes throughout the finished film. The debris pattern consists of droplets remaining within the hole; for the thinnest film, the metal droplets do not appear to have moved outwardly at all.

(*iii*) As the film thickness is increased above 400 Å, the hole areas machined at given power decrease inversely with thickness (viz. also Fig. 19). The debris from holes machined close to threshold consists of one large droplet lying within a thick peripheral rim. At high powers, the droplets lie outside the hole, and the rim thickness is nonuniform.

Figure 17 provides a more detailed look at hole opening in "optimally-thick" bismuth films. Holes of incomplete circumference (viz. Figs. 17a and 17b) have been formed by scanning the machining beam along a line making a shallow angle with a sharply defined edge in 400 Å-thick film. Their debris patterns are similar to those from nearly complete (Fig. 17c) or complete (Fig. 17d) holes, in that the rims of the holes are formed into *azimuthally discrete segments*. The number of segments increases with hole size, but the average segment length is roughly constant. Regularly spaced metal droplets of approximately constant size either congeal out directly onto the peripheral lip (Fig. 17d) or are flung radially beyond it, often remaining attached by a thin thread. Sometimes the thread breaks, producing detached, comet-like droplets.

The regularity of the azimuthal patterns in Fig. 17 tends to invalidate a "brush-fire" hole enlargement model (such as might be kindled by the 250Å data of Fig. 16) in which the melt surrounding one or more hole nucleation sites contracts into independent droplets which random-walk over the film surface and disappear by consolidating into a cooler rim. Figures 1a and 17 hint instead at a propensity of *peripheral* material in thicker films to form beads, implying that the rim is the source of droplets rather than the reverse. The rim of an expanding hole is "superheated" since it contains molten material which originally lay closer to beam center. The initial accretion of material to form a rim may have involved the surface-tension-controlled, iris-like withdrawal of molten metal away from the site of hole nucleation—an hypothesis which is compatible with a widely observed tendency of small holes to be smooth-edged and circular. However, further expansion must proceed by drawing cooler liquid metal into the rim from an



(a)

(b)

(c)

(d)

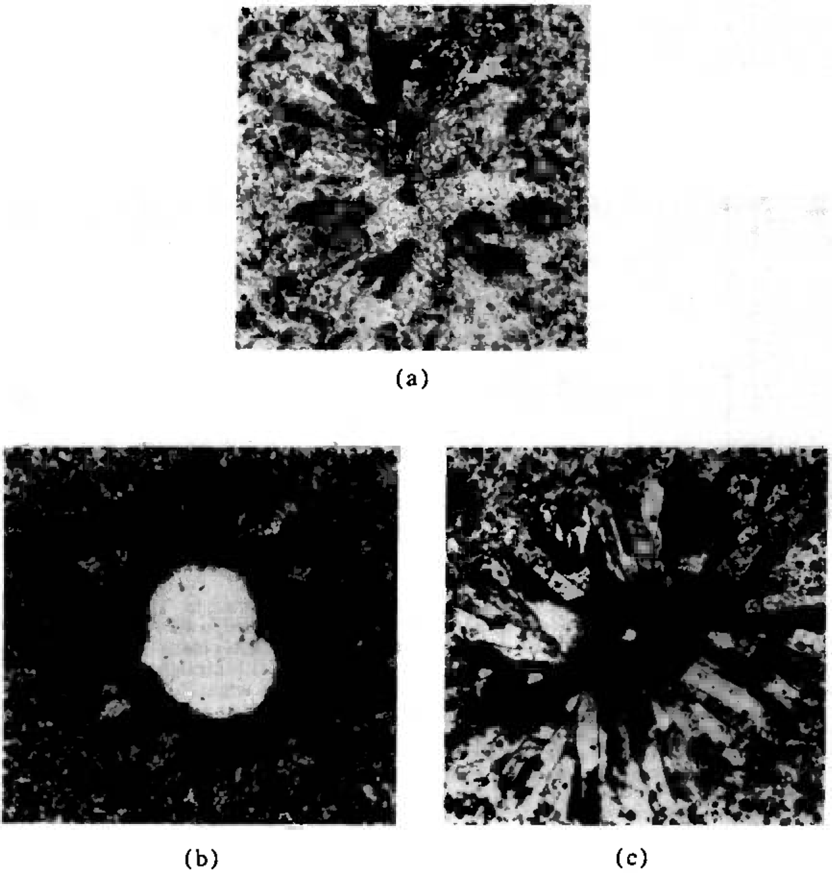
Fig. 17—Regularities in the debris patterns of holes machined in "optimally-thick" bismuth films. The following descriptions refer to holes machined in 400Å-thick bismuth evaporated onto Mylar.* A 30-ns-duration, cavity-dumped, YAG-laser beam of 6.8-μm diameter at $1/e^2$ of maximum intensity was employed. (a)–(c) A straight film boundary was sharply defined by masking the evaporation with a knife edge. Holes of incomplete circumference were formed by translating the machining beam at a shallow angle to the boundary. Vapor-pressure buildup could not have occurred with these holes. Segmentation of the hole rims and the regular spacing of metal drops around them are clues to the hole opening kinematics. (d) Regularly spaced beads are visible in the rim of a complete hole.

* Registered trademark of E. I. DuPont de Nemours.

incrementally larger radius and redistributing it circumferentially into beads whose spacing and maximum size are characteristic of the film thickness. Once it acquires spherical shape, a bead is somewhat insulated both from the incident beam and from the rest of the film and may persist briefly as a superheated liquid droplet after its surroundings have resolidified.

5.2 Film oxidation

The preceding sections hint that the laser-machining shelf life of thin metal films is limited primarily by the growth of mechanically rigid oxide structures which must be melted or ruptured before hole nucleation and enlargement can proceed. *Melting* the oxide may permit hole nucleation in pinhole-free, thick films subject only to surface oxidation; but *rupture* by attainment of high metal vapor pressures is required to nucleate holes in films which contain a high



(a)

(b)

(c)

density of oxide-filled pinholes or whose preparation has incorporated oxygen into the crystalline grain boundaries. Evidence for the types of surface and bulk oxide structures present in thin metal films, and their effects on hole machining, is given in Figs. 18 through 21.

5.2.1 Surface oxidation

Figure 18 consists of transmission electron micrographs of laser-irradiated areas of 540Å-thick, free-standing bismuth film. The area shown in Fig. 18a was irradiated with two or three laser pulses, each of energy sufficient to melt the metal but not the oxide; the metal is seen to resolidify after each pulse into large, radially oriented crystallites while the oxide (faint pattern within the grains) retains the structure of the original, as-evaporated metal crystallites. If the area is subjected, as in Fig. 18b, to a total of 20 such pulses, a hole is eventually produced, the opening being bridged by an intact skeleton which can be identified as a surface oxide.

The practical effect of this surface oxide on laser-machined miniature images is revealed by Figs. 18c, 18d, 18e, and 19. Unmelted surface oxide surrounds the holes in Figs. 18c and 18d. These holes were

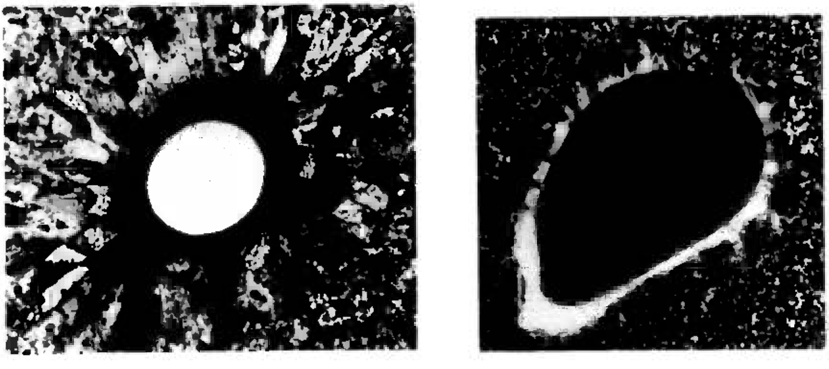


Fig. 18—Resolidification in bismuth films at various machining energies, illustrating nucleation delay caused by surface oxidation. Transmission electron micrographs of 540Å-thick, free-standing bismuth film irradiated with a cavity-dumped He-Ne beam for the pulse-power conditions described in the text. Very similar data are found for 700Å-thick films. The areas that have melted and resolidified are defined by the large, radially oriented gains. The low-contrast tracery in (a) and (b) establishes the existence of a thin surface oxide whose structure corresponds to the crystalline matrix of the as-evaporated metal. This tracery remains visible in (c) and (d), which depict holes machined just above threshold. Hole diameters near threshold are much smaller than the diameter to which the film was melted; i.e., because the hole opening is restrained by the surface oxide, the temperatures underneath beam center were able to reach values significantly higher than the 660°C equivalent temperature corresponding to the enthalpy to melt bismuth. Large holes (e) consume a much larger fraction of the maximum liquidus diameter. Note that the annulus surrounding the hole of (e) has recrystallized by heat flow perpendicular to the hole boundary.

produced by single pulses at energies just above machining threshold, and they are very much smaller than the diameter to which the metal was melted. In contrast, hole diameters approach the molten diameter at high machining energies, as shown in Fig. 18e. This tendency finds quantitative expression in the "two-diameter" picture of hole machining suggested by Fig. 19. Data points lying within the dotted region of this figure represent the hole "areas" (hole diameters squared) machined in numerous bismuth films ranging in thickness from 420 to 980Å as functions of volume energy density (i.e., absorbed machining-energy-density per unit film thickness). Film temperatures have been associated with the abscissa by using the experimental result (obtained

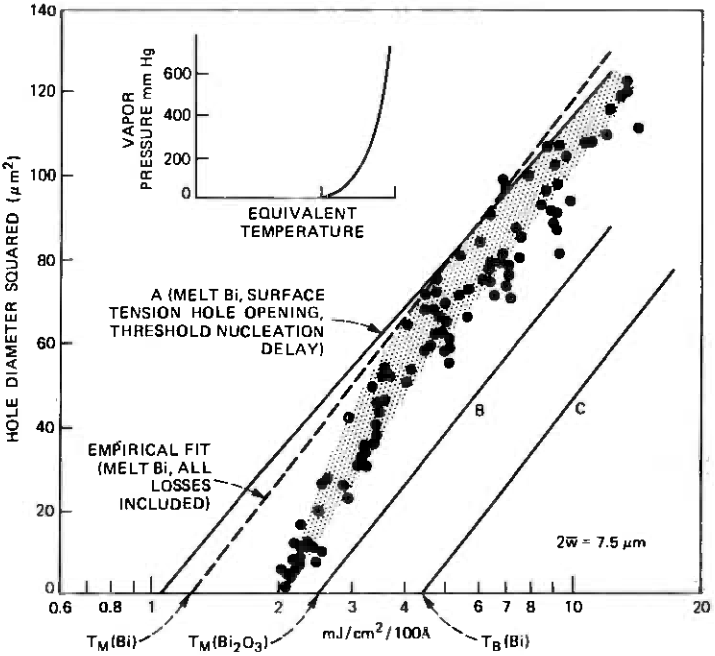


Fig. 19—Dependence of machined-hole areas in surface-oxidized bismuth films on absorbed volume-energy-density: hypothetical and measured dependence. The ordinate is the square of the hole diameter in units of $(\mu m)^2$. The abscissa is the logarithm of absorbed volume-energy-density, in units of millijoules/cm^2 in the focused machining beam per 100Å of bismuth thickness. A temperature scale derived from film-resistivity data (see text) has been fitted to the abscissa. The equivalent-temperature scale of the vapor pressure (inset) corresponds to this abscissa. The data (dotted region) represent the diameter squared of holes machined in numerous bismuth films, ranging in thickness from 350Å to 1000Å, by a pulsed YAG-laser beam of 7.5-μm diameter at $1/e$-intensity and of 30-ns duration. Incident beam power was varied with neutral-density filters. Film optical absorptions were measured in independent experiments. Curve *A* was calculated from eqs. (2) and (3) of the text for a 900Å-thick film, assuming τ = 300 ns thermal-relaxation time and hole-nucleation delay time shorter than the laser pulse duration. Lines *B* and *C* were obtained from eq. (2) evaluated for $\tau = \infty$. Line *B* corresponds to hole diameters determined by the enthalpy needed to reach the Bi_2O_3 melting point, line *C* to that needed to reach the Bi boiling point.

by monitoring film resistance during machining, viz. Section 5.3.1) that the pulse energy needed to nucleate holes in 500Å-thick bismuth film is approximately twice that needed to melt the metal (including heat of fusion) and hence is enough to raise the film to 900°C, slightly above the Bi_2O_3 melting point. The semilogarithmic straight lines A, B, and C in Fig. 19 are the predicted[46] machining curves if the final hole diameters in 500Å-thick bismuth were determined, at all pulse energies, by freezeback of the bismuth liquidus, by freezeback of the Bi_2O_3 liquidus, or by attainment of the bismuth boiling temperature, respectively.

The rapid dependence of hole area on machining energy at the transition between melt-oxide and melt-metal machining explains the observation that gray-level tonality is very hard to control near the "black" end of the contrast scale in unprotected films of low-melting-temperature metals (such as evaporated bismuth) and in alloys which do not form intermetallic compounds (such as the Bi/Pb-alloy). Such films usually lose the fight against surface oxidation within a few hours, and their gray-scale reproducibility is especially sensitive to film aging. Indeed, definition of the final hole diameter in terms of the metal melting enthalpy at all pulse energies—i.e., curve A of Fig. 19—is observed in the case of bismuth only for *in vacuo* machining of freshly evaporated metal. Also, machining and SEM data on metallic Ni and electroless nickel[47] indicate that these slow-oxidizing films remain in the "melt-metal" mode for about 10^5 hours at room temperature. Further, hole formation in certain multiconstituent films which remain free of oxygen throughout most of their lives can be described in terms of a chemically stable "melt-compound" mode analogous to the "melt metal" mode of curve A. This machining process, described in Section 5.4, applies to fairly thick composite films, such as Se/Bi, which can form tightly bound interfacial compounds (Bi_2Se_3) that prevent bulk oxidation of the bismuth.

5.2.2 Bulk oxidation

5.2.2.1 Pinhole-Rich Films. The deliberate fabrication of pinhole-rich films has been proposed as a means of artificially seeding the film with hole nucleation sites and lowering the threshold machining energy. The discussion of Fig. 16 shows that this scheme won't work. The film thickness for maximum machining efficiency coincides with a critical value—dependent on substrate surface roughness and on film deposition conditions—below which the layer uniformity and continuity of bismuth evaporated onto Mylar* are poor, resulting in high pinhole density in the finished film. Mechanically stiff oxides develop readily

* Trademark of E. I. DuPont de Nemours Company.

within the volume of pinhole-rich films and impair machining. Electrical conductivity measurements support the hypothesis that extensive oxidation at the metal-substrate interface, centered around the foot of each pinhole, lifts metal from the substrate because of the increased volume displaced by the oxide. This exposes an appreciable area of fresh metal around the base of each pinhole to reaction with oxygen and increases the size of the foot.

Sufficiently thin free-standing films are also pinhole rich. Figure 20 consists of SEM and TEM micrographs of 340Å-thick, free-standing bismuth films laser-machined at different pulse energies. These suggest that hole nucleation may occur at several locations near beam center, perhaps caused by independent melting of the oxide pillars lining the pinholes. Figures 20b and 20c show that liquid metal released by

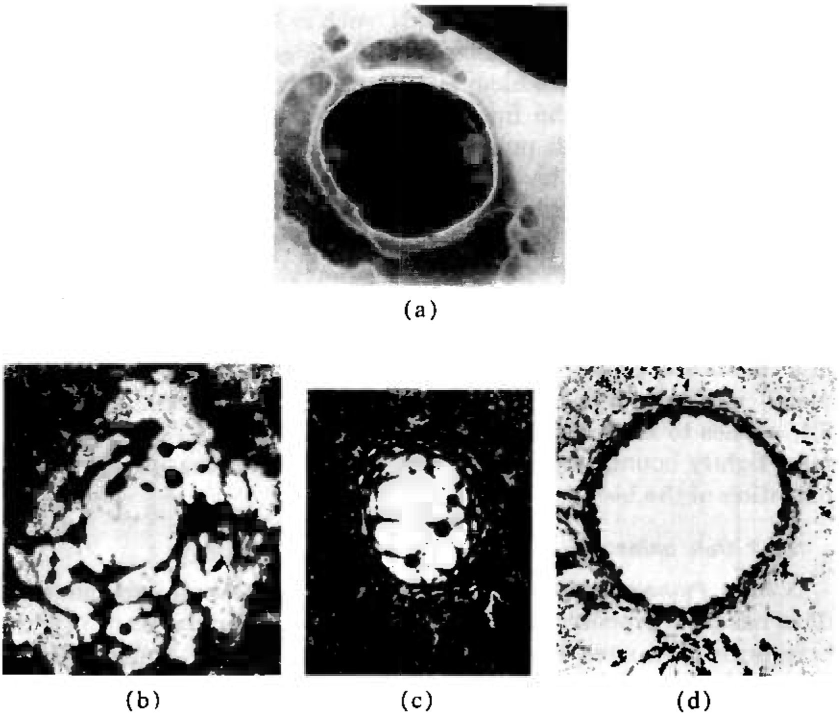


Fig. 20—Hole machining in presence of bulk oxidation. The micrographs for this figure were obtained from 340Å-thick, free-standing bismuth film machined by 25-ns-duration pulses from a cavity-dumped He-Ne laser. Heavy oxidation of this film can be presumed because of the exposure of both its surfaces to atmospheric oxygen and because of the presence of numerous pinholes. (a) SEM micrograph suggesting an oxide membrane surrounded by resolidified metal and perforated by several holes frozen in the process of coalescing into one hole. (b)–(d) TEM micrographs of areas irradiated at progressively higher powers. See text.

independent meltings beads up into nodes connected by oxide threads. These threads contract toward the oxide liquid boundary, pulling the nodes after them and producing the clustering seen in Fig. 20c. The outer resolidification boundaries in Figs. 20b, 20c, and 20d are identified with the bismuth liquidus, and the inner boundaries with the Bi_2O_3 liquidus; a similar identification is suggested by Fig. 20a. The machined hole diameters are seen to equal, at most, the diameter to which the oxide can be melted.

5.2.2.2 Heavily Oxidized Thick Films. The grain sizes of sputtered metal films tend to be much smaller than those of evaporated films. Complete oxide penetration even of fairly thick films occurred readily in sputtered films of tin and indium, producing individual metal grains imbedded in an oxide honeycomb. Melting the oxide changes this structure from a solid to a liquid honeycomb, but it does not nucleate holes. Experimentally, hole nucleation requires that the honeycomb be ruptured mechanically by heating the metal to approximately its boiling temperature. A hole initiated near beam center then enlarges to the boundary defined by the metal boiling point and, if the oxide structure is strong, stops there. The machining curves of Fig. 21 indicate that hole formation in sputtered tin and indium can be described in terms of this "boiling-point" mode.

However if the oxide structure is weak, it may be torn away when the hole opens, causing the machining to switch from the boiling-point mode near threshold to an oxide mode at high powers. This appears to be the case for sputtered bismuth, as is consistent with electron microscopy data which show that its crystallites are appreciably larger than those of sputtered tin and indium and present correspondingly less surface area for oxidation.

Despite their higher threshold machining energies, a certain interest persists in films of the above type. Since the sputtered coatings represent an extreme state of oxidation, they have the virtue of being comparatively stable against further loss of machining efficiency with age, and can be considered whenever sufficiently high-pulse-power lasers are available. The inclusion in sputtered films of a soluble element to increase the vapor pressure has been proposed as a means of lowering their machining threshold. However, a trial of this proposal on sputtered Bi/Sb film produced no improvement in machining over sputtered bismuth.

Strong oxide honeycombs which require boiling-point machining can be formed in other ways. The oxide Bi_2O_3 has been investigated as an antireflection coating between the transparent substrate and the deposited bismuth film, for use when the machining laser beam is incident onto the metal through the substrate ("back machining," viz. Section 5.4). The Bi_2O_3 is vacuum-evaporated in a high oxygen am-

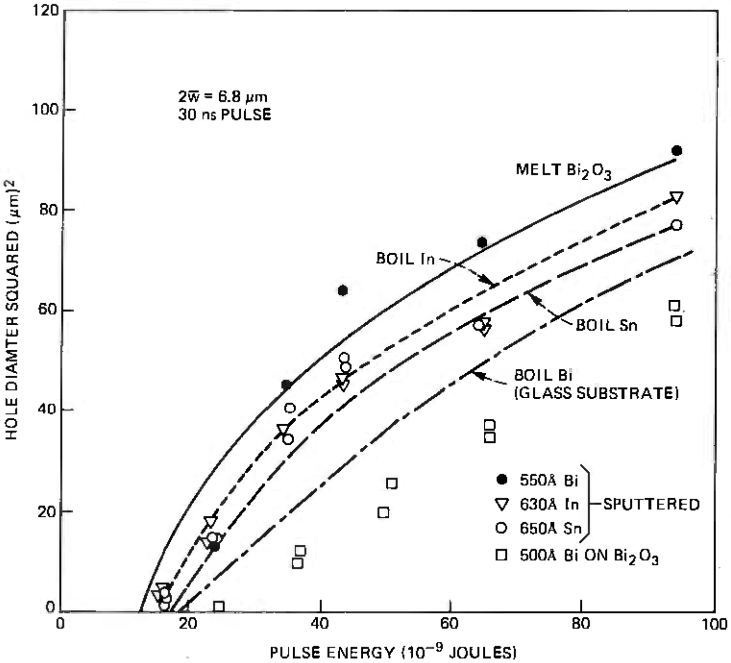


Fig. 21—Machining curves for oxidized films of bismuth, indium, and tin. Squares of machined-hole diameters are shown as functions of incident pulse energy for 1-percent transmitting films of bismuth (●), (550Å thick, 400Å grain size), indium (▽), (630Å thick, 150Å grain size), and tin (○), (650Å thick, 150Å grain size), sputtered onto polyethylene teraphalate substrate, and for bismuth evaporated onto Bi_2O_3 without breaking vacuum (□). The 30-ns-duration pulse was obtained from a cavity-dumped YAG laser; the focussed machining-beam diameter was 6.8 μm at the 1/e-intensity level. The curves were calculated from the simplified hole-opening models, eqs. (2) and (3) of the text, with Rayleigh velocities estimated from listed surface-tension values, and thermal-relaxation times scaled from the 300-ns value for 500Å-thick bismuth and the specific heats of the various metals. Comparison between these calculated curves and the machining data indicates that the temperatures of small-grain size, sputtered, indium and tin films must be raised to the metal boiling point to nucleate holes, but the temperature of large-grain-size, sputtered bismuth need be raised only up to the Bi_2O_3 melting point. Forces even higher than the vapor pressures obtainable at the boiling point are required to break the tenacious Bi/Bi_2O_3 bonds.

bient. If bismuth is evaporated onto this surface without breaking the vacuum, tenacious bonds surround the metal grains so tightly that the bismuth must be brought to its boiling point to rupture them. As shown in Fig. 21, it is difficult to machine large holes in such films until substantially more energy has been supplied than the boiling point energy.

5.2.3 Control of surface energy and thermal impedance at the metal-substrate interface

The polyethylene teraphalate substrate (Mylar or Celanar*) evolves impurities at room temperature that tend to oxidize the bismuth metal

* This polyester is a trade name of the Celanese Company.

deposited on it. This evolution eventually forces the machining into an oxide-dominated mode, even though the upper surface of the metal may be protected from atmospheric oxygen, and is believed to be the principal aging mechanism in dielectric-overcoated metal films (such as Se/Bi) deposited directly onto the substrate. It is possible to block the evolution of oxygen and to modify the surface energy (both favorably and unfavorably) by overcoating the substrate with selected, thin, transparent films prior to evaporating the metal. For example, poly-isobutyl methacrylate (IBM)* flourinated ethylene polymer (FEP), TiO_2 exposed to air, and cross-linked vinyltrimethylsilane (VTMS) have low surface energies and affect machining favorably; high-surface-energy polyamide (PA), polyvinylidine chloride (PC), and fresh Bi_2O_3 coatings on the substrate produce the opposite effect. A comparison of bismuth machining curves made with various undercoatings is shown in Fig. 22. All these coatings reduce the evolution rate of substrate impurities.

Substrate overcoatings can modify the thermal impedance for heat flow between the metal and the bulk substrate and can also modify the film-substrate interface bonding energy. Even though a free liquid surface may exist at the metal-air interface, the metal may be so strongly bound to the substrate as to delay nucleation or to reduce the hole opening velocity significantly, allowing appreciable resolidification to occur before the final hole diameter is reached. Thus, changes in bonding energy should primarily affect the hole diameters which can be attained at energies somewhat above threshold, whereas increases in thermal impedance—i.e., in the thermal relaxation time τ—should both lower the threshold energy and increase the final hole diameter. (However, for cases in which hole nucleation involves lifting the metal from the substrate, reductions in substrate surface energy can improve the ease with which a dome or blister can form, via the reduced adhesion.) Adhesion and scratch resistance must obviously be considered in any effort to improve machining by reducing the bonding energy. The TiO_2 (and also the VTMS) overcoating probably increased both the characteristic thermal decay time and the hole opening velocity. In correlation, metals deposited on these two overcoatings have extremely poor adhesion to the substrate, whence the improved machining is without practical value. An approximate factor-of-two reduction in machining threshold, relative to bismuth on Mylar, is gained by the use of an IBM overcoating of the substrate, as seen in Fig. 22. The apparent effect of IBM is to permit the nucleation of small holes in bismuth at machining energies close to the "melt-metal" limit

* The IBM overcoating is prepared by solution-dipping Mylar in a 5-percent concentration of IBM in methyl ethyl ketone.

(curve A of Fig. 19); larger holes, already restrained by this limit, are less affected. Adequacy of the adhesion of the weakest interface in a complicated layer combination such as Se/Bi/IBM/Mylar can be determined on a "pass-fail" basis by the "Scotch-Tape-test." Scratch resistance, obtained quantitatively by varying the load on a 25-μm radius sapphire stylus, is sensitive[48] both to adhesion and to the

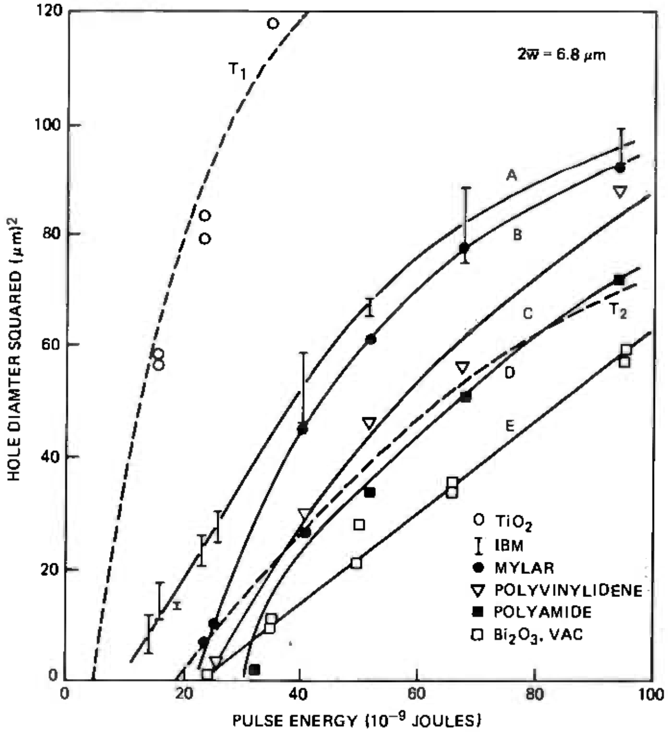


Fig. 22—Effect of substrate overcoatings on machining of bismuth films. Cavity-dumped YAG-laser beam of 30-ns duration, focused to 6.8-μm diameter at 1/e-intensity. ○—Back-machining data for 500Å-Bi evaporated onto 100Å-thick TiO_2 on glass; curve T_1 calculated for melt-bismuth, zero-heat-loss limit. Similar data have been measured for VTMS overcoatings on Mylar. I—Front-machining data for 600Å-Bi evaporated onto IBM on Celanar* (Celanar solution—dipped in 5-percent concentration of isobutylmethacrylate in MEK); smooth curve A connects data points. ●—No overcoating; 600Å-thick Bi on Mylar, front-machined; smooth curve B drawn through data points. ▽, ■—Front-machining data for 600Å-Bi evaporated onto polyvinylidene (▽) or polyamide (■) on Mylar; smooth curves *C* and *D* connect data points. □—500Å-Bi evaporated onto 3600Å-Bi_2O_3 on 7059 glass and front-machined without breaking vacuum. Curve T_2 calculated for bismuth-boiling-point, zero-heat-loss machining; smooth curve *E* connects data points.

* Trademark of Celanese Corporation.

coefficient of friction at the selenium surface. The weakest link in the above film combination is found to occur at the IBM/Mylar interface. Recent data indicate that this interface can be strengthened satisfactorily, and the machining efficiency at high energies improved slightly, if the IBM is copolymered with a vinyl acetate. Results of adhesion and scratch tests are summarized in Table I.

5.3 Machining limitations set by laminar heat losses and optical absorption

The machining data of Figs. 16, 19, and 21 indicate that the incident intensities in Gaussian profile beams must equal 50–60 mj/cm^2 (at 1.06-μm wavelength) to produce the 900°C temperatures needed to nucleate holes in 500Å-thick, surface-oxidized bismuth evaporated onto Mylar, and must approach 200 mj/cm^2 to provide the range of hole sizes needed for high-quality microimage machining. These requirements on the laser are representative of the most favorable which can be achieved with single-element metal films, as is shown in this section.

Present GaAs lasers cannot machine practical microimages in single-element metal films even with the assistance of low-surface-energy undercoatings (viz. Fig. 22 and Table I). The film modifications needed to bring microimage machining within reach of GaAs lasers are described in Section 5.4.

5.3.1 Thermal conduction effects

The effects of laminar heat conduction on hole machining have been studied for surface-oxidized bismuth on Mylar and for the Se/Bi/Mylar system, using cavity-dumped gas lasers whose focused beams can be adequately approximated by Gaussian intensity profiles. Machining curves (hole-diameter-squared versus beam-energy or peak-power) were measured with pulse duration T as a parameter. They were compared with a machining model in which the film temperature at time T and at a radius equal to (or at least correlated with) the final hole radius was assigned a physically definite, characteristic value, e.g., the bismuth or Bi_2Se_3 melting temperature. The film temperatures

Table I—Adhesion and scratch test summary

Film Structure	Scratch Resistance (grams)	Scotch Tape Test (pass/fail)	Machining Performance with GaAs Laser
Bi/Mylar	2	fail	unusable
Se/Bi/Mylar	100	pass	good
Se/Bi/IBM/Mylar	10–15	fail	excellent
Se/Bi/IBM/Mylar*	~ 100	pass	best

* IBM copolymered with vinyl acetate.

were calculated[2] from the time-dependent heat equation,[3] assuming zero thermal impedance at the metal/Mylar interface and optical absorption only in the metal. (Resolidification diameters in the free-standing films of Figs. 18 and 20 were predicted accurately by assuming that all heat delivered to the film is redistributed by radial thermal conduction and that no optical power is lost by premature hole opening.)

The hole sizes machined in 600Å-Se/600Å-Bi/Mylar film by a cavity-dumped He-Ne laser beam of 3.2 μm diameter at $1/e$ intensity and 25 ns to 5000 ns pulse duration are plotted in Figs. 23a and 23b as functions of total pulse energy and of peak pulse power. A progressive increase in the total energy required to machine a given hole size occurs as the pulse duration is lengthened. This failure of "reciprocity" was observed for all hole sizes, and is illustrated in Fig. 23c for holes of 16-$(\mu\text{m})^2$ area. Figure 23c shows that the machining energies required as functions of pulse duration T are compatible with calculations from the laminar-flow heat equation (normalized to the data at $T = 25$ ns). Earlier data[2] on the Bi/Mylar system also agree with the laminar flow calculation, similarly normalized at $T = 25$ ns. The overall energy efficiency, including heat loss during the pulse as well as the effects of refreezing, is illustrated in Fig. 23d for holes whose diameter approximately equals the laser beam $1/e^2$-intensity diameter. The machining efficiency of 100-ns-duration pulses is about 60 to 65 percent of the efficiency obtained with very short (25 ns) pulses.

The close agreement between experiment and theory in Fig. 23c has implications for hole-opening kinematics in the fairly thick films pertinent to these data. It appears that machining a hole of diameter d depends primarily upon attaining a physically definite temperature profile in unruptured regions of the film at the end of the laser pulse, no matter what the pulse duration. The data presented in Fig. 19 show that the diameter d of holes machined at energies somewhat above threshold in films of 500Å thickness and greater approaches the diameter to which the metal melts.

The time evolution of the molten diameter (liquidus) is estimated here by approximating the effects of film heat loss with a thermal relaxation time τ. The heat content at radius r and time t, per unit area of (unruptured) film, for a film whose absorptance at the laser wavelength is A and for a Gaussian profile beam of power P and pulse duration T whose $1/e$-intensity diameter is $2\bar{w}$, may be written

$$u = \begin{cases} (AP/\pi\bar{w}^2)e^{-(r/\bar{w})^2}\tau(1 - e^{-t/\tau}), & 0 \leq t \leq T \\ u_T e^{-(r/\bar{w})^2} e^{-(t-T)/\tau}, & t \geq T, \end{cases} \tag{1}$$

in which $u_T = A\tau(1 - e^{-T/\tau})(P/\pi\bar{w}^2)$ is the heat content per unit area at $r = 0$, $t = T$. The value $\tau = 300$ ns has been estimated for 500Å-thick

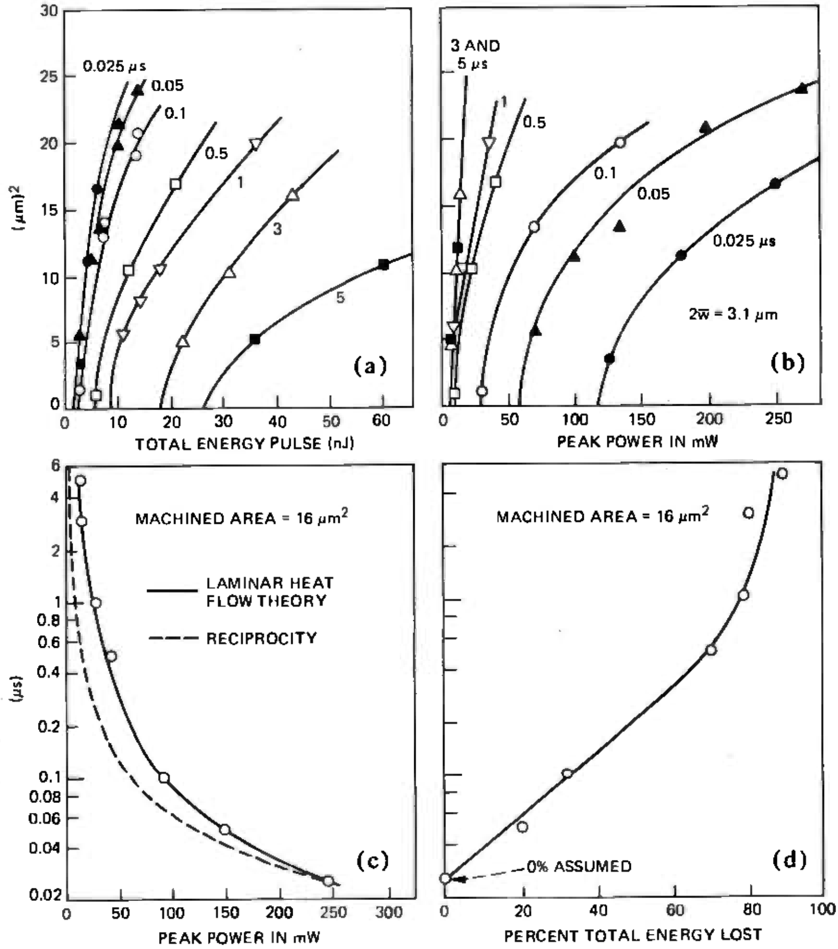


Fig. 23—The effect of pulse duration on hole machining by a Gaussian-profile laser beam. A cavity-dumped, He-Ne, laser beam at 0.63-μm wavelength incident onto Se/Bi film was used to obtain these data. A 1/e-intensity full width of $2\bar{w} = 3.1$ μm for the focused beam was estimated from the diameter at the entrance pupil of the 8-mm-focal-length writing lens. The smooth curves in (a), (b), and (d) have no theoretical significance. (a) Square of machined hole diameter versus total pulse energy (in units of 10^{-9} joules) for pulse durations from 0.025 to 5 μs. (b) The same data plotted versus peak pulse power. (c) The pulse duration versus peak pulse power required to machine a 16-μm² hole. The dashed curve is calculated assuming that reciprocity holds; the solid curve is obtained from the laminar heat-conduction model of Ref. 2. Both curves are normalized to the 0.025-μs data point. (d) Machining inefficiency of long-duration pulses, relative to the 0.025-μs pulse.

bismuth evaporated onto Mylar by measuring resistance changes in a 3-μm × 3-μm neck of film heated by a 25-ns-duration YAG laser pulse. (The beam diameter was about 8 μm at 1/*e*-intensity, and construction of the neck was such that only laminar heat flow to the substrate was important in the measurement.)

The position of the liquidus as a function of time is obtained by equating $u(r, t)$ to the enthalpy h_M needed to melt a unit area of the film starting from room temperature. The liquidus radius is given, in the approximation $T/\tau \ll 1$, by

$$(r/\bar{w}) = \begin{cases} 0, \quad t/T \leqslant h_M/u_T \\ [\log(u_T/h_M) + \log(t/T)]^{1/2}, \quad (h_M/u_T) \leqslant t/T \leqslant 1 \\ [\log(u_T/h_M) - (t-T)/\tau]^{1/2}, \quad t/T \geqslant 1. \end{cases} \tag{2}$$

(The inequalities $T/\tau < 0.1$ and $T/\tau < 0.3$ apply to practical micŕoimage machining by cavity-dumped lasers and by GaAs lasers respectively.) The film melts underneath beam center at the instant $t_M = (h_M/u_T)T$. The liquidus expands to a maximum radius $(r_c/\bar{w}) = [\log(u_T/h_M)]^{1/2}$ at $t = T$, and then begins to contract as resolidification propagates inward. This growth and contraction are illustrated in Fig. 24 for a metal and its oxide, assuming $\tau/T = 3$. The example is chosen to correspond to bismuth, for which the enthalpy to reach the Bi_2O_3 melting temperature is 1.8 times the enthalpy to melt bismuth.

Equation (2) does not include the radial redistribution of heat which occurs when hot (possibly superheated) melt is transported away from beam center as the hole opens, nor are the possible effects of viscosity considered. Indeed, a kinematical description of hole opening which incorporates these effects is forbiddingly complicated. Fortunately, as is apparent from Fig. 24, the details of hole nucleation and enlargement are irrelevant in determining the hole size if τ is large enough. If the film stays hot for a long time, the hole diameter will always approach the molten metal diameter (oxide structure permitting). Alternatively, τ need not be large if hole nucleation and enlargement are prompt. The hole-opening loci in Fig. 24 were drawn by supposing, for the sake of definiteness, that hole nucleation occurs in the neighborhood of beam center after a specified delay, and that the hole opens outward like an iris at a radial speed estimated by equating the kinetic energy of the displaced material to the released surface energy. (Composite films whose viscosity has been increased[7] by the incorporation of arsenic or selenium might be expected best to exemplify iris-like hole opening. However, data presented in Fig. 26 below suggest that the hole opening dynamics in selenium-containing films do not differ greatly from those in slightly oxidized bismuth films of the same thickness.) If the surface energy of the hole debris is neglected, this speed has the constant value (Rayleigh velocity)[40] $v_R = \sqrt{4\gamma/\sigma}$, expressed in terms of the surface tension γ (erg/cm^2) and film surface density σ (gm/cm^2), and leads to the expression

$$(r/\bar{w}) = (v_R/\bar{w})(t - t_N) \tag{3}$$

for the hypothetical hole rim loci in Fig. 24. The calculated Rayleigh

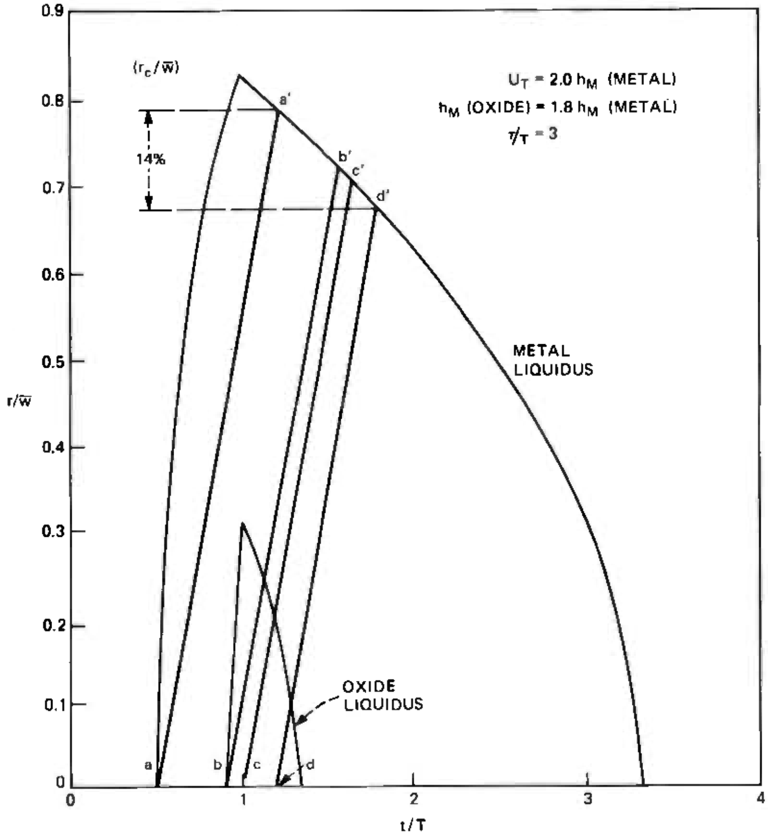


Fig. 24—Generalized kinematical hole-opening model. This figure applies to the case in which threshold is determined by melting a weak surface oxide and large holes are limited by the molten metal diameter. It is intended to show that the final hole diameter is comparatively insensitive to kinematical details if the thermal relaxation time is long and the hole enlargement speed high. The radius r of the metal or oxide liquid-solid boundary is given in units of $\bar{w}$, the laser beam $1/e$-intensity radius. The laser pulse begins at time $t = 0$ and ends at time $t = T$. The metal- and oxide-liquidus loci were obtained from eq. (2) for a case in which the laser pulse energy u_T equals twice the enthalpy $h_M(m)$ needed to melt the metal, and $\tau/T = 3$. The enthalpy $H_M(ox)$ needed to reach the oxide melting point was taken equal to 1.8 $H_M(m)$, corresponding to m = Bi, ox = Bi_2O_3. Solid lines are loci of hypothetical hole peripherys, with points a' through d' representing final hole sizes as determined by freezeback of the metal. The lines were obtained from eq. (3) for $(v_R T/\bar{w}) = 1.15$, corresponding to $\bar{w} = 3.4$ μm, $T = 61 \times 10^{-9}$ s, v_R (Rayleigh velocity) = 5.3×10^3 cm/s. Line aa' represents the limit of zero nucleation-delay time; i.e., hole opening initiated by melting the metal; line bb' depicts hole opening initiated by melting the oxide; line cc' corresponds to nucleation at the end of the laser pulse; and line dd' illustrates an opening delayed nearly until the onset of oxide freezeback. Note that this wide range of physical conditions produces only a 14-percent spread in hole diameters, and that the predicted diameters are 80 to 85 percent of the maximum molten-metal diameter.

velocities of 1-percent-visible-wavelength-transmitting bismuth, tin, and indium films range from 53 to 67 $\mu m/\mu s$, whence released surface energy is sufficient to enlarge the rim diameter to 6 μm within 50 ns of the hole nucleation instant. If nucleation occurs at the end of the laser pulse, this diameter will intersect the resolidification front (i.e., the final hole diameter will equal 6 μm) of an irradiated area melted to 7.1-μm diameter. This type of result is compatible with the data in Fig. 19 for holes machined well above threshold.

5.3.2 Limitations set by the optical absorption of single-element films

Since laser-machined images are intended for projection viewing by transmitted visible light, the visible-wavelength transmission of unmachined areas must be limited to some generally acceptable maximum to obtain good "black-white" contrast. The transmittance[49] of a specularly reflecting metal film depends on its complex refractive index, $N = n - jk$, evaluated for the wavelengths in question, on its thickness b, and on the incidence angle. When a transmittance value such as 1 percent is assigned for normally incident visible light, n and k become functionally related, with film thickness b as a parameter, as shown in Fig. 25a. (Projection light is incident over angles ranging up to 30 degrees, so the practical transmission of these films is less than 1 percent.) By assuming that this same functional relationship applies at the machining laser wavelength, the absorptance A of these films for laser radiation can be plotted versus n, as is done in Fig. 25b for $\lambda = 0.89$ μm.

The temperature to which any of the above films, whose visible-wavelength transmittance is 1 percent, can be heated by a given intensity of laser radiation depends on its *specific* absorption, A/b, at the laser wavelength and on appropriate film enthalpies evaluated on a per-unit-volume basis. The maximum possible value which can exist for the specific absorption of a 1-percent-transmitting film of given thickness,* as obtained from the curves of Fig. 25b, is plotted in Fig. 25c. The specific absorptions† and thicknesses of 1-percent-transmitting films of various low-melting-point metals have been calculated from bulk optical constants,[50] and it is seen that specularly reflecting films of bismuth, zinc, cadmium, and antimony approach the (A/b) limiting curve at thicknesses in the range 500 to 600Å. (Higher specific absorption limits could be realized with thinner films. But even if metals with the required n and k values existed, Fig. 16 indicates that they would probably be pinhole-rich and would not machine well

* The films must obey the further restriction $n \leq 10$, $k \leq 10$.

† The measured absorptance of 500Å-thick bismuth films is about 20 percent smaller than the absorptance calculated from bulk optical constants in the wavelength range 0.63 to 1.15 μm.

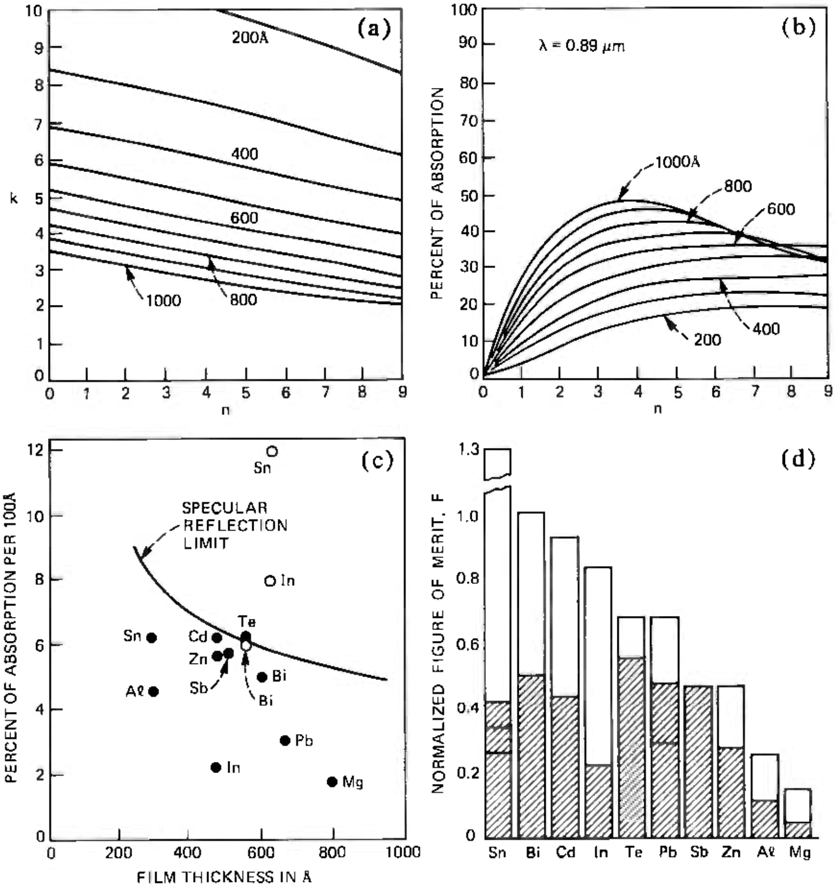


Fig. 25—Optical and thermodynamic limitations on the laser machining of high-contrast microimages in single-element metal films. The descriptions which follow refer to films whose optical transmission at 0.63-μm wavelength equals 1-percent. (a) Real (n) vs. imaginary (k) refractive indices at 0.89-μm wavelength for 1-percent transmitting films of different thickness. (b) The percentage of incident flux at 0.89-μm wavelength which is absorbed in 1-percent-transmitting films of different thickness, as a function of real refractive index, n. (c) The smooth curve is the maximum percentage absorption per 100Å of film thickness, as obtained from the locus of absorption maxima in (b), versus film thickness. The specific absorptions of 1-percent transmitting films of various single-element metals, as calculated from bulk values of their optical constants, are shown as solid dots. The measured specific absorptions of several sputtered films, whose reflection is believed *not* to be specular, are indicated by open circles. (d) The machining figure-of-merit F for single element metals is defined as the absorption per unit thickness divided by the enthalpy per unit volume needed to reach hypothetical hole-nucleation conditions, as follows; values of F for melting the metal (including heat of fusion) are shown by the open bars, normalized to the F-value for melting bismuth; values of F for reaching the oxide melting temperature (excluding the oxide heat of fusion) are shown by the shaded bars, with multiple values indicated for tin and lead, which have more than one possible oxide.

unless they happened to be non-oxidizing.) Figure 25c indicates that specularly reflecting, single-metal film systems are limited to efficiencies of about 35 percent in their ability to absorb light. The measured (A/b) values of *diffusely* reflecting, sputtered tin and indium, shown in Fig. 25c, exceed the specular-reflection limit, a possibility for films whose structure can trap radiation; however, erroneously large values of (A/b) could also have been caused by failure to detect all the reflected light.

Machining figures-of-merit, F, can be defined for actual metals as $F = A_b/(\Delta H_v \cdot b)$, in which A_b is the fraction of laser radiation absorbed in a film of thickness b, obtained either from direct measurement or from appropriate optical constants, and ΔH_v is the enthalpy per unit volume either to melt the metal or to raise its temperature to the oxide melting point. Figure 25d indicates the figures of merit calculated from bulk optical constants at 0.89-μm wavelength, for the various low-melting-point metals of Fig. 25c, normalized to the melt-bismuth F-value. The thickness of each film was chosen so the transmitted intensity was 1 percent at 0.63-μm wavelength, and it is seen that none appears strongly preferable to bismuth and its oxide with respect to the incident radiation needed to satisfy the enthalpy requirements for nucleating and enlarging holes. The nonspecular, sputtered films of indium and tin are found to have machining efficiencies similar to, but not much better than, those of evaporated bismuth (viz. Fig. 22). Since usually more than half the enthalpy needed to melt a metal (starting from room temperature) is associated with the heat of fusion, amorphous alloys, which require no latent heat to melt, may be energetically preferable to their crystalline counterparts even though their specific heats are higher. Electroless nickel[47,51] (an amorphous alloy of nickel and phosphorus), despite its high melting point, has exhibited comparable machining efficiency (and vastly superior scratch resistance and oxidation stability) to bismuth. The As-Te thin films for which excellent laser-machining characteristics have been reported[7] are also amorphous.

Unfortunately, the machining energy requirements for facsimile microimage application of evaporated and sputtered single metal systems, and of all the amorphous alloys investigated, lie a factor of 2 or 3 beyond the practical reach of present GaAs lasers. Figures 23d and 25c indicate that the incident energy for machining could be reduced about a factor of 4 from its value for bismuth (viz., Fig. 19) if the optical absorption were increased to 100 percent and the thermal conductivity to zero without, in the process, adding to the hole formation enthalpy. In fact, very long laminar-heat-loss relaxation times probably *have* been achieved with TiO_2 and vinyltrimethylsilane (VTMS) undercoatings (viz., Fig. 22), but in both cases film adhesion to the substrate was reduced well below acceptable limits. The most important *practical* machining energy improvements have resulted from the modifications in optical absorption described in Section 5.4.

5.4 Practical multilayer films: the bismuth-selenium system

Antireflective overcoatings can increase the optical absorption of a single-metal film and reduce its rate of oxidation by the atmosphere, but improvements in machining may be thwarted by the added mass which must be melted and displaced, and by possible complications in the hole-opening dynamics. The most obvious counter to this difficulty—substrate-incident machining with an antireflection coating sandwiched between substrate and metal—has proved to be only marginally practical. It has been attempted with bismuth, whose oxide Bi_2O_3 is transparent to visible light, and is a fairly good, near-infrared, antireflection match[52] between the metal and a substrate of 1.5 refractive-index value. Improved machining was obtained by exposing an evaporated Bi_2O_3 layer to air prior to evaporating bismuth, but such films did not age reproducibly.* The combination Bi/Se/Bi on Mylar can also be made antireflective for substrate-incident machining if the bismuth layer adjacent to the Mylar is very thin (75Å). Substrate-incident machining of such films was comparable, but not superior, to the air-incident machining of Se/Bi on Mylar near threshold, and was poorer well above threshold. Further, when substrate-incident machining was attempted with printers in which the film substrate was tensioned over a transparent film gate, position-dependent optical interference occurred between the film gate and the contacting surface of the substrate, producing unintended hole-size modulations similar to those caused by mode destabilization (Section 4.2.3) and preventing gray-scale rendition. The antireflection structures described[41,53] for use with optically recorded video disks are inapplicable when the machined image is viewed by transmitted light.

The largest step toward better machining efficiency has been made by recognizing that the mass added by overcoating can be used to advantage if it consists of an element or compound which mixes exothermically with the metal when both are molten. Low-melting-point dielectrics with high refractive indices comprise one group of overcoating possibilities. For example, a 660Å thickness of selenium[54] (n = 2.56 at 0.89-μm wavelength) deposited on 600Å-thick bismuth trebles the optical absorption at 0.89-μm wavelength from 30 to about 90 percent. But since 2.3 times as much energy is now needed to melt both constituents, the resulting film figure-of-merit (Fig. 25d), consid-

* Sequential evaporation of Bi_2O_3 and bismuth without breaking vacuum produced very stable films but extremely high machining energies, as shown in Figs. 21 and 22.

ering just these factors, would improve only about 29 percent. However, a large exothermic reaction occurs[8,55] when the molten bismuth and selenium mix, as would be expected for two components which tend to form strongly bound intermediate phases. The released energy maximizes at 5.4 Kcal/mole corresponding to the composition Bi_2Se_3. The maximum temperature which the film system can achieve is self-limited by vaporization of selenium. The machining threshold figure-of-merit calculated for this system at stiochiometry is about five times that of a single-layer bismuth film, assuming that there is no selenium vaporization.

A measured comparison between the machining curves of 1-percent-visible-light transmitting bismuth and Se/Bi film is shown in Fig. 26; the energy in 30-ns-duration YAG-laser pulses required to reach machining threshold in Se/Bi has been decreased about a factor of 3.5 from that at the 1.06-μm wavelength for which the optical absorption of the Se/Bi combination was maximized. Uncertainty in the quantity of selenium vaporized and relatively minor departures from exact stoichiometry are sufficient to account for observed differences between the expected and the measured machining performance of the Se/Bi system. The photographic montage of Fig. 26 illustrates a debris pattern consisting of balled-up droplets (presumably Bi_2Se_3) inside and outside the hole, and of beads strung around the peripheral rim. More debris appears inside the holes than was found for the thick bismuth films of Fig. 16, a result suggesting that hole opening in this relatively thick film is not dominated by the viscosity of the molten material. (But compare Ref. 7.) The increase in debris is not enough to detract seriously from microimage quality in transmitted light.

This rather dramatic result prompted the investigation of many two- and three-layer film systems. The machining performance of the most interesting of these is shown in Fig. 27; data on single-layer bismuth, with and without an IBM overcoating, are included from Figs. 22 and 26 for comparison. These curves represent holes produced by YAG-laser pulses of 30-ns duration and variable peak power. The various multilayer film compositions were adjusted to maximize the optical absorption of 1-percent-transmitting film at this wavelength. Although several of these compositions machine better than Se/Bi on polyethylene teraphalate, the Se/Bi system remains the most attractive since both of its components can be evaporated sequentially without breaking vacuum, thereby producing a film of controlled composition with the bismuth well-protected from the atmosphere. In contrast, indium layers of multiconstituent films are difficult to deposit, and the stoichiometry of Sb_2S_3 is difficult to assure.

The fractional improvement in machining brought about by an IBM undercoating of the Se/Bi/Mylar system is essentially the same as its

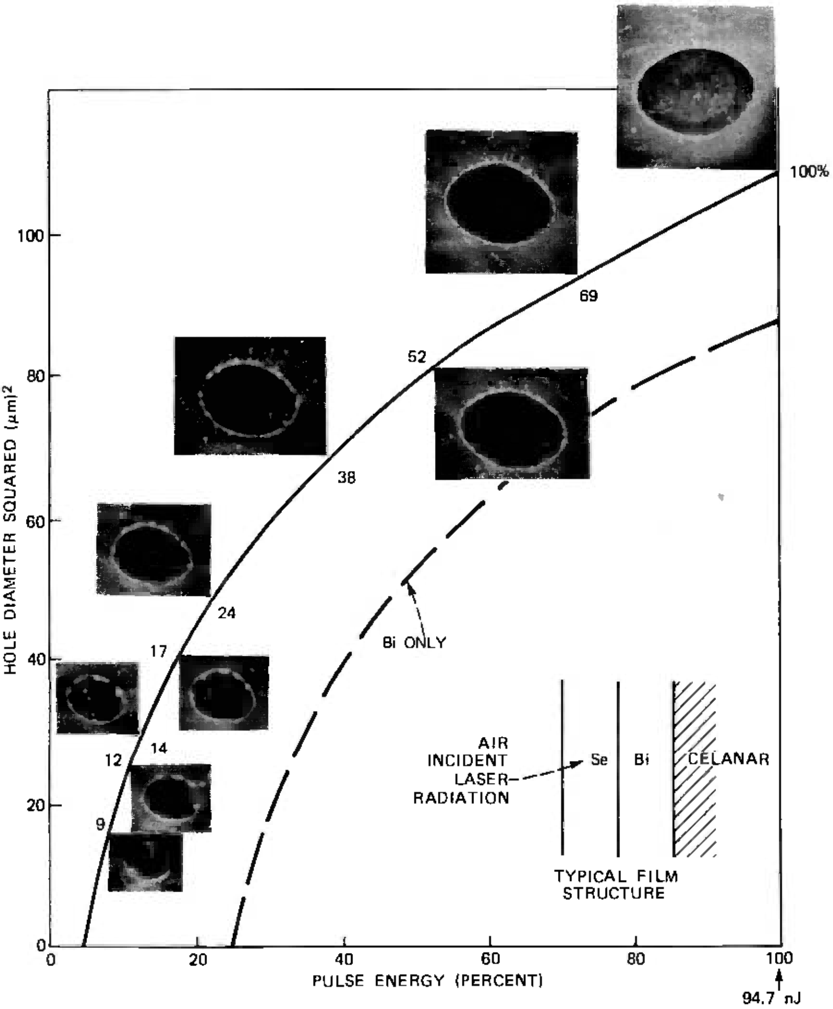


Fig. 26—Size and appearance of holes machined in Se/Bi film as function of laser pulse energy; comparison with bismuth machining curve. The montage of SEM photographs shows the appearance of holes machined in the air/860Å Se/750Å-Bi/substrate film system by the beam from a cavity-dumped YAG laser. The machining pulse duration was 30 ns, and the focused beam diameter was 6.8 μm at 1/e-intensity. The incident pulse energy is given as a percentage of 94.7 × 10^{-9} joules. The Se/Bi film was less than 0.3-percent transmitting at 0.63-μm wavelength, and the 500Å-thick bismuth film was about 1.5-percent transmitting. The Se/Bi machining threshold power is 7.5 nJ, and 6-μm-diameter holes are machined at 16 nJ.

fractional improvement of the Bi/Mylar system (viz. Fig. 22). The net effect of this undercoating on Se/Bi machining is to shift the Se/Bi curve of Fig. 27 down to the neighborhood of the Se/Bi/In curve. Thus, compared to a single layer of bismuth on Mylar (curve *k*), the

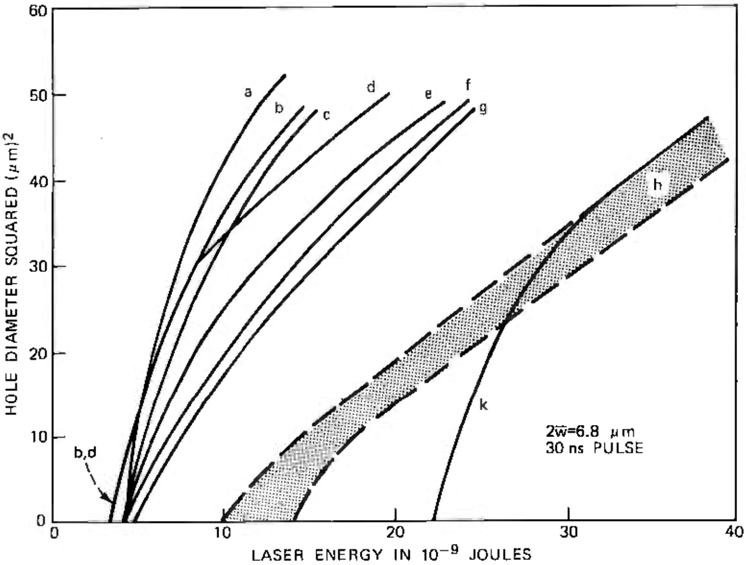


Fig. 27—Machining curves of various two- and three-layer, antireflective film systems; comparison with bismuth. The substrate is 100-μm-thick polyethylene teraphalate (Mylar or Celanar), and the 6.8-μm, 1/e-intensity-diameter, cavity-dumped, YAG-laser beam is incident from air. (a) Air/Se/Bi/IBM/substrate. (b) Air/Se/Bi/In/substrate. (c) Air/Se/In/substrate. (d) Air/Sb_2S_3/Bi/substrate. (e) Air/Se/Bi/substrate. (f) Air/Se/In/Bi/substrate. (g) Air/Bi_2S_3/Bi/substrate. (h) Air/Bi/IBM/substrate. (k) Air/Bi/substrate. The stippled band (*h*) includes data from many film samples.

machining threshold energy of Se/Bi/IBM/Mylar has been lowered a factor of 6, and the energy needed to machine 6-μm-diameter holes has been lowered a factor of 4. This improvement places the Se/Bi/IBM system well within the pulse-power capabilities of the GaAs laser described in Sections III and IV.

5.5 Film chemical stability; shelf life; archival life

The life of unmachined film ends when either its visible-wavelength opacity or its machining efficiency decrease below assigned limits. Optical absorptance declines monotonically with time in low-melting-temperature, single-metal films because opaque metal is gradually converted to optically transparent metal oxide. This increased transparency reduces the image contrast achievable with the film. It also reduces the heat generated by the incident beam, adding to the difficulty of melting the oxide whose stiffness impairs hole nucleation and enlargement. In the earliest stages of aging (identified with planar oxidation of surface crystallites), a gradual increase in film electrical resistivity accompanies the increased optical transmission, and near-threshold machining exhibits the transition from the "melt-metal" to the "melt-oxide" mode (data trend in Fig. 19). Gradual, monotonic increases in optical transmission continue as the single-metal film ages. Eventually, however, a penetration of oxygen into the film is produced by creepage of oxide tentacles between grain boundaries and by evolution of oxidizing impurities from the substrate. This interrupts the continuity of the metal, causing an abrupt resistivity increase and forcing the machining from the "melt-oxide" into the "boiling-point" mode (curves B and C of Fig. 19).

Changes in the optical transmission at 0.63-μm wavelength of 20 single-metal and multi-constituent film systems have been studied in aging runs conducted at temperatures between 40° and 125°C. Some of these elevated-temperature runs have lasted up to 4000 hours. Both monotonic and nonmonotonic changes in optical transmission with aging time have been observed. End of life in the monotonic cases has been defined as the number of hours required for transmission to increase by a factor of 1.5 from its initial value of 1 percent. Plots of this longevity versus the reciprocal of aging temperature are given in Fig. 28 for bismuth and for eight multiconstituent film systems. In a few cases, the data define straight lines, inviting extrapolation to suggest a value of room-temperature shelf life. No theoretical justification exists for this procedure, since the aging chemistry is generally unknown. Aging results exist for several other systems: thus, sputtered bismuth and indium age about a factor of 10 less rapidly than evaporated bismuth, sputtered tin ages much more rapidly, and aging effects have been undetectable in nickel and Ni_xP_y films.

The optical transmission of multilayers such as Se/Bi whose constituents can react chemically with one another is not monotonic with aging time and does not correspond monotonically to machining efficiency. Chemical reactions at the layer interfaces can increase opacity even while the machining efficiency is decreasing, and a number of reactions may be possible. The following reactions are capable of affecting the shelf life of unmachined Se/Bi/IBM films:

(*i*) Transition[56] of as-evaporated amorphous selenium to crystalline selenium.

(*ii*) Formation of an interfacial layer of Bi_2Se_3 (and/or BiSe).

(*iii*) Glass transition of poly-IBM (at about 70°C).

(*iv*) Formation of Bi_2O_3 by impurities evolved from substrate.

Shelf lives of the Se/Bi and Se/Bi/IBM systems have been studied by measuring changes in machining efficiency and optical transmission. Inclusion of the IBM coating appears not to change film aging significantly. Two different formats for presenting aging data on the Se/Bi/IBM system are used in Figs. 29 and 30. (The high machining threshold of Se/Bi/IBM in Fig. 30 is an artifact caused by the aging conditions: the layer thicknesses were optimized for $\lambda 0.89$ μm but were machined at $\lambda 1.06$ μm during the aging runs.) Quite as expected, the percentage

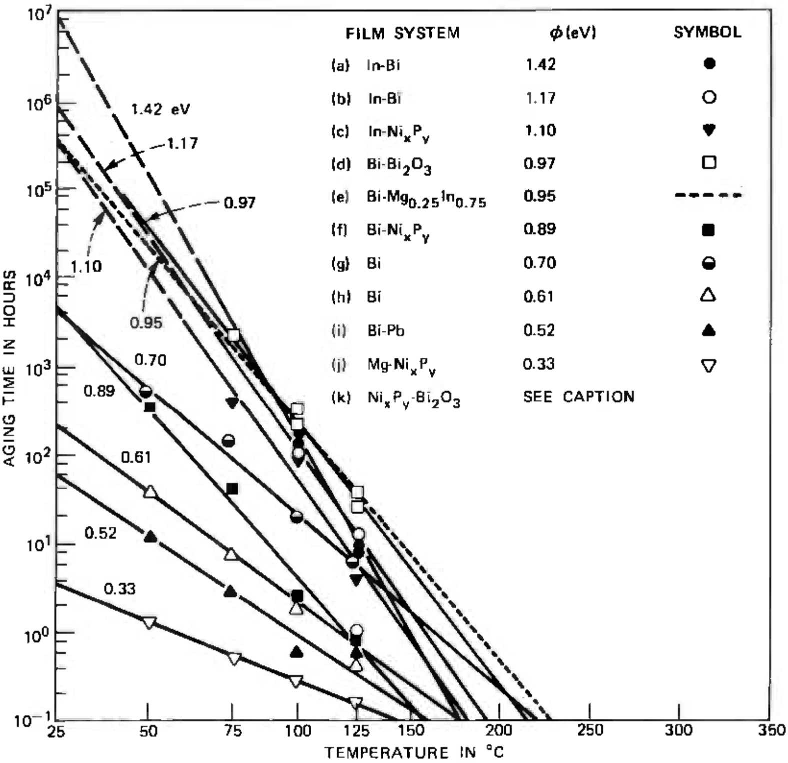


Fig. 28—Longevity of nonreactive, multiconstituent film systems at elevated temperatures. The ordinate is the number of hours required, at the chosen elevated temperature (abscissa), for the film optical transmission at 0.63-μm wavelength to increase by 50-percent from its initial value. Description of the film systems are: (a) In-Bi/Celanar: metals co-evaporated from separate sources to 500Å total thickness; machines similarly to Bi. (b) In-Bi/Celanar: alternate layers evaporated to a total of 5 Bi layers, 4 In layers, 500Å total thickness; requires higher machining power than (a). (c) Air/In/Ni_xP_y/Celanar: 10-percent-transmitting, evaporated Ni_xP_y, overcoated by 1.7-percent-transmitting, evaporated In; intended for substrate-incident machining where it performs better than (a). (d) Air/Bi/Bi_2O_3/Celanar: 500Å-Bi, evaporated onto 3600Å-Bi_2O_3; requires more than boiling point enthalpy to machine (similar to Fig. 22, curve *E*). (e) Air/Bi/$Mg_{0.25}In_{0.75}$/Celanar: $Mg_{0.25}In_{0.75}$ alloy sputtered to 86-percent transmission, then Bi sputtered to 1-percent transmission; intended for substrate-incident machining where it performs comparably to Bi/IBM (Fig. 22, Curve *A*). (f) Air/Bi/Ni_xP_y/Celanar: 10-percent-transmitting Ni_xP_y, 1-percent-transmitting Bi, constructed similarly to (c); machines similarly to (a). (g) Air/Bi/Celanar: 500Å-thick, fast-evaporated Bi (data from one evaporation run). (h) Air/Bi/Celanar: same as (g), but average of three evaporation runs. (i) Bi-Pb/Celanar: prepared similarly to (a); machines better than (a), but ages very rapidly. (j) Air/Mg/Ni_xP_y/Celanar: prepared similarly to (c) and (f); machines better than (a), but ages very rapidly. (k) Air/Ni_xP_y/Bi_2O_3/Celanar: machines like Bi and does not age; the oxide Bi_2O_3 is 3400Å thick (3 λ/4), and Ni_xP_y is 1.4-percent transmitting. Estimates of activation energy ϕ, in eV, are listed.

change in hole area caused by film aging is large near machining threshold, but decreases monotonically as the hole area increases. Specifically, Fig. 30 shows that the Se/Bi/IBM system machines along a single semilogarithmic curve and that 7000 hours of aging at 25°C

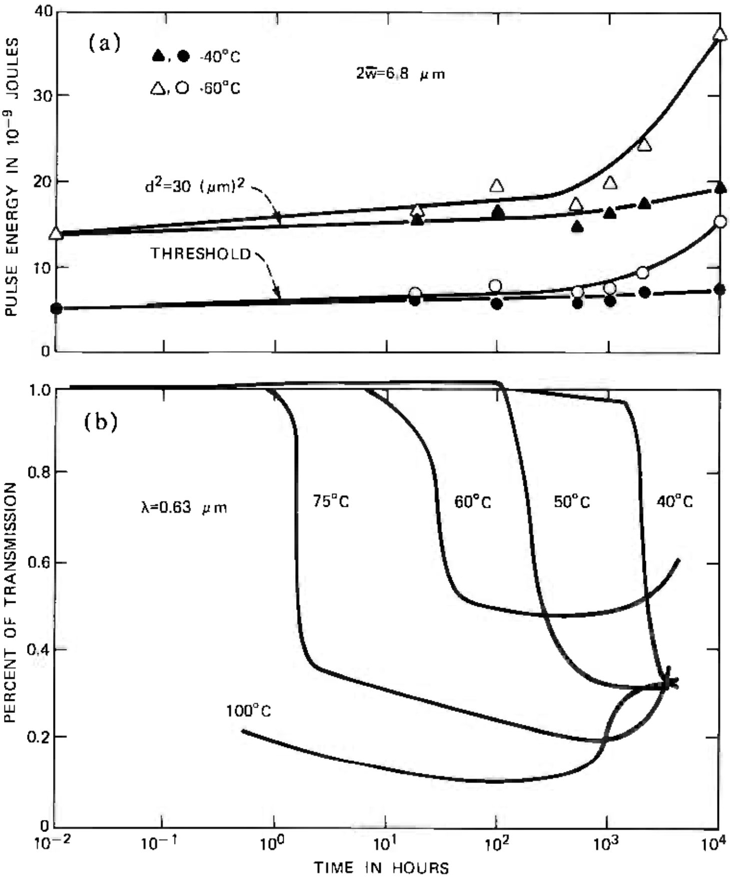


Fig. 29—Longevity of Se/Bi/IBM film system at elevated temperatures. (a) The pulse energies (ordinate) required to reach machining threshold and to machine 5.5-μm-diameter holes were measured at the indicated times (abscissa) on film samples maintained at 40° and 60°C. The film system is air/Se/Bi/IBM/Mylar. (b) Optical transmissions at 0.68-μm wavelength were measured as functions of time for the above film samples at temperatures up to 100°C. The transmission changes are nonmonotonic and are not closely correlated with changes in machining behavior.

causes a 30-percent increase in the energies required to machine large holes. Film shelf life depends, in a practical situation, on the gray-scale requirements of the printer usage and also on the effort expended to automate the electronic control of gray scale. As is shown in Section VIII, effective room temperature shelf life exceeding 10 years can be predicted if gray-scale capability is restricted or forfeited.

Once an image has been machined in Se/Bi film, it is protected and stabilized by the formation of Bi_2Se_3. Images of the IEEE Test Chart (viz. Section VIII) subjected to aging at 125°C for 2000 hours show only slightly degraded readability and gray-scale quality. Close exam-

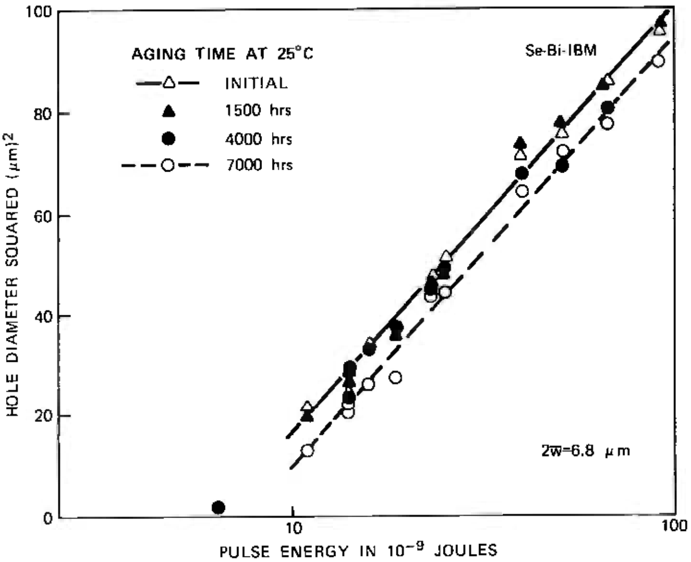


Fig. 30—Changes in machining curves of the Se/Bi/IBM film system with age at room temperature. The apparently high machining threshold for this film occurred because the Se and Bi thicknesses were optimized for use at $\lambda = 0.89\ \mu$m, but the machining data were acquired at $\lambda = 1.06\ \mu$m. The rapid dependence on aging time of hole size near threshold indicates that this film has less than one year shelf life for reproducible machining of gray scale. However, it is quite suitable for high-contrast use for much longer than one year (see Fig. 47).

ination reveals that the degradation has not occurred in the metal image, but instead that crystallites have grown on the Mylar substrate. The several millenia of room-temperature archival life for gray scale in Se/Bi images, suggested by these tests, is not unexpected. Whereas a single-metal film decreases in optical opacity as it ages so that images machined therein eventually lose all contrast, the Se/Bi bilayer increases in optical opacity (Fig. 29b). Hence the contrast between machined and unmachined areas of Se/Bi improves with age, and the gray scale, which is controlled by the fractional machined area, is preserved.

VI. GENERAL OPTICAL CONSIDERATIONS; MACHINING PERFORMANCE WITH GaAs

Generally speaking, the beam-acquisition, guidance, and machining optics in a metal-film-machining printer must be corrected to ~λ/4-net-optical-path difference from laser to film and must include cylindrical correction of beam astigmatism and ellipticity in printers employing GaAs lasers. Optics whose spherical aberration is equivalent to that of good commercial microscope objectives appear to be adequate, whereas medium-to-large-aperture photographic lenses and projection lenses, even of the best commercial quality, will not deliver a properly focused beam onto the film; they produce poorly shaped holes with low machining efficiency and are only marginally usable even with the high powers provided by cavity-dumped YAG and ArII lasers. The optics may contain more than two dozen air-glass interfaces which, because of the limited GaAs laser power, must have high overall transmission. Typical overall optical throughputs of 75 to 80 percent have been obtained with the aid of single-layer, dielectric, antireflection coatings.

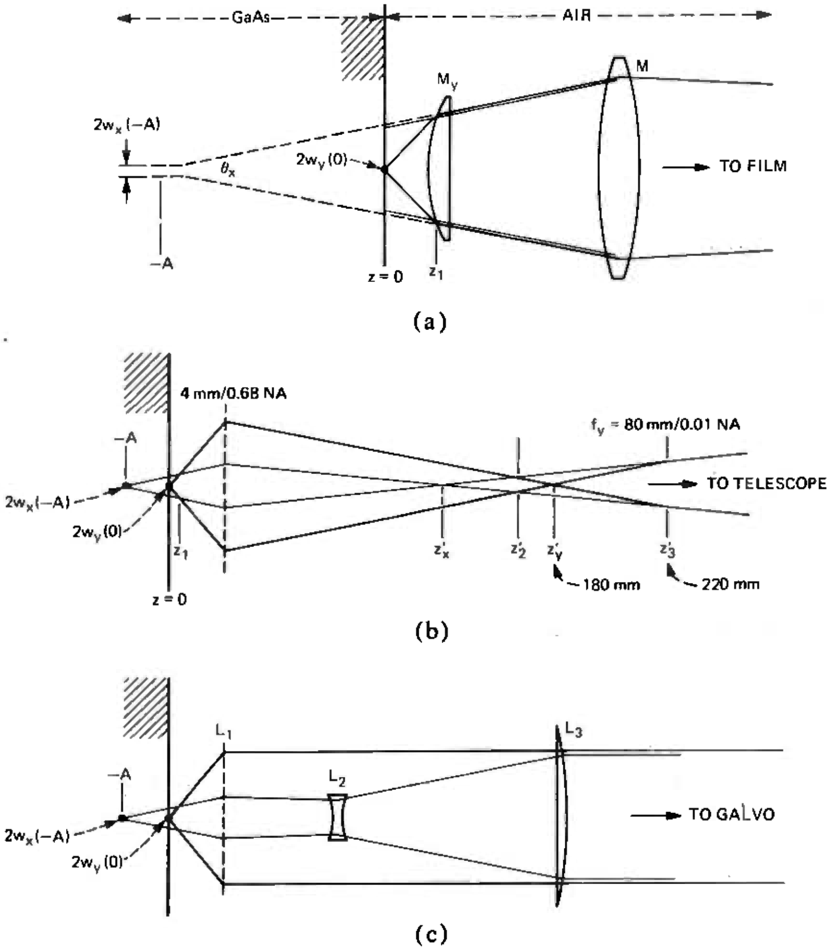
6.1 Beam-acquisition methods

Referring to Fig. 31, consider the astigmatic GaAs laser as an idealized optical source consisting of two displaced orthogonal, cylindrical beam waists with a common principal ray propagating along the $+z$-axis. The waist located at $z = 0$ has a small width, $2w_y(0) \approx 1.5\ \mu\text{m}$, from which rays diverge in the $y - z$ plane; the waist located at $z = -A$ has a larger width, $2w_x(-A) \approx 3\ \mu\text{m}$, from which rays diverge in the $x - z$ plane.

Although the far-field radiation pattern from these waists is sometimes treated according to the propagation law for cylindrical Gaussian beams,[30] actually neither waist is Gaussian. More realistic physical approximations to the cavity mode spatial dependence have been used to calculate[29] far-field, $y - z$ plane, angular distributions from the Kirchoff-Huygens formalism which agree fairly well with measurement. But the complicated secondary lobe patterns of mode profiles measured parallel to the junction plane (viz. Fig. 11) suggest that a similar estimate for the $x - z$ plane will not be easy. A convincing theory of the junction-plane fields is not presently available.

6.1.1 Astigmatism correction at first-crossover point

The beams diverging from the two sources have a common diameter $2w(z_1)$ in a plane $z = z_1$ just beyond the laser mirror ($z_1 \approx 6$ to $8\ \mu\text{m}$), as seen in Fig. 31(a). This plane is in the far field of either source. Hence a cylindrical thin lens of focal length $f_y(z_1) = (1 + z_1/A)z_1$ located at z_1 will form a virtual image of the $z = 0$ source at $z = -A$ with magnification $M_y = 1 + A/z_1$. Since z_1 is a plane of beam circularity, this magnification will produce the equality $2w_y(0)\cdot M_y = 2w_x(-A)$. A system of spherical optics placed beyond plane z_1 and designed to image point $z = -A$ onto the film with magnification M will produce a circular machining waist of diameter $2w_F = (2w_x)\cdot M$. (Thus, a machining waist of 7.8-μm diameter at $1/e^2$-intensity represents an $M = 2.6$ magnification of the junction plane virtual waist $2w_x(-A) = 3\ \mu\text{m}$).



(a)

(b)

(c)

Fig. 31—Methods of acquiring GaAs laser beam. (a) Microoptical system: Intensity profile is circular in plane z_1, close to laser mirror; cylindrical lens M_y of appropriate focal length causes rays in $y - z$ plane to appear to diverge from point z = −A, whence spherical lens M can be used to form nonastigmatic, circular beam waist at film. (b) Microscope objective: Intensity profile is circular in plane z_3', conjugate to plane z_1, and plane z_x' is conjugate to plane $z = -A$. Appropriate cylindrical lens at z_3' causes rays in $y - z$ plane to appear to diverge from z_x'; subsequent spherical optics form nonastigmatic, circular, beam waist at film. (c) Two-cylinder telescope: Lens L_1 is spherical; lenses L_2 and L_3 refract rays in the x–z plane.

Cylindrical *thick* lenses of ~20-μm diameter and refractive index $n = 1.65$, made by drawing glass fibers from a hemicylindrical boule, can be positioned to yield approximately the same astigmatism and ellipticity corrections as the cylindrical thin lens. The *angular*-spherical-aberration coefficients[57] of such lenses are extremely large for the rapidly diverging ray fan perpendicular to the junction. But the lens focal length is extremely small, and it turns out that the net lateral spherical aberration in the virtual image at $z = -A$ is a small fraction of the total image size. Rays emerge from the thick lens in an F/3 cone which, when introduced along the optical axis of a parabolic-graded-index glass-fiber lens,[58] can be focused into a ~7-μm-diameter spot on the film.

A printer utilizing these optics[12] has been constructed by aligning a GaAs laser and the two fiberoptic lenses just mentioned along a rigid, low-mass arm which was oscillated about the galvanometer axis. The laser emission acquired by the cylindrical lens was focused by the graded-index lens onto a curved section of film with ~75 μm of free working clearance. Microimages of 8-mm to 9-mm width were successfully back-machined in Se/Bi/Mylar film at ~260-kHz pulse repetition rate. The film radius of curvature was 19 mm.

6.1.2 Astigmatism correction at third-crossover point

A beam-acquisition system of smaller spherical aberration than that of the above micro-optical configuration can be obtained by using a large-NA microscope objective and a weak cylindrical lens, as shown in Fig. 31b. The objective is positioned to produce a real image of the waist $2w_y(0)$ at the design conjugate distance $z = z'_y = 180$ mm. The waist $2w_x(-A)$ will then be imaged at a distance in the range 120 mm $\leqslant z'_x \leqslant$ 140 mm if the value of A is in the range 27 μm $\leqslant A \leqslant$ 47 μm. Ray path calculations[59] for a modified, Bausch and Lomb, 4-mm focal length, 0.68-NA, microscope objective* show that optical path differences are smaller than $\lambda/4$ for conjugate points in the range 170 mm $\leqslant z'_y \leqslant$ 190 mm when the lens is used at full aperture and at 0.9-μm wavelength. When the lens is used at 0.2 NA, the path differences are smaller than $\lambda/3$ for conjugate distances in the range 120 to 140 mm. Hence the microscope objective contributes little spherical aberration to the waists at z'_y and z'_x.

Unfortunately, a numerical aperture of 0.68 is insufficient to accommodate the rapidly diverging ray fan in the laser $y - z$ plane, and intensity profiles scanned in this direction generally exhibit some Fresnel diffraction (viz. Fig. 12b). High-transmission objectives with numerical apertures larger than 0.68 usually have very small working distances. They cannot be used here because the typical 8- to 10-degree beam-steering angle creates a mechanical interference between the laser mounting stud and the lens.

Two planes, z'_2 and z'_3, exist in the image field of the microscope objective at which the orthogonal ray fans have common diameters.

* The elements in this objective were specially spaced to minimize spherical aberration at 0.885-μm wavelength.

The plane z'_3 is conjugate to the first-crossover plane z_1, and a cylindrical lens of focal length $f_y(z'_3) = (z'_3 - z'_y)(z'_3 - z'_x)/(z'_y - z'_x)$ placed there will superimpose a virtual image of the waist at z'_y onto the real image z'_x, producing a nonastigmatic, circular beam beyond the lens.[13] The numerical aperture of this lens can be ~0.01, in contrast to the very large numerical aperture required for the microcylinder in Section 6.1.1. Hence its aberrations are extremely small; also its focal-length value is rather uncritical. The beam emerging from this lens is enlarged to a suitable diameter (viz. Section 7.1) and collimated by an F/7 Galilean telescope prior to deflection by the galvanometer mirror.

Beam correction can also be effected in less axial distance by using a cylindrical lens of focal length $f_x(z'_2) = (z'_2 - z'_x)(z'_y - z'_2)/(z'_y - z'_x)$, located at the second-crossover position z'_2, which superimposes a real image of the waist at z'_x onto the waist at z'_y. However, this lens requires critical tolerances on its focal length and a significantly larger aperture than 0.01, sacrificing the advantages cited in the previous paragraph.

6.1.3 Two-cylinder beam corrections

The path length from laser to galvanometer mirror is $z_G = 320$ mm when beam correction at point z'_3 is followed by a Galilean collimating telescope. This length can be decreased by using two cylindrical lenses for beam correction. One such scheme, which shortens the laser-to-galvanometer distance to $z_G = 150$ mm, is outlined in Fig. 31c. The ray fan perpendicular to the junction plane ($y - z$ plane) is collimated to the desired final beam diameter by a specially constructed, large-working-distance, infinity-corrected, microscope objective. A cylindrical Galilean telescope is used to enlarge the ray fan in the $x - z$ plane and to collimate it to this same diameter.

For the sake of definiteness, suppose that the spherical elements form a beam waist in the $y - z$ plane at their exit pupil. Astigmatism will be eliminated if the Galilean telescope is adjusted to produce a waist in the $x - z$ plane at this same location. The lens of Fig. 31c is provided with mechanical adjustments for correcting astigmatism over a large range, $0 \leqslant A \leqslant 100$ μm. However, the cylinders must be large enough to accept the final beam diameter and must be carefully made to maintain aberrationless performance.

6.2 Machining focus and machining performance

Following its acquisition, the enlarged, collimated beam is deflected by a moving galvanometer mirror and focused onto the Se/Bi film by a scanning lens. This lens converts angular displacements of the collimated beam into a translation of the focused beam over a line in a flat focal plane. If the actual beam obeyed the propagation law for

Gaussian beams, the design of an optimal printer (i.e., one which machined a given size frame with minimum optical power) would be straightforward. With GaAs laser beams of the behavior illustrated by Figs. 9, 11, and 12a, deflection and machining apertures must be determined empirically, and even though the principal lobe of the GaAs beam often machines like a Gaussian profile beam, focusing of this lobe usually requires numerical apertures about double those needed for the Gaussian case.

6.2.1 Focusing and machining performance of gallium-arsenide laser beams

The numerical aperture requirements for machining with GaAs beams have been estimated by measuring focused beam profiles parallel and perpendicular to the laser junction plane with a mechanically scanned slit/photomultiplier arrangement for several focus conditions. The beam was enlarged and collimated to either 10-mm or 12-mm nominal diameter by an F/7 Galilean telescope and was then focused by lenses of either 37-mm or 50-mm focal length, nominally corrected to ~$\lambda/4$ optical path difference at 0.885-μm wavelength. The effect of departures from good collimation were estimated by comparing profiles obtained with the telescope first adjusted to produce a beam waist at 15 m and then readjusted to produce the waist at (nominal) infinity. Small increases in the secondary mode amplitudes were noted for the 15-m position, but there was no change in the main (machining) lobe between the two adjustments. The beam-acquisition scheme of Sec. 6.1.2 was used ahead of the telescope with the astigmatism correction checked both visually and by profile scanning. The data shown in Fig. 32 illustrate the beam profiles for a particular focus condition (10-mm nominal collimated-beam diameter, 37-mm-focal-length machining lens) at several different laser currents.

Comparison of Figs. 12 and 32 supports our description of the focused, astigmatism-corrected beam as a superposition of the laser real and virtual waists, $w_y(0)$ and $w_x(-A)$. However, the relative amplitudes of secondary modes in the focused beam are functions of the convergence F-number (a fact not illustrated by Fig. 32) and are also influenced by particulars of the focusing optics. Variations within the secondary structure as the current is changed also depend obscurely on details of the focusing arrangement. The large secondary-mode background in Fig. 32 precludes use of the $1/e^2$-intensity width as a meaningful predictor of machining performance. However, the 1/e-intensity width, $2\bar{w}$, usually lies above this background.

The values of $2\bar{w}$ measured parallel and perpendicular to the junction at 1.5-A pulse current, and also the geometric mean of these values, are plotted in Fig. 33 as functions of the focusing F-number for the four focusing conditions mentioned above. The 1/e-intensity full

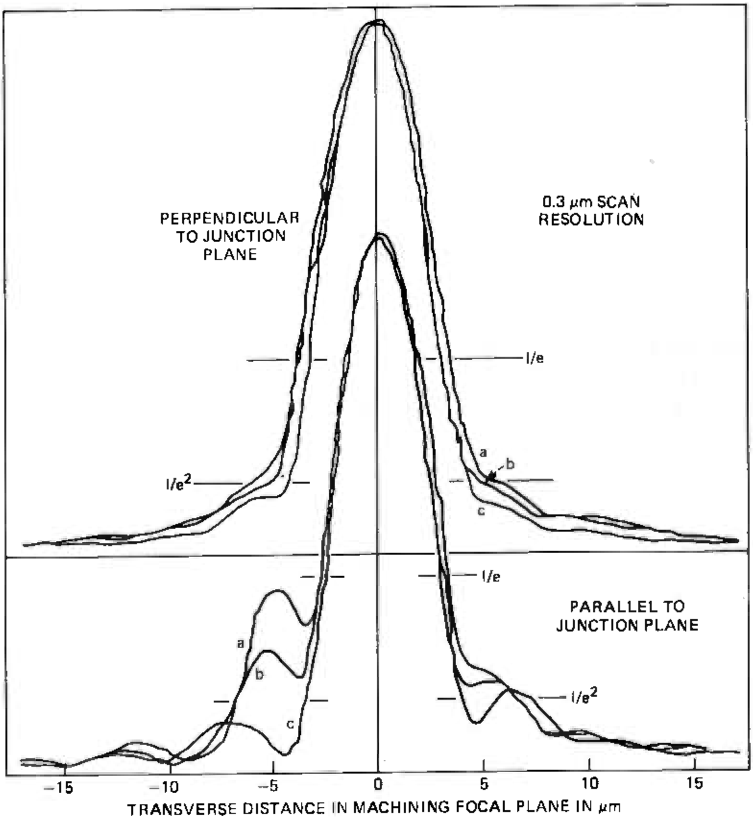


Fig. 32—Machining-beam profiles from GaAs laser. Profiles were measured in machining focal plane (i.e. plane of film) with mechanically scanned slit. Profiles a: 1.7-A laser pulse current, 100-ns duration. Profiles b: 1.5-A laser pulse current, 100-ns duration. Profiles c: 1.0-A laser pulse current, 100-ns duration. The machining beam was accurately corrected for astigmatism with an $f_y = 80$-mm cylindrical lens located at plane z_4' of Fig. 31b. Its residual ellipticity in the machining focal plane is 1.2:1 at the 1/e-intensity level. The beam was enlarged and collimated to 10-mm nominal diameter by an F/7 Galilean telescope and focused onto the film with a 37-mm-focal-length machining lens.

widths which would be produced by the same *F*-numbers with a Gaussian beam are also shown. We see that the numerical-aperture requirements for focusing the GaAs beams are approximately double those calculated from the Gaussian-beam expansion law.

A focusing arrangement which provides the required apertures over a large image field is illustrated in Fig. 34. The galvanometer mirror is positioned in the entrance pupil of a flat-field scanning lens. A telecentric lens design[60] (i.e., principal rays in the image space lie roughly parallel to the optic axis) was used to obtain enough spacing between the entrance pupil and the lens to keep the incident beam from

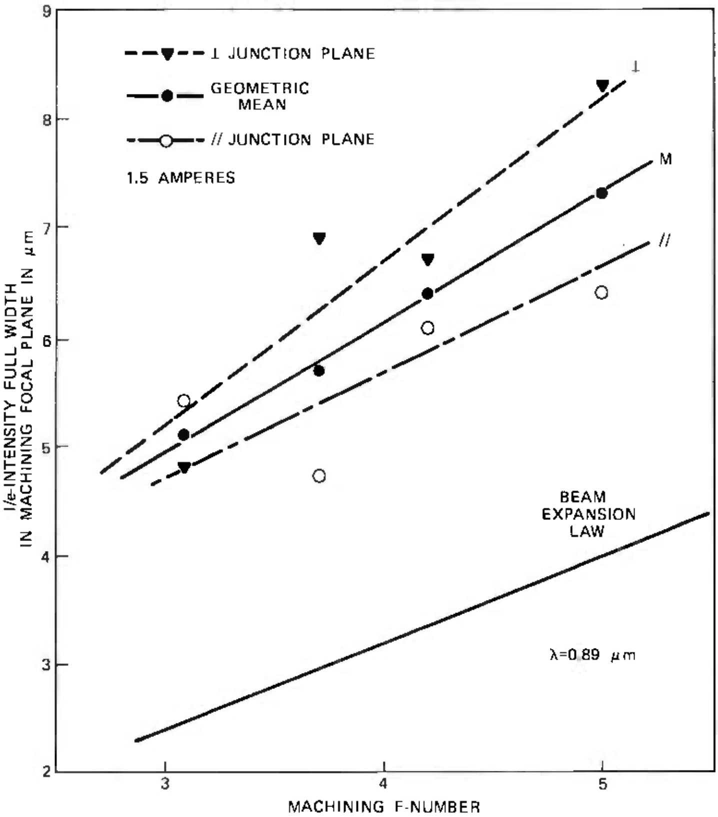


Fig. 33—Numerical-aperture requirements for focusing GaAs beam. The 1/e-intensity full widths of the machining-beam waist measured parallel and perpendicular to the junction plane are plotted versus the machining beam F-number (i.e., the ratio machining-lens-focal-length/collimated-beam-nominal-diameter) for 1.5-A laser current. Predictions of the Gaussian-beam expansion law at 0.89-μm wavelength are shown for comparison.

truncating on the lens housing. Tangent distortion was provided to produce images whose size scales linearly with galvanometer deflection angle. Several recent GaAs printers have employed a telecentric lens of focal length 36.7 mm, corrected to ~λ/4 optical path difference over a 14.7-mm-diameter focal plane and over a chromatic bandwidth of ±100Å centered at 0.887-μm wavelength. The entrance pupil diameter is 12.6 mm, spaced 18 mm from the lens. Figure 34, which is scaled to the above dimensions, suggests that a mirror-bias-angle $\psi = 40°$ should provide comfortable clearance between the lens body and the incoming collimated beam. The 14.7-mm field permits frames up to 9 × 11.6 mm to be optically raster-scanned. A lens designed to accommodate only the 9-mm-wide line scan could be physically much smaller.

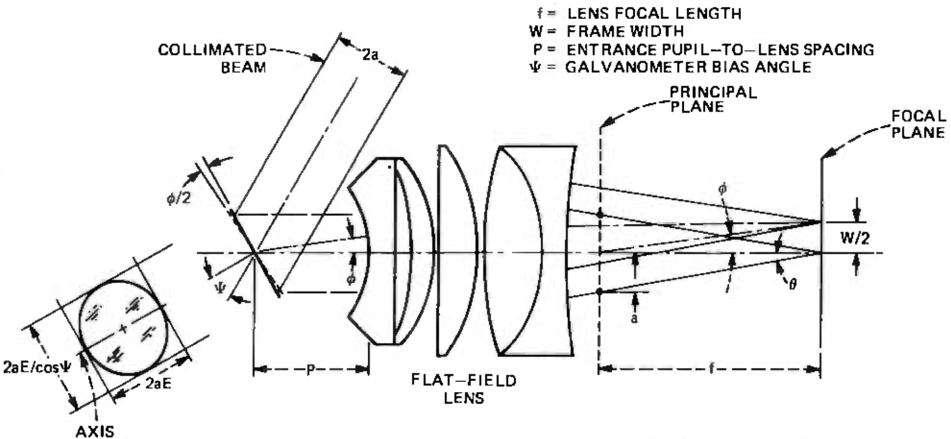


Fig. 34—Flat-field machining optics. This sketch is scaled to the dimensions mentioned in Section 6.2.1, with $P = 18$ mm and $f = 36.7$ mm. The machining F-number, $f/2a$, is limited to a minimum value of $(12.6/36.7)^{-1} = 2.9$. The image-field angular half width, $\phi = \tan^{-1}(W/2f)$, is limited to a maximum value of 11.3 degrees by the lens design. The lens-housing outside diameter equals 51 mm.

6.2.2 Gallium-arsenide machining curves

The scaling behavior for machining of unoxidized films with Gaussian profile beams of different diameters follows from eqs. (1) and (2): the power P_{th} required to reach machining threshold and the slope of the semi-logarithmic machining curve should both increase proportionally to the area $\pi\bar{w}^2$ of the focused beam. Figure 35 indicates that the machining produced by focusing the cavity-dumped beam from a He-Ne laser onto fresh Se/Bi film with various microscope objectives conforms at least qualitatively to these expectations. This figure compares the He-Ne data with machining curves measured for the GaAs laser and the four focusing conditions of Fig. 33. The GaAs machining results (with one exception) vary with width of the main machining lobe roughly as required by the above scaling laws; they exhibit higher machining thresholds than He-Ne curves of approximately the same slope, as expected from the diversion of power into secondary, nonmachining modes.

The comparison between GaAs and He-Ne laser machining results is continued in Fig. 36. The GaAs data were obtained with F/3 machining optics (viz. Figs. 32, 33) on 40 lasers selected from the same LPE-grown wafer. Figure 36 contains a sample of these data for nine lasers, and it repeats several He-Ne machining curves of Fig. 35. The nine GaAs curves would be coincident if the wafer produced identical lasers. The observed differences are attributed to differing proportions of power wasted in nonmachining secondary modes and to variations in width of the machining main lobe. The slopes of the GaAs curves fall roughly into two classes. Those curves having the smaller slope

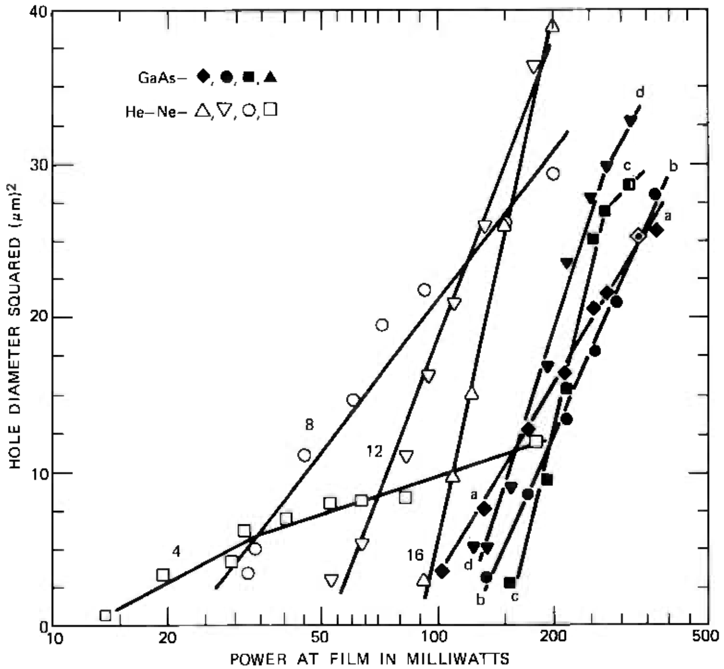


Fig. 35—Effect of focusing optics on machining; comparison of He-Ne and GaAs. Machined-hole diameter squared is plotted versus power incident onto film for the four GaAs-beam focus conditions of Fig. 33, and also for a cavity-dumped, 0.63-μm wavelength, He-Ne beam focused by various microscope objectives. The Se/Bi/Mylar film samples were separately optimized for machining by 0.89-μm- and 0.63-μm-wavelength beams. Solid symbols refer to machining of 600Å-Bi/600Å-Se film by 100-ns-duration GaAs pulses, with focused beam diameters $2\bar{w}$ at 1/e-intensity as follows: (◆, a)-5.1 μm; (●, b)-5.7 μm; (■, c)-7.3 μm; (▼, d)-6.4 μm. Open symbols refer to machining of 600Å-Bi/446Å-Se film by 100-ns-duration He-Ne pulses, with beam focused by microscope objectives of focal lengths as follows: (△, 16 mm), (▽, 12 mm), (○, 8 mm), (□, 4 mm).

value (shown dashed in Fig. 36) also have the lowest machining-threshold powers, whence the scaling laws suggest that these lasers have the smallest main-lobe width. Curves having the larger slope not only have larger threshold powers but also have a fairly large spread in thresholds; here, the scaling laws suggest a larger main-lobe width and varying fractions of power in nonmachining modes. Although some of the GaAs curves overlap He-Ne machining data, the machining thresholds for GaAs curves of given slope are never smaller than those for Ne-He curves of approximately the same slope.

A machining curve obtained by varying the pulse duration at a constant laser-pulse-current is shown in Fig. 37. (The data refer to the laser and the F/3 focusing conditions of Fig. 33.) The focused-spot profile is the same for each point on this curve, unlike the data of Fig.

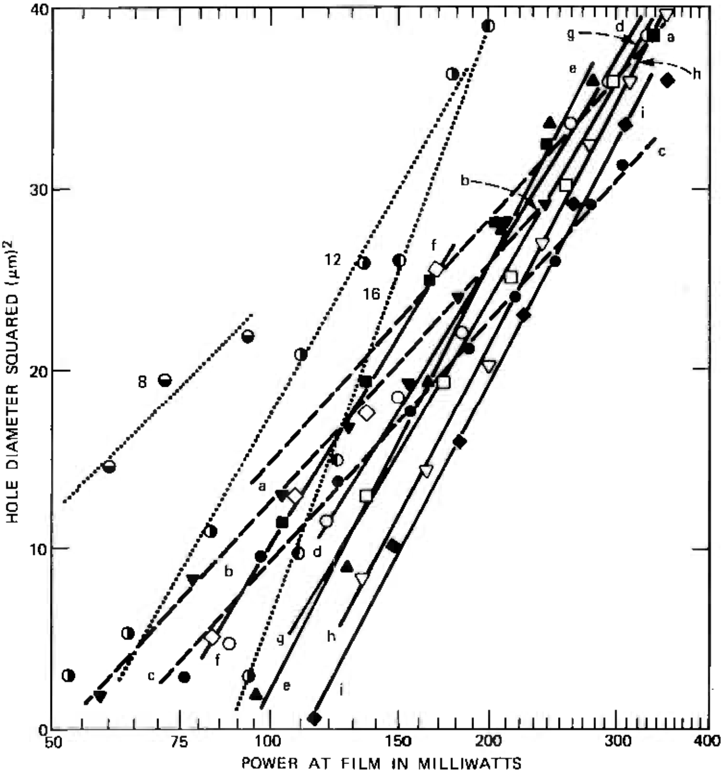


Fig. 36—Variation in machining performance of GaAs lasers; comparison with He-Ne machining. The F/3.0-focusing conditions of Fig. 33 were used to obtain machining data from a sample of nine GaAs lasers cleaved from the same wafer. Variations exist in the focused beam sizes and secondary-mode amplitudes produced by these lasers, and hence in the slopes and threshold power of their semi-logarithmically-plotted machining curves. (a, ■); (b, ▼); (c, ●)—GaAs lasers whose machining curves have small slopes and low thresholds, suggestive of a small focused-beam diameter. (d, ○); (e, ▲); (f, ◇); (g, □); (h, ▽); (i, ◆)—GaAs lasers whose machining curves have large slopes and high thresholds, suggestive of larger focused-beam diameter. ◐; ◑; ◒ He-Ne conditions of Fig. 35 for machining with 16-, 12-, and 8-mm-focal-length microscope objectives.

32 in which the secondary-mode amplitudes and the primary (machining)-mode width are both current dependent. The principal motives for considering pulse-width modulation in preference to current modulation were discussed in Section IV.

VII. PRINTER DESIGN CONSIDERATIONS

The design of a laser machining printer starts with knowledge of the machining curve and assigned values of frame width and frame writing time. These quantities determine the machining lens and the line-scan-galvanometer requirements. If suitable allowances are made for

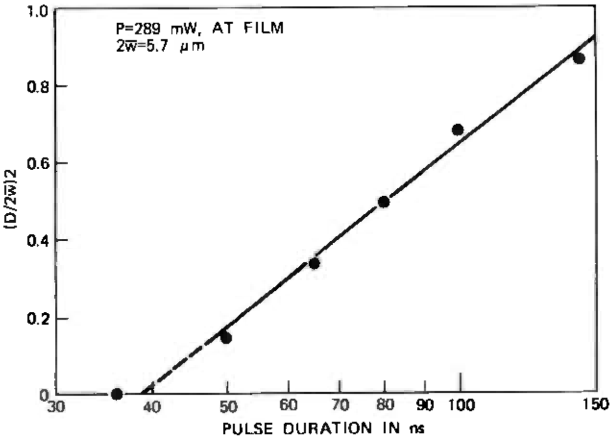


Fig. 37—Machined hole sizes versus GaAs-laser pulse duration. The laser peak-pulse current and power were kept constant, and the pulse duration was varied. The hole diameter D has been normalized to the beam diameter $2\overline{w}$ in plotting the data. Since the beam junction-plane displacements and steering-angle shifts stabilize within 30 ns of the pulse turn-on instant, all holes were machined with the same size and shape beam. The existence of secondary-mode amplitudes can still produce undesirable effects; thus, the slope of this maching curve is significantly smaller than that of a Gaussian-beam machining curve of the same threshold plotted in the normalized $(D/2\bar{w})^2$ units.

the larger apertures needed to focus GaAs laser beams, useful designs can be carried out in terms of a hypothetical Gaussian beam.

7.1 Scanning rate limitations of galvanometer mirrors

Referring to the geometry of Fig. 34, consider a frame of width W, with N contiguous machining hole sites per line, and a line-scan repetition rate ν. We assume that the beam profile is Gaussian, of diameter $2\bar{w}$ at 1/e-intensity and that the machined-hole area is defined by a single semilogarithmic curve of type A or B in Fig. 19. Then the optical power needed to machine the hole diameter W/N is a *minimum* if $2\bar{w} = W/N$, as follows from Section 5.2.1 and Ref. 2. Let ϕ be the peak-to-peak mechanical-deflection angle of the galvanometer, so that 2ϕ is the field angle subtended by W at the machining-lens principal plane. Let $2a$ be the $1/e^2$ intensity diameter of the beam entering the lens. The values of $\bar{w}$ and a are related by the expansion law for a Gaussian beam of wavelength λ, i.e.,

$$a \tan \phi = (\lambda/\pi\sqrt{2})(W/2\bar{w}) = (1/\sqrt{2}\pi)\lambda N. \tag{4}$$

Specifically, for $\lambda = 0.885$ μm and $N = 1600$, the line-scanning galvanometer requirements for "optimal" machining are defined by the condition $a \tan \phi = 0.32$ mm and by the scanning rate ν.

The field angle ϕ must be chosen such that the resulting mirror dimensions yield a moment of inertia, and hence a loaded resonant frequency ν_{0L}, which satisfies the fundamental scan-linearity requirement[9]

$$\nu_{0L} \geqslant 2.5\ \nu.$$

The mirror is an ellipse whose dimensions parallel and perpendicular to the galvanometer axis equal $2aE$ and $2aE/\cos\psi$, as labeled in Fig. 34. The numerical factor E ($1.2 \leq E \leq 1.3$) represents the ratio of aperture width to incident-beam $1/e^2$-intensity width which is needed to avoid serious reduction of focused beam intensity by Fresnel diffraction at the aperture.[61] It is unfortunate for the realization of fast scan rates that the mirror moment of inertia depends most strongly on the major axis dimension.

Design values $W = 8$ mm frame width, $N = 1600$ hole sites/line, $\phi/2 = \pm 3.5$ degrees mirror mechanical-deflection angle, and $\psi = 40$ to 45 degrees mirror bias angle have been used in several laser-machining printers. Mirror minor and major axes of 6.8 and 9.6 mm would be needed to achieve this design with a *Gaussian* beam at 0.885-μm wavelength, and the moment of inertia of a corresponding 1.5-mm thick glass ellipse would be $I_m = 0.0114$ gm cm^2. The General Scanning Co. G-108 galvanometer,* rated for ±4-degree mechanical deflection, has an intrinsic moment of inertia $I_u = 0.011$ gm cm^2 and an unloaded resonant frequency $\nu_{0u} = 1500$ Hz. If this galvanometer is used with the above mirror, the deflection-system loaded resonant frequency is $\nu_{0L} = [I_u/(I_m + I_u)]^{1/2}\nu_{0u} = 1050$ Hz, whence acceptable scan linearity would be expected for line repetition frequencies smaller than 420 Hz. Circular mirrors of 8-mm diameter were employed in early printers which machined images at 500-Hz line rates using the TEM$_{00}$ beam from a cavity-dumped He-Ne laser. Good scan linearity was very hard to achieve at this rate, presumably because the limit $\nu_{0L} \geqslant 2.5\nu$ was being pressed too closely, but the linearity at and below 400 Hz was satisfactory.

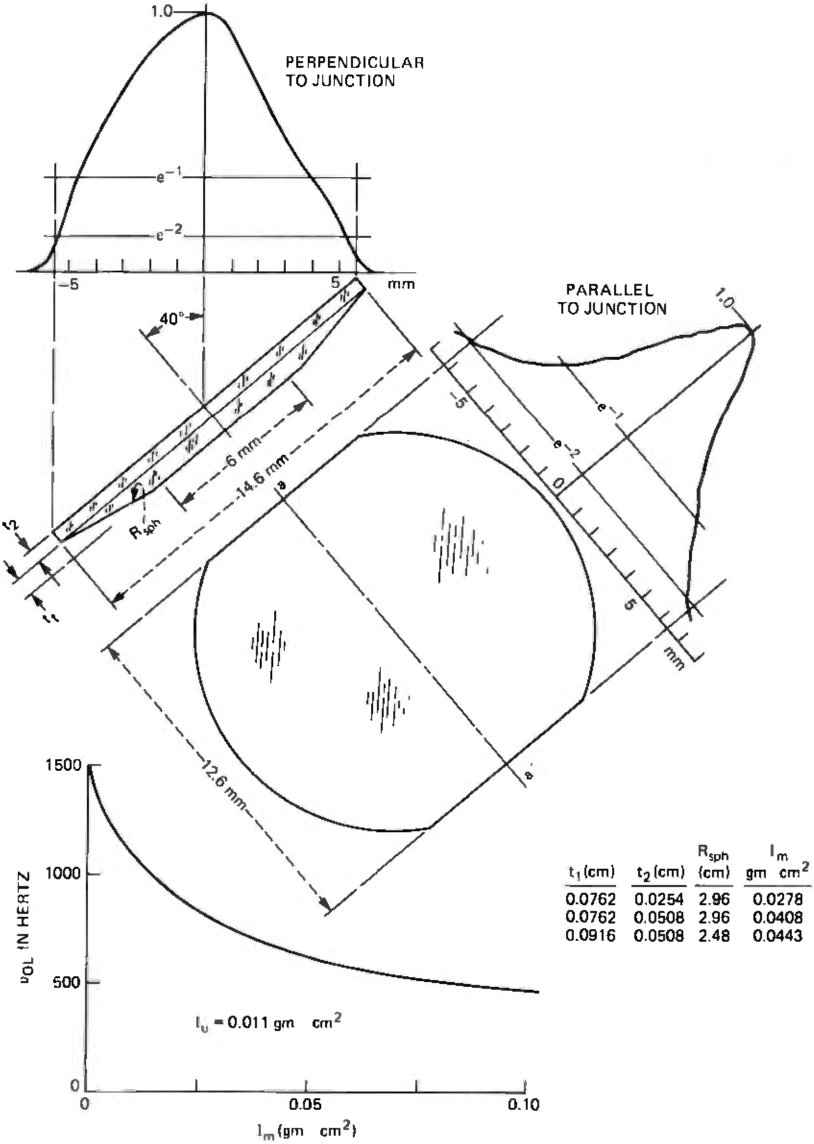
Profiles of the collimated GaAs beam used with the 37-mm-focal-length machining lens of Fig. 34 (the F/3.6 aperture of Fig. 33) have been measured with a mechanically scanned slit. These profiles are shown in Fig. 38 superimposed on a mirror which can deflect the beam at a bias angle $\psi = 40$ degrees with some allowance for diffraction losses, assuming that the laser junction plane is oriented parallel to the 12.6-mm mirror height. The deflection-system loaded resonant frequency, when a G-108 galvanometer is used with a 1.5-mm thick mirror, is measured to be 555 Hz; hence linear sawtooth scanning should be available up to ~220-Hz scan rate, corresponding to 9-s

* General Scanning Co., Watertown, Mass., 02172.

writing time for a 2000-line frame. Acceptable scan linearity has been observed with this mirror at 166-Hz line rates (12-s frames). The mirror has also been used at this rate at F/3.0 aperture by tolerating more diffraction loss. However, its 12.6-mm height is insufficient to handle large, current-dependent beam shifts parallel to the junction. Indeed, as follows from Section 4.2.1, a mirror with this capability is impractical, and the limitation must be removed by the use of pulse-duration rather than pulse-amplitude modulation.

The model G-108 galvanometer, which has been used in nearly all our printers, can produce a linear sawtooth scan at 600-Hz rates when the external load is small compared to the galvanometer intrinsic moment of inertia. The external load (mirror plus mounting seat) must be kept below 0.025 gm cm^2 to maintain linearity at sawtooth scan rates of 500 Hz. The mirror principal-axis of inertia should be aligned with the galvanometer axis to achieve even loading of the shaft bearings. The mirror moment of inertia can be reduced significantly by tapering, as illustrated by the dimensions t_1 and t_2 in Fig. 38. The taper is formed prior to grinding the front of the mirror optically flat. No trouble has been encountered in maintaining flatness during sawtooth deflection up to 500-Hz line rate, but considerable art is needed to avoid flatness distortions when the mirror is affixed, either by epoxy or by mechanical clamping, to the galvanometer shaft. Such distortions can easily exceed $\lambda/2$, and are among the most frequent causes of low machining efficiency and poor images. A total thickness $(t_1 + t_2) \geqslant 1.5$ mm is usually needed to preserve $\lambda/4$-quality flatness after mounting when mirrors larger than about 10-mm diameter are used. Hence a moment of inertia of about 0.05 gm cm^2 is expected for a practical mirror (viz., Fig. 38) used with a GaAs laser in the machining geometry of Fig. 34. Factor-of-2 reductions in the moment of inertia of glass mirrors have been obtained, apparently without impairing the mirror flatness, by honey-combing the back of the glass blank, but such mirrors have not yet been used in a fast-scan-rate printer.

The mirror width needed to accommodate a given beam diameter can be reduced by 30 percent, and the mirror moment of inertia by nearly a factor of 3, if the machining geometry of Fig. 34 is modified so that the beam is normally incident onto the mirror ($\psi = 0°$). This is readily accomplished by using the scanning lens in a double-pass mode[62] as shown in Fig. 39. This usage exploits the fact that focus is flat over a plane, so that rays which appear to diverge from a point such as a′, located in the focal plane an appropriate distance above or below the scanned line, can be collimated by the lens, reflected from the galvanometer mirror at a small bias angle ψ, and refocused onto the line being scanned. A smaller scanning lens can be used because the entrance-pupil-to-lens distance need now be only a few millimeters,



t_1 (cm)	t_2 (cm)	R_{sph} (cm)	I_m gm cm^2
0.0762	0.0254	2.96	0.0278
0.0762	0.0508	2.96	0.0408
0.0916	0.0508	2.48	0.0443

Fig. 38—Galvanometer-mirror properties. The mirror is deflected about axis aa', and the beam is incident onto it at an angle of 40 degrees to the normal. The laser junction plane is oriented parallel to axis aa'. Profiles of the beam, as enlarged and collimated by an F/7 Galilean telescope preparatory to machining by a 37-mm-focal-length machining lens, are shown. The mirror moment-of-inertia I_m is strongly influenced by the thicknesses t_1 and t_2. Difficulty in maintaining λ/4 flatness of this mirror surface is encountered if $(t_1 + t_2)$ is smaller than 1.5 mm. The deflection-system loaded resonant frequency is also shown as a function of I_m for a G-108 galvanometer, whose unloaded moment-of-inertia I_u equals 0.011 gm cm^2.

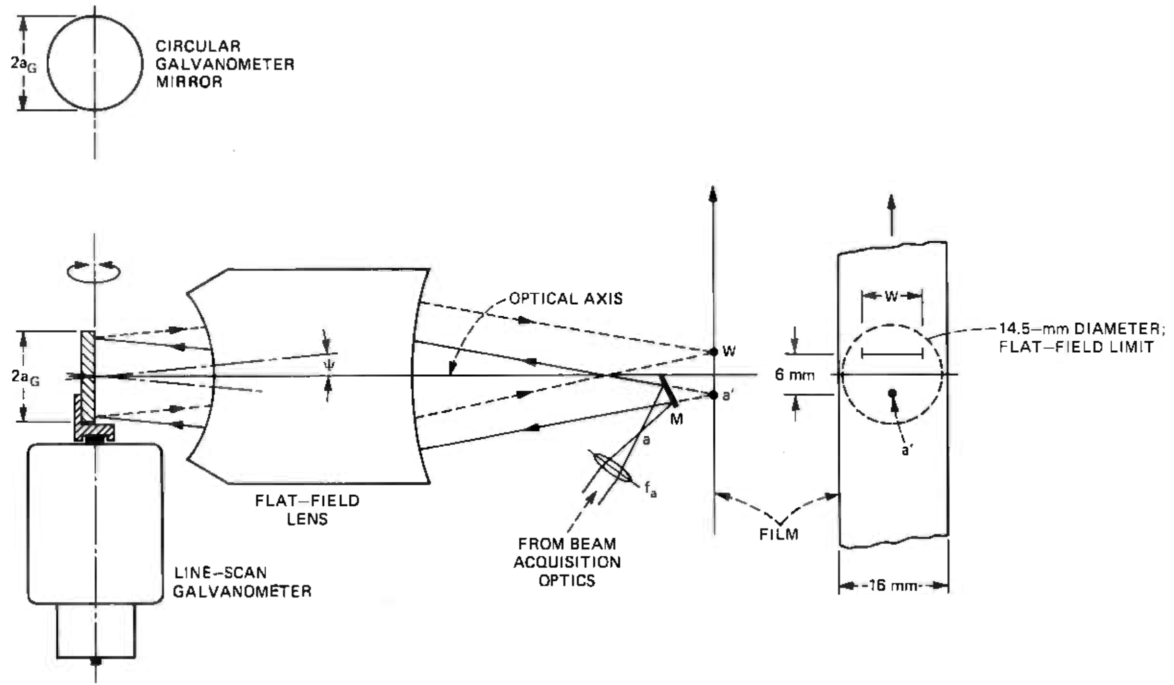

2a_G
CIRCULAR GALVANOMETER MIRROR
2a_G
OPTICAL AXIS
ψ
W
a'
M
a
f_a
FLAT–FIELD LENS
LINE–SCAN GALVANOMETER
FROM BEAM ACQUISITION OPTICS
FILM
6 mm
W
14.5–mm DIAMETER; FLAT–FIELD LIMIT
a'
16 mm

Fig. 39—Double-pass machining optics for higher scan rates. Lens f_a is a microscope objective chosen to produce a beam waist at point a equal in size to the desired machining beam waist. The small steering mirror M produces a virtual image of point a at point a' located in the film plane about 3 mm below the flat-field-lens optic axis. The beam diverging from M is collimated by the flat-field lens and made incident onto the galvanometer mirror at an angle $\psi = 4.6$ degrees to its normal. After deflection by the moving galvanometer mirror, the beam is refocussed by the flat-field lens to the appropriate point on the scanned focal line in the film. The scanned line of length W is located 3 mm above the optic axis.

and symmetries of the double-pass geometry may permit other simplifications in the lens design. Although assembly of a double-pass machining system has required some extra care, its machining quality with GaAs lasers at 160-Hz line scanning rate has been the same as that of the Fig. 34 single-pass system.

Marginally acceptable scan linearity was obtained for images machined at 500-Hz line rate with the double-pass geometry when a 9-mm-diameter mirror, tapered similarly to the example in Fig. 38, was used, with approximately 35 percent of the 2-ms line period allotted to flyback. Unfortunately, the image quality at this higher printing speed was unacceptable for the reasons mentioned in Section 4.4.2: namely, nonuniform hole sizes were machined in correlation with the laser line-averaged duty factor. At high pulse-repetition rates, the varying facsimile-information content produced temperature variations at the p-$n_{0.02}$ junction which prevented the laser from operating along a unique or predictable machining curve. The laser mounting stud was cooled thermoelectrically to maintain the p-active region at 20°C at 1-MHz pulse rate and prolong laser life, but the heat sinking was ineffective in damping temperature fluctuations produced on a millisecond time scale. Examples of such hole size nonuniformity are shown in Section 8.3.

7.2 Frame-scan, film-gate, and frame-projection considerations

A very high degree of smoothness in the frame scan (slow scan) is essential for image uniformity. Hesitations, slippages, or jerkiness which produce abrupt changes as small as a few microns in the spacing between adjacent scanned lines are highly visible in the final image—the more so if the discontinuities occur periodically along the frame length. A close match of the film to the focus of the line scan is similarly essential. A tilt of the film by one milliradian with respect to the line-scan focus can cause the beam size at the film to vary sufficiently over an 8-mm frame width to produce noticeable grayscale variations from side to side in the final image.

Early printers that used a flat-field scanning lens were designed for stationary-frame, real-time viewing as sketched in Fig. 40a. The film was tensioned over a rectangular, frame-size opening in a smooth metal plate, nominally coincident with the focal plane of the scanning lens, and was vacuum-chucked around the perimeter of this opening. The vibrating-diaphragm pump which produced the film hold-down vacuum was also used to draw off the metal "tektites" which are sprayed out from the film during machining in a well-aligned printer. The line-scanning galvanometer was mounted in a gimbal whose axis passed through the center of the mirror orthogonally to the galvanometer axis and to the machining-lens optical axis. The gimbal axis was direct-

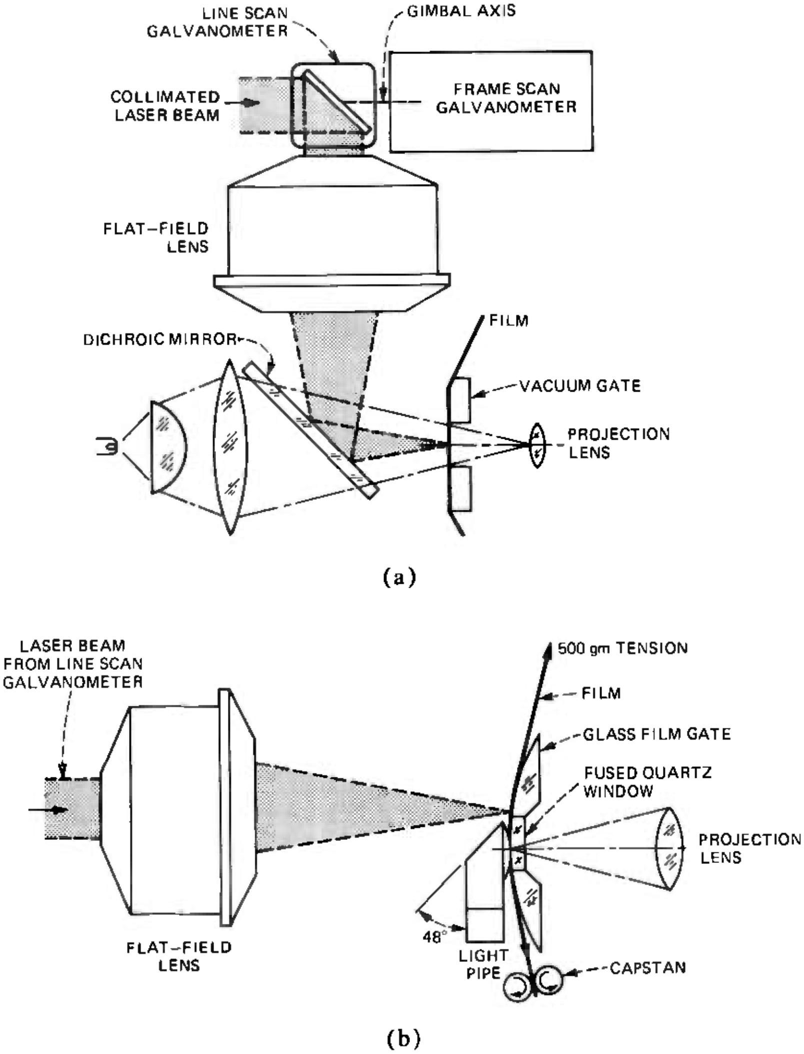


Fig. 40—Stationary-frame and moving-film machining systems. (a) The film is free-standing in the gate opening; the vacuum system which clamps it down around the perimeter of the opening is not shown. This vacuum system also draws away the metallic "tektites" which are usually evolved during machining and prevents their redeposition on the film. The line scan occurs in the plane of the drawing. (b) The line scan is perpendicular to the plane of the paper. An elastic putty is used as the dashpot fluid for damping the film-transport capstan.

coupled to the shaft of a heavy-duty, slow-scan galvanometer (General Scanning Co. G-300 series). A $\lambda/4$-flat dichroic mirror between lens and film reflected the infrared machining beam onto the film and simultaneously transmitted visible light from a projection lamp to the film. The beam was focused by adjusting the machining lens. The gimbal-drive galvanometer slowly scanned the machining line down the length of the frame, producing a stationary final image which became visible instantly as it was machined.

Such real-time viewing has putative esthetic value but numerous engineering problems: the free-standing frame area must remain coincident with the machining focal plane in the face of temperature variations caused by its continuously changing absorption of projection light, severely limiting the tolerable projection brightness; the slow-scan galvanometer cannot be damped heavily enough to be free of microphonic vibrations and jerkiness; the scanning lens must be designed with an awkwardly long working distance and the projection-lamp optics with large diameters to accommodate the dichroic mirror; and the mirror itself is expensive and must be provided with twist and tilt adjustments. The microphonic vibration problem was particularly intransigent.

These defects can be avoided by abandoning the attempt to write in strict real time and by using instead a heavily damped film transport for the frame scan, with the film stretched at high tension over a glass cylindrical surface, as depicted in Fig. 40b. A rectangular, optically flat, fused-quartz window was set into this detail, tangent at its edges to the two cylindrical sections, whose radius of curvature was 50 mm. Machining occurs along a line close to the boundary between one cylindrical surface and the flat window, and the machined portion of film is transported over the window as frame machining proceeds. The distance between the machining line focus and the position at which the image first becomes visible in the projection field is approximately one-tenth of a frame height. Following this delay, the written frame moves steadily into the projection field and comes to rest, centered in the field, a short time after the last line has been machined. The film was tensioned at 500 gm, which forced it to be flat within $\pm 2\ \mu$m across the machining line (well within the focused-beam-waist field depth) but maintained it free-standing about 25 μm away from the optical flat. This latter spacing is sufficient to wash out the effects of color fringes which would occur in the projected image if the film were in more intimate optical contact with the flat. The damping dashpot responsible for the continuity and linearity of the frame scan used Dow Corning bouncing putty as the viscous fluid, the friction disk being driven by a 300-ounce-inch synchronous motor.

A projection illumination flux of 7.5 lumens was obtained from a fan-cooled, 80-W, quartz-iodine projection lamp with an integral focusing reflector. A fused-quartz light pipe was used to convey this light from the rear of the printer enclosure to the film. A prism reflector and a 20-mm-focal-length lens at the end of the light pipe (viz. Fig.

40b) directed the light through the film and into an F/2 projection lens of 19-mm focal length, causing an image of the frame to be rear-projected onto an 8.5 × 11 in., frosted-glass screen via two path-folding mirrors. The brightness of the final projected image, including a gain factor of 1.5 from the forward-scattering frosted screen, was 17 foot-lamberts. The screen was highly directional despite its low gain, providing satisfactory viewing contrast at forward angles even with a 160-W fluorescent lighting fixture operating directly overhead. Projection-flux power densities up to 150 mW/cm^2 at the film have caused no problems with film flatness or film overheating, in contrast with the film buckling which afflicts stationary-frame printers at high projection-lamp brightness. The scanning lens in the moving-film printer was cantilevered from the film-gate support by Invar rods (not shown) to maintain machining focus against ambient temperature changes. The machining and projection optics for either the stationary-frame or the moving-film systems can easily be fitted into the confines of conventional, desk-top, microfiche or microfilm viewers. A photograph of a GaAs laser printer utilizing the moving-film system and the projection optics just described is shown in Fig 41.

VIII. MACHINED-IMAGE QUALITY

The ultimate quality of laser-machined images, limited by the number of hole sites allotted to each frame and by the reproducibility of the machining curve, can be very high. The quality achieved in practice is determined for high contrast black-and-white images primarily by the linearity and smoothness of the raster scan, and for gray-scale images both by scan linearity and by the shape and reproducibility of the machining curve near threshold. Attempts have not been made in the present stage of printer developments to match the video-amplifier gain curve dynamically to the overall machining curve of a GaAs laser and Se/Bi film sample. Hence, only the expected result can be reported: the contrast, boldness, and edge-sharpening adjustments which produce best high-contrast performance differ, in general, from those which produce best gray scale. Consequently, the images machined in Se/Bi film in a printer assigned to general facsimile use are the result of compromise. If the printer is committed to black-white graphics use, there is no need to compromise.

8.1 Facsimile images

Facsimile signals were generated from original documents by two methods. In the first method, the page was raster-scanned by a He-Ne-laser, double-galvanometer, flying-spot scanner which used a photomultiplier to detect the document red-light reflectivity. In the second method, the length of the page was mechanically scanned by an

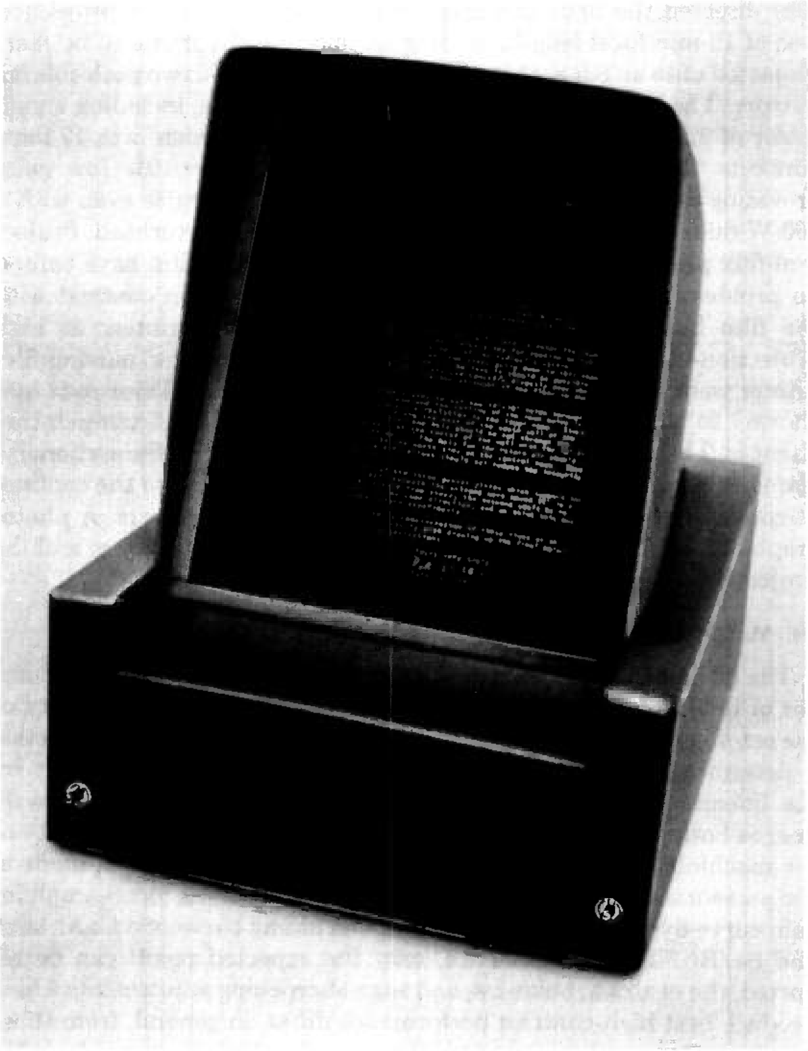
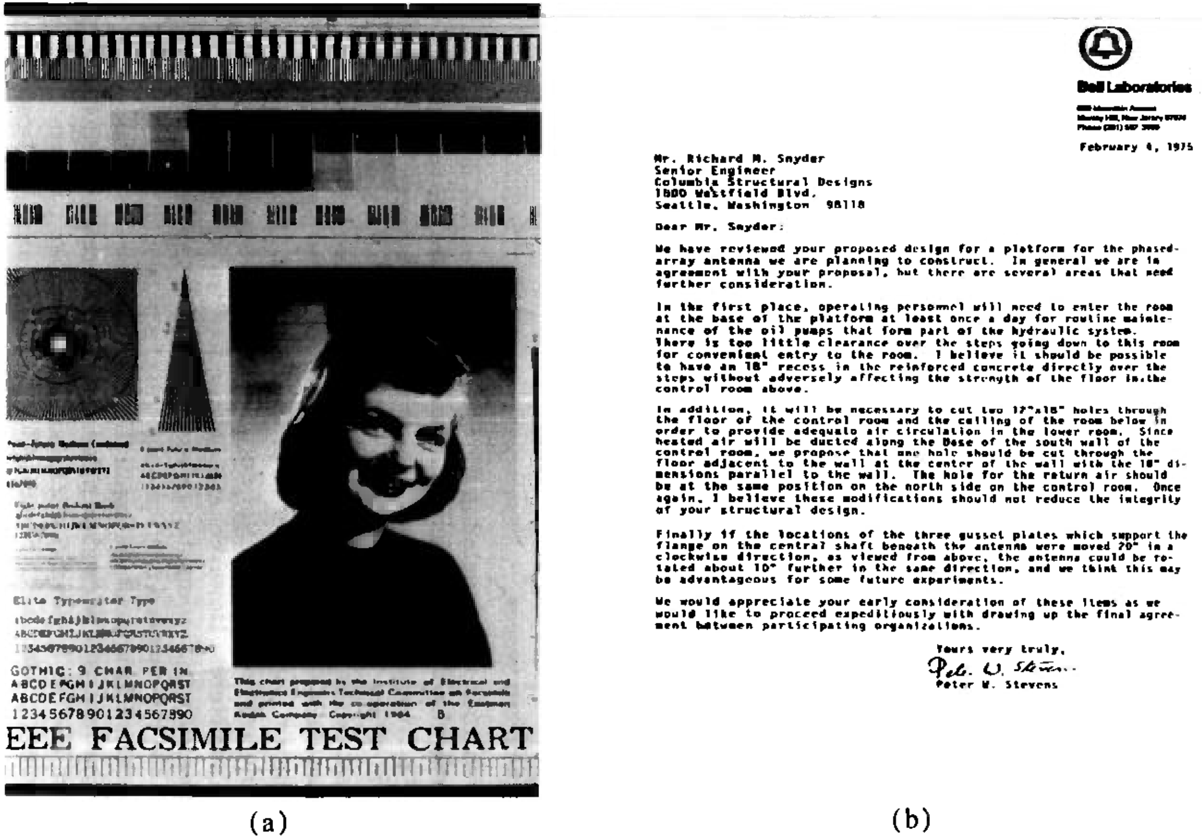


Fig. 41—Desk top gallium-arsenide laser printer. The viewing screen size is 8.5 × 11 inches.

assembly which provided white-light illumination and employed a 1728-element, linear, charge-coupled photodetector* to effect the line scan. Figures 42a and 42b are photographic enlargements of 7.7-mm-wide frames machined on Se/Bi/IBM film by a GaAs-laser printer. This printer acquired the beam with the optics of Fig. 31b and

* CCD 121 Linear Image Sensor, manufactured by Fairchild Semiconductor Components Group.


Elite Typewriter Type
GOTHIC: 9 CHAR PER IN
ABCDEFGHIJKLMNOPQRST
ABCDEFGHIJKLMNOPQRST
12345678901234567890
EEE FACSIMILE TEST CHART
(a)
(b)

Bell Laboratories

[illegible]

February 4, 1975

Mr. Richard M. Snyder
Senior Engineer
Columbia Structural Designs
1800 Westfield Blvd.
Seattle, Washington 98118

Dear Mr. Snyder:

We have reviewed your proposed design for a platform for the phased-array antenna we are planning to construct. In general we are in agreement with your proposal, but there are several areas that need further consideration.

In the first place, operating personnel will need to enter the room at the base of the platform at least once a day for routine maintenance of the oil pumps that form part of the hydraulic system. There is too little clearance over the steps going down to this room for convenient entry to the room. I believe it should be possible to have an 18" recess in the reinforced concrete directly over the steps without adversely affecting the strength of the floor in the control room above.

In addition, it will be necessary to cut two 12"x18" holes through the floor of the control room and the ceiling of the room below in order to provide adequate air circulation in the lower room. Since heated air will be ducted along the base of the south wall of the control room, we propose that one hole should be cut through the floor adjacent to the wall at the center of the wall with the 18" dimensions parallel to the wall. The hole for the return air should be at the same position on the north side on the control room. Once again, I believe these modifications should not reduce the integrity of your structural design.

Finally if the locations of the three gusset plates which support the flange on the central shaft beneath the antenna were moved 20° in a clockwise direction, as viewed from above, the antenna could be rotated about 10° further in the same direction, and we think this may be advantageous for some future experiments.

We would appreciate your early consideration of these items as we would like to proceed expeditiously with drawing up the final agreement between participating organizations.

Yours very truly,

Pete W. Stevens

Peter W. Stevens

Fig. 42—Quality of laser-machined facsimile images. Photographic enlargements of facsimile images of two types of subject matter as machined in Se/Bi/IBM films by a GaAs-laser printer are shown. The laser pulse repetition rate was approximately 0.33 MHz and the line scan rate approximately 160 Hz (with 30-percent allowance for flyback time). The machined frames are 7.7-mm wide, and their facsimile signals were generated by scanning with a linear charge-coupled detector.

machined with the stationary-frame optics of Fig. 40a, utilizing pulse-intensity modulation rather than pulse-duration modulation to vary the machined-hole size. Facsimile signals for these 2000-line frames were obtained from the charge-coupled photodetector operated at 160-Hz line repetition rate. The ramp for the printer line-scan galvanometer was synchronized to this photodetector electronic scan at the start of each line.

Resolution and gray scale both suffer in the photographic processing of Fig. 42, relative to the image quality perceptible in ×20 rear-projection enlargements of the microimages onto a ground-glass screen. Six-point type can just be read in the machined IEEE test charts of Fig. 42a, whereas four-point material can be read in the rear-projection enlargements, with only minor difficulties caused by line-scan jitter, edge raggedness, and missing holes. Gray-scale tonality in the GaAs-produced frames ranges up to seven distinguishable shades when viewed under rear-projection conditions. (Twelve distinguishable shades have been attained on 10.3-mm-wide frames which were machined by a cavity-dumped YAG-laser printer, using similar subject matter scanned by the He-Ne flying-spot scanner.) Image defects are noticeable even in casual inspections of Figs. 42a and 42b, and their number increases as the inspection becomes more detailed and critical. Examples of the most frequently occurring defects are presented in Section 8.3.

8.2 Synchronized images

The resolution of fine detail in the facsimile images just illustrated falls short of the ultimate which could be expected of frames containing 3.3×10^6 picture elements. Lack of synchronization between the laser pulse timing and the galvanometer-mirror instantaneous position is the largest factor in this shortcoming. The examples below demonstrate that such synchronization is inappropriate in general facsimile usage, since it runs the danger of producing a beating phenomenon (moiré fringes) in the images of documents containing finely-spaced periodic structures. However, synchronization can be employed to great advantage with computer-generated images, as shown in Fig. 43.

Figure 43a represents a microphotograph of a portion of a machined image obtained in a facsimile application. It illustrates the edge raggedness characteristic of random pulse timing with respect to the line-scan start and of a galvanometer employing neither position- nor velocity-sensing feedback. Figure 43b illustrates an ASCII-coded graphics application. Alphanumeric graphics were generated by converting 8-bit ASCII code into a 16x20-element matrix of machining hole sites; a row of ASCII characters was translated by a minicomputer into machining instructions for 20 rows of holes which were transmitted to the printer synchronously with a staircase sweep applied to a line-

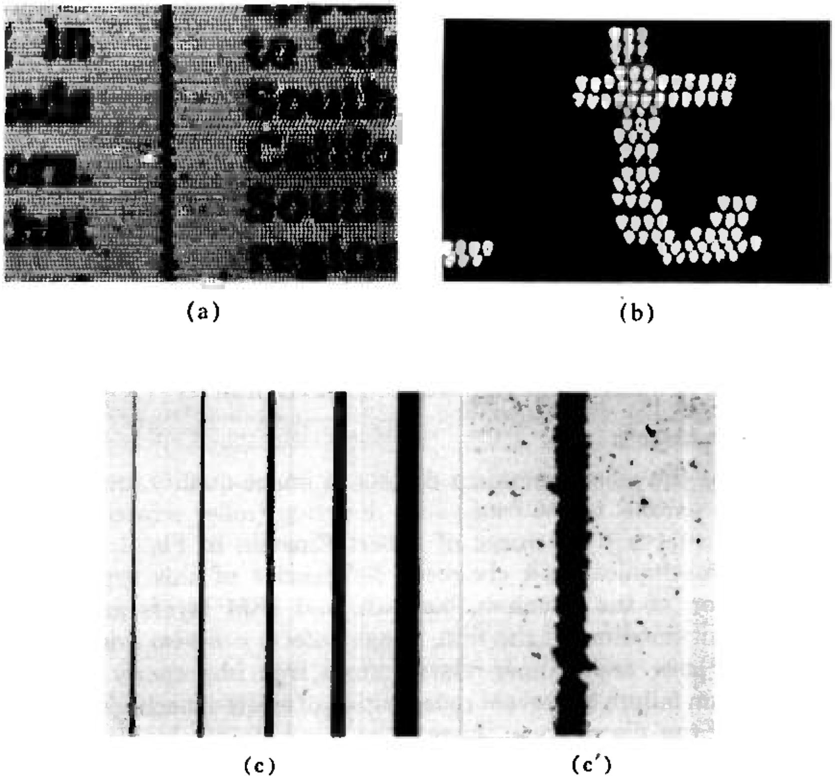


Fig. 43—Effects of pulse synchronization on the definition of vertical edges in facsimile use and in computer-graphics use. (a) Laser pulses free-running with respect to start of line scan; no position- or velocity-sensing feedback used with line scan. (b) An alphanumeric graphic character from moving-film printer. The laser pulses are synchronized to the edges of a 20-Hz-line-rate staircase sweep applied to the galvanometer, using position-sensing feedback. Vertical edges are defined to within about ±0.5-hole diameter. (c) Laser pulses synchronized to scanned-beam instantaneous position; the direction of line scan is from left to right. Photograph (c′) is an enlargement of the narrowest line in (c) and represents a line exactly one hole-diameter wide. Vertical edges are defined to within ±0.1-hole diameter.

scanning galvanometer. Well-shaped characters were obtained by synchronizing the laser pulses to edges of the staircase and by using a position-sensing feedback capacitor to assure galvanometer deflection reproducibility from line to line. Figure 43b is an enlarged microphotograph of one such character.

The best vertical edge definition yet achieved in a laser-machining printer geometry is depicted in the microphotographs of Figs. 43c and 43c′. These vertical lines were machined in a raster-scanned application in which a secondary laser beam, collinear with the machining beam*

* The machining beam for Figs. 43c, 43c′ was obtained from a cavity-dumped ArII laser.

but differing from it in wavelength, was deflected by the galvanometer, separated out by a dichroic mirror, and made incident onto a ruled transmission grating located in a (mirror-image) focal plane of the scanning lens. The pulse count generated in a detector placed behind the grating measured the instantaneous beam position and was used to synchronize the output information from a minicomputer. The resolution provided by the focused secondary-beam size and the grating ruling density was higher than 1600 sites per line and, with suitable signal processing, permitted the machining of long straight lines exactly one hole-diameter wide, as in Fig. 43c. This width corresponds to 0.005 in. on an 8.5-in. wide page, smaller than the line width of two-point characters, and suggestive of a capability for handling computer-generated mathematical text and extracondensed print.

8.3 Image defects

Perhaps the most prominent defects in image quality are the long striations parallel to the frame-scan direction (roller scratches), such as those affecting the image of Albert Einstein in Fig. 1c. Beyond obvious mechanical and chemical deficiencies of this type in the preparation of the selenium, bismuth, and IBM layers and in the subsequent handling of the film, image defects can also arise from a variety of laser- and scanner-related causes, from improperly corrected optics, from failure to prevent redeposition of ejected machining debris, and from the discreteness inherent in the hole machining method. Some of these are illustrated in Figs. 44, 45, and 46 for GaAs-laser facsimile images.

Fig. 44

(a) Frame-scan discontinuities (D)—microphonics, caused by rapping the printer table, affected the gimballed frame-scanning galvanometer; this highly visible defect is not seen with strongly damped, film-transport scans.

(b) Line-scan nonlinearity—the line-scanning galvanometer is moving faster (F) and slower (S) than its average speed in the indicated regions; the defect is difficult to correct only when the scan rate exceeds 40 percent of the galvanometer loaded resonant frequency.

(c) Line-scan wobble—bad galvanometer bearings caused waviness in the scan lines, most easily seen here in region VV; local bunching of lines (bands W) produces short, horizontal, dark lines in the image which are highly visible.

(d) Frame-scan nonlinearity—regions of faster- and slower-than-average frame speed, indicated by (F) and (S), occur both with gimballed galvanometer and with film-transport frame scans.

Fig. 45

(a),(a′) Moiré effect—a position-sensing line-scan galvanometer was used, with machining synchronized to the start of the scan, for facsimile

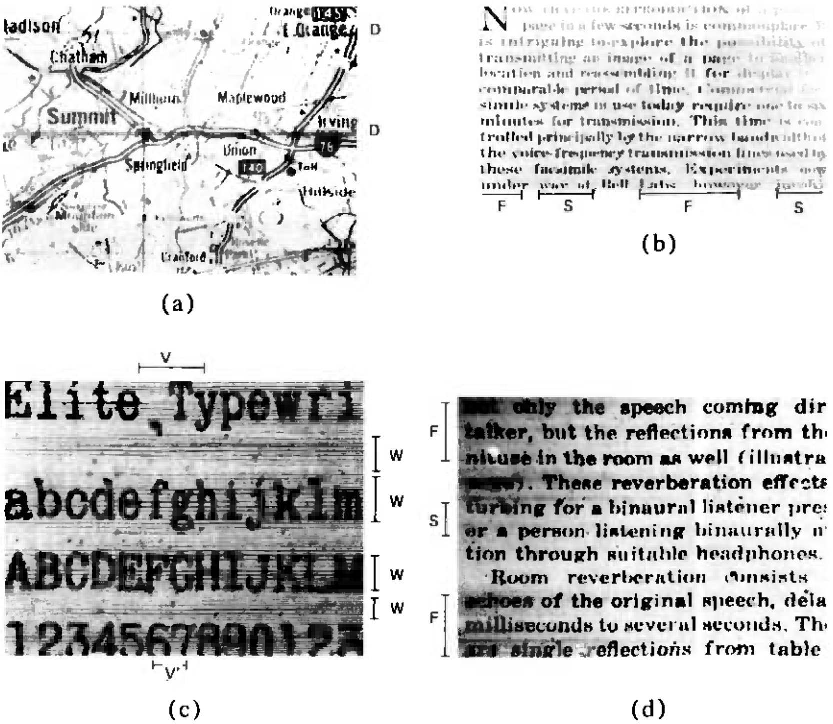


Fig. 44—Image defects. Various defects caused by line- and frame-scan irregularities, as described in text.

image (a); a strong moiré beat results from the attempt to reproduce the dot array in the original (a′) with the regular-hole array of the facsimile.

(b) Moiré effect—galvanometer did not utilize position- or velocity-sensing feedback, and laser pulses were unsynchronized with start of line scan; moiré beat in halftone background (Atlantic Ocean) is weaker than in Fig. 45a.

(c) Bad machining in center of frame—film buckled out of the machining focal plane during "real-time" printing due to overheating by projection illumination; a limitation with the stationary-frame printer of Fig. 40a, but not with the moving-film printer of Fig. 40b.

(d) Badly machined holes—poor transmission in "white" areas, caused by balled-up metal remaining in center of each hole (confirmed by secondary electron microscopy); typical of machining produced by poorly corrected optics and/or a nonplanar galvanometer mirror.

Fig. 46

(a),(a′) Duty-factor-dependent machining—given a 500-Hz line-scan rate and the indicated directions of line and frame scans, note that

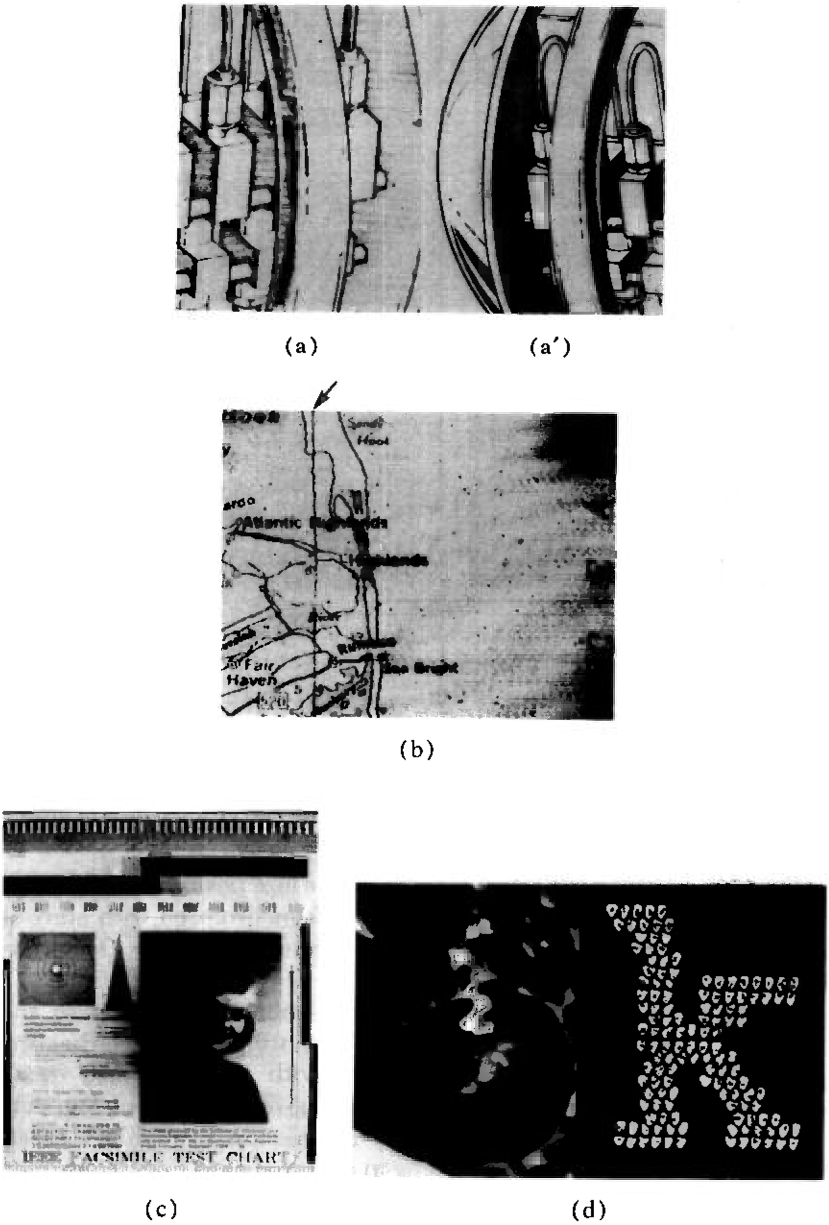
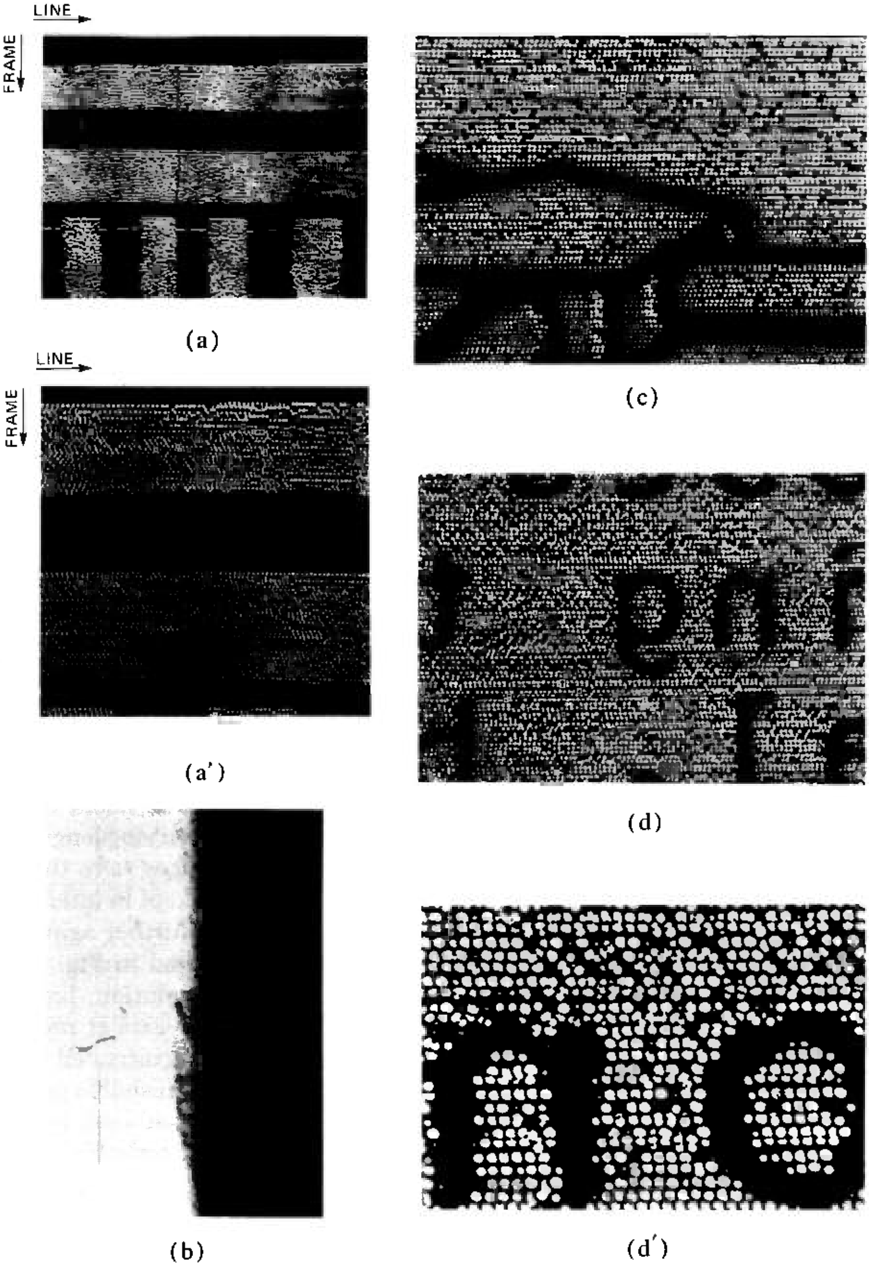


Fig. 45—Image defects. (a) Moiré effects in facsimile. (a′) Original of (a). (b) Moiré effect. (c) Film buckling. (d) Debris in holes caused by poor-optical-quality machining beam.



(a) (a') (b) (c) (d) (d′)

Fig. 46—Image defects. (a), (a′) Laser overheating at high printing speed. (b) Redeposition of machining debris. (c), (d), (d′) Erratic machining caused by reflection-induced instabilities.

holes produced at 1-MHz pulse-repetition rate slowly decrease in size and quality along the length of a scanned line. The laser temperature does not completely recover during flyback time at the 500-Hz line-scan rate, so the hole-quality degradation accelerates with successive scans of high-duty-factor machining. However, if the laser is left off for several lines, the temperature recovers, and large, well-shaped holes reappear when machining resumes.

(b) Redeposition of machined bismuth—metal, evolved from Se/Bi/IBM film during machining at 0.3-MHz repetition rate, is shown redeposited on uncoated margin of film. This effect can impair image quality if redeposition is permitted to occur within the frame, as in "Madison-Summit" area of map, Fig. 44a. Its effect can be more serious at 1-MHz repetition rate where the bismuth, when allowed to redeposit ahead of the raster scan, has raised the machining threshold enough to prevent machining.

(c),(d),(d′) Reflection-induced beam destabilization—this effect, described in Section 4.2.3, can cause holes to be omitted (c), or to be both omitted and misshapen (d), (d′).

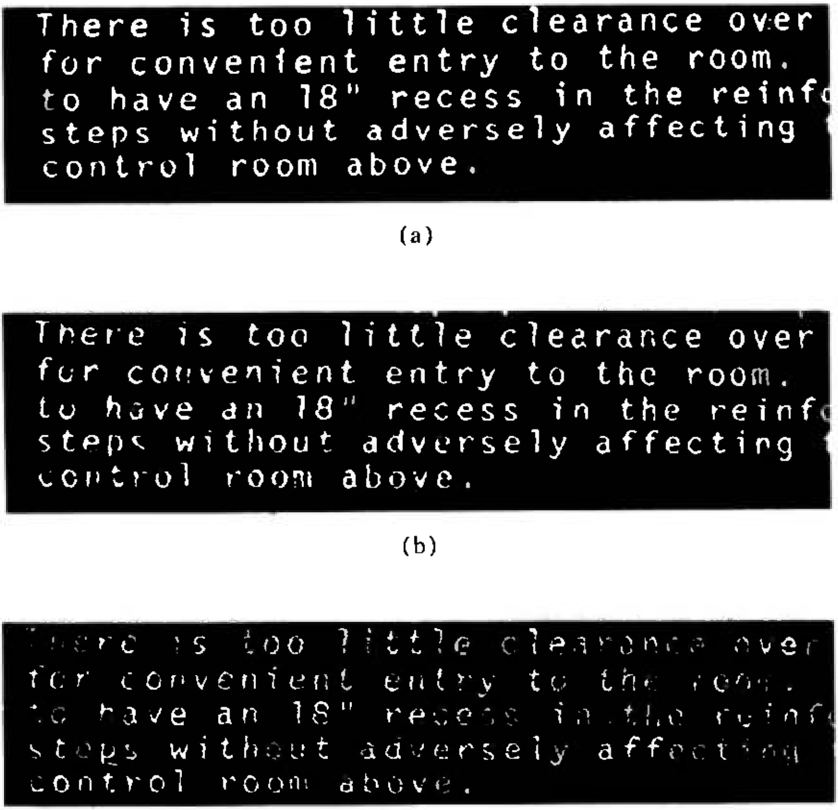
Means for curing or avoiding most of the defects illustrated in Figs. 44, 45, and 46 have been described in the preceding sections.

8.4 Image quality of aged film in high contrast use

Extensive time-temperature stress tests conducted on the Se/Bi/IBM/Mylar film system over the temperature range 25° to 60°C yield an activation-energy value of 24 Kcal/mole for the aging process responsible for degraded film machining, assuming that the process corresponds to a single activation energy at these temperatures. Under this assumption, one week of life at 60°C is equivalent to about one year of life at 25°C. Film samples were aged at 60°C for varying lengths of time up to 20 weeks, and high-contrast negative images were then simultaneously machined in them. (Aged samples were kept in interim storage at 4°C, a temperature which strongly inhibits further aging.) The results for 2-, 10-, and 20-week samples, as depicted in Fig. 47, provisionally indicate that the film can yield high-resolution, high-contrast, negative images after a time equivalent to 10 years at room temperature. A second test, made in a printer which was intended for high contrast use and which produced only a few distinguishable gray levels, showed that three-year-old films of Se/Bi which had been kept at room temperature provided the same machining thresholds and image quality as did freshly prepared film.

IX. SUMMARY AND DISCUSSION

Miniature archival images have been machined in Se/Bi film with the galvanometer-deflected beam from a pulsed-current GaAs laser.



There is too little clearance over
for convenient entry to the room.
to have an 18" recess in the reinf
steps without adversely affecting
control room above.

(a)

There is too little clearance over
for convenient entry to the room.
to have an 18" recess in the reinf
steps without adversely affecting
control room above.

(b)

There is too little clearance over
for convenient entry to the room.
to have an 18" recess in the reinf
steps without adversely affecting
control room above.

(c)

Fig. 47—Aging of Se/Bi/IBM film in high-contrast use. Figures (a), (b), and (c) refer to Se/Bi/IBM Mylar film samples which had been aged at 60°C for 2, 10, and 20 weeks, respectively, prior to machining of computer-generated negative images by the moving-film printer. Given an activation energy of 24 Kcal/mole for the film-degradation process, image (b) corresponds to high-contrast machining on film aged for 10 years at 25°C.

High resolution is readily available with this image-forming method. Frames containing 3×10^6 machining hole sites and demonstrating up to seven distinguishable gray levels have been written in 12 seconds. Machined white background areas of the film are about 50 percent transparent to visible projection illumination, and unmachined areas are less than 1 percent transparent. Information can be added to unmachined frame areas at later times, up to a limit set by the shelf life of the film. The images can be viewed instantly as they are written.

Polymer-undercoated Se/Bi is the most sensitive metal film system for the GaAs wavelength which exhibits mechanical integrity and good shelf life. At present, the room-temperature shelf life of this film for high-contrast (black-white) printing exceeds 10 years, and the archival life of machined images probably exceeds several centuries. Further development effort on Se/Bi film might concentrate on extending shelf life and improving the compromise between low machining threshold energy and adhesion. Rather sketchy aging data suggest that present GaAs machining lasers can have commercially adequate life when used at 12-s frame rates. However, attempts to reduce the frame machining time to 4 seconds have foundered on laser thermal diffusivity limitations, and this printing speed also strains the capabilities of the line-scan galvanometer.

Junction-plane secondary modes, which do not contribute to hole enlargement, are usually present in the pulsed output of proton-delineated, stripe-geometry, double-heterostructure GaAs lasers when their current exceeds lasing threshold by 10 or 20 percent. As a result, the total power which must be drawn from present lasers to obtain enough peak intensity at the Se/Bi film to machine 8×10mm images often exceeds by 50 percent or more the power that would suffice if only the fundamental beam mode were present. Also because of the junction-plane modes, the numerical apertures required in the deflection and writing optics to focus the main GaAs machining lobe into a given size at the film are larger than would be calculated from the Gaussian beam propagation law. Recent progress in high-power buried-stripe heterostructures provides hope that most of these laser-related printing problems will eventually yield to satisfactory technical compromise.

X. ACKNOWLEDGMENTS

The development of laser-machining facsimile receivers started in 1970 with initial contributions from D. Maydan and R. Sard, and has been concentrated since 1973 on the Bi-Se film system and the GaAs laser. A large number of coworkers have been associated with the project since its beginning.

Contributions to metal film studies have been made by D. J. Bartelink, E. W. Chase, P. S. Henry, and D. Y. Lou, with assistance from C. Van Hise and W. Q. McKnight. Contributions to the GaAs laser development were made by B. C. DeLoach, B. W. Hakki, R. T. Hepplewhite, D. C. Krupka, T.L. Paoli, N. C. Schumaker, and D. L. Van Haren. The experimental work on early GaAs printers was done by R. G. Chemelli, D. D. Cook, and D. C. Krupka, assisted by R. B. Lawry and E. W. Chapman. Work on the most recent printer was done by T. G. Melone, J. S. Wagner, A. D. White, and D. L. White. A prototype YAG-laser printer was investigated by D. Cheng, with assistance from M. C. Page; a compact, self-contained version of this printer was assembled and made operational by R. G. Chemelli and E. D. Urbanek. The charge-coupled-device page scanner was produced by D. Cheng, E. W. Chase, R. G. Chemelli, and E. D. Urbanek. A large fraction of the diffraction-limited printer optics was designed either in prototype or final form by A. D. White. Optics fabrication was in the hands of W. Gronros and T. G.Melone. Joseph Hill was responsible for the mechanical design details of several GaAs printers.

The authors thank J. H. Brúning for permission to use results from his ArII mask-machining facility, J. S. Courtney-Pratt and his associates for the He-Ne flying-spot scanner, R. G. Chemelli for help with the preparation of this paper, and R. W. Dixon for valuable comments on the manuscript.

REFERENCES

1. H. A. Watson, Bell Laboratories Record, *53* (March, 1975), p. 163.
2. D. Maydan, "Micromachining and Image Recording on Thin Films by Laser Beams," B.S.T.J., *50,* No. 6 (July-August 1971), pp. 1761–1789.
3. H. S. Carslaw and J. C. Jaeger, *Conduction of Heat in Solids,* 2nd ed, New York: Oxford, 1959, Ch. XII.
4. C. O. Carlson, E. Stone, H. L. Bernstein, W. K. Tomita, and W. C. Myers, Science, *154* (December 23 1966), p. 1550.
5. M. B. Panish and I. Hayashi, *Appl. Solid State Science*, Vol. 4, ed. R. Wolf, New York: Academic Press, 1974, p. 235.
6. L. A. D'Asaro, J. Luminescence, *7* (1973), p. 310.
7. M. Terao et al, Proceedings of IEEE/OSA Conf. on Laser Eng. and Applic., Wash. D.C., June, 1979.
8. T. Maekawa, T. Yokokawa, and K. Niwa, J. Chem. Therm., *4* (1972), p. 873.
9. P. J. Brosens, Electro-Optical Systems Design, 4/71, p. 21.
10. D. Marcuse, *Light Transmission Optics*, New York: Van Nostrand Reinhold, 1972, Ch. 6.
11. D. D. Cook and F. R. Nash, J. Appl. Phys., 46 (1975), p. 1660.
12. R. G. Chemelli, D. D. Cook, and R. C. Miller, U.S. Patent 4,030,122 (June 14, 1977).
13. R. G. Chemelli and R. C. Miller, U.S. Patent 3,974,507 (Aug. 10, 1976).
14. T. L. Paoli, B. W. Hakki, and B. I. Miller, J. Appl. Phys., *44* (1973), p. 1276.
15. J. C. Dyment, L. A. D'Asaro, J. C. North, B. I. Miller, and J. E. Ripper, Proc. IEEE Lett., *60* (1972), p. 726.
16. H. C. Casey, Jr. and M. B. Panish, *Heterostructure Lasers; Part B: Materials and Operating Characteristics*, New York: Academic Press, 1978, Ch. 8.
17. B. W. Hakki and C. J. Hwang, J. Appl. Phys., *45* (1974), p. 2168.
18. D. C. Krupka, IEEE J. Quantum Electronics, *QE-11* (1975), p. 390.
19. B. W. Hakki, J. Appl. Phys., *46* (1975), p. 292.
20. W. B. Joyce and S. H. Wemple, J. Appl. Phys., *41* (1970), p. 3818.
21. W. O. Schlosser, "Gain-Induced Modes in Planar Structures," B. S. T. J., *52*, No. 6 (July-August 1973), pp. 887–905.
22. F. R. Nash, J. Appl. Phys., *44* (1973), p. 4696.
23. D. Marcuse, Ref. 10, Ch. 7.
24. G. H. B. Thompson, D. F. Lovelace, and S. E. H. Turley, Solid-State and Electron Devices, *2* (1978), p. 12.
25. T. L. Paoli, IEEE J Quantum Electronics, *QE-13* (1977), p. 662.
26. R. W. Dixon, F. R. Nash, R. L. Hartman, and R. T. Hepplewhite, Appl. Phys. Lett., *29* (1976), p. 372.
27. T. L. Paoli, IEEE J. Quantum Electronics, *QE-12* (1976), p. 770.
28. R. W. Dixon, W. B. Joyce, and R. C. Miller, J. Appl. Phys., *50* (1979), p. 1128.
29. H. C. Casey Jr., M. B. Panish, and J. J. Merz, J. Appl. Phys., *44* (1973), p. 5470.
30. T. H. Zachos and J. C. Dyment, IEEE J Quantum Electronics, *QE-6* (1970), p. 317.
31. J. P. J. Heemskerk et al, Proceedings of IEEE/OSA Conf. Laser Eng. and Appl., Wash. D.C. (June, 1979).
32. W. T. Tsang, R. A. Logan, and M. Ilegems, Appl. Phys. Lett., *32* (1978), p. 311.
33. M. Born and E. Wolf, *Principals of Optics*, 4th ed. New York: Pergamon Press, 1970, Section 1.6. Theoretical TE and TM reflectivity-versus-angle curves were calculated from the one-layer and two-layer M-matrix formalism.

34. F. K. Reinhardt, I. Hayashi, and M. B. Panish, J. Appl. Phys., *42* (1971), p. 4466.
35. T. Ikegami, IEEE J. Quantum Electronics, *QE-8* (1972), p. 470.
36. E. I. Gordon, IEEE J. Quantum Electronics, *QE-9* (1973), p. 772.
37. B. W. Hakki and F. R. Nash, J. Appl. Phys., *45* (1974), p. 3907.
38. R. C. Miller and W. B. Joyce, Appl. Phys. Lett., *31* (1977), p. 764.
39. W. B. Joyce and R. W. Dixon, J. Appl. Phys., *46* (1975), p. 855. A junction-temperature increase of 34° C per watt of input electrical power was calculated by Joyce for our laser structure.
40. Lord Rayleigh, Sci. Papers *3* (1902), p. 441.
41. A. E. Bell and F. W. Spong, IEEE J. Quantum Electron., *QE-14* (July, 1978), p. 487.
42. A. J. deVries, Rec. Trav. Chim., *77* (1958), pp. 383 and 441.
43. J. Bohdansky, J. Chem. Phys., *49* (1968), p. 2982.
44. J. W. Rutter in *Liquid Metals and Solidification*, Chicago: Am. Soc. for Metals, 1958, p. 243.
45. Yu F. Komnik, Fiz. metal. metalloved. *25* (1968), p. 817.
46. R. Hultgren, R. L. Orr, P. D. Anderson, and K. K. Kelley, *Selected Values of Thermodynamic Properties of Metals and Alloys*, Am. Soc. Metals, 1973; also R. C. Weast, ed., *Handbook of Chemistry and Physics*, 52nd ed., Chemical Rubber Publishing Co., 1971.
47. R. Berry, P. Hall and M. Harris, *Thin Film Technology*, New York: Van Nostrand, 1968, p. 265.
48. D. Benjamin and C. Weaver, Proc. Roy. Soc., *A254* (1960a), p. 163.
49. F. Abeles, *Advances in Optical Technology*, A. C. S. Van Heel, ed., Amsterdam: North Holland, 1967, p. 144.
50. Values of bulk optical constants may be obtained from the following sources:
    Bi: A. P. Lenham, D. M. Treherne, and R. J. Metcalfe, J. Opt. Soc. Am., *55* (1965), p. 1072; J. N. Hodgson, Proc. Phys. Soc. (London), *B67* (1954), p. 269.
    Sb: Lenham et al, ibid.
    Cd, Zn, Mg: R. H. W. Graves and A. P. Lenham, J. Opt. Soc. Am., *58* (1968), p. 58.
    Pb: A. I. Golorashkin, JETP (USSR), *48* (1965), p. 825; Soviet Phys. JETP, *21* (1965), p. 548.
    In: I. N. Shkylyarevskii and R. G. Yarovaya, Optika i Spek., *16* (1964), p. 85; Optics and Spectrosc., *16* (1964), p. 45.
    Sn: A. I. Golorashkin and G. P. Motulevich, JETP (USSR), *47* (1964), p. 64.
    Te: J. Stuke and H. Keller, Phys. Stat. Sol., *7* (1964), p. 189.
    Al: L. G. Schult, J. Opt. Soc. Am., *44* (1954), pp. 357 and 362; J. N. Hodgson, Proc. Phys. Soc. (London), *B68* (1955), p. 593.
51. M. Hansen, *Constitution of Binary Alloys*, New York: McGraw Hill, 1958.
52. Our refractive-index dispersion curve for evaporated Bi_2O_3 is intermediate between the results of P. B. Clapham, Brit. J. Appl. Phys., *18* (1967), p. 363, for sputtered Bi_2O_3 and of J. Halford and H. Hacker, Jr., Thin Solid Films, *4* (1969), p. 265, for evaporated Bi_2O_3.
53. R. A. Bartolini, A. E. Bell, and F. W. Spong, Proc. IEEE/OSA Conf. on Laser Eng. and Appl., Wash., D. C. (June 1979).
54. The bulk optical constants of selenium may be obtained from C. Froissart, C. R. Acad. Sc. Paris, *270B* (1970), p. 1544.
55. P. Chaudhari and B. Beaver, Trans. Met. Soc. *239* (1967), p. 501.
56. R. G. Crystal, J. Poly. Sci. (A-2) *8* (1970), p. 2153.
57. W. J. Smith, *Modern Optical Engineering*, New York: McGraw-Hill, 1966, Sec. 13.6.
58. H. Kita, I. Kitano, T. Uchida, and M. Furukawa, NEC Research and Development, *25* (1972), p. 21.
59. The optical path calculations were carried out by P. E. Naufnagel, Bausch and Lomb Optics Center, Rochester, New York.
60. W. J. Smith, Ref. 57, p. 131, Fig. 6.9.
61. A. L. Buck, Proc. IEEE, *55* (1967), p. 448.
62. W. J. Smith, Ref. 57, p. 434, Fig. 14.15.

Copyright © 1979 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 58, No. 9, November 1979
Printed in U.S.A.

Sequencer Designs for Scanning-Beam Satellites

By W. L. ARANGUREN, R. E. LANGSETH, and C. B. WOODWORTH

(Manuscript received April 24, 1979)

We present conceptual designs for the controlling sequencer that would be required on board a communication satellite employing a scanning phased-array antenna. Two basic forms are discussed: one providing separate pointing control for each time slot in the TDMA *frame and the other providing a compact storage arrangement when the typical beam dwell time is several time slots. The update interface to the telemetry link is discussed briefly. The impending availability of radiation-resistant* CMOS *memories makes the control part of a large (100-element) scanning array feasible in the 1980 time frame, with an estimated power consumption under 50 watts.*

I. INTRODUCTION

Recently D. O. Reudink and Y. S. Yeh[1] proposed a scanning spot-beam antenna to provide high-gain, area coverage in a satellite communication system. The most efficient method of implementing the scanning function appears to be by way of a phased array together with electronically-controlled phase shifters (separate sets for transmit and receive). These phase shifters are scanned in a cyclic fashion to provide the necessary interconnection between ground stations.

If the scanning were sufficiently slow, it would be possible to transmit the ongoing phase shifter information to the satellite from a control earth station. However, the downlink beam may shift position every microsecond or so.[1] With a 100-element array and 3-bit phase shifters per element, about 300 megabits per second would be required for real-time phase control. Obviously, then, the satellite must store the necessary information to control the basic scanning with only *changes* transmitted by the control station. In this paper we describe efficient methods for organizing this storage on board the satellite, with consideration given to power consumption by some representative memory devices currently available. Additionally, we indicate how selective phaser stepping (or dithering), which may be required for initial beam forming, can be accomplished.

II. BASIC CONTROLLER

Figure 1 shows the basic N element phased-array beam former and phase controller; separate shifters and control memories will be necessary for transmit and receive functions. As described in Ref. 1, there might be M ($\approx$100) distinct beams, each requiring a distinct set of N phases, θ_n.

Let us assume that each phase, θ_n, must be stored using b bits, to provide for acceptable sidelobe degradation as compared to purely analog phases; three bits may be adequate for this purpose.[1] Thus, $N \cdot M \cdot b$ bits of onboard storage must be provided to define the basic beam-number-to-phase mapping. For reasons of reliability, this storage module should be arranged as N subunits of Mb bits each, where each subunit is used to control the phase shifters of one element.

The control function is thus reduced to selecting, in some sequence, the Nb bits which form a given beam, maintaining these bits on the phase shifters for the desired duration of the given up- or downlink beam, moving onto the next beam by selecting a new set of Nb bits, and so on. Thus, a basic controller has the general form sketched in Fig. 2, wherein a sequencer, to be described shortly, controls $\log_2 M$ binary address lines to permit selection among the M possible beams. Note that because uplinks and downlinks scan separately over the M

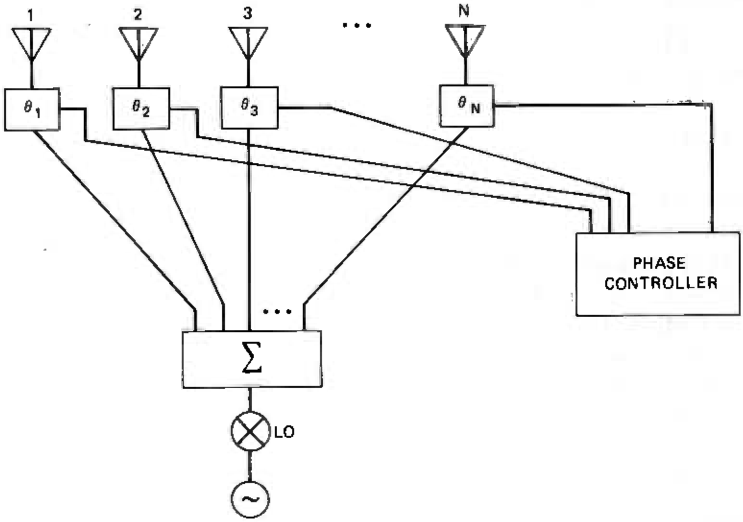


Fig. 1—A phased array with digital phase shifters.

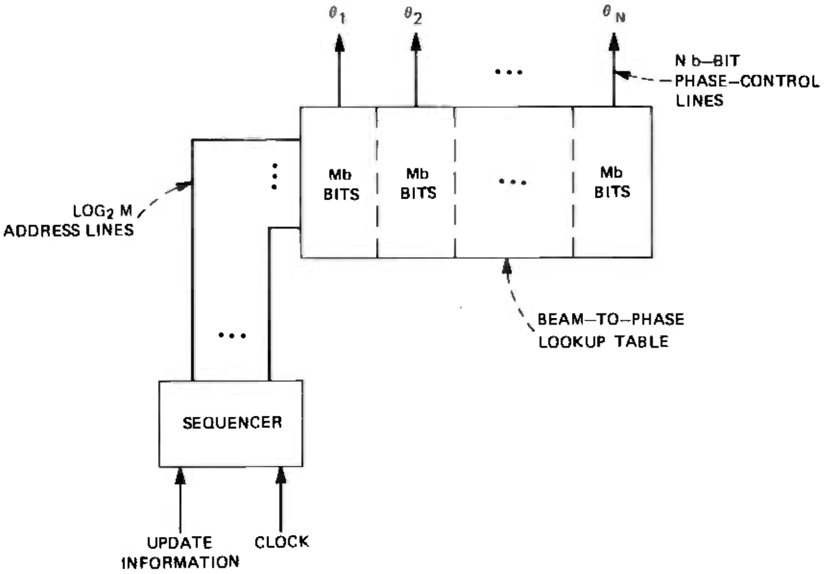


Fig. 2—Generic phase controller.

beams, two such controllers will be on the satellite, each with its own beam-to-phase lookup memory and sequencer. The update input provides for reconfiguration, which may provide for demand assignment, if desired.

The heart of the controller is thus the sequencer which contains the information to control the actual duration of each beam and the sequence with which they occur (the up- and downlink sequences being distinct). By updating the sequencers, variations in traffic patterns can be accommodated. In the next section, three basic forms of the sequencer are presented. The first form is designed to permit fine-grain timing control, with a resolution of one basic time slot. The second form is convenient when each beam position is maintained for several time slots, which permits a compact storage arrangement for the sequence of durations. This will probably be primarily true of the uplink, which remains fixed while the downlink is distributing data among as many as $M - 1$ other beams. The complexity of the two methods are similar for downlink control. The third design provides improved demand-assignment flexibility with limited memory size.

III. SEQUENCERS

The first form of the sequencer provides a from-to assignment for each individual satellite circuit. That is, both the uplink and downlink are capable of being reconfigured on a per-time-slot basis. This allows the system to be used in a fixed assignment, a demand assignment, or a mixed mode of operation.

Figure 3 presents a block diagram of this sequencer. A single ÷ K counter drives a look-up table which provides the beam number for each time slot. Thus, each step of the counter defines a single time slot. For an M-beam system, the memory requirement for each table would be a block of K words of $\mathrm{Log}_2 M$ bits each. For example, seven memory ICs of 16,384 × 1 bits each would be capable of distributing 16,384 circuits among 128 beams.

This configuration of using one memory IC per bit of output word lends itself to block error-correcting codes. That is, we are assuming the failure mode would involve a single bit at a time (not entire words), which would be the case of a single memory IC failure. Four additional bits, and thus four additional memories, would be required to use a Hamming single error-correcting code, for example.

In the second form of sequencer implementation, we observe that the sequence of beams is uniquely determined by (i) how long to stay with the current beam and (ii) which beam is next. In the downlink case, these numbers are also a function of which beam is currently active on the uplink. Consider first the uplink sequencer. With M beams, there can be $M - 1$ "next beams," each of which can be coded by $\log_2 (M - 1) \approx \log_2 M$ bits. Thus, a memory of M words, each of $\log_2 M$ bits suffices to define the sequence of "next beams." For a satellite system with up to $M = 128$ beams, a 128 × 7 memory suffices. This memory can be implemented as one integrated circuit 128 × 8 RAM (random access memory). This same memory organization can be used for the duration of each beam, providing for a 128:1 range of duration.

This sequencer is sketched in Fig. 4. Initially, we may latch beam "zero" into the next beam latch, and its duration into the down

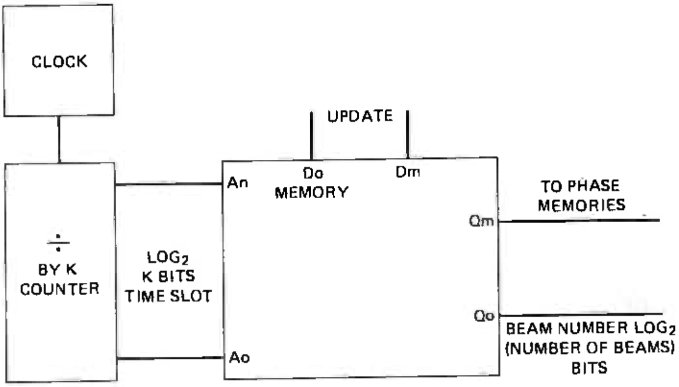


Fig. 3—A sequencer that maps individual time slots to desired beam number.

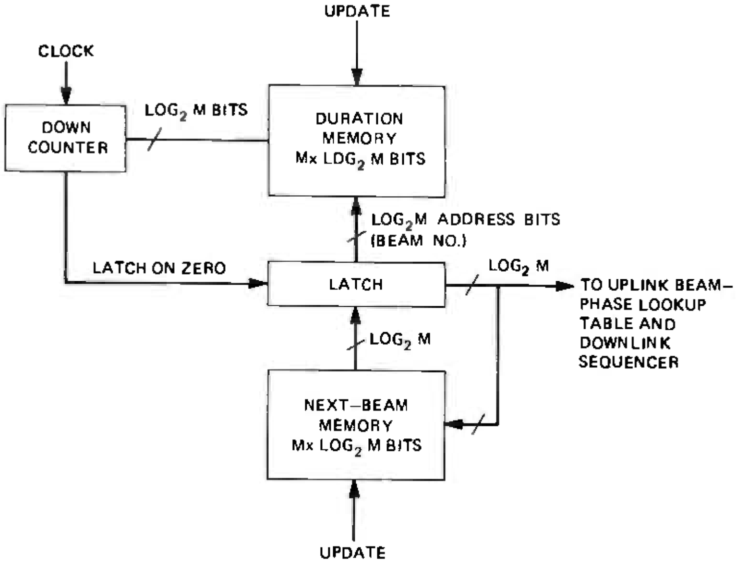

UPDATE
CLOCK
DOWN COUNTER
LOG_2 M BITS
DURATION MEMORY Mx LOG_2 M BITS
LOG_2M ADDRESS BITS (BEAM NO.)
LATCH ON ZERO
LATCH
LOG_2 M
TO UPLINK BEAM-PHASE LOOKUP TABLE AND DOWNLINK SEQUENCER
LOG_2 M
NEXT-BEAM MEMORY Mx LOG_2 M BITS
UPDATE

Fig. 4—Uplink sequencer.

counter. The contents of the next beam latch are used to address the beam-phase lookup table. As the counter decreases, the same beam is maintained until the counter reaches zero. At this time, the binary-coded address of the next beam number is latched, the new duration is loaded into the counter, and the process repeats.

A similar sequencer may be used to form the downlink beams, but it will have to be M times larger, since the sequence of downlink "next-beams" and durations will probably depend on which uplink beam is active. Thus, instead of $\log_2 M$ address lines for these two memories, $2 \log_2 M$ lines are required, where $\log_2 M$ of them came from the uplink sequencer. This configuration is sketched in Fig. 5. Note that the output of this sequencer is still only $\log_2 M$ bits to address the M sets of Nb bits defining the M downlink beams. With $M = 128$, each memory of Fig. 5 would be 16,384 × 7 bits, which can be implemented with seven chips. In practice, one would probably use the method of Fig. 3 for the downlink, since it could be implemented with a single set of seven 16K × 1 memories.

In a demand assignment situation, it may be desirable to be able to return to a beam number several times within a superframe. For example, there may be a fixed block of time slots assigned at the beginning of the superframe followed by a pool for demand assignment. As a result, a ground station may be bursting up several times within a superframe.

With the previous implementation of the sequencer, it is impossible

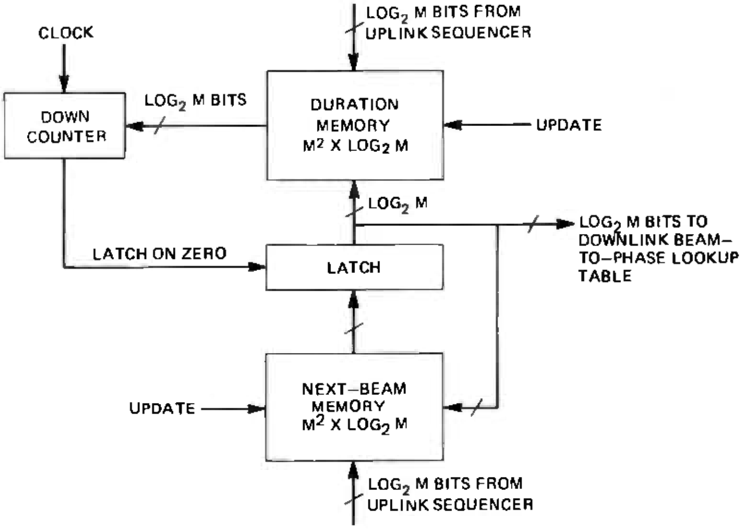

CLOCK
DOWN COUNTER
LOG_2 M BITS
LOG_2 M BITS FROM UPLINK SEQUENCER
DURATION MEMORY M^2 X LOG_2 M
UPDATE
LOG_2 M
LOG_2 M BITS TO DOWNLINK BEAM–TO–PHASE LOOKUP TABLE
LATCH ON ZERO
LATCH
UPDATE
NEXT–BEAM MEMORY M^2 X LOG_2 M
LOG_2 M BITS FROM UPLINK SEQUENCER

Fig. 5—Downlink sequencer.

to return to any beam number within a superframe without becoming caught in a loop. Figure 6 represents an impossible sequence. When beam number 6 is latched the second time, the next beam number should be 2 to continue in the sequence. This is impossible, since the next beam memory is addressed by the current beam number which is 6, and this location already has a next beam of 1. Thus, the remaining sequence would be beam 1 and beam 6 repeated until the end of the superframe. Also, with the previous proposal, the number of bits in the data word of the duration memory sets the maximum number of time slots per beam per superframe, since any beam number can only be entered once in the next beam memory.

A sequencer that can have repeated beam numbers in the beam memory, thus allowing demand assignment and unlimited time for each beam, is shown in Fig. 7. Instead of addressing the next beam memory and duration memories by the current beam number, a sequence number addresses a current beam memory and a duration memory. At the beginning of a superframe, the sequence counter is reset. This addresses the first beam number which is latched, and the duration number corresponding to that beam number is loaded into a counter which acts as a timer. On timeout, the sequence number is incremented and the next beam number, which may be the same as the first, is latched for a new duration. Figure 8 displays a typical sequence in operation. Here 50 units represent the maximum duration; beams requiring more time can be repeated, as in the case of beams 0 and 6 in the figure.

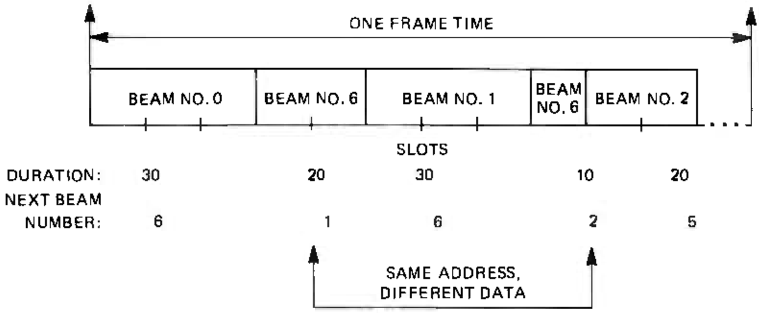


Fig. 6—Impossible frame picture from sequencer in Fig. 4. Next beam number and duration for beam 6 are not unique.

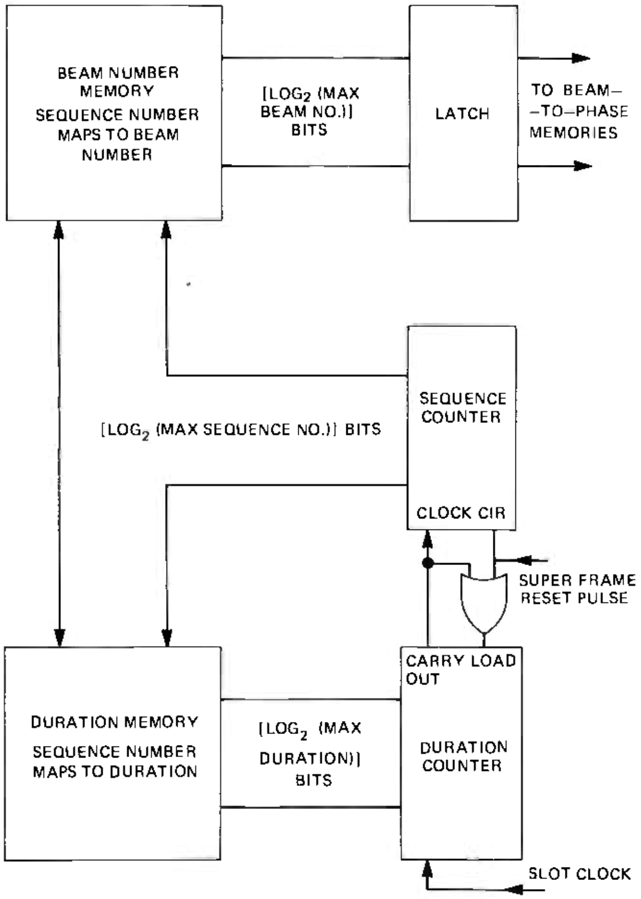

BEAM NUMBER MEMORY
SEQUENCE NUMBER MAPS TO BEAM NUMBER
[LOG2 (MAX BEAM NO.)] BITS
LATCH
TO BEAM–TO–PHASE MEMORIES
[LOG2 (MAX SEQUENCE NO.)] BITS
SEQUENCE COUNTER
CLOCK CIR
SUPER FRAME RESET PULSE
CARRY OUT
LOAD
DURATION MEMORY
SEQUENCE NUMBER MAPS TO DURATION
[LOG2 (MAX DURATION)] BITS
DURATION COUNTER
SLOT CLOCK

Fig. 7—New uplink sequencer which allows repeated beam numbers.

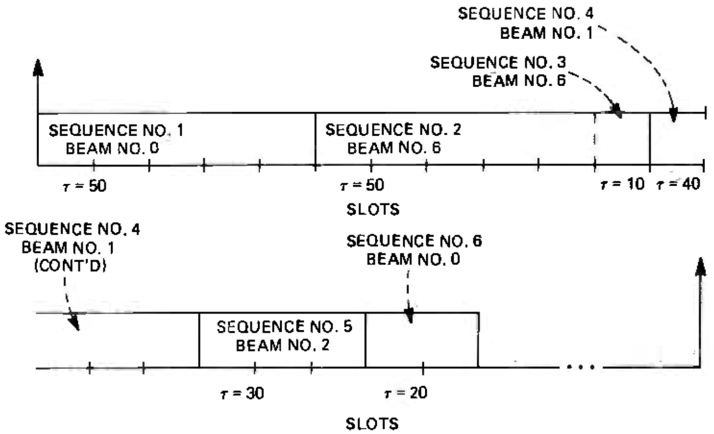


Fig. 8—A typical frame picture of the new uplink sequencer in Fig. 7.

The bus size of the beam memory is $\log_2$ of the maximum sequence number by $\log_2$ of the maximum number of beams. The duration memory must be $\log_2$ of the maximum sequence number by $\log_2$ of the maximum duration per sequence number.

Just as the "uplink" sequencer of Fig. 4 was generalized to the downlink sequencer of Fig. 5, so the sequencer of Fig. 7 can be generalized to a downlink counterpart, as shown in Fig. 9. Here the contents of the uplink sequence counter are used to form a portion of the address for the downlink duration memory and sequence-to-beam number memory. A separate duration counter is incremented by the basic slot clock; its carry-out is used to increment the downlink sequence counter. Each time the uplink sequence counter is incremented by the carry from the uplink duration counter, the downlink sequence counter is reset to begin counting the downlink sequence for the new uplink beam number, using a new set of duration and beam numbers.

Given several possible sequencer designs, it is natural to ask when it is reasonable to use each design. The answer basically depends on two things: the number of slots per TDMA frame, and the number of beams in the satellite system. Consider a TDMA frame length of 25 ms, such as is discussed in Ref. 2 and 3. A basic slot may be around 2.5 μs in length, permitting capacity to be assigned in units of from one voice circuit (a slot repeated once per frame) on up. Thus, there are 10,000 slots per frame. Let us consider two exemplary systems, one with 100 distinct beam positions and one with 20.

In the former case, there would be, on average, 100 slots assigned to each uplink beam with one slot assigned per downlink for each uplink. In this case, the compact memory designs of Fig. 4 or Fig. 7 would be

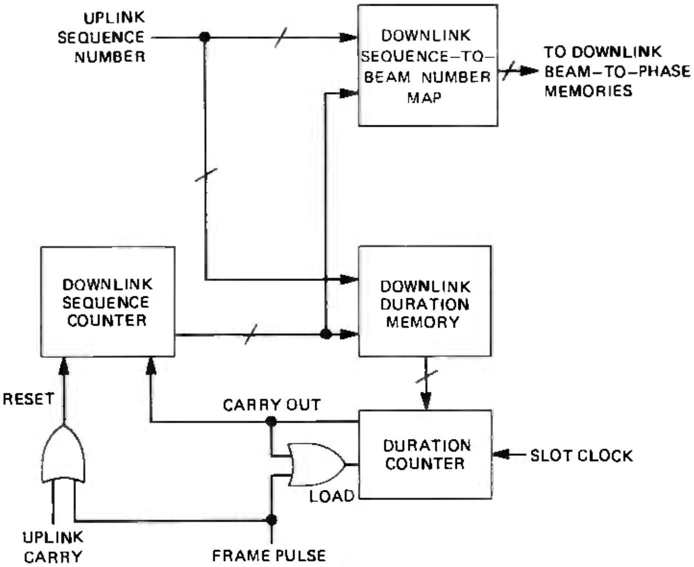


Fig. 9—Downlink sequencer allowing repeated beam numbers.

used for the uplink sequencer using a 7-bit sequence counter while the expanded design of Fig. 3 could be used on the downlink using a 14-bit slot counter. In the latter situation of a system using 20-beam positions, around 500 slots would be assigned to each uplink beam, with about 25 slots assigned to each downlink beam for each uplink beam. In this case, both up- and downlink sequencers could be of the compact designs, either Fig. 4 or Fig. 7 for the uplink, and Fig. 5 or Fig. 9 on the downlink. Here we would need, perhaps, a 5-bit sequence counter together with a 9-bit uplink duration memory, along with a 5-bit downlink sequence counter and a 5-bit downlink duration memory and counter in Fig. 9.

No matter what sequencer design is chosen, it must permit modification of the memory contents to permit changes in the scanning sequence to accommodate changes in traffic patterns; these changes may be seasonal, daily, hourly, or perhaps more frequent. In any case, the sequencers must be able to receive update information without interrupting the sequence in progress. This involves writing update information into memories that may be accessed each time slot. One way of accomplishing this would be to provide two copies of each memory, one of which is active at any time. Update information would be written into the second copy, with the second taking over upon command, probably at the beginning of some frame.

Alternatively, update can be accomplished by either using a read-modify-write cycle or by separate read-and-write cycles for each time slot. A read-modify-write cycle has the advantage of being faster, thus allowing smaller slot times, but it requires that the update address match the current sequencer or phase memory address. State-of-the-art memories are sufficiently fast for separate read-and-write cycles per slot at slot times presently proposed. Dual access allows any location to be updated anytime and has simpler timing requirements.

Since the beam-to-phase memories also must be updated without interruption, the same technique can also be used to update both the sequencers and the beam-to-phase memories. Figure 10 shows an interlaced update scheme using the dual access technique. While the sequencer memories are doing a beam lookup, the beam-to-phase memories are in an update memory cycle. Once the lookup is complete, the beam number is now available to access the beam-to-phase memories. At this time, the sequencer can receive update information. Using this method, the sequencer could be completely updated in a single frame if necessary. With this flexibility, demand assignment

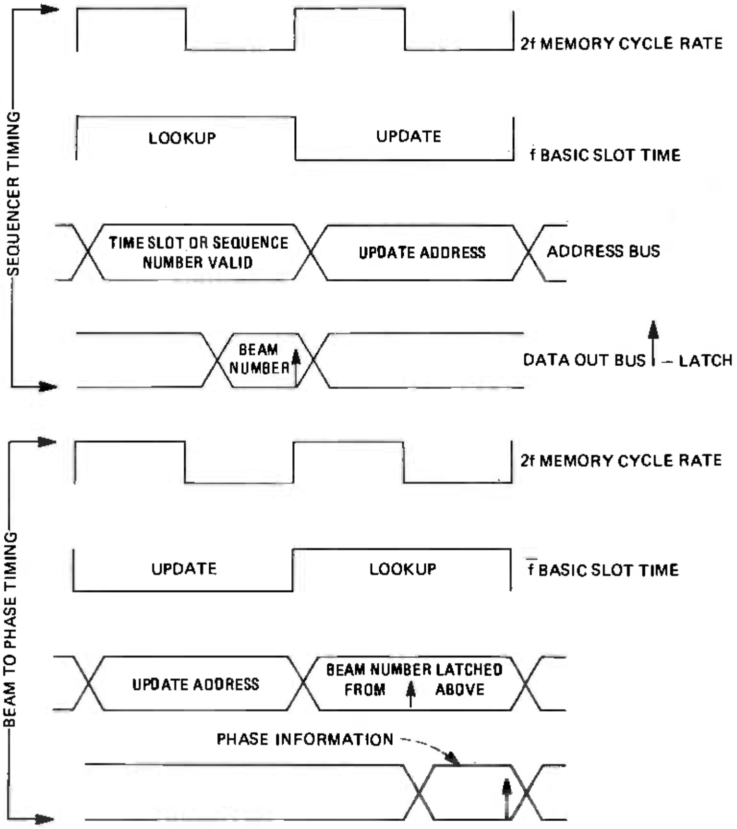


Fig. 10—Interlaced update timing diagram.

could be done on a call-by-call basis if desired, provided the uplink telemetry channel had adequate capacity.

IV. POWER REQUIREMENTS

Power on board a satellite is a precious commodity. We are thus concerned about the total power required by the scanning controller. From previous discussion, it is apparent that the beam-to-phase lookup table will probably be the largest power consumer, since it could be constituted of up to 100 separate memories. Typical devices that are useful here are 256 × 4 bit CMOS RAMs. These devices have access times under 0.5 μs, which is adequate, since a typical slot-time in the satellite system's transmission frame would be around 1 to 2 μs. Typical power requirements for these devices appears to be in the 50- to 100-mW range. Thus, the basic lookup table for 100 elements should require between 5 and 10 W, not including a few miscellaneous MSI and SSI circuits, whose power consumption would amount to another watt or so.

The second major component is the sequencer, for which several designs have been discussed. Consider Fig. 3 first. The slot to beam-number mapping sequencer might be implemented with seven 16K × 1 RAMs, providing for up to 128 beams. The actual satellite transponder capacity may be around 10,000 slots (voice circuits), assuming 64 kb/s per voice circuit and a transponder bit rate of around 600 Mb/s. Available 16K × 1 static RAMs have maximum power dissipations of around 500 mW. Considering that such memories require some additional SSI circuitry for address decoding, etc., we may allocate around 5 W for this form of sequencer, which would most likely be used in the downlink. The design of Fig. 9 would be similar.

Moving on to Fig. 4 or Fig. 7, the major components here are the duration and next beam memories. For up to 128 beams, and 256:1 range of uplink beam durations, each of these could be implemented as single 128 × 8 MOS static RAM, with a maximum power requirement of around 600 mW. Adding the usual SSI peripheral chips suggests that this sequencer would consume between 1.5 and 2 W.

We may, thus, construct the following estimated power budget for both up- and down-link phase controllers:

Beam-to-phase Lookup (2)	20 watts
Uplink sequencer (Fig. 4 or Fig. 7)	2
Downlink sequencer (Fig. 3)	5
	27 watts

Reliability considerations suggest that redundancy must be added to the sequencer control memories. Although this has not been investigated in detail, we know, for example, that by adding eight additional 16K × 1 RAMs to the seven already proposed for the downlink sequencer, the sequencer can be made immune to the failure of any two of the memories.[4] (This assumes, of course, that a highly reliable error-correcting decoder converts the 15 redundant bits to the desired seven control bits.) Thus, we may estimate that a phase controller with redundancy will require more like 30 to 35 W of dc power. This amount of power would appear to place no major additional requirement on the overall satellite power generation.

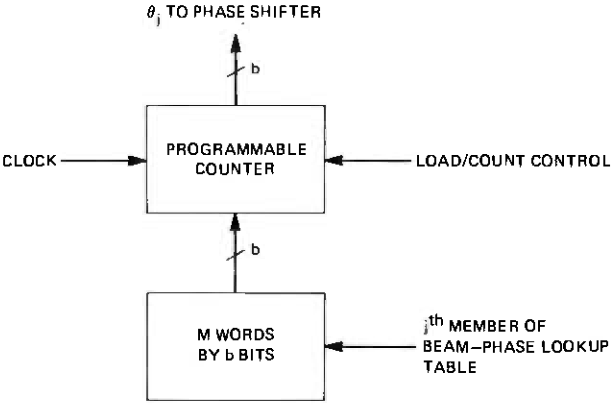
Since we are placing rather heavy reliance on CMOS memories, care must be taken in their design to ensure survivability in the radiation environment of geostationary orbit. At the present time, CMOS radiation tolerance is insufficient for a 7-year on-orbit lifetime. However, considerable effort is under way in the aerospace industry to improve this situation,[5] so that it is highly likely that adequate CMOS memories will be available in a short time.

V. PHASE DITHER

It may be necessary to provide some closed-loop means by which it can be ascertained that a given beam in fact points where it is supposed to. Reasons for this concern include the slow, uncorrected satellite motion that occurs between station-keeping maneuvers, which may cause small but significant pointing error, particularly for beamwidths under one degree. To track this motion, the basic beam-to-phase mapping will require occasional updating. As suggested by Y. S. Yeh, the update information can be obtained by sequentially tagging the array elements with a phase dither, by which means some (or perhaps all) the earth stations can measure the relative element phases. By transmitting the information to a control station, corrective updating of the satellite downlink beam-phase lookup can be effected as necessary. (A similar updating of the uplink lookup table can be done with measurements made on the satellite.)

The dither approach basically requires that, during a preassigned subframe, the phase shifter of one of the array elements be periodically stepped, providing a stepwise linear phase progression, essentially equivalent to a frequency offset of that element. At the same time, a fixed reference element is stepped in the opposite direction. By proper processing at an earth station, the relative element phase can be estimated.

A simple method of adding this capability to the phase controller of Fig. 1 is to follow each of the N array-element phase memories with a so-called programmable up/down counter, as shown in Fig. 11 for a typical element. During normal operation, the counting is inhibited, and the normal data from the memory are "programmed" through the



θ_j TO PHASE SHIFTER
b
CLOCK
PROGRAMMABLE COUNTER
LOAD/COUNT CONTROL
b
M WORDS BY b BITS
j^{th} MEMBER OF BEAM–PHASE LOOKUP TABLE

Fig. 11—How to provide phase dither.

counter and simply appear on the output for normal operation. By selectively enabling the counter mode, the phase on a given element will be stepped by the clock, implementing the desired tagging.

VI. CONCLUSIONS

We have discussed several designs for the controller sequencer that would be required on board a communication satellite employing a scanning spot beam. By making use of LSI memory technology, compact and reasonably low power controllers are feasible. Two different memory organizations were described, one providing for a different beam position in each time slot of the TDMA frame, while the other provides a more compact memory arrangement for those situations wherein typical beam dwell times are many slots. For large arrays (about 100 elements), it is necessary to use a low-power technology such as CMOS to store the individual element phases; it appears that a radiation-resistant version of this technology should be available by the late 1980s.

REFERENCES

1. D. O. Reudink and Y. S. Yeh, "A Rapid Scan Spot-Beam Satellite System," B.S.T.J., *56*, No. 8 (October 1977), pp. 1549–60.
2. D. O. Reudink and Y. S. Yeh, "The Organization and Synchronization of a Switched Spot-Beam System," Conf. Record of 4th Internat'l Conf. on Dig. Sat. Comm., Montreal, October 1978, pp. 191–6.
3. A. S. Acampora and R. E. Langseth, "Baseband Processing in a High-speed Burst Modem for a Satellite Switched TDMA System," Conf. Record of 4th Internat'l Conf. on Dig. Sat. Comm., Montreal, October 1978, pp. 131–8.
4. S. Lin, *An Introduction to Error-Correcting Codes*, Englewood Cliffs, N.J.: Prentice-Hall, 1970, Section 6.1.
5. P. J. Klass, "New Microcircuits Resist Radiation," Aviation Week and Space Technology, October 10, 1977, pp. 58–61.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 58, No. 9, November 1979
Printed in U.S.A.

Some Theoretical Observations on Spread-Spectrum Communications

By J. E. MAZO

(Manuscript received May 25, 1979)

We consider a simple direct-sequence model of spread-spectrum communications over a bandwidth W using signals of duration T. In addition to the dimension $n = 2WT$ of signal space, the other parameters of interest are the number $(u + 1)$ of simultaneous users of the system, the error rate $Pe(u)$, and the number M of subscribers, or potential users. We investigate the relationships between these parameters, and, in particular, study the validity of the usual Gaussian approximation often used to compute $Pe(u)$. Basically, we conclude that, if M is larger than $n^2/2$, then the Gaussian approximation is not a guaranteed error rate for $(u + 1)$ users, but rather an average over all possible $(u + 1)$ users. If M is somewhat less than this number (the exact value is not known), codes can be assigned so that a uniform performance at least as good as $Pe(u)$ can be obtained, where, again, $Pe(u)$ is calculated from the Gaussian approximation. Uniform performance guarantees are given for any value of M, but they (for M large) permit fewer simultaneous users that the Gaussian approximation predicts. These bounds explicitly use the maximum cross-correlation between the signals of the different subscribers. This quantity played no role in the Gaussian approximation.

I. INTRODUCTION

In general, spread-spectrum communications refers to a class of modulation methods by which the information-bearing signal is transmitted via a modulated signal having much greater bandwidth. Two common methods are used to accomplish the spreading. In one method, direct-sequence modulation, the information signal is multiplied by a rapidly varying waveform. This waveform, which the receiver is required to know, may be thought of as having a pseudo-random character. For practical reasons, it has finite duration and is repeated in time. We refer to any particular suitable waveform, or a collection of such, as a code. The other method for spectrum spreading is frequency hopping. In this case, the available transmission band is divided into a large number of disjoint frequency intervals, and the information is conveyed by hopping from one such frequency to another. The information may be transmitted by phase-shift-keying each frequency or by using a particular set of frequencies for a particular symbol, etc. Again the code for frequency hopping must be known to the receiver.

The initial motivation for introducing such a scheme appears to be in its military use as an anti-jamming device. The jammer, not knowing the transmitter's code for spectrum spreading, must thus blanket all codes. Most of the jammer's power is wasted in codes that are orthogonal to the one in actual use.

An additional application of more commercial interest was introduced by Costas.[1] His idea was to use spread spectrum as a way to make a large bandwidth, W, available as a communication resource to many potential users without preassigning frequency divided channels (FDM) (and thus overlimiting the number of potential users) and without having a dynamic assignment of FDM (thus incurring the need and cost of external control). In modern work, this is usually accomplished, or imagined to be accomplished, by assigning "almost orthogonal" codes, or code vectors, to different users as a means to limit the mutual interference between users.

Very recently, attention has been drawn to spread spectrum as a possible modulation method for cellular mobile radio systems.[2] Our interest was drawn to this area by Henry's subsequent criticism[3] of the analysis of Cooper and Nettleton.[2] While the particular question in this controversy appears to have been resolved (in Henry's favor), our own survey of the situation has brought out some deeper questions relating to assumptions made in the analysis by Cooper, Nettleton, and Henry, and often made elsewhere. Specifically, lip service is often paid to making the codes approximately orthogonal. Yet when the performance analysis is finally made for these digital systems, one typical user is considered and all other users which are simultaneously using the channel are treated as interfering Gaussian noise having uniform power spectrum over the band of interest. Nowhere does any measure of the approximate orthogonality of the signals of different users enter the performance estimates. Our objective, then, is to examine what validity can be given to performance curves calculated using the Gaussian approximation and what relation, if any, this approximation has to the idea of approximate orthogonality of different users' signals.

To gain some insight into these questions, we examine a simple direct-sequence system designed to transmit binary data and give upper bounds or the error rate $\mathrm{Pe}(u)$ that a user will experience when there are u other users on the channel. Assuming that the bounds yield a good description, we can reach some easily stated conclusions.

Let there be M subscribers or potential users of the system and let there be $(u + 1)$ simultaneous users, $u + 1 \leq M$. Also denote the calculated error rate by $\mathrm{Pe}(u)$. Then, by a random coding argument, we conclude for $M = u + 1$ that there is a code such that the $(u + 1)$ users each have error rate $\mathrm{Pe}(u)$, where $\mathrm{Pe}(u)$ is calculated via the Gaussian approximation. This, however, does not take into account the possibility that the $u + 1$ users may be selected from $M > u + 1$ subscribers. The random coding argument also suggests that, in fact, no limit need be placed on M if $\mathrm{Pe}(u)$ is estimated from the Gaussian approximation. However, the interpretation given to $\mathrm{Pe}(u)$ immediately changes. It is no longer an error rate for each of $(u + 1)$ users, but is an *average* error rate where the average is taken over all ways $(u + 1)$ users are selected from the subscriber population.

Next, the question of a *guaranteed* error rate for any $(u + 1)$ users is taken up. That is, we present a performance bound valid for each of the $(u + 1)$ users which is independent of how the $(u + 1)$ users are selected from the population of size M. The results show that if, in our n-dimensional signal space, we are packing so many unit energy signal vectors (corresponding to different subscribers) that the cross-correlations (cosines between vectors) are required to be as large as $(t/n)^{1/2}$ in magnitude, where $t > 1$ is simply a convenient parameter, then the number of simultaneous users is reduced by a factor of t compared to what the Gaussian approximation would predict. Finally, recent bounds by Kabatyovskii and Levenshtein for sphere-packing problems are applied to give upper bounds on M so that the $(t/n)^{1/2}$ bound on the cross-correlations can be met. For $t = 1$, the upper bound is $n^2/2$ vectors, which, for n large, is quite generous. How closely this can be approached is not known.

II. MODEL AND ANALYSIS

We consider the following simple model of direct-sequence, binary, spread-spectrum communication. We have a bandwidth W over which $u + 1$ independent users simultaneously communicate binary information with a central station at the rate $R = 1/T$ b/s. The individual information rates are small compared with the available bandwidth; thus, $TW >> 1$. There are M subscribers or potential users of the system and M may be large compared with $u + 1$, the maximum number of simultaneous users allowed. Each potential user is permanently assigned a coded carrier (or code vector). The ith subscriber's carrier $c_i(t)$ is written as

$$c_i(t) = \sum_{k=1}^{n} x_k^{(i)} \psi_k(t), \tag{1}$$

where $\psi_k(t)$ are orthonormal basis functions in the $n = 2WT$ dimensional space of functions approximately limited to W Hz in bandwidth and to T s in duration.

The entire system operates synchronously and every T seconds the ith user square-wave modulates his carrier by ±1, the independent, identically distributed (i·i·d) binary data that he wishes to transmit.

We may consider the n real number $x_k^{(i)}$, $k = 1, \cdots, n$ forming a vector $\mathbf{x}^{(i)}$ in real n space and the collection of vectors $\{\mathbf{x}^{(i)}\}_1^M$ belonging to the different subscribers is sometimes called a code. We only consider equal energy codes and set

$$E = \int_0^T c_i^2(t)\, dt = \sum_{k=1}^{n} (x_k^{(i)})^2. \tag{2}$$

In the analysis, we distinguish the user whose performance we will be interested in by the subscript $i = 1$; the other users are designated by $i = 2, \cdots, u + 1$. Thus, in a typical T-second interval the received signal will be a time translate of

$$b_1 c_1(t) + \sum_{i=2}^{u+1} b_i c_i(t), \tag{3}$$

the b_i being the binary data, i·i·d for each user and also between users.† We assume correlation detection of (3), and thus the receiver bases its decision of b_1 on the sign of

$$b_1 \int_0^T c_1^2(t)\, dt + \sum_{i=2}^{u+1} \int_0^T c_1(t) c_i(t)\, dt = Eb_1 + E \sum_{i=2}^{u+1} b_i \rho_{1i}. \tag{4}$$

In (4) we have introduced the normalized cross-correlation,

$$\rho_{ij} = \frac{\int_0^T c_i(t) c_j(t)\, dt}{E} = \sum_{k=1}^{n} x_k^{(i)} x_k^{(j)}. \tag{5}$$

This equals the cosine of the angle between $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(j)}$ in n-space. Thus, when $|\rho_{ij}|$ is small, the vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ are almost orthogonal.

Assume that $b_1 = 1$ in (4). Then the probability of error, Pe, is

$$\text{Pe} = \Pr\left[1 + \sum_{i=2}^{u+1} b_i \rho_{1i} < 0 \right]. \tag{6}$$

† Such a system where the "spreading" function is modulated by the data is usually termed a direct-sequence system. Frequency hopping is another spread-spectrum technique. The relative practicality of the two techniques depends on particular circumstances.

This depends on the distribution of the noise-like quantity

$$q = \sum_{i=2}^{u+1} b_i \rho_{1i}, \tag{7}$$

where (assuming unit energy)

$$\rho_{1i} = \sum_{k=1}^{n} x_k^{(i)} x_k^{(1)}. \tag{8}$$

First let us consider not a specific code, but rather an average over all possible codes assuming all the unit vectors $\mathbf{x}^{(i)}$ are uniformly and independently distributed over the unit sphere.† In this case, note (letting $\langle\,\rangle$ denote an average)

$$\langle \rho_{ij}^2 \rangle = \frac{1}{n} \quad i \neq j, \tag{9}$$

which follows from noting that, for a unit vector, the sum of the squares of its n direction cosines add to 1, while each must have the same average.

Using the random coding assumption, the moment-generating function $M_q(s)$ of q may be shown to be

$$M_q(s) \equiv \langle e^{sq} \rangle$$

$$= \langle e^{sv} \rangle^u = \left[\frac{1}{B(1/2, (n-1)/2)} \int_{-1}^{1} e^{sv}(1 - v^2)^{n-3/2}\, dv \right]^u, \tag{10}$$

where, in (10), $B(1/2, (n-1)/2)$ is the beta function and

$$p(v) = \frac{1}{B(1/2, (n-1)/2)} (1 - v^2)^{(n-3)/2} \tag{11}$$

is the probability density of any direction cosine v of a vector uniformly distributed over a sphere in n-dimensions. A saddle point evaluation of (10) yields, for $n \gg 3$ and s/n small,

$$M_v(s) \approx e^{s^2/2n}. \tag{12}$$

Since the Chernoff bound[4] states that, for any random variable q,

$$\Pr[q < A] \leq e^{-sA} M_q(s) \quad \text{for any} \quad s < 0, \tag{13}$$

then

$$\mathrm{Pe} = \Pr[q < -1] \leq e^{s} M_q(s) \approx e^{((s^2 u/2n)+s)}. \tag{14}$$

† We could also assume that $x_k^{(i)} = \pm 1$, i·i·d, and obtain similar numbers.

Optimizing the inequality in (12) over s yields[†] $s_{\text{opt}} = -n/u$, and thus

$$\text{Pe}(u) \le e^{-(n/2u)}, \tag{15}$$

where $\text{Pe}(u)$ is the error rate for $(u + 1)$ simultaneous users.

We begin our discussion of (15) by reconsidering the performance question from a different point of view. Assume that the interfering power resulting from users 2 through $u + 1$ is uniformly distributed over the band W. If each user has power P, then the one-sided power spectral density N_0 thus obtained is

$$N_0 = \frac{uP}{W} = \frac{uE}{TW}. \tag{16}$$

Further assume this noise is Gaussian. As is well known, the error rate for antipodal signals, each of energy E (this is what our transmitters are using), is, in white noise of spectral-density N_0, given by

$$\begin{gathered} \text{Pe} = Q\left(\sqrt{\frac{2E}{N_0}}\right) < e^{-(E/N_0)}, \\ Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-y^2/2}\, dy. \end{gathered} \tag{17}$$

The bound in (17) is exponentially correct. Using (16), the Gaussian approximation yields

$$\text{Pe}(u) \le e^{-(TW/u)} = e^{-(n/2u)}, \tag{18}$$

where we have introduced the dimension $n = 2WT$. This is precisely the same as the random coding bound (15). We use the random coding argument to interpret the result of the standard Gaussian approximation. The first interpretation is that there must be a code of $(u + 1)$ vectors so that each of the $(u + 1)$ users has error rate $\text{Pe}(u)$.[‡] A second interpretation is that, to achieve $\text{Pe}(u)$, no limit need be placed on the number M of subscribers. They can, in fact, be assigned codes at random. The average error rate a user sees with u other users present is (15). But (15) clearly then refers to a $\text{Pe}(u)$ averaged over all possible combinations of u users; it is not an error rate that can be met for any set of u other users. Some combinations of $u + 1$ users will give very bad error rates.

To see what the guaranteed level of performance can be for any $u + 1$ users selected out of a subscriber population of size M, reconsider

[†] The fact that $s_{\text{opt}} = -n/u$ justifies treating s/n small in (10).

[‡] More precisely, one can guarantee that a code exists so that a fraction $(1 - (1/\alpha))$ of the users has an error rate no larger than $\alpha \cdot \text{Pe}(u)$.

(6) with the ρ_{1i} fixed. Then the standard Chernoff bound yields

$$\mathrm{Pe} \le \exp\left[-\frac{1}{2}\left(\sum_{i=2}^{u+1} \rho_{1i}^2\right)^{-1}\right]. \tag{19}$$

Suppose we have designed our code so that, out of the entire population of M subscribers,

$$\max_{i,j,i\neq j} |\rho_{ij}| = \rho_{\max}.$$

Then from (19),

$$\mathrm{Pe} \le e^{(-1/2u\rho_{\max}^2)}. \tag{20}$$

If $\rho_{\max}^2 = 1/n$, the Gaussian approximation again results. In random coding $\langle\rho_{ij}^2\rangle = 1/n$, but nothing was said about $\rho_{\max}$ and thus nothing could be said about a guaranteed error rate.

Given M unit vectors in n space, how small can $\rho_{\max}$ be? One result in this direction is due to Welch.[5] He states that

$$\rho_{\max}^2 \ge \frac{1}{M-1}\left[\frac{M}{n} - 1\right]. \tag{21}$$

If n is large and M large compared to n, then (21) already states $\rho_{\max}^2 \ge \frac{1}{n}$.[†]

Rephrasing our question, given the dimension n and $\rho_{\max}$, how large can M be? That is, how many vectors $\mathbf{u}_i$ can we put on the unit sphere so that

$$|\rho_{ij}| = |\mathbf{u}_i \cdot \mathbf{u}_j| \le \rho_{\max}? \tag{22}$$

This sphere-packing problem is different from the conventional one which requires

$$\rho_{ij} = \mathbf{u}_i \cdot \mathbf{u}_j \le \rho_{\max}. \tag{23}$$

The number of vectors M will be much smaller under condition (22) than under condition (23). Luckily, our sphere-packing problem (22) was one of the packing problems recently considered by Kabatyonskii and Levenshtein.[6] Precise values of M are not known, but upper bounds are.

An extension of their work provides us with the following. For $k = 0, 1, 2, \cdots$,

[†] If $(n+1)$ vectors are the vertices of the regular simplex in n-space, then $\rho_{\max}^2 = 1/n^2$. Then (21) is exact for $M = n$ and $M = n + 1$.

$$M \le \frac{(1-\rho_{\max}^2)}{(2k+1)-(n+2k)\rho_{\max}^2}(n+2k)\frac{(n/2)_k}{(1/2)_k}, \qquad (24)$$

provided the denominator in (24) is positive.[†] In (24), we have used the notation

$$(\alpha)_k = \alpha(\alpha+1)\cdots(\alpha+k-1)$$
$$(\alpha)_0 = 1. \qquad (25)$$

It is useful to write $\rho_{\max}^2$ as

$$\rho_{\max}^2 = \frac{t}{n}. \qquad (26)$$

If we set $t = k =$ integer and use the same k in (24), and further assume $k^2 \ll n$, we have approximately

$$M \lesssim \frac{1}{(k+1)}\frac{n^{k+1}}{(2k-1)!!}. \qquad (27)$$

Equation (24) is plotted in Fig. 1 for $n = 100$ as a function of t. The part of the curve for smallest t uses $k = 0$, the next set of t values uses $k = 1$, and so on.

In general, the value of the upper bound at $t = 1$ (where the Gaussian assumption agrees with random coding) is given by $n^2/2$. How closely this upper bound can be achieved is not clear. If, however, $t = 4$, a result of van Lint and interpreted for the present situation by Welch,[5] indicates that at least 10^4 binary (± 1) waveforms are available, compared with the bound of 4×10^5. The point is indicated by the small circle in Fig. 1.

III. CONCLUSIONS

Consideration of our simplified model has provided the following insights. Treating many other users as background Gaussian noise for a particular user is a good approximation in that a code can be found which (at least approximately) provides the calculated error rate for each simultaneous user. This is no bargain, however, for the orthogonal code would be even better in performance in the present problem, permitting more simultaneous users.[‡] Evaluated as a pure modulation

[†] The bound (24) is valid for a real vector space. For a complex space, the following larger bound applies ($\rho = \rho_{\max}$):

$$M_{\text{complex}} \le \frac{1-\rho^2}{k+1-(n+k)\rho^2}\frac{(n)_{k+1}}{k!}.$$

[‡] More explicitly, (18) states that, for a small error rate, we require $2u \ll n = 2WT$, or $u \ll WT = W/R$. However, in this model, FDM with double-sideband modulation allows about W/R users with zero error rate.

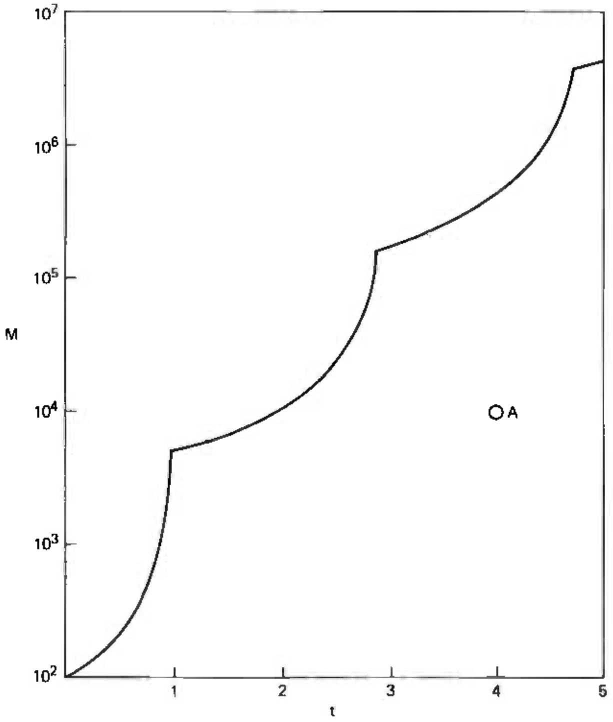


Fig. 1—Upper bound on number of signals M vs $t = n\rho^2_{\max}$ for $n = 100$. The point A is known to be achievable using binary codes.

method for accommodating (u + 1) known and fixed users in a white-noise channel, spread spectrum has little to offer over SSB-FDM. The only advantage for spread spectrum on the type of channel that we have considered would be as a multi-access scheme for M potential users.[†] Selecting codes at random for this situation still permits us to use the Gaussian approximation with u "other" simultaneous users to calculate an error rate Pe(u), but only in an average sense. Namely, it is an average over all possible (u + 1) users selected out of the M subscribers. To guarantee a level of performance for any (u + 1) users, the departure from strict orthogonality must not be too severe, and this puts a definite restriction on the number M of potential users.

[†] The potential absence of channel assignment for spread spectrum is the basis of the Cooper-Nettleton (Ref. 2) proposal for spread spectrum for mobile radio. Other considerations may make spread spectrum an attractive alternative. For example, it can provide frequency diversity for a frequency-hopped system when there is fading. Such a scheme has been proposed by Goodman et al. (Ref. 7) for mobile radio, modifying Viterbi's (Ref. 8) proposal for satellites. Also in satellite systems, the Doppler shift can be large compared to data rates for individual users. Viterbi (Ref. 8) has suggested that spread spectrum would not need the large guard bands between channels that FDM would require.

While upper bounds on M were given in the text, the exact value is not known, nor, in general, did we discuss explicit construction of suboptimum codes.

IV. ACKNOWLEDGMENTS

It is a pleasure to acknowledge several helpful general discussions on spread-spectrum communications with H. E. Rowé. The author is particularly indebted to N. J. A. Sloane for alerting him to Ref. 6.

APPENDIX

Derivation of Eq. (24)

A brief outline is presented to guide the reader who wishes to rederive (24) from the results of Ref. 6.

We first summarize the relevant results of Ref. 6, given on pages 10 and 11 of that work. We will be concerned with the interval $-1 \le t \le 1$ and expansions in Jacobi polynomials $P_i^{\alpha\beta}(t)$, i denoting the degree of the polynomial. The parameters α and β depend on the dimensionality n and type (real or complex) of space that we are considering. For a real space $\alpha = (n - 3)/2$, $\beta = -1/2$ $(n \ge 3)$, while for a complex space $\alpha = n - 2$, $\beta = 0$ $(n \ge 2)$. Let s be a real number $-1 \le s < 1$ and denote by $R(\alpha, \beta, s)$ the set of polynomials

$$f(t) = \sum_{i=0}^{l} f_i P_i^{\alpha\beta}(t) \tag{28}$$

of degree $l = 1, 2, \cdots$ such that

$$(i) \qquad f_i \ge 0, \qquad i = 0, \cdots, k \quad \text{where} \quad f_0 > 0. \tag{29a}$$

$$(ii) \qquad f(t) \le 0 \quad \text{for} \quad -1 \le t \le s. \tag{29b}$$

Then the maximum number M of unit vectors $\mathbf{u}_i$ in n-space such that

$$|\mathbf{u}_i \cdot \mathbf{u}_j| \le \rho$$

satisfies

$$M \le \inf_{f(t)\in R(\alpha,\beta,2\rho^2-1)} \frac{f(1)}{f_0}. \tag{30}$$

The evaluation of $f(1)/f_0$ for any particular allowed $f(t)$ further upper bounds (30). We choose the polynomials

$$f(t) = (t - s)(t + 1)^k \qquad k = 0, 1, \cdots. \tag{31}$$

The verification of (29b) is thus trivial, as is the evaluation of $f(1)$. The verification of (29a) as well as the evaluation of f_0 is a direct calculation. We evaluate f_i using the orthogonality properties of the Jacobi polynomials $P_i^{\alpha\beta}(t)$ with respect to the weight function $w(t) = (1 - t)^\alpha(1 + t)^\beta$. The required normalization may be found in Ref. 9, p. 262, formula 1. If we rewrite (31) as

$$f(t) = (t + 1)^{k+1} - (1 + s)(t + 1)^k, \tag{32}$$

then the integrals needed to evaluate f_i may be calculated from Ref. 8, p. 263, formula 3. One then directly verifies that $f_i > 0$ $(i > 0)$ is positive whenever f_0 is, and the indicated evaluation of f_0 results in (24) for the real case.

REFERENCES

1. J. P. Costas, "Poisson, Shannon and the Radio Amateur," Proc. IRE, *47* (December 1959), pp. 2058–2068.
2. G. R. Cooper and R. W. Nettleton, "A Spread-Spectrum Technique for High-Capacity Mobile Communications," IEEE Trans. Vech. Tech., *VT-27*, No. 4 (November 1978), pp. 264–275.
3. P. S. Henry, "Spectrum Efficiency of a Frequency-Hopped-DPSK-Mobile Radio System," 29th IEEE Vehicular Technology Conference Record, March 1979, pp. 7–12.
4. Robert G. Gallager, *Information Theory and Reliable Communication*, New York: John Wiley, 1968. See pp. 126–128 for the Chernoff bound.
5. L. R. Welch, "Lower Bounds on the Maximum Cross Correlation of Signals," IEEE Trans. Info. Theory, *IT-20*, No. 3 (May 1974), pp. 397–399.
6. G. A. Kabatyanskii and V. I. Levenshtein, "Bounds for Packings on a Sphere and in Space," Problems of Information Transmission, *14*, No. 1 (January–March 1978), pp. 1–17.
7. D. J. Goodman, P. S. Henry, and V. K. Prabhu, "Preliminary Assessment of Multilevel FSK for Mobile Radio," unpublished work.
8. A. J. Viterbi, "A Processing Satellite Transponder for Multiple Access by Low-Rate Mobile Users," Proc. Digital Satellite Commun. Conf., Montreal, October 23–25, 1978.
9. Peter Beckmann, *Orthogonal Polynomials for Engineers and Physicists*, Boulder: Golem Press, 1973.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 58, No. 9, November 1979
Printed in U.S.A.

Transistor Surface Effects Responsible for Anomalous Third-Order Intermodulation Distortion in Undersea Cable Telephone System

By D. G. DUFF and H. M. OLSON

(Manuscript received June 12, 1979)

Anomalous intermodulation distortion has been observed in the SG submarine cable repeater amplifier. The anomalous distortion was found to be dependent upon the crystal orientation in the SG bipolar transistor. This fact suggested that transistor surface effects might be responsible for the anomalous distortion. To test this hypothesis, computer models of the transistor with surface effects were used to compute the third-order modulation coefficient M_3. Surface effects in the collector-base junction alone were modeled by a junction-controlled MOS capacitor with fast surface states. The MOS capacitor model then was added to a modified Gummel-Poon transistor model to simulate surface effects in the complete transistor. It was found that intermodulation contributed by the surface effects was dominant in the case of ⟨111⟩ orientation, and the behavior of the computed intermodulation was similar to the observed behavior. This paper describes the transistor modeling and the intermodulation computations developed in testing the hypothesis. The principal conclusion confirms that the anomalous distortion from the SG transistor is due to surface effects, in particular, to a high surface-state density.

I. INTRODUCTION

The deviation from perfect linearity in a repeater amplifier for a high-capacity, long-haul system is very small by normal standards. However, when the signal passes through many repeaters, the effect of nonlinear distortion can accumulate to a significant value. The resulting modulation noise of many signals is similar to white noise and is an important factor in the specification of such system parameters as repeater spacing and power-handling capability. Permissible third-order harmonic distortion at the output of a repeater, for a milliwatt of fundamental, is typically 10^{-10} to 10^{-13} mW.

Measurements of the third-order intermodulation product for the SG submarine cable amplifier showed unusual behavior as power was increased.[1] The deviation from the expected 3-dB increase in third-order power for a 1-dB increase in the three fundamental powers was largest for the case where the fundamental frequencies were slightly lower than the product frequency, as illustrated in Fig. 1. Distortion measurements on the open loop amplifier produced similar results, so the feedback network was not a cause. When the bias voltage and current were varied separately for each of the signal-carrying bipolar transistors in the amplifier, the collector voltage nonlinearity showed

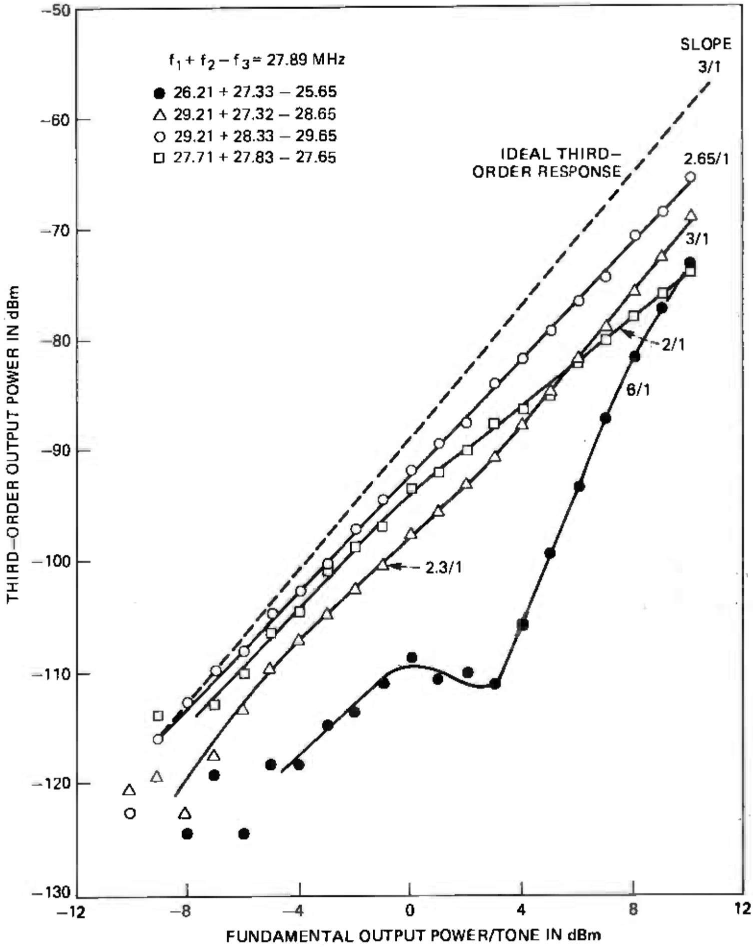


Fig. 1—Distortion in the closed loop SG amplifier.

the most unusual behavior. Two types of silicon transistors with identical geometry but different epitaxial crystal orientations, ⟨111⟩ and ⟨100⟩, were tested in the amplifier. The transistors with ⟨111⟩ orientation showed the unusual distortion characteristics described above, while the transistors with ⟨100⟩ orientation showed classical performance. The distortion performance of the collector-base junction alone showed unusual performance for the ⟨111⟩ orientation and normal performance for the ⟨100⟩ orientation, as shown in Fig. 2.

These measurements suggested that the collector-base junction was the prime source of the unusual distortion. The strong dependence of modulation coefficient M_3 on crystal orientation suggested surface effects as the basic cause of the unusual intermodulation, since surface charge is known to be dependent upon crystal orientation. The SG transistor chip is shown in Fig. 3. The base contact metalization overlaps the collector region with an oxide layer for isolation. This forms a pn junction-controlled MOS capacitor. It is known that the MOS capacitor across the base-collector junction can produce a degradation of transistor distortion.[2] However, the measured junction capacitance for the SG transistors was well-behaved and showed none of the anomalous ripples as a function of voltage, as is expected from large concentrations of surface states. The M_3 measurement of the collector-base junction is a more sensitive test than capacitance measurement and clearly showed anomalous behavior for ⟨111⟩ junctions.

This paper shows that the unusual distortion measured for the collector-base junction of the SG transistor is accurately modeled for both the ⟨111⟩ and ⟨100⟩ crystal orientations by a junction-controlled MOS capacitor with fast surface states. The MOS capacitor model is added to a modified Gummel-Poon transistor model, including the effects of base pushout, lateral spreading, and internal heating to simulate the measured distortion for an SG transistor.

The principal conclusion of this work has been to confirm the hypothesis that the anomalous distortion from the SG submarine cable

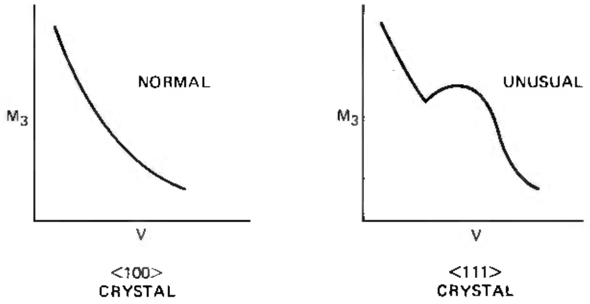


Fig. 2—M_3 vs reverse bias for the collector-base junctions of SG transistors.

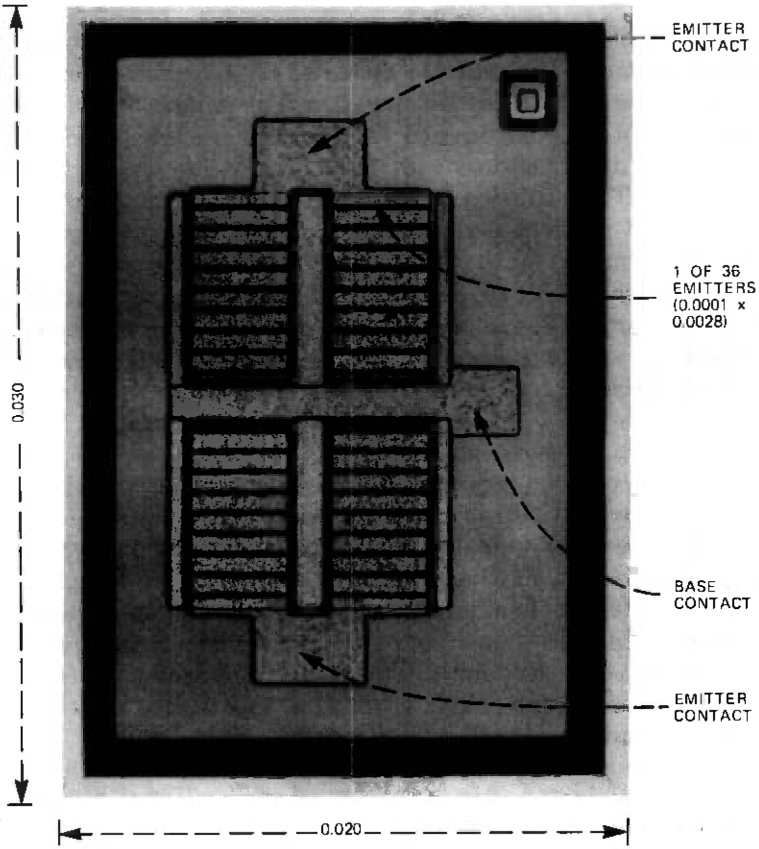


Fig. 3—Transistor chip.

transistor is due to surface effects, in particular to a high surface-state density. Consequently, it is important in designing transistors for applications where intermodulation distortion is critical to take special pains to control and minimize surface effects. This may involve using silicon wafers that are ⟨100⟩ oriented as well as minimizing the collector-base overlap capacitance. Or it may involve special processing steps such as:

(*i*) Routine characterization of transistor surfaces for surface-state and fixed charge densities.

(*ii*) Special annealing or gettering steps to reduce the number of surface states.

Appendix E contains a list of symbols used throughout the paper.

II. DESCRIPTION OF COLLECTOR-BASE JUNCTION MODEL

The transistor is made up of two base regions separated by a base pad, as shown in Fig. 3. Insulated by an SiO_2-SiN layer, the base pad extends over the collector up to the collector-base junction on each side, as can be seen in Fig. 4. This forms a junction-controlled MOS capacitor with the voltage applied to the junction appearing also across the MOS capacitor, as shown in Fig. 5.[3]

2.1 Nonlinear small-signal ac model

A rather complicated ac circuit model for a junction-controlled MOS capacitor[4] is shown in Fig. 6. C_j represents the junction capacitance between the p-base and n-collector, C_{ox} the oxide capacitance, and C_d the capacitance between the semiconductor-oxide interface and the n-type bulk. C_s models charge storage in the fast interface states at the oxide-semiconductor interface. The resistances characterize the following physical processes:

R_{ns} and R_{ps}—The transition of electrons from the conduction band and holes from the valence band into surface states.

R_{nD} and R_{pD}—The flow of electrons and holes through the surface space-charge layer.

R_{pB}—The generation (or recombination) of holes and their flow rate

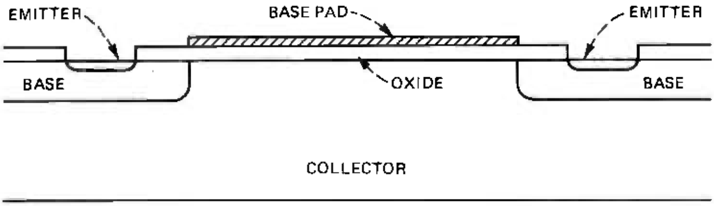


Fig. 4—Cross section of SG transistor.

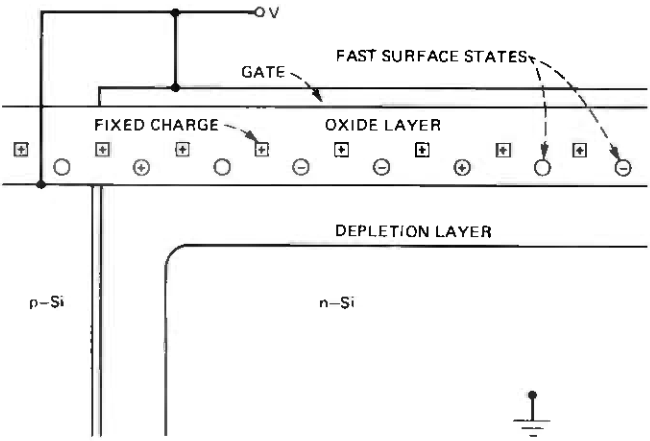


Fig. 5—Schematic of junction-controlled MOS capacitor.

in the region of quasi-neutrality in the bulk adjacent to the space-charge layer.

R_0—The flow of electrons and holes through the bulk of the semiconductor. For an n-type collector, R_I models the average drop in voltage from the base to the semiconductor-oxide interface and helps define the surface quasi-fermi level for holes ψ_{Fp}. The inversion charge is stored in C_I and minority carriers can flow to the surface from the bulk through R_p or through R_I. Holes may be trapped by surface states through R_{ps} and stored in C_s. However, for an n-type collector with positive fixed charge in the oxide, the action of the pn junction on the MOS capacitor is such as to inhibit the semiconductor surface from becoming inverted.

Since electrons are the important carriers for operation in the depletion region, the equivalent circuit shown in Fig. 6 has been

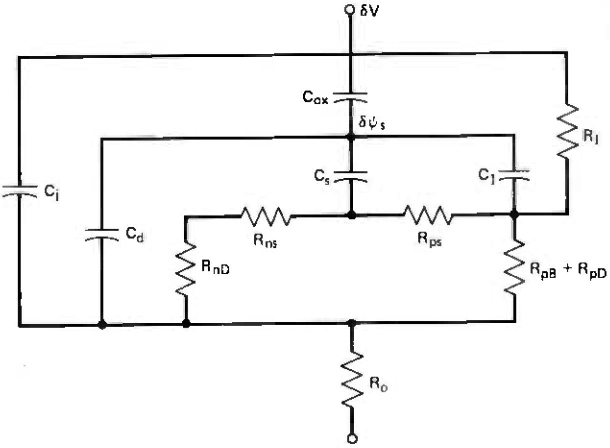


Fig. 6—Alternating-current circuit model for a junction-controlled MOS capacitor on n-type semiconductor.

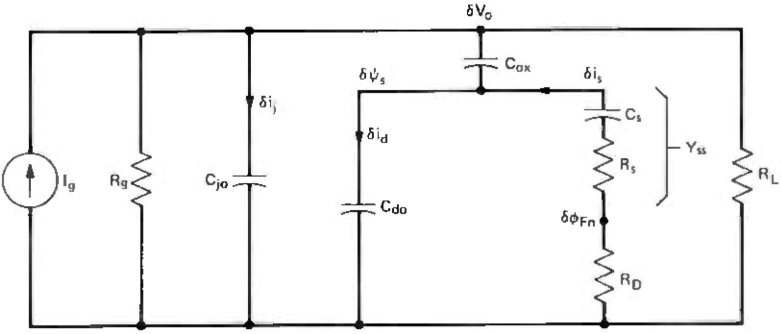


Fig. 7—Simplified ac circuit model of a junction-controlled MOS capacitor.

simplified to that in Fig. 7. Electrons are supplied to the surface from the bulk through R_D which is given by Ref. 4:

$$R_D = \frac{1}{q\mu_n A}\int_0^d \frac{dx}{n_0(x)}, \tag{1}$$

where

μ_n = electron mobility
n_0 = dc electron concentration
d = depletion layer width
A = surface area.

The R_s, C_s network in Fig. 7 is a simple small-signal model for capture and emission of electrons by surface states all located at one energy level within the bandgap. In reality, they are spread throughout the bandgap, and the admittance for a continuum of states of density $N_s(E)$ in the bandgap is given by Ref. 5:

$$Y_{ss} = j\omega\frac{q}{kT}\int_{E_V}^{E_C}\frac{N_s(E)f_0(1-f_0)}{1+\dfrac{j\omega f_0}{c_n n_{s0}}}dE, \tag{2}$$

where

$f_0 = [1 + g_s\exp(u - u_F)]^{-1}$ is the fermi function at potential u for a quasi-fermi potential u_F expressed in units of the thermal voltage kT/q
g_s = ground state degeneracy (= 2 for donors, = 1/4 for acceptors)
c_n = electron capture probability
n_{s0} = electron density at the silicon surface established by the bias
E_V, E_C = energy levels at the edges of the valence and conduction bands.

Some authors have argued that the integrand of (2) peaks sharply about the quasi-fermi level with a width of about kT.[6] This would make it easy to integrate (2) if $N_s(E)$ and c_n do not vary much with E over a range of kT. However as is shown in Appendix A, this is the case only at low frequency and high semiconductor doping levels to which these intermodulation computations are not restricted. To preserve accuracy, therefore, integrals of the type exemplified by (2) have been evaluated numerically in this work.

In depletion, the time-constant dispersion measured for MOS capacitors is much broader than that predicted from (2). Good agreement is obtained if one modifies (2) by assuming a statistical fluctuation of surface potential in the plane of the interface.[6] As a first attempt to study intermodulation due to surface states, the surface admittance is rederived to retain nonlinear terms up to third order for a continuum of surface states. The statistical fluctuation of surface states is taken into account via correction terms derived in Appendix B.

III. COMPUTER ALGORITHMS FOR INTERMODULATION CALCULATIONS

3.1 Volterra series representation

We are interested in computing the third-order modulation coefficient for the circuit of Fig. 7. A convenient way of doing this is by expressing the current through the nonlinear element as a Volterra series of the voltage across the element.[7] If the element is memoryless (purely resistive), the Volterra series simplifies to a power series:

$$i(t) = h_1 v(t) + h_2[v(t)]^2 + h_3[v(t)]^3 + \cdots. \tag{3}$$

If the element has memory (capacitive), the first three terms of the Volterra series are:

$$i_1(t) = \int_0^t h_1(t-\tau)v(\tau)\,d\tau \tag{4}$$

$$i_2(t) = \int_0^t\int_0^t h_2(t-\tau_1,\, t-\tau_2)\prod_{i=1}^{2} v(\tau_i)\,d\tau_i \tag{5}$$

$$i_3(t) = \int_0^t\int_0^t\int_0^t h_3(t-\tau_1,\, t-\tau_2,\, t-\tau_3)\prod_{i=1}^{3} v(\tau_i)\,d\tau_i. \tag{6}$$

The h's in the integrands are known as Volterra kernels of the first, second, and third degree. In the Fourier transform domain, these relations become

$$I_1(\omega) = H_1(\omega)V(\omega) \tag{7}$$

$$I_2(\omega_1, \omega_2) = H_2(\omega_1, \omega_2)\prod_{i=1}^{2} V(\omega_i) \tag{8}$$

$$I_3(\omega_1, \omega_2, \omega_3) = H_3(\omega_1, \omega_2, \omega_3)\prod_{i=1}^{3} V(\omega_i). \tag{9}$$

The transformed Volterra kernels $H_1(\omega)$, $H_2(\omega_1, \omega_2)$, and $H_3(\omega_1, \omega_2, \omega_3)$ may be found from a knowledge of the physics of the device as described in the following sections.

3.2 Volterra kernels for the surface-state admittance in equilibrium (nonlinear case)

The following derivation is strictly valid only for equilibrium conditions, which is a good approximation for an MOS capacitor in depletion with no surface inversion layer. The incremental current flowing from the conduction band for an n-type semiconductor into donor or acceptor surface states within the energy band dE is given by

$$di_s(t) = -\{qN_s(E)c_n[1 - f(t)]n_s(t) - qN_s(E)e_nf(t)\}\, dE, \tag{10}$$

where

$i_s(t)$ = time-varying current flowing into the surface states
q = electron charge
c_n = electron capture probability
$f(t)$ = fermi distribution function
$n_s(t)$ = surface concentration of electrons
e_n = electron emission probability from the surface state.

The incremental current is also equal to the change in charge per unit time

$$di_s(t) = -qN_s(E)\frac{df}{dt}\, dE. \tag{11}$$

Substitute (11) into (10) to get

$$\frac{df}{dt} = c_n(1 - f)n_s - e_nf. \tag{12}$$

This is a nonlinear equation for f in terms of n_s with memory due to the time derivative. Let n_s be the forcing function. Then one desires the small signal components of f about a dc operating point f_0 for small variations in n_s about n_{s0}. So let

$$f = f_0 + \delta f, \quad n_s = n_{s0} + \delta n_s. \tag{13}$$

Substituting (13) into (12) produces a dc equation and an ac equation.

$$e_n = \frac{c_n(1 - f_0)n_{s0}}{f_0} \quad \text{(dc)}. \tag{14}$$

$$\frac{d}{dt}\delta f = c_n(1 - f_0)\delta n_s - \frac{c_nn_{s0}}{f_0}\delta f - c_n\delta f\delta n_s \quad \text{(ac)}. \tag{15}$$

Since the nonlinearity (third term on the right) has memory, a Volterra series may be used to express δf.

$$\delta f = H_1(\omega_a)\ o\ \delta n_s + H_2(\omega_a, \omega_b)\ o\ \delta n_s^2 + H_3(\omega_a, \omega_b, \omega_c)\ o\ \delta n_s^3, \tag{16}$$

where the driving signal δn_s is assumed to consist of three terms:

$$\delta n_s = a_1 \cos(\omega_1 t + \phi_1) + a_2 \cos(\omega_2 t + \phi_2) + a_3 \cos(\omega_3 t + \phi_3) \tag{17}$$

and the circle operator o indicates that the magnitude and phase of each term in δn_s^n are changed by the magnitude and phase of $H_n(\omega_1, \ldots, \omega_n)$.

The Volterra kernels H_1, H_2, H_3 are derived in Appendix C. The current into the surface states is computed by converting (11) into the frequency domain by using the theorem in Appendix D for differentiating Volterra series and substituting (16) for δf to get

$$di_s(\omega) = -qN_s(E)[j\omega_a H_1(\omega_a) \text{ o } \delta n_s + j(\omega_a + \omega_b)H_2(\omega_a, \omega_b) \text{ o } \delta n_s^2 + j(\omega_a + \omega_b + \omega_c)H_3(\omega_a, \omega_b, \omega_c) \text{ o } \delta n_s^3]\, dE. \quad (18)$$

The total current $i_s(\omega)$ is found by integrating di_s over the band gap, or

$$\begin{aligned} i_s(\omega) &= -j\omega_a q \int_{E_V}^{E_c} N_s(E)H_1(\omega_a)\, dE \text{ o } \delta n_s \\ &\quad - j(\omega_a + \omega_b)q \int_{E_V}^{E_C} N_s(E)H_2(\omega_a, \omega_b)\, dE \text{ o } \delta n_s^2 \\ &\quad - j(\omega_a + \omega_b + \omega_c)q \int_{E_V}^{E_C} N_s(E)H_3(\omega_a, \omega_b, \omega_c)\, dE \text{ o } \delta n_s^3 \\ &\equiv G_1(\omega_a) \text{ o } \delta n_s + G_2(\omega_a, \omega_b) \text{ o } \delta n_s^2 + G_3(\omega_a, \omega_b, \omega_c) \text{ o } \delta n_s^3. \end{aligned} \quad (19)$$

In order to use eq. (19), the surface electron concentration δn_s must be known. It may be computed as a nonlinear function of the surface potential ψ_s and the surface quasi-fermi level for electrons ψ_{Fn} to be

$$n_s = n_i \exp[\psi_s - \psi_{Fn})/V_T]. \quad (20)$$

Using a Taylor series expansion

$$\delta n_s = b_1(\delta\psi_s - \delta\psi_{Fn}) + b_2(\delta\psi_s - \delta\psi_{Fn})^2 + b_3(\delta\psi_s - \delta\psi_{Fn})^3, \quad (21)$$

where

$$b_1 = \frac{n_{s0}}{V_T},\ b_2 = \frac{b_1}{2V_T},\ b_3 = \frac{b_2}{3V_T}.$$

Note that the nonlinearity does not have memory, so a Taylor series expansion is sufficient to describe δn_s. Since $i_s = G(n_s)$ and $n_s = B(\psi_s - \psi_{Fn})$, where B and G are functions from (19) and (20), then $i_s = G[B(\psi_s - \psi_{Fn})] \equiv D(\psi_s - \psi_{Fn})$, where the composite function D has a Volterra series as shown below.[7]

$$i_s = D_1(\omega_a) \text{ o } (\delta\psi_s - \delta\psi_{Fn}) + D_2(\omega_a, \omega_b) \text{ o } (\delta\psi_s - \delta\psi_{Fn})^2 + D_3(\omega_a, \omega_b, \omega_c) \text{ o } (\delta\psi_s - \delta\psi_{Fn})^3, \quad (22)$$

where

$$D_1(\omega_a) = G_1(\omega_a) b_1 \tag{23}$$

$$D_2(\omega_a, \omega_b) = G_1(\omega_a + \omega_b) b_2 + G_2(\omega_a, \omega_b) b_1^2 \tag{24}$$

$$\begin{aligned} D_3(\omega_a, \omega_b, \omega_c) &= G_1(\omega_a + \omega_b + \omega_c) b_3 \\ &\quad + 2\overline{G_2(\omega_a, \omega_b, + \omega_c)}\, b_1 b_2 + G_3(\omega_a, \omega_b, \omega_c) b_1^3 \end{aligned} \tag{25}$$

and

$$\begin{aligned} \overline{G_2(\omega_a, \omega_b + \omega_c)} &= \frac{1}{3} [G_2(\omega_a, \omega_b + \omega_c) + G_2(\omega_b, \omega_a + \omega_c) \\ &\quad + G_2(\omega_c, \omega_a + \omega_b)]. \end{aligned}$$

3.3 Volterra kernels for the collector-base junction capacitance and surface depletion capacitance

These nonlinear capacitances can be represented as Taylor series expansions of the voltage δV_C across the capacitance.

$$C(\delta V_C) = C_0 + C_1\, \delta V_C + C_2 \delta V_C^2. \tag{26}$$

The current i_c through the capacitor can be expressed as

$$i_c = \frac{d}{dt}\left(C_0 \delta V_C + \frac{1}{2}\, C_1 \delta V_c^2 + \frac{1}{3}\, C_2 \delta V_C^3 \right). \tag{27}$$

Converting i_c to the frequency domain gives the Volterra series description

$$\delta I_c = j\omega\left(C_0 \delta V_C + \frac{1}{2}\, C_1 \delta V_C^2 + \frac{1}{3}\, C_2 \delta V_C^3 \right). \tag{28}$$

The depletion capacitance C_d can be found by differentiating the expression for charge in the depletion region (34) with respect to the surface potential. This procedure leads to the following expressions for the depletion capacitance Volterra kernels:

$$H_1(\omega) = j\omega\, \frac{dQ(0)}{dv} = -j\omega K H \tag{29}$$

where

$$H = \exp\,(u_s - u_{Fn}) - \exp\,(u_{Fp} - u_s) + 2 \sinh u_{Fn} \tag{29a}$$

$$K = \frac{q n_i L_D}{G^{1/2}} \tag{29b}$$

$$\begin{aligned} G &= \exp\,(u_s - u_{Fn}) - \exp\,(-u_{Fn}) + \exp\,(u_{Fp} - u_s) - \\ &\quad \exp u_{Fn} + 2u_s \sinh u_{Fn} \end{aligned} \tag{29c}$$

and

$$L_D = \left[\frac{kT}{q}\frac{\epsilon_s}{2qn_i}\right]^{1/2}$$

$$H_2(\omega_1, \omega_2) = j(\omega_1 + \omega_2)\frac{K}{V_T^2}\left[\frac{H^2}{2G} - \exp(u_s - u_{Fn}) - \exp u_{Fp}\right] \quad (30)$$

$$H_3(\omega_1, \omega_2, \omega_3) = -\frac{K}{2V_T}\left\{\frac{3}{4}\frac{H^3}{G^2} - \frac{3}{2}\frac{H}{G}\right.$$

$$[\exp(u_{Fn} - u_s) + \exp(u_s - u_{Fp})] \quad (31)$$

$$\left. + \exp(u_s - u_{Fn}) - \exp(u_{Fp} - u_s)\right\},$$

where V_T is the thermal voltage kT/q.

3.4 Finding the dc solution for the MOS capacitor

Finding the dc solution for the MOS capacitor is simply a matter of finding the potential at the oxide-semiconductor interface corresponding to a prescribed dc bias voltage. The charges existing in the neighborhood of the interface can be expressed in terms of this surface (interface) potential. Since the entire MOS capacitor is electrically neutral, the summation of these charges must vanish. That is,

$$Q_g + Q_s + Q_{ss} + Q_o = F(\psi_s) = 0. \quad (32)$$

The charge on the gate Q_g is given by

$$Q_g = C_{ox}(V - \psi'_{MS} - \psi_s), \quad (33)$$

where V is the applied voltage, ψ'_{MS} the metal-semiconductor work function, ψ_s the surface potential, and C_{ox} the oxide capacitance.

The charge in the semiconductor (space-charge region) is given by Ref. 3:

$$Q_s = 2qn_iL_D[\exp(u_s - u_F) - \exp(-u_F) + \exp(u_F - u_s - v)$$

$$- \exp(u_F - v) + 2u_s \sinh u_F]^{1/2}, \quad (34)$$

where u_s, u_F, and v are the surface potential, electron quasi-fermi potential, and applied voltage expressed in units of the thermal voltage kT/q.

The amount of charge in the surface states depends on the type of state (donor or acceptor), the surface state distribution, and the occupancy of the state as described by the fermi function. Figure 8 shows the charge on the surface states. We have made the conventional assumption that the states above mid-gap are acceptors and those

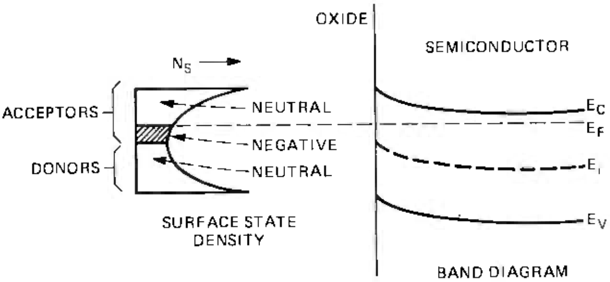


Fig. 8—Charge in surface states.

below mid-gap are donors.[8,9] Acceptors are negative and donors are neutral when filled by an electron. For simplicity, we have assumed that all states below the quasi-fermi level are occupied by electrons and all above are vacant. Thus the charge in the surface states Q_{ss} has been approximated by integrating the surface-state density N_s between u_F and u_s assuming all these states to be charged.

$$Q_{ss} = q \int_{u_F}^{u_s} N_s \, du, \tag{35}$$

Q_o is the fixed charge in the oxide.

The value of u_s which satisfies eq. (32) is the dc solution.

3.5 Computing circuit intermodulation

The modulation coefficient M_3 measured in decibels is defined by Ref. 10:

$$M_3 = 10 \log \frac{P_0(\omega_1 + \omega_2 - \omega_3)}{P_0(\omega_1)P_0(\omega_2)P_0(\omega_3)} - 15.6, \tag{36}$$

where P_0 is the output power measured in milliwatts for an input signal of the form

$$I_g = I_1 \cos \omega_1 t + I_2 \cos \omega_2 t + I_3 \cos \omega_3 t. \tag{37}$$

The perturbation method is used to compute the circuit response at the various combined frequencies as outlined below.

(*i*) Solve for the first-order response at ω_1 by replacing the nonlinear elements in the equivalent circuit (Fig. 7) with the first-order terms of the respective Volterra series, that is, $C_{d0} \; o \; \delta\psi_s$ for δi_d, $C_{j0} \; o \; \delta V_0$ for δi_j, and $D_1 \; o \; \delta\psi_s$ for δi_s. The small signal response δV_0 is then computed for the resulting linear circuit for the input $I_1 \cos \omega_1 t$.

(*ii*) Solve for the first-order response at ω_2 and ω_3 in a similar fashion. Note that D_1 is frequency-dependent, so the first-order equivalent circuit is different for each frequency.

(*iii*) Solve for the second-order response at $\omega_1 + \omega_2$ by removing the I_g generator and inserting intermodulation current generators evaluated from the second-order terms in parallel with the linear-controlled current sources due to the first-order terms (see Fig. 9a). For example, for C_j an intermodulation generator of value $H_{j2} \circ \delta V_{02}^2$ shunts a capacitor of value C_{j0} and for δi_s an intermodulation generator $D_2 \circ \delta\psi_{s2}^2$ is in parallel with the linear generator $D_1 \circ \delta\psi_{s2}$, where both D_1 and D_2 are evaluated at $\omega_1 + \omega_2$. The second-order Volterra kernels require the values of the circuit node voltages at ω_1, ω_2, ω_3 computed in steps (*i*) and (*ii*).

(*iv*) Solve for the response at $\omega_1 - \omega_3$ and $\omega_2 - \omega_3$, as in step (*iii*).

(*v*) Solve for the response at $\omega_1 + \omega_2 - \omega_3$. The second-order intermodulation generators used in steps (*iii*) and (*iv*) are replaced by third-order intermodulation generators (see Fig. 9b). Each third-order generator consists of a pure third-order term plus three second-order interaction terms computed from the responses at $\omega_1 + \omega_2$, $\omega_1 - \omega_3$, $\omega_2 - \omega_3$, ω_1, ω_2, ω_3.

There are two kinds of contributors to the third-order generator currents. The first kind expresses the interaction of a second-order product with a fundamental signal to produce a third-order product [eq. (8)]. The second kind expresses the interaction of three fundamental signals to form a third-order product [eq. (9)]. In terms of

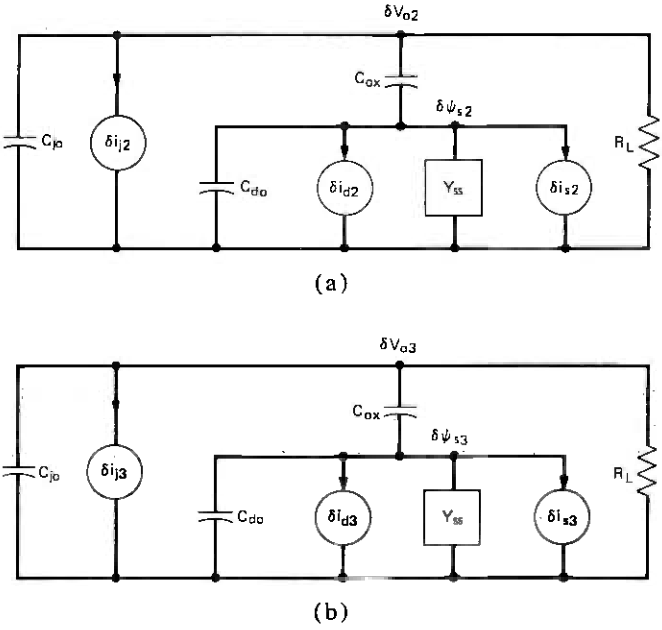


Fig. 9—Equivalent circuits for distortion analysis. (a) Circuit for finding second-order voltages. (b) Circuit for finding third-order voltages.

the output voltages in volts across load resistor R_L, M_3 becomes

$$M_3 = 20 \log \frac{|\delta V_{03}(\omega_1 + \omega_2 - \omega_3)| \, 2 \times 10^{-3} R_L}{|\delta V_{01}(\omega_1)| \, |\delta V_{01}(\omega_2)| \, |\delta V_{01}(\omega_3)|} - 15.6. \qquad (38)$$

The output voltages found in steps (*i*), (*ii*), and (*v*) above are inserted into (38) to find the value of M_3.

IV. DETERMINING SURFACE PARAMETERS

Several parameters are used in this model to characterize the oxide-semiconductor interface. These describe:

(*i*) The distribution of surface states across the semiconductor bandgap.

(*ii*) The capacitance per unit area of the oxide.

(*iii*) The impurity concentration at the surface of the semiconductor.

(*iv*) The electron capture coefficient of the surface states.

(*v*) The fixed charge density in the oxide.

Although for these parameters the literature provides order-of-magnitude values that are representative of this kind of interface, some values are strongly dependent on the specific processing used to grow the oxide. This is particularly true of the surface-state distribution. Techniques for measuring surface-state distribution can be found in the literature. The surface-state density can be inferred, for example, from measurements of the ac conductance of an MOS capacitor.[6]

Measurements of capacitance and conductance were available for MOS capacitors fabricated with the identical oxidation process used for the transistors. These data covered wide ranges of frequency (0.1, 1, 5, 10, and 30 MHz) and bias voltage (from 0 volts into inversion). If appropriate values of the parameters were used, the ac and dc surface model developed for the intermodulation analysis should be able to predict these measurements quite accurately. An optimization program has been used to adjust the parameter values in this model to make it fit the measured admittance data. The surface-state density has been represented by a sixth-degree power series and each of the other parameters by a single average value.

Two small additions had to be made to the equivalent circuit of Fig. 7 to achieve a good fit of the data. A conductance was added in shunt with the oxide capacitance, and a series resistance was inserted at one of the terminals.

The fitting of the data required the use of an adaptive nonlinear least-squares algorithm.[11] A set of 12 parameters, seven of which characterized the surface-state distribution, were optimized by the algorithm. A thirteenth parameter, the electron capture cross section, was held fixed at 5×10^{-16} cm^2. Figures 10 and 11 show the fitting of the conductance and capacitance data for the ⟨111⟩-oriented capacitors by the model. Figures 12 and 13 show the fitting for ⟨100⟩-oriented capacitors. The surface-state distributions resulting from this data fitting appear in Fig. 14; the rest of the parameters are listed in Table I. The surface-state distributions of Fig. 14 appear somewhat higher and have narrower valleys than is usually found for SiO_2-Si interfaces. It is known that the deposition of SiN layers over the SiO_2 layers can modify the surface-state distributions in this fashion.[12]

The values of oxide conductance in Table I are quite consistent with recently reported oxide conductance measurements of T. M. Nazar.[13] Although the frequency range of Nazar's measurements was below ours, his measurements show the conductance to be roughly proportional to frequency over a wide range. From this behavior, he suggests that the conductance is due to a hopping process of conduction. However, no mention was made of the orientation of his silicon substrates. In our model, it would probably have been better to

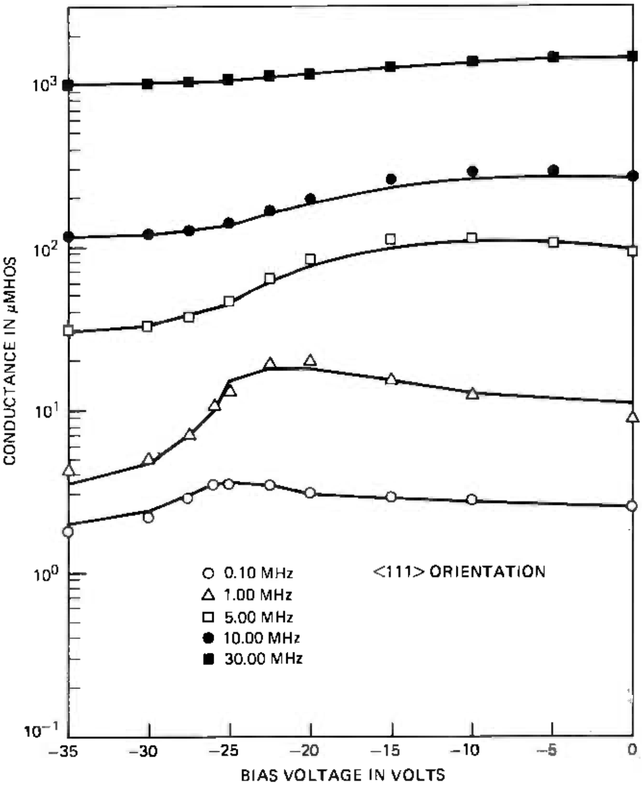


Fig. 10—Conductance of MOS capacitor.

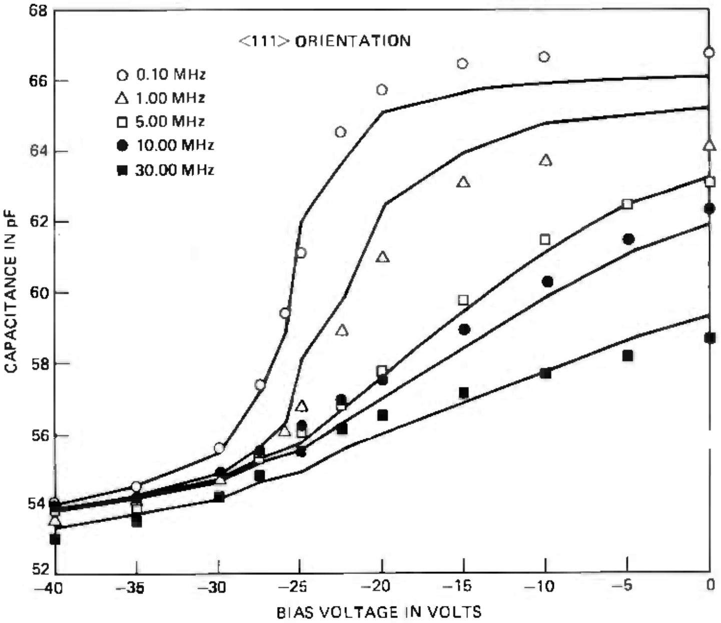


Fig. 11—Capacitance of MOS capacitor.

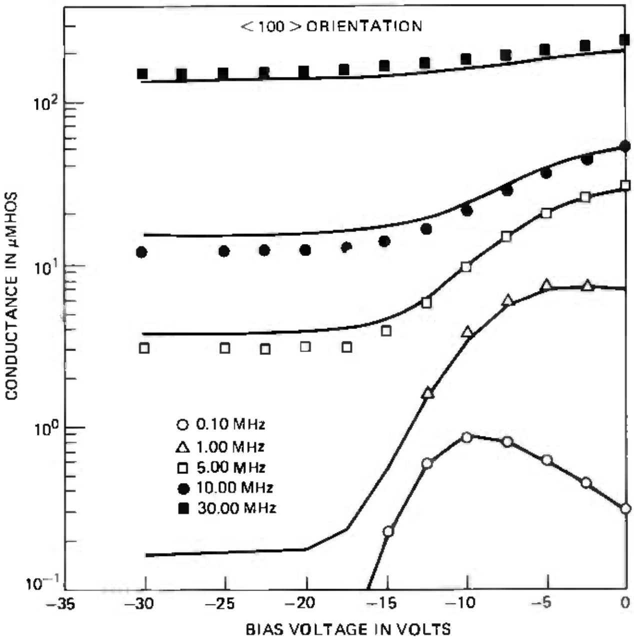


Fig. 12—Conductance of MOS capacitor.

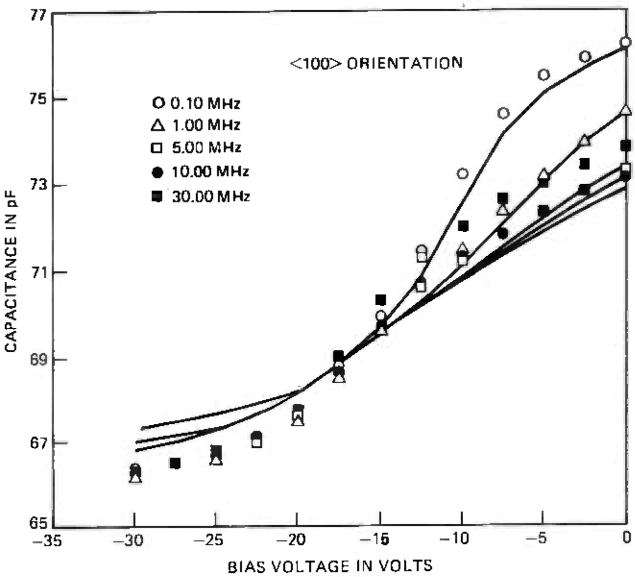


Fig. 13—Capacitance of MOS capacitor.

represent the conductance by a simple frequency dependence than to assume it to be constant, as we did. In any case, this has little effect on the computed modulation coefficient.

V. DISCUSSION OF RESULTS OBTAINED WITH THE COLLECTOR-BASE JUNCTION-CONTROLLED MOS CAPACITOR MODEL

Using the algorithms described above, values of M_3 have been computed for two collector-base junctions, each with a different junction-controlled MOS capacitor. These capacitors are characterized by the parameters given in Table I, which are representative of two different silicon crystal orientations, ⟨111⟩ and ⟨100⟩, with high and low surface-state densities. The collector doping profile of the SG transistor is far from flat near the collector-base junction. Hence, to find the junction capacitance, we preferred to use an algorithm that could calculate junction capacitance for an arbitrary doping profile. H.J.J. De Man has described a fast algorithm for such a calculation.[14] The algorithm used in this model for the collector-base junction capacitance is based on De Man's work.

R_D turned out to be more than two orders of magnitude smaller than R_s for these simulations and so was ignored. Moreover, in (22) $\delta\psi_{Fn}$ was assumed to be negligible in comparison with $\delta\psi_s$.

Computed and measured values of M_3 as a function of collector-base bias voltage are plotted in Figs. 15 and 16. Figure 15 represents the case of ⟨111⟩ crystal orientation, and Fig. 16 represents ⟨100⟩. The

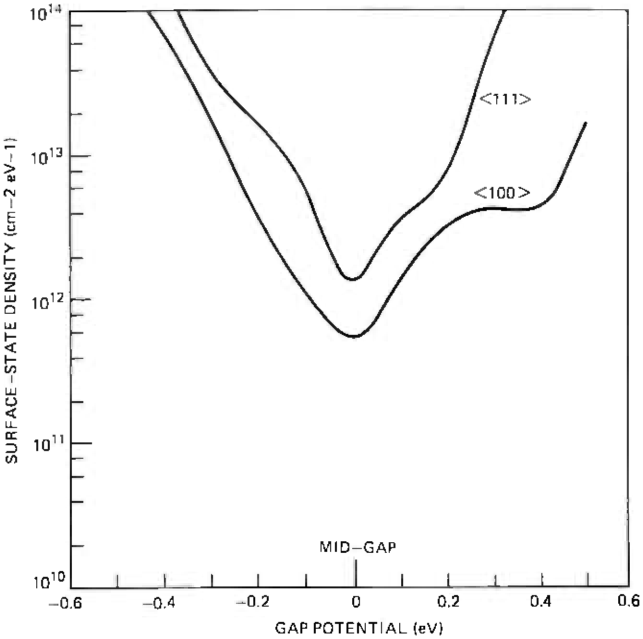


Fig. 14—Distribution of surface states.

Table I—Values of surface parameters

Parameter	Orientation ⟨111⟩	Orientation ⟨100⟩
Oxide capacitance per unit area, nF cm^{-2}	3.83	4.19
Surface impurity concentration, 10^{15}cm^{-3}	1.08	4.67
Fixed charge density, 10^{11}/cm^2	5.85	3.15
Oxide conductance, micromhos/cm^2	75.	0.32
Series resistance, ohms	17.	1.6

computed curve for the ⟨111⟩ junction agrees qualitatively with the measured data in that it shows the hump that is characteristic of ⟨111⟩ junctions. The appearance of the hump at higher voltage in the computed curve than in the measured data is, we believe, due to inaccuracies in determining the surface parameters used for the simulation and to approximations used in the modeling.

The position and height of the hump are sensitive to the values of fixed charge and average surface-state density used in the simulation, as shown in Fig. 17. Even though the MOS capacitors used to determine the surface parameters all came from a single ⟨111⟩-oriented silicon slice and a single ⟨100⟩ slice, there was considerable variation in the measured admittance data for the ⟨111⟩ samples. This, in turn, produces a spread in the surface-state parameters derived from fitting the

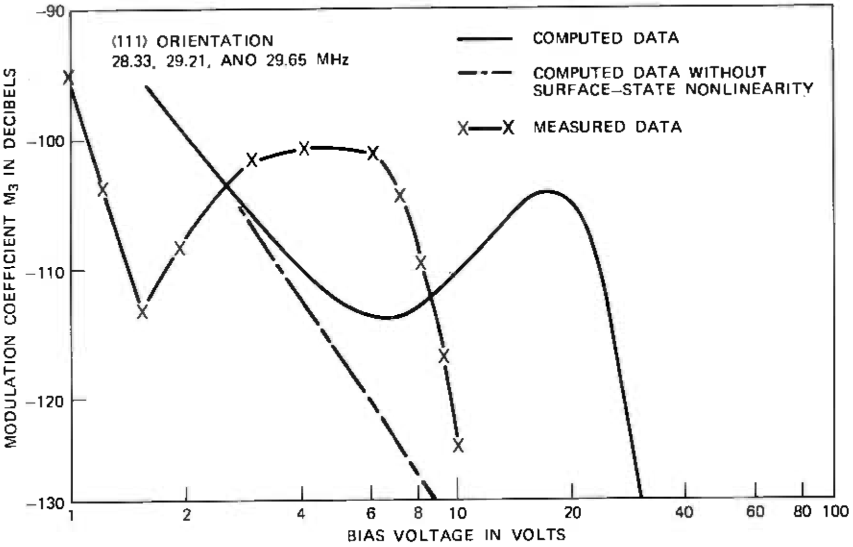


Fig. 15—Intermodulation characteristic.

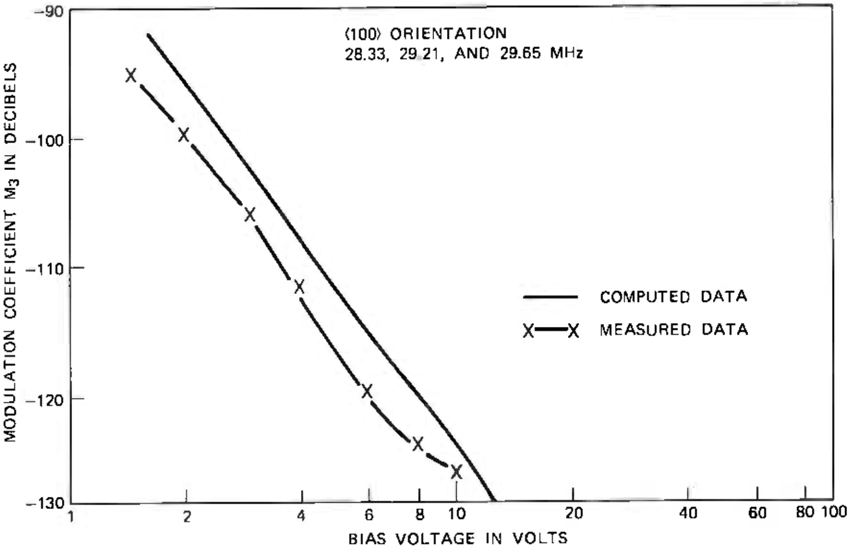


Fig. 16—Intermodulation characteristic.

admittance data. And this spread undoubtedly contributes to the difference between the computed and measured M_3.

As discussed earlier, the dominant surface effect on intermodulation would be expected to arise from the region under the base connecting pad. And so we have restricted our surface modeling to this portion of

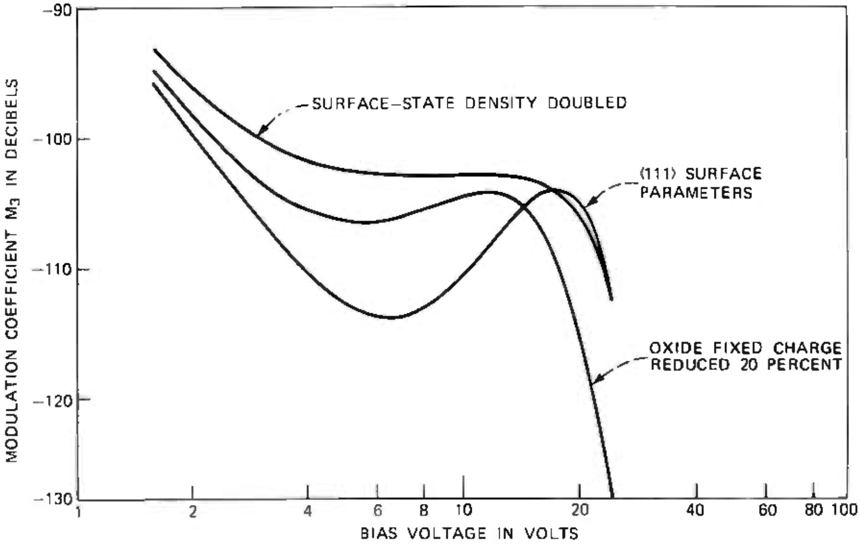


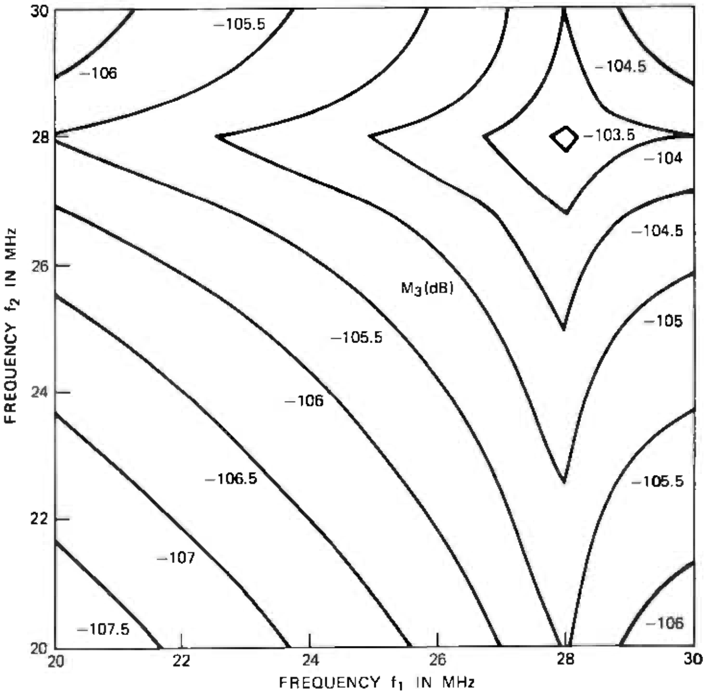
Fig. 17—Effects of changing surface parameters on intermodulation characteristic.

the intersection of the collector-base junction with the surface. However, there are probably also contributions from the extensive region not under the base pad where the collector-base junction intersects the surface. This region would be much more difficult to model, and we have not attempted to do so. This omission is probably responsible for some of the difference in the computed and measured M_3, also.

The computed M_3 for the ⟨100⟩ orientation agrees well with measured data. Note the almost complete absence of a hump in these curves. This is because of the low surface-state and fixed charge densities for this orientation.

If, in computing M_3, for the ⟨111⟩ case, the surface-state nonlinearity is omitted, the computed M_3 comes out to be that shown by the dashed line in Fig. 15. This shows clearly that the surface states are the primary source of the difference in intermodulation produced by the ⟨100⟩ and ⟨111⟩-oriented transistors.

Contour maps of M_3 in the f_1, f_2 frequency plane for a fixed value of the product frequency $(f_1 + f_2 - f_3)$ reveal those combinations of frequency that give rise to high intermodulation. These maps have been computed using the collector-base junction-controlled MOS capacitor model. Figures 18 and 19 show M_3 maps for the ⟨111⟩ junctions at bias voltages corresponding to the peaks of the characteristic M_3-bias humps (see Fig. 15). These bias points (17 V for the simulation and 5 V for the measurements) have been chosen because surface effects dominate the distortion at these points. The predicted levels of M_3 and the general shape of the contours agree reasonably with



BIAS VOLTAGE = 17.6 V, PRODUCT FREQUENCY = 27.89 MHz
⟨111⟩ ORIENTATION

Fig. 18—Computed contour map of modulation coefficient.

measurements. However, the measurements show a stronger frequency dependence.

Contour maps of M_3 for ⟨100⟩ junctions are shown in Fig. 20 for the model and Fig. 21 for measured data. In the absence of any hump in the M_3-bias curves, a bias voltage of 5 V was chosen for the comparison. Again the predicted map agrees in magnitude and general shape with the measured map. The M_3 is seen to be less frequency-dependent than for the ⟨111⟩ case, except at frequencies very close to the product frequency.

It should be pointed out that the nonlinear model only includes terms up to third order. This approximation is valid only if the power in the fundamental driving tones is sufficiently low. The measured data shown in Figs. 15, 16, 19, and 21 have been measured with power in each fundamental P_0 set at −5 dBm. Lower power was not used because the resulting distortion power is near the test set noise limit of −140 dBm when the bias voltage is near 5 V. The power dependence of M_3 is shown in Fig. 22 for both ⟨111⟩ and ⟨100⟩ crystal orientations for a bias voltage of 5 V.

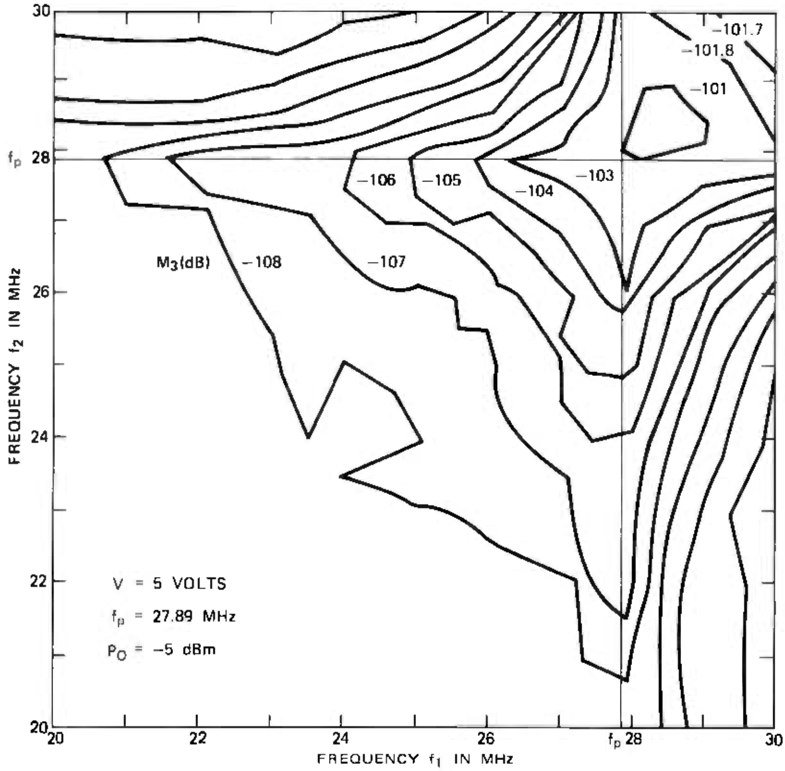
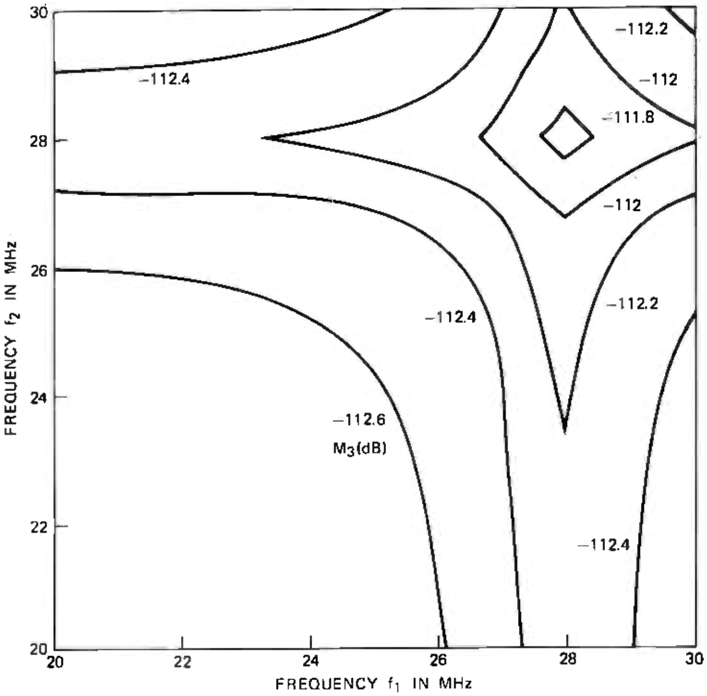


Fig. 19—Measured M_3 frequency map for ⟨111⟩ collector-base junction.

The noticeable rise in M_3 as P_0 is reduced indicates that the distortion mechanism is not accurately described by only second- and third-order terms when P_0 is greater than −15 dBm. This is because a pure third-order nonlinearity does not yield a power-dependent M_3. For this reason, the model of the collector-base junction does not predict M_3 variation with power. The lack of higher-order terms in the model may also explain the low predicted sensitivity of M_3 to the choice of fundamental frequencies.

The use of Volterra series and the perturbation method to describe a complex nonlinear system becomes rather unwieldly when terms higher than third-order are included. A simpler analytical method is needed to simulate the higher-order nonlinear effects in a collector-base junction-controlled MOS capacitor. Simulations using the present model suggest that the dominant source of surface-state distortion is due to the exponential dependence of surface electron density n_s on surface potential ψ_s as given in eq. (20). The G_2 and G_3 Volterra kernels are negligible contributors to the D_1, D_2, and D_3 composite Volterra kernels. If a higher-order nonlinear model were to be developed for



BIAS VOLTAGE = 5.0 V, PRODUCT FREQUENCY = 27.89 MHz
⟨100⟩ ORIENTATION

Fig. 20—Computed contour map of modulation coefficient.

surface-state intermodulation, perhaps only the n_s nonlinearity would need to be included.

These computations have shown how important the contribution of surface states can be to the nonlinear performance of collector-base junctions. If intermodulation distortion is critical in the application of these junctions, it is important to control the distribution of surface states at the Si-SiO$_2$ interface. This is not only a matter of the choice of crystal orientation, but a matter of the oxidation process used and also of post-oxidation processing.

VI. MODELING INTERMODULATION DISTORTION IN THE COMPLETE SG TRANSISTOR

Transistor models used in circuit analysis programs such as the Gummel-Poon model[15] and Extended Ebers-Moll model[16] are adequate for low-current modeling where temperature effects, base pushout, and lateral spreading are not important. Low-current operation means that, for a fixed collector-emitter voltage, the collector current is below the currents for which f_T or β peak. High-frequency circuits require

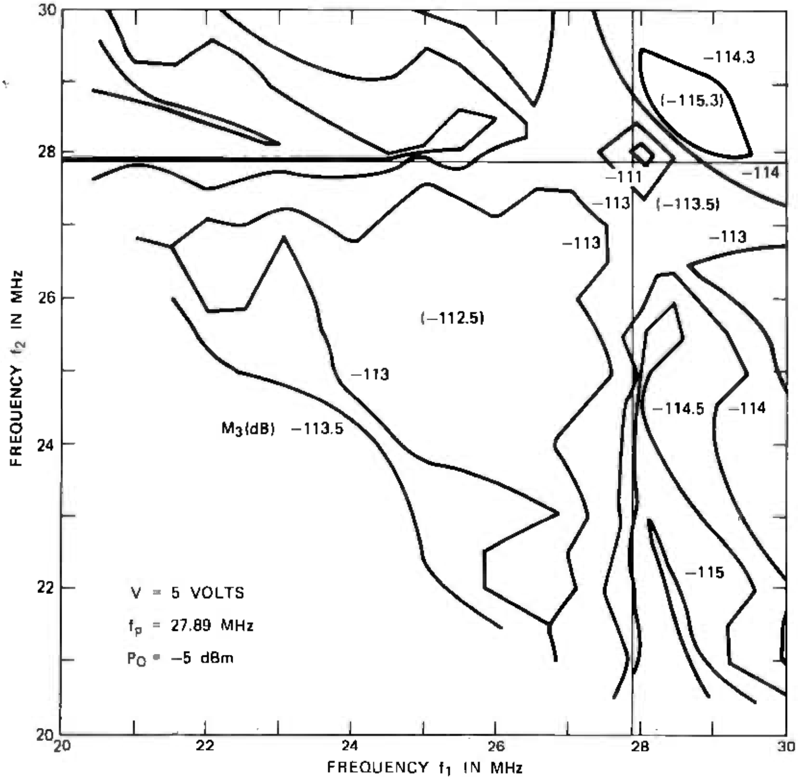


Fig. 21—Measured M_3 frequency map for (100) collector-base junction.

the highest f_T possible for the critical transistors, so the transistors must be biased with high currents.

The high current model presented in this paper extends the Gummel-Poon model by defining two base charge terms. One is the actual base charge and the other is a pseudo-base charge that allows Gummel's equation to describe the dominant component of collector current even when two-dimensional effects are important. In the limiting case of a one-dimensional structure, both charges are identical.

A number of high-current base-charge models have been proposed,[17-20] but they are based on device physics and need device geometry and material parameters as inputs. These models handle one-dimensional high current effects accurately, but only approximate two-dimensional effects. The base charge and the pseudo-base charge discussed in this paper are not derived from device physics but use empirical formulas as first suggested by Choma.[21] The equations for base charge given by Choma are found to model the quasi-base charge adequately, but a new set of equations is used to model the total base charge.

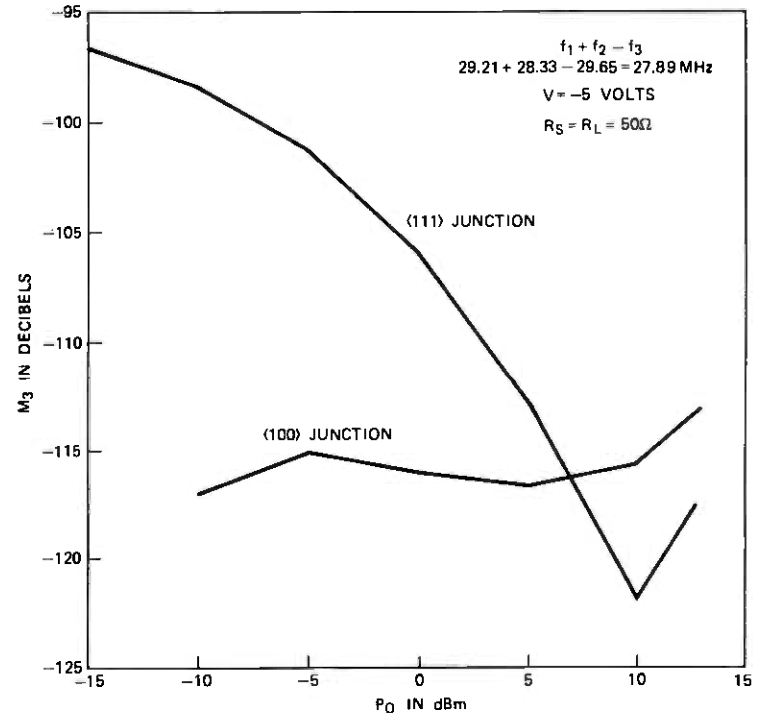


Fig. 22—Measured M_3 vs fundamental power for collector-base junctions.

Model parameter extraction is done with a computer program that uses measured currents and voltages of the test transistor as an input. A two-pass operation is needed to get a complete model. The first pass computes low current model parameters and plots the base charge multiplication factor. The user must define a functional form that can approximate the base charge multiplication factor as a function of collector current and collector-base voltage. The functional form may differ between transistor types due to different doping and structure. The user then computes initial guesses at the model parameters. The second pass uses an optimization program to refine the approximate model parameters. The use of a model parameter extraction program insures unique model parameters because the parameters are determined in precisely the same order with the same assumptions for every extraction run.

6.1 Transistor model description

The cross section of a typical bipolar device is shown in Fig. 23. Two components of dominant collector current I_{CCI}, I_{CCS} and base charge

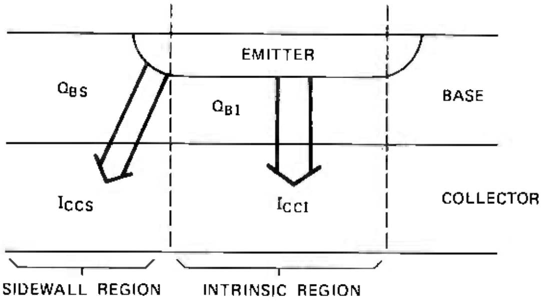


Fig. 23—Transistor cross section.

Q_{BI}, Q_{BS} associated with the intrinsic region directly under the emitter and the sidewall region are defined. The intrinsic region has current flow in only one dimension, and Gummel's equation applies where

$$I_{CCI} = \frac{I_S Q_{B0}[\exp(V_{BE}/V_T) - \exp(V_{BC}/V_T)]}{Q_{BI}}. \tag{39}$$

Of course, this assumes that lateral base current flow has a negligible effect. Current flow in the sidewall region is two-dimensional, and no rigorous analog of Gummel's equation exists to describe I_{CCS}. However, if one sacrifices rigor, (39) can be modified to approximate I_{CC}.

$$I_{CC} = I_{CCI} + I_{CCS} = \frac{I_S Q_{B0}[\exp(V_{BE}/V_T) - \exp(V_{BC}/V_T)]}{Q_{BQ}}, \tag{40}$$

where Q_{BQ} is defined as the quasi-base charge which equals Q_{BI} for a purely intrinsic device. At low collector currents, $Q_{BQ} \approx Q_{BI}$. However, at high currents both base pushout and lateral spreading occur, and Q_{BQ} must change; so (40) models $I_{CCI} + I_{CCS}$. Q_{BQ} can be expressed as shown below.

$$Q_{BQ} = Q_{B0}\left[1 + \frac{Q_E}{C_{JE0}V_{B0}} + \frac{Q_C}{C_{JC0}V_{A0}} + \frac{T_{F0}}{Q_{B0}} B_X(V_{BE}, V_{BC}) I_{CC}\right] \tag{41}$$

Q_{B0} = approximate zero bias total majority base charge
Q_E = base-emitter junction majority charge
Q_C = base-collector junction majority charge
V_{A0} = forward Early voltage
V_{B0} = reverse Early voltage
T_{F0} = extrapolated low-current transit time for $V_{CB} = 0$
$B_X(V_{BE}, V_{BC})$ = base pushout and lateral spreading function.

For an intrinsic device, $C_{JE0}V_{B0} = C_{JC0}V_{A0} = Q_{B0}$. However, this is not true for a real two-dimensional device. The emitter junction charge Q_E and the associated capacitance C_E are given below.

$$Q_E = \int_0^{V_{BE}} C_E(V)\, dV \tag{42}$$

$$C_E = \begin{cases} C_{E\max} & V_{BE} > P_E(1 - B_E) \\ C_{JE0}\left(\dfrac{V_{BE}}{P_E}\right)^{-M_E} & V_{BE} < P_E(1 - B_E) \end{cases} \tag{43}$$

$$B_E = \left(\frac{C_{JE0}}{C_{E\max}}\right)^{1/M_E}. \tag{44}$$

The zero-bias capacitance is C_{JE0}, whereas the peak value of the depletion capacitance in forward bias is $C_{E\max}$. B_E defines the voltage at which C_E becomes constant at $C_{E\max}$, and P_E is the built-in voltage. The collector charge and capacitance are defined with similar equations and model parameters C_{JC0}, P_E, M_C. An empirical function for B_X is described by Choma[21] and is given below:

$$B_X(V_{BE}, V_{BC}) = B_{X0} + (1 - B_{X0})\, e^{-f} \tag{45}$$

$$B_{X0} = 1 + a e^{V_{BC}/b} \tag{46}$$

$$f = c(e^{V_{BE}/d} - 1), \tag{47}$$

where a, b, c, and d are model parameters.

This function is adequate for test devices studied in this paper, but other functions may be necessary for different device structures.

Equation (41) defines Q_{BQ} in terms of I_{CC}; to define Q_{BQ} in terms of only V_{BE} and V_{BC}, substitute (40) into (41) and solve for Q_{BQ} to get

$$\frac{Q_{BQ}}{Q_{B0}} = \frac{1}{2}\left(1 + \frac{Q_E}{C_{JE0} V_{B0}} + \frac{Q_C}{C_{JC0} V_{A0}}\right) + \frac{1}{2}\left[\left(1 + \frac{Q_E}{C_{JB0} V_{B0}} + \frac{Q_C}{C_{JC0} V_{A0}}\right)^2 + \frac{4T_{F0}}{Q_{B0}} B_X I_S \left(e^{V_{BE}/V_T} - e^{V_{BC}/V_T}\right)\right]^{1/2}. \tag{48}$$

The base-pushout and lateral spreading function B_X as defined by (45), (46), and (47) is only valid for the active or quasi-saturation regions of operation. A more complete description of B_X is needed to model inverse operation.

The description of I_{CC} given above is combined with the other elements of the Gummel-Poon model[15] to describe base current and charge storage as shown in Fig. 24. The base currents injected into the collector I_{BC} and emitter I_{BE} and the base resistance R_B are given by

$$I_{BE} = I_1 e^{V_{BE}/V_T} + I_2 e^{V_{BE}/N_E V_T} \tag{49}$$

$$I_{BC} = I_3 e^{V_{BC}/V_T} + I_4 e^{V_{BC}/N_C V_T} \tag{50}$$

$$R_B = R_{BX} + R_{BI}/(Q_{BQ}/Q_{B0}). \tag{51}$$

The base resistance due to the external base is modeled with a constant resistance R_{BX} and a conductivity-modulated resistance R_{BI}.

Self-heating of a device is an important consideration for high-current modeling, so an auxiliary model is included. A current generator equal to the power being dissipated in the device drives the thermal resistance R_T and capacitance C_T. The voltage across R_T equals the instantaneous temperature of the device. For low-frequency circuit simulations, the temperature of the device must be included as a dynamic variable in all the model equations. However, for high frequency simulations, only the dc operating point must be computed with temperature as a variable, since the time constant of the $R_T C_T$ network is long.

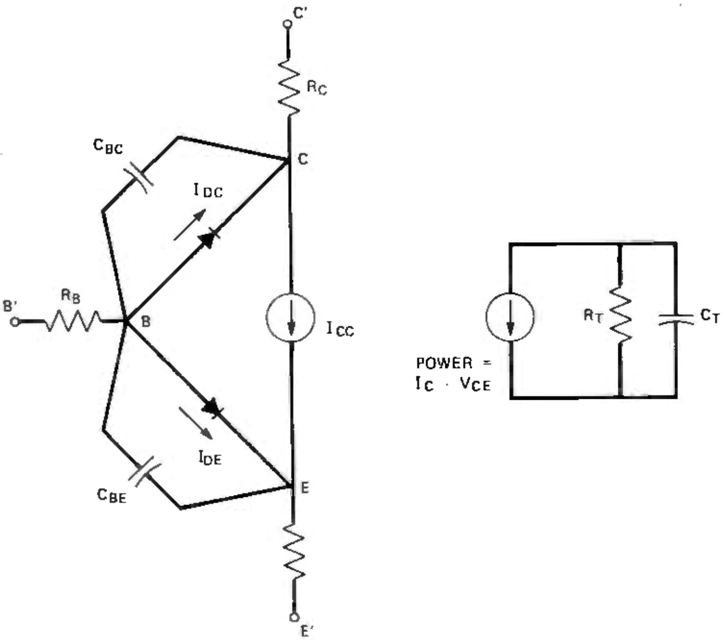


Fig. 24—Transistor model topology.

Charge storage is modeled by capacitors C_{BE} and C_{BC} where

$$C_{BE} = \frac{\partial Q_B}{\partial V_{BE}}, \qquad C_{BC} = \frac{\partial Q_B}{\partial V_{BC}} \tag{52}$$

and Q_B is the actual charge in the base not to be confused with Q_{BQ}. This definition of C_{BE} and C_{BC} assumes the quasi-static approximation[22] for charge distributions in the base. The base charge distribution at time t_0 is given by the dc charge distribution evaluated for $V_{BE} = V_{BE}(t_0)$ and $V_{BC} = V_{BC}(t_0)$. The excess phase in the β versus frequency measurement is caused by delay in the base and in the collector space-charge regions. The quasi-static approximation does not model the diffusion of carriers in the base or drift of carriers in the space-charge region; hence, excess phase is not modeled. However, at frequencies that are low compared to the f_T of the device, the quasi-static model is useful.

The actual base charge consists of a constant charge due to doping Q_{B0}, depletion charge Q_E, Q_C, and diffusion charge, and is defined in a similar fashion as Q_{BQ} to be

$$Q_B = Q_{B0} + Q_E + Q_C + T_{F0}B_Y(I_{CC}, V_{BC})I_{CC}. \tag{53}$$

Note that Q_{B0} is not really necessary in (53) since only the derivatives of Q_B are used. The function $B_Y(V_{BE}, V_{BC})$ accounts for the effects of base pushout and lateral spreading. It is necessary to use a different functional form for B_Y than used for B_X. The f_T vs I_C data cannot be matched closely at both low and high V_C if B_Y has the same form as B_X. However, with B_Y given by the function in (54), the f_T data could be matched over the entire range of V_{CB} as shown in Fig. 25. The dc β vs I_C curve corresponding to Fig. 25 is shown in Fig. 26.

$$B_Y = AI_{CC}^n + B_{Y0} \tag{54a}$$

$$n = 1 + n_0e^{-I_C/n_1} \tag{54b}$$

$$B_{Y0} = B_1e^{V_{BC}/B_2} + B_3 \tag{54c}$$

$$A = A_1V_{BC}^2 + A_2V_{BC} + A_3 \tag{54d}$$

$$n_0 = C_1 + C_2V_{BC} \tag{54e}$$

$$n_1 = C_3 + C_4V_{BC} \tag{54f}$$

with model parameters A_1, A_2, A_3, B_1, B_2, B_3, C_1, C_2, C_3, C_4. Equation (54) represents a function of considerable complexity. However, each subfunction (54b) to (54f) is determined from plots of A, n, and B_{Y0} vs V_{BC} obtained from measured data. The exact forms of these equations may differ for devices with completely different structures.

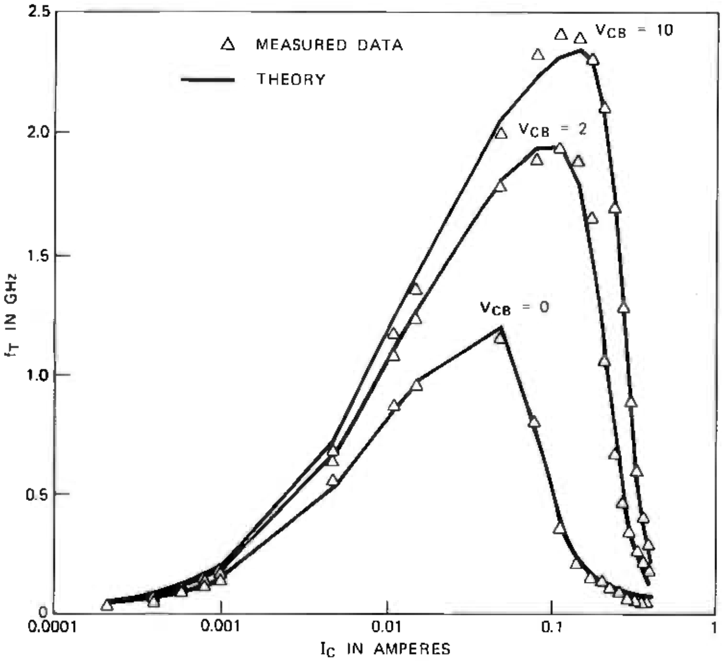


Fig. 25—f_T-current characteristics.

VII. COMPUTATIONS USING THE TRANSISTOR MODEL

The SG transistor has been the product of years of effort to develop a highly linear transistor for submarine cable use.[23,24] As a result, extreme flatness of the I_{BE} versus V_{BE} and β versus V_{BC} and I_C characteristics has been achieved.[23,24]

We have tried to model the SG transistor carefully, so that the model would reproduce measured distortion accurately when anomalous distortion is absent. However, in this attempt, it has become apparent that the distortion predicted by the transistor model per se without surface effects has been excessive.

The primary source of the excess transistor distortion from the model is its inability to adequately simulate the extreme linearity of the SG transistor. To accomplish this involves fitting not only f_T and β accurately over a wide range of voltage and current bias, but the second and third derivatives of these quantities as well. Poon[24] has shown that, in the high frequency, limit M_3 is strongly related to the second derivative of f_T. At lower frequencies, β becomes important. Unfortunately, the modeling equations are not capable of matching β and f_T over a wide bias range. To improve the model, the functions

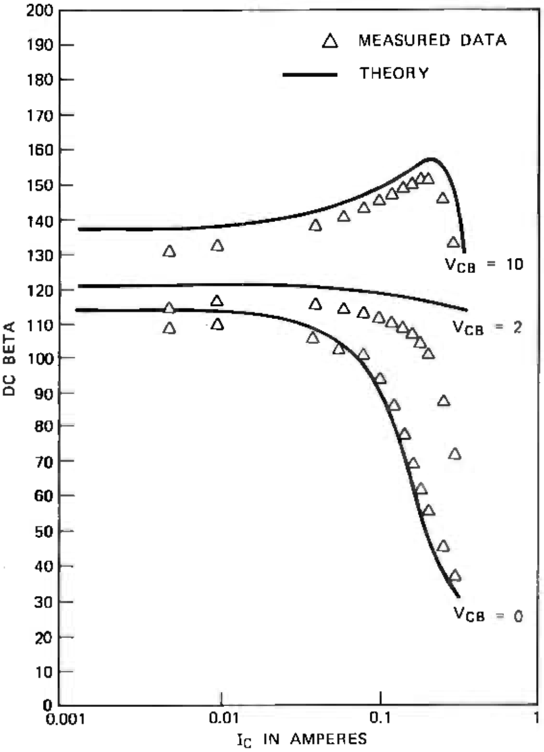


Fig. 26—Direct-current beta-current characteristics.

describing base charge B_X (47) and B_Y (54) must be increased in complexity so as to fit the higher derivatives of charge.

Replacing the collector-base diode with the full transistor model produced no fundamentally new problems to the intermodulation computation itself. The principal effect was simply to increase the complexity of the computations as a result of the increase in equivalent circuit complexity.

Intermodulation analysis requires derivatives up to third order for all nonlinear elements. These are easily obtained explicitly except for the normalized base charge Q_B, which is a function of V_{BE} and V_{BC}. Derivatives of Q_B have been computed numerically to avoid derivation mistakes and to allow quick changes in the equations for Q_B without rederiving the derivatives.

For an operating bias of $V_{CB} = 10$ V, the internal junction temperature is elevated about 75°C. The inclusion of junction heating in the transistor model lowers the simulated M_3 by some 4 dB at the higher voltages. This occurs because of the slight reduction of exponential modeling factors of the form $e^{qV/kT}$.

Computed and measured plots of M_3 versus collector-base bias are shown in Fig. 27. The measured data show the characteristic hump for the ⟨111⟩-oriented transistor. These plots are very similar to those for just the collector-base junction (Figs. 15 and 16), the principal difference being that the transistor amplifies the distortion by about 10 dB.

For the simulated M_3 curve, the β transistor nonlinearities were artificially reduced in accordance with the comments above. Without this reduction, the M_3 curve from the transistor alone would be about 14 dB higher at the high-voltage end and would override any contribution from surface effects. The effect of the surface-state contribution to M_3 is seen to be a hump at approximately the same level as the measured hump.

Although the transistor model does not allow exact prediction of

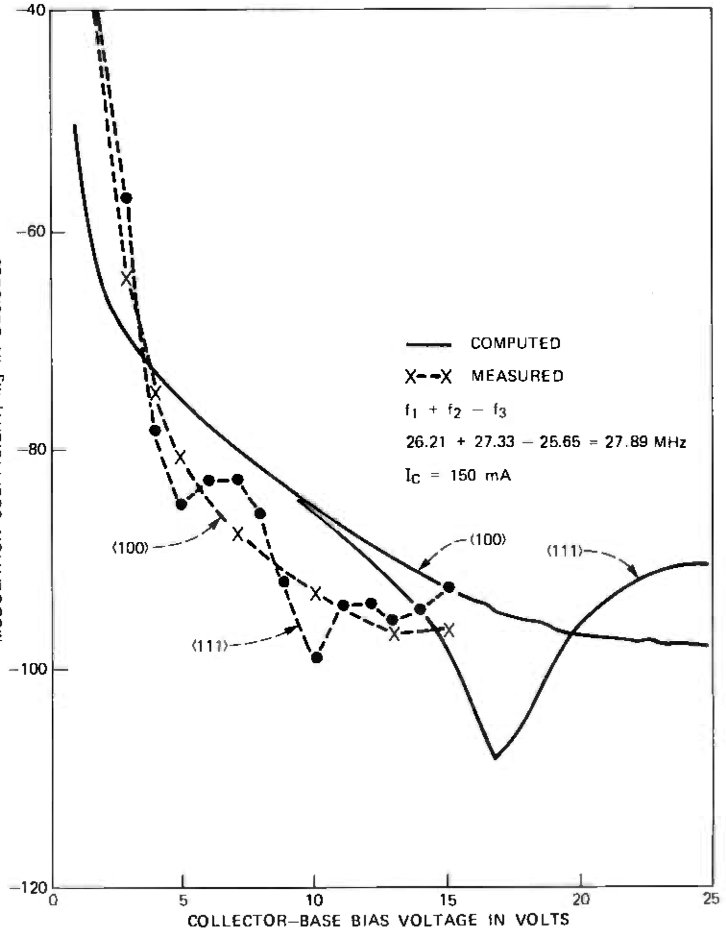


Fig. 27—Transistor intermodulation-voltage characteristics.

distortion, simulations were made to show variations of M_3 with collector current. As shown in Fig. 28, these follow the same trend as measured data. The general shape of the curves is a result of transistor nonlinearities, not surface effects. Base current is composed of two terms.

$$I_B = \frac{I_C}{\beta} + \frac{dQ_B}{dt}. \tag{55}$$

At high currents where β is constant, Q_B is dominated by depletion charge, which is nonlinearly related to I_C. As I_C increases, Q_B becomes dominated by diffusion charge, and Q_B is proportional to I_C/f_T. The first minimum in M_3 is reached at the current for which $\partial f_T/\partial I_C = 0$, that is, at the peak of the $f_T - I_C$ curve. The next local minimum in M_3

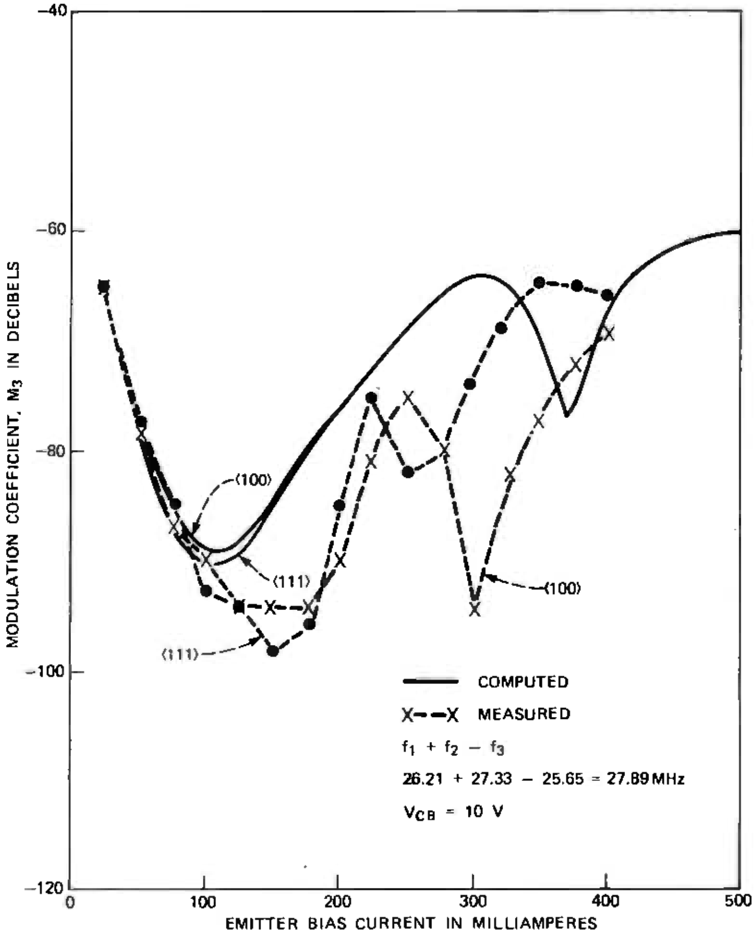


Fig. 28—Transistor intermodulation-current characteristics.

occurs where $\partial^2 f_T/\partial I_c^2 = 0$, that is, at the inflection point in the $f_T - I_C$ curve. This simplified discussion is valid at high frequency where dQ_B/dt dominates in (55). For lower frequencies, the effect of nonlinearities in β becomes important.

VIII. ACKNOWLEDGMENTS

We would like to express our appreciation to L. E. Miller for having originally proposed the hypothesis that surface effects are responsible for the anomalous SG distortion and for his enthusiastic support of this work. We are also indebted to D. G. Ross for making available to us his extensive collection of experimental intermodulation data. M. Chruney made the capacitance and conductance measurements on the MOS capacitors, which were provided by W. M. Fox. L. C. Kaufman supplied the algorithm for adjusting the model parameters to fit these measurements.

APPENDIX A

Integrating Volterra Kernels over the Surface-State Distribution

As can be seen from equation (19) and Appendix C, to find the total current flowing into the surface states involves a sum of integrals containing the factor

$$\frac{f_0(1-f_0)}{1+j\omega f_0/c_n n_{s0}}, \tag{56}$$

where $f_0 = [1 + g_s \exp(u - u_F)]^{-1}$.

Now

$$f_0(1-f_0) = \frac{g_s \exp(u-u_F)}{[1+g_s \exp(u-u_F)]^2} = \frac{df_0}{du} \tag{57}$$

is sharply peaked about $u = u_F$. This is readily seen by plotting $f_0(u)$ and noticing that it slopes steeply only near $u = u_F$. Since df_0/du is peaked, so are $f_0(df_0/du)$ and $f_0^2(df_0/du)$. If $\omega/c_n n_{s0} \ll 1$, that is, at low frequency and high semiconductor doping, (56) reduces approximately to (57). Therefore, provided that N_s is essentially constant through this peak, the integrals in (19) can be readily evaluated in the manner of Nicollian and Goetzberger.[7]

However, as $\omega/c_n n_{s0}$ becomes comparable with or greater than unity, i.e., at high frequency or low doping, the peak broadens and shifts away from u_F toward the conduction band. The location of the peak can be found by setting the derivative of (56) with respect to u equal to zero and solving for u. The integrals could then be evaluated as above with N_s taken out of the integral and set equal to the value at the peak location. But broadening the peak makes this approach rather unpalatable, and we have preferred to evaluate the integrals numerically.

APPENDIX B

Taking into Account Surface Potential Fluctuations in Calculating Surface-State Volterra Kernels

Following Nicollian and Goetzberger,[7] assume that the dc surface potential is uniform over a small region of characteristic area α. The probability that the dc surface potential on a particular one of these regions lies between u_s and $u_s + du_s$ is $P(u_s)\, du_s$. Then, to find the first-order Volterra kernel $D_1(\omega_1)$, we multiply (23) by $P(u_s)\, du_s$ and integrate over all possible surface potentials.

$$\int_{-\infty}^{\infty} D_1(\omega_1)P(u_s)\, du_s = \int_{-\infty}^{\infty} j\omega_1 \frac{q}{V_T} \int_{u_V}^{u_C} N_s(u,\, u_s)H_1(\omega_1)\, du \cdot n_{s0}P(u_s)\, du_s. \quad (58)$$

We recognize that f_0 is not a function of u_s and assume $P(u_s)$ to be sharply peaked at the average surface potential $\bar{u}_s$. In particular, we assume $P(u_s)$ to be Gaussian.

$$P(u_s) = (2\pi\sigma_s^2)^{-1/2} \exp\left[\frac{-(u_s - \bar{u}_s)^2}{2\sigma_s^2}\right], \quad (59)$$

where

$$\sigma_s = \frac{A}{(C_{ox} + C_d)\, V_T}\left[\frac{q\bar{Q}}{\alpha(D_d)}\right]^{1/2}$$

and $\bar{Q}$ is the mean oxide charge density in C/cm^2.

Then inverting the order of integration, we assume that $N_s(u,\, u_s)$ varies slowly with u_s and can be taken outside the integral with respect to u_s. We then have

$$\int_{-\infty}^{\infty} D_1(\omega_1)P(u_s)\, du_s$$

$$= \frac{j\omega_1 q}{V_T} \int_{u_V}^{u_C} N_s(u,\, \bar{u}_s) f_0(1 - f_0) \int_{-\infty}^{\infty} \frac{P(u_s)\, du_s}{1 + R(u_s)}\, du, \quad (60)$$

where $\bar{u}_s$ is the average value of the surface potential and

$$R(u_s) = \frac{j\omega_1 f_0}{c_n n_i} \exp\left[-(u_s - u_F)\right].$$

Now expand $[1 + R(u_s)]^{-1}$ in a Taylor's series about $\bar{u}_s$,

$$F(u_s) = \frac{1}{1 + R(u_s)} = F(\bar{u}_s) + F'(\bar{u}_s)(u_s - \bar{u}_s) + \frac{F''(\bar{u}_s)}{2}(u_s - \bar{u}_s)^2 + \cdots.$$

Then

$$\int_{-\infty}^{\infty} \frac{P(u_s)}{1 + R(u_s)}\, du_s \approx F(\bar{u}_s) \int_{-\infty}^{\infty} P(u_s)\, du_s$$

$$+ F'(\bar{u}_s) \int_{-\infty}^{\infty} (u_s - \bar{u}_s) P(u_s)\, du_s$$

$$+ \frac{F''(\bar{u}_s)}{2} \int_{-\infty}^{\infty} (u_s - \bar{u}_s)^2 P(u_s)\, du_s$$

$$\approx F(\bar{u}_s) + \frac{\sigma_s^2}{2} F''(\bar{u}_s). \tag{61}$$

The second integral on the right vanishes. Hence,

$$\int_{-\infty}^{\infty} D_1(\omega_1) P(u_s)\, du_s = \frac{j\omega_1 q}{V_T} \int_{u_V}^{u_C} N_s f_0 (1 - f_0) \hat{F}_1(\omega_1, \bar{u}_s)\, du, \tag{62}$$

where $\hat{F}_1(\omega_1, \bar{u}_s)$ is the quantity on the right side of (61).

Treating the second- and third-order Volterra kernels similarly, we obtain

$$\int_{-\infty}^{\infty} D_2(\omega_1, \omega_2) P(u_s)\, du_s \approx -\frac{j(\omega_1 + \omega_2) q}{2V_T^2} \int_{u_V}^{u_C} [\hat{F}_1(\omega_1 + \omega_2, \bar{u}_s) f_0$$

$$- \hat{F}_2(\omega_1, \omega_2, \bar{u}_s) f_0^2] N_s(u)(1 - f_0)\, du, \tag{63}$$

where, letting $\omega_1 + \omega_2 - \omega_3 = \omega_p$,

$$\hat{F}_2(\omega_1, \omega_2, \bar{u}_s) = F(\omega_1, \omega_1 + \omega_2, \bar{u}_s) + F(\omega_2, \omega_1 + \omega_2, \bar{u}_s)$$

$$+ \frac{\sigma_s^2}{2} [F''(\omega_1, \omega_1 + \omega_2, \omega_p, \bar{u}_s)$$

$$+ F''(\omega_2, \omega_1 + \omega_2, \omega_p, \bar{u}_s)]$$

$$\int_{-\infty}^{\infty} D_3(\omega_1, \omega_2, \omega_3) P(u_s)\, du_s \approx -\frac{j(\omega_p) q}{V_T^3} \int_{u_V}^{u_C} \left[\frac{\hat{F}_1(\omega_p, \bar{u}_s)}{6} f_0 \right.$$

$$- \overline{\hat{F}_2(\omega_1, \omega_2 + \omega_3, \bar{u}_s)} f_0^2$$

$$\left. + \frac{\hat{F}_3(\omega_1, \omega_2, \omega_3, \bar{u}_s)}{6} f_0^3 \right]$$

$$\cdot (1 - f_0) N_s(u)\, du, \tag{64}$$

where the bar is used as in (25) and

$$\begin{aligned}\hat{F}_3(\omega_1, \omega_2, \omega_3) = {} & F(\omega_1, \omega_1 + \omega_2, \omega_p, \bar{u}_s) + F(\omega_2, \omega_1 + \omega_2, \omega_p, \bar{u}_s) \\ & + F(\omega_1, \omega_1 + \omega_3, \omega_p, \bar{u}_s) + F(\omega_3, \omega_1 + \omega_3, \omega_p, \bar{u}_s) \\ & + F(\omega_2, \omega_2 + \omega_3, \omega_p, \bar{u}_s) + F(\omega_3, \omega_2 + \omega_3, \omega_p, \bar{u}_s) \\ & + \frac{\sigma_s^2}{2} [F''(\omega_1, \omega_1 + \omega_2, \omega_p, \bar{u}_s) \\ & + F''(\omega_2, \omega_1 + \omega_2, \omega_p, \bar{u}_s) \\ & + F''(\omega_1, \omega_1 + \omega_3, \omega_p, \bar{u}_s) + F''(\omega_3, \omega_1 + \omega_3, \omega_p, \bar{u}_s) \\ & + F''(\omega_1, \omega_2 + \omega_3, \omega_p, \bar{u}_s) + F''(\omega_3, \omega_2 + \omega_3, \omega_p, \bar{u}_s)].\end{aligned}$$

The definitions of F and F'' are as follows:

$$F(\omega_1, \omega_2, \cdots, w_n) = \prod_{i=1}^{n} \frac{1}{1 + R(\omega_i)}$$

$$\begin{aligned}F''(\omega_1, \omega_2, \cdots, \omega_n) = {} & F(\omega_1, \omega_2, \cdots, \omega_n) \\ & \left[\left(\sum_{i=1}^{n} R(\omega_i)F(\omega_i)\right)^2 - \sum_{i=1}^{n} R(\omega_i)F^2(\omega_i)\right].\end{aligned}$$

APPENDIX C

Derivation of Surface-State Volterra Kernels

A Volterra series for f in terms of n_s, eq. (12), describing the time dependence of the Fermi level is rewritten as

$$\frac{df}{dt} = c_n(1 - f)n_s - e_n f. \tag{65}$$

First break (56) into a dc part and an ac part. Let

$$f = f_0 + \delta f, \qquad n_s = n_{s0} + \delta n_s$$

$$\frac{d}{dt} \delta f = c_n(1 - f_0 - \delta f)(n_{s0} + \delta n_s) - e_n(f_0 + \delta f).$$

Direct current equation:

$$0 = c_n(1 - f_0)n_{s0} - e_n f_0.$$

Therefore

$$e_n = \frac{c_n(1 - f_0)n_{s0}}{f_0}$$

or

$$\frac{c_n n_{s0}}{f_0} = c_n n_{s0} + e_n.$$

Alternating current equation:

$$\frac{d}{dt}\,\delta f = c_n(1 - f_0)\delta n_s - (c_n n_{s0} + e_n)\delta f - c_n \delta f \delta n_s$$

or

$$\frac{d}{dt}\,\delta f = c_n(1 - f_0)\delta n_s - \frac{c_n n_{s0}}{f_0}\,\delta f - c_n \delta f \delta n_s. \tag{66}$$

It is desirable to express δf as a Volterra series

$$\delta f = H_1(\omega_a)\ o\ \delta n_s + H_2(\omega_a, \omega_b)\ o\ \delta n_s^2 + H_3(\omega_a, \omega_b, \omega_c)\ o\ \delta n_s^3, \tag{67}$$

where the o operator indicates that the magnitude and phase of each term in δn_s^n is to be changed by the magnitude and phase of $H_n(\omega_a, \cdots, \omega_n)$. To find the Volterra kernels $A_n(\omega_1, \cdots, \omega_n)$ substitute (58) into the ac eq. (57) using a theorem for higher order Fourier transforms of derivatives (76) to get:

$$\begin{aligned}
&j\omega_a H_1(\omega_a)\ o\ \delta n_s + j(\omega_a + \omega_b)H_2(\omega_a, \omega_b)\ o\ \delta n_s^2 \\
&\qquad + j(\omega_a + \omega_b + \omega_c)H_3(\omega_a, \omega_b, \omega_c)\ o\ \delta n_s^3 \\
&\qquad = c_n(1 - f_0)\delta n_s - \frac{c_n n_{s0}}{f_0}\,[H_1(\omega_a)\ o\ \delta n_s + H_2(\omega_a, \omega_b)\ o\ \delta n_s^2 \\
&\qquad + H_3(\omega_a, \omega_b, \omega_c)\ o\ \delta n_s^3] \\
&\qquad - c_n\{[H_1(\omega_a)\ o\ \delta n_s]\delta n_s + [H_2(\omega_a, \omega_b)\ o\ \delta n_s^2]\delta n_s \\
&\qquad + [H_3(\omega_a, \omega_b, \omega_c)\ o\ \delta n_s]\delta n_s\}.
\end{aligned}$$

Consider $\delta n_s = a_1 \cos(\omega_1 t + \phi_1) + a_2 \cos(\omega_2 t + \phi_2) + a_3 \cos(\omega_3 t + \phi_3)$ as the forcing function and ignore the $[H_3(\omega_a, \omega_b, \omega_c)\ o\ \delta n_s^3]\delta n_s$ term since it contributes fourth-order terms and small second-order terms. Also ignore the small contribution to the first-order terms due to $[H_2(\omega_a, \omega_b)\ o\ \delta n_s^2]\delta n_s$. Equate terms of the same order to get three equations:

$$j\omega_a H_1(\omega_a)\ o\ \delta n_s = c_n(1 - f_0)\ o\ \delta n_s - \frac{c_n n_{s0}}{f_0}\,H_1(\omega_a)\ o\ \delta n_s \tag{68}$$

$$\begin{aligned}
j(\omega_a + \omega_b)H_2(\omega_a, \omega_b)\ o\ \delta n_s^2 = &-\frac{c_n n_{s0}}{f_0}\,H_2(\omega_a, \omega_b)\ o\ \delta n_s^2 \\
&- c_n[H_1(\omega_a)\ o\ \delta n_s]\delta n_s
\end{aligned} \tag{69}$$

$$\begin{aligned}
j(\omega_a + \omega_b + \omega_c)H_3(\omega_a, \omega_b, \omega_c)\ o\ \delta n_s^3 = &-\frac{c_n n_{s0}}{f_0}\,H_3(\omega_a, \omega_b, \omega_c)\ o\ \delta n_s^3 \\
&- c_n[H_2(\omega_a, \omega_b)\ o\ \delta n_s^2]\delta n_s.
\end{aligned} \tag{70}$$

Solving (68) for H_1 gives

$$H_1(\omega_a) = \frac{c_n(1 - f_0)}{j\omega_a + (c_n n_{s0}/f_0)}.$$

Solving (69) for H_2 using (77) gives

$$j(\omega_a + \omega_b)H_2(\omega_a, \omega_b) \text{ o } \delta n_s^2 = -\frac{c_n n_{s0}}{f_0} H_2(\omega_a, \omega_b) \text{ o } \delta n_s^2$$

$$- \frac{c_n}{2}[H_1(\omega_a) + H_1(\omega_b)] \text{ o } \delta n_s^2.$$

So

$$H_2(\omega_a, \omega_b) = -\frac{c_n[H_1(\omega_a) + H_1(\omega_b)]}{2\left[j(\omega_a + \omega_b) + \dfrac{c_n n_{s0}}{f_0}\right]}.$$

Solving (70) for H_3 using (78) gives

$$j(\omega_a + \omega_b + \omega_c)H_3(\omega_a, \omega_b, \omega_c) \text{ o } \delta n_s^3 = -\frac{c_n n_{s0}}{f_0} H_3(\omega_a, \omega_b, \omega_c) \text{ o } \delta n_s^3$$

$$- \frac{c_n}{3}[H_2(\omega_a, \omega_b) + H_2(\omega_a, \omega_c)$$

$$+ H_2(\omega_b, \omega_c)] \text{ o } \delta n_s^3.$$

So

$$H_3(\omega_a, \omega_b, \omega_c) = -\frac{c_n[H_2(\omega_a, \omega_b) + H_2(\omega_a, \omega_c) + H_2(\omega_b, \omega_c)]}{3\left[j(\omega_a + \omega_b + \omega_c) + \dfrac{c_n n_{s0}}{f_0}\right]}.$$

APPENDIX D

Using Volterra Series

Given a nonlinear system with an input $u(t)$ and output $y(t)$, one may describe the system using a functional series of the form[2]

$$y(t) = y_1(t) + y_2(t) + \cdots y_n(t), \tag{71}$$

where

$$y_n(t) = \int \cdots \int h_n(t - \tau_1, \cdots, t - \tau_n)u(\tau_1), \cdots, u(\tau_n)\, d\tau_a, \cdots, d\tau_n.$$

The $y_1(t)$ term is the ordinary convolution integral that is used for linear system analysis and $h_1(t)$ is the impulse response. The kernel

$h_n(t)$ is the general impulse response. If the input $u(t)$ is changed by a gain factor ϵ, then the new output is

$$\hat{y}_n(t) = \int \cdots \int h_n(t-\tau_1, \cdots, t-\tau_n)\epsilon u(\tau_1), \cdots, \\ \epsilon u(\tau_n)\, d\tau_1, \cdots, d\tau_n = \epsilon^n y_n(t).$$

The system can be thought of as a parallel combination of first, second, third, $\cdots$ nth order subsystems.

The following theorems deal with Volterra series in the frequency domain (7) to (9) using the circle operator defined in (17).
Let

$$u = A(\omega_1)\ o \cos \omega_1 t + \cdots + A(\omega_3)\ o \cos \omega_3 t,$$

then

$$\begin{aligned} H_1(\omega_a)\ o\ u &= H_1(\omega_1)A(\omega_1)\ o \cos \omega_1 t + H_1(\omega_2)A(\omega_2)\ o \cos \omega_2 t \\ &\quad + H_1(\omega_3)A(\omega_3)\ o \cos \omega_3 t \\ &= H_1(\omega_a)A(\omega_a)\ o \cos \omega_a t \quad a = 1, 2, 3 \end{aligned} \tag{72}$$

$$\begin{aligned} &H_2(\omega_a, \omega_b)\ o\ u^2 \\ &\quad = \text{dc terms} \\ &\qquad + H_2(\omega_1, \omega_1)A^2(\omega_1) \quad [½] \quad o \cos 2\omega_1 t \\ &\qquad + H_2(\omega_2, \omega_2)A^2(\omega_2) \quad [½] \quad o \cos 2\omega_2 t \\ &\qquad + H_2(\omega_3, \omega_3)A^2(\omega_3) \quad [½] \quad o \cos 2\omega w_3 t \\ &\qquad + H_2(\omega_1, -\omega_2)A(\omega_1)A(-\omega_2) \quad o \cos(\omega_1 - \omega_2)t \\ &\qquad + H_2(\omega_1, \omega_2)A(\omega_1)A(\omega_2) \quad o \cos(\omega_1 + \omega_2)t \\ &\quad = H_2(\omega_a, \omega_b)A(\omega_a)A(\omega_b) \quad [½] \quad o \cos(\omega_a + \omega_b)t, \end{aligned} \tag{73}$$

where [½] is used for dc and second-order harmonics only.

$$\begin{aligned} &H_3(\omega_a, \omega_b, \omega_c)\ o\ u^3 \\ &\quad = \text{1st-order terms} \\ &\qquad + H_3(\omega_1, \omega_1, \omega_1)A^3(\omega_1) \quad [¼] \quad o \cos 3\omega_1 t \\ &\qquad + H_3(\omega_2, \omega_2, \omega_2)A^3(\omega_2) \quad [¼] \quad o \cos 3\omega_2 t \\ &\qquad + H_3(\omega_3, \omega_3, \omega_3)A^3(\omega_3) \quad [¼] \quad o \cos 3\omega_3 t \\ &\qquad + H_3(\omega_1, \omega_1, -\omega_2)A^2(\omega_1)A(-\omega_2) \quad [¾] \quad o \cos(2\omega_1 - \omega_2)t \\ &\qquad + H_3(\omega_1, \omega_1, \omega_2)A^2(\omega_1)A(\omega_2) \quad [¾] \quad o \cos(2\omega_1 + \omega_2)t \\ &\qquad \vdots \\ &\qquad + H_3(\omega_1, \omega_2, \omega_3)A(\omega_1)A(\omega_2)A(\omega_3) \quad [³⁄₂] \quad o \cos(\omega_1 + \omega_2 + \omega_3)t \\ &\qquad + H_3(\omega_1, -\omega_2, \omega_3)A(\omega_1)A(-\omega_2)A(\omega_3) \quad [³⁄₂] \quad o \cos(\omega_1 - \omega_2 + \omega_3)t \\ &\qquad \approx H_3(\omega_a, \omega_b, \omega_c)A(\omega_a)A(\omega_b)A(\omega_c)\ [\text{Fraction}]\ o \cos(\omega_a + \omega_b + \omega_c)t. \end{aligned} \tag{74}$$

This last formula is useful for all the third-order terms but not the first-order terms.

$$[H_1(\omega_a)\ o\ u]^2 = \tfrac{1}{3}[H_1(w_a)H_2(\omega_b, \omega_c) + H_1(\omega_b)H_2(\omega_a, \omega_c)$$
$$+ H_1(\omega_c)H_2(\omega_a, \omega_b)]\ o\ u^3$$
$$= \overline{H_1(\omega_a)H_2(\omega_b, \omega_c)}\ o\ u^3, \tag{75}$$

where the bar is used as in (25).

Differentiating Volterra series,

$$\frac{d}{dt} y_K(t) \longleftrightarrow (j\omega_1 + \cdots + j\omega_K)H_K(\omega_1, \cdots, \omega_K) \tag{76}$$

$$[H_1(\omega_a)\ o\ u] \times u = a_1 |H_1(\omega_1)| \cos(\omega_1 t + \phi_1 + \phi_{\omega 1})$$
$$+ a_2 |H_1(\omega_2)| \cos(\omega_2 t + \phi_2 + \phi_{\omega 2})$$
$$+ a_3 |H_1(\omega_3)| \cos(\omega_3 t + \phi_3 + \phi_{\omega 3})] \times u.$$

For harmonic terms, i.e.,

$$\omega_a = \omega_1 \quad \omega_b = \omega_1 \rightarrow 2\omega_1$$

or

$$\omega_a = \omega_1 \quad \omega_b = -\omega_a \rightarrow \text{dc}$$

$$[H_1(\omega_a)\ o\ u] \times u = \tfrac{1}{2} A(\omega_a)A(\omega_b)H_1(\omega_a)\ o\ \cos(\omega_a + \omega_b)t.$$

For sum and difference terms,

$$[H_1(\omega_a)\ o\ u] \times u$$
$$= \tfrac{1}{2} A(\omega_a)A(\omega_b)[H_1(\omega_a) + H_1(\omega_b)]\ o\ \cos(\omega_a + \omega_b)t.$$

In general,

$$[H_1(\omega_a)\ o\ u] \times u = \frac{1}{2} A(\omega_a)A(\omega_b)[H_1(\omega_a)$$
$$+ H_1(\omega_b)]\left[\frac{1/2}{1}\right]\ o\ \cos(\omega_a + \omega_b)t$$
$$= \frac{1}{2}[H_1(\omega_a) + H_1(\omega_b)]\ o\ u^2 \tag{77}$$

$$[H_2(\omega_a, \omega_b)\ o\ u^2] \times u = \tfrac{1}{3}[H_2(\omega_a, \omega_b) + H_2(\omega_a, \omega_c)$$
$$+ H_2(\omega_b, \omega_c)]\ o\ u^3. \tag{78}$$

APPENDIX E

List of Symbols

B_X, B_Y	Base pushout and lateral spreading functions.
C_d	Surface depletion capacitance.

C_I	Surface inversion capacitance.
c_n	Electron capture probability.
C_{ox}	Oxide capacitance.
C_s	Surface-state capacitance.
C_T	Thermal capacitance.
d	Depletion layer width.
D_n	Complete Fourier-transformed surface-state Volterra kernel of order n.
E	Energy level.
e_n	Electron emission probability.
f_0, $f(t)$	Direct-current and time-dependent fermi distribution functions
G_n	nth order component of Fourier-transformed surface-state Volterra kernel.
g_s	Ground state degeneracy.
h_n	Time-dependent Volterra kernel of order n.
H_n	Fourier-transformed Volterra kernel of order n.
I_{BC}	Base current injected into the collector.
I_{BE}	Base current injected into the emitter.
I_{CCI}, I_{CCS}	Intrinsic and sidewall components of dominant collector current.
$i_s(t)$	Time-dependent surface-state current.
k	Boltzmann's constant.
L_D	Intrinsic Debye length.
M_3	Third-order modulation coefficient.
n_i	Intrinsic electron density.
N_s	Surface-state density.
n_{s0}, $n_s(t)$	Direct-current and time-dependent electron surface concentrations.
P_0	Output power.
q	Electronic charge.
Q_B	Actual charge in base.
Q_{BI}, Q_{BS}	Intrinsic and sidewall components of base charge.
Q_{B0}	Total majority base charge at zero bias.
Q_C	Base-collector majority charge.
Q_E	Base-emitter majority charge.
Q_g	Gate charge.
Q_I	Surface inversion charge.
Q_s	Semiconductor charge.
Q_{ss}	Surface-state charge.
Q_0	Fixed oxide charge.
$\bar{Q}$	Mean oxide charge density in C/cm^2.
R_B	Base resistance.
R_D	Surface depletion layer resistance.
R_I	Surface inversion layer resistance.

R_{nD}	Surface depletion layer resistance for electron flow.
R_{ns}	Surface-state resistance for electron flow.
R_{pB}	Resistance to hole flow in quasi-neutral region adjacent to depletion layer.
R_{pD}	Surface depletion layer resistance for hole flow.
R_{ps}	Surface-state resistance for hole flow.
R_T	Thermal resistance.
t, t_0	Time.
T	Temperature.
T_{F0}	Low-current transit time for $V_{CB} = 0$.
u_{Fn}, u_{Fp}	Fermi potentials for electrons and holes in units of the thermal voltage.
u_s	Surface potential in units of the thermal voltage.
V_{A0}	Forward Early voltage.
V_{BC}	Base-collector voltage.
V_{BE}	Base-emitter voltage.
V_{B0}	Reverse Early voltage.
V_T	Thermal voltage (kT/q).
Y_s	Surface-state admittance.
δ	Alternating current component of.
μ_n	Electron mobility.
ω	Angular frequency.
ψ_{Fn}	Electron quasi-fermi potential in electron volts.
ψ'_{MS}	Metal-semiconductor work function.
ψ_s	Surface potential in volts.

REFERENCES

1. S. T. Brewer et al., "SG Undersea Cable System: Requirements and Performance," B.S.T.J., *57,* No. 7 (September 1978), pp. 2319–2354.
2. H. E. Abraham and R. G. Meyer, "Transistor Design for Low Distortion at High Frequencies," IEEE Trans. on Electron Devices, *ED-23,* No. 12 (December 1976), pp. 1290–1297.
3. A. S. Grove and D. J. Fitzgerald, "Surface Effects on p-n Junctions: Characteristics of Surface Space-Charge Regions Under Non-Equilibrium Conditions," Solid-State Electronics, *9* (August 1966), pp. 783–806.
4. K. Lehovec and A. Slobodskoy, "Impedance of Semiconductor-Insulator-Metal Capacitors," Solid-State Electronics, 7 (January 1964), pp. 59–79.
5. K. Lehovec, "Frequency Dependence of the Impedance of Distributed Surface States in MOS Structures," Appl. Phys. Lett., *8,* No. 2 (January 15, 1966), pp. 48–50.
6. E. H. Nicollian and A. Goetzberger, "The Si-SiO$_2$ Interface—Electrical Properties as Determined by the Metal-Insulator-Silicon Conductance Technique," B.S.T.J., *46,* No. 6 (July–August 1967), pp. 1055–1133.
7. S. Narayanan, "Transistor Distortion Analysis Using Volterra Series Representation," B.S.T.J., *46,* No. 5 (May–June 1967), pp. 991–1024.
8. M. H. White and J. R. Cricchi, "Characterization of Thin Oxide NMOS Memory Transistors," IEEE Trans. on Electron Devices, *ED-19,* No. 12 (December 1972), pp. 1280–1288.
9. U. Kämpf and H. G. Wagemann, "Radiation Damage of Thermally Oxidized MOS Capacitors," IEEE Trans. Electron Devices, *ED-23,* No. 1 (January 1976), pp. 5–10.
10. Members of Technical Staff, Bell Laboratories, Inc., *Transmission Systems for Communication,* Winston-Salem: Western Electric, 1971, p. 247.

11. J. E. Dennis, D. M. Gay, and R. E. Welsch, "An Adaptive Nonlinear Least-Squares Algorithm," NBER Working Paper No. 196 (1977).
12. L. I. Popova, P. K. Vitanov and B. Z. Antov, "Interface States in MNOS Systems," Thin Solid Films, *51*, No. 3 (June 15, 1978), pp. 305–309.
13. F. M. Nazar, "Low-Frequency Conductance of Thermally Grown SiO_2 Films," Int. J. of Electronics, *44*, No. 6 (June 1978), pp. 657–662.
14. H. J. J. De Man, "A Fast Algorithm for the Calculation of Junction Capacitance and Its Application for Impurity Profile Determination," Solid-State Electronics, *15* (February 1972), pp. 177–187.
15. H. K. Gummel and H. C. Poon, "An Integral Charge Control Model of Bipolar Transistors," B.S.T.J., *49*, No. 5 (May–June 1970), pp. 327–852.
16. L. W. Nagel and D. O. Pederson, "SPICE (Simulation Program with Integrated Circuit Emphasis)," Electronics Research Laboratory, University of California, Berkeley, Memorandum No. ERL-M382, 1973.
17. D. G. Duff, "Modeling High Current Effects in Bipolar Transistors for Use in Computer-Aided Distortion Analysis," D. Eng. Thesis, University of California, Berkeley, 1976.
18. D. G. Duff and R. G. Meyer, "An Extended Integral Charge Control Transistor Model for High Current DC, AC and Distortion Analysis," Proc. 1977 IEEE Internat. Symp. Circuits and Systems, pp. 19–22.
19. R. J. Whittier and D. A. Tremere, "Current Gain and Cutoff Frequency Falloff at High Currents," IEEE Trans. on Electron Devices, *ED-16*, No. 1 (January 1969), pp. 39–57.
20. A. Van der Ziel and D. Agouridis, "The Cutoff Frequency Falloff in UHF Transistors at High Currents," Proc. IEEE, *54*, No. 3 (March 1966), pp. 411–412.
21. J. Choma, "A Curve-Fitted Circuits Models for Bipolar Transistor f_T Roll-Off at High Injection Levels," IEEE J. Solid-State Circuits, *SC-11*, No. 2 (April 1976), pp. 346–348.
22. P. Gray, D. DeWitt, A. Boothryd, and J. Gibbons, *Physical Electronics and Circuit Models of Transistors*, New York: John Wiley, 1964, p. 141.
23. W. M. Fox et al., "SG Undersea Cable System: Semiconductor Devices and Passive Components," B.S.T.J., *57*, No. 7 (September 1978), pp. 2405–2434.
24. H. C. Poon, "Implication of Transistor Frequency Dependence on Intermodulation Distortion," IEEE Trans. on Electron Devices, *ED-21*, No. 1 (January 1974), pp. 110–112.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 58, No. 9, November 1979
Printed in U.S.A.

The Effects of Several Transmission Systems on an Automatic Speaker Verification System

By C. A. McGONEGAL, A. E. ROSENBERG, and L. R. RABINER

(Manuscript received May 29, 1979)

In an earlier report, the effects of several transmission systems on speaker verification by human listeners were investigated. It was shown that the transmission system played a significant role in the speaker verification process. In this paper, we show the effects of the transmission system on an existing automatic speaker verification system in which the measured features are pitch and gain as a function of time for a specified utterance. In this experiment, there were 10 male and 10 female customers and 40 male and 40 female impostors. Fifty utterances were recorded using a conventional telephone connection over a period of two months. All utterances were post-processed by an ADPCM *coding system and* LPC *vocoding system. When the reference and test utterances were subjected to different transmission systems, no significant difference in the verification accuracy of this automatic system was found. This result verifies that pitch and gain are robust features for use in a speaker verification system.*

I. INTRODUCTION

The automatic speaker verification problem has two aspects—the creation of a reference pattern and the determination of similarity between a test and a reference pattern. When verification is performed over dialed-up telephone lines, the transmission system used in the telephone plant is an additional factor that must be considered. In a recent subjective experiment,[1] the effects of adaptive differential pulse code modulation (ADPCM) coding and linear predictive vocoding (LPC) on the speaker verification accuracy of human listeners was investigated. It was shown that the verification task was easiest (most accurate) when homogeneous systems were used (i.e., the test and reference utterances were transmitted over the same system) and significantly more difficult (higher error rate) when mixed systems (i.e., different transmission systems) were used for the test and reference utterances. In this paper, we investigate how these same transmission systems affect machine verification accuracy using a system that has been studied for the past few years at Bell Laboratories.[2-6]

The automatic system is based on the analysis of fixed sentence-long utterances in which the verification features are the time variations (contours) of the pitch and gain (intensity) of the utterance. A training set of utterances (both customer and impostor) is required to establish a reference pattern and to choose weights and measurements for the verification process. Following time alignment of the reference and test contours, a combination of weighted Euclidian distances between a set of test and reference measurements is compared with a threshold to determine whether to accept or reject an identity claim.

In an extensive investigation of the automatic speaker verification system over dialed-up telephone lines, Rosenberg obtained an average verification accuracy of about 91 percent.[2] Rosenberg also found that some talkers tended to perform significantly worse than average, and some significantly better than average.

To investigate the behavior of the automatic speaker verification system on different transmission systems, a new data base of customer and impostor utterances was created. Dialed-up telephone connections were used in all recordings. Since the earlier work on human verification used wideband, high-quality recordings, the experiments with human listeners were repeated using the new data base. Following this, a series of experiments was run with the automatic verification system.

The key results of this study are:

(*i*) Human verification accuracy on the telephone speech was essentially the same as previously reported for the high-quality speech, i.e., the highest verification scores were obtained when the reference and test utterances were transmitted over the same system, and significantly lower verification scores were obtained when different transmission systems were used for the test and reference utterances.

(*ii*) Machine verification accuracy on the telephone speech was essentially independent of the transmission system used for the test and reference utterances.

These results tend to confirm the notion that pitch and gain are robust features for verification and hence are suitable for many applications.

The organization of this paper is as follows. The automatic speaker verification system is described in Section II. Section III describes the experimental procedure used to evaluate the automatic verification system. This section includes a description of the speech transmission systems, as well as the data base used in the evaluation. The machine verification and the human verification results are presented in Sections IV and V. Finally, in Section VI the main results of the experiment are discussed.

II. THE AUTOMATIC SPEAKER VERIFICATION SYSTEM

Although the operation of the automatic speaker verification system has been described previously,[2-6] a brief review is given here. A block diagram of the overall verification system is shown in Fig. 1. Two inputs are provided to the system. These are an identity claim which retrieves reference data associated with the claimed identity and a sentence-long sample utterance. The sample utterance is analyzed to extract time functions or contours of specified features which are compared with (previously obtained) reference contours. Reference contours are obtained by averaging and combining sets of contours obtained from training utterances from the individual whose identity is claimed. The features used in this experiment are the intensity (gain) and pitch period. The gain contour is normalized so that its peak

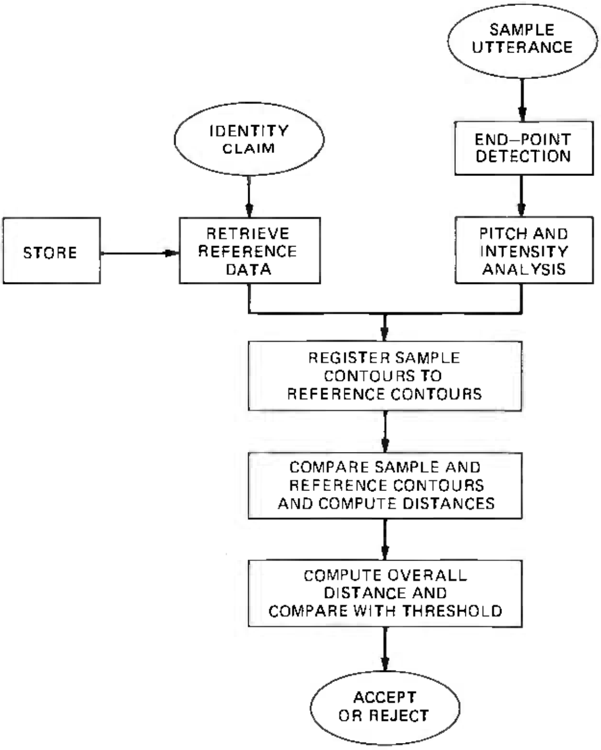

SAMPLE UTTERANCE
IDENTITY CLAIM
END–POINT DETECTION
STORE
RETRIEVE REFERENCE DATA
PITCH AND INTENSITY ANALYSIS
REGISTER SAMPLE CONTOURS TO REFERENCE CONTOURS
COMPARE SAMPLE AND REFERENCE CONTOURS AND COMPUTE DISTANCES
COMPUTE OVERALL DISTANCE AND COMPARE WITH THRESHOLD
ACCEPT OR REJECT

Fig. 1—Flow diagram of the verification system.

over the entire utterance is a fixed value and is low-pass filtered to give a smooth contour.

Before comparing sample and reference contours, time registration is carried out. Using a dynamic programming technique, the sample intensity contour is time-warped so that corresponding events in the sample and reference contours are aligned in time. The resulting time-warping function is also applied to the sample pitch period contour in order to align it to the reference contour. Following registration, the contours are divided into 20 equal length segments. In each segment, a set of measurements is applied to both the sample and reference contours. A squared difference is calculated specifying the dissimilarity between contours for each measurement and weighted inversely by a variance which is calculated from the set of training contours used to construct the reference. The effect of using the variance is to weight more heavily those segments in which a particular measurement is consistent over the set of training contours.[4] The various segment-by-segment measurements characterize the shape of the contours. In addition, the system computes distances based on the overall cross correlation of sample and reference contours (after time alignment) and distances based on the amount of warping requried to register the sample contours to the reference contours.

These distances are combined into an overall distance in two different ways. The first, the "overall distance" procedure, is a simple (unweighted) average over the entire set of individual distances. In the second procedure, the "selected distance" is a simple average calculated over a prespecified speaker-dependent subset of the entire set of distances. The subset is obtained as part of the training procedure by selecting those distances which are most effective in separating populations of customer and impostor utterances.

For either procedure, the combined distance is compared with a speaker-dependent threshold to determine whether to accept or reject the identity claim. The threshold distance, obtained from the training set and included in the reference data, is estimated from the overall distance distribution of distances from customer and impostor training utterances. Normally, a threshold is chosen to equalize the false (impostor) acceptance and false (customer) rejection rates. In the absence of direct knowledge of the costs of rejecting a customer or accepting an impostor, setting the threshold to give equal error rates yields the minimum cost. This is called the equal error criterion in this paper. In many real-world applications, the costs of these two types of errors would not be equal—e.g., in a banking situation the cost of rejecting a customer would be lower than the cost of accepting an impostor. In such cases, the threshold would be adjusted appropriately.

III. EXPERIMENTAL EVALUATION

To evaluate the automatic verification system of the previous section, a speech data base of utterances was created. The experimental setup for creating this speech data base is shown in Fig. 2. The speech was recorded in a sound booth over conventional dialed-up telephone lines. The signal was bandlimited from 100 to 3200 Hz (the nominal telephone bandwidth) and digitized at a 10-kHz rate. Both the reference and the test utterances were processed by one of the following three transmission systems:

(*i*) Clear channel—i.e., no additional processing.

(*ii*) Adaptive differential pulse code modulation (ADPCM) coding.

(*iii*) Linear predictive vocoding (LPC).

The ADPCM coder used in this experiment was a simulation of the coder built by Bates,[7] based on the work of Cummiskey et al.[8] Figure 3 is a block diagram of the ADPCM system. Since the required sampling rate for the ADPCM coder was 6 kHz, a sampling rate conversion system was used to convert from 10 to 6 kHz at the input to the coder.[9] The signal bandwidth was reduced to 2600 Hz for the ADPCM coder by using a 100- to 2600-Hz bandpass filter in the sampling rate conversion system. In the coder, a 4-bit adaptive quantizer was used to code the difference signal ($\delta(n)$ in Fig. 3), giving an overall bit rate of 24 kbits/s for the coder. The step-size multiplier of the quantizer ranged over a 41-dB range (i.e., the ratio between the smallest step size was 114 to 1). A first-order predictor was used with a multiplier coefficient of $\alpha = 0.9375$. Signal levels were chosen so that the coder was operating at approximately the optimum range.[8]

A block diagram of the LPC vocoder is given in Fig. 4. The implementation was based on the autocorrelation method of linear prediction.[10-12] Pitch detection and voiced-unvoiced decision were performed

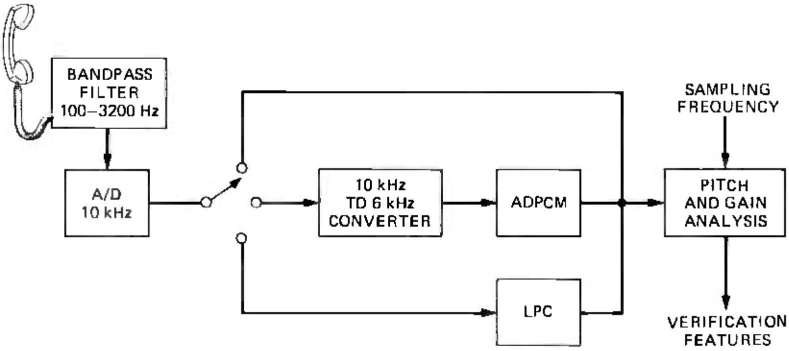


Fig. 2—Block diagram of the data collection system.

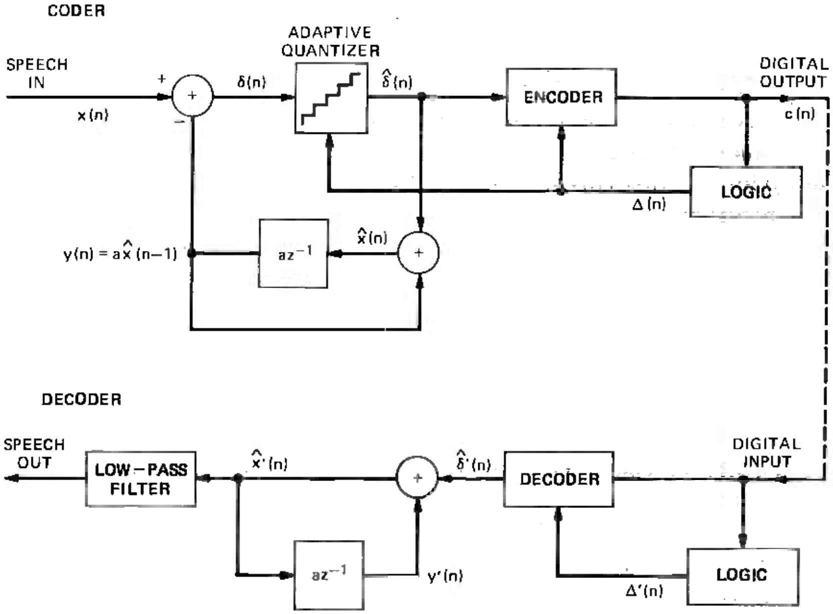


Fig. 3—Block diagram of the ADPCM coder.

using the modified autocorrelation pitch detector of Dubnowski et al.[13] A 12-pole LPC analysis was performed using a pitch-adaptive, variable frame size, at a rate of 100 frames per second.[14] No quantization of the LPC parameters was used in this experiment.

To evaluate the effects of the three transmission systems on verification accuracy, a data base was designed which included:

(*i*) Fifty recordings made by each of 20 experienced talkers (10 male and 10 female) over a period of two months. The first 10 recordings were made once a day; the remaining 40 were made twice a day (morning and afternoon). These talkers were designated "customers."

(*ii*) One recording made by each of 80 naive talkers (40 male and 40 female). These talkers were designated "impostors." There was no attempt to mimic the "customers."

Two all-voiced sentences were used in the recordings. The males used the sentence, "We were away a year ago," and the females used the sentence, "I know when my lawyer is due." In previous studies, only the first sentence was used.[2-5]

Since the automatic speaker verification system used pitch and intensity contours as features, these contours were measured once and stored on disk for later retrieval in the experiment.

3.1 Reference construction

For each customer and each transmission system, pitch and intensity contours from 10 of the 50 utterances were used to construct "refer-

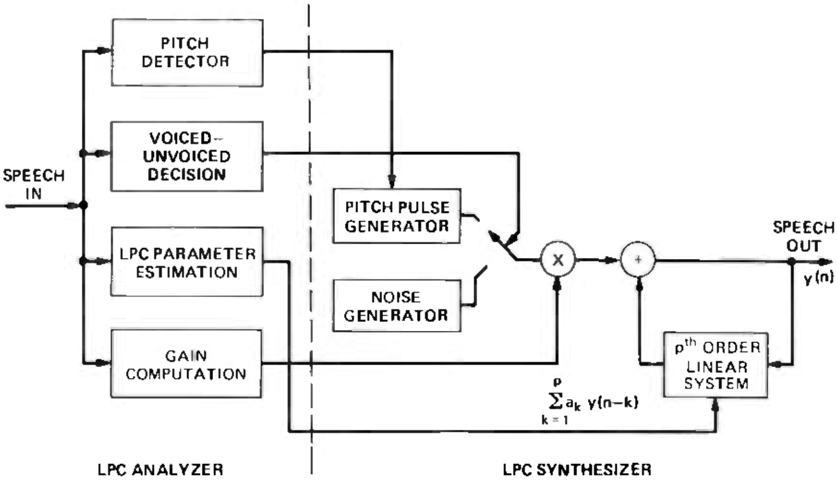


Fig. 4—Block diagram of the LPC vocoder

ence" contours. As discussed previously, in order to complete the reference construction (i.e., to get weights, thresholds, and to choose selected distances), 10 additional utterances of the 50 were used, along with the pitch and intensity contours of 15 impostors of the same sex as the customer. Two different sets of reference files were created for each customer—one obtained from the first 20 consecutive recordings of the customer (method 1) and the other obtained by using two of every five utterances recorded (method 2).

Figure 5 is a series of plots of relative cumulative frequency distributions of customer and impostor samples as a function of combined distance. Each column contains the results for two female and two male customers using the (method 1) training data. In each plot, the customer sample distributions (on the left) show the fraction of samples with distances greater than the abscissa value while the impostor sample distributions (on the right) show the fraction of samples with distances less than the abscissa value. The first column shows the results for clear channel utterances; the second column shows results for LPC vocoded utterances, and the third column shows results for ADPCM coded utterances. The decision threshold is chosen as the distance where the cumulative distributions cross (i.e., the equal error threshold), or, in the case when the distributions are separated, the point midway between the ends of the separate distributions. Since there was only a small number of training utterances, the distributions for the two types of errors are poorly defined and only a rough estimate of the decision threshold is obtained. It should be clear that the equal error threshold will vary for each pair of transmission systems being compared, as well as for different talkers. For the four talkers shown in Fig. 5, the *worst* equal error threshold indicates a 10-percent error

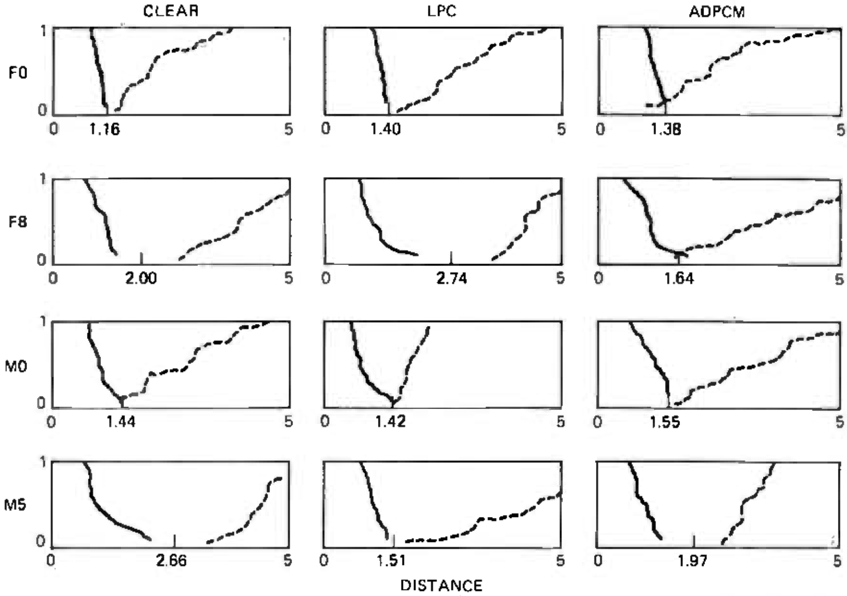


Fig. 5—Plots of the cumulative distributions of the errors for four talkers and three transmission systems based on the training set of utterances.

rate; however, for the 20 talkers, the average equal error rate over all conditions was about 1 percent. Figure 5 also shows that the value of the threshold distance varies considerably (2.5 to 1) across talkers, as does the shape of the cumulative distributions. These results tend to indicate that the training data are inadequate for obtaining a good estimate of the equal error rate threshold.

IV. RESULTS ON AUTOMATIC SPEAKER VERIFICATION

To test the automatic speaker verification system for each transmission system, each customer reference was compared to the 30 customer utterances and the 25 impostor utterances which were not used in the training set. For each set of comparisons, a Type 1 (customer rejection) and a Type 2 (impostor acceptance) error score was measured. If we denote the Type 1 error scores as E_1 and the Type 2 error scores as E_2, then E_1 and E_2 are functions of:

(*i*) The transmission system used in the training, *i*, where $i = 1$ denotes the clear channel, $i = 2$ denotes the LPC vocoder, and $i = 3$ denotes the ADPCM coder. The mnemonics *C*, *V*, and *A* are used in the plots to denote clear channel, LPC vocoder, and ADPCM coder, respectively.

(*ii*) The transmission system used in the testing, *j*, where $j = 1$, 2, and 3 are identical to $i = 1$, 2, and 3.

(*iii*) The training method, k, where $k = 1$ denotes training method 1 and $k = 2$ denotes training method 2.

(*iv*) The type of measurements used in the verification distance, l, where $l = 1$ is selected measurements (speaker specific), and $l = 2$ is overall measurements (speaker independent).

(*v*) The customer, m, where $m = 1$ to 10 for the 10 male customers and $m = 11$ to 20 for the female customers.

Since different sentences were used for male and female customers, results are presented separately for each subset of the customers.

To illustrate some of the results, Fig. 6 shows plots of E_1 and E_2 as a function of the training system, testing system pair (i, j), and talker (m), for selected distance measurements $(l = 1)$, and training method 2 $(k = 2)$. Figures 6a and 6b are E_1 scores for male customers $(m = 1$ to $10)$, and female customers $(m = 11$ to $20)$, and Figs. 6c and 6d are E_2 scores for male and female customers. A bar graph denotes the error score for each condition. The reader should note that within each group there are 10 bars, some of which are 0 indicating zero error. From this figure the following observations can be made:

(*i*) E_2 scores are significantly smaller than E_1 scores, indicating that the distance threshold obtained from the training set for equal errors $(E_1 = E_2)$ was not a stable point.

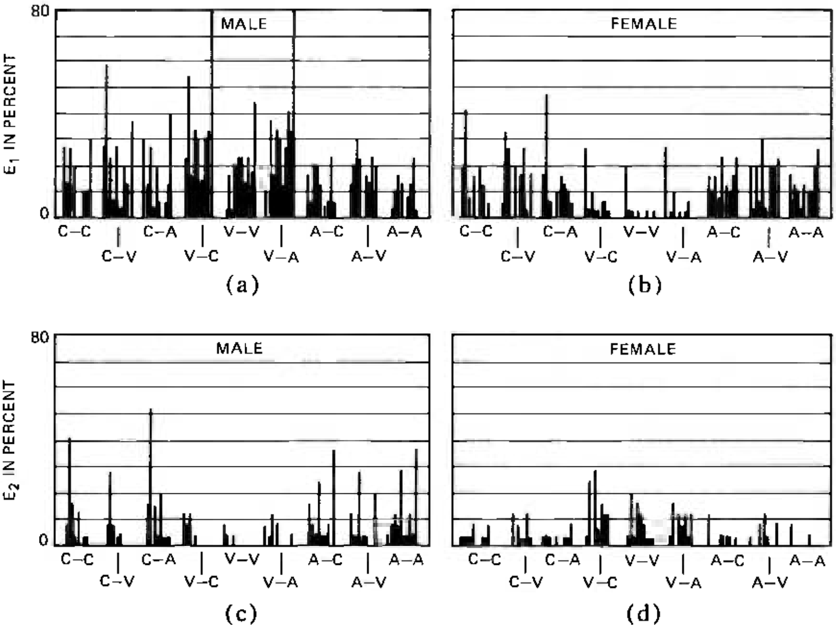


Fig. 6—Plots of E_1 and E_2 versus system pair and talker for male and female talkers (training method 2 using selected distance measurements).

(*ii*) A high degree of variability in scores (for both E_1 and E_2 and for male and female customers) exists among customers for each pair of transmission systems. This type of result was also obtained earlier by Rosenberg.[2]

(*iii*) Error scores for female customers are somewhat smaller than error scores for male customers.

(*iv*) The variability between scores for pairs of transmission systems is smaller than the variability of scores within a pair of transmission systems.

Based on the above observations, it is clear that the E_1 and E_2 scores cannot simply be averaged over customers because of the high variability among customers. Thus two data processing procedures were tried, one to use the median over customers, the other to statistically eliminate the extremes of the distribution of error scores (using a recently described method[15]) and then to average the scores of the remaining customers. Both data processing methods yielded essentially the same results and hence only the results of taking medians are presented here.

Table I gives values of the medians of E_1 and E_2 over (male and female) customers for each (i, j) pair, and for $k = 1$ and 2, and $l = 1$ and 2. Also included in this table is the median of the quantity

$$E_3 = (E_1 + E_2)/2,$$

which is the average error rate of the system. Since E_1 and E_2 were significantly different, E_3 provides a better measure of the overall performance of the verification system than either E_1 or E_2. It can be seen from Table I and statistically verified at the 0.001 level that training method 2 provides significantly better scores than training method 1, and that using selected measurements for the distance score provides significantly better scores than the overall measurements. As such, we restrict our discussion to this case only, i.e., selected measurements from training method 2.

Figure 7 is a plot of the E_3 median scores for each pair of transmissions systems. Although there is some variability in score among these systems, the variability is statistically insignificant (at the 0.01 level). Thus the major result of the testing is that the verification accuracy is relatively insensitive to the transmission system used for training and testing. This result is very different from the one obtained when verification is performed by human listeners as discussed previously. To ensure that the human verification accuracy for this new data base remained the same, the perceptual verification experiment was repeated, and the results are given in the next section.

Before the results of the perceptual experiment are described, Fig. 8 illustrates the variability of the distance threshold in the testing.

Table 1—Median over customers of the error rates

	E_1		E_2		E_3	
	male	female	male	female	male	female
C-C	16.66	11.66	20.0	8.0	17.83	14.01
C-V	25.00	10.00	16.0	8.0	24.50	14.00
C-A	15.00	11.66	20.0	8.0	18.83	15.84
V-C	35.00	18.33	8.0	16.0	23.50	20.16
V-V	25.00	5.00	8.0	8.0	15.83	7.67
V-A	30.00	11.66	6.0	8.0	19.00	12.17
A-C	18.33	18.33	16.0	6.0	23.00	13.01
A-A	20.00	10.00	8.0	4.0	22.83	12.16
A-V	18.33	16.67	12.0	6.0	22.66	13.51

(a) Overall Measurement, Training Method 1

	E_1		E_2		E_3	
	male	female	male	female	male	female
C-C	23.33	16.67	8.0	4.0	17.83	9.16
C-V	28.33	20.00	2.0	2.0	16.67	13.17
C-A	21.66	15.00	6.0	2.0	14.34	7.84
V-C	35.00	15.00	4.0	4.0	19.66	12.33
V-V	28.33	6.67	4.0	4.0	16.17	5.17
V-A	36.67	13.33	4.0	4.0	20.33	9.33
A-C	23.33	13.33	4.0	4.0	19.66	9.99
A-A	26.66	13.33	4.0	2.0	17.33	11.16
A-V	30.00	15.00	4.0	4.0	23.83	10.17

(b) Selected Measurement, Training Method 1

	E_1		E_2		E_3	
	male	female	male	female	male	female
C-C	11.66	6.67	22.0	12.0	21.50	9.83
C-V	18.33	10.00	14.0	6.0	16.16	9.00
C-A	20.00	8.30	18.0	12.0	17.33	8.16
V-C	30.00	11.67	8.0	14.0	24.00	11.01
V-V	18.33	5.00	4.0	12.0	13.00	9.16
V-A	33.33	8.33	8.0	10.0	22.83	11.17
A-C	20.00	15.00	8.0	10.0	17.83	13.34
A-V	21.66	18.33	6.0	12.0	17.00	10.00
A-A	16.67	15.00	8.0	8.0	16.16	11.50

(c) Overall Measurement, Training Method 2

	E_1		E_2		E_3	
	male	female	male	female	male	female
C-C	11.66	11.66	4.0	4.0	10.17	8.00
C-V	21.66	16.67	2.0	4.0	14.51	10.34
C-A	13.33	10.00	4.0	4.0	12.50	7.17
V-C	30.00	3.33	0.0	10.0	16.66	8.50
V-V	20.00	3.33	0.0	4.0	11.01	3.84
V-A	31.66	3.33	2.0	6.0	16.66	5.67
A-C	10.00	15.00	6.0	4.0	11.51	8.33
A-V	18.33	20.00	4.0	2.0	11.84	10.00
A-A	10.00	11.66	8.0	0.0	11.67	6.67

(d) Selected Measurement, Training Method 2

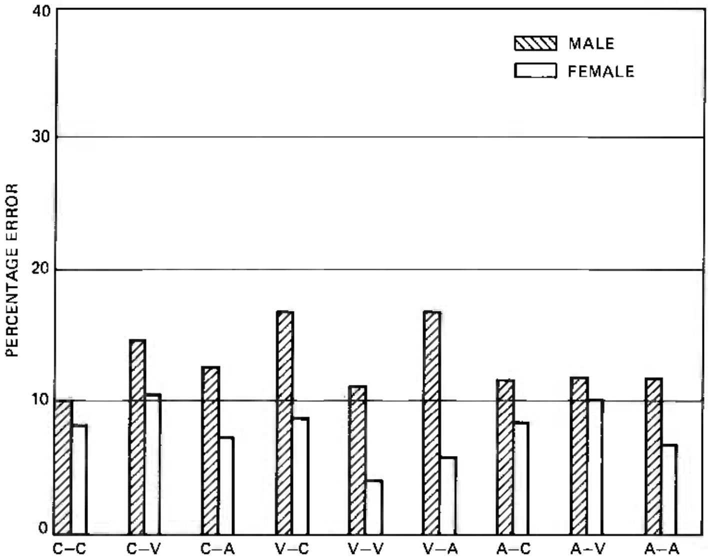


Fig. 7—Overall median error rate as a function of system pair for male and female talkers.

This figure shows the sum of the cumulative error distributions for the testing results for several different cases. Each part of the figure contains three vertical lines. The solid vertical line is the *a priori* distance threshold which gives an equal error based on the training data. The dashed vertical line is the *a posteriori* distance threshold that gives equal error based on the testing data. The dotted vertical line is the threshold that minimizes the total error E_3 for the testing data. In the ideal case, all three thresholds would be equal. However, because of the inadequacy of the training data and the unusual shapes of the cumulative error distribution, thresholds were different, as seen in this figure. The variability in both distance thresholds and error scores is fairly large (typically, between 10 and 30 percent in most cases).

V. HUMAN VERIFICATION TEST RESULTS

To verify that the human verification accuracy was strongly affected by the speech transmission system when using the telephone recordings, the experiment performed in Ref. 1 was repeated exactly. The results of these tests are given in Figs. 9 and 10. Figure 9 shows false alarm (customer rejection or Type 1 error) and miss (impostor acceptance or Type 2 error) rates for the male and female customers as a function of the pair of transmission systems used in the comparison. Only the first eight customers (female and male) were used to corre-

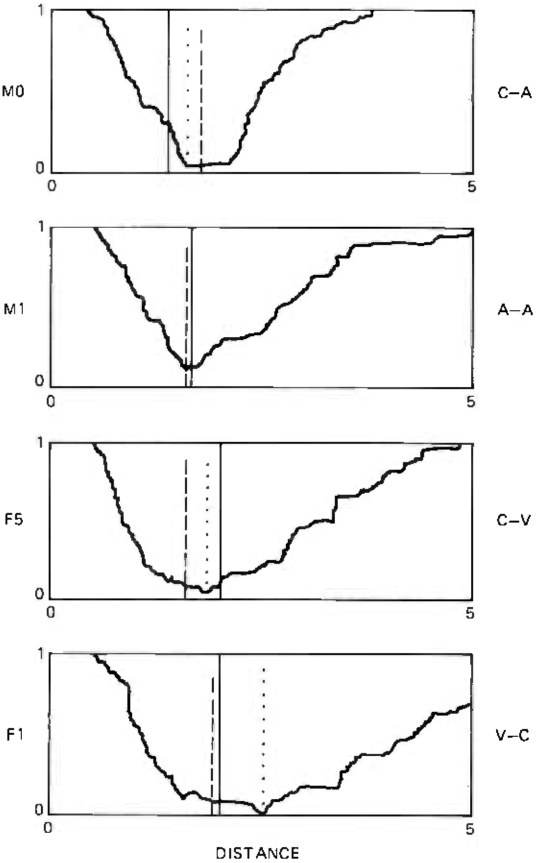


Fig. 8—Total error ($E_1 + E_2$) as a function of distance threshold for four conditions for the testing data.

spond to the eight customers used in the earlier experiment. This figure again shows the high degree of variability of the scores among customers. Thus, to combine customer scores, a median was used instead of averaging. Figure 10 shows the median false alarm rate, miss rate, and overall error rate for each pair of speech transmission systems.

The results shown in Figs. 9 and 10 are essentially identical to those reported previously,[1] namely, that the false alarm rates for transmission pairs which were the same were statistically significantly lower than for mixed systems, whereas the miss rates for homogeneous systems were larger than for mixed systems. Statistical comparisons showed that, in this experiment, the results were not statistically significantly different from those of the earlier experiment in any category.

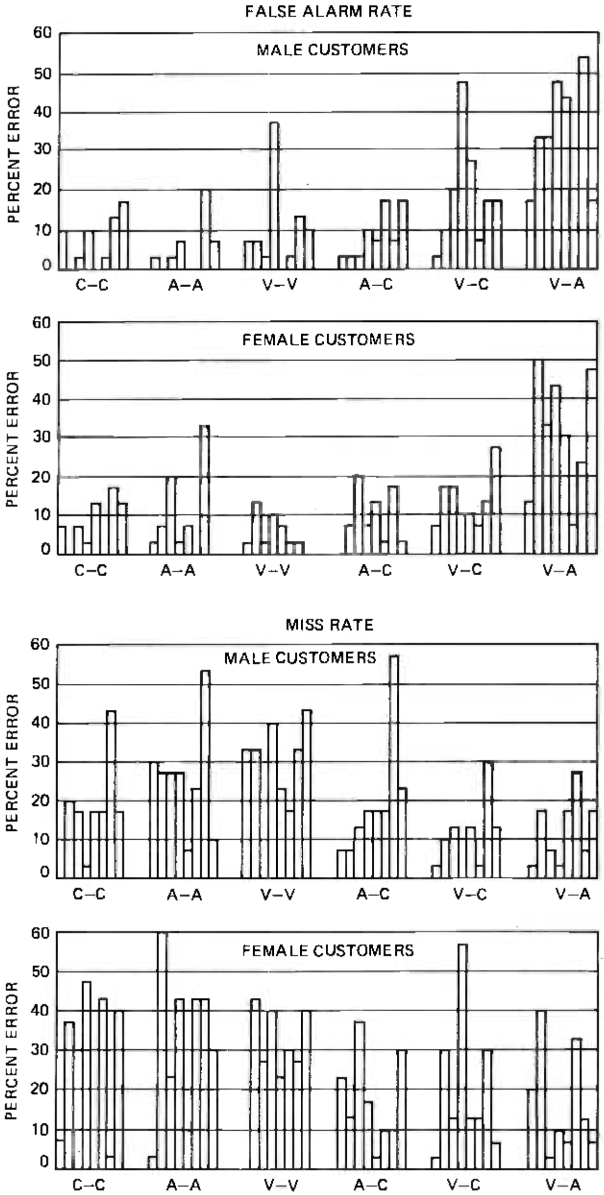


Fig. 9—False alarm and miss rates as a function of the system pair for male and female customers for the human verification experiment.

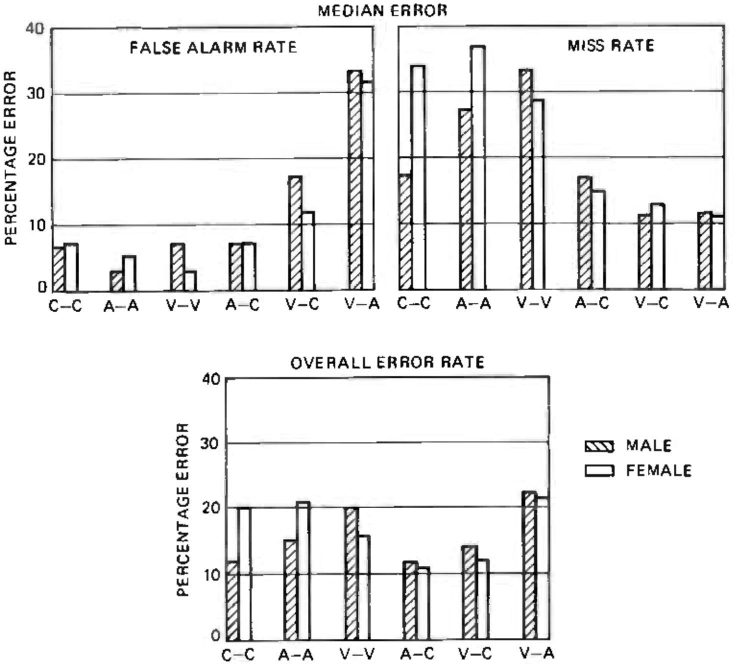


Fig. 10—Median false alarm, miss, and overall error rates as a function of the system pair for male and female customers for the human verification experiment.

VI. DISCUSSION

The major result of this investigation was the finding that, for an automatic speaker verification, accuracy was statistically insensitive to either ADPCM coding or LPC vocoding of the speech utterance for either (or both) the reference and test utterances. For this same data base, the verification accuracy by human listeners was significantly affected by the different transmission systems in the manner described in Ref. 1. The conclusion drawn from these results is that pitch and gain are reasonably robust to the distortions of ADPCM coding and LPC vocoding, thereby enabling the automatic system to be insensitive to these transmission systems.

The overall verification accuracy in this system was about 12 percent for male talkers and 8 percent for female talkers. These verification rates are comparable to those obtained by Rosenberg in a large experiment over dialed-up telephone lines.[2] As in the earlier work, considerable variability in verification scores among talkers was found, again indicating that the variability of pitch and gain for some talkers is large, and thus for these talkers other feature sets should be considered for verification. Recent unpublished investigations by Furui indicate significantly smaller (on the order of 0.5 percent) error rates when the features are cepstral coefficient contours rather than pitch and gain. Whether such feature sets are robust to coding and vocoding remains to be investigated.

It was found in this investigation that the training methods used were inadequate for giving stable reference contours and reliable distance thresholds. This effect was previously noted by Rosenberg,[2] Furui,[16] and Furui et al.,[17] who showed that long-time variability in feature contours had to be taken into consideration to obtain stable reference data for verification.

Two other effects were noted during the course of this study. First, it was found that selected distances provided significantly better scores than overall distances. Rosenberg also noted that, once a reasonable amount of training data was obtained, selected distances were better than overall distances.[2] Thus the results here indicate that 10 training utterances are sufficient for selected distance scores to be superior to overall distance scores. The second point concerned the lower error rates for female talkers than for male talkers. Rosenberg found no statistically significant differences between verification scores for males and females.[2] Thus, this result may be due to the difference in sentence used in the verification task. If this is true, then the implication is that the test utterance chosen may provide small but consistent improvements in the verification scores.

VII. SUMMARY

In this paper, we have shown that, whereas the false alarm and miss rates for verification by human listeners are strongly affected by the pair of transmission systems used for the reference and test utterances, the false alarm and miss rates for an automatic verification system based on pitch and gain are relatively insensitive to the transmission system in the case of ADPCM coding and LPC vocoding. Although the average overall error rate for this system was around 10 percent, the robustness of pitch and gain to transmission systems makes them attractive features for automatic speaker verification systems.

REFERENCES

1. C. A. McGonegal, L. R. Rabiner, and B. J. McDermott, "Speaker Verification by Human Listeners Over Several Speech Transmission Systems," B.S.T.J., *57*, No. 8 (October 1978), pp. 2887–2900.
2. A. E. Rosenberg, "Evaluation of an Automatic Speaker Verification System over Telephone Lines," B.S.T.J., *55*, No. 6 (July–August 1976), pp. 723–744.
3. G. R. Doddington, "A Computer Method of Speaker Verification," Ph.D. Dissertation, Department of Electrical Engineering, University of Wisconsin, 1970.
4. R. C. Lummis, "Speaker Verification by Computer Using Speech Intensity for Temporal Registration," IEEE Trans. Audio and Electroacoust., *AU-21* (April 1973), pp. 80–89.
5. A. E. Rosenberg and M. R. Sambur, "New Techniques for Automatic Speaker Verification," IEEE Trans. Acoustics, Speech and Signal Processing, *ASSP-23* (April 1975), pp. 169–176.

6. A. E. Rosenberg, "Automatic Speaker Verification: A Review," Proceedings of the IEEE, *64,* No. 4 (April 1976), pp. 475–487.
7. S. L. Bates, "A Hardware Realization of a PCM-ADPCM Code Converter," M.I.T. M.S. Thesis, Dept. of Elec. Eng. and Comp. Sc., January 1976.
8. P. Cummiskey, N. S. Jayant and J. L. Flanagan, "Adaptive Quantization in Differential PCM Coding of Speech," B.S.T.J., *52,* No. 7 (September 1973), pp. 1105–1118.
9. R. E. Crochiere and L. R. Rabiner, "Optimum FIR Digital Filter Implementations for Decimation, Interpolation, and Narrow-Band Filtering," IEEE Trans. on Acoustics, Speech and Signal Proc., *ASSP-23,* No. 5 (October 1975), pp. 444–456.
10. J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech,* New York: Springer-Verlag, 1976.
11. J. D. Markel and A. H. Gray, Jr., "A Linear Prediction Vocoder Simulation Based Upon an Autocorrelation Method," IEEE Trans. on Acoustics, Speech and Signal Proc., *ASSP-22,* No. 2 (April 1974), pp. 124–134.
12. B. S. Atal and S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," J. Acoust. Soc. Am., *50* (1971), pp. 637–655.
13. J. J. Dubnowski, R. W. Schafer, and L. R. Rabiner, "Real-Time Digital Hardware Pitch Detector," IEEE Trans. Acoust., Speech, Signal Proc., *ASSP-24* (February 1976), pp. 2–8.
14. L. R. Rabiner, "On the Use of Autocorrelation Analysis for Pitch Detection," IEEE Trans. Acoust. Speech, and Signal Proc., *ASSP-25,* No. 1 (February 1977), pp. 24–33.
15. H. J. Chen, unpublished work.
16. S. Furui, "Effects of Long-Term Spectral Variability on Speaker Recognition," Paper NNN28, Acoustical Society Meeting, Honolulu, 1978.
17. S. Furui, F. Itakura, and S. Saito, "Personal Informantion in the Long-Time Averaged Speech Spectrum," Review of Elec. Comm. Lab, *23,* No. 9–10 (September–October 1975), pp. 1133–1141.

Copyright © 1979 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 58, No. 9, November 1979
Printed in U.S.A.

Periodic Sequences that Facilitate Data Set Spectrum Measurements

By D. C. RIFE

(Manuscript received May 3, 1979)

It is often desirable to measure the spectrum of the modulated baseband pulse of a quadrature amplitude modulated data set. The usual method, which is not very satisfactory, is to send a pseudorandom data sequence into the data set and measure the transmitted signal with a spectrum analyzer. Here we present a new technique, based on the properties of so-called "perfect sequences," which facilitates stable and accurate spectrum measurements.

I. INTRODUCTION

The modulated pulse spectrum (MPS), defined below, of a data set's signal is a basic characteristic of the data set that is of interest, for example, in determining how much bandwidth the signal requires. One can usually calculate an MPS, but it is often necessary to confirm the calculation by experiment. The usual practice in such experiments is to apply a binary pseudorandom sequence (also called maximum-length sequence[1]) to the input of a data set and measure the resulting output signal with a spectrum analyzer. This technique usually leads to a picture that shows the MPS approximately, but not exactly and not in a repeatable manner; the reasons are given in Section II.

The output of the spectrum analyzer is usually not constant at any one frequency setting, and the relative level at each frequency is not proportional to the MPS. An example is shown on Fig. 1.

If a data set modulator is a member of the quadrature amplitude modulation (QAM) class, which includes the differential phase shift keyed modulators, then there exist easily generated periodic constellation point sequences whose resulting signals have line spectra with amplitudes that are proportional to the modulator's MPS, and thus facilitate accurate experimental measurement of the MPS. The derivation and application of these sequences are the subject of this paper.

The next section establishes the properties required of the constel-

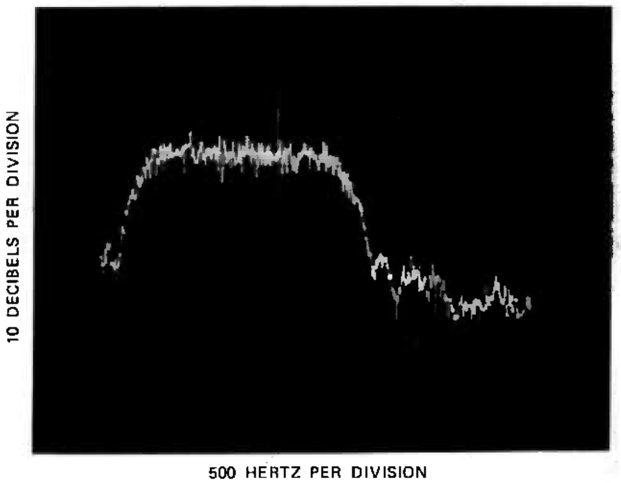


Fig. 1—Measured spectrum of a data set using the usual pseudorandom data method.

lation point sequences. Section III shows how the desired sequences can be derived from pseudorandom sequences and presents an example with experimental verification. Section IV presents another example and Section V contains the concluding remarks.

II. DATA SET SIGNALS

A QAM data set generates a signal that is often written as the real part of a complex signal, say $z(t)$, given by

$$z(t) = \sum_n C_n y(t - nT) e^{j\omega_c t}, \tag{1}$$

where ω_c is the carrier frequency, T is the intersymbol time, and $y(t)$ is a complex pulse waveform.[2,3] Each C_n is a complex constant selected from a finite set of allowable values by an algorithm that operates on the binary data to be transmitted (the input data stream). In general, several consecutive input bits determine each C_n. For example, the allowable real and imaginary parts of C_n may be ±1 and ±3, which gives 16 possible values (constellation points).

The real part of $z(t)$ is often passed through an analog filter, perhaps an equalizer or a low-pass filter, before being sent to the channel. Thus, if $f(t)$ is the impulse response of the filter, then the output signal is

$$s(t) = \mathrm{Re}\{s_0(t)\}, \tag{2}$$

where

$$s_0(t) = z(t) * f(t). \tag{3}$$

If the data set uses an analog modulator, then $y(t)$ is a real baseband pulse.

If the modulator is one that generates a step approximation to the corresponding continuous signal (digital implementation with D/A converter), then

$$y(t) = e^{-j\omega_c t} \sum_k h_k e^{jk\omega_c \tau} p(t - k\tau), \tag{4}$$

where the h_k are samples of a corresponding real baseband pulse, τ is the step width, and

$$p(t) = \begin{cases} 1, |t| < \tau/2 \\ ½, |t| = \tau/2 \\ 0, |0| > \tau/2. \end{cases} \tag{5}$$

We assume T/τ is an integer.

The Fourier transform of $s_0(t)$ is

$$S_0(\omega) = F(\omega) Y(\omega - \omega_c) C(\omega - \omega_c), \tag{6}$$

where $F(\omega)$ and $Y(\omega)$ are the Fourier transforms of $f(t)$ and $y(t)$ and

$$C(\omega) = \sum_n C_n e^{-jnT\omega}. \tag{7}$$

In the step approximation case,

$$Y(\omega - \omega_c) = P(\omega) H(\omega - \omega_c), \tag{8}$$

where

$$P(\omega) = \frac{\tau \sin(\omega\tau/2)}{(\omega\tau/2)} \tag{9}$$

and

$$H(\omega) = \sum_k h_k e^{-jk\tau\omega}. \tag{10}$$

We define the modulated pulse spectrum (MPS), for either case, to be

$$B(\omega) = F(\omega) Y(\omega - \omega_c). \tag{11}$$

Then

$$S_0(\omega) = B(\omega) C(\omega - \omega_c). \tag{12}$$

The Fourier transform of $s_0(t)$ is the product of two parts, the MPS and a function of the transmitted constellation sequence. The MPS is closely related to the spectral density function of the signal, under suitable assumptions about the statistics of random C_n. For an example, see Ref. 4. When characterizing a data set signal, it is important to know the magnitude of the MPS and to be able to measure it.

If $\{C_n\}$ is a periodic sequence of period M, then $C(\omega)$ becomes

$$C(\omega) = A(\omega) \sum_{m=-\infty}^{\infty} \omega_0 \delta(\omega - m\omega_0), \tag{13}$$

where

$$\omega_0 = 2\pi/MT \tag{14}$$

and

$$A(\omega) = \sum_{n=0}^{M-1} C_n e^{-jnT\omega}. \tag{15}$$

Thus $s(t)$ consists of tones at frequencies $\omega_c + m\omega_0$ and levels given by $(1/MT) | B(\omega_c + m\omega_0) A(m\omega_0) |$. Therefore, if points along $| B(\omega) |$ are to be generated, one needs to select $\{C_n\}$ so that $| A(m\omega_0) |$ is a constant for all m. Notice that the periodic set of $A_m = A(m\omega_0)$ is the discrete Fourier transform (DFT) of $\{C_n\}_{n=0}^{M-1}$.

If the input binary data sequence is periodic and the usual types of scramblers and coders are used in the data set, then the resulting C_n sequence will be periodic. However, the corresponding A_m seldom have equal magnitude. That is one reason a spectrum analyzer gives a measurement of the sort shown in Fig. 1. The other reason is that the period, M, is usually so large that several of the resulting tone frequencies are in the narrow passband of the spectrum analyzer at the same time. The envelope of such a signal is modulated, giving an unstable spectrum analyzer output.

In the next section, we show how to overcome these problems.

III. PERFECT SEQUENCES

The autocorrelation function of a periodic sequence $\{C_n\}$, with period M, is:

$$R_c(k) = \sum_{n=0}^{M-1} C_{k+n} C_n^*. \tag{16}$$

If

$$R_c(k) = \begin{cases} \alpha_1, & k = 0 \text{ modulo } M \\ \alpha_2, & \text{otherwise}, \end{cases} \tag{17}$$

where α_1 and α_2 are constants, then the sequence $\{C_n\}$ is called a *perfect sequence.*[5]

One can show that, if

$$A_m = \sum_{n=0}^{M-1} C_n e^{-j2\pi nm/M}, \tag{18}$$

then

$$|A_m|^2 = \sum_{n=0}^{M-1} R_c(n) e^{-j2\pi nm/M}. \tag{19}$$

If $\{C_n\}$ is a perfect sequence, then

$$|A_m|^2 = \begin{cases} \alpha_1 + (M-1)\alpha_2, & m = 0 \text{ modulo } M \\ \alpha_1 - \alpha_2, & m \neq 0 \text{ modulo } M. \end{cases} \tag{20}$$

Thus the desired property of equal magnitude A_m can almost be achieved if $\{C_n\}$ is a perfect sequence (the exception is A_0). We next show how to construct perfect C_n sequences.

There are many perfect binary sequences.[6] Two are:

11011100010

1101000001000

One can also show that pseudorandom binary sequences are perfect sequences. An example is the sequence generated by

$$X_n = X_{n-3} \oplus X_{n-5}, \tag{21}$$

where ⊕ denotes modulo 2 addition. The period of this sequence is 31.

If X_n is a pseudorandom sequence of period M, then its autocorrelation function has two values:

$$R_x(k) = \begin{cases} \dfrac{M+1}{2}, & k = 0 \text{ modulo } M \\ \dfrac{M+1}{4}, & \text{otherwise.} \end{cases} \tag{22}$$

A perfect C_n sequence could be generated by simply using $C_n = X_n$. However, most QAM data sets do not include the value zero in their constellation. Thus, it is necessary to apply a linear transformation to a perfect binary sequence, $\{X_n\}$, to obtain a perfect sequence $\{C_n\}$ of allowable values.

Suppose α and β are two complex constants and $C_n = \alpha X_n + \beta$, then the autocorrelation function of $\{C_n\}$ is

$$\begin{aligned} R_c(k) &= \alpha\alpha^* R_x(k) + (\alpha^*\beta + \alpha\beta^*) \sum_{n=0}^{M-1} X_n + M\beta\beta^* \\ &= \alpha\alpha^* R_x(k) + (\alpha^*\beta + \alpha\beta^*) \frac{M+1}{2} + M\beta\beta^*. \end{aligned} \tag{23}$$

If $R_x(k)$ has the two levels given by (22), then

$$R_c(k) = \begin{cases} \alpha\alpha^* \dfrac{M+1}{2} + (\alpha^*\beta + \alpha\beta^*)\dfrac{M+1}{2} + M\beta\beta^*, & k = 0 \\ \\ \alpha\alpha^* \dfrac{M+1}{4} + (\alpha^*\beta + \alpha\beta^*)\dfrac{M+1}{2} + M\beta\beta^*, & k \neq 0. \end{cases} \tag{24}$$

Suppose, for example, we want each C_n to have one of two of the 16 possible values given in the 16-point constellation example above, say, $-3 - j3$ and $3 + j3$. Then we can choose $\alpha = 6 + j6$ and $\beta = -3 - j3$. The corresponding values of A_m, with $M = 31$, are

$$|A_m|^2 = \begin{cases} 18, & m = 0 \\ \\ 576, & m \neq 0. \end{cases} \tag{25}$$

Consider a data set with a carrier frequency of 1800 Hz and a symbol rate of 2400 baud. If the C_n sequence just described modulates the data set, the output signal will consist of tones at frequencies of

$$1800 + \frac{2400}{31} m \text{ Hz}$$

and the carrier level (corresponding to A_0) will be about 15 dB below the level of its two neighboring tones.

This example has been simulated by V. B. Lawrence using an experimental modem. The resulting measured spectrum is illustrated in Fig. 2. This can be compared to the calculated MPS for that data set,

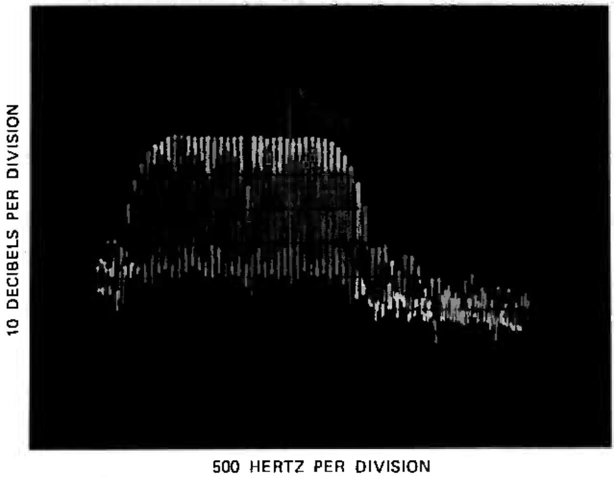


Fig. 2—Measured spectrum of a data set using the proposed method.

shown on Fig. 3. Notice that the carrier level on Fig. 2 is about 15 dB down from the level of its neighbors.

This technique can be applied to other types of QAM data sets by selecting binary sequences with desired periods and by choosing suitable values for the constants α and β.

IV. ANOTHER EXAMPLE

With some types of QAM data sets, it is feasible to find an input data sequence that will cause the data set to generate a perfect C_n sequence. An example is described below.

A differential PCM data set is on the market that uses unity-valued $C_n = e^{j\theta_n}$ and encodes consecutive pairs of bits (dibits) according to Table I.

One can cause the data set to generate a perfect sequence of C_n with the values 1 and -1 by an appropriate dibit sequence. To use a simple example, start with the 7-bit pseudorandom sequence 1110010. The corresponding periodic C_n sequence is:

$$1, 1, 1, -1, -1, 1, -1.$$

The periodic sequence of phase shifts must be

$$\pi, 0, 0, \pi, 0, \pi, \pi.$$

From the table, the corresponding periodic input data sequence is 11 10 10 11 10 11 11.

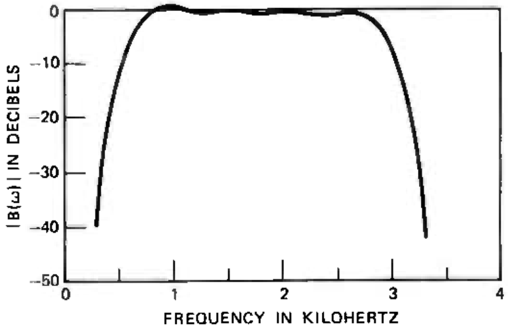


Fig. 3—Computed MPS of example data set.

Table I

Dibit	$\theta_n - \theta_{n-1}$
10	0
01	$\pi/2$
11	π
00	$-\pi/2$

If this periodic sequence were the input to the data set, the bits would be grouped by the data set into the desired dibit sequence or the following sequence, depending upon initial conditions:

$$11 \quad 11 \quad 01 \quad 01 \quad 11 \quad 01 \quad 11.$$

This sequence will produce a C_n sequence of period 28 that is not a perfect sequence. The spectrum of the line signal would be easily distinguished from the desired spectrum.

V. CONCLUSIONS

We have presented a simple method of generating a constellation point sequence such that, when it is applied to a QAM data set, the peaks (or envelope) of the resulting spectrum will be very nearly the MPS. The procedure is twofold: (*i*) select a period, M, small enough that the resulting tones can be separated by the spectrum analyzer and (*ii*) force $\{C_n\}$ to be a perfect sequence. The advantage of the method is that the spectrum can be accurately and repeatably measured with a spectrum analyzer. The accuracy is such that the observed spectrum will represent the MPS to within a few tenths of a decibel if the modulator is working properly. Modulator defects that would be obscured by the usual MPS estimation method are clearly evident when the proposed method is used.

REFERENCES

1. W. W. Peterson, *Error Correcting Codes*, Cambridge, Mass.: M.I.T. Press, 1961, p. 147.
2. R. D. Gitlin and J. F. Hayes, "Timing Recovery and Scramblers in Data Transmission," B.S.T.J., *54*, No. 3 (March 1975), pp. 569–593.
3. D. D. Falconer, "Adaptive Equalization of Channel Nonlinearities in QAM Data Transmission Systems," B.S.T.J., *57*, No. 7 (September 1978), pp. 2589–2611.
4. W. R. Bennett and J.R. Davey, *Data Transmission*, New York: McGraw-Hill, 1965, pp. 324–325.
5. S. W. Golomb, *Shift Register Sequences*, San Francisco: Holden-Day, 1967, p. 88.
6. L. D. Baumert, *Cyclic Difference Sets*, New York: Springer-Verlag, 1971, pp. 148–158.

Copyright © 1979 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 58, No. 9, November 1979
Printed in U.S.A.

A Shared Resource TDMA Approach to Increase the Rain Margin of 12/14-GHz Satellite Systems

By A. S. ACAMPORA

(Manuscript received May 10, 1979)

A method to increase the rain margin of a satellite system by as much as 10 dB is presented. The technique does not require site diversity, larger antennas, or greater satellite radiated power, all of which are quite inefficient because they represent additional system resources which are only infrequently called upon during rain attenuation events. Rather, the technique creates a small pool of time division multiple access (TDMA) time slots, shared among all earth terminals in the network, to be used only when needed by the sites experiencing rain attenuation above the power margin. Coding techniques are employed during rain events to extract additional margin from these pooled resources with little bandwidth penalty. The hardware complexity to enable TDMA operation with coding during rain events is assessed and found to be quite modest. Although the transponder data rate may be in the neighborhood of 600 Mbits/s, the decoder operates at a much lower speed by virtue of the low TDMA duty cycle associated with a given ground station.

I. INTRODUCTION

The current trend in communication satellites appears to be increasingly toward the use of the 12/14-GHz frequency bands and the use of digital modulation formats with time division multiple access (TDMA) techniques. The former provides freedom from existing 4/6-GHz terrestrial interference and also provides higher antenna gain and narrower beams for a given size aperture, while digital transmission in conjunction with TDMA provides for more efficient utilization of the available satellite system resources.

A major drawback associated with 12/14-GHz systems is the signal attenuation associated with rainfall.[1,2] In general, attenuation at these frequencies is an increasing function of rain rate, with the result that, over a large portion of the United States, significant power margin must be provided to prevent excessive outage due to rain fades. Standard techniques that might be employed to provide rain margin include (*i*) increasing the radiated power of the satellite and earth stations, (*ii*) improving the noise figure of the receivers, (*iii*) installing larger ground station antennas, and (*ii*) providing site diversity. Unfortunately, all these techniques are costly in that permanently dedicated system resources are used only infrequently, i.e., when it rains. Looking at this another way, the system has been tremendously overdesigned for the clear air conditions that might exist more than 99.9 percent of the time at any particular ground station location if, say, a 15- or 20-dB rain margin is required to achieve the desired rain outage.

A much more efficient technique to achieve the desired outage would be to reduce the rain margin to a lower (and more reasonable) value of, say, 5 to 10 dB, and provide a common pool of resources to be shared among all ground stations and allocated as needed to increase the rain margin only at those stations which might be experiencing a fade depth greater than the built-in margin. Clearly, such a shared approach cannot be applied to ground station facilities. However, the satellite resources can readily be shared among all users (demand-assigned TDMA is only one example of pooled resources). To increase the rain margin for a particular user for a short interval, we might, for example, consider providing onboard batteries to increase the radiated power of the transponder serving that user experiencing excess rain attenuation. Such an approach, however, is not very attractive because it requires dual-mode final power amplifiers and excess battery back-up power, and also because the interference produced in adjacent transponders by this high power mode of operation might tend to become excessive.

In this paper, a different approach is taken to provide a pool of assignable resources used to increase the rain margin. For this scheme, the pooled resources consist of TDMA burst packet slots. Consider the downlink. In each TDMA frame, some small number of slots are reserved for use by any ground stations experiencing downlink rain attenuation. Then, for each slot occupied by a receiving ground station experiencing a fade, three or four slots would be assigned from the pool and would be used by encoding at the transmitting ground station using, for example, a rate $r = \frac{1}{3}$ convolutional code. Thus, three channel symbols are created for each information symbol. These additional symbols are transmitted during the pooled time slots; there is no increase in the channel data rate. Such an approach might provide 8 to 10 dB additional fade margin with no increase in either satellite power or satellite or ground station antenna diameter. The overhead associated with the pooled reserve time slots would be about 3 percent for a system containing 100 ground terminals. Such an approach can also be used for up-link fades, but here it is more desirable to employ up-link power control, as is described later.

In Section II, the concepts of this approach are developed in greater detail, and a new interpretation of outage due to rain attenuation is developed. Section III contains a description of the modest digital hardware required at each ground station to implement this concept.

II. SHARED RESOURCE-CODING CONCEPT

The concept to be presented is equally applicable to area coverage systems,[3] fixed multiple spot beam systems,[4] and singly[5] or multiply[6,7] scannable spot beam systems. For simplicity, let us consider a single scanning spot beam system (or, more precisely, a system containing a single scanning uplink beam and a single scanning downlink beam). The satellite contains a single 500-MHz transponder. The service area is divided into N spot beam footprints, labeled F_1 through F_N, as shown in Fig. 1. Each footprint typically contains several ground stations.

Figure 2 shows a typical TDMA switching sequence performed at the satellite to interconnect the various footprints. Within each frame are dedicated time slots used to establish a two-way signaling channel between a master ground station and each remote station in the network.[8] The signaling channels are used to enable TDMA synchronization, distribute system status information, handle new requests for service, assign time slots, etc. Except for the signaling slots, all other

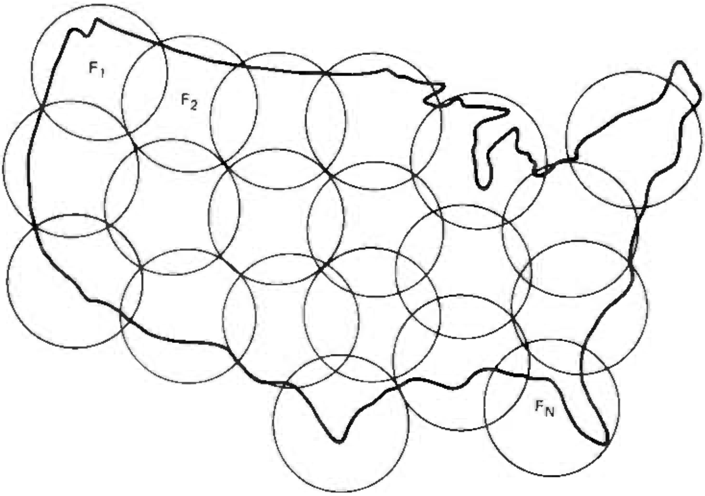


Fig. 1—A typical subdivision of the United States into N spot beam footprints labeled F_1 through F_N.

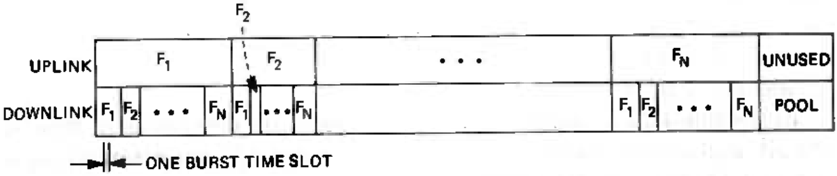


Fig. 2—Typical TDMA switching frame showing the interconnections between the N spot beam footprints and showing an unused pool of time slots.

slots can, if needed, be assigned upon demand. Also shown (at the end of the frame) is a pool of unused time slots. As will be described, these slots are to be made available to service ground stations experiencing downlink rain attenuation. For a multiple spot beam system with onboard switching,[6,7] a similar pool of unused time slots might be reserved for each transponder.

These slots can also be made available to any ground stations experiencing uplink rain attenuation. However, a more attractive means for combating uplink fades is via uplink power control. For this approach, the uplink power during rain events is adjusted such that a constant incident power is maintained at the satellite. When the rain attenuation exceeds the margin provided by the maximum ground station transmitter power, fading occurs on the uplink. Since uplink power is usually not at a premium, the maximum transmitter power can often be set to provide the desired outage. Thus, uplink power control represents a very attractive means for combating uplink loss of signal while maintaining a constant signal-to-interference ratio at the satellite (this latter concern arises with multibeam systems and also as far as coexistence with other satellite systems is concerned). A hybrid approach incorporating both shared satellite resources and uplink power control is also possible. It is the resource-limited downlink where rain attenuation presents a more serious problem.

When a downlink fade occurs, the carrier-to-noise ratio at the receiving terminal experiencing the fade is no longer sufficient to maintain the desired bit error rate. Thus, the capacity into that terminal is reduced. Suppose, for example, the rain attenuation is such that the signal level falls 8 dB below the value required to maintain a voice grade bit error rate (BER) equal to 10^{-3}. The channel error rate for Gaussian noise is then about 0.1; a lower bit error rate would result if both Gaussian noise and peak-limited interference set the error rate. The BER can still be maintained at 10^{-3} or lower if the bit interval is increased by a factor of 7 (i.e., 8 dB) via allocating seven times as much of the TDMA frame and by restructuring the receiver to accept the longer bit interval. Such an approach is unattractive, however, because not only does the longer bit interval involve a great deal of complexity at the receiver, but it also wastes valuable TDMA frame time through inefficient use of the available bandwidth.

Rather, when power measurements indicate that downlink attenuation exceeding the built-in power margin is imminent, let us use the signaling link to notify the master ground station, as well as all transmitting stations communicating with the fade site, that a fade is about to occur. Then, we borrow time slots from the reserve pool of Fig. 1 and use them as follows. Suppose that, before the fade, the fade site is using the time slot equivalent of V voice circuits. Let us borrow the equivalent of, say, 3V additional time slots from the pool, thereby providing the equivalent of 4V voice circuits into that ground station. At the originating ground station for each voice circuit, we employ a rate $r = ⅓$ convolutional code which produces three channel bits for each information bit. For both single and multiple spot beam systems, the switching sequence at the satellite is then modified via the signaling link such that each voice circuit packet is transmitted as four contiguous packets which contain the encoded channel bits (transmitted at the original full bandwidth data rate) plus an extended preamble containing 7 to 10 times the clear air number of bits required to enable carrier and clock recovery at a carrier-to-noise ratio as much as 8 to 10 dB below system margin (the bandwidth of the carrier and clock recovery circuits at the receiver are correspondingly reduced by a factor of 7 to 10). At the receiver, the entire extended burst for each voice circuit is serially detected by either a soft decision or hard decision detection device[9] and stored in a high-speed buffer. Since the duty cycle of burst arrivals is small, we read out of the buffer during the time interval between burst arrivals and process the detected channel bits by a relatively slow speed decoder to recover the original information bits.

Figure 3 shows the BER vs channel symbol-to-noise ratio (E_S/N_o) curves for (*i*) a constraint length $K = 8$, $r = ⅓$ code used in conjunction with hard-decision Viterbi decoding, and (*ii*) a $K = 4$, $r = ⅓$ code used with 3-bit soft decision decoding. Both curves assume that Gaussian noise is the only system impairment. We note that, without coding, $e_s/N_o = 7$ dB is required to provide a BER = 10^{-3}. Thus, the $K = 8$ code can maintain a BER = 10^{-3} with 7.5 dB less power; the $K = 4$ code maintains a BER = 10^{-3} with 9 dB less power. These margins generally increase if both Gaussian noise and peak-limited interference (e.g., cochannel and/or intersymbol interference) are present. Thus, by sharing a small number of TDMA slots among all users, it is possible to provide an additional 8 to 10 dB fade margin at no cost in terms of satellite power, satellite antenna gain, or earth station antenna diameter. We see from Fig. 3 that the additional fade margin provided by coding increases if the system BER threshold is reduced to a value less

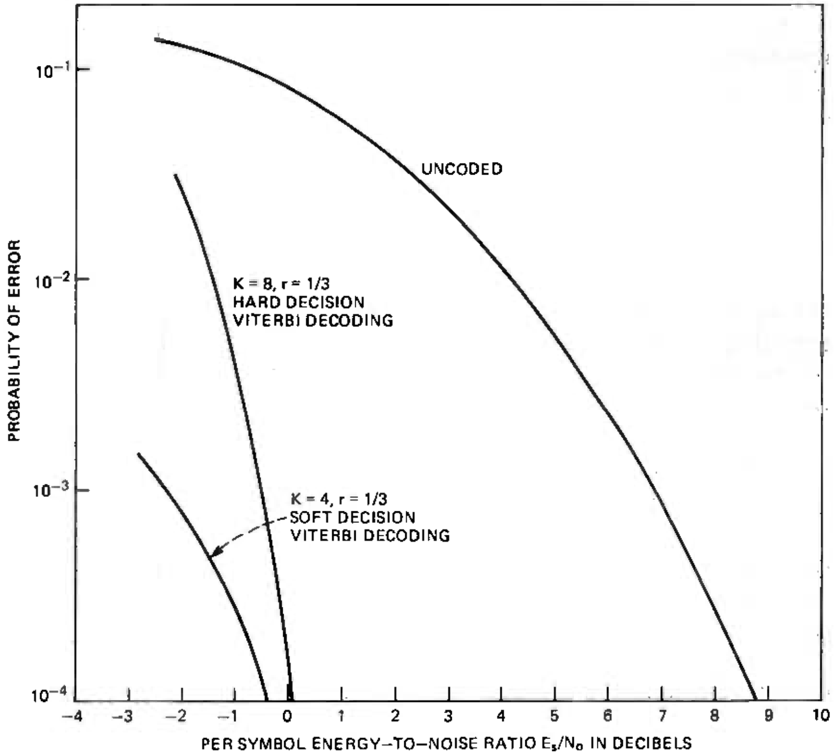


Fig. 3—BER performance for coded and uncoded transmission.

than 10^{-3}. Of course, other convolutional or block codes might be employed to achieve the same or similar results.

When the fade has passed, the extra time slots are returned to the pool to be reassigned as needed to other ground stations in the network.

The primary virtue of this approach is that a relatively small number of equivalent voice circuits can be shared among a large number of users to provide additional rain margin when needed. The additional resources are not wasted by merely retransmitted uncoded data a number of times, but rather the entire transponder bandwidth is exploited to provide additional gain through redundancy coding. Other, lower-rate codes might be used to increase the fade margin still further.

The TDMA time slots reserved for rain fades can be allocated to nonfade sites during periods of high system demand. This possibility provides for an interesting interpretation of rain outage. During clear air conditions, each ground terminal in the network presents an instantaneous demand for some number of equivalent voice circuit packets; the capacity of the satellite is, however, fixed at C two-way voice circuits. Call blockage occurs whenever the total offered load exceeds C. A ground station which uses M one-way voice circuit and experiences a fade now demands additional one-way circuits to remain operational; the number of additional circuits required increases with the fade depth, and coding is employed in an attempt to minimize the additional demand. Provided that the additional circuits are available, outage will not occur. Thus, rain attenuation can be interpreted as placing additional demands upon the voice circuit resources of the satellite, and outage is interpreted in terms of demand exceeding capacity, i.e., blocked calls. Rain outage, then, is more likely to occur during the busy hour, and would be virtually nonexistent at other times of the day.

For practical reasons, it might be desirable to limit the excess demand for voice circuits due to rain attenuation to a factor of 4 or 5 above the clear air demand. Then outage occurs when the attenuation exceeds the additional rain margin provided by these extra circuits. Thus, when designing the network, the offered traffic must be contained to a level such that the desired rain outage and call blockage probability can be achieved by the satellite capacity C. Factors affecting this design would include the rain statistics at the various sites, the built-in rain margin, the number of ground stations, the clear air Erlang load of each ground station, and the statistical dependence of rain attenuation in excess of the built-in margin at the various ground stations.

Let us examine the TDMA overhead associated with reserving time slots for rain events. Suppose that S ground stations are in the network, and a total of N one-way voice circuits are available. We reserve R of these for rain events. Thus, on the average, each station uses $(N - R)/S$ one-way circuits. The value R is determined by noting that, for each circuit into a given ground station, we need three additional circuits to provide the additional rain margin of 10 dB. We will provide a reserve pool sufficient to accommodate M simultaneous fades. We then obtain the relationship

$$3M(N - R)/S = R \Rightarrow R = 3MN/(S + 3M).$$

Thus, the TDMA inefficiency η is given by:

$$\eta = \frac{3M}{S + 3M}.$$

Thus for 100 sites and allowing for two simultaneous fades, the inefficiency or cost is under 6 percent, assuming that all ground stations carry approximately equal traffic.

If the ground stations of a satellite network exhibit large traffic imbalances, then the rain outage objective for a few high traffic ground stations might be achieved by more conventional approaches such as larger antennas or site diversity, with pooled time slots reserved for the exclusive shared usage among a large number of somewhat lower traffic ground stations. In this manner, the shared resource approach can still be applied to efficiently provide the outage objective for most of the ground terminals, without requiring a large overhead penalty to protect a small number of high traffic users.

III. IMPLEMENTATION

The equipment needed to implement the pooled resource approach to combat rain fades consists for the most part of digital electronics which operate at a rate much less than the full transponder data rate of 600 Mbits/s. The bit rate reduction is achieved by virtue of the small duty cycle TDMA mode of operation. Consider the transmit function, shown in Fig. 4a. Data arriving at a ground station are formatted for transmission via satellite and the appropriate preamble is attached. Each burst is temporarily stored in a high-speed buffer. Then, infrequently throughout the TDMA frame, the buffer is quickly read out to the modulator at a data rate of 600 Mbits/s. The timing of these events is such that the bursts from the various ground stations arrive at the satellite on a nonoverlapping basis. The receive operation, shown in Fig. 4b, is analogous. The receiver carrier and clock recovery circuits operate continuously on the incoming full bandwidth packet stream, supplying recovered carrier and clock to the demodulator and bit detector. Via the signaling links, each receiver is informed of which bursts are intended for local reception. Allowing some guard time on each side, the correct bursts are rapidly read into a cache memory; between burst arrivals, the memory is slowly offloaded for subsequent slow-speed processing.

The clear air burst packet structure is shown in Fig. 5. Each burst consists of six fields containing the following information:[10]

(1) Carrier and clock recovery preamble.
(2) Unique word signifying start of burst.
(3) Destination code.
(4) Source code.
(5) Identification of signaling or data burst.
(6) Text or signaling data.

Fields 1 to 5 collectively contain 67 bauds, and field 6 contains 400 bauds. Modulation is 4ϕ-CPSK.

The function of each field is elaborated upon in Ref. 10. For now, we describe the modification needed to assemble extended rain attenuation bursts. By using the $K = 8$, $r = 1/3$ code of the previous section, the system must be capable of operation at channel error rates as high as 0.1 to provide 7.5 dB of extra rain margin. The extended burst also is divided into six fields, each serving the same function as before.

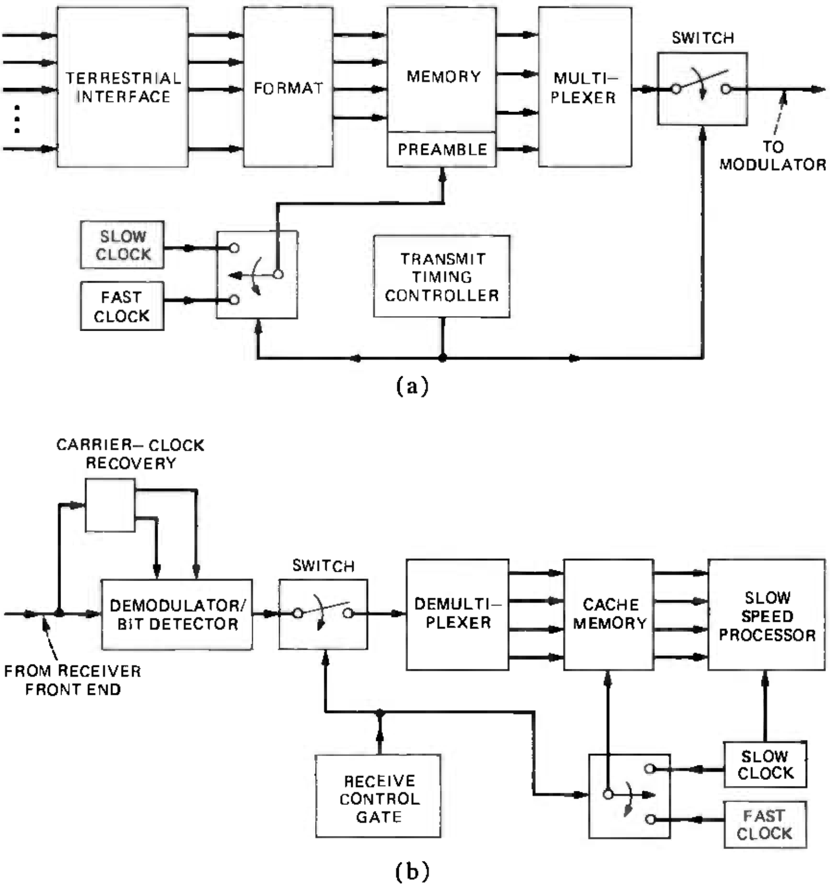


Fig. 4—Essentials of TDMA burst modem data processing. (a) Transmit function. (b) Receive function.

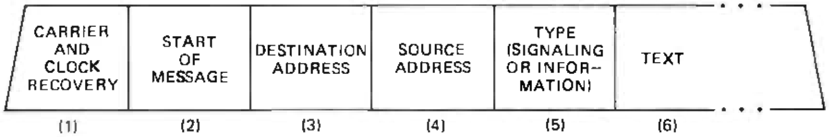


Fig. 5—TDMA burst structure.

However, field 1 must be extended by a factor of 6 to provide the same accuracy of carrier and clock recovery. This represents about 200 bauds. Also, the start-of-burst unique word must be extended to enable identification under degraded channel conditions. Discussion of this unique word extension and processing will be deferred until later. The data of fields 3 through 6 are transmitted in coded form. Encoding for the $K = 8$, $r = ⅓$ code appears in Fig. 6a, and that for the $K = 4$, $r = ⅓$ code appears in Fig. 6b. Data are read into the shift register one bit at a time; each time a shift occurs, three encoded bits are produced at the outputs of the modulo-2 adders. These encoded bits are augmented by the preamble and start-of-burst unique word, and the entire assembled burst is stored in a buffer awaiting transmission onto the channel. The length of the buffer is about 4000 bits, to be compared against about 1000 bits for clear air bursts.

At the receiver, carrier and clock recovery, demodulation, and bit-

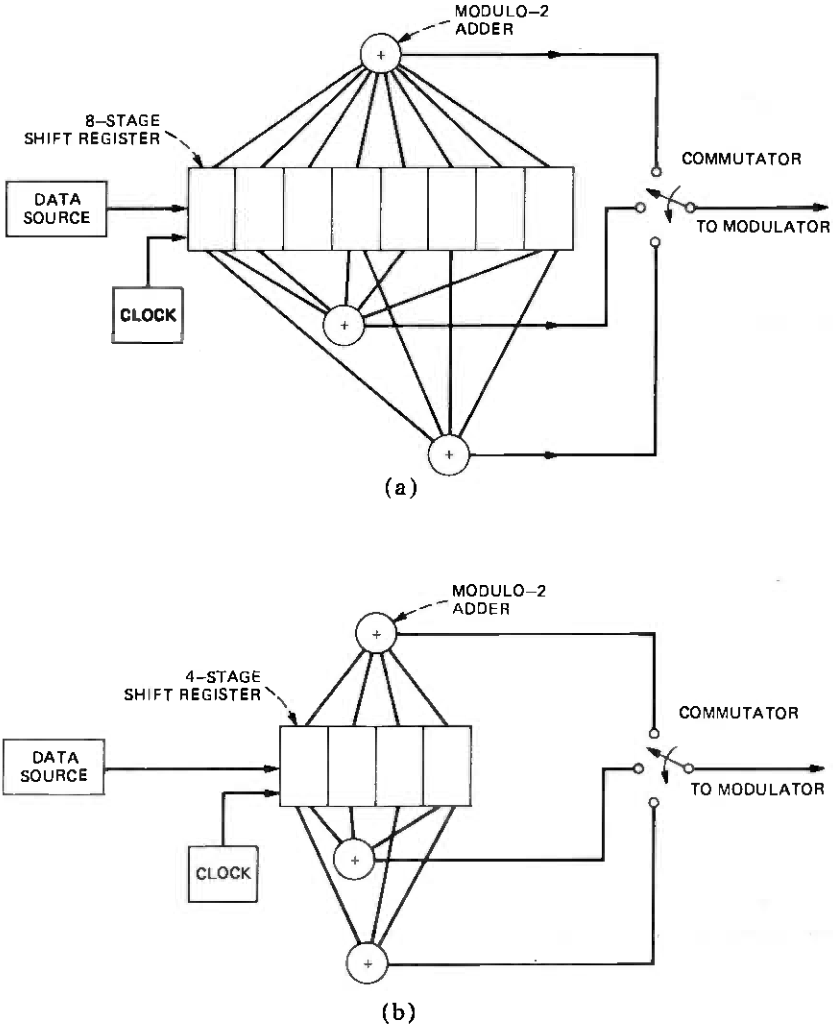


Fig. 6.—Convolutional encoders. (a) $K = 8$, $r = ⅓$ encoder. (b) $K = 4$, $r = ⅓$ encoder.

by-bit detection (either hard-quantized or soft-quantized) are performed as shown in Fig. 4b. Upon command of the timing circuitry (see Ref. 10 for details), a window is opened to seize only intended bursts. Fields 2 through 6 of the burst are demultiplexed into perhaps eight parallel rails for storage in a cache memory. The size of this memory is about 4000 bits for hard decisions and about 12,000 bits for 3-bit soft decisions. Between burst arrivals, a relatively slow-speed unique word processor, shown in Fig. 7, locates the beginning of intelligible data; only the sign bit of soft decision 3-bit words is used for this function. The processor consists of a digital correlator followed by a comparator which compares the number of coincidences between the contents of the correlator and the known unique word bit pattern against some preset threshold. The length of the unique word and the threshold (required number of coincidences) are such that reliable detection is possible on a degraded channel. For example, suppose the channel error rate is 0.1, there are 50 bits in the unique word, and we require 30 or more coincidences. Then the probability of missing the start-of-burst unique word is about 4×10^{-9}.

Having identified the start-of-burst sequence, the remainder of the cache memory is slowly read to the Viterbi decoder. The principles of Viterbi decoding are well known,[11] and the required operations will be only briefly described here. For simplicity, we consider the $K = 3$, $r = \frac{1}{2}$ code shown in Fig. 8a. The decoder is segmented into $2^{K-1} = 4$ states, corresponding to the four possible contents of the initial two states of the shift register. Upon entry of a new data bit into the encoder, permissible state transitions, and the corresponding channel

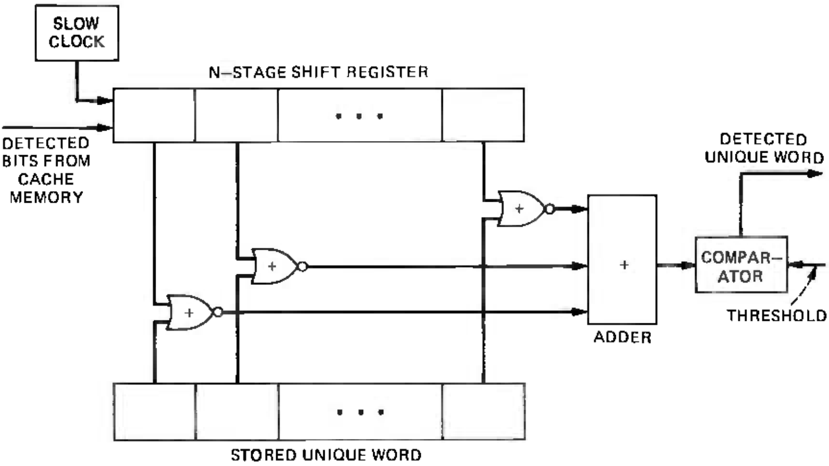


Fig. 7—Standard correlator to detect occurrence of a unique word. Detection occurs whenever the contents of the shift register agree with the unique word above a value set by the threshold.

bits generated, are as shown in Fig. 8b. Decoding is in accordance with Fig. 8c. The decoder must correlate the two received words with the channel bits generated for each possible transition, add the appropriate correlation to a metric representing the likelihood of each initial state, and choose which of two merging paths for each state is most likely. The metric of the surviving path for each state is retained and becomes the initial metric for subsequent calculations. Also stored are the surviving paths into each state, to a depth of four or five constraint lengths (about 40 bits for a $K = 8$ code). Thus, to implement hard-decision Viterbi decoding for the $K = 8$, $r = \frac{1}{3}$ code, an add-compare-store module to operate on the one-bit received words and a 40-bit memory must be provided for each state. For soft-decision decoding, the add function consists of adding three 3-bit words (appropriately weighted by ±1) to the old metric (perhaps a 5-bit word); the compar-

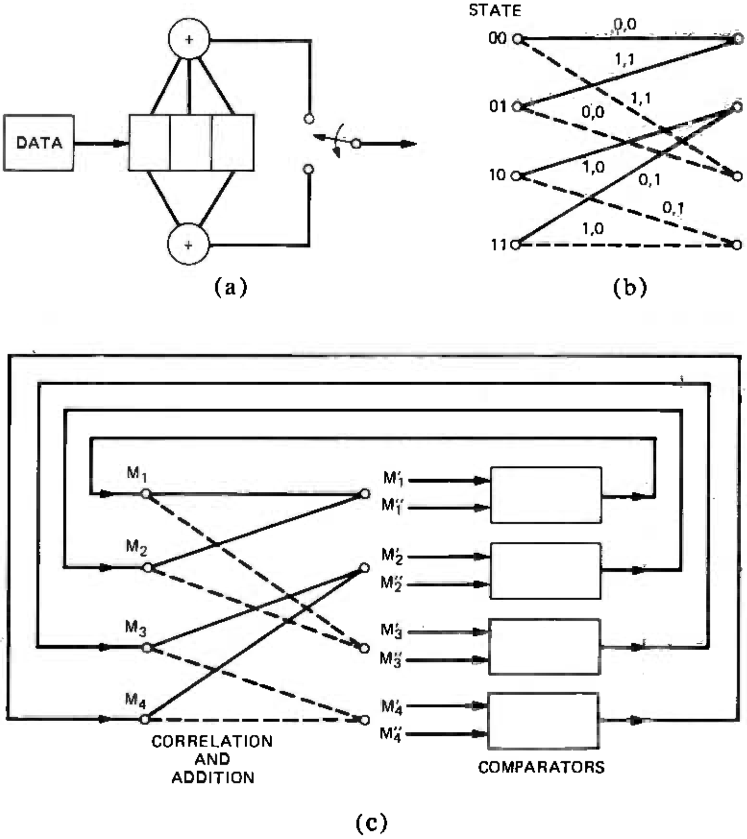


Fig. 8—Viterbi decoding for $K = 3$, $r = \frac{1}{2}$ convolutional code. (a) $K = 3$, $r = \frac{1}{2}$ encoder. (b) State transition diagram and generated channel symbols. (c) Decoder path metric update.

ison is between two 5-bit words, and the storage is a 5-bit word. For the $K = 8$ code, 128 such add-select-store modules, each operating upon 1-bit data, must be provided; for the $K = 4$ code, this number is reduced to 8 modules, but arithmetic operations must be performed upon multibit words. Thus, Viterbi decoding can be accomplished with about one to three cards of digital electronics.

For the TDMA mode envisioned here, each burst will start with the encoder initially in the all-zeros state, and $K - 1$ zero bits will be stuffed into the encoder after the final data bit; the overhead is small (21 channel bits out of about 4000 for the $K = 8$ code), and ambiguity is prevented at the decoder.

The operating speed of the Viterbi decoder is readily estimated by dividing the satellite transponder data rate by the number of ground terminals; this is an estimate of the average bit rate to a given user. Thus, for a 600-Mbit/s transponder and for 100 users, the required decoding speed is on the order of 6 Mbits/s. TTL decoders which can operate at rates up to 10 Mbits/s are readily available. High rates are also possible with TTL; alternatively, ECL decoders at speeds up to 50 Mbits/s are possible. Another option for increasing the data rate, if necessary, might be to parallel several low-speed decoders.

To maintain frame synchronization during rain attenuation conditions, a second, extended frame marker is inserted into each frame marker burst. Recalling that the initial, short frame marker was provided only to enable rapid acquisition during clear-air conditions, we can readily use the second, extended frame marker to maintain synchronization after initial acquisition. The function of the second frame marker is analogous to the extended start-of-burst word described earlier, namely, to permit identification via a slow-speed, correlation threshold as depicted in Fig. 7. Since the entire frame marker burst is stored in cache memory, we can slowly read from this memory into the correlator to find the frame marker. Then, by counting the number of elapsed bits until the frame marker is encountered, we can maintain frame synchronization.

IV. CONCLUSION

A shared-resource TDMA approach to increase the downlink rain margin of a 12/14-GHz digital satellite system by as much as 10 dB was described. The motivation for such an approach is the observation that dedicated resources (higher transmitter power, larger antennas, etc.) are a costly means for increasing rain margin since, during the overwhelming clear-air fraction of time, the system would be tremendously overdesigned. By contrast, we can view the TDMA resources of the satellite as representing a common pool of resources which can be dynamically assigned as needed to increase the rain margin at only those sites where they are instantaneously needed. By assigning additional TDMA packets to sites experiencing excessive rainfall and using coding to efficiently utilize the additional bandwidth (time slots) so made available, we can increase the fade margin by as much as 10 dB with no increase in downlink radiated power or satellite or terrestrial antenna size.

Additional margin on the uplink is achieved by uplink power control. This is attractive since it allows constant incident power at the satellite from all users, thereby avoiding severe interference from nonfade users. Also, terrestrial power is not nearly as precious as space platform power.

The overhead associated with reserving TDMA bursts for fade conditions was shown to vary with the number of ground stations served. If we wish to reserve sufficient resources to provide 10 dB extra fade margin simultaneously at two sites, then the overhead is about 1.2 percent for a network containing 500 sites and about 6 percent for a network containing 100 sites. It is not, however, necessary to reserve these slots; rather, during peak demand periods, all resources can be assigned if needed, but the additional fade margin is lost. During nonpeak hours, excess capacity is available and fade margin can be provided.

The additional hardware needed to implement this approach is mostly digital electronics and consists of a convolutional encoder at each ground station and a decoder at those ground stations expected to require additional rain margin. The decoder operates at data rates much lower than the transponder bit rate, since desired bursts arrive at a low duty cycle and are buffered; between bursts, the decoder can slowly read from the buffer and process the data. The decoding speed might be as high as 20 Mbits/s for a 20-ground-station network, or as low as 1 Mbits/s for a 600-ground-station network. Such decoders are readily available using TTL logic; if higher rates are required, two or more decoders can be operated in parallel, or a faster decoder using ECL can be used.

In addition, a quite modest amount of digital hardware is needed to maintain frame synchronization, identify start of burst indicators, and provide recovered clock and carrier. The cost of the additional electronics is by far smaller than the cost associated with either antennas exhibiting 10 dB more gain or satellites with 10-dB more radiated power.

REFERENCES

1. D. C. Hogg and T. S. Chu, "The Role of Rain in Satellite Communications," Proc. IEEE, *63* (September 1975).
2. A. J. Rustako, Jr., "An Earth-Space Propagation Measurement at Crawford Hill Using the 12-GHz CTS Satellite Beacon," B.S.T.J., *57*, No. 5 (May–June, 1978), pp. 1431–1448.

3. J. D. Barnla and F. R. Zitzmann, "Digital Communications Satellite System of SBS," EASCON '77 Conf. Rec., Paper 7.2, Washington, D.C. (September, 1977).
4. D. O. Reudink, 6th AIAA Communications Satellite System Conference, 1976, Montreal, Canada.
5. D. O. Reudink and Y. S. Yeh, "A Scanning Spot Beam Satellite System," B.S.T.J., *56,* No. 8 (October, 1977), pp. 1549–1560.
6. A. S. Acampora and B. R. Davis, "Efficient Utilization of Satellite Transponders via Time-Division Multibeam Scanning," B.S.T.J., *57,* No. 8 (October, 1978), pp. 2901–2914.
7. A. S. Acampora, C. Dragone, and D. O. Reudink, "A Satellite System with Limited-Scan Spot Beams," IEEE Trans. Comm., Oct. 1979.
8. D. O. Reudink and Y. S. Yeh, "The Organization and Synchronization of a Switched Spot-Beam System," Fourth Int'l Conference on Digital Satellite Communications, Montreal, Canada, Oct. 1978.
9. K. S. Gilhausen, et al., "Coding Systems Study for High Data Rate Telemetry Links," Linkabit Corp., San Diego, available through NTIS-N71-27786, January, 1971.
10. A. S. Acampora and R. E. Langseth, "Baseband Processing in a High-Speed Burst Modem for a Satellite-Switched TDMA System," Fourth International Conference on Digital Satellite Communications, Montreal, Canada, October, 1978.
11. A. J. Viterbi, "Convolutional Codes and Their Performance in Communication Systems," IEEE Trans. Comm. Tech., *COM-19,* No. 5 (October, 1971).

Copyright © 1979 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 58, No. 9, November 1979
Printed in U.S.A.

An Ordering Scheme for Facsimile Coding

By F. W. MOUNTS, E. G. BOWEN, and A. N. NETRAVALI

(Manuscript received June 19, 1979)

We give simulation results using one of our ordering algorithms for the coding of eight CCITT test documents. These algorithms do not make any approximations and therefore can reproduce the documents exactly at the receiver. The code design is similar to the one-dimensional modified Huffman code that has been proposed by the CCITT as a standard. One-dimensional run-length coding using the modified Huffman code results in 445,316 bits per document on the average, which can be transmitted in 92.77 seconds using a transmission rate of 4800 bits per second. Our ordering algorithm, which is two-dimensional in nature, requires, on the average, 264,632 bits per document, or 55.13 seconds per document for the same transmission rate. Thus, the ordering scheme reduces the transmission time by approximately 41 percent, compared to one-dimensional run-length coding.

I. INTRODUCTION

This paper presents simulation results using one of our ordering schemes[1,2] for efficient coding of two-level (black-and-white) facsimile pictures using the eight CCITT (International Telegraph and Telephone Consultative Committee) test documents* as the source data. The scheme, we consider, allows exact reproduction of the documents at the receiver. Specifically, we give results on compression factors using a seven-element predictor, and a modified Huffman code for coding of the ordered data. It is assumed that every kth ($k = 2, 4, \infty$) line is encoded by a one-dimensional run-length code to limit the vertical propagation of the transmission errors. In our previous papers,[1-3] we gave entropies of run lengths of the ordered data using predictors and

* The eight CCITT study group XIV test documents are those used for the Graphics Coding Contest of the 1976 Picture Coding Symposium. Each image contains 2128 lines with 1728 pels per line. The scanning density in both directions is approximately 8 dots/mm.

ordering parameters specifically optimized for each document. In this paper, we give results when the predictor, the ordering parameters, and the Huffman code are obtained by averaging the relevant statistics for all eight CCITT documents.

In the basic ordering scheme, we make a prediction of the present element using the surrounding previously transmitted picture elements, and classify it as "good" or "bad," depending upon the probability of the prediction being in error, conditioned on the specific values of the surrounding picture elements. We then change the relative order of the prediction errors corresponding to picture elements along a scan line, using the "goodness" of the prediction in such a way as to increase the average run length of the black ("1") and/or white ("0") elements and then transmit the run lengths.

Several variations of the ordering algorithm have existed for some time.[4-6] We describe one specific variation of the algorithm, which we believe is simple to implement and can be easily extended to the coding of two-level pictures dithered to give the appearance of gray-level.[2] The results of computer simulations are given in Section III. They indicate that, on the average, a CCITT document can be transmitted with 264,632 bits, which would require 55.13 seconds using a transmission channel at 4800 bits per second and $k = \infty$. A one-dimensional modified Huffman code, on the other hand, would require 445,316 bits, on the average, and 92.77 seconds per document for the same transmission rate.* Thus, use of the two-dimensional ordering algorithm decreases the transmission time by 41 percent. Use of $k = 4$, however, increases the average bits per document to 307,310 and the average transmission time to 64.02 seconds.

II. CODING ALGORITHM

In this section, we describe our coding algorithm in detail. The coding algorithm consists of four steps: (*i*) prediction, (*ii*) ordering, (*iii*) dropping a specific sequence from transmission, and (*iv*) run-length coding of ordered data. We give details of each of these steps below.

2.1 Prediction algorithm

The first step in the ordering algorithm consists of making a prediction of the present picture element using the already-transmitted surrounding picture elements. We define a state S_i using the seven surrounding picture elements, A,B,C,D,E,F, and G, as shown in Fig. 1. There are 128 states. The predictor is developed in a standard way[1] as

* Without any compression techniques, each document requires 3,702,720 bits and can be transmitted in 12.86 minutes using a transmission rate of 4800 bits per second.

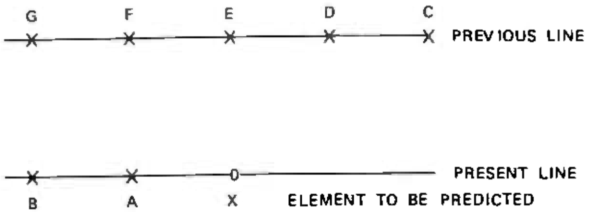


Fig. 1—The configuration of the seven elements constituting the state which is used to construct the predictor used for forward ordering.

the one which minimizes the probability* of making an error, given that a particular state has occurred. Thus, the prediction $C(S_i)$, for state S_i, is given by:

$$C(S_i) = \text{“black,” if } P(X = \text{“black”} \,|\, S = S_i) \geq 0.5$$
$$= \text{“white,” otherwise,}$$

where $P(.\,|\,.)$ is the experimentally determined conditional probability. An error in the prediction is denoted by "1" (black) and no error is denoted by "0" (white). A boundary of "0" is assumed wherever necessary to develop this prediction. This predictor varies from picture to picture. We have chosen the predictor to be the average over all the eight CCITT documents. Table I gives the relevant statistics and the predictor for each of the 128 states.

2.2 Coding

The ordering algorithm can be illustrated by considering a memory of 1728 cells (equal to the number of elements per line). Suppose the cells of this memory are numbered from 1 to 1728; we classify the states used for prediction into two categories, "good" or "bad." "Good" states are those for which the probability of the prediction being in error, conditional on that state, is less than or equal to a given threshold. All the other states are "bad." The classification of states into "good" and "bad" was based on the average over the eight CCITT documents using a goodness threshold of 0.90. This classification is given in Table I.

In the process of ordering, if the first element of the present line has a state classified as "good," then we put the value of the prediction error corresponding to it in memory cell 1; if, on the other hand, the state is classified as "bad," then we put the prediction error corresponding to it in memory cell 1728. We continue in this manner; the prediction error for the *i*th element of the present line is put in the unfilled memory cell of the smallest or the largest index, depending on

* This is computed by taking sample means over the eight CCITT documents.

Table I—Prediction, probability of correct prediction, and the "goodness" of the prediction for each state

Both forward and reverse ordering are considered with the appropriate definition of the state.

	Forward Ordering				Reverse Ordering			
No.	State Configuration	Prediction	Probability of Correct Prediction	Goodness (G or B)	State Configuration	Prediction	Probability of Correct Prediction	Goodness (G or B)
0	00000 00X	0	0.999	G	00000 X00	0	0.999	G
1	00000 01X	1	0.834	B	00000 X10	1	0.833	B
2	00000 10X	0	0.982	G	00000 X01	0	0.983	G
3	00000 11X	1	0.767	B	00000 X11	1	0.770	B
4	00001 00X	0	0.965	G	10000 X00	0	0.963	G
5	00001 01X	1	0.931	G	10000 X10	1	0.917	G
6	00001 10X	0	0.918	G	10000 X01	0	0.928	G
7	00001 11X	1	0.924	G	10000 X11	1	0.900	G
8	00010 00X	0	0.776	B	01000 X00	0	0.745	B
9	00010 01X	1	0.951	G	01000 X10	1	0.939	G
10	00010 10X	0	0.673	B	01000 X01	0	0.724	B
11	00010 11X	1	0.960	G	01000 X11	1	0.959	G
12	00011 00X	0	0.833	B	11000 X00	0	0.846	B
13	00011 01X	1	0.985	G	11000 X10	1	0.983	G
14	00011 10X	0	0.787	B	11000 X01	0	0.783	B
15	00011 11X	1	0.973	G	11000 X11	1	0.955	G
16	00100 00X	0	0.602	B	00100 X00	0	0.607	B
17	00100 01X	1	0.948	G	00100 X10	1	0.979	G
18	00100 10X	0	0.617	B	00100 X01	0	0.540	B
19	00100 11X	1	0.936	G	00100 X11	1	0.945	G
20	00101 00X	0	0.629	B	10100 X00	0	0.640	B
21	00101 01X	1	0.793	B	10100 X10	1	0.878	B
22	00101 10X	1	0.571	B	10100 X01	1	0.545	B
23	00101 11X	1	0.972	G	10100 X11	1	0.886	B
24	00110 00X	1	0.819	B	01100 X00	1	0.784	B
25	00110 01X	1	0.992	G	01100 X10	1	0.994	G
26	00110 10X	1	0.623	B	01100 X01	1	0.656	B

Table I—(*Continued*)

	Forward Ordering				Reverse Ordering			
No.	State Configuration	Prediction	Probability of Correct Prediction	Goodness (G or B)	State Configuration	Prediction	Probability of Correct Prediction	Goodness (G or B)
27	00110 11X	1	0.967	G	01100 X11	1	0.976	G
28	00111 00X	1	0.649	B	11100 X00	1	0.639	B
29	00111 01X	1	0.996	G	11100 X10	1	0.998	G
30	00111 10X	1	0.524	B	11100 X01	1	0.529	B
31	00111 11X	1	0.985	G	11100 X11	1	0.987	G
32	01000 00X	0	0.971	G	00010 X00	0	0.949	G
33	01000 01X	1	0.522	B	00010 X10	0	0.507	B
34	01000 10X	0	0.906	G	00010 X01	0	0.860	B
35	01000 11X	1	0.636	B	00010 X11	1	0.646	B
36	01001 00X	0	0.934	G	10010 X00	0	0.899	B
37	01001 01X	1	0.503	B	10010 X10	1	0.753	B
38	01001 10X	0	0.556	B	10010 X01	0	0.800	B
39	01001 11X	1	0.751	B	10010 X11	1	0.692	B
40	01010 00X	0	0.854	B	01010 X00	0	0.885	B
41	01010 01X	1	0.583	B	01010 X10	1	0.635	B
42	01010 10X	0	0.684	B	01010 X01	1	0.529	B
43	01010 11X	1	0.813	B	01010 X11	1	0.878	B
44	01011 00X	0	0.910	G	11010 X00	0	0.864	B
45	01011 01X	0	0.516	B	11010 X10	1	0.730	B
46	01011 10X	0	0.645	B	11010 X01	1	0.526	B
47	01011 11X	1	0.837	B	11010 X11	1	0.796	B
48	01100 00X	0	0.661	B	00110 X00	0	0.533	B
49	01100 01X	1	0.973	G	00110 X10	1	0.983	G
50	01100 10X	0	0.895	B	00110 X01	0	0.692	B
51	01100 11X	1	0.759	B	00110 X11	1	0.838	B
52	01101 00X	0	0.783	B	10110 X00	0	0.697	B
53	01101 01X	1	0.868	B	10110 X10	1	0.941	G
54	01101 10X	0	0.765	B	10110 X01	0	0.596	B
55	01101 11X	1	0.729	B	10110 X11	1	0.827	B

Table I—(*Continued*)

	Forward Ordering				Reverse Ordering			
No.	State Configuration	Prediction	Probability of Correct Prediction	Goodness (G or B)	State Configuration	Prediction	Probability of Correct Prediction	Goodness (G or B)
56	01110 00X	1	0.726	B	01110 X00	1	0.739	B
57	01110 01X	1	0.997	G	01110 X10	1	0.998	G
58	01110 10X	0	0.594	B	01110 X01	1	0.565	B
59	01110 11X	1	0.950	G	01110 X11	1	0.976	G
60	01111 00X	1	0.530	B	11110 X00	1	0.502	B
61	01111 01X	1	0.996	G	11110 X10	1	0.996	G
62	01111 10X	0	0.768	B	11110 X01	0	0.615	B
63	01111 11X	1	0.961	G	11110 X11	1	0.975	G
64	10000 00X	0	0.996	G	00001 X00	0	0.996	G
65	10000 01X	1	0.768	B	00001 X10	1	0.637	B
66	10000 10X	0	0.997	G	00001 X01	0	0.999	G
67	10000 11X	0	0.648	B	00001 X11	0	0.665	B
68	10001 00X	0	0.982	G	10001 X00	0	0.986	G
69	10001 01X	1	0.757	B	10001 X10	1	0.689	B
70	10001 10X	0	0.988	G	10001 X01	0	0.991	G
71	10001 11X	1	0.614	B	10001 X11	1	0.607	B
72	10010 00X	0	0.902	G	01001 X00	0	0.890	B
73	10010 01X	1	0.625	B	01001 X10	1	0.682	B
74	10010 10X	0	0.717	B	01001 X01	0	0.843	B
75	10010 11X	1	0.880	B	01001 X11	1	0.828	B
76	10011 00X	0	0.924	G	11001 X00	0	0.937	G
77	10011 01X	1	0.748	B	11001 X10	1	0.767	B
78	10011 10X	0	0.906	G	11001 X01	0	0.927	G
79	10011 11X	1	0.826	B	11001 X11	1	0.789	B
80	10100 00X	0	0.879	B	00101 X00	0	0.878	B
81	10100 01X	1	0.660	B	00101 X10	1	0.719	B
82	10100 10X	0	0.531	B	00101 X01	1	0.595	B
83	10100 11X	1	0.882	B	00101 X11	1	0.930	G
84	10101 00X	0	0.897	B	10101 X00	0	0.768	B

Table I—(*Continued*)

	Forward Ordering				Reverse Ordering			
No.	State Configuration	Prediction	Probability of Correct Prediction	Goodness (G or B)	State Configuration	Prediction	Probability of Correct Prediction	Goodness (G or B)
85	10101 01X	1	0.739	B	10101 X10	1	0.700	B
86	10101 10X	1	0.623	B	10101 X01	1	0.611	B
87	10101 11X	1	0.846	B	10101 X11	1	0.860	B
88	10110 00X	0	0.551	B	01101 X00	0	0.510	B
89	10110 01X	1	0.912	G	01101 X10	1	0.889	B
90	10110 10X	1	0.637	B	01101 X01	1	0.505	B
91	10110 11X	1	0.921	G	01101 X11	1	0.881	B
92	10111 00X	0	0.712	B	11101 X00	0	0.767	B
93	10111 01X	1	0.959	G	11101 X10	1	0.900	G
94	10111 10X	1	0.561	B	11101 X01	1	0.528	B
95	10111 11X	1	0.959	G	11101 X11	1	0.933	G
96	11000 00X	0	0.991	G	00011 X00	0	0.983	G
97	11000 01X	1	0.795	B	00011 X10	1	0.812	B
98	11000 10X	0	0.997	G	00011 X01	0	0.998	G
99	11000 11X	0	0.711	B	00011 X11	0	0.709	B
100	11001 00X	0	0.978	G	10011 X00	0	0.965	G
101	11001 01X	1	0.795	B	10011 X10	1	0.826	B
102	11001 10X	0	0.982	G	10011 X01	0	0.989	G
103	11001 11X	0	0.576	B	10011 X11	0	0.600	B
104	11010 00X	0	0.894	B	01011 X00	0	0.892	B
105	11010 01X	1	0.686	B	01011 X10	1	0.625	B
106	11010 10X	0	0.678	B	01011 X01	0	0.808	B
107	11010 11X	1	0.703	B	01011 X11	1	0.664	B
108	11011 00X	0	0.942	G	11011 X00	0	0.958	G
109	11011 01X	1	0.810	B	11011 X10	1	0.735	B
110	11011 10X	0	0.914	G	11011 X01	0	0.951	G
111	11011 11X	1	0.617	B	11011 X11	1	0.602	B
112	11100 00X	0	0.934	G	00111 X00	0	0.885	B
113	11100 01X	1	0.922	G	00111 X10	1	0.941	G

Table I—(*Continued*)

	Forward Ordering				Reverse Ordering			
No.	State Configuration	Prediction	Probability of Correct Prediction	Goodness (G or B)	State Configuration	Prediction	Probability of Correct Prediction	Goodness (G or B)
114	11100 10X	0	0.975	G	00111 X01	0	0.969	G
115	11100 11X	1	0.806	B	00111 X11	1	0.805	B
116	11101 00X	0	0.951	G	10111 X00	0	0.900	G
117	11101 01X	1	0.885	B	10111 X10	1	0.910	G
118	11101 10X	0	0.920	G	10111 X01	0	0.908	G
119	11101 11X	1	0.748	B	10111 X11	1	0.753	B
120	11110 00X	0	0.756	B	01111 X00	0	0.724	B
121	11110 01X	1	0.975	G	01111 X10	1	0.972	G
122	11110 10X	0	0.849	B	01111 X01	0	0.813	B
123	11110 11X	1	0.929	G	01111 X11	1	0.926	G
124	11111 00X	0	0.830	B	11111 X00	0	0.835	B
125	11111 01X	1	0.957	G	11111 X10	1	0.953	G
126	11111 10X	0	0.848	B	11111 X01	0	0.847	B
127	11111 11X	1	0.989	G	11111 X11	1	0.989	G

whether the state corresponding to the *i*th element is "good" or "bad." When the memory is filled, its cells are read in numerical order and the contents are run-length encoded. It is easy to see that the present line can be easily reconstructed from the knowledge of the run lengths of the ordered data, since the ordering sequence is known explicitly to the receiver.

2.3 Dropping of the sequence

The first sequence of the prediction errors of the type (0, 0, 0, ⋯, 0, 1) for each line is dropped. The following run is assumed to start with a "0." If it is indeed a run of "0," then it is assigned a proper run-length code. On the other hand, if the following run starts with a "1," a run of zero length of "0"s is transmitted first and then the code corresponding to the run of "1"s is transmitted. Two special cases ought to be mentioned. In one, if the prediction errors for a line generate a sequence 0, 0, 0, ⋯, 0, 1 (i.e., 1727 "0"s followed by a "1"), then this sequence is dropped and a special code word* (101011) is

* Identical to the second make-up code for the "0" codebook.

sent. In the second case, if the prediction errors for a line generate all "0"s, then the sequence is dropped and no special code word is transmitted except in a situation where the ambiguity with the end of message exists (i.e., occurrence of six consecutive end-of-line codes). In such a case, a special code word* (00110) is transmitted after the fifth consecutive end-of-line code. The length of the dropped sequence can be derived from the length of the rest of the decoded data in a line since the number of samples in a line is fixed and known.

2.4 Coding of run lengths

Each line is ordered from either left to right (forward) or from right to left (reverse), depending upon which direction requires the least number of coded bits. In the forward as well as the reverse ordering mode, the prediction and the "goodness" of a state are based on the average over all eight CCITT documents when ordered from left to right or right to left, respectively. The prediction and the "goodness" of a state for the reverse ordering mode are shown in Table I. A flag bit is used to define the direction of ordering to the receiver and is transmitted as the first bit of each line of coded data. The code for the run lengths for both modes of operation is based on the average over all the eight CCITT documents using the forward ordering mode. Only two sets of codes are used for coding the run lengths, one for run lengths of "0" and another for run lengths of "1." The structure of the codebook for the runs of "0" is similar to the standard one-dimensional code agreed upon by the CCITT. It uses a make-up code [MU] (when necessary) followed by a single terminating code [TC]. There are 64 terminating code words and 27 make-up code words. This codebook is given in Table II. The codebook for the runs of "1" consists of 10 terminating code words and a single make-up code word, which may be repeated a number of times depending on the length of the run being coded. This codebook is described in Table III. Both of these codebooks are constructed so that no combination of the code words of runs of "0" and "1" can result in a sequence of coded bits identical to the end-of-line code, which is taken to be the same 12-bit word specified by the CCITT (000000000001).

III. SIMULATION RESULTS

In this section, we give results of computer simulations using the algorithm described in the previous section, along with its variations. Also, for comparison, we give results for the standard one-dimensional Huffman code proposed by the CCITT. This is given in Table IV. Table IV shows that, on the average, a CCITT document can be transmitted

* Identical to the first make-up code for the "0" codebook.

Table II—Make-up and terminating codes for runs of "0"

A run of "0" is coded by using a make-up code (whenever necessary) and a terminating code, depending on the run length.

Terminating Codes for "0"

"0" Run Length	Code Word	"0" Run Length	Code Word
0	01110111	32	00101010
1	11	33	10100110
2	010	34	10101000
3	100	35	011110101
4	0001	36	011110010
5	1011	37	011101101
6	01101	38	011001110
7	00111	39	011001010
8	011111	40	011001001
9	011100	41	011001011
10	000001	42	011001000
11	101000	43	000000101
12	0111010	44	000011010
13	0110000	45	000000001
14	0000100	46	001001000
15	0010111	47	001011001
16	1010010	48	001000011
17	01111000	49	101001111
18	01100110	50	001000010
19	01100010	51	001011000
20	00001100	52	0111101110
21	00001011	53	0111101111
22	00001010	54	101010011
23	00000001	55	101001110
24	00000011	56	0111101101
25	00100111	57	0111100111
26	00100110	58	101010110
27	00100101	59	101010010
28	00100011	60	0111101001
29	00100000	61	0111101000
30	00100010	62	101010111
31	00101011	63	001011011

Make-Up Codes for "0"

"0" Run Length	Code Word	"0" Run Length	Code Word
64	00110	960	0010110101
128	101011	1024	01100111110
192	0000111	1088	01100111101
256	0010100	1152	01100111111
320	01100011	1216	01100111100
384	10101010	1280	000000000111
448	011101100	1344	000000000100
512	000000100	1408	000000000110
576	001001001	1472	0000000001011
640	0111101100	1536	00000000010100
704	0111100110	1600	000000000101011
768	0000110110	1664	0000000001010101
832	0000110111	1728	0000000001010100
896	0010110100		

Table III—Make-up and terminating codes for runs of "1"

Only one make-up code and 10 terminating code words are used. A make-up code may be repeated several times, depending upon the run length being coded.

Runs of "1" are coded in the form of

$$[\text{MU}], \cdots [\text{MU}]^i, [\text{TC}], \qquad 0 \le i \le 172,$$

where i represents the number of times the make-up code [MU] is transmitted. For a given run length (RL), the value of i is chosen such that

$$10\,i < \text{RL} \le 10\,(i + 1).$$

The terminating code is the word assigned to (RL-10i) as follows:

(RL-10i)	Terminating Code Word
1	1
2	01
3	001
4	0001
5	00001
6	0000010
7	00000110
8	0000011110
9	00000111110
10	00000111111

The make-up code is assigned the binary word (000001110).

with 445,316 bits, which would take 92.77 seconds at a transmission rate of 4800 bits per second.*

The results for the two variations of the ordering scheme are given in Table V. In both, so as to limit the propagation of transmission errors in the reconstructed image, we employ the technique of coding every kth line of the original data using the standard one-dimensional modified Huffman code. This kth line may be coded by omitting the first sequence of 0, 0, 0, $\cdots$, 0, 1 or it can be coded in its entirety. Table V gives the results when the first sequence is omitted, whereas Table VI gives the results when the first sequence is transmitted.

Additional results are given in Tables V and VII for the case when the transmission time of each line is constrained to be at least 0 ms, 5 ms, or 10 ms; i.e., 0, 24, or 48 coded bits per line using a transmission rate of 4800 bits per second. When the transmission time per line is lower bounded, for example, by 5 ms and if the number of coded bits (including the end-of-line bits) for any particular line is less than 24, then fill bits are used to make up the difference. The fill bits are taken to be a string of all "0"s and are inserted prior to the end-of-line code. Table VII gives results when the best ordering direction is switched between left to right or right to left depending on which requires the smaller number of coded bits, and this switching is specified to the

* These figures include the bits required for the end-of-line code (=2128 × 12).

Table IV—Compression results using one-dimensional run-length coding with the modified Huffman code proposed by CCITT

Total coded bits and the time required to transmit include the end-of-line code and start- and end-of-message codes. No constraint on the transmission time for each line is imposed.

CCITT Image No.	Total Coded Bits	Time Required (in Seconds) to Transmit Using a 4800-bps Rate
1	220745	45.99
2	238521	49.69
3	417306	86.94
4	682195	142.12
5	430259	89.64
6	353650	73.68
7	757804	157.88
8	462051	96.26

Average total bits per document = 445,316.
Average transmission time per document = 92.77 seconds.

receiver by a flag bit transmitted as a first bit of coded data. A flag bit is also used for every kth line.

The simulation results show that both the forward and reverse ordering give approximately the same compression, with a slight preference for the forward ordering when $k = 4$ and $k = \infty$ and the reverse ordering when $k = 2$. Adaptive ordering, on the other hand, does better than either the forward or the reverse ordering for all values of k and all cases of fill that we studied. The approximate improvement is close to a second of transmission time per document by using adaptive ordering. Using adaptive ordering and 0-ms constraint on the transmission time, $k = 4$ results in a 16-percent increase in the transmission time, whereas $k = 2$ results in a 32-percent increase in the transmission time compared to $k = \infty$. However, with adaptive ordering and $k = 2$, there is still a 21-percent decrease in transmission time compared to one-dimensional run-length coding using the modified Huffman code. When $k = 2$ or $k = 4$, the scan lines that are coded by a simple run-length code may be transmitted with or without the first sequence of the form $0, 0, 0, \cdots, 0, 1$. In the case of adaptive ordering, dropping of the first sequence for the kth line decreases the transmission time by approximately one second compared to not dropping the first sequence. Obviously, this decrease is greater for $k = 2$ than for $k = 4$. The constraint on the transmission time for each line increases the overall transmission time of the document. Using the constraints of 10 ms, the overall transmission time is increased over the case of 5 ms by 4.1 seconds, for $k = \infty$.

Table V—Compression results with forward and reverse ordering

The numbers corresponding to the transmission time constraint of 0 ms do not include the end-of-line codes, but they do include bits necessary for start and end of message; 25,536 bits are necessary for transmitting the end-of-line code for each document. The first sequence of the form 0,0,0,· · ·,0,1 is dropped in all cases.

	Forwarding Ordering			Reverse Ordering		
	0 ms	5 ms	10 ms	0 ms	5 ms	10 ms
Image 1:						
$k = 2$	160196	199922	232725	159072	198811	231720
$k = 4$	146431	186202	219729	144999	184795	218433
$k = \infty$	132156	171939	206220	130257	170088	204542
Image 2:						
$k = 2$	140564	171076	187869	140778	171328	188192
$k = 4$	111380	142486	162718	111284	142446	162959
$k = \infty$	82405	114134	137590	81777	113597	137564
Image 3:						
$k = 2$	289670	320459	336055	289130	320005	335721
k = 4	242191	273657	292645	242606	274195	293293
k = ∞	195198	227315	249424	195792	228084	250465
Image 4:						
$k = 2$	578680	609552	623370	573570	604490	618202
$k = 4$	540474	571256	585478	535034	565900	580016
$k = \infty$	503548	534304	549021	496238	527082	541684
Image 5:						
$k = 2$	309949	340302	354977	308335	338726	353319
$k = 4$	267979	299010	316769	265914	297040	314716
$k = \infty$	224792	256533	277408	222687	254598	275230
Image 6:						
$k = 2$	217758	248646	263847	217713	248664	263591
$k = 4$	167956	198841	215302	167673	198689	214851
$k = \infty$	118565	149499	167399	118460	149596	167137
Image 7:						
$k = 2$	632662	661074	669233	634047	662572	670568
$k = 4$	587417	615803	623959	590930	619562	627638
$k = \infty$	542808	571182	579348	548127	576883	585092
Image 8:						
$k = 2$	294094	321536	329539	298915	324360	332304
$k = 4$	225967	253686	263471	231735	259451	269049
$k = \infty$	157859	185779	197170	166726	194666	205892
	Average Total Bits per Document					
$k = 2$	327947	359071	374702	327695	358620	374202
$k = 4$	286224	317618	335009	286272	317760	335119
$k = \infty$	244666	276336	295448	245008	276824	295951
	Average Transmission Time per Document (in Seconds)					
$k = 2$	68.32	74.81	78.06	68.27	74.71	77.96
$k = 4$	59.63	66.17	69.79	59.64	66.20	69.82
$k = \infty$	50.97	57.57	61.55	51.04	57.67	61.66

The least bits are required for the ordering scheme with $k = \infty$, 0-ms transmission time constraint, and adaptive ordering. This results in 55.13 seconds of transmission time, on the average, which is about 41 percent less than that required for one-dimensional run length with the modified Huffman code. However, a reasonable alternative, considering the effect of transmission errors, would be for $k = 4$; and, in this case, the average transmission time is 64.02 seconds, which is 31 percent less than the one-dimensional run-length code. It should be pointed out that the effect of transmission errors has not been evaluated.

Table VI—Compression results for forward, reverse, and adaptive ordering

Each case does not include bits necessary for the end-of-line code (2128 × 12) but does include bits necessary for start and end of message. There is no constraint on the minimum transmission time for each line. Adaptive ordering results include the bits necessary for specifying the direction of the ordering. The first sequence of the form 0,0,0, · · · ,0,1 is not dropped for the kth line ($k = 2, 4$).

	Forward Ordering	Reverse Ordering	Adaptive Ordering
Image 1:			
$k = 2$	163476	162432	163252
$k = 4$	148027	146630	146747
$k = \infty$	132156	130257	129683
Image 2:			
$k = 2$	147603	147247	145915
$k = 4$	114940	114530	111501
$k = \infty$	82405	81777	77120
Image 3:			
$k = 2$	293211	293067	292352
$k = 4$	243960	244562	241912
$k = \infty$	195198	195792	191804
Image 4:			
$k = 2$	583396	579372	578165
$k = 4$	542996	537947	534577
$k = \infty$	503548	496238	491533
Image 5:			
$k = 2$	315514	314772	313800
$k = 4$	270788	269107	266576
$k = \infty$	224792	222687	218542
Image 6:			
$k = 2$	223402	223495	222072
$k = 4$	170847	170636	167751
$k = \infty$	118565	118460	113778
Image 7:			
$k = 2$	638451	641133	635789
$k = 4$	590304	594525	585217
$k = \infty$	542808	548127	535221
Image 8:			
$k = 2$	297418	300717	295936
$k = 4$	227666	233657	225470
$k = \infty$	157859	166726	155089
	Average Total Bits per Document		
$k = 2$	332809	332779	330910
$k = 4$	288691	288949	284969
$k = \infty$	244666	245008	239096
	Average Transmission Time per Document (in Seconds)		
$k = 2$	69.34	69.33	68.94
$k = 4$	60.14	60.20	59.37
$k = \infty$	50.97	51.04	49.81

Table VII—Compression results with adaptive ordering

The first run of the form 0,0,0,· · ·,0,1 of each line is dropped. The adaptive ordering includes the bits necessary to specify the direction of the ordering. In the case of 0 ms, bits necessary for end-of-line codes are not included.

	0 ms	5 ms	10 ms
Adaptive Ordering, Total Bits			
Image 1:			
$k = 2$	159426	197893	230952
$k = 4$	144889	183421	217286
$k = \infty$	129683	168244	202963
Image 2:			
$k = 2$	137856	168016	185731
$k = 4$	107433	138182	159785
$k = \infty$	77120	108503	133731
Image 3:			
$k = 2$	286949	317318	333089
$k = 4$	239228	270222	289416
$k = \infty$	191804	223424	245902
Image 4:			
$k = 2$	571272	601723	615573
$k = 4$	531166	561568	575811
$k = \infty$	491533	521523	536679
Image 5:			
$k = 2$	306399	336372	351292
$k = 4$	262874	293520	311645
$k = \infty$	218542	249897	271152
Image 6:			
$k = 2$	215791	246276	261758
$k = 4$	164517	195059	211925
$k = \infty$	113778	144418	162820
Image 7:			
$k = 2$	627383	655650	663814
$k = 4$	580995	609353	617502
$k = \infty$	535221	563682	571841
Image 8:			
$k = 2$	291264	318521	326718
$k = 4$	223086	250612	260626
$k = \infty$	155089	182816	194551
Average Number of Bits per Image			
All Images:			
$k = 2$	324543	355221	371116
$k = 4$	281774	312742	330500
$k = \infty$	239096	270313	289955
Average Transmission Times (in Seconds)			
All Images:			
$k = 2$	67.61	74.00	77.32
$k = 4$	58.70	65.15	68.85
$k = \infty$	49.81	56.32	60.41

REFERENCES

1. A. N. Netravali, F. W. Mounts, and E. G. Bowen, "Ordering Techniques for Coding of Two-Tone Facsimile Pictures," B.S.T.J., *55*, No. 10 (December 1976), pp. 1539–1552.
2. A. N. Netravali, F. W. Mounts, and J.-D. Beyer, "Techniques for Coding Dithered Two-Level Pictures," B.S.T.J., *56*, No. 5 (May–June 1977), pp. 809–819.
3. F. W. Mounts, A. N. Netravali, and K. A. Walsh, "Some Extensions of the Ordering Techniques for Compression of Two-Level Facsimile Pictures," B.S.T.J., *57*, No. 8 (October 1978), pp. 3057–3067.

4. D. Preuss, "Comparison of Two-Dimensional Facsimile Coding Schemes," International Conference on Communications, June 1975, pp. 7.12–7.16.
5. Y. Yasuda and K. Arai, "Facsimile Data Compression by Rearranging Picture Elements," 1977 Picture Coding Symposium, Tokyo, Japan.
6. Y. Ueno, F. Ono, T. Semasa, S. Tomita, and R. Ohnishi, "Comparison of Facsimile Data Compression Methods," ICC '78 Conference Record, *3*, June 4–7, 1978, pp. 48.4.1–48.4.6.

Contributors to This Issue

Anthony S. Acampora, B.S.E.E., 1968, M.S.E.E., 1970, Ph.D., 1973, Polytechnic Institute of Brooklyn; Bell Laboratories, 1968—. Mr. Acampora initially worked in the fields of high power microwave transmitters and radar system studies and signal processing. Since 1974, he has been studying high capacity digital satellite systems. His current research interests are modulation and coding theory, time division multiple access methods, and efficient frequency re-use techniques. Member Eta Kappa Nu, Sigma Xi, IEEE.

William L. Aranguren, B.S., Monmouth College; M.S., Rutgers University; Bell Laboratories, 1970—. Mr. Aranguren is presently in the Voiceband Data Communications Department. Prior to his recent work with voiceband data systems, his interests included phased array combining and null steering techniques. Member, Eta Kappa Nu.

E. G. Bowen, Assoc. Deg. (E.E.), 1963, Vermont Technical College, B.S. (E.E.), 1971, Newark College of Engineering, Bell Laboratories, 1963—. Mr. Bowen has worked on analog-to-digital magnetic coders, plasma display panels, and coding of video signals.

L. A. D'Asaro, B.S. (Physics), 1949, M.S. (Physics), 1950, Northwestern University; Ph.D. (Physics), 1955, Cornell University; Bell Laboratories, 1955—. At Bell Laboratories, Mr. D'Asaro has been concerned with development of semiconductor devices including junction lasers. At the present time, he is participating in the development of microwave gallium arsenide field-effect transistors. Member, IEEE, American Physical Society, Sigma Xi, Phi Beta Kappa, Gamma Alpha.

Donald G. Duff, B.S. (E.E.), 1970, University of Arizona; M.S. (E.E.), 1971, Stanford University; D. Engr. (E.E.), 1977, University of

California at Berkeley; Bell Laboratories, 1970—. Mr. Duff has worked on device modeling for computer-aided circuit design and distortion analysis. He is currently concerned with systems planning for long-distance fiber-optic communication systems.

Martin Feldman, B.S. (Physics), 1957, Rensselaer Polytechnic Institute; Ph.D. (Experimental Physics), 1962, Cornell University; Assistant Professor, University of Pennsylvania, Philadelphia, 1963–1968; Bell Laboratories, 1968—. At the University of Pennsylvania, Mr. Feldman worked in high-energy nuclear physics. At Bell Laboratories, he has worked on holographic memories, display systems, and laser machining systems. Currently he is working on automatic inspection and alignment systems for integrated circuit lithography. Member, American Physical Society, IEEE.

Rollin E. Langseth, B.S., M.S., Ph.D. (all Electrical Engineering), University of Minnesota; Bell Laboratories, 1968—. At Bell Laboratories, Mr. Langseth first worked on problems in UHF mobile telephony. He has most recently been a member of the Satellite Systems Research department concerned with the design and application of digital satellite systems. Currently, he is a supervisor in the Communications Methods Research Department. Member Eta Kappa Nu, Tau Beta Pi, IEEE.

J. E. Mazo, B.S. (Physics), 1958, Massachusetts Institute of Technology; M.S. (Physics), 1960, and Ph.D. (Physics), 1963, Syracuse University; Research Associate, Department of Physics, University of Indiana, 1963–1964; Bell Laboratories, 1964—. At the University of Indiana, Mr. Mazo worked on studies of scattering theory. At Bell Laboratories, he has been concerned with problems in data transmission and is now working in the Mathematical Research Center. Member, American Physical Society, IEEE.

Carol A. McGonegal, B.S. (cum laude) (mathematics), 1974, Fairleigh Dickinson University; M.S. (computer science), 1977, Stevens Institute of Technology; Bell Laboratories, 1967—. Ms. McGonegal is a member of the Acoustics Research Department, where she has worked on problems in digital filter design, digital speech processing, computer voice response, and speaker verification.

Richard C. Miller, B.A. (Mathematics), 1949, University of Chicago; Ph.D. (Physics), 1957, University of Illinois; Bell Laboratories, 1957—. At Bell Laboratories, Mr. Miller has worked on electron device designs, diagnostic studies of gaseous laser systems, and laser microrecording systems. He is currently investigating photovoltaic device applications.

F. W. Mounts, E.E., 1953, M.S., 1956, University of Cincinnati; Bell Laboratories, 1956—. Mr. Mounts has been concerned with research in efficient methods of encoding pictorial information for digital television and graphics systems. Member, Eta Kappa Nu; Senior Member, IEEE.

Arun N. Netravali, B. Tech. (Honors), 1967, Indian Institute of Technology, Bombay, India; M.S., 1969, and Ph.D. (E.E.), 1970, Rice University; Optimal Data Corporation, 1970–1972; Bell Laboratories, 1972—. Mr. Netravali has worked on problems related to filtering, guidance, and control of the space shuttle. At Bell Laboratories, he has worked on various aspects of signal processing, and is presently Head of the Visual Communication Research Department. He is also an adjunct professor of Electrical Engineering at Rutgers University. Member, Tau Beta Pi, Sigma Xi; Senior Member, IEEE.

Hilding M. Olson, B.S.E.E., 1948, Northwestern University; M.S. (E.E.), 1954, Stevens Institute of Technology; D.E.E., 1959, Polytechnic Institute of New York; Mr. Olson has worked on electron tubes and semiconductor devices for microwave applications. He is currently working on problems of bipolar transistors. Member, Tau Beta Pi, Eta Kappa Nu, Sigma Xi.

Lawrence R. Rabiner, S.B. and S.M., 1964, Ph.D. (electrical engineering), 1967, Massachusetts Institute of Technology; Bell Laboratories, 1962—. From 1962 through 1964, Mr. Rabiner participated in the cooperative plan in electrical engineering at Bell Laboratories. He worked on digital circuitry, military communications problems, and problems in binaural hearing. Presently, he is engaged in research on speech communications and digital signal processing techniques. He is coauthor of *Theory and Application of Digital Signal Processing* (Prentice-Hall, 1975) and *Digital Processing of Speech Signals* (Prentice-Hall, 1978). Former President, IEEE G-ASSP Ad Com;

former Associate Editor, G-ASSP Transactions; former member, Technical Committee on Speech Communication of the Acoustical Society. Member, G-ASSP Technical Committee on Speech Communication, IEEE Proceedings Editorial Board, Eta Kappa Nu, Sigma Xi, Tau Beta Pi. Fellow, Acoustical Society of America and IEEE.

David C. Rife, B.S.E.E., 1960, University of Washington; M.E.E., 1962, New York University; Ph.D. (E.E.), 1973, Polytechnic Institute of Brooklyn; Bell Laboratories, 1960—. Mr. Rife has worked on a data carrier system and has supervised numerous development projects including automatic calling units, data test sets, and the Automatic Data Test System. He is currently supervising the registration of data station equipment. Senior member, IEEE; member, Tau Beta Pi, Phi Beta Kappa, Sigma Xi.

Aaron E. Rosenberg, S.B. (E.E) and S.M. (E.E.), 1960, Massachusetts Institute of Technology; Ph.D. (E.E.), 1964, University of Pennsylvania; Bell Laboratories, 1964—. Mr. Rosenberg is presently engaged in studies of systems for man-machine communication-by-voice in the Acoustics Research Department at Bell Laboratories. Member, Eta Kappa Nu, Tau Beta Pi, Sigma Xi; fellow, Acoustical Society of America; member, IEEE and IEEE Acoustics, Speech, and Signal Processing Group Technical Committee on Speech Processing.

H. A. Watson, B.A.Sc. (Engineering Physics), 1948, University of Toronto; M.Sc. (Applied Mathematics), 1949, McGill University; Ph.D. (Nuclear Physics), 1952, Massachusetts Institute of Technology; Bell Laboratories, 1952—. At Bell Laboratories, Mr. Watson has worked on a number of electron device, imaging device, and display device development projects, as well as digital transmission system development. He is now head of the Integrated Circuit Technology Evaluation Department.

R. H. Willens, B.S., M.S., Ph.D. (Physics and Engineering Science), California Institute of Technology; Bell Laboratories, 1966—. At Bell Laboratories, Mr. Willens has been involved in basic research in materials, including studies in superconductivity, magnetic properties, and band structure of solids. He is a member of the Physical Metallurgy and Ceramics Research Department. Member, American Physical Society, Metallurgical Society of AIME, American Society for Metals.

Clark B. Woodworth, B.S. (Electrical Engineering), Monmouth College; Bell Laboratories, 1977—. Mr. Woodworth is in the Satellite Systems Research Department which develops new concepts for future generation digital satellites. Member, Eta Kappa Nu, Sigma Pi Sigma, IEEE.

Papers by Bell Laboratories Authors

PHYSICAL SCIENCES

Absorption Curve of the 607-μm Line in Benzene. J. Stone, Appl. Opt. *17* (1978), p. 2876.

Anisotropic Thermal Expansion Behavior in Transcrystalline Polyethylene. T. T. Wang, T. Nishi, and H. M. Zupko, J. Mater. Sci., *13* (1978), pp. 2524–6.

Brillouin Scattering from Polymer Blends. G. D. Patterson, Polym. Preprints, *19*, No. 1 (1978), pp. 149–51.

Characterization of a Nonlinear Cochlear Model by Wide Band Stimuli. J. L. Hall, J. Acoust. Soc. Amer., *64* (1978), p. S133.

Chemical Kinetics of the Reactions of Silicon Tetrachloride, Silicon Tetrabromide, Germanium, Tetrachloride, Phosphorus Chlorate, and Boron Trichloride with Oxygen. W. G. French, L. J. Page, and V. A. Foertheyer, J. Phys. Chem., *82* (1978), pp. 2191–4.

Classical Trajectory Calculations of the Dissociation of Hydrogen on Copper (III) The Effect of Surface Roughness. A. Gelb and M. J. Cardillo, Surf. Sci., *75* (July 1978), pp. 199–214.

Comments on the Elucidation of Stereochemical Control in Alpha-Olefin Polymerization via Carbon-13 NMR Spectroscopy of Ethylene-Propylene Copolymers. A. E. Tonelli, Macromolecules, *11* (1978), pp. 634–6.

Comparison of Two Independent Determinations of the Structure of Calcium Ascorbate Dihydrate. S. C. Abrahams, ACTA Crystalline, *34* (Oct. 15, 1978), pp. 2981–5.

Compatibility and Physical Properties of Blends of Poly(Vinyl Chloride) and PB 3041, A Polymeric Plasticizer. H. E. Bair, D. Williams, T. K. Kwei, and F. J. Padden, Polym. Preprints, *19* No. 1 (Mar 1978), pp. 143–8.

Considerations in Dynamic Time Warping Algorithms for Discrete Word Recognition. L. R. Rabiner, A. E. Rosenberg, and S. E. Levinson, IEEE Trans. Acoust. Speech and Signal Processing, *ASSP-26*, No. 6 (December 1978) pp. 575–82.

Continuous Colorimetric Monitoring of Vapor-Phase Urea and Dynamics. L. E. Trimble and R. J. H. Voorhoeve, Analyst, *103* (1978), pp. 759–65.

Conversion of Nitric Oxide to Isodyanic Acid and Ammonium Dyanate Over Palladium, Iridium, and Platinum—10% Rhodium Catalysts. R. J. H. Voorhoeve and L. E. Trimble, J. Cata, *54* (1978), pp. 269–80.

Correlation Between Elastic Constants and Flow Behavior in Metallic Glasses. H. S. Chen, J. Appl. Phys., *49*, No. 1 (Jan 1978), pp. 143–8.

A Coulometric Analysis of Iron II in Ferrites Using Chlorine. P. K. Gallagher, Amer. Ceram. Soc. Bull., *57* (June 1978), pp. 578–8.

Defects in Indium Phosphide Homoepitaxial Layers. S. Mahajan, K. J. Bachmann, D. Brasen and E. Buehler, J. Appl. Phys., *49*, No. 1 (Jan 1978), pp. 245–8.

Depolarized Rayleigh Scattering from Benzonitrile. G. R. Alms and G. D. Patterson, J. Chem. Phys., *68* (1978), pp. 3440.

Diffusion in a Palladium Copper-Silicon Metallic Glass. H. S. Chen, L. C. Kemerling, J. M. Poate, and W. L. Brown, Appl. Phys. Lett., *32*, No. 8 (Apr 1978), pp. 461–3.

The Diffusion of Water in Optical Fibers. K. Nassau, Mat. Res. Bull., *13*, No. 1 (Jan 1978), pp. 67–76.

Digital Processing of Speech Signals. L. R. Rabiner and R. W. Schaefer, Englewood Cliffs: Prentice-Hall, 1978.

Dislocation—Free and Low Dislocation Quartz Prepared by Hydrothermal Crystallization. R. L. Barns, P. E. Freeland, E. D. Kolb, R. A. Laudise, and J. P. Patel, J. Cryst. Growth, *43*, No. 8 (July 1978), pp. 676–86.

A Dispersive Phase Shifter for Use in Nonlinear Optical Interference Experiments. C. R. Crane and J. G. Bergman, J. Appl Phys., *50*, No. 1 (1979), pp. 47–8.

Dynamics of the Charge Density Wave II. Long-Range Coulomb Effects in an Array of Chains. P. A. Lee and H. Fukuyama, Phys. Rev. B, *17*, No. 2 (Jan 15, 1978), pp. 542–8.

Dynamics of the Roughness Transition. S. T. Chui and J. D. Weeks, Phys. Rev. Lett., *40* (1978), pp. 733–6.

Effect of Crystallographic Factors on Spherulitic Morphology, as Evidenced in Unidirectionally Solidified Polyamides. A. J. Lovinger, Bull. Amer. Phys. Soc., *23*, No. 3 (Mar 1978), p. 420.

Effect of Drawing Tension on Residual Stresses in Clad Glass Fibers. C. R. Kurkjian and U. C. Paek, J. Amer. Ceram. Soc., *61*, No. 3–4 (Mar/Apr 1978), pp. 176–7.

Effects of Chemical Treatment on Loss Quality of Microwave Dielectric Ceramics. H. M. O'Bryan, J. Thompson, and J. K. Poourde, Ber. Deut. Chem. GE8, *55*, No. 7 (1978), pp. 348–51.

Electrical Properties of Binary Amorphous Alloys. J. J. Hauser and J. Tauc, Phys. Rev. B, *17*, No. 8 (Apr 1978), pp. 3371–80.

Electron Irradiation of Polymers and Its Application to Resists for Electron Beam Lithography. M. J. Bowden, Critical Rev. in Solid State Science, *8* (1979), p. 223.

Electronic Orbitals of Se Bonded to Ni(100). H. Froitzheim and H. D. Hagstrum, J. Vac. Sci. Technol., *15* (1978), pp. 485–487.

Enhanced Photoelectrochemical Solar Energy Conversion by Gallium Arsenide Surface Modification. B. A. Parkinson, A. Heller, and B. Miller, Appl. Phys. Lett., *33*, No. 6 (Sept 15, 1978), pp. 521–3.

EPR Spectra of Carbalkoxycarbenes, Geometric Isomerism in Ground State Triplets. R. S. Hutton and H. D. Roth, J. Amer. Chem. Soc., *100* (1978), pp. 4324–5.

ESR Studies of Sulfur-Based Donor Heterocycles: Sulfur-33 Couplings. F. B. Beamwell, R. C. Haddon, F. Wudl, and M. L. Kaplan, J. H. Marshall, J. Amer. Chem. Soc., *100* (1978), pp. 4612–4.

Examination of Aluminum-Copper Films During the Galvanostatic Formation of Anodic Oxide—Part II, Rutherford Backscattering and Depth Profiling. H. H. Streblow, C. M. Melliar-Smith, and W. M. Augustyniak, J. Electrochem. Soc., June 1978, pp. 915–8.

Filter Characteristics of Codirectionally Coupled Waveguides with Weighted Coupling. R. C. Alferness and R. S. Cross, IEEE J. Quantum Electron., *14* (1978), pp. 843–7.

FIR System Modeling and Identification in the Presence of Noise and Band-limited Inputs. L. R. Rabiner, R. E. Crochiere, and J. B. Allen, IEEE Trans. Acoust. Speech and Signal Processing (Aug 1978), pp. 319–33.

Fluorinated Molecule as a Tracer: Difluoroserotonin in Human Platelets Mapped by Electron Energy Loss Spectroscopy. J. L. Costa, D. C. Joy, D. M. Maher, K. Kirk, and S. W. Hui, Science, *200* (1978), pp 537–9.

The Formation of Compounds with Nitrogen-Carbon Triple Bond During Reduction of Nitric Oxide Over Rhodium and Platinum Catalysis. R. J. H. Voorhoeve, C. N. K. Patel, L. E. Trimbole, P. J. Kerl, J. Catal, *54* (1978), pp. 102–5.

Furnace-Drawn Silica Fibers With Tensile Strength Greater Than 3.5 Gn/M(2) (500 kpsi) in 1 Km Lengths. F. V. DiMarcello and A. C. Hart, Electron. Lett., *14*, No. 18 (Aug 1978), pp. 578–9.

Guest Editorial for Special Issue on Emerging 900-MHz Technologies. W. H. Chriss and J. J. Milulski, IEEE Trans. Veh. Technol., *27*, No. 4 (Nov 1978), p. 173.

Hydrolysis Reactions of Transition Metal Hexafluorides in Liquid Hydrogen Fluoride, Oxonium Salts with Platinum, Iridium and Ruthenium. H. Selig, W. A. Sunder, F. A. DiSalvo, and W. E. Falconer, J. Fluor. Chem *11* (1978), pp. 39–50.

Induced Magnetic Form Factor of Samarium in Mixed-Valance Compounds. R. M. Moon, W. O. Koehler, D. B. McWhan, and F. Holtzberg, J. Appl. Phys., *49* (1978), pp. 2107–12.

Interaction Parameter in Polymer Mixtures. T. K. Kwei and H. L. Frisch, Macromolecule, *11* (1978), p. 267.

Intercalation of Halogen Fluorides into Graphite. H. Selig, W. A. Sunder, Vasile M. J. F. A. Stevie, P. K. Gallagher, and L. B. Ebert, J. Fluor. Chem., *12* (1978), pp. 397–412.

The Liquid Phase Epitaxy of Aluminum (Y) Gallium (1-Y) Arsenic (1-X) Antimony (X) and the Importance of Strain Effects Near the Miscibility Gap. R. E. Nahory, M. A. Pollack, E. D. Beebe, D. E. Winter, and J. C. Ilegems., J. Electrochem. Soc., *125* (1978), pp. 1053–1118.

Low Temperature Magnetic Susceptibilities of Cubic Langreinite Transition Metal Sulfates. J. M. Waszczak, Magn. Lett., *1*, No. 4 (1971), pp. 97–101.

Magnetic and Structural Properties of Some Amorphous Rare Earth Transition Metal Thin Films. A. G. Dirks and H. J. Leanhny, IEEE Trans. Magn., *14* (1978), pp. 835–7.

Magnetic Properties of Manganese Silicon-Cobalt Silicon Solid Solution Alloys. K. Motoya, H. Yasudka, Y. Nakamura, and J. H. Wernick, J. Phys. Soc., *44*, No. 5 (1978), 1525–32.

Magnetic Susceptibility of an Amorphous Spin Glass: Gold, Silicon, Manganese. J. J. Hauser and J. V. Waszczak, Phys. Rev. B, *8* (Dec 1978), pp. 6206–12.

A Method for Evaluating Viscosities of Metallic Glasses from the Rates of Thermal Transformations. H. S. Chen, J. Crystal Solids, *27* (1978), pp. 31–4.

Model Study of Zwicker's Masking Period Patterns. J. L. Hall, J. Acoust. Soc. Amer., *64* (Aug 1978), pp. 473–7.

Molecular Orbital Theory of Polyenes Implicated in the Dehydrochlorination of Poly(vinyl Chloride), II. Electronic Structures, Equilibrium Geometries, and Energetics of the Ground States of Polyene Cations and Neutral Polyenes. R. C. Haddon and W. H. Starnes, Jr., Amer. Chem. Ser., *169* (1978), pp. 333–62.

Neutron Scattering at High Pressure in Singlet-Triplet Systems. D. B. McWhan in *High-Pressure and Low-Temperature Physics*, C. W. Chu and J. A. Woollam, New York: Plenum Press, 1978, pp. 87–96.

A New Self-Consistent Greens Function Approach to the Electronic Structure of Localized Defects. G. A. Baraff and M. Schluter, Phys. Rev. Lett., 1978, No. 2, pp. 892–5.

Nonlinear Electrooptic Fabry-Perot Devices Using Reflected Light Feedback. P. W. Smith, E. H. Turner, and B. B. Mumford, Opt. Lett., *2*, No. 3 (Mar 1978), pp. 55–7.

Numerical Evaluation of the Impulse Response of Multimode Optical Fibers. J. Arnaud, Fiber Integ. Opt, *1*, No. 1 (1977), pp. 77–100.

Observations of Anomalous Intensities in the Lines of the Formyl Ion (+) Isotopes. W. D. Langer, R. W. Wilson, P. S. Henry, and M. Guelin, Astrophys. J., *225*, (Nov 1, 1978), pp. 2139–42.

Optical Radiation from Low Energy Hydrogen Atomic and Molecular Ion-Surface Collisions. S. Y. Leung, N. H. Tolk, W. Heiland, J. C. Tully, J. S. Kraus, and P. Hill, Phys. Rev. A, *18* (1978), pp. 447–51.

On the Orientation of Hydromagnetic Waves in the Magnetosphere. A Hasegawa and L. J. Lanzerotti, Rev. Geophys. Space Phys. *16*, No. 2 (May 1978), pp. 263–6.

Parametric Excitation of Drift Waves with the Pump Near the Ion Cyclotron Frequency in a Two-Ion Species Plasma. M. Ono, M. Porkolab, and R. P. H. Chang, Phys. Lett., *67A* (1978), p. 379.

Parametric Lower Hybrid Instability Driven by Modulated Electron Beam Injection. G. R. Allen, D. K. Owens, S. W. Seiler, M. Yamada, H. Ikezi, and M. Porkolab, Phys. Rev. Lett., *41*, No. 15 (Oct 1978), pp. 1045–8.

Particle Kinetics of Selective Excitation Spectroscopy. R. A. Stern, Phys. Fluids, *21*, No. 8 (Aug 1978), pp. 1287–94.

Phenalene[1,9-CD]Dithiolyl: The First Example of a Monomeric, Coplanar, Carbon-Based Free Radical. R. C. Haddon, F. Wudl, M. L. Kaplan, J. H. Marshall, and F. R. Bramwell, J. Soc. Chem. Commun. (London) (1978), pp. 429–30.

Photocollection Efficiency and Interface Changes of MBE Grown Abrupt P-(Gallium Arsenide)-Nitrogen (Aluminum (0.33) Gallium (0.67) Arsenic) Hete-

rojunctions. H. Kroemer, W. Y. Chien, H. C. Casey, and A. Y. Cho, Appl. Phys. Lett., *33* (1978), p. 749.

Photoemission Studies of 2P Core Levels of Pure and Heavily Doped Silicon. W. Eberhardt, G. Kalkoffen, C. Kunz, D. E. Aspnes, and M. Cardona, Phys. Status Solidi, *(b)88* (1978), pp. 135–43.

Pinning and Conductivity of Two-Dimensional Charge Density Waves in Magnetic Fields. H. Fukuyama and P. A. Lee, Phys. Rev. B, *18* (1978), p. 6245.

Plastic Deformation of Vanadium (3) Silicon Single Crystals at Elevated Temperatures. S. Mahajan, J. H. Wernick, G. Y. Chin, S. Nakahara, and C. T. H. Geballe, Appl. Phys. Lett., *33*, No. 11 (Dec. 1, 1978), pp. 972-74.

Polarimetry of Specular and Nonmultiple Scattered Electromagnetic Radiation from Selectively Roughened Silicon Surfaces. M. D. Williams and D. E. Aspnes, Phys. Rev. Lett., *41* (1978), pp. 1667–70.

Polarization Characteristics of Noncircular Core, Single-Mode Fibers. V. Ramaswamy, W. G. French, and R. D. Standley, Appl. Opt., *17*, No. 18 (Sept. 15, 1978), pp. 3014–7.

Poly(vinylidene) Fluoride Used for Piezoelectric Coin Sensors. G. R. Crane, IEEE Trans. Sonic and Ultrasonic, *25*, No. 6 (1978), pp. 393–5.

Positronium Formation at Surface. A. P. Mills, Phys. Rev. Lett., *41* (1978), pp. 1828–31.

A Practical Electron Spectrometer for Chemical Analysis. D. C. Joy and D. M. Maher, J. Microscopy, *114* (Nov. 1978), pp. 117–29.

Properties of Polycrystalline Aluminum Arsenide/Gallium Arsenide on Graphite Heterojunctions for Solar Cell Applications. W. D. Johnston and W. M. Callahan, J. Electrochem. Soc., *125* (June 1978), pp. 977–83.

Pseudo-Three-Dimensional Turbulence in Magnetized Nonuniform Plasma. A. Hasegawa and K. Mina, Phys. Fluids, *21*, No. 1 (Jan 1978), pp. 87–92.

Pulsewidth Dependence on Intracavity Bandwidth in Synchronously Mode-Locked CW Dye Lasers. C. P. Ausschnitt and R. K. Jain, Appl. Phys. Lett., *32*, No. 11 (Jun 1, 1978), pp. 727–30.

Questions About Rotating Superfluid Dynamics: Problems of Pulsar Astrophysics Accessible in the Laboratory. P. W. Anderson, D. Pines, M. Ruderman, and J. Shaham, J. Low Temp. Phys. *30* (1978), p. 839.

Real-Time Extraction of Bubble Chamber Tracks Using a Single Vidicon. C. E. Roos, J. O. Limb, and E. G. Bowen, Nucl. Instr. and Methods, *153* (1978), pp. 383–8.

A Refraction Correction for Rheed Patterns from Amorphous Surfaces. K. J. Matysik, J. Appl. Phys., *49*, No. 6 (June 1978), pp. 3600–1.

Room Temperature Threshold Current Dependence of Gallium Arsenide-Aluminum (X) Gallium (1-X) Arsenide Double Heterostructure Lasers on X and Active Layer Thickness. H. C. Casey, J. Appl Phys., *49*, No. 7 (Jul 1978), pp. 3684–92.

The Rotating Sample Technique for Measurement of Random Backscattering Yields from Crystals and Its Application to Beta-Aluminia. P. Blood, L. C. Feldman, G. L. Miller, and J. P. Remeika, Nucl. Instr. and Methods, *149* (1978), pp. 225–8.

Rotational Excitation of Molecules by Electrons in Interstellar Clouds. P. Goldsmith, Astron. Astrophys. *54* (1978), p. 645.

Self-Consistent APW Electronic Structure Calculations for the A-15 Compounds Vanadium (3X) and Niobium (3X), X = Aluminum, Gallium, Silicon, Germanium, and Tin. B. M. Klein, L. L. Boyer, D. A. Papaconstantopoulos, and L. F. Mattheiss, Phys. Rev. B., *18* (Dec 15, 1978), pp. 6411–38.

A Simple High Vacuum Evaporation System with Low Temperature Substrate. J. P. Garno, Rev. Sci. Instrum., *49*, No. 8 (Aug 1978), pp. 1218–20.

Simultaneous Observation of Nuclear Spin Polarization and Line Broadening, Evidence for an Alternative Polarization Mechanism. M. L. M. Schilling and H. D. Roth, J. Amer. Chem. Soc., *100* (1978), pp. 6533–5.

Single Crystal Aluminum Schottky Barrier Diodes Prepared by Molecular Beam Epitaxy. A. Y. Cho and P. D. Dernier, J. Appl. Phys, *49* (June 1978), pp. 3328–32.

Some Viscoelastic Crack Growth Relations for Orthotropic and Prestrained

Media. G. S. Brockway and R. A. Schapery, Eng. Fract. Mech., *10* (June 1978), pp. 453–68.

Spatially Controlled Crystal Regrowth of Ion Implanted Silicon by Laser Irradiation. G. K. Celler, J. M. Poate, and L. C. Kimerling, Appl. Phys. Lett., *32* (1978), pp. 464–6.

Spatially Resolved Photoluminescence Characterization and Optically Induced Degradation of Indium (1-X) Gallium (X) Arsenic (Y) Phosphorus (1-Y) DH Laser Material. W. D. Johnston, G. Y. Epps, R. E. Nahory, and M. A. Pollack, Appl. Phys. Lett., *33* (Dec 1978), pp. 992–4.

Specific Heat and Phonon Dispersion in Liquid Helium-4. D. S. Greywall, Phys. Rev. B, *18* (Sept 1, 1978), pp. 2127–44.

Spin-Wave Excitations and Low Temperature Magnetization in the Amorphous Metallic Ferromagnets (Iron (X) Nickel (1-X)) (75) Phosphorus (16) Boron (6) Aluminum (3). R. J. Birgeneau, J. A. Tarvin, G. Shirane, E. M. Gyorgy, R. C. Sherwood, H. S. Chen, and C. L. Chien, Phys. Rev. B, *18* (Sept 1978), pp. 2192–5.

Sputtering of Platinum Silicide. Z. L. Liau, J. M. Mayer, W. Brown, and J. M. Poate, J. Appl Phys., *49*, No. 10 (Oct 1978), pp. 5295–305.

The Stability of Interstellar Clouds Containing Magnetic Fields. W. D. Langer, Astrophys. J., *225* (1978), p. 95.

Statics and Dynamics of Incommensurate Lattices. G. Theodorou and T. M. Rice, Phys. Rev. B, *18* (1978), pp. 2840–56.

Stationary Spectrum of Pseudo-Three-Dimensional Turbulence in Magnetized Plasmas. A. Hasegawa, T. Inamura, K. Mina, and T. Taniuti, J. Phys. Soc. Jap., *45*, No. 3 (Sept 1978), pp. 1005–10.

Stereoregulation Energies in Propene Polymerization. A. Zambelli, P. Locatelli, G. Zannoni, and P. A. Bovey, Macromolecules, *11* (1978), p. 923.

Structural Aspects of Polymerization in Crystalline Ionic Monomers: Comparison of Barium Methacrylate Monohydrate and Anhydrate. M. J. Bowden, C. H. L. Kennard, J. H. O'Donnell, R. D. Sothman, N. W. Isaacs, and G. Smith, J. Macromolecules, Sci. Chem., *12* (1978), pp. 63.

Structural Defects in Annealed Silicon Containing Oxygen. A. Staudinger, J. Appl. Phys. *49*, No. 7 (July 1978), pp. 3870–4.

Structure of the Electron Gas at the Surface of Liquid Helium. C. C. Grimes, J. de Phys. *III* (1978), pp. 1352–7.

Studies of Adsorbate Electronic Structure Using Ion-Neutralization of Photoemission Spectroscopies. H. D. Hagstrum in *Electron and Ion Spectroscopy of Solids,* L. Freimans, J. Vennik, and W. Dekeysa, eds. New York: Plenum Press, 1978, pp. 273–323.

Studies of the Silicon-Silicon Dioxide Interface by MEV Ion Scattering. L. C. Feldman, I. Stensgaard, P. J. Silverman, and T. E. Jackman in *The Physics of Silicon Dioxide and Its Interfaces,* S. T. Pantelides, ed., New York: Pergamon Press, 1978, pp. 844–51.

A Study of Surface Reproducibility and Impurity Segregation Using Ion Neutralization and Photoemission Spectroscopies. T. Sakurai and H. D. Hagstrum, Surf. Sci., *70* (1978), pp. 617–28.

Study of Zone Folding Effects on Phonons in Alternating Monolayers of Gallium Arsenide-Aluminum Arsenide. A. S. Barker, J. L. Merz, and A. C. Gossard, Phys. Rev. B, *17*, No. 8 (Apr 15, 1978), pp. 3181–96.

Subpicosecond Spectroscopy. E. P. Ippen and C. V. Shank, Phys. Today, *41*, No. 7 (May 1978).

Surface Properties of $TiSe_2$. W. Fabian, C. H. Chen, and F. C. Brown, Bull. Amer. Phys. Soc., *23* (1978), p. 390.

Surface Segregation in Copper-Nickel Copper Potassium Alloys. A Comparison of Low-Energy Ion-Scattering Results with Theory. H. H. Brongersma, M. J. Sparnaay, and T. M. Buck, Surf. Sci., *71*, No. 1 (1978), pp. 657–78.

Surface Studies of Tungsten by MEV He Scattering. R. A. Zuhr, L. C. Feldman, R. L. Kauffman, and P. J. Silverman, Nucl. Instr. and Methods, *149* (1978), pp. 349–54.

Temperature Dependence of Exchange Narrowing in the One-Dimensional Antiferromagnet Nitrogen (Carbon Hydrogen (3)) (4) Manganese Chlorine (3) (TMMC). T. T. P. Cheung, Z. G. Soos, R. E. Dietz, and F. R. Merritt, Phys. Rev. B, *17* (1978), pp. 1266–76.

Thermal Expansion of Amorphous Metal Alloys. H. A. Brooks, J. Appl. Phys., *49,* No. 1 (Jan 1978), pp. 213–4.

Transport Properties and the Phase Transition on Titanium (1-X) M(X) Selenium (2) (M = Tantalum or Vanadium). F. J. DiSalvo and J. V. Waszczak, Phys. Rev. B, *17* (1978), pp. 3801–7.

The Tunneling Density of States of Superconductive Niohium-Tin. D. F. Moore, M. R. Beasley, and J. M. Posell, J. de Phys., *39,* No. 6 (1978), p. 1390.

Tunneling of Solitons and Charge Density Waves Through Impurities. A. I. Larkin and P. A. Lee, Phys. Rev. B, *17* (1978), p. 1596.

The Two-Dimensional Electron Gas in a Strong Magnetic Field. H. Fukuyama, P. M. Platzman, and P. W. Anderson, Surf. Sci., *73* (1978), p. 374.

Use of Ion Beam Techniques to Characterize Thin Plasma Grown GaAs and GaAlAs Oxide Films. R. L. Kauffman, L. C. Feldman, and R. P. H. Chang, Nucl. Instr. Methods, *149* (1978), pp. 619–22.

Use of the Parallel Plate Plastometer to Characterize Glass-Reinforced Resins: I. Flow Model. C. J. Bartlett, J. Elastomers Plast., *10* (Oct. 1978), pp. 369–76.

Use of Thin Silicon Crystals in Backscattering Channeling Studies of the Silicon-Silicon Dioxide Interface. L. C. Feldman, P. J. Silverman, J. Williams, and T. Jackman, I. Stensgaard, Phys. Rev. Lett., *20* (Nov 13, 1978), pp. 1396–9.

Vacuum Interlock System for Molecular Beam Epitaxy. J. W. Robinson and M. Ilegems, Rev. Sci. Instrum., *49,* No. 2 (Feb 1978), pp. 205–7.

Vortex Formation by Baroclinic Vector in Laser-Pellet Interaction. A. Hasegawa, M. Y. Yu, and P. K. Shukla, Phys. Rev. Lett., *41,* No. 24 (Dec 1978), p. 1656.

Weak Beam Electron Microscopy of Faulted Dipoles in Deformed Silicon. A. T. Winters, S. Mahajan, and D. Brasen, Phil. Mag. Part A, *37,* No. 3 (Mar 1978), pp. 315–26.

MATHEMATICS

Evaluation of a Determinant. D. Slepian, SIAM Rev., *20,* No. 3 (July 1978), p. 594.

Extremal Self-Dual Lattices Exist only in Dimensions 1–8, 12, 14, 15, 23, and 24. J. H. Conway, A. M. Odlyzko, and N. J. A. Sloane, Mathematika, *25* (1978), pp. 36–43.

Hamiltonian Decompositions of Products of Cycles. M. F. Foregger, Discrete Math., *24* (December 1978), pp. 251–60.

A Note on the Eigenvalues of Hermitian Matrices. D. Slepian and J. Landau, SIAM J. Math. Anal., *9,* No. 2 (Apr 1978), pp. 291–7.

Self-Dual Codes over GF(U). F. J. MacWilliams, A. M. Odlyzko, N. J. A. Sloane, and N. M. Ward, J. Comb. Theory A, *25* (1978), pp. 288–318.

Two- and 3-Dimensional Localization of Random Phase Waves. K, Mina and A. Hasegawa, Rocky Mt. Math. J., *8,* No 1–2 (1978), pp. 309–22.

ENGINEERING

On Creating Reference Templates for Speaker Independent Recognition of Isolated Words. L. R. Rabiner, IEEE Trans. Acout. Speech Signal Processing, *26,* No. 1 (February 1978), pp. 34–42.

The Data Encryption Standard Retrospective and Prospects. R. Morris, IEEE Trans. Commun., *16,* No 6 (November 1978), pp. 11–14.

Device Physics of Dual-Dielectric, Charge-Storage Cells. K. K. Thornber and D. Kahng, J. Electrochem. Soc, *78,* No. 2 (1978), pp. 616–7.

Electroless Copper Plating of Multilayer Printed Circuit Boards. C. J. Bartlett, R. D. Rust, and R. J. Rhodes, Plat. Surf. Finish., *65* (July 1978), pp. 36–41.

Ensuring Contact-Finish Quality of Selectively Plated Connectors. D. L. Hindermann, R. L. Meek, B. T. Kerns, M. E. Terry, and D. B. Bixler, W. Elec. Eng., *22* (Apr. 1978), pp. 5–11.

Long-Term Mechanical Behavior of Optical Fibers Coated with a UV-Curahle Epoxy Acrylate. T. T. Wang and H. M. Zupko, J. Mater. Sci., *13* (1978), pp. 2241–8.

New Grids for Improved Polarization Diplexing of Microwaves in Reflector Antennas. C. Dragone, IEEE Trans. Ant. Propag. *26,* No. 3 (May 1978), pp. 459–63.

SOCIAL AND LIFE SCIENCES

Industrial Health Education in Morristown, New Jersey. M. N. Wagner, N. Bryant, and R. B. Bauer, Public Health Rep., *91,* No. 3 (May–June 1978), pp. 273–4.

B. T. Matthias—Scientist and Humanist. A. M. Clogston, N. B. Hannay, and C. K. N. Patel, J. Less Common Metals, Matthias Issue, *62* (Nov–Dec 1978), p. vii.

Syllables as Concatenative Phonetic Units. O. Fujimura and J. B. Lovins in *Syllables and Segments,* A. Bell and J. B. Hooper, eds., Amsterdam: North-Holland, 1978, pp. 107–120.

MANAGEMENT AND ECONOMICS

Science and Technology as Society Resources. N. B. Hannay, N. J. Bell J., *1,* No. 2 (Fall 1978).