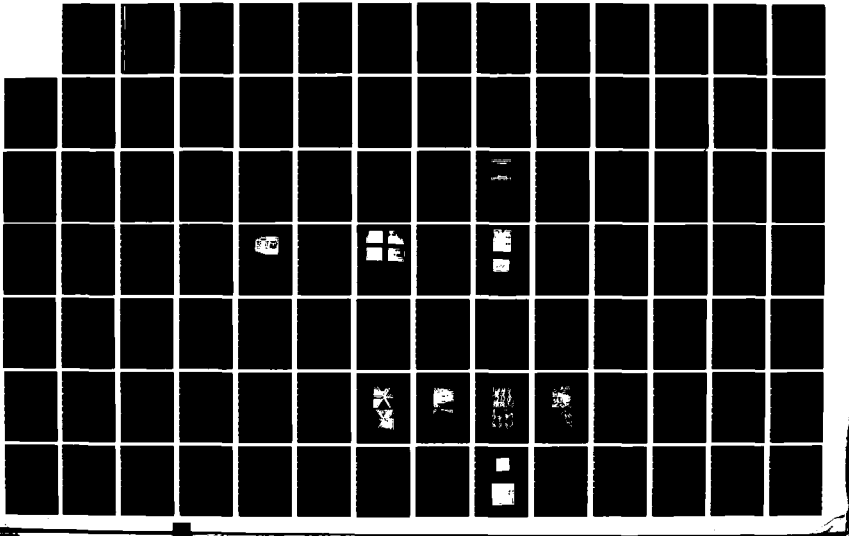
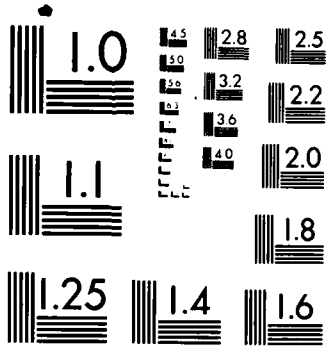
AD-A162 272 OPTICAL COMPUTING RESEARCH(U) STANFORD UNIV CA 1/2
INFORMATION SYSTEMS LAB J W GOODMAN ET AL 01 JUN 85
ISL-L722-10 AFOSR-TR-85-1037 AFOSR-83-0166

UNCLASSIFIED F/G 9/3 NL





AFOSR
INFORMATION SYSTEMS LABORATORY

STANFORD ELECTRONICS LABORATORIES
DEPARTMENT OF ELECTRICAL ENGINEERING
STANFORD UNIVERSITY · STANFORD, CA 94305

OPTICAL COMPUTING RESEARCH

Joseph W. Goodman
Raymond Kostuk
Ellen Ochoa
Bradley Clymer

June 1985

Annual Technical Report Number L722-10

Research supported by the Air Force Office of Scientific Research, Air Force Systems command, USAF, under Grant No. AFOSR 83-0166. The United States Government is authorized to reproduce and distribute reprints for governmental purposes, notwithstanding any copyright notation hereon.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

AD-A162272

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION: UNCLASSIFIED	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT: Approved for public release; distribution unlimited
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S): L722-10	5. MONITORING ORGANIZATION REPORT NUMBER(S): AFOSR-TR- 85-1037
6a. NAME OF PERFORMING ORGANIZATION: Stanford University; 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION: ~~AFOSR-TR- 85-1037~~ AFOSR/NE
6c. ADDRESS (City, State and ZIP Code): Stanford, CA 94305	7b. ADDRESS (City, State and ZIP Code): Bolling Air Force Base, D.C. 20332
8a. NAME OF FUNDING/SPONSORING ORGANIZATION: Air Force Office of Scientific Research; 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER: AFOSR-83-0166
8c. ADDRESS (City, State and ZIP Code): Bolling Air Force Base, D.C. 20332	10. SOURCE OF FUNDING NOS. — PROGRAM ELEMENT NO.: 61102F; PROJECT NO.; TASK NO.; WORK UNIT NO.
11. TITLE (Include Security Classification): Optical Computing Research	
12. PERSONAL AUTHOR(S): J. W. Goodman	
13a. TYPE OF REPORT: Annual; 13b. TIME COVERED: FROM 5/18/84 TO 5/17/85	14. DATE OF REPORT (Yr., Mo., Day): June 1, 1985; 15. PAGE COUNT: 111
16. SUPPLEMENTARY NOTATION	

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB. GR.	

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

This document contains information on the research accomplished under AFOSR Grant No. AFOSR 83-0166 during the time period 18 May 1984 through 17 May 1985. The work covers several different areas of optical computing and related topics. The primary emphasis of the work is on applications of optics to interconnections in the area of microelectronics. A second area of investigation is the inversion of images and wavefronts using photorefractive crystals, and the applications of such nonlinear operations to spatial filtering problems. Work has also been completed on the wavefront inversion using holograms and on the suppression of speckle in coherently formed images. Publications during the last year arising out of the grant are also attached.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT: UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. ☐ DTIC USERS ☐	21. ABSTRACT SECURITY CLASSIFICATION: UNCLASSIFIED
22a. NAME OF RESPONSIBLE INDIVIDUAL: ~~Prof. J. W. Goodman~~ Dr Files	22b. TELEPHONE NUMBER (Include Area Code): 415/497-3304 (202) 767-4933; 22c. OFFICE SYMBOL: NE

DD FORM 1473, 83 APR EDITION OF 1 JAN 73 IS OBSOLETE.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

ABSTRACT

This document contains information on the research accomplished under AFOSR Grant No. AFOSR 83-0166 during the time period 18 May 1984 through 17 May 1985. The work covers several different areas of optical computing and related topics. The primary emphasis of the work is on applications of optics to interconnections in the area of microelectronics. A second area of investigation is the inversion of images and wavefronts using using photorefractive crystals, and the applications of such nonlinear operations to spatial filtering problems. Work has also been completed on the wavefront inversion using holograms and on the suppression of speckle in coherently formed images. Publications during the last year arising out of the grant are also detailed.

I. INTRODUCTION

This document is an interim annual report on the research accomplished under the sponsorship of Air Force Office of Scientific Research Grant No. AFOSR 83-0166 during the time period May 18, 1984 through May 17, 1985. The research performed covers four different areas: (1) Chip-to-chip communication using optics, with emphasis on the properties of thick reflective holographic optical elements for realizing imaging interconnections; (2) Optical distribution of a clock to an integrated circuit chip; (3) Image inversion and nonlinear optical filtering in real time using photorefractive crystals; and; (4) wavefront inversion using holograms and suppression of speckle in coherently formed images (two projects completed in the past year). We summarize the progress in each of these areas (Sections II through V). In addition we present other information pertinent to the grant in the final section (Section VI).

II. Chip-Chip Interconnections Using Reflective Holographic Optical Elements

Interconnections are among the most challenging problems facing the electronics industry today. Optics is being successfully applied to the problem of machine-to-machine interconnections (fiber-optic local area networks), but there are interconnection problems at many other levels of electronic architectures, from high-speed busses within a single machine to board-to-board, chip-to-chip, and even within-chip communications. The most difficult problems in many respects are those at the lowest levels of architecture (chip-to-chip and within-chip), and it is at these levels that the majority of our work is focused.

The uniqueness of our work lies in the emphasis placed on "imaging" interconnections, for which a connection channel is established by means of a holographic imaging element that images a small source onto one or more small detectors. The holograms involved must be as efficient as possible, and must achieve resolutions that are commensurate with the minimum detector sizes anticipated in practice. If the source is modulated at high speed, then data is delivered at that speed to all detector sites to which the image of the source has been placed. The delivery of data from one source to many detectors is referred to as "fan-out".

Our activities on this project during the past grant year have been aimed at two goals: (1) a quantitative comparison of optical and electronic interconnections in the chip-to-chip communication problem, and (2) development of understanding and experimental know-how in the area of thick reflective holograms suitable for providing interconnections at the chip-to-chip level.

In the first area, a first-order analysis of the speed limitations of metalized interconnections and optical interconnections at the chip-to-chip level has been performed. The results of this study have been accepted for publication in *Applied Optics*, and are scheduled to appear in the September 1 issue. A preprint of this work is attached as Appendix A. The main conclusion arrived at in this analysis is that the speed limitations of optical and metalized interconnections are about the same, and therefore the prime advantage of optical interconnects rests in the immunity from cross-talk and interference, not in any speed advantage. If the interconnect lengths are longer than the few centimeters postulated

in our studies, then optics may have a speed advantage, provided that the capacitance of the metalized lines becomes comparable with the bonding pad capacitances. Bonding pad capacitances dominated the speed limitations in the problem analyzed by us.

In the second area of investigation, holographic optical elements, some important advances were made. On the experimental side, we have developed techniques that allow holographic reflectors with a single focal point to be realized in silver halide materials with diffraction efficiencies in excess of 50%. While this efficiency is much less than can be realized with dichromated gelatin, nonetheless in a research environment silver halide materials are much easier to work with, and the achievement of diffraction efficiencies in excess of 50% makes such elements quite useful. Secondly, we have been paying substantial attention to the problem of making holographic optical elements with multiple focal points, since such holograms are needed when interconnections must have fan-out. Considering the simplest case of a hologram with two foci, the recording process requires that either two successive exposures be made, each with the same reference beam and a different single object beam, or that a single exposure be made with all three beams present. For reflective elements, a recording with all three beams present results in the superposition of two reflective gratings, each arising from interference of one object beam and the reference beam, and a single transmission grating, resulting from interference of the two object beams. If the geometry is not properly chosen, then the undesired transmission grating can rob energy from the desired reflection orders, reducing the diffraction efficiency or one or both of the reflection gratings. Detailed experimental results have quantified

these processes, and an analysis is under way. Also under analysis are the relative advantages of sequential vs simultaneous recording of the multiple elements. A paper on this subject has been submitted for oral presentation at the Annual meeting of the Optical Society of America in Washington DC in October 1985. A major effort will be continued towards developing a complete analytical understanding of beam coupling in multi-element reflective holographic gratings.

III. Optical Clocking of an Integrated Circuit Chip

The speed at which integrated circuit chips can be run is sometimes limited by the so-called "clock-skew" problem, in which the electronic clock that maintains synchronism of operations across the chip is subject to significant propagation delays across the chip, with the result that different parts of the chip can no longer be kept in synchronism. The clock lines on a chip must supply signals to virtually the entire chip, and for this reason they are very heavily loaded with device capacitances and very long in length. The capacitive loading limits the speed at which the clock can be run, and the long propagation delays lead to synchronism problems.

Our experience with optical clock distribution to date has been limited to the design and simulation of relatively simple circuits with integrated detectors suitable for extracting and amplifying an optically supplied clock. The simulations have yielded very encouraging results. We have found that substantially lower capacitances must be driven by the detectors and amplifiers than by a direct clock distribution line, due to the fact that the optical signal is easily fanned out for delivery at several separated sites on the chip. Thus the number

of devices driven by a single detector is far smaller than the number of devices driven by a conventional clock distribution line. As a consequence, simulations indicate that clock speeds can be increased by a factor of four if the optical distribution method is employed, rather than the conventional clock distribution method. In addition, the optical method requires distribution of electronic clock signals to a much smaller region of the chip, and therefore clock skew problems will be negligible.

The next step in the development of this work is the fabrication of some simple CMOS circuits and the experimental testing of the ideas. As our facility with simple circuits improves, more complex IC's will be fabricated.

The financial resources supplied by AFOSR under this grant are insufficient to fund much further work on this topic. In addition, we have been told that, because of the interest of other funding agencies in optical interconnections, AFOSR does not wish to support such work heavily. For this reason we expect the optical clock distribution work to be entirely funded by the Army Research Office starting sometime in the Summer of 1985.

IV. Image Inversion and Non-linear Filtering Using Photorefractive Crystals

For the past two years we have been pursuing the use of photorefractive crystals for image inversion, wavefront inversion, and non-linear filtering using coherent optics. This work has now yielded experimental results of major interest. First, during the past year we have successfully demonstrated the real-time inversion of intensity in images obtained by four-wave mixing. Such

inversion is obtained in a four-wave mixing experiment by imaging the object onto the crystal, and using a reference beam intensity that is weak by comparison with the object beam intensity. Under such conditions it can be shown that the local diffraction efficiency from the crystal is inversely proportional to the local object intensity. The physics of the process is more fully explained in Appendix B, which is a preprint of a paper that has been accepted for publication in *Applied Optics*, and is scheduled to appear in June. Also presented in that paper are photographs of a variety of images that have been inverted by this imaging process.

More interesting than image inversion are the non-linear filtering operations that can be performed when the photorefractive crystal lies in the Fourier domain, rather in the image domain. We have utilized this process to perform enhancement of defects in periodic structures, as explained in more detail in Appendix C, which is a preprint of a letter submitted to *Optics Letters*. A brief explanation of how defect enhancement takes place now follows. In the Fourier domain, the spectrum of the periodic structure consists of a series of bright and sharp Fourier components, corresponding to the harmonics of the periodic object. Also present in the Fourier domain are much broader and much weaker contributions of light corresponding to small defects in the periodic structure. The inten-

inversion process suppresses the strength of the bright spikes from the periodic components, but does not suppress the strength of the weaker light from the defects. Returning from the Fourier domain to an image plane, we find an image consisting primarily of the defects, with the periodicities missing. Experimental results are remarkably good, as shown in Appendix C. It should be noted

that this enhancement process takes place in real time. The mask is introduced at the input, and within some milliseconds an image of the defects appears. To date the smallest defects detected are 100 microns by 10 microns in size, but the optical system is now being substantially improved, and detection of much smaller defects is believed to be possible. The applications of this technique are many, ranging from the industrial inspection of meshes used in chemistry and biology, to the inspection of shadow grids used in color TV tubes, and possibly to the inspection of integrated circuit masks, provided the resolution of the technique can be brought to the sub-micron level. The potential commercial applications of these techniques appear very exciting. A patent application has been filed by Stanford University covering this invention.

V. Projects Completed in the Past Year

In the early portion of this grant year, we completed two projects that had been under way for several years. One project was concerned with the diagonalization and inversion of circulant matrices using coherent optics. This project was successfully concluded with the publication of a paper in *Applied Optics* on wavefront inversion using holography, and its application to matrix inversion. A reprint of this paper is attached as Appendix D, to which the reader is referred for further details.

A second completed project concerns the use of digital image processing techniques for the suppression of speckle in coherently formed images. A wide variety of techniques were investigated and compared quantitatively for the first time. The results of this work have yet to be published, but at least one paper

should be completed this summer.

Both of the above projects resulted in the granting of Ph.D. degrees during the past year.

VI. Miscellaneous Information

This section contains miscellaneous information regarding the personnel and publications associated with this AFOSR grant. The following individuals contributed to the research output of the grant:

1. Professor Joseph W. Goodman, Principal Investigator
2. Ms. Ellen Ochoa, Graduate Student Research Assistant
3. Mr. Raymond Kostuk, IBM Doctoral Fellow
4. Mr. Bradley Climer, Graduate Student Research Assistant

The publications supported in whole or in part by this grant during the last 12-month grant period are listed as follows:

Published Works

1. J.W. Goodman, F.J. Leonberger, S-Y. King and R.A. Athale, "Optical interconnections for VLSI systems", *Proc. I.E.E.E.*, Vol. 72, pp. 850-866 (1984).
2. Q. Cao and J.W. Goodman, Wave-front inversion using thin phase holograms: a computer simulation", *Applied Optics,* Vol. 23, pp. 4575-4587 (1984).

Under Submission

3. M. Nazarathy and J.W. Goodman, "Systolic lattice processing and ultrafast pulse shaping by optical coupled-wave device arrays", *Proc. S.P.I.E.*, Vol. 517.

4. E. Ochoa, L. Hesselink, J.W. Goodman, "Real-time intensity inversion using two-wave and four-wave mixing in photorefractive $Bi_{12}GeO_{20}$", Accepted for publication in *Applied Optics*.

5. R.K. Kostuk, J.W. Goodman, and L. Hesselink, "Optical imaging applied to microelectronic chip-to-chip interconnections", Accepted for publication in *Applied Optics*.

6. E. Ochoa, J.W. Goodman and L. Hesselink, "Real-time enhancement of defects in a periodic mask using photorefractive $Bi_{12}SiO_{20}$", *Submitted to Optics Letters*.

In addition to the above written descriptions of work, 3 oral presentations were on research results from this grant were presented at the 1984 Annual Meeting of the Optical Society of America, and one such presentation at the Optical Computing Conference of the OSA.

We are indebted to Prof. Lambertus Hesselink for the advice he has given us on both the work with photorefractive crystals and the work involving holographic optical elements.

APPENDIX I

Optical Imaging Applied to Microelectronic Chip-to-Chip Interconnections

Raymond K.Kostuk

Joseph W. Goodman

Lambertus Hesselink

Electrical Engineering Department

Stanford University

Stanford,CA.94305

An imaging system is proposed as an alternative to metalized connections between integrated circuits. Power requirements for metalized interconnects and electro-optic links are compared. A holographic optical element (*HOE*) is considered as the imaging device. Several experimental systems have been constructed which have visible LED's as the transmitters and PIN photodiodes as the receivers. Signals are evaluated at different source-detector separations. Multiple exposure holograms are used as a means of optical fan out allowing one source to simultaneously address several receiver locations. Limitations of this technique are also discussed.

1. Introduction:

A limitation of increasing importance in VLSI electronic integrated circuit design is the interconnections between devices and systems. Restrictions of conventional interconnects arise from a) increased space allocated to wiring, b) propagation delays with increased line lengths and RC time constants, c) inductive noise between lines, d) dominance of line capacitance over other sources of capacitance as line lengths increase, and e) degrading electromigration effects on wiring materials.[1,2,3] Since different optical signals can propagate through the same spatial volume without interference, the possibility of using optical methods to alleviate this space restriction is attractive.[4,5,6] In this paper we discuss a number of aspects of optical imaging which are applicable to the electronic interconnect problem and evaluate an experimental system.

2. Comparison of Optical and Electronic Interconnections

Figure 1 shows a typical VLSI microelectronic circuit mounted and bonded to a package which can be connected to other electronic systems. There are several thousand gates on this circuit and several hundred output pins which allow communication to other systems. Two levels of interconnection can be identified: One connects two or more devices on a common chip, and another connects an integrated system or chip to another chip.

There are a number of ways to compare the performance and capability of different types of interconnections.[1,3] Consider one such criterion, the reactive power required of one electronic inverter to trigger another inverter. Reactive

power is given by:

$$P = \frac{CV^2}{2\tau} \tag{1}$$

where C is the capacitance of the line and attached devices, V is the device threshold level (assumed 1 volt) , and τ is the clocking period (assumed 1 nsec).

Figure 2 illustrates gate-to-gate connection.[7] The gate capacitance of two devices and the metal line connecting them must be charged to the threshold potential for the gate. The gate capacitance is given by

$$C_g = \frac{\epsilon_r \epsilon_o A}{d} \tag{2}$$

where ϵ_r=3.9 for SiO_2 , ϵ_o=8.854 $\times$ $10^{-14} F/cm$, A is the device area,and d is the oxide thickness layer. Projected VLSI device lengths and oxide layer thickness are 0.5μm and 0.02μm, respectively. This gives a gate capacitance of

$$C_g = 50 fF/device.$$

The capacitance of the line joining two devices is

$$C_l = \epsilon_r \epsilon_o \frac{w}{h} l$$

where l is the line length and w the line width. The width/height ratio is restricted by fringing field effects to a minimum value of about 2. For a typical VLSI circuit the average length is approximately 1mm long. This gives a line capacitance of

$$C_l = 70 fF$$

The total capacitance of this link is then

$$C_t = 2C_g + C_l = 170fF$$

and the corresponding reactive power

$$P_g = 85\mu W$$

Figure 3 shows a chip to chip connection.[7] To minimize propagation delays, gate capacitances are gradually increased in size until the device capacitance is comparable to that of a bonding pad. A voltage pulse from a logic element must have sufficient power to charge these gates, two bonding pads, the line connecting them, and a receiving gate to the device threshold level. The total capacitance of this link is

$$C_t = 2C_b + C_l + 2C_g$$

where C_b is the bonding pad capacitance. For a pad area of approximately $100\mu m^2$, and assuming a SiO_2 dielectric, this capacitance is about 0.4pF. Lines connecting the pads are 25μm in diameter, and are assumed to be 500μm above the ground plane. When a number of chips are connected on the same substrate, a typical length separating a nearby pair is on the order of one centimeter. At this distance transmission line standing wave effects are not significant (i.e.λ = 30cm).

The line capacitance in this case is only 4.5fF. The total capacitance becomes

$$C_t = 0.8pF + 0.0045pF + 0.1pF = 0.9pF$$

and the switching power

$$P_c = 430\mu W.$$

Next consider a simple electro-optic link consisting of a semiconductor source and detector. Initially it is assumed that all of the light from the source is focused on the detector. The detector circuit model is shown in Figure 4. The current generated is a function of the physical parameters of the junction and the illumination ,[8]

$$i_p = \Phi \frac{q(1-r)}{h\upsilon}(1 - e^{-\alpha_0 z})$$

where i_p is the photocurrent,Φ is the optical flux, q is electronic charge, r is the Fresnel reflection coefficient of the detector surface, $h\upsilon$ is photon energy, α_0 the semiconductor absorption coefficient at λ, and z the absorption width.

Typical responsivity for a silicon device is 0.4 A/W.

The usual condition of low series and large shunt resistance simplifies the model to a capacitance shunting a current source. Current from the detector must charge the gate to its threshold level in a time less than the clocking period τ . If no preamplifier is assumed, all current must originate from electrons generated from the incident optical flux Φ . For a 2μm thick, 25μm square active area detector, the junction capacitance is

$$C_d = 32.5 fF$$

Since the detector must charge the capacitance of a gate, the total capacitance is

$$C_t = C_d + C_g = 82.5 fF$$

For a threshold voltage of approximately 1 volt

$$V = \frac{Q}{C_t}$$

$$Q = \int_0^{\tau} i dt \simeq i\tau$$

$$\tau = 1ns.$$

With 200 μW of incident optical power, 80μA of current

can be generated in the detector, and can produce 80 femtocoulombs of charge. This is sufficient to produce the 1 volt threshold value. Assuming a laser diode electrical to optical conversion efficiency of 30%, the electro-optic link will require about 670μW of electrical power.

These first order considerations indicate that with currently available electro-optic technology, the power required for an electro-optic link is of the same order of magnitude as that necessary for the electrical chip to chip interconnect, and would not suffer from the problems previously outlined for conventional interconnections. The electro-optic link compares less favorably with gate to gate connections on the same chip.

3. Optical Chip-to-Chip Layout

The chip to chip interconnect problem can be formulated in more specific terms as shown in Figure 5. One or more integrated systems are mounted on a common substrate separated by distances of about 1cm. As mentioned previously, at these lengths, and frequencies of 1GHz, transmission line effects are not significant. Bonding pads are assumed to be 100μm square, and separated by 100μm. Several hundred bonding pads must be connected. Each transmission point should be able to address several receiver locations, it is also desireable for channels to cross without interference.

An imaging system can provide this connection mechanism. Consider the

arrangement of Figure 6. A semiconductor emitter illuminates a holographic optical element, coded to distribute radiation to one or more image points. Photodiodes convert optical to electrical signals, which are then decoded by a digital electronic circuit.

Advantages of using holographic elements include their adaptability to decentered layouts by using off-axis recording geometries, and to fanout by using sequentially exposed multiple holograms.

A number of factors must be considered in a practical system of this type. The most attractive sources and detectors are those made from materials which are compatible with integrated electronics. Semiconductor sources developed for optical communications have emission wavelengths from 780nm to 1.6μm. To date only a few holographic recording materials are responsive at these wavelengths, and these are not very sensitive.[9]

Other considerations are the emission profile and polarization characteristics of the source. Laser diodes have an emission profile corresponding to the diffraction pattern of the junction geometry. Planar stripe junction diodes have transverse mode divergence angles which have typical values of 60° by 10°. Therefore only a portion of the volume above the source will be illuminated. The hologram need only occupy this region above the source to be effective.

The polarizations of these two directions are orthogonal. Kogelnik [10] has shown that polarization vectors oriented in the plane of incidence of the grating produces a reduced coupling constant and diffraction efficiency which results in lower image intensity.

An LED is also a potential semiconductor source. It has the advantage of being a surface emitter and is much easier to fabricate than a laser diode. In addition they can be made to emit in the visible by introducing traps in the band gap. However, they are inefficient in comparison to laser diodes and have spectral bandwidths of about 20nm. Also they emit unpolarized light which results in lower diffraction efficiency for the reason mentioned above. Their intensity emission profile is cosinusoidal in angle, and therefore illuminates a larger region of a hologram than would a laser diode. Image reconstructions with this type of emission profile are brighter when the hologram occupies large solid angles relative to the source.

4. HOE Characteristics

The requirement for a compact system implies that the element must have a small f-number. This also improves flux collection. The meridional angles for an f/1 and f/3.5 element are 26.5° and 8.1° in air. A model for diffraction efficiency must be valid for grating vectors covering this angular range. A relatively simple description of grating diffraction efficiency is Kogelnik's coupled two-wave treatment.[10] The expression of effeciency for reflection holograms with absorption is given by

$$\eta = \left\{ \xi/\upsilon + \left(1 + \frac{\xi^2}{\upsilon^2} \right)^{1/2} \coth(\upsilon^2 + \xi^2)^{1/2} \right\}^{-1}$$

where η is the diffraction efficiency,

$$\upsilon = j\pi n_1 \frac{d}{\lambda (c_r\, c_s)^{1/2}} \ ,$$

$$\xi = 1/2\ D_0 (\ 1 - c_r / c_s\ ) \ ,$$

$$D_0 = \frac{\alpha d}{cos\ \theta_0} \ ,$$

n_1 is the refractive index modulation, d the grating thickness, c_s and c_r are obliquity factors, α is the absorption/ length, and θ_0 is the Bragg angle. The planar grating treatment can be extended to curved surfaces by assuming that the surface is locally plane in the region where the ray intersects the grating.[11]

A number of planar volume phase holograms were formed in bleached photographic photographic film. The thickness, refractive index, and post-bleached absorption were measured to obtain average values for these parameters. The results were then used in Kogelnik's model to predict the diffraction efficiency curves, and were compared to measured curves. Although slight changes to the values for absorption and emulsion thickness change had to be used, the agreement was very good. Figure 7 shows two measured diffraction efficiency curves from gratings with $\vec{K}$ orientations approximately equal to the meridional angles of f/1 and f/3.5 systems. High diffraction efficiency is maintained over a large range of playback angles. The HOE field of view is essentially this angular range and is about 30° for 25° grating slant angles, and 60° for 10° slant angles corresponding to the f/3.5 system.

A single grating element can interconnect a number of sources and their conjugate receiver locations over the angular range of high efficiency. When source

reconstruction coordinates differ significantly from formation positions, hologram image aberrations reduce image irradiance. Aberrations can be evaluated with ray tracing techniques. For thick holograms these expressions may be derived from the reflected ray components which are perpendicular and tangent to the grating vector

$$\bar{r} = (\vec{K}\cdot\bar{r})\,\vec{K} - \vec{K} \times (\vec{K} \times \bar{r})$$

where $\bar{r}$ is a unit vector along the reconstruction ray, and $\vec{K}$ is the grating vector given by

$$\vec{K} = \bar{r}_o - \bar{r}_c$$

and $\bar{r}_o$ and $\bar{r}_c$ are unit vectors along the object and reference ray directions respectively. The reconstructed or image ray is

$$\bar{r}' = -(\vec{K}\cdot\bar{r})\vec{K} - \vec{K} \times (\vec{K} \times \bar{r})$$

The spot diagram generated by ray tracing should be adjusted for the variation in efficiency at different locations in the aperture of the volume HOE. However it has been shown that a close relationship exists between the observed image field and the density of rays traced through the element.[12] Figure 8 shows the spot diagram of rays from a source point displaced 0.5 cm perpendicular to the axis, and 0.1 cm along the axis from the source formation positions for an f/1 and f/3.5 element. It is clear that off axis imaging degrades much more rapidly for smaller f/# HOE's.

A computer program coding the grating equation can be used to generate a spot diagram at any desired image plane. When used in conjunction with Kogelnik's efficiency model, both the aberrations and the efficiency of the rays

forming the image can be determined. This gives a better indication of the distribution of flux at the receiver location and the detector current produced from a source of given size, output power, and location relative to the HOE. Such a program is currently under development in our lab for use with multiple image reflection hologram design.

The effective HOE aperture and reflection losses also restrict the useable source power. The solid angle subtended by the HOE relative to a source point is

$$\Omega = (\pi \frac{D^2}{4}) \frac{cos\,\theta}{r^2}$$

where D is the diameter of the hologram aperture, θ is the angle from the source point to the center of the hologram, and r is the distance from the source point to the hologram center.

If the source is a Lambertian emitter, the flux collected by the aperture of the HOE is

$$\Phi = (I_o \, cos\,\theta)\Omega$$

When the source and optical element are on axis, 12.5% of the available source power is collected with an f/1 system,and only about 1.0% for an f/3.5 system.

If the hologram recording medium is not index matched to the source and detector surfaces, Fresnel reflection losses also reduce the flux entering the grating. The recording medium used has a refractive index of 1.64, resulting in transmitted intensities ranging from 91 to 93 % for incident angles of 0° to 30°. Therefore 7 to 9% of the available source power is lost by reflection. If a fixed amount of flux Φ_{dt}, is required at the detector then axially located sources must

have output powers Φ_{st} exceeding this value by

$$\Phi_{st} = \mu \, \Phi_{dt}$$

with

$\mu = 1/(.125*.08) = 9$; for an f/1 system, and

$\mu = 1/(.01*.08) = 109$; for an f/3.5 system.

Therefore considerable power is required from a Lambertian source even when a 100% efficient hologram is used.

The divergence angle from a laser diode is approximately matched to the meridional angle of an f/1 element (about 30° for the laser and 26° for the optical element). This implies that all of the power from a laser diode can be collected by a smaller aperture than for a Lambertian source. A laser diode can therefore have much lower input power and still produce the required detector current and perform the switching task.

After considering the above HOE and semiconductor source characteristics, three types of hologram configurations appear to offer a solution. The first arrangement is a large aperture reflecting lens with one or multiple gratings (Figure 9). This element is relatively easy to fabricate and position, and uses a point source for reconstruction. Multiple grating formation allows a single reconstruction source to address several locations simultaneously. It does however restrict the locations of sources and detectors to positions along diameters which pass through the optical axis, and fan out can only be accomplished in an invariant pattern. This restriction may preclude this arrangement from practical

application, but it is important for optical system evaluation. The second and third configurations utilize the multifacet or aperture partioning concept recently discussed by Case[13] for transmission holograms, and require directed beam reconstruction either from a laser diode or from a directed LED emission pattern. In one of these arrangements a mask with the address pattern serves as the object wave and a converging beam as the reference wave (Fig 10). This method has the attractive aspect of having an IC compatible technique (i.e. mask making) used for generating an address pattern. The draw back of this arrangement are the intermodulation terms which limit the efficiency of the reconstruction images.[14] It is not obvious where this becomes restrictive for this application. In the last hologram configuration proposed each facet is illuminated sequentially with a number of diverging object beams and a converging reference wave (Fig 11). The positions of the object beams can be moved automatically with a computer controlled stepper motor drive and beam ratios can be adjusted for maximum diffraction efficiency. This configuration appears to offer the most flexible arrangement for fabricating an interconnect pattern since it satisfies requirements for both large number of independent channels and spatially variant fan out. The difficulty with this HOE fabrication technique is the mechanical complexity of the mount; however there appears no fundamental restriction to its implementation.

5. Experimental Results

To evaluate some of the above ideas a number of experimental systems were fabricated and tested. Only the first hologram design described above is discussed

here. The other two hologram types will be presented in future papers.

The effects of image degradation and power loss were determined by mounting a number of sources and detectors at increasing separations, and measuring the received detector photocurrent. The operating characteristics of the sources and detectors are given in Table 1. The sources are surface emitting GaP LEDs. The primary reason for using these devices is their peak emision in the visible (635 and 655nm) making them compatible with a number of available holographic recording materials. They have about a 20nm spectral bandwidth and a near Lambertian intensity emission profile. Their main disadvantage is their poor electrical to optical conversion efficiency. Measured efficiency of both the 635nm and the 655nm LED's is about 0.5%. Sources and detectors used were in chip form with cross sectional dimensions of the same order of magnitude as the size of the bonding pads (see Figure 12).

Two source-detector mounts were used. On the first, the devices were set on the common conducting plane of a dual in-line IC package. This arrangement allowed evaluation of both electrical coupling and direct optical scattering on the detector signal received from the source image. The second mount had source and detector on different substrates and was optically isolated to allow examination of the effects of image degradation and aperturing at large source-detector separations.

Figure 13 is a plot of the ratio of photodiode current with the image of the source focused onto the detector to the current with the image focused just off the detector. Response with source-detector separations from 86μm to [illegible]nm were

obtained with the source and detector mounted on the same conducting substrate. It appears that optical scattering and electrical coupling greatly reduce the effective signal response at separations less than a 100μm. At separations from 2 - 4 mm, contrast ratios increase more slowly than at closer separations. With source and detectors on separate substrates and isolation from optical scatter, the contrast ratio improves by an order of magnitude at 1.0cm distances, then falls by a factor of two as separation increases to 2cm.

The image of the source was also observed on a CCD line scanner to directly evaluate the image irradiance pattern. Figure 14 shows these profiles when the 635nm LED is 0.45cm , 0.60cm ,1.00cm ,and 1.50cm from the line scanner. The hologram used for these measurements has a diameter of 1.5cm and is 3.10cm from the source detector plane. The CCD scans indicate that the fall off in effective signal response at large separations results from an increase in the image area and a corresponding decrease in image irradiance illuminating the detector.

A number of multiple exposure holograms were made to examine the potential of optical fanout. Elements were made with the arrangement shown in Figure 15. A converging and diverging wavefront overlap to form an on-axis reflecting lens type hologram. The film plane is then translated in this overlap region to form a number of holographic lenses with their optical axes displaced by the amount of translation. A single reconstruction source has a different displacement from the optical axis of each encoded element, and therefore images the source at a different position in space. Figures 16a,b show images produced from two such elements. In the first, film translations of 0.7cm by 0.25cm were used, while in

the second 0.5mm movements were made. Both situations give well resolved images with full width at half intensity maxima (FWHM) of about 300μm. The LED emission surface is 150μm in length.

6. Conclusions and Future Reasearch

Reactive power considerations indicate that with current electro-optic technology an optical chip-to-chip interconnect requires approximately the same amount of power to transmit high speed signals as electrical connections, but without the need to devote large sections of the circuit substrate to communication channels. This would allow the use of more input-output ports and increase the information capacity of the IC. It could also reduce electrical coupling difficulties of conventional interconnect schemes. The chip-to-chip interconnect can be recast in terms of an optical imaging system with semiconductor sources as signal transmitters and photodiode detectors as receivers.

The diffraction efficiency characteristics of reflection volume holograms have sufficient angular response to accomodate source detector separations of a few centimeters. These separations also require that the holographic element be located a comparable distance above the circuit substrate. Other practical considerations are Fresnel reflection losses, and flux collection characteristics of a particular f/# element and source emission profile. Serious limitations also exist in the lack of compatibility between efficient semiconductor sources nd holo-

graphic recording materials. A match between these components would allow use of much more efficient sources and greatly improved flux collection geometries.

Initial experiments indicate that electrical and optical coupling are serious problems when sources and detectors are less than 100μm apart, and that image blurring causes the fall off in detector irradiance at separations of a few centimeters and greater.

Experiments also indicate that sequentially exposed holograms have sufficient resolution to address a number of receivers spaced from several hundred micrometers to centimeters. This could be used to implement a number of very flexible interconnect patterns, without the drawbacks of conventional electrical systems.

Acknowledgements:

This work was supported by the Air Force Office of Scientific Research. One of us (R.K.K) would especially like to thank IBM for fellowship support during this period.

References:

1. R.W.Keyes, Proc.I.E.E.E., Vol.69, No.2, (1981).

2. A.J.Rainal, A.T.&T. Bell Lab. Tech. J., Vol.63, No.1, (1984).

3. P.M.Solomon, Proc.I.E.E.E., Vol.70, No.5, (1982).

4. J.W.Goodman, F.J.Leonberger, S.Y.Kung, R.A.Athale, Proc.I.E.E.E., Vol.72, No.7, (1984).

5. B.K.Jenkins and T.C.Strand, Proc. Int. Conf. on Computer Generated Holography, SPIE Proc., Vol.437, pp. 110-118, (1983).

6. A.A.Sawchuk and T.C.Strand, Proc.I.E.E.E., Vol.72, No.7, (1984).

7. C.Mead and L.Conway, *Introduction to VLSI Systems,* Addison-Wesley,Menlo Park,CA.,1980.

8. S.E.Miller and A.G.Chynoweth,*Optical Fiber Telecommunications,* Academic Press,New York,1979.

9. H.M.Smith, ed., *Holographic Recording Materials,* Springer-Verlag, Berlin Heidelberg, 1977.

10. H.Kogelnik,Bell Syst. Tech.J.Vol.48,2909 (1969).

11. R.R.A.Syms,L.Solymar,Opt.Quan.Elect.Vol.13,415(1981).

12. R.Ferrante,M.P.Owen,L.Solymar,J.Opt.Soc.Am.,Vol.71,No.11,(1981).

13. P.R.Hagen, H.Bartelt, S.K.Case, Appl.Opt., Vol.22, No.18, (1983).

14. J.Upatnieks and C.Leonard, JOSA, Vol.60, No.3, (1970).

FIGURE 1. VLSI circuit (manufactured by Honeywell) with approximately 3000 gates and 150 bonding pads. Interconnections exist between gates on a common substrate and from bonding pads to other circuits and outside systems.

FIGURE 2. Schematic of gate to gate connection for two inverters. The line between gates is modeled as a single capacitor.

FIGURE 3. Schematic of chip to chip connection. Inverters have gates with increasing capacitance to minimize signal delay. Lines are assumed short enough so as not to be influenced by transmission line effects.

FIGURE 4. Detector circuit model.The space charge region of the junction results in a capacitance shunting a photon induced current source.The series resistance is typically a few ohms and can be neglected.The parallel resistance is on the order of 10^9 ohms and can be assumed to be an open circuit.

FIGURE 5. Geometrical layout of a chip to chip connection. Two integrated circuits are mounted on a common substrate with L = 1 to 5cm, w = 1cm, and bonding pad widths and separations = 100μm.

FIGURE 6. Imaging system for chip to chip communication. Light emitting sources and detectors replace transmitting and receiving bonding pads. A

hologram is used as the imaging element. Design must include f/# or D/l ratio, intensity emission profile of the source, and source-detector separation.

FIGURE 7. Measured diffraction efficiency curves for gratings with $\vec{K}$'s approximating those formed by the meridional rays in : (o) an f/1 system,i.e 25° ; and (x) an f/3.5 system,i.e.10°. Significant efficiency exists over anangular range of 30° to 60°.

FIGURE 8. Spot diagrams for f/1 (a), and f/3.5 (b) systems with a reconstruction source point 0.5cm from the axis of the element at x=0,y=0. Computations are based on the grating vector equation.(+) optical axis ; (o) source point.

FIGURE 9. Simplest holographic configuration for imaging interconnects. Light from a point source is imaged to a diametrically opposite point. Several sequential exposures can be encoded and used to produce an invariant pattern of images. This can be used for invariant fan-out configurations.

FIGURE 10. Combined multi-facet hologram and variable image mask. A separate hologram facet is formed with each fan-out pattern encoded on the mask. The mask and hologram are translated with respect to each other. Each hologram is formed with a converging reference wave to allow play back with an expanding beam.

FIGURE 11. Multi-facet hologram formed with selective object source points. Source points are encoded in sequential fashion. This is the most flexible configuration, but also the most difficult to implement.

FIGURE 12. Photograph of Littronix LED with $250\mu m^2$ emission area, and a Hewlett-Packard photodiode from an electro-optic coupler with $400\mu m^2$ active area. The separation of the two chips is about 60μm.

FIGURE 13. Plot of the ratio of photodiode current with image focused on the detector to the current with the image focused off the detector. The equipment used did not allow measurements with source detector separations from 4mm to 10mm. (X) indicates measurements obtained with sources and detectors on the same substrate ; (0) on separate substrates.

FIGURE 14. CCD line scan traces of images of the 635nm LED produced with the f/1.9 HOE. The CCD has 256, 13μm elements. Oscilloscope scale is 330μm per 1cm. Source-CCD separations are a)0.45cm ; b)0.60cm ; c)1.00cm ; and d)1.50cm.

FIGURE 15. Schematic of hologram construction arrangement to form a multiple image with a single reconstruction source. Film plane is translated through fixed construction beams. The resulting element (b), is in effect a set of reflecting lenses

with displaced optical axes which image the source relative to their respective axes. The lenses in (b) are shown unfolded for clarity.

FIGURE 16. a)Photograph of multiple images formed with an element having 0.25cm horizontal and 0.70cm vertical displacements using an LED reconstruction source. The diode is 1cm from the center of the image pattern. b) Photograph of a CCD line trace of the LED imaged by an HOE with three 500μm translations. Scale is 330μm per 1cm.

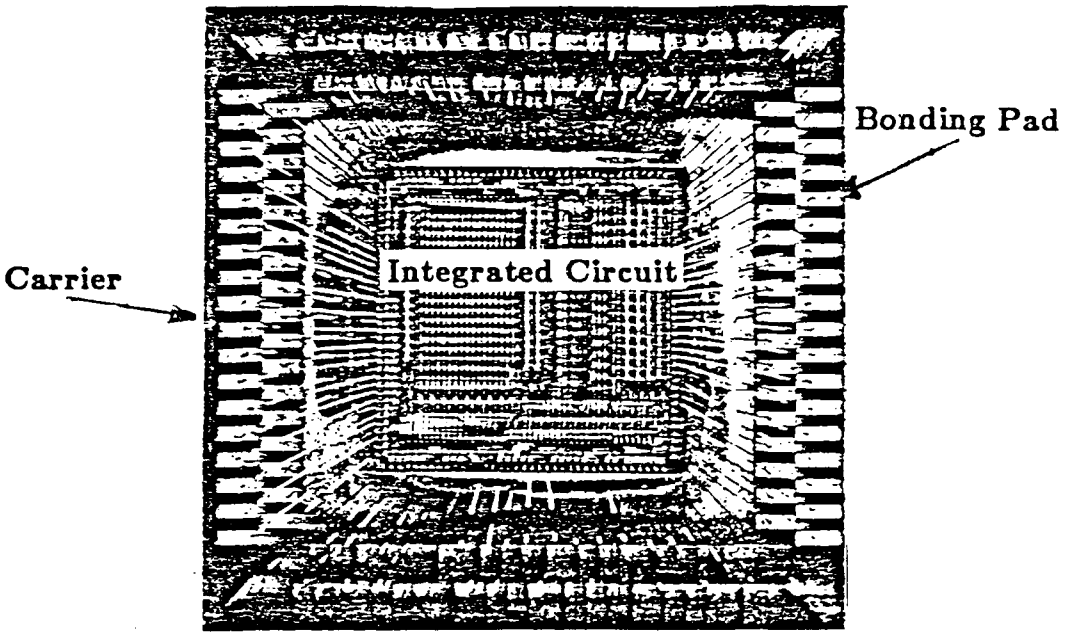


FIGURE 1

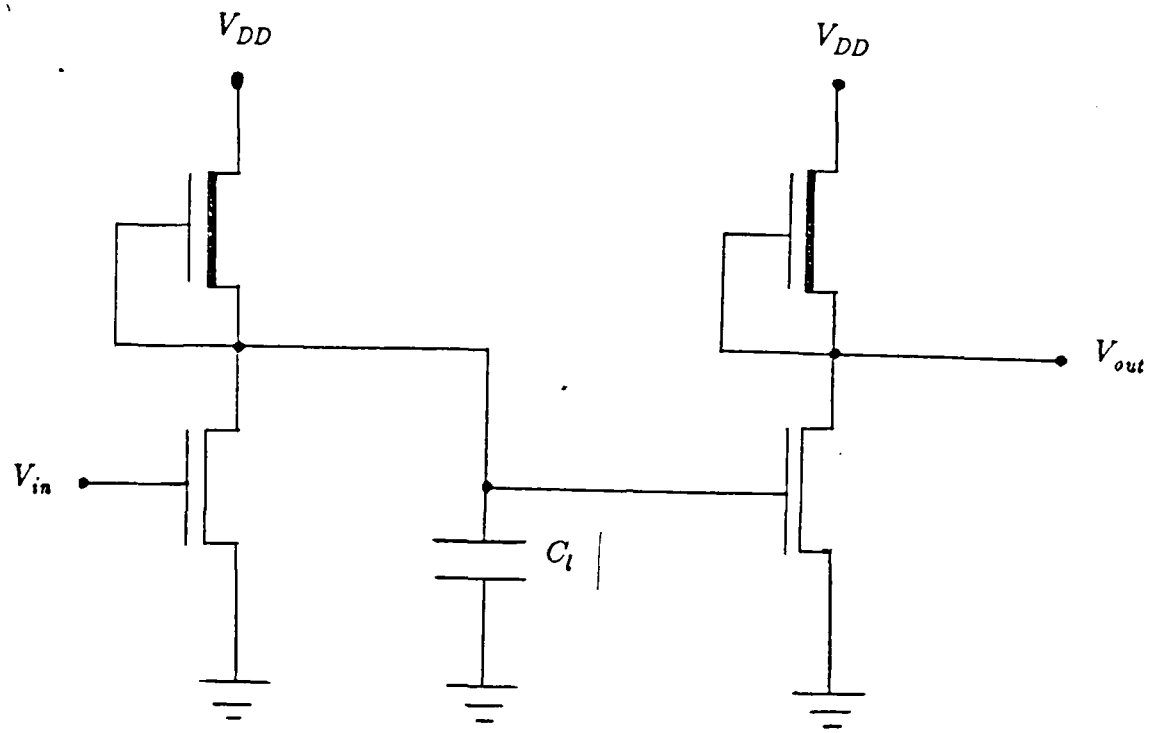


FIGURE 2

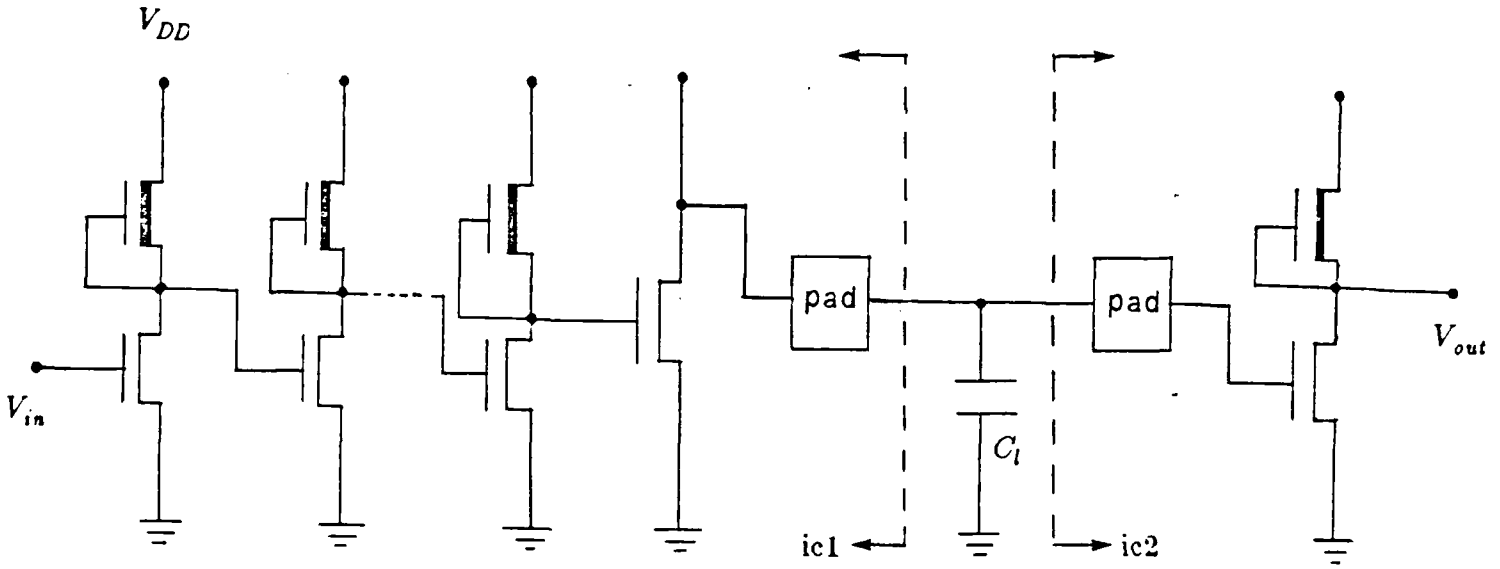


FIGURE 3

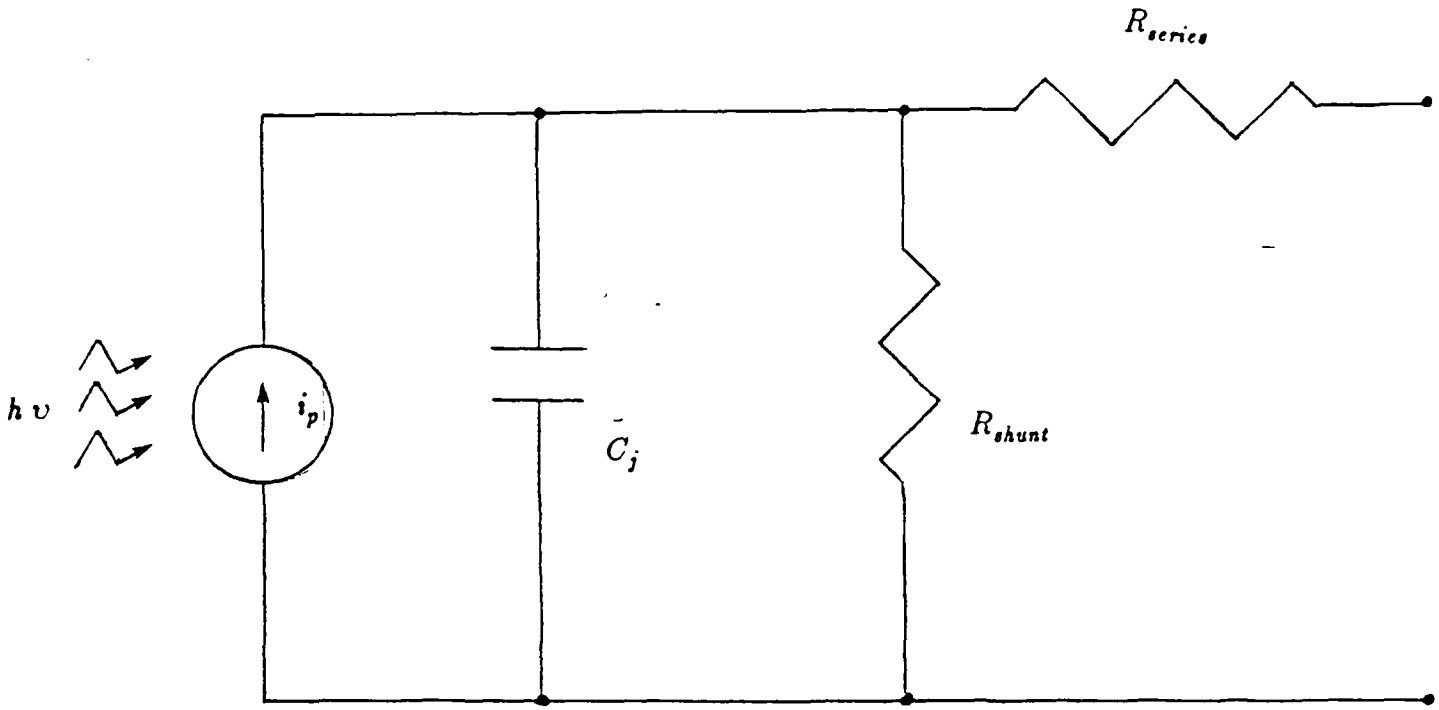


FIGURE 4

pattern. This restriction may preclude this arrangement from practical

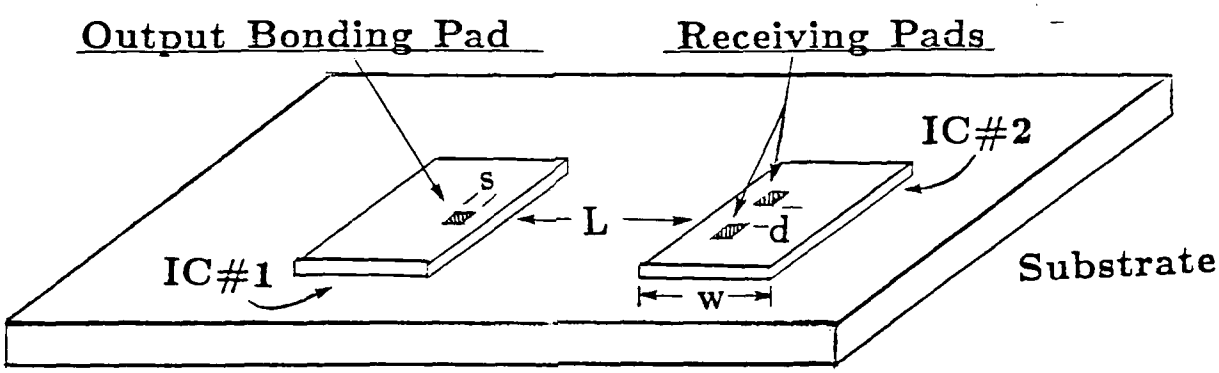


FIGURE 5

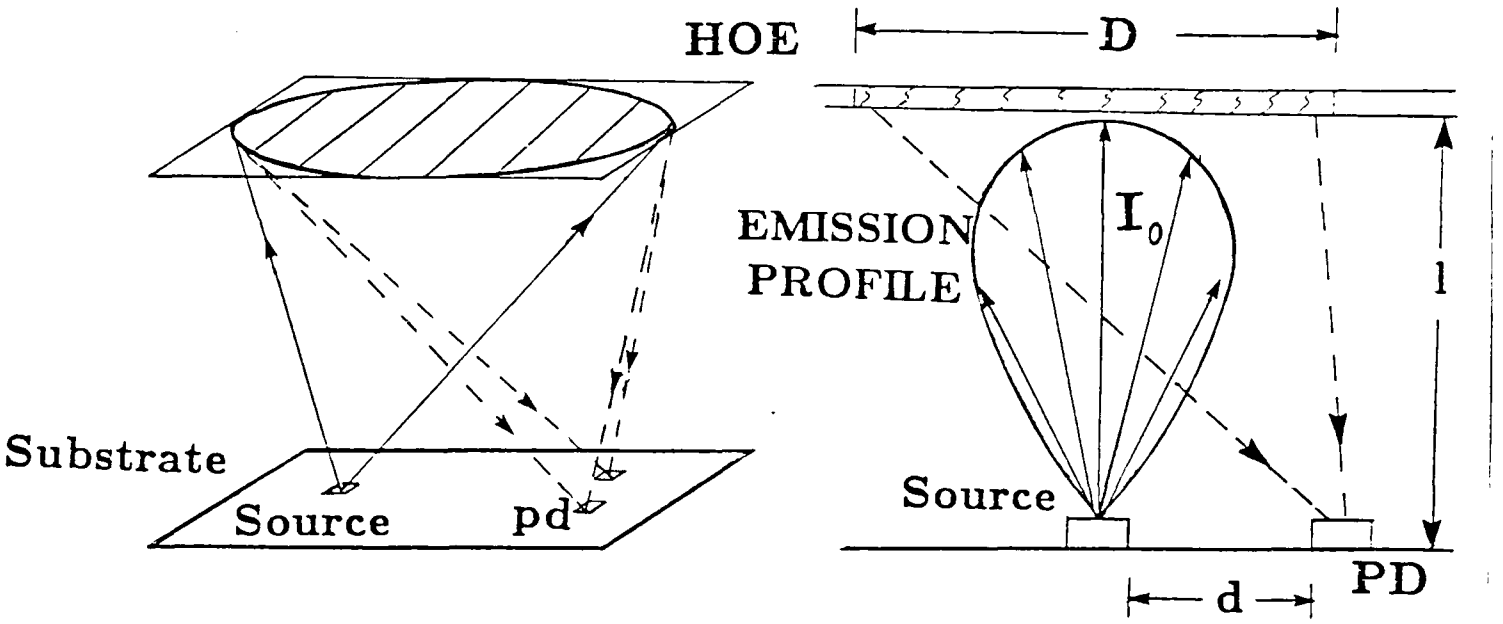
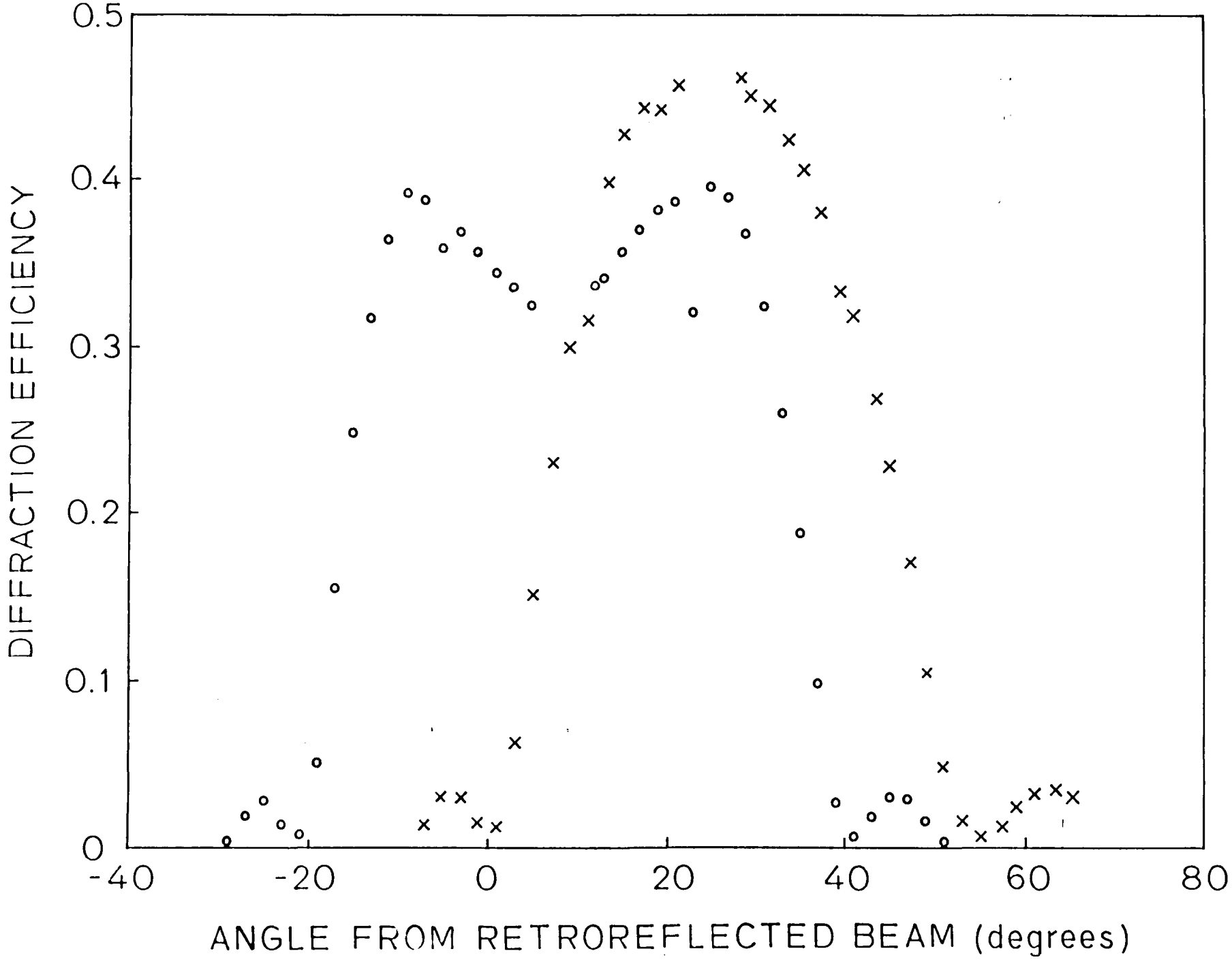
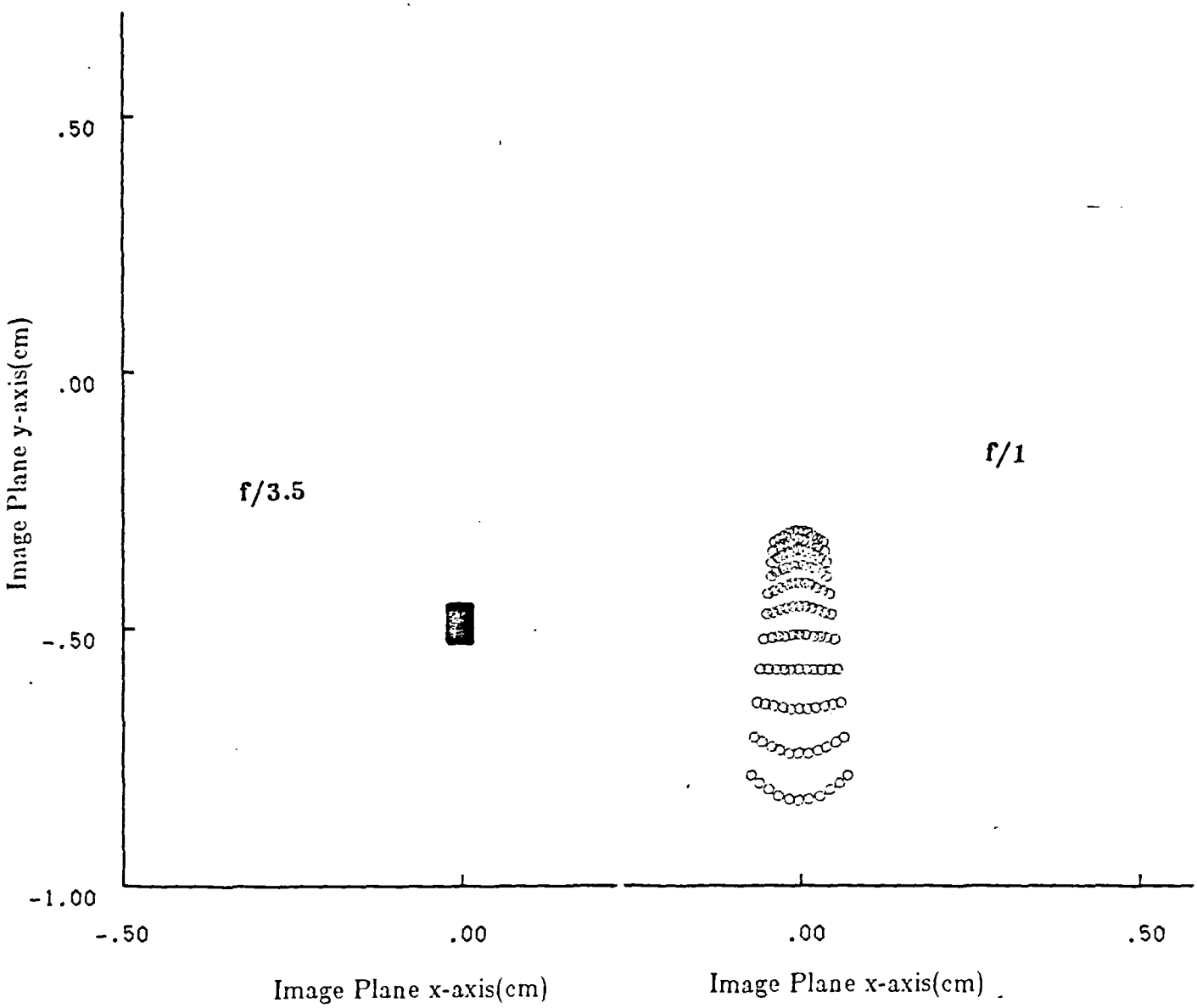


FIGURE 6


0.5
0.4
0.3
0.2
0.1
0
DIFFRACTION EFFICIENCY
-40
-20
0
20
40
60
80
ANGLE FROM RETROREFLECTED BEAM (degrees)


.50
.00
-.50
-1.00
Image Plane y-axis(cm)
f/3.5
f/1
-.50
.00
.00
.50
Image Plane x-axis(cm)
Image Plane x-axis(cm)

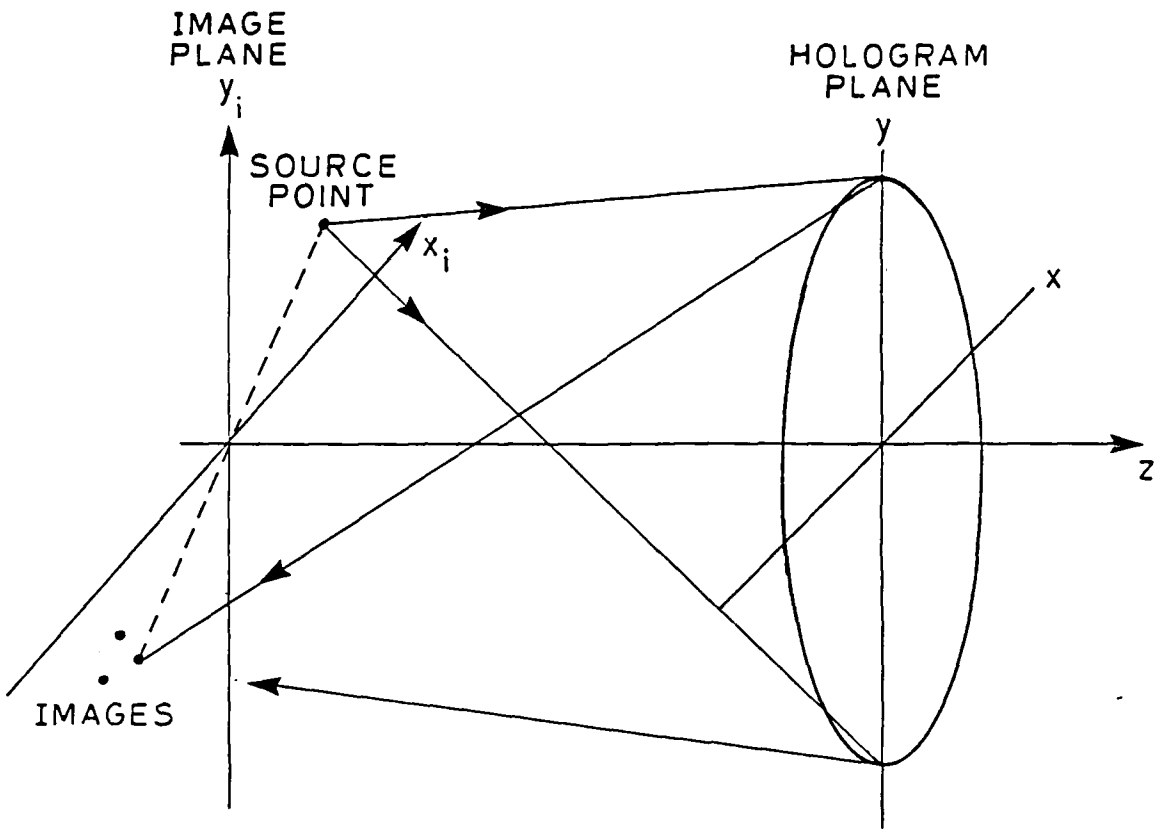
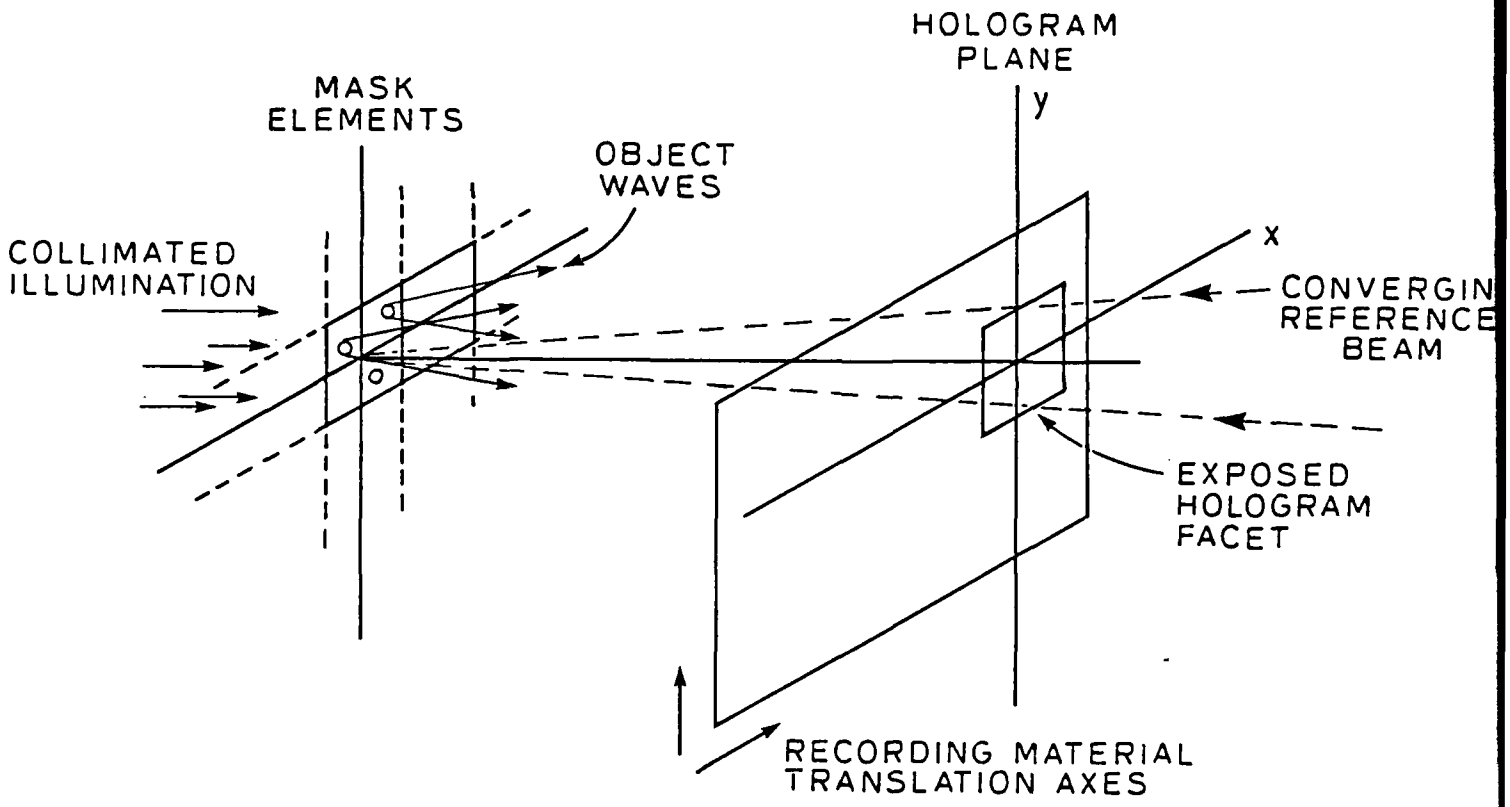
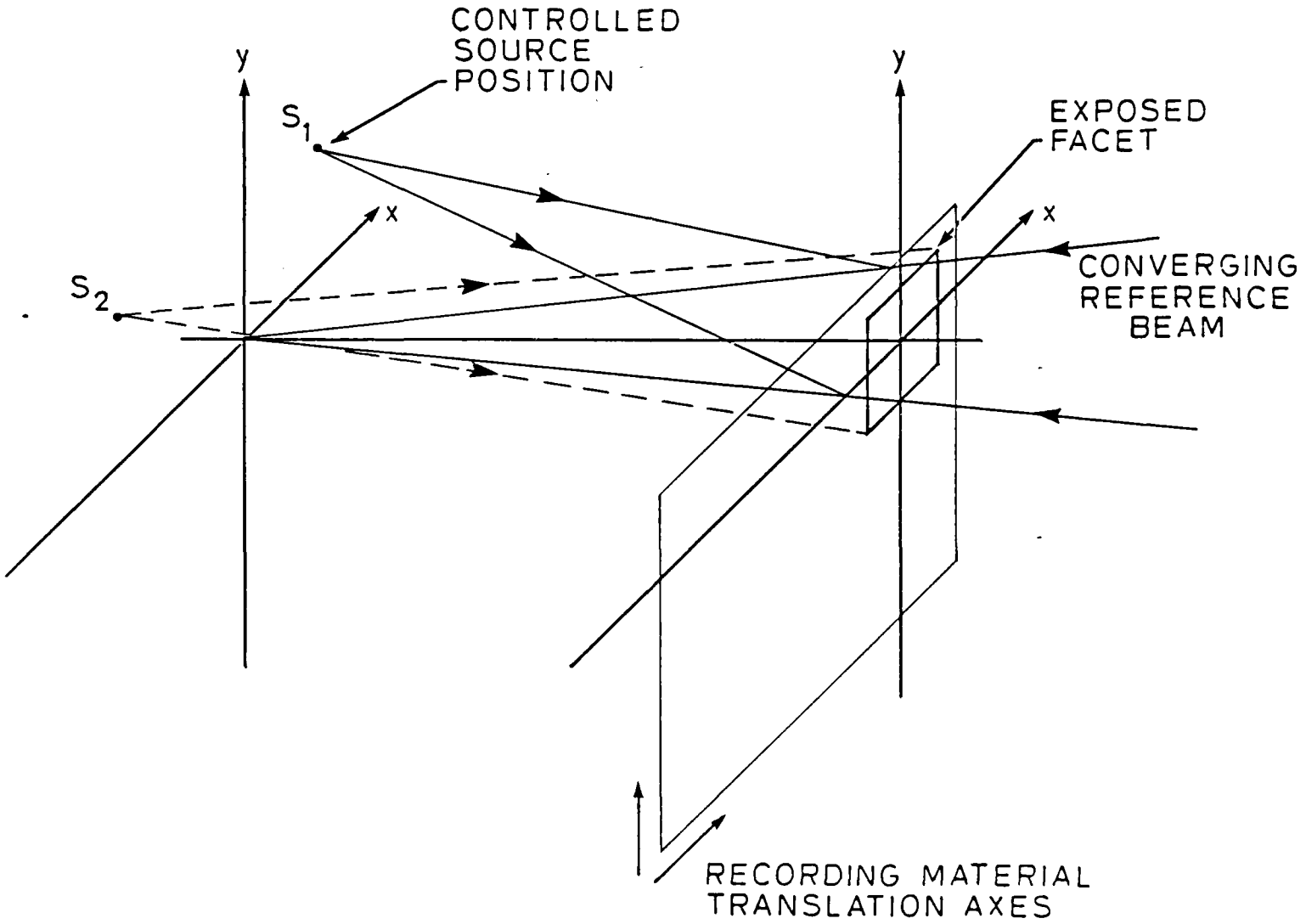
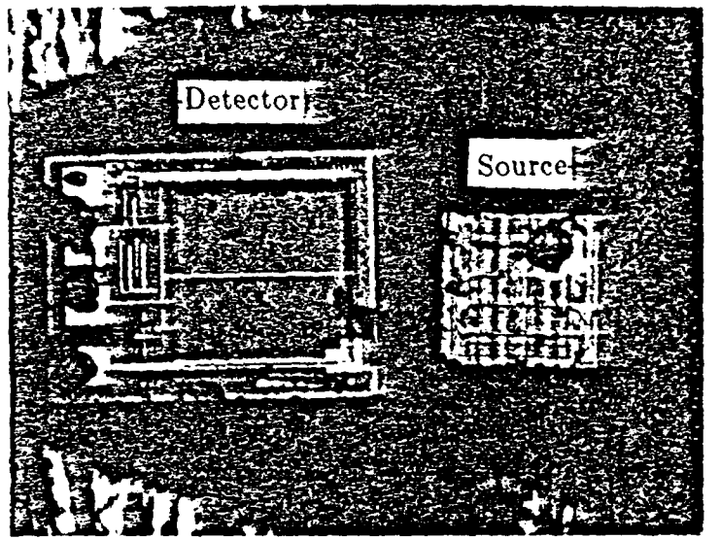
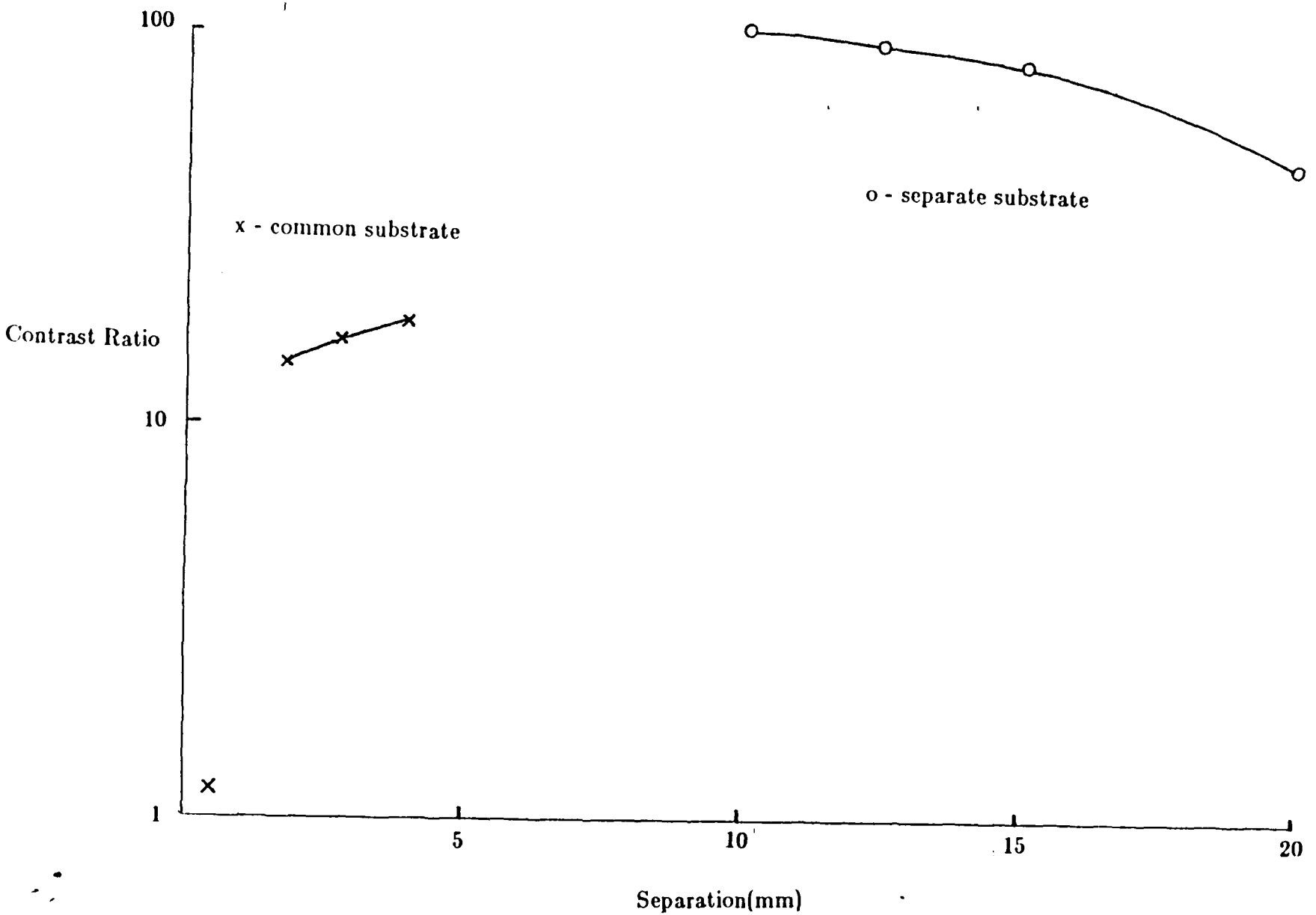


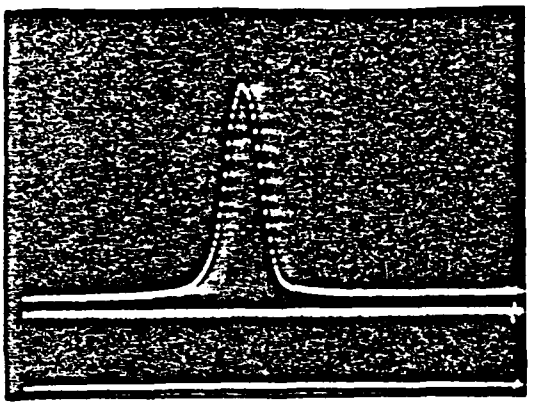
IMAGE
PLANE
y_i
SOURCE
POINT
x_i
HOLOGRAM
PLANE
y
x
z
IMAGES


HOLOGRAM
PLANE
y
MASK
ELEMENTS
OBJECT
WAVES
COLLIMATED
ILLUMINATION
x
CONVERGIN
REFERENCE
BEAM
EXPOSED
HOLOGRAM
FACET
RECORDING MATERIAL
TRANSLATION AXES

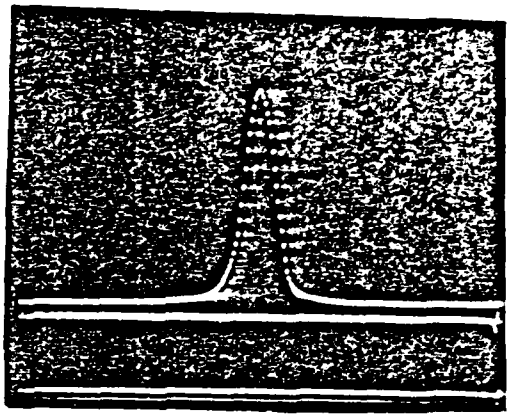

CONTROLLED
SOURCE
POSITION
y
S_1
x
S_2
y
EXPOSED
FACET
x
CONVERGING
REFERENCE
BEAM
RECORDING MATERIAL
TRANSLATION AXES


Detector
Source

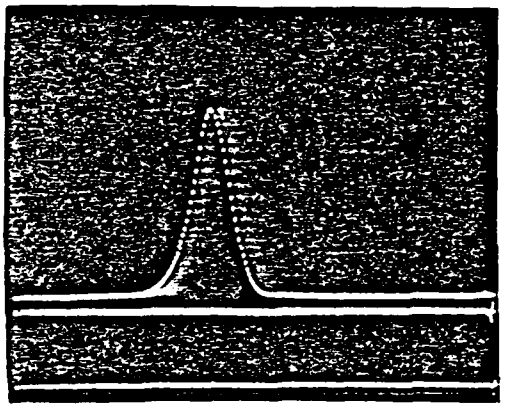

100
10
1
Contrast Ratio
x - common substrate
o - separate substrate
5
10
15
20
Separation(mm)



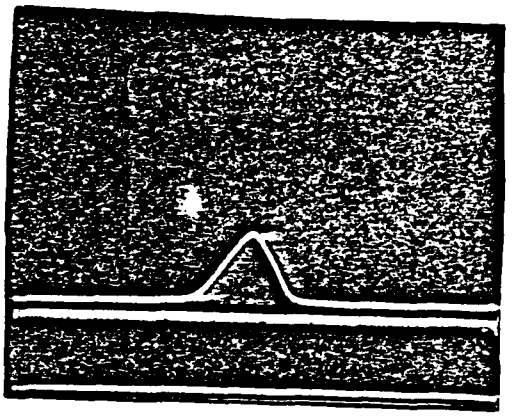
(a)



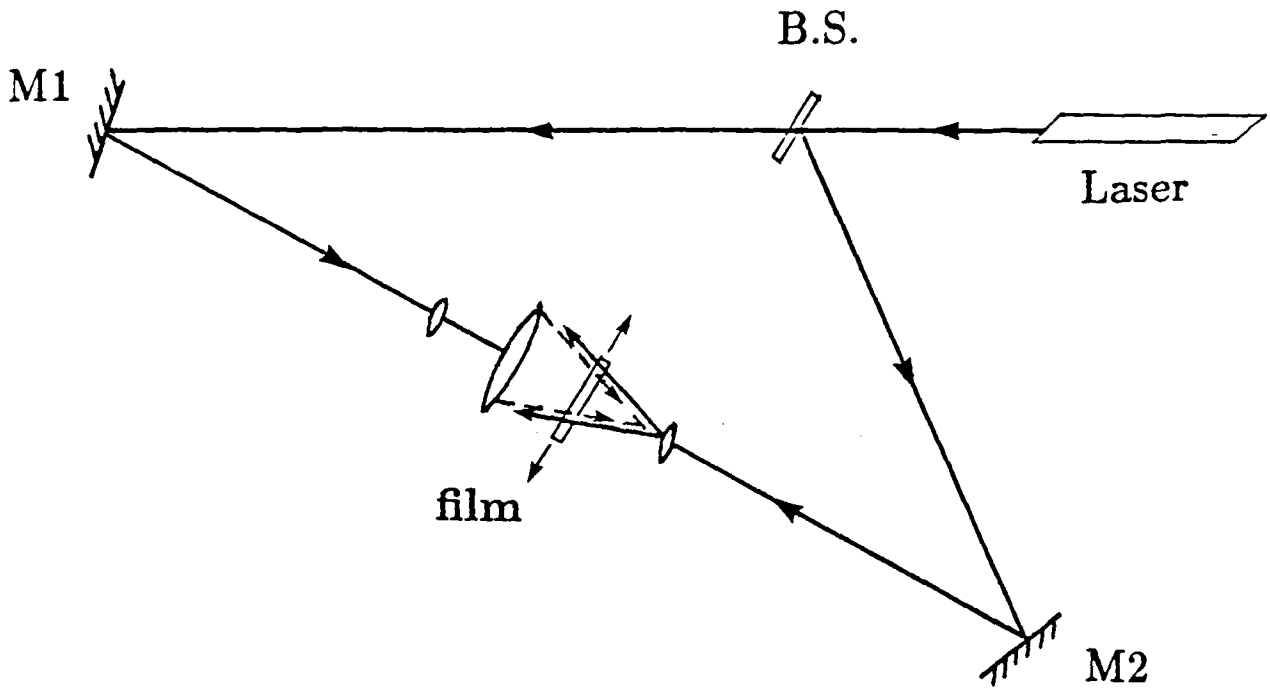
(b)

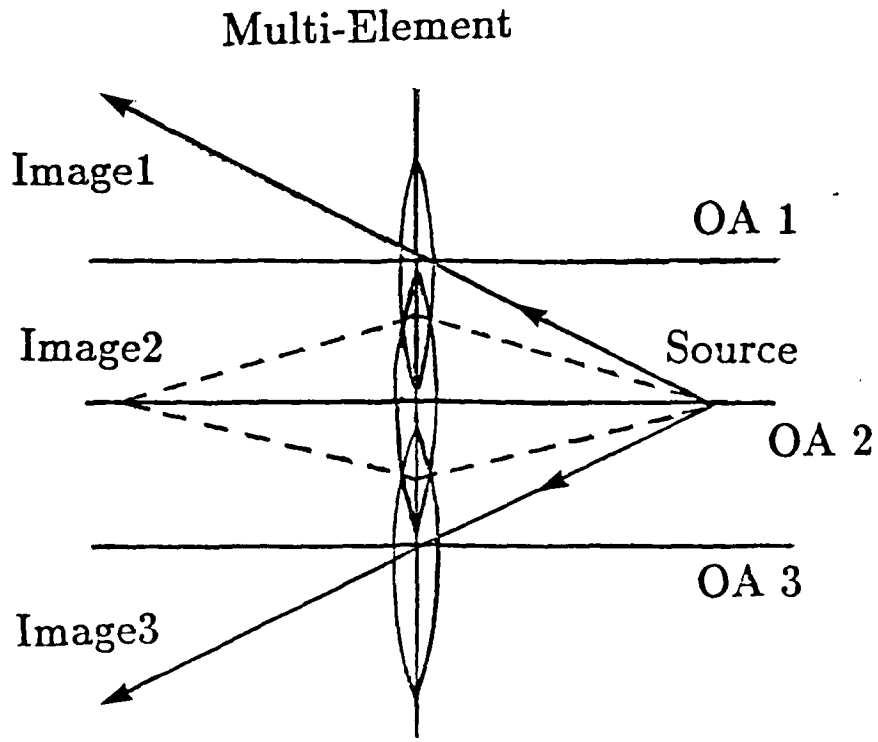


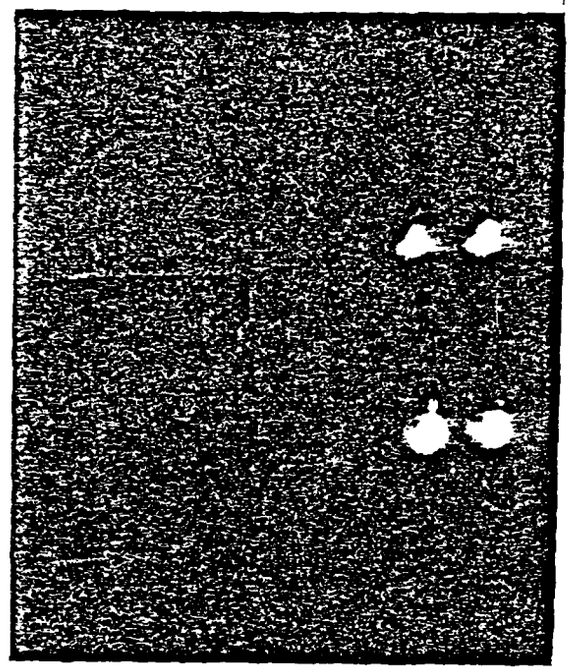
(c)



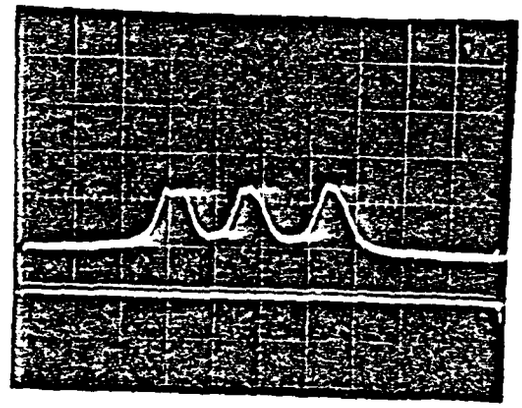
(d)


B.S.
M1
Laser
film
M2


Multi-Element
Image1
OA 1
Image2
Source
OA 2
OA 3
Image3



(a)



(b)

TABLE 1: SOURCE and DETECTOR CHARACTERISTICS

Sources	Littronix	Hewlett-Packard(HP)
Input Power	155mW	140mW
Output Power	70μW/str	80μW/str
Intensity Profile	Lambertian	Lambertian
Size	250μm^2	150μm^2
λ_{peak}	660nm	635nm
$\Delta\lambda$	20nm	20nm

Detectors	HP EO Coupler PD	HP4205 PIN
Size	400μm^2	200μm(diameter)
Responsivity(630nm)	0.1A/W	0.4A/W

Real-time Intensity Inversion using Two-Wave and Four-Wave Mixing in Photorefractive $Bi_{12}GeO_{20}$

Ellen Ochoa, Lambertus Hesselink, and Joseph W. Goodman

Stanford University
Department of Electrical Engineering
Informations Systems Laboratory
Stanford, CA 94305

I. Introduction

Photorefractive crystals show considerable promise as suitable materials for real-time optical processing. Their ability to perform phase conjugation in a degenerate four-wave mixing configuration [1] prompted a flurry of investigation into their properties. In addition, several types of optical processing using photorefractive crystals have been demonstrated, including convolution and correlation [2], interferometry [3], beam amplification [4], image subtraction [5], edge enhancement [6, 7], and image division [8].

The last two operations took advantage of an approximate dependence of diffraction efficiency on the modulation depth of the grating formed in the crystal. The authors of those papers noted that only the beam ratio of the two writing beams and not the total beam intensity determined the output. For the proper beam ratios, this results in an intensity inversion; that is, areas of an object beam which are originally less intense than other sections become more intense in the output. However, this dependence was not investigated in any depth nor was the effect of other experimental parameters discussed. In this paper, we present an expression for the beam ratio dependence of diffraction efficiency which leads to a more complete theoretical understanding of the inversion process and its dependence on crystal and experimental parameters.

The regime of validity of the theory is discussed. In particular, the range of beam ratios over which inversion occurs (the inversion dynamic range) is investigated both theoretically and experimentally, using a $Bi_{12}GeO_{20}$ (BGO) crystal in a two-wave mixing configuration. Results of inversion of gray-scale objects are presented. It should be noted that, in contrast to the real-time division demonstrated by Ja in a reflection grating mode [8], we are utilizing the transmission grating formed within the BGO crystal, as well as an applied electric field. Most significantly, our inversion of gray-scale objects is accomplished in an imaging configuration while Ja's system apparently operated on the Fresnel diffraction pattern of a two-gray-level mask.

Our interest in inversion lies in its potential application to many practical problems. While photorefractive crystals have been used to perform phase distortion correction, the added property of inversion could correct for amplitude distortion as well. Because inversion is basically a division process, using the inversion properties of the crystal in the Fourier domain could allow deconvolution of two images to be performed; in particular, a deblurring system could be envisioned [9].

II. Theory

Suppose two plane waves of wavelength λ_1 are incident on a photorefractive crystal surface, as shown in Fig. 1. An index of refraction grating is formed within the crystal, and a probe beam of wavelength λ_2, incident at the Bragg angle appropriate for its wavelength, can diffract off this grating. All polarizations are perpendicular to the plane of incidence. The diffraction efficiency may be found using the standard coupled-wave analysis for a volume phase hologram [10]:

$$\eta = e^{-\frac{\alpha_2 d}{\cos\theta_2}} \sin^2\left(\frac{\pi \Delta n d}{\lambda_2 \cos\theta_2} \right) \tag{1}$$

where α_2 is absorption per unit length of the probe beam, d is the crystal thickness, and θ_1 and θ_2 refer to the angles of the writing and probe beams as measured inside the crystal. The appropriate expression for Δn is found by considering the physical processes involved in photorefraction. Incident light intensity excites electrons from a donor/trapping level into the conduction band where they migrate under the influence of drift and diffusion. The electrons are subsequently retrapped and a space-charge field develops. Through the electro-optic effect, a change in the refractive index occurs which is proportional to the space-charge field. In one-dimension, the charge transport equations are

$$\frac{d\rho_c}{dt} = (AI + \beta)\rho_d - B\rho_c - \frac{dJ}{dx} \tag{2}$$

$$\frac{d\rho_d}{dt} = B\rho_c - (AI + \beta)\rho_d \tag{3}$$

$$\frac{dE}{dx} = \frac{1}{\epsilon}(\rho_c + \rho_d - \rho_o) \tag{4}$$

$$J = \mu \rho_c E - \frac{\mu k_B T}{q} \frac{d\rho_c}{dx} \tag{5}$$

where

ρ_c = charge density in conduction band
ρ_d = charge density in donor/trapping level
ρ_o = total background charge density of donor/trap sites
E = electric field inside crystal
I = incident optical intensity
J = current density
A = excitation rate per unity intensity
B = spontaneous decay rate

$\beta =$ spontaneous excitation rate

$\mu =$ carrier mobility

$T =$ temperature of crystal

$k_B =$ Boltzmann' s constant

$\epsilon =$ static dielectric constant

and $-|q| =$ charge on an electron

If the two writing beams are designated I_o and I_r (object and reference), then

$$I = I_T(1 + m\cos Kx) \tag{6}$$

where

$$I_T = I_o + I_r$$

$$m = \frac{2\sqrt{I_o I_r}}{I_o + I_r} \text{ , the modulation depth}$$

$$\text{and } K = \frac{2\pi}{\Lambda} = \frac{2\pi}{(\lambda/n)/2\sin\theta_1} \text{ , the grating spatial frequency.}$$

We are interested in steady-state conditions. In addition, when the induced photocurrent is well above the dark current ($AI_T >> \beta$), and when ρ_d is nearly ρ_o, the approximation

$$\rho_c \cong (\frac{AI}{B})\rho_o \tag{7}$$

may be used. Substituting this expression in Eq. (5) and solving for E yields

$$E = \frac{C}{1 + m\cos Kx} - \frac{\frac{k_B TK}{q} m\sin Kx}{1 + m\cos Kx} \text{ .} \tag{8}$$

where C is a constant to be determined. Since

$$V = \int_0^d E \ ds \text{ ,} \tag{9}$$

C may be solved for in terms of V:

$$C = \frac{V}{d}\sqrt{1-m^2} \quad . \tag{10}$$

Labeling

$$E_a \equiv \frac{V}{d} \quad , \text{ the applied field}$$

$$\text{and} \quad E_d \equiv \frac{Kk_B T}{q} \quad , \text{ the diffusion field,}$$

E may be rewritten as

$$E = \frac{E_a\sqrt{1-m^2}}{1 + m\cos Kx} - \frac{E_d\, m \sin Kx}{1 + m\cos Kx} \quad . \tag{11}$$

The coefficient of E at the fundamental frequency is [11]:

$$E_1 = \left(\frac{\sqrt{1-m^2}-1}{m}\right)(E_a - iE_d) \quad . \tag{12}$$

This equation essentially duplicates the expression found by Moharam, et al [11]. It is enlightening to rewrite this equation in terms of beam ratio R $(\equiv I_o/I_r)$.[1] Thus,

$$m = \frac{2\sqrt{R}}{1+R}$$

and

$$E_1 = -(E_a - iE_d)\cdot \begin{cases} \sqrt{R} & : R < 1 \\ \dfrac{1}{\sqrt{R}} & : R > 1 \end{cases} \quad . \tag{13}$$

Note that for large R, $1/\sqrt{R}$ is approximately equal to $m/2$. The magnitude of the change in refractive index at the fundamental frequency is

[1] Note that this definition for beam ratio is the reciprocal of the definition usually found in the holography literature. We have found our definition convenient since we are concerned with regimes where $I_o > I_r$.

$$\Delta n = \frac{1}{2}\, n^3 r \cdot \frac{1}{\sqrt{R}} (E_a^2 + E_d^2)^{\frac{1}{2}} \quad , \tag{14}$$

for $R > 1$. Here, n is the unperturbed index of refraction and r is the electro-optic coefficient appropriate to the geometry of the set-up. Since the argument of Eq. (1) is generally quite small, the $\sin x \cong x$ approximation is valid; hence, for $I_o > I_r$,

$$\eta \approx \frac{I_r}{I_o} \quad . \tag{15}$$

It is possible for I_T to be small enough such that it is not accurate to neglect β in favor of AI_T. If we define

$$\delta_o \equiv \frac{\beta}{AI_T} \quad \text{and} \quad m' \equiv \frac{m}{1+\delta_o} \quad ,$$

then in the new expression for E_1, each m is replaced by m'. In this case, E_1 is no longer strictly proportional to $1/\sqrt{R}$. Furthermore, in an experimental situation, I_o will vary as I_r stays constant. This simulates the change in beam ratio that would occur if I_o varied over an image. I_T thus varies with beam ratio, and δ_o is not a constant in a diffraction efficiency vs. beam ratio analysis. In this case, δ_o should be rewritten as

$$\delta_o \equiv \frac{\delta}{1+R} \quad \text{where} \quad \delta \equiv \frac{\beta}{AI_r} \quad . \tag{16}$$

The other approximation used in Eq. (7) ($\rho_d \cong \rho_o$) may be violated when the space-charge field becomes large. To avoid this, upper limits must be placed on the values of the applied field and the modulation depth of the grating. An extension of this theory which discusses these limits in depth will be presented in a future paper [12]. However, it can be said that for applied fields of 2 kV or less, the preceding theory is valid for modulation depths up to very close to

unity.

III. Experimental Set-Up

The two-wave mixing system shown in Fig. 2 was used to obtain data to compare to the expression developed in Section II. An Argon ion laser ($\lambda = 514.5$ nm) was collimated and split to form the two writing beams. The combination of a half-wave plate, a polarizing cube beamsplitter (PCB) and a second half-wave plate allows the beam ratio to be changed while keeping the polarizations the same. The probe beam ($\lambda = 632.8$ nm) entered the crystal, from the opposite side as the writing beams, at the Bragg angle. Care was taken to reduce the probe beam to such an intensity such that it did not affect the grating formed by the writing beams. The BGO crystal, of size 10 x 10 x 10 mm^3, was oriented with the x-direction shown in Fig. 2 along the [001] axis. Gold was evaporated onto two sides of the crystal which were then placed in contact with copper electrodes. The diffracted red beam was directed off a beamsplitter into a power meter, after passing through a filter which blocked green light. Diffraction efficiency was calculated by comparing the output diffracted power with the total power in the red beam after it had traversed the crystal when no grating was present. To obtain diffraction efficiency vs. beam ratio data, I_r was fixed and I_o (and thus R) was varied by the insertion of neutral density filters.

Performing inversion of a gray-scale object requires that the probe beam simultaneously satisfy the Bragg conditions for all the gratings formed in the crystal, since the object can be considered as a superposition of plane waves at different angles. This can only be accomplished by a probe beam of the same wavelength as the writing beams. When all beams emanate from the same laser and so are related in phase, a degenerate four-wave mixing system is obtained.

The experimental configuration used is depicted in Fig. 3. The variable attenuator/beamsplitter (VA/BS) regulates the amount of light in the probe beam and preserves the polarization of light in both legs. The probe beam was aligned to be counterpropagating to the reference beam. A lens imaged the object onto the crystal, and the output was detected by a TV camera. To improve the signal-to-noise ratio, a polarizer was placed in front of the output to reduce the scattered light [13]. The crystal was placed in a Freon environment in order to avoid dielectric breakdown at higher applied fields.

IV. Results

The two-wave mixing data are plotted as $log(\eta \cdot R)$ vs. $log(R)$ so that inversion is displayed as a horizontal line. The inversion dynamic range may then easily be found from the graphs. In the experiment, the limit on increasing beam ratio is due both to intensity limitations of the laser and to the ability to detect the decreasing output signal.

The following values were used to compute the theoretical plots:

$$
\begin{aligned}
d &= 1 \text{ cm} \\
\alpha_2 &= 0.38 \text{ cm}^{-1} \text{ (measured in our crystal)} \\
\theta_1 &= 2^o \\
n &= 2.6 \\
r_{41} &= 3.5 \cdot 10^{-12} \text{ m/V}
\end{aligned}
$$

Fig. 5 shows both experimental and theoretical curves for zero applied voltage (V_a) and for $V_a = 2$ kV. The experimental data have been adjusted in diffraction efficiency in order to compare the curve shapes; the actual diffraction efficiency is approximately one order of magnitude less than the theory predicts. This can be attributed to reflection and scattering of the input light as well as the optical activity of BGO [14], neither of which are included in the theoretical

expression. Since these effects should reduce the output in the same manner for all beam ratios, each point was adjusted by the same multiplicative factor.

A nonzero value for δ is necessary in order to fit the data to the theoretical curves. For the case of Fig. 4, $\delta = 0.25$ and is the same for both values of V_a, as is expected since δ is independent of E_a. Hence, for moderate fields, the inversion dynamic range is independent of the applied field, except that the overall increased diffraction efficiency may permit detection of the inverted output at higher beam ratios. Due to insertion loss at the electrode contacts, the actual value of V_a is less than the applied value. The best fit to our data gave a 30% loss, leading to an effective $V_a = 1.4$ kV rather than 2 kV. The I_T dependence of δ is verified in Fig. 5 where curves for two values of I_r are shown. The ratio of δ_1/δ_2 is approximately inversely proportional to I_{r_1}/I_{r_2}. Note that both curves approach the same diffraction efficiency at large beam ratios and only diverge significantly for beam ratios less than 10:1. Consequently, for all other parameters fixed, it is desirable to have a higher incident intensity to maximize the linear dynamic range.

Results using images in the four-wave mixing set-up are shown in Figures 6, 7, 8, and 9. In Figure 6a, the phase-conjugate image of an exposure selection mask is shown. A field of 10 kV was applied across the crystal. The transmittances of the sectors of the mask range from 86% to 7%, and the corresponding reference-to-object beam ratio ranges from 4.4 to 54. Note that these beam ratios lie in the typical phase-conjugate imaging regime. The beam ratios (I_o/I_r) used to form the inverse, shown in Fig. 6b, vary from 18 (for the sector which was originally the most transmitting and is now the darkest) to 1.5 (for the sector which is now the brightest). Edge enhancement is apparent at transitions between bright and dark regions because there is some point where

the beam ratio is approximately unity and diffraction efficiency is maximized.

A black-and-white 35mm negative of a face is the object used in the next set of photographs. The phase-conjugate image is shown in Fig. 7a; the inverted image is shown in Fig. 7b, both obtained using an applied field of 10 kV. The dimensions of the portion of the negative shown are 7mm x 4mm, but the imaging lens demagnifies the object by a factor of 2. The transmittance of the negative, measured with a microdensitometer, varies from 21% (for the hair region) to 8% (for the forehead), yielding reference-to-object beam ratios of 5.7 and 14.9, respectively. In the inverted image, corresponding object-to-reference beam ratios varied from 9.7 to 3.7.

The final two objects were portions of a slide made from an M.C. Escher print entitled "Day and Night." The applied field was 10 kV. For both, the most transmitting regions allowed 72% of the light to pass while the darkest regions had a transmittance of 8%. The image of the birds scene, shown in Fig. 8a, had reference-to-object beam ratios ranging from 3.3 to 30. Fig. 8b, the inverse, was produced with object-to-reference beam ratios of 16.3 to 1.8. The corresponding figure for the river and town scene are 4.2 to 38 for the image, seen in Fig. 9a, and 15 to 1.7 for the inverse in Fig 9b. Note that the high contrast of the image produces edge enhancement at almost all boundaries.

Transmittances of the above images and inverses were measured to determine the degree of inversion. For Figures 6, 8, and 9, the ratio of transmittances of the lightest and darkest regions in the images were inversely proportional to the ratio of transmittances of the corresponding regions in the inverses, to within 2%. For Fig. 7, however, this "contrast" was 20% less in the inverse, because the dark regions of the images did not fully invert.

V. Discussion

In assessing the process of real-time optical intensity inversion, it is instructive to analyze the results in terms of parameters such as dynamic range, speed, and resolution. The available inversion dynamic range in the optical system, investigated in the two-wave mixing experiments discussed earlier, can be three or more orders of magnitude at optimal conditions. Our experimental set-up yielded close to two orders of magnitude and could be increased by using a laser of higher output power so that larger beam ratios could be obtained without decreasing I_r and I_p to unacceptably low values. In addition, increasing the electric field increases the diffraction efficiency and allows the output to be detected for a larger range of beam ratios. Of course, there is a limit on the maximum space-charge field that can exist in the crystal, which depends on the concentration of donor/trap sites. Digital techniques are almost unlimited in dynamic range, but this can be a curse as well as a blessing. Many digital inversion schemes boost low-intensity noise to high levels and can seriously degrade the resulting image. In our optical scheme, the very nature of the process precludes this from occurring since only light of an intensity near to or greater than the reference beam intensity participates in inversion.

The inherent parallel nature of optical processing means that the limit on the speed of the system is imposed only by the time constant of the photorefractive crystal. The time constant decreases as the light intensity incident on the crystal increases. In our experiments, the time constant was roughly 50-500 msec. However, with the use of a pulsed laser system, the time constant could be decreased to a value on the order of nanoseconds.

The resolution is theoretically limited by the available donor/trapping sites. This figure is estimated to be 10^{16} cm^{-3} for our crystal, which implies an average

distance between traps of less than 50 nm [15]. This should allow images with spatial frequencies of up to 2000 line pairs/mm to be adequately recorded. More practically, the position and diameter of the imaging lens usually limits the modulation transfer function of the system. This constraint reduced the resolution capability of our system to 100 line pairs/mm. With a crystal surface area of 1 cm^2, this yielded a field of 1000 x 1000 resolution elements.

In the degenerate four-wave mixing system, gratings are actually formed between all pairs of beams, not just the object and reference beams. In particular, a grating along the z-direction is formed between the object and probe beam which can be read out by the reference beam, resulting in a diffracted beam in the same direction as the output we have been considering. However, under our experimental conditions, the diffraction efficiency of this "reflection" grating is much smaller than the diffraction efficiency of the transmission grating. This is a consequence of both the absence of an applied field along the z-direction as well as the presence of optical activity in the crystal which results in an angular difference between the polarizations of the object and probe beams. A rigorous analysis of a degenerate four-wave mixing system must also take into account beam coupling amongst all the beams present. The results of the two-beam coupling analysis presented in Section II are nevertheless sufficient to describe the diffraction efficiency dependence of our scheme since we are working under conditions of weak beam interaction.

However, the applied field used to enhance the diffraction efficiency of the transmission grating also reduces the validity of the small space-charge field approximation discussed earlier, particularly for m close to 1. The effect is a more pronounced fall-off from linearity than what is seen in Figs. 4 and 5, which occurs for smaller m (larger R) as E_a is increased. As a result, it is necessary to

be at higher beam ratios in order to be operating in a true inversion regime. In our four-wave mixing experiments, however, we were also using larger values of I_r, on the order of a few mW/cm^2, which partially offset the increased falloff by decreasing the effective δ.

BGO (and its isomorph BSO) are most often used for optical processing because of their availability in large sizes of good optical quality. However, in terms of optimizing inversion, it would be advantageous to use crystals with a larger electro-optic coefficient. The diffraction efficiency would then be increased at all beam ratios, and it would be possible to realize inversion over a wider dynamic range.

VI. Conclusion

We have measured the range of beam ratios over which real-time intensity inversion occurs and compared it to a theoretical model. The experimental data verified the inversion dynamic range dependence of writing beam intensity, for the moderate light level case. For moderate applied field levels (< 2 kV), the dynamic range is unaffected by the applied field, except that the overall increased diffraction efficiency at higher fields may permit detection of the inverted output at higher beam ratios. Using a BGO crystal in a four-wave mixing configuration, we have demonstrated real-time optical intensity inversion of gray-scale objects. Therefore, it appears feasible to use inversion in applications such as feature enhancement, deconvolution and amplitude correction.

Acknowledgments

This work was supported by the NSF-MRL program through the Center for Materials Research at Stanford University, the Air Force Office of Scientific Research, and an IBM fellowship (E.O.). Stimulating discussions with Fred

Vachss are gratefully acknowledged.

References

1. J.P. Huignard, J.P. Herriau, P. Aubourg, and E. Spitz, "Phase-conjugate wavefront generation via real-time holography in $Bi_{12}SiO_{20}$ crystals," *Optics Letters*, vol. 4, p. 21, 1979.

2. Jeffrey O. White and Amnon Yariv, "Real-time image processing via four-wave mixing in a photorefractive medium," *Applied Physics Letters*, vol. 37, p. 5, 1980.

3. J.P. Huignard and J.P. Herriau, "Real-time double-exposure interferometry with $Bi_{12}SiO_{20}$ crystals in transverse electrooptic configuration," *Applied Optics*, vol. 16, p. 1807, 1977.

4. J.P. Huignard and A. Marrakchi, "Coherent signal beam amplification in two-wave mixing experiments with photorefractive BSO crystals," *Optics Communications*, vol. 38, p. 249, 1981.

5. Y.H. Ja, "Real-time image subtraction in four-wave mixing with photorefractive BGO crystals," *Optics Communications*, vol. 42, p. 377, 1982.

6. J.P. Huignard and J.P. Herriau, "Real-time Coherent Object Edge Reconstruction with $Bi_{12}SiO_{20}$ Crystals," *Applied Optics*, vol. 17, p. 2671, 1978.

7. Jack Feinberg, "Real-time Edge Enhancement using the Photorefractive Effect," *Optics Letters*, vol. 5, p. 330, 1980.

8. Y.H. Ja, "Real-time image division in four-wave mixing with photorefractive BGO crystals," *Optics Communications*, vol. 44, p. 24, 1982.

9. Y.H. Ja, "Real-time image deblurring using four-wave mixing," *Optical and Quantum Electronics*, vol. 15, p. 457, 1983.

10. Herwig Kogelnik, "Coupled Wave Theory for Thick Hologram Gratings," *The Bell System Technical Journal*, vol. 48, p. 2909, 1969.

11. M.G. Moharam, T.K. Gaylord, R. Magnusson, and L. Young, "Holographic grating formation in photorefractive crystals with arbitrary electron transport lengths," *Journal of Applied Physics*, vol. 50, p. 5642, 1979.

12. Ellen Ochoa, Frederick Vachss, and Lambertus Hesselink, "A Higher-Order Analysis of the Photorefractive Effect for Large Modulation Depths," *JOSA A*, (submitted).

13. J.P. Herriau, J.P. Huignard, and P. Aubourg, "Some polarization properties of volume holograms in $Bi_{12}SiO_{20}$ crystals and applications," *Applied Optics*, vol. 17, p. 1851, 1978.

14. Frederick Vachss and Lambertus Hesselink, "Holographic Beam Coupling in Generally Retarding Media," *JOSA A*, (submitted).

15. M. Peltier and F. Micheron, "Volume hologram recording and charge transfer process in $Bi_{12}SiO_{20}$ and $Bi_{12}GeO_{20}$," *Journal of Applied Physics*, vol 48, p. 3683, 1977.

Fig. 1

Geometry of grating formation and read-out

Fig. 2

Two-wave experimental set-up

CL: collimating lens; PCB: polarizing cube beamsplitter; BS: beamsplitter; N.D.: neutral density.

Fig. 3

Four-wave experimental set-up

VA/BS: variable attenuator/beamsplitter; CL: collimating lens; PCB: polarizing cube beamsplitter; BS: beamsplitter; IL: imaging lens.

Fig. 4

Dependence of inversion dynamic range on applied field

The solid curves represent the theoretical model using $\delta = 0.25$ and applied voltages of 0 and 1.4 kV. The points are experimental data taken using the effective applied voltages of 0 and 1.4 kV.

Fig. 5

Dependence of inversion dynamic range on writing beam intensity

The solid curves represent the theoretical model using the $\delta = 0.25$ and $\delta = 0.40$. The experimental points were taken at $I_r = 0.5$ mW/cm^2 and $I_r = 0.3$ mW/cm^2. The applied voltage was zero.

Fig. 6

Phase-conjugate image and inverse of an exposure selection mask

(a) shows the image at reference-to-object beam ratios ranging from 4.4 for the brightest sector to 5.1 for the darkest sector. (b) is the inverse with object-to-reference beam ratios ranging from 18 (for the sector which was originally the brightest and is now the darkest) to 1.5 (for the sector which is now the brightest).

Fig. 7

Phase-conjugate image and inverse of a face

(a) Image of a 35mm negative with I_r/I_o from 5.7 to 14.9. (b) Inverse, with corresponding I_o/I_r from 9.7 to 3.7.

Fig. 8

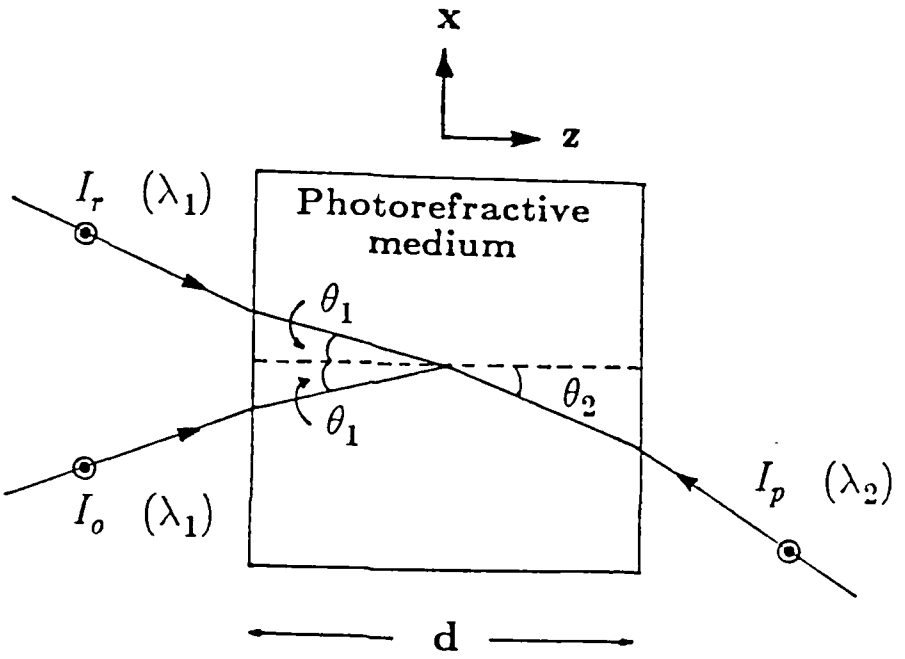
Phase-conjugate image and inverse of a scene

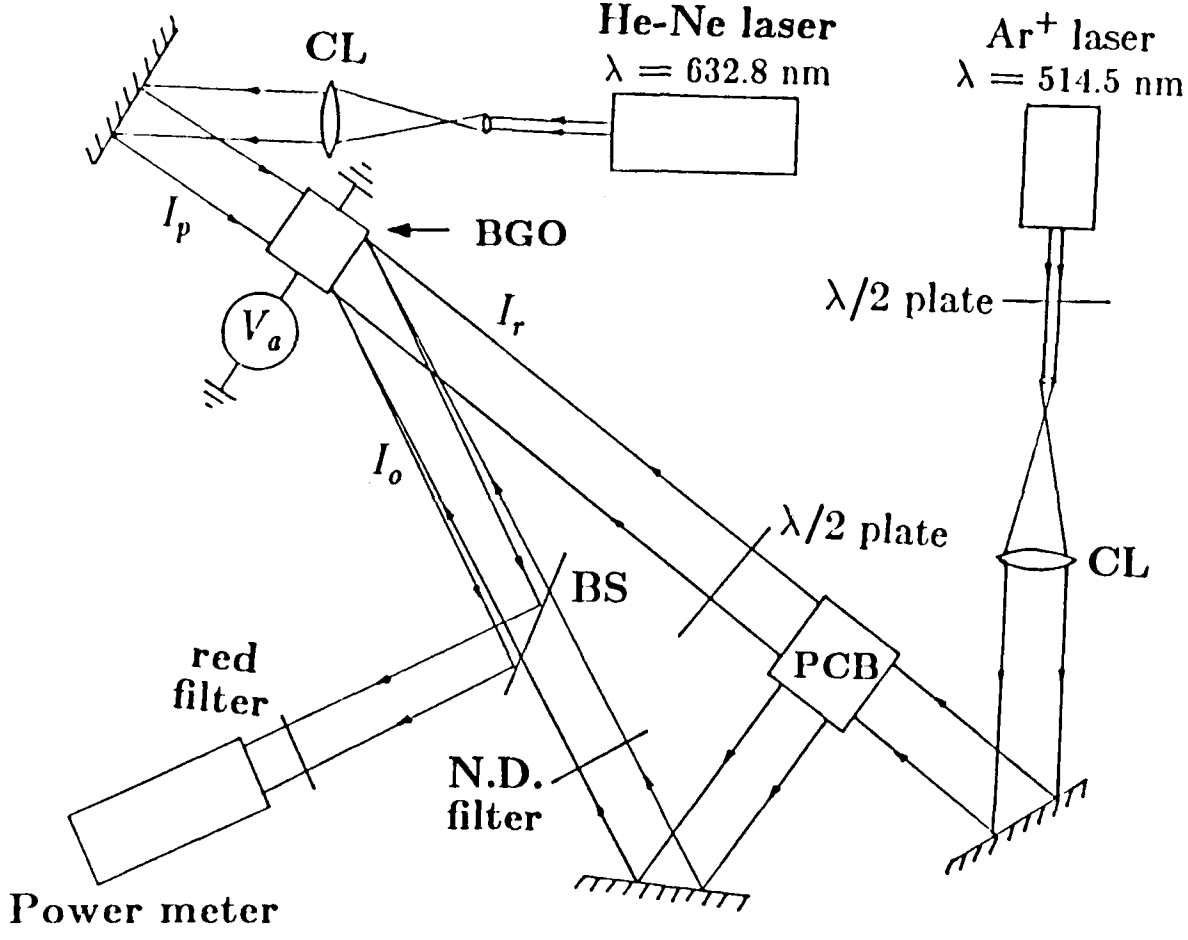
(a) Image of a portion of a slide made from an M.C. Escher print. Reference-to-object beam ratios were 3.3 to 30. (b) Inverse obtained with object-to-reference beam ratios of 16.3 to 1.8.

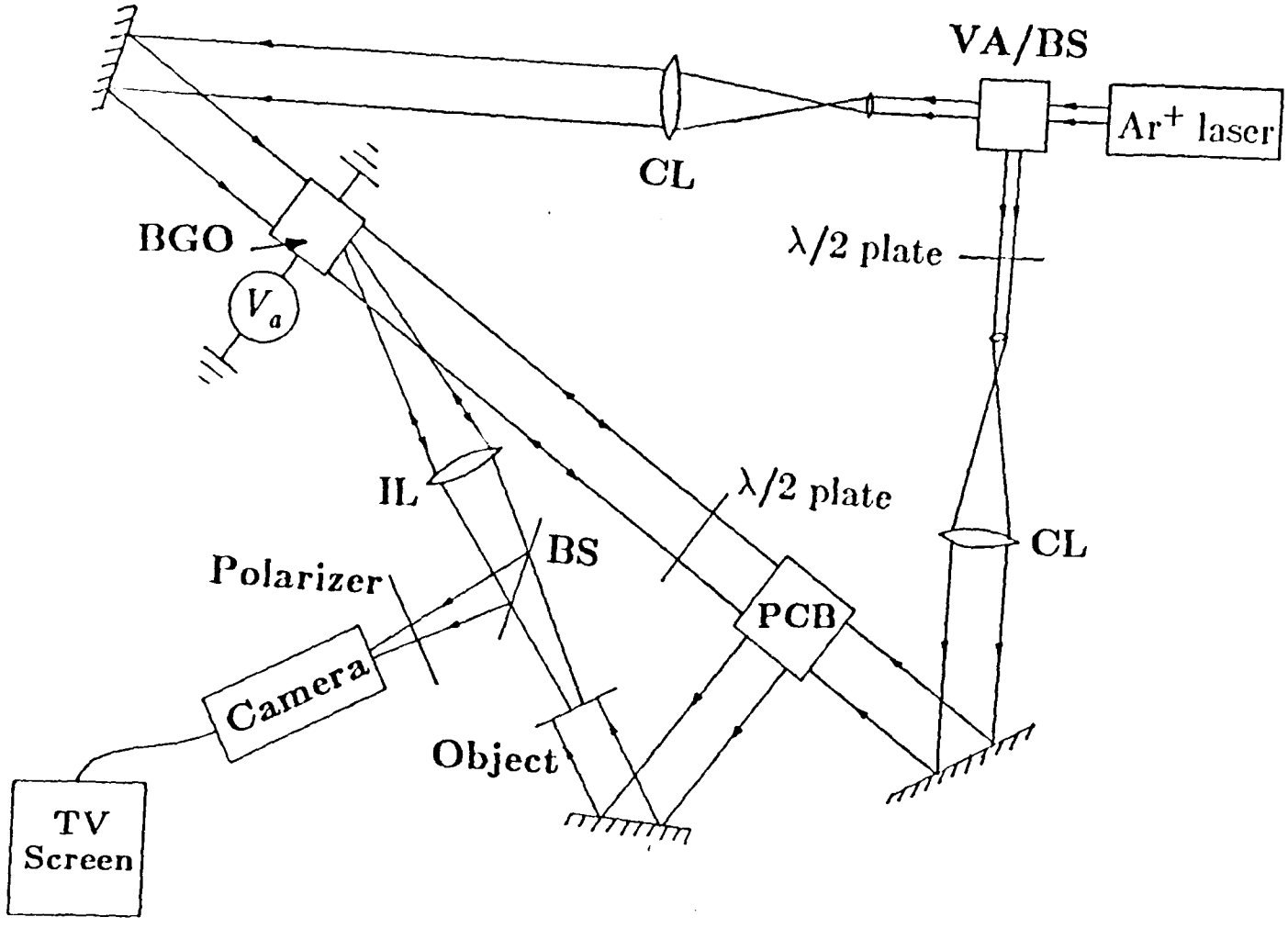
Fig. 9

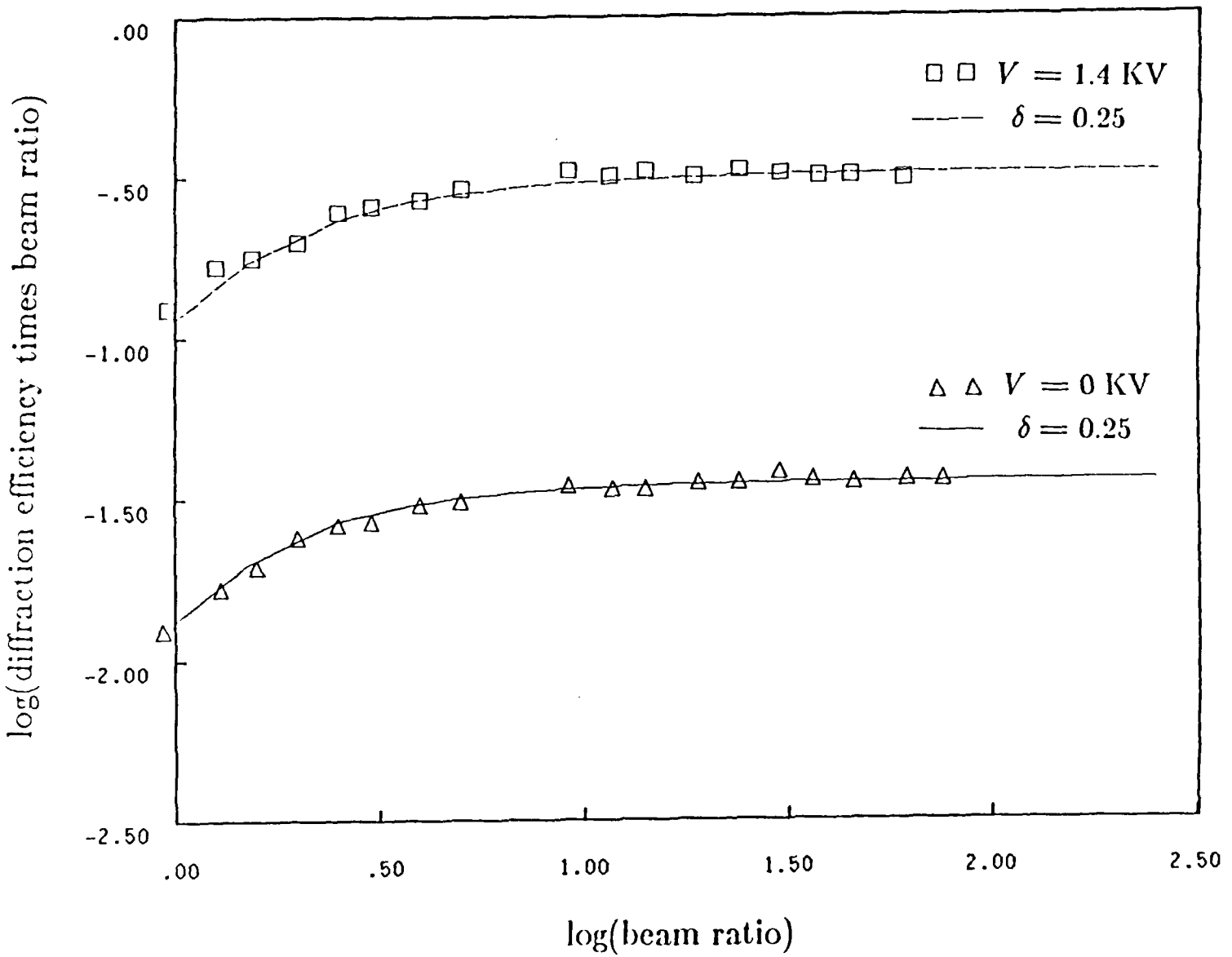
Phase-conjugate image and inverse of a scene

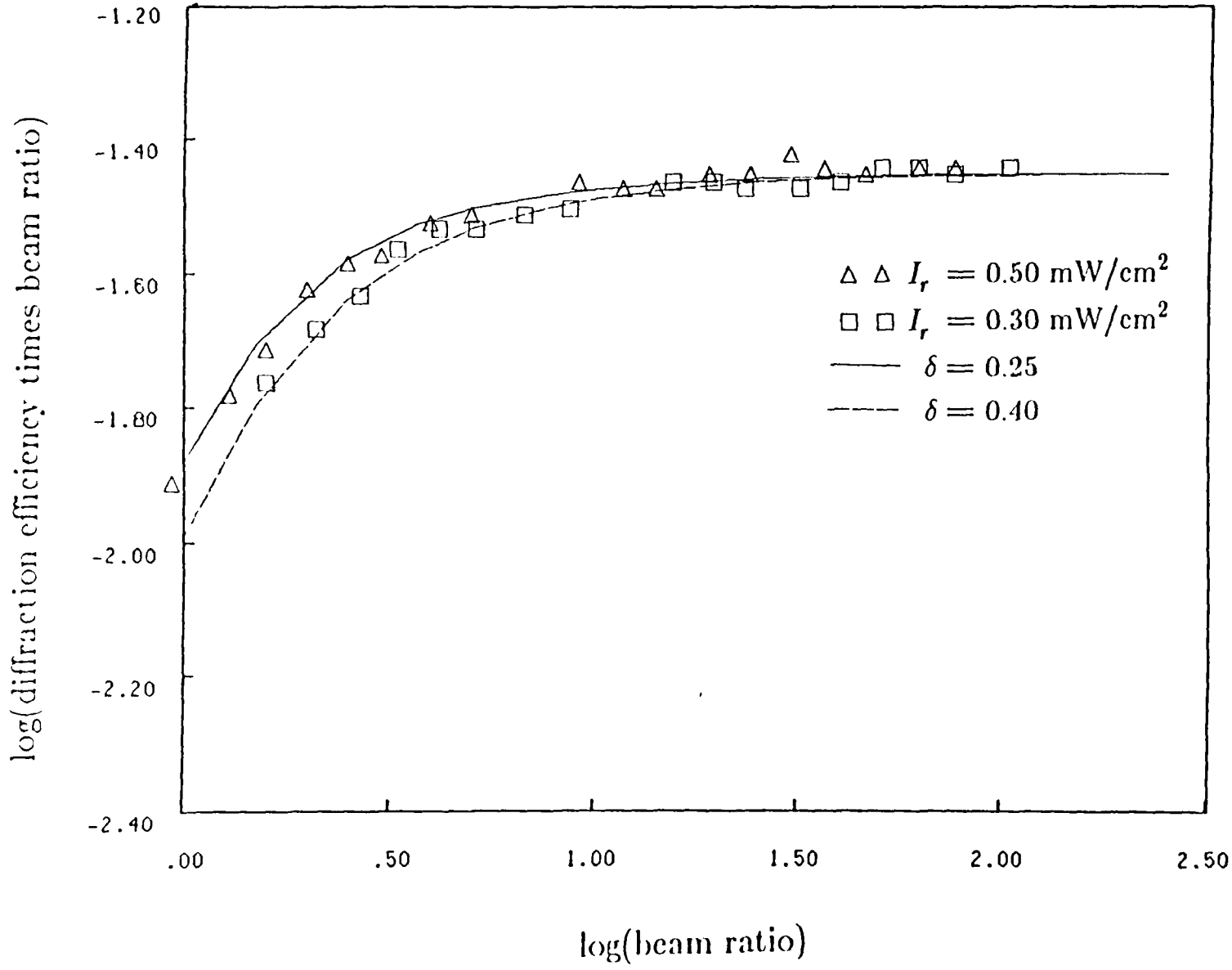
(a) Image from a different portion of the same Escher slide, using I_r/I_o from 4.2 to 38. (b) Inverse, with I_o/I_r from 15 to 1.7.

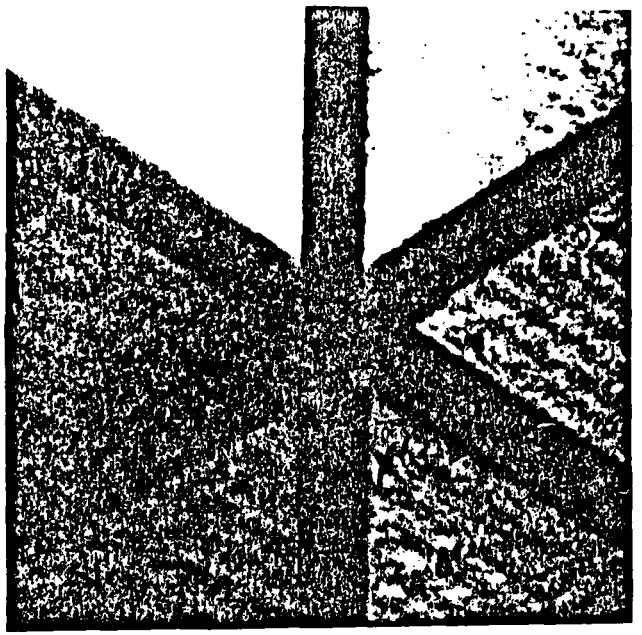

x
z
I_r (λ_1)
Photorefractive medium
θ_1
θ_2
θ_1
I_o (λ_1)
I_p (λ_2)
d


CL
He-Ne laser
λ = 632.8 nm
Ar+ laser
λ = 514.5 nm
I_p
BGO
V_a
I_r
λ/2 plate
I_o
λ/2 plate
CL
BS
PCB
red
filter
N.D.
filter
Power meter

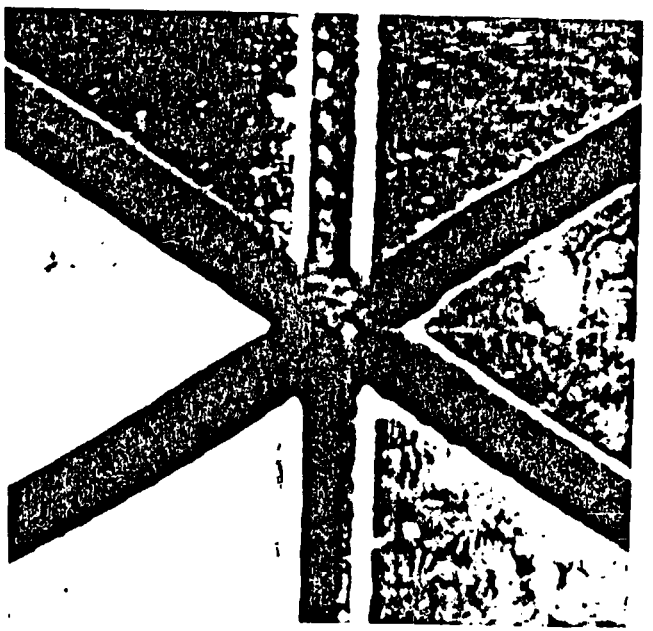

VA/BS
Ar+ laser
CL
BGO
λ/2 plate
Va
IL
λ/2 plate
CL
BS
Polarizer
PCB
Camera
Object
TV
Screen


.00
-.50
-1.00
-1.50
-2.00
-2.50
.00
.50
1.00
1.50
2.00
2.50
□ □ V = 1.4 KV
---- δ = 0.25
△ △ V = 0 KV
—— δ = 0.25
log(diffraction efficiency times beam ratio)
log(beam ratio)

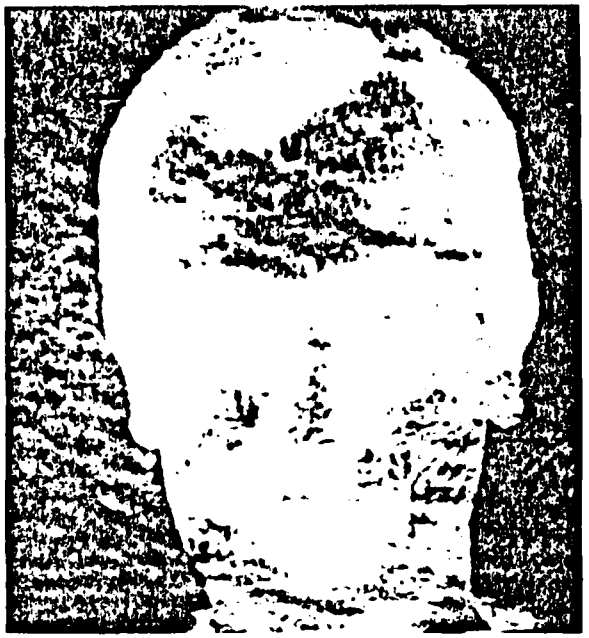

-1.20
-1.40
-1.60
-1.80
-2.00
-2.20
-2.40
.00
.50
1.00
1.50
2.00
2.50
log(diffraction efficiency times beam ratio)
log(beam ratio)
△ △ $I_r = 0.50$ mW/cm^2
□ □ $I_r = 0.30$ mW/cm^2
—— $\delta = 0.25$
---- $\delta = 0.40$



(a) IMAGE

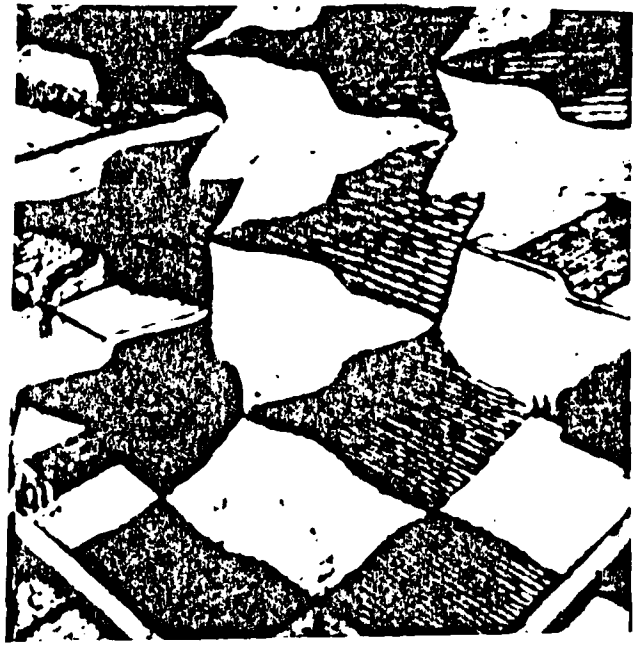


(b) INVERSE



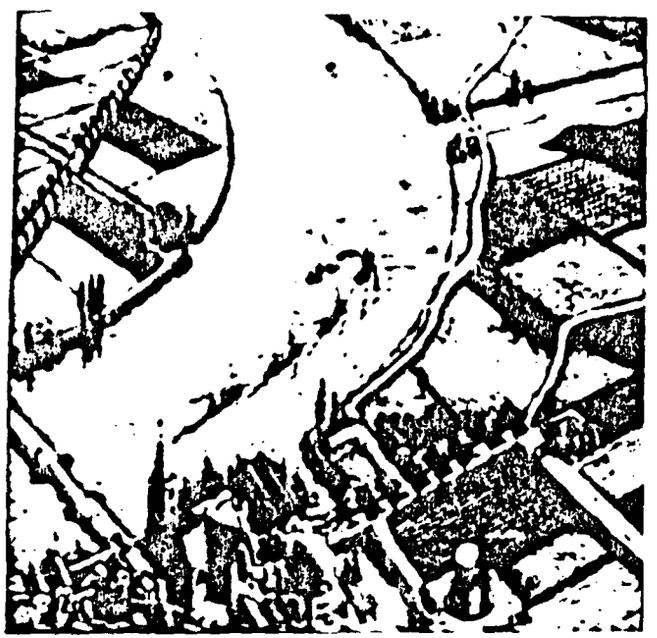
(a) IMAGE

(b) INVERSE



(a) IMAGE

(b) INVERSE



(a) IMAGE

(b) INVERSE

Real-time Enhancement of Defects in a Periodic Mask using Photorefractive $Bi_{12}SiO_{20}$

Ellen Ochoa, Joseph W. Goodman, and Lambertus Hesselink

Stanford University
Department of Electrical Engineering
Informations Systems Laboratory
Stanford, CA 94305

ABSTRACT

The first experimental results of real-time optical defect enhancement of a periodic mask are reported. A low-intensity reference wave interferes with the Fourier transform of an object beam to form a hologram in a photorefractive crystal. The nonlinear properties of the crystal perform a filtering operation, and phase-conjugate read-out results in a defect-enhanced image. Defects of size 10 x 100 μm^2 have been easily detected with high signal-to-noise ratio, and a discussion of performance limitations is presented.

Real-time Enhancement of Defects in a Periodic Mask using Photorefractive $Bi_{12}SiO_{20}$

Ellen Ochoa, Joseph W. Goodman, and Lambertus Hesselink

Stanford University

Department of Electrical Engineering

Informations Systems Laboratory

Stanford, CA 94305

In this letter, we consider the problem of selectively enhancing defects in a mask which consists of mostly periodic structure. This type of problem in image processing occurs, for example, in the inspection of integrated circuit masks. Digital techniques for inspection of a two-dimensional field, generally utilizing a dual scanning microscope system and sophisticated algorithms for comparison and detection, are complicated and time-consuming [1-6]. Optical systems, however, offer the advantage of parallel processing. Furthermore, there is no excessive requirement on accuracy in the output in terms of the actual intensity at each point. It is sufficient that the signal associated with the defect be much larger than the signal associated with the surrounding periodic structure, so that, for example, a thresholding operation can be used to determine the defect location.

Optical spatial filtering techniques to perform defect enhancement have been examined in the past with regard to such applications as inspection of the electron beam collimating grid and the silicon diode array target for a television camera tube as well as photomasks used in the manufacture of integrated circuits [7-

9]. These systems used a filter in the Fourier plane to attenuate the discrete spatial frequencies of the periodic portion of the mask, so that, upon retransformation, only images of defects were present in the output. Though the results of such systems were promising, the usefulness of the technique was limited by the fabrication time or difficulty of the filter, and by the need to use high-quality, low f# lenses when inspecting objects of large dimensions. Recently, the second constraint has been removed by employing holographic recording of the output combined with phase-conjugate read-out [10]. While this method has been used to detect submicron defects, it requires two processing steps: for each mask to be inspected, a new hologram must be recorded, and for each different type of mask, a new photographic filter must be made.

We present a method to enhance defects in real-time, using a photorefractive crystal. Use of the crystal allows holographic recording, filtering, and phase-conjugate read-out processes to be performed simultaneously. The mask to be inspected is placed in the input plane, and the defect-enhanced image appears at the output plane, in a time limited only by the time constant of the photorefractive material. This time constant, which depends on the material used and the incident light intensity, ranged from about 50-250 msec for our experimental parameters. This method also differs from that described above in that all operations are carried out in the Fourier domain. To our knowledge, this work is the first demonstration of a real-time system for enhancing small defects in a periodic mask.

The technique for performing real-time defect enhancement is based on two observations. The first is that the Fourier transform of a periodic object is an array of discrete spikes whose width depends inversely on the input field size and whose spacing depends inversely on the period of the mask. In contrast, the

Fourier transform of a small defect is a continuous function which is several orders of magnitude less intense than the periodic spikes. The second observation is that the diffraction efficiency of a volume phase hologram formed in a photorefractive medium is maximized when the intensities of the two writing beams are approximately equal and decreases as the difference in intensity increases. For a reference plane wave more intense than the object beam, the output is proportional to the object beam intensity; for an object beam more intense than the reference beam, the output is proportional to the intensity inverse of the object beam. A typical diffraction efficiency versus beam ratio curve is plotted in Fig. 1 on a log-log scale, assuming that beam ratio R ($\equiv I_o / I_r$) is varied by changing I_o while keeping I_r fixed [11]. Therefore, a defect can be enhanced by focusing the Fourier transform of the mask onto the photorefractive crystal and matching the intensity of the peak spectral component due to the defect to the intensity of the reference beam. The beam ratio of the defect spectrum intensity to the reference beam intensity should fall between R_1 and 1 (see Fig. 1). The intensity of the spikes due to the periodic structure will be so much greater than the reference beam intensity ($R > R_2$), that the corresponding diffraction efficiency will be very small. Thus the refractive index pattern formed inside the crystal performs both recording and filtering operations.

A Fourier optics analysis can be used to describe the propagation of light from the object to the crystal. Suppose the mask is rectangular with dimensions W x L and a small transparent rectangular defect, located at (x_o, y_o), has dimensions w x l. Let $p(x, y)$ represent one unit cell of the periodic structure, which is spaced at intervals of length a . The intensity of the Fourier transform at the crystal, assuming $W, L >> a$ and unit illumination, is

$$| T(u,v) |^2 \cong \frac{1}{(\lambda f)^2} \left[(WL)^2 \frac{1}{a^4} \sum_n \sum_m P^2(\frac{n}{a},\frac{m}{a}) \mathrm{sinc}^2(W(u-\frac{n}{a})) \mathrm{sinc}^2(L(v-\frac{m}{a}) \right.$$
$$\left. + (wl)^2 \mathrm{sinc}^2(wu) \mathrm{sinc}^2(lv) \right] ,$$

where the sinc function is as defined in Bracewell [12]. The spatial frequencies variables are related to spatial variables as $u = x/\lambda f$ and $v = y/\lambda f$ and $P(u,v)$ is the Fourier transform of $p(x,y)$. $P(0,0)$ represents the transmitting area of one period of the pattern, and $P(0,0)/a^2$ is the fraction of the mask area that is transmitting. If D is defined as the dynamic range of the periodic portion, which we wish to record in the $R > R_2$ region, then the ratio of the intensities of the mask portion to the defect portion of $| T(u,v) |^2$, assuming the reference beam intensity is matched to the latter, should satisfy

$$\frac{(WL)^2 \frac{P^2(0,0)}{a^4}}{(wl)^2} \geq R_2 D \quad .$$

For most objects of interest, this inequality is easily satisfied.

The experimental set-up used to obtain defect enhancement is shown in Fig. 2. An Argon ion laser ($\lambda = 514.5$ nm) was collimated and split to form the two writing beams, as well as the probe (read-out) beam. A $Bi_{12}SiO_{20}$ (BSO) crystal, of size 8 x 8 x 8 mm^3, was oriented with the x-direction along a [110] axis. A lens of focal length 381 mm and diameter 78 mm was used to perform the Fourier transform, and the output was detected by a CCD camera. To improve the signal-to-noise ratio, a polarizer was placed in front of the output to reduce the scattered light [13].

The object mask consisted of a 36 x 36 array of squares, each with sides of 150 μm. The spacing between the squares was 100 μm, so the period a was equal to 250 μm. The total mask size was 9 x 9 mm^2. Within this array were

placed seven transmitting defects of sizes 100 x 100 μm^2, 100 x 50 μm^2 (2 times), 100 x 25 μm^2 (2 times), and 100 x 10 μm^2 (2 times), as shown in Fig. 3a. The output of the optical system is shown in Fig. 3b, obtained using an applied voltage of 4 kV. The periodic background has been quite effectively suppressed, leaving the defects clearly visible. Fig. 4 shows an intensity scan of one line of the output image, illustrating the worst-case signal-to-noise ratio obtained. The defect represented is one of the two 10 x 100 μm^2 spots; thus, the system appears easily capable of detecting smaller defects.

In recording the hologram, the object beam intensity at the mask (I_{o_1}) was 16 mW/cm^2 and the reference beam intensity was 3.0 mW/cm^2 which led to beam ratios at the crystal (for the defect alone) of 0.014 to 0.00014, depending on the size of the defect. Thus, the experimental results indicate that enhancement occurs even for values of R much less than one. As a result, the incident beam ratio at the mask needed to obtain a defect-enhanced output (I_{o_1}/I_r) did not vary much in response to a range of defect areas of more than two orders of magnitude. Because the inverse properties shown in Fig. 1 were derived under conditions of plane wave illumination, the filtering properties of the crystal cannot be described by simply a beam ratio dependence. Further investigation into the actual behavior of the crystal is currently being undertaken. The ratio of the peak of the periodic portion of the transform to the peak of the defect portion ranged from 8.5 x 10^6 to 8.5 x 10^8.

The resolution obtained in the output was constrained by two factors. The primary constraint was the size of the crystal. Given the f# of the system, the crystal captured only the central fifth of the primary lobe of the sinc function due to the smallest defect; therefore, the output of the system produced the defect convolved with a smoothing function. Even for the largest defect present,

only the central lobe plus one lobe on each side were within the crystal boundaries. Thus, reducing the f# of the optical system (and using a crystal of larger dimensions) will greatly improve the resolution capability. The second constraint on the resolution was the size of the imaging elements of the CCD camera, each of which measured 23 μm x 13.4 μm.

In summary, a method to enhance defects in a periodic mask in real-time has been presented. A photorefractive crystal is used to perform holographic recording, filtering, and read-out processes simultaneously. Preliminary experimental results show detection of defects down to 10 x 100 μm^2 in size. Detection of smaller defects should be possible by using an optical system with a smaller f# and a camera with smaller resolution elements.

This work was supported by the NSF-MRL program through the Center for Materials Research at Stanford University and the Air Force Office of Scientific Research. The assistance of Mike Smith and Zora Norris in the mask preparation is greatly appreciated.

References

1. J.D. Knox, P.V. Goedertier, D. Fairbanks, and F. Caprari, "Inspecting IC Masks with a Differential Laser Scanning Inspection System," *Solid State Technology*, vol. 20, p. 48, May 1977.

2. Kenneth Levy, "Automated Equipment for 100% Inspection of Photomasks," *Solid State Technology*, vol. 21, p. 60, May 1978.

3. KLA Instruments Corp., "Automatic Reticle/Mask Inspection System for VLSI," *Solid State Technology*, vol. 26, p. 45, January 1983.

4. Donald B. Novotny and Dino R. Ciarlo, "Automated Photomask Inspection: Part 1," *Solid State Technology*, vol. 21, p. 51, May 1978.

5. R.A. Simpson, D.E. Davis, "Detecting submicron pattern defects on optical photomasks using an enhanced EL-3 electron-beam lithography tool," *Proceedings of SPIE - Optical Microlithography*, vol. 334, p. 230, 1982.

6. Bunjiro Tsujiyama, Kunio Saito, and Kenji Kurihara, "A Highly Reliable Mask Inspection System," *IEEE Transactions on Electron Devices*, vol. ED-27, p. 1284, 1980.

7. L.S. Watkins, "Inspection of Integrated Circuit Photomasks with Intensity Spatial Filters," *Proceedings of the IEEE*, vol. 57, p. 1634, 1969.

8. Norman N. Axelrod, "Intensity Spatial Filtering Applied to Defect Detection in Integrated Circuit Photomasks," *Proceedings of the IEEE*, vol. 60, p. 447, 1972.

9. Robert A. Heinz, Richard L. Odenweller,Jr., Robert C. Oehrle, and Laurence S. Watkins, "Tool and Product Inspection by Optical Spatial Filtering of Periodic Images," *The Western Electric Engineer*, vol. 17, p. 39, 1973.

10. R.L. Fusek, K. Harding, L.H. Lin, and S.C. Gustafson, "Holographic optical

processing for submicron defect detection," *Proceedings of the SPIE*, vol. 523, January 1985.

11. Ellen Ochoa, Lambertus Hesselink, and Joseph W. Goodman, "Real-time Intensity Inversion using Two-Wave and Four-Wave Mixing in Photorefractive $Bi_{12}GeO_{20}$," *Applied Optics*, to be published June 15, 1985.

12. Ronald N. Bracewell, *The Fourier Transform and its Applications*, McGraw-Hill, Inc., New York, 1978.

13. J.P. Herriau, J.P. Huignard, and P. Aubourg, "Some polarization properties of volume holograms in $Bi_{12}SiO_{20}$ crystals and applications," *Applied Optics*, vol. 17, p. 1851, 1978.

Fig. 1

Diffraction efficiency vs. beam ratio

Shown is a typical curve for photorefractive BSO or BGO.

Fig. 2

Experimental set-up

VA/BS: variable attenuator/beamsplitter; BS: beamsplitter; CL: collimating lens; PCB: polarizing cube beamsplitter; FTL: Fourier transform lens.

Fig. 3

Input mask and output defect-enhanced image

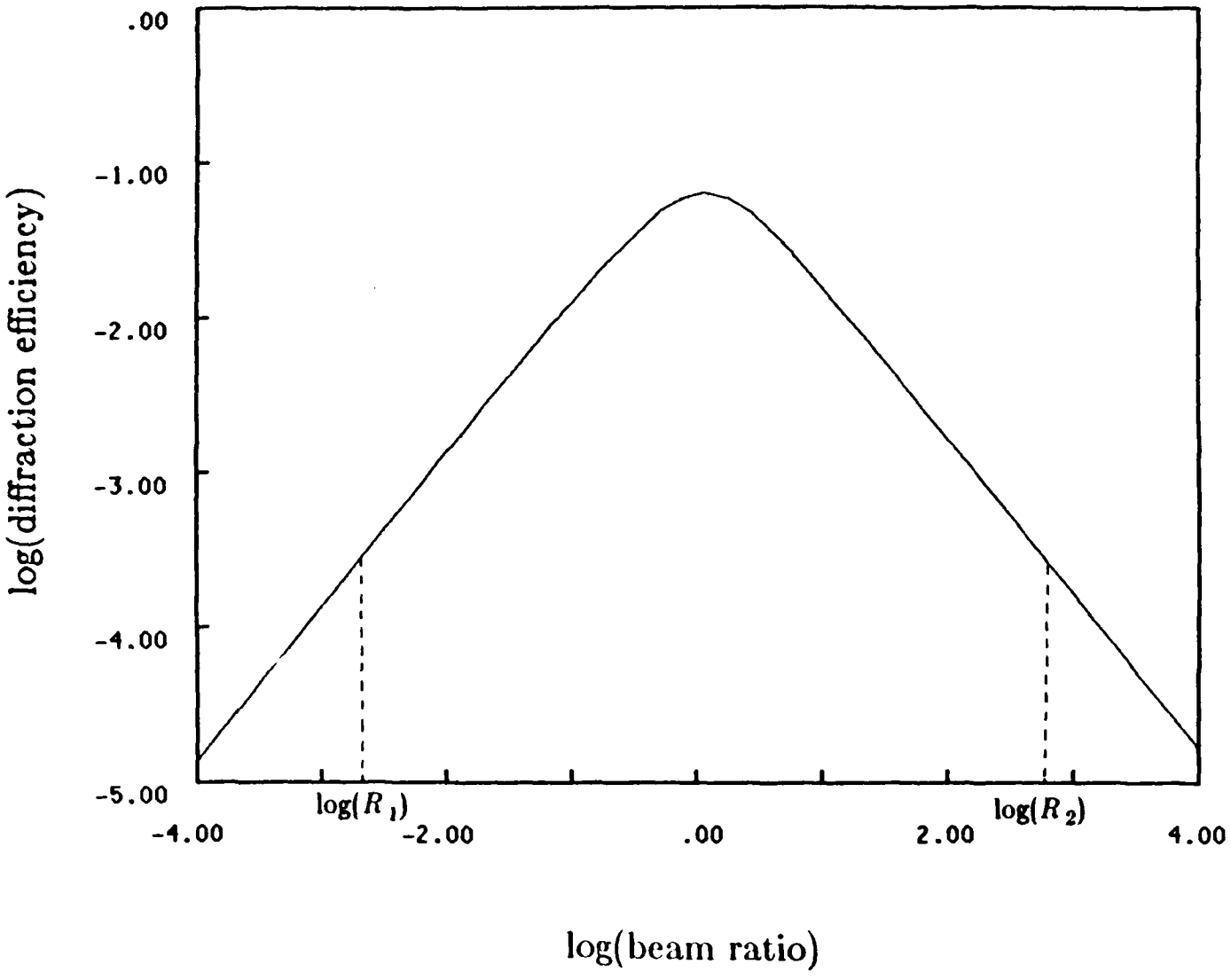
The coordinates of the seven defects, measured in units of numbers of squares and taking the center of the lower left square to be (0,0) are:

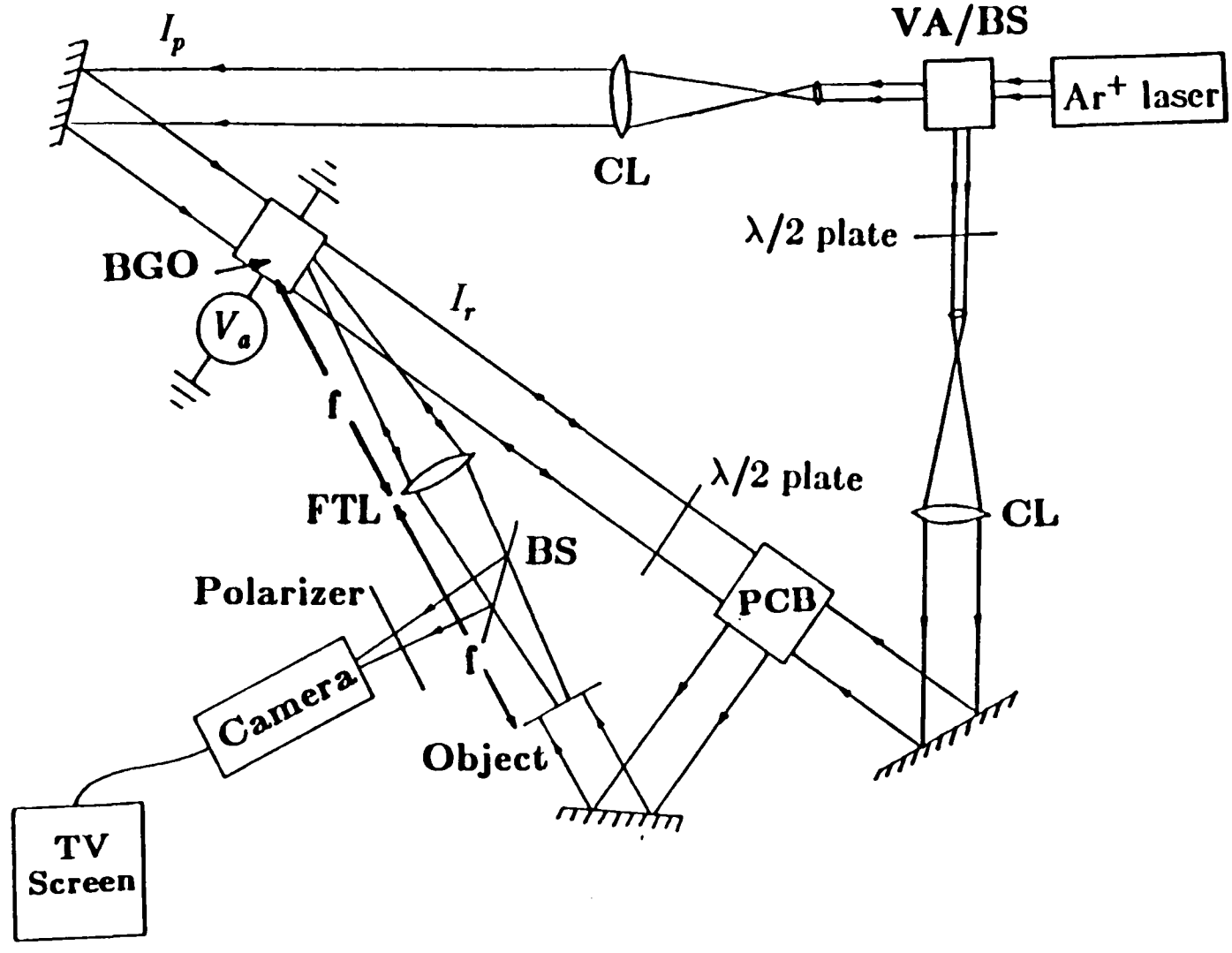
Defect size (μm^2)	Coordinates (hor,vert)
100 x 100	(22,7.5)
50 x 100	(10,13.5)
100 x 50	(12.5,25)
10 x 100	(15,20.5)
25 x 100	(24,27.5)
100 x 25	(25.5,22)
100 x 10	(25.5,15)

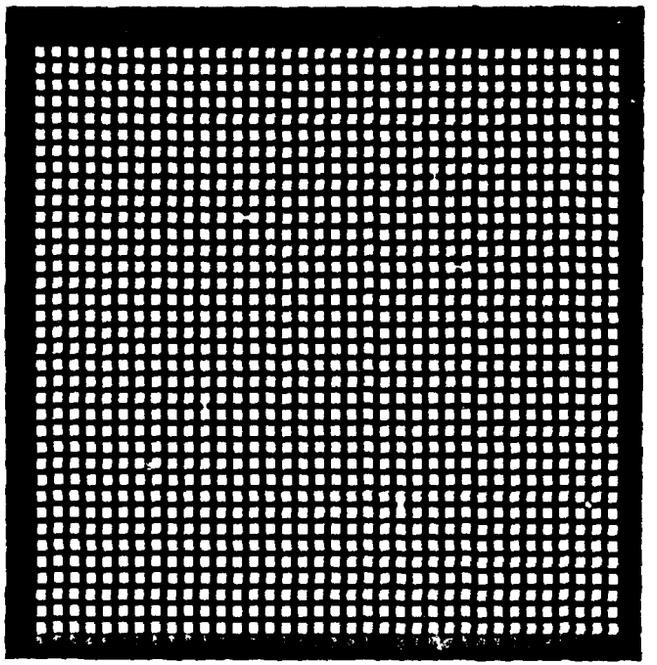
Fig. 4

Intensity line scan of 10 x 100 μm^2 defect

Graph illustrating the signal-to-noise ratio obtained for the smallest defect.


.00
-1.00
-2.00
-3.00
-4.00
-5.00
log(diffraction efficiency)
log(R 1)
log(R 2)
-4.00
-2.00
.00
2.00
4.00
log(beam ratio)


I_p
VA/BS
Ar^+ laser
CL
λ/2 plate
BGO
V_a
I_r
f
FTL
λ/2 plate
CL
BS
Polarizer
PCB
f
Camera
Object
TV
Screen



(a)

(b)

APPENDIX I✓

Wave-front inversion using a thin phase hologram: a computer simulation

Qizhi Cao and Joseph W. Goodman

a reprint from **Applied Optics**
volume 23, number 24, December 15, 1984

Wave-front inversion using a thin phase hologram: a computer simulation

Qizhi Cao and Joseph W. Goodman

It has been demonstrated previously that a thin phase hologram, recorded in a weak reference condition, is capable of inverting a complex field. Using a computer simulation of the properties of a thin phase hologram, we find the operating conditions and dynamic range for wave-front inversion. The conclusions of the simulation are used for designing an experiment to invert a circulant matrix and the results of the experiment well support the analysis.

I. Introduction

Wave-front inversion by means of holography has been studied by researchers for realizing image deblurring, and three types of optical deblurring filter have been developed.[1] One is the so-called sandwich structure filter which consists of a thin absorption hologram and an absorption transparency.[2,3] The absorption hologram is used to reconstruct an optical field representing the conjugate of the transfer function of a blurring system, while the absorption transparency has an amplitude transmittance proportional to the reciprocal of the squared modulus of the transfer function. Similar to the first one in the sense of still using an absorption hologram, the second type of deblurring filter[4] leaves out the absorption transparency from the sandwich structure by placing an absorption mask before the object field while the hologram is recorded. While the structure of this filter is simpler, it still requires a two-step process and has the same dynamic range as the previously described type.

The third type of deblurring filter, originally demonstrated by Ragnarsson[5] and further studied by Tichenor,[6] is formed with a single thin phase hologram, inverting the wave front by diffraction instead of absorption. This method, compared with the sandwich structured filter, is considerably simpler in construction; namely, it is a one-step method. In addition, it has a larger dynamic range.[1] Moreover, unlike a conventional hologram, Ragnarsson's deblurring holographic filter is recorded in the condition of a very weak reference beam (much weaker than the object beam).

Although the demonstration of Ragnarsson's method for wave-front inversion is quite impressive, the analytical explanation of the mechanism behind this method is not complete. To understand how it works, one has to adopt several postulations,[7] which may not always be true in practice.

In this paper we reformulate the analysis of this type of hologram, explicitly pointing out the inevitable phase distortion that accompanies the desired inversion. Then a computer simulation of this type of hologram is carried out based on the theoretical analysis. From the results of the computer simulation, some important quantitative information and instructive conclusions can be obtained; e.g., the dynamic range of the inversion with respect to the beam ratio; the relation between phase distortion and the brightness of the output, etc. These results are used for designing an experiment to invert a circulant matrix.

The method of inverting a circulant matrix by a coherent optical system has been described in our previous paper[8] (abbreviated as pre-paper). A brief review of the basic idea is given as the following. A circulant matrix is one for which each successive row is a simple circular shift of the row above by a single element. For example, in the matrix C below, the numbers 1, 2, 3, and 4 stand for four distinct elements; the organization of those elements in the circulant matrix is as follows:

$$C = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 4 & 1 & 2 & 3 \\ 3 & 4 & 1 & 2 \\ 2 & 3 & 4 & 1 \end{bmatrix}.$$

A remarkable property of circulant matrices is that they are diagonalized by the discrete Fourier transform (DFT). The resulting diagonal elements are the complex eigenvalues of the original matrix. By means of

The authors are with Stanford University, Department of Electrical Engineering, Stanford, California 94305.

Received 14 August 1984.
0003-6935/84/244575-13$02.00/0.

photographic film, a circulant matrix can be encoded on a film negative that consists of many cells with different transmittances representing the values of the matrix elements. After suitable modification, the coherent optical system performs the DFT of the circulant matrix presented at its input, generating an optical field representing the diagonalized matrix. If an inversion operation and an inverse DFT follow this step, we will finally obtain the inverse of the circulant matrix. While an experimental result for the diagonalized process has been presented in the previous paper, the next two steps, and in particular the inversion of the diagonalized matrix, will be discussed here.

In Sec. II, as a background, the characteistics of bleached photographic film are briefly reviewed, and the principle of generating a wave-front inversion by a thin phase hologram is discussed. Section III first illustrates the outline and flow chart of the computer simulation and then presents conclusions arising from the results of the simulation. While the regime of inverse reconstruction is our main concern, for comparison purpose the behavior of a conventional hologram is also briefly studied. In Sec. IV the experimental results which support the simulation are described: the measured intensity or the profile of the optical fields representing the inverse eigenvalue matrix and the inverse matrix are presented, respectively. Finally, Sec. V summarizes the conclusion of this work.

II. Background

2.1 Characteristics of Bleached Photographic Film

As discussed in the literature,[9,10] one of the most important photosensitive properties of photographic film can be illustrated by a plot of D, optical density, vs $\log E$, the logarithm of exposure, namely, the Hurter-Driffield curve (or briefly, the H-D curve). The density D is defined as

$$D = \log\left(\frac{1}{\tau}\right),$$

where τ is the intensity transmittance of a film after exposure and development. It is known that D so defined is proportional to the silver mass per unit area of the developed transparency. The exposure E is defined as

$$E = IT,$$

i.e., the product of the light intensity I to which the film is exposed and T, the time duration of the exposure.

Figure 1 illustrates a typical H-D curve for a developed negative film. When the exposure is below a certain level, the density is independent of exposure and equal to a minimum value referred to as gross fog. In the toe of the curve, density begins to increase with increasing exposure. There follows a region of the curve in which density is linearly proportional to logarithmic exposure; the slope of this linear region of the curve is called the film contrast, γ. Finally, the curve saturates in a region called the shoulder, and again there is no change in density with increasing exposure. In the

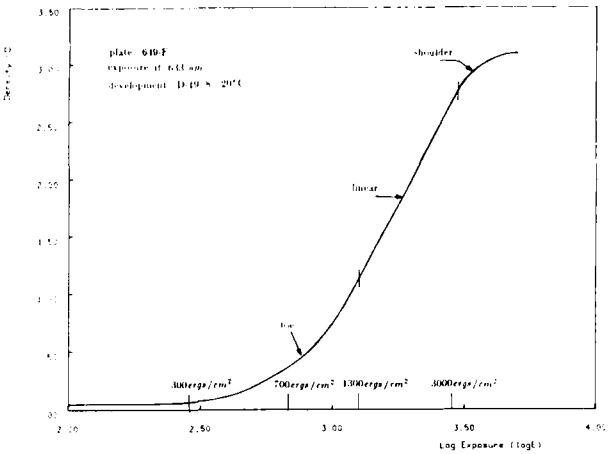


Fig. 1. Typical H-D curve.

linear region of the H-D curve, the relation between D and E can be written as

$$D = D_0 + \gamma \log E, \tag{1}$$

where D_0 is a constant determined by the intercept of the extended straight line region.

When a film negative is subjected to a bleaching process, the opaque metallic silver grains return to a transparent silver halide compound. The variation of concentration of silver salt result in a variation of refractive index and a surface relief. Therefore, the exposure variation across the film is now mapped to a phase modulation of transmitted light. It is reasonable to think that the concentration of silver salt is proportional to that of the original silver grains, and so to the photographic density. With this as a main assumption, the variation of phase and density can be proved to have a linear correspondence[11]; i.e.,

$$\frac{\Delta\phi}{\Delta D} = C, \tag{2}$$

where C is a constant depending on the salt polarizability, wavelength, etc. Its units are radians/density unit, abbreviated here as rad/den. Experiments carried out with plates of Agfa 10E75 and several bleaching chemicals[11] show that C lies in the range from 0.575 to 0.960 rad/den when the phase relief is eliminated with the aid of a liquid gate, and from 1.43 to 2.39 rad/den when the relief is included.

From Eq. (2) it follows that a plot of ϕ vs $\log E$ should be similar to the H-D curve. The $\phi - \log E$ curve, or equivalently, the H-D curve together with Eq. (2), basically characterizes a bleached photographic film.

To this point, we have tacitly assumed a lack of correlation between film response and the spatial frequency of an input. In practice, the density variations produced in the developed negative grow smaller as the frequency increases and essentially vanish beyond a limiting frequency. This property is conventionally characterized by a modulation transfer function [12] $M(f)$. However, the modulation transfer function of the films used remain flat within a large frequency range, see Fig. 2. Therefore, the frequency dependence can usually

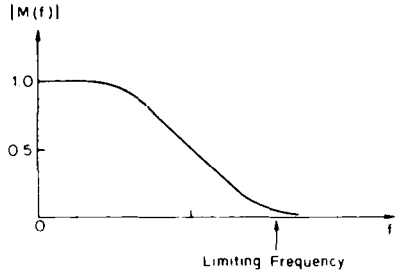


Fig. 2. Typical modulation function of film.

be neglected if the spatial frequencies of the input are constrained to be much lower than the limiting frequency.

2.2 Prinicple of Generating Wave-front Inversion by a Thin Phase Hologram

When a hologram is formed by an object beam, $O\exp(j\phi_o)$ and a reference beam, $R\exp(j\phi_r)$, the exposure on the plate is given by

$$\begin{aligned} E &= I \cdot T \\ &= \{|O|^2 + |R|^2 + OR^* \exp[j(\phi_o - \phi_r)] \\ &\quad + O^*R \exp[-j(\phi_o - \phi_r)]\} \cdot T, \end{aligned}$$

where T is the exposure time. Let $E_0 = |O|^2 \cdot T$, $E_r = |R|^2 \cdot T$, and $E_{or} = E_0 + E_r$; then

$$E = E_{or}[1 + v\cos(\phi_o - \phi_r)], \tag{3}$$

where

$$v = \frac{2\sqrt{E_0 E_r}}{E_o + E_r} \tag{4}$$

is the visibility of the fringe pattern. When the exposure E falls within the linear region of the H-D curve of the film, from Eqs. (1) and (2) it follows that the phase modulation is

$$\Delta\phi = C\gamma \log E.$$

Substitution of Eq. (3) for E yields

$$\Delta\phi = C\gamma \log|E_{or}[1 + v\cos(\phi_o - \phi_r)]|.$$

Then the amplitude transmittance of the phase hologram is

$$\begin{aligned} t &= \exp(j\Delta\phi) \\ &= \exp(jC\gamma \log E_{or} \cdot \exp|jC\gamma \log[1 + v\cos(\phi_o - \phi_r)]|. \end{aligned} \tag{5}$$

Now let us apply the weak reference condition; i.e., $E_r \ll E_o$; them from Eq. (4)

$$v \sim 2\sqrt{\frac{E_r}{E_o}} \ll 1,$$

$$\log[1 + v\cos(\phi_o - \phi_r)] \sim v\cos(\phi_o - \phi_r). \tag{6}$$

Therefore, from Eq. (5) it follows that

$$\begin{aligned} t &\approx \exp(jC\gamma \log E_{or}) \cdot \exp\{|jC\gamma v\cos(\phi_o - \phi_r)|\} \\ &\approx \exp(jC\gamma \log E_{or}) \cdot \sum_{k=-\infty}^{\infty} (j)^k J_k(C\gamma v)\exp|jk(\phi_o - \phi_r)|, \end{aligned} \tag{7}$$

where the expansion

$$\exp(jM\cos\alpha) = \sum_{k=-\infty}^{\infty} J_k(M)\exp(jk\alpha)(j)^k$$

is used in the derivation and $J_k(M)$ is the kth-order Bessel function of the first kind.

From (7) it can be seen that this phase hologram produces an infinite set of diffracted waves, each of which has the amplitude $J_k(C\gamma v)$. In addition, $C\gamma v$ can be very small because v is very small due to the weak reference condition and $C\gamma$ can be small, e.g., ~1, by controlling the chemical process. Therefore,

$$J_1(C\gamma v) \sim C\gamma v, \tag{8}$$

due to the nature of the Bessel function for small arguments.

Substitution of this formula for (7) shows that amplitudes of the ±1st orders of the diffracted waves from the hologram are proportional to v. Furthermore, when the hologram is illuminated by the conjugate of the original reference beam, the −1-order diffracted wave can be expressed as

$$\begin{aligned} t_{-1} &\approx \exp jC\gamma \log E_{or} \cdot C\gamma v \exp(-j\phi_o) \\ &\approx \exp(j\phi_d)\frac{|O||R|}{|O|^2 + |R|^2}\exp(-j\phi_o), \end{aligned} \tag{9}$$

where $\phi_d = C\gamma \log E_{or}$, and unimportant constants have been dropped. If the reference beam has uniform amplitude and the weak reference condition is applied again, (9) can be written as

$$t_{-1} \approx \begin{cases} \exp(j\phi_d)\dfrac{1}{|O|}\exp(-j\phi_o) & |O| \neq 0, \\ 0 & |O| = 0. \end{cases} \tag{10}$$

We now see that a pseudoinversion process is accomplished except for a phase distortion factor $\exp(j\phi_d)$.

III. Computer Simulation of a Thin Phase Hologram

The analytical solution in the previous section depicts the behavior of a thin phase hologram recorded in a weak reference condition. However, there are still some aspects of the problem that need to be clarified. For example, we saw that the weak reference is a key condition for an inverse reconstruction, but it is more important to know: what is the appropriate beam ratio in order to gain an acceptably good inversion? In addition, it can be seen that the phase distortion ϕ_d shown in (10) will be reduced if small C and γ are used. However, a decrease of C and γ results in a sacrifice of the diffraction efficiency or the brightness of output. Therefore, it will be very helpful to obtain a quantitative estimate of how these parameters are correlated and to determine how to make a compromise when an experiment is performed. These are the motivations for a computer simulation of a thin phase hologram. In this section, an outline of the computer simulation and the program flow chart is given. The conclusions coming from the results will be presented in the next section.

3.1. Outline and the Program Flow Chart of the Computer Simulation

We started with a typical H-D curve for the 649-F holographic plate shown in Fig. 1. This curve was obtained experimentally except for the shoulder that was later added to comply with theory. However, this will not hurt the generality of the discussion. The experiment for this curve was performed in the following conditions: exposure at wavelength 633 nm, and development for 8 min at 20°C in developer D-19. The curve is linear when exposure E is between ~1300 and 3000 ergs/cm^2 or 3.11–3.50 on the log scale, with a contrast γ of ~4.5. The toe portion ranges from 500 to 1300 ergs/cm^2 and the shoulder portion above 3000 ergs/cm^2. Later we will see that the behavior of this curve very much affects the properties of the first-order diffracted wave, particularly when an inverse reconstruction is concerned.

By using formula (2), a density variation is mapped to a phase variation. The constant C depends on the bleaching process as mentioned in the foregoing section and we chose the value 1 rad/den in the simulation.

Now let us consider a hologram formed by two plane waves (Fig. 3). After bleaching, the amplitude transmittance of the hologram has a pure periodic phase factor as shown in Eq. (5). $(\phi_o - \phi_r)$ can now be expressed as ωx, where x is the spatial coordinate on the axis perpendicular to the grating fringes (see Fig. 3); ω is the angular frequency of the grating and is determined by the angle between the two beams. The values of ωx can be anywhere between 0 and 2π, which determines the complex amplitude transmittance when E_{or}, C, and γ are known. Sampling the complex transmittance by sampling ωx uniformly and performing a FFT of the sampled transmittance, we are able to obtain its spectrum and extract information about the first-order diffracted wave, such as its modulus and its phase. Fixing the reference beam and varying the object beam to many different values, we can also find out the relation between the diffraction efficiencies and the phases of the first-order wave and its corresponding input object wave.

With all this information in hand, we can then characterize the resulting phase hologram by plotting eff1, the efficiency of the first order, vs E_o, with E_r as a parameter. A list of E_o, eff1, k, and α with a fixed C and E_r will also be produced, where k is the beam ratio, and α is the phase distortion. A program flow chart is shown in Table I.

Repeating the job illustrated in Table I twenty times and collecting the various eff1's, we are able to plot an eff1 vs E_o curve, etc.

3.2. Results

Although the main goal of our simulation is to find the regime for an inverse reconstruction, we will report results of both conventional and and inversion cases. A conventional hologram usually is expected to generate an image similar to the original object during reconstruction; i.e., the reconstructed image field is expected to be linearly proportional to the input object field.

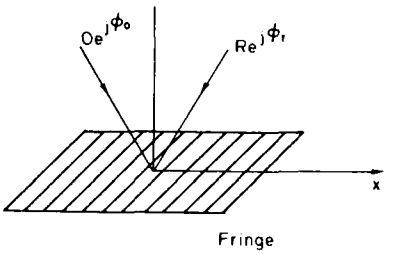
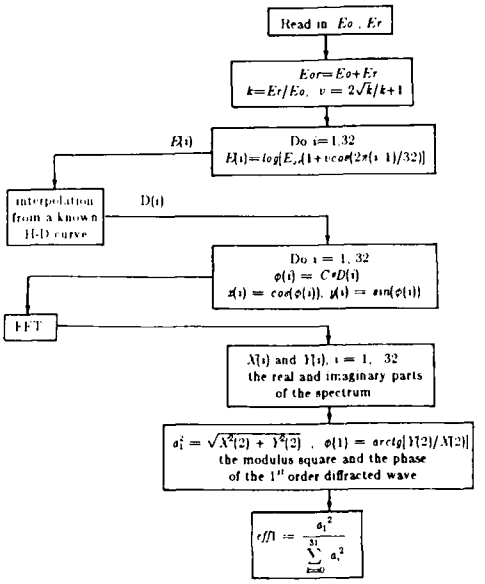


Fig. 3. Hologram formed by two plane waves.

Table I. Program Flow Chart



This requires exposure in the condition $E_r > E_o$. In what follows, we will present the results for both cases. All the tables and figures illustrating the results appear at the end of this section.

A. $E_r > E_o$—*Conventional Hologram*

The data listed in Tables II(a)–(c) together with the three graphs shown in Figs. 4(a)–(c) are the results of the three sample trails. In each of the tables there are five columns containing the following data: (1) the energy of the object beam, E_o; (2) the percentage efficiency of the first-order diffracted wave, eff1(%); (3) the beam ratio k; (4) the visibility v; and (5) α, the radian phase angle associated with the first-order diffracted wave. At the top of the tables, the coefficient C and the fixed reference beam energy E_r are given.

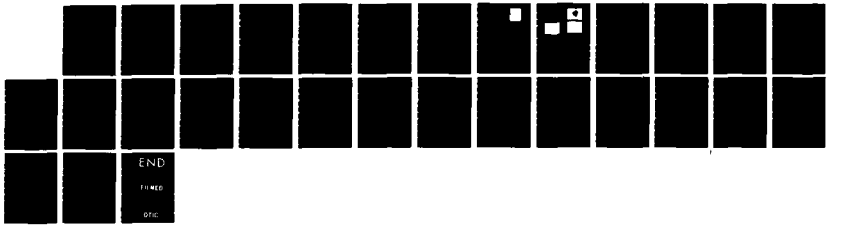
From the data shown, some observations and conclusions can be made.

1. Linear reconstruction occurs when the average exposure E_{or} falls in the toe portion of the H-D curve.

From Fig. 4(b) it can be seen that the reconstruction will be very linear when the reference beam is 600 ergs/cm^2 and the object beam is not over 200 ergs/cm^2. That is to say, the optimum average exposure is ~700 ergs/cm^2, which falls in the toe portion of the H-D curve (Fig. 1) and is just at the center of the t–E curve. This

AD-A162 272 OPTICAL COMPUTING RESEARCH(U) STANFORD UNIV CA 2/2
INFORMATION SYSTEMS LAB J W GOODMAN ET AL 01 JUN 85
ISL-L722-10 AFOSR-TR-85-1037 AFOSR-83-0166

UNCLASSIFIED F/G 9/3 NL



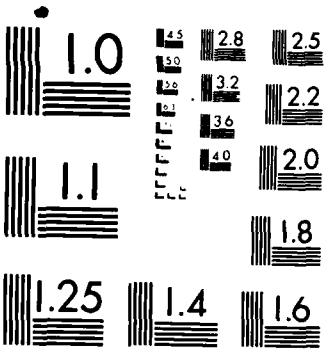


Table II. Output Data from Simulation: (a) $E_r > E_o, C = 1$ rad/den, E_r = 500 ergs/cm²; (b) $E_r > E_o, C = 1$ rad/den, E_r = 600 ergs/cm²; (c) $E_r > E_o, C = 1$ rad/den, E_r = 700 ergs/cm²

E_o (ergs/cm²)	eff1(%)	k	v	α (radian)
1	.01	500.00	.09	1.76
10	.35	50.00	.28	1.78
20	.69	25.00	.38	1.80
30	1.01	16.67	.46	1.82
40	1.34	12.50	.52	1.84
50	1.67	10.00	.57	1.86
60	2.01	8.33	.62	1.88
70	2.39	7.14	.66	1.90
80	2.79	6.25	.69	1.92
100	3.63	5.00	.75	1.96
120	4.50	4.17	.79	1.99
140	5.41	3.57	.83	2.03
160	6.42	3.13	.86	2.07
180	7.43	2.78	.88	2.11
200	8.46	2.50	.90	2.14
220	9.48	2.27	.92	2.17
240	10.47	2.08	.94	2.20
260	11.46	1.92	.95	2.23
280	12.41	1.79	.96	2.26
300	13.38	1.67	.97	2.29

(b)

E_o (ergs/cm²)	eff1(%)	k	v	α (radian)
1	.00	600.00	.08	1.85
10	.56	60.00	.25	1.87
20	1.08	30.00	.35	1.89
30	1.60	20.00	.43	1.90
40	2.16	15.00	.49	1.93
50	2.73	12.00	.53	1.95
60	3.33	10.00	.57	1.97
70	3.92	8.57	.61	2.00
80	4.51	7.50	.64	2.02
100	5.72	6.00	.70	2.06
120	6.96	5.00	.75	2.10
140	8.18	4.29	.78	2.14
160	9.38	3.75	.82	2.18
180	10.55	3.33	.84	2.22
200	11.70	3.00	.87	2.25
220	12.80	2.73	.89	2.28
240	13.86	2.50	.90	2.31
260	14.89	2.31	.92	2.34
280	15.89	2.14	.93	2.37
300	16.83	2.00	.94	2.39

(c)

E_o (ergs/cm²)	eff1(%)	k	v	α (radian)
1	.07	700.00	.08	1.95
10	.73	70.00	.24	1.97
20	1.53	35.00	.33	1.99
30	2.38	23.33	.40	2.01
40	3.21	17.50	.45	2.01
50	4.01	14.00	.50	2.06
60	4.85	11.67	.51	2.08
70	5.68	10.00	.57	2.11
80	6.48	8.75	.61	2.13
100	8.08	7.00	.66	2.17
120	9.59	5.83	.71	2.21
140	10.99	5.00	.75	2.25
160	12.32	4.38	.78	2.29
180	13.59	3.89	.81	2.32
200	14.79	3.50	.83	2.35
220	15.92	3.18	.85	2.38
240	16.99	2.92	.87	2.41
260	18.01	2.69	.89	2.44
280	18.97	2.50	.90	2.46
300	19.88	2.33	.92	2.49

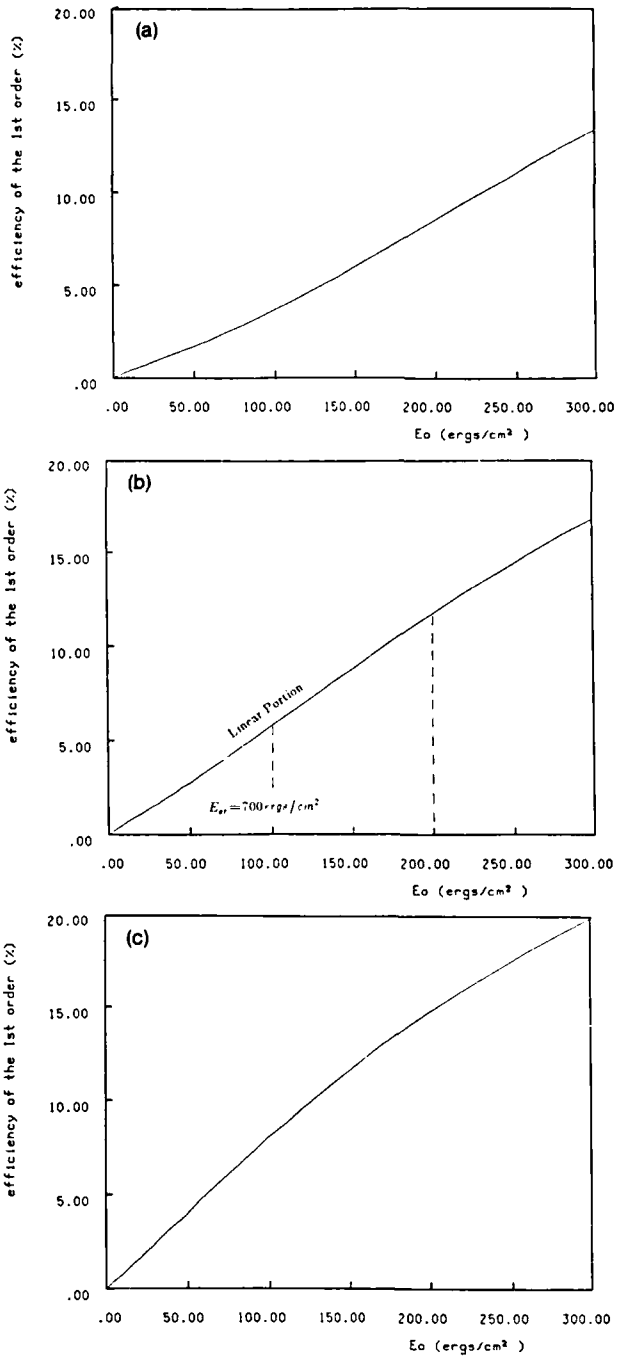


Fig. 4. eff1 vs E_0 for $E_r > E_o$: (a) (E_r = 500 ergs/cm², C = 1 rad/den); (b) (E_r = 600 ergs/cm², C = 1 rad/den); (c) (E_r = 700 ergs/cm², C = 1 rad/den).

is the same condition that would normally be required when an absorption hologram is recorded.[14]

2. Linear reconstruction occurs when the beam ratio is not less than 3/1.

As pointed out above, linear reconstruction takes place when the reference beam is 600 ergs/cm² and the object beam not over 200 ergs/cm². This fact can be stated in terms of beam ratio, i.e., linear reconstruction occurs if the beam ratio is not less than 3 or the visibility is not higher than 0.87 [Table II(b)]. This is also similar to the condition required for a good absorption hologram.[14]

3. The diffraction efficiency is higher than would be obtained with an absorption hologram.

The diffraction efficiency will be up to ~10% [Table II(b)] for a linear reconstructed image. This is much higher than that for an absorption hologram.

4. The phase distortion is very small.

When the hologram is formed by two plane waves, from Eq. (3) the log exposure can be expressed as

Table III. Output Data from Simulation: (a) $E_r \ll E_o, C = 1$ rad/den, $E_r = 1$ ergs/cm²; (b) $E_r \ll E_o, C = 1$ rad/den, $E_r = 10$ ergs/cm²; (c) $E_r \ll E_o, C = 1$ rad/den, $E_r = 100$ ergs/cm²; (d) $E_r \ll E_o, C = 1$ rad/den, $E_r = 1.3$ ergs/cm²

(a)

$C = 1$ rad/den, $E_r = 1$ erg/cm²					
E_o (ergs/cm²)	eff1(%)	k	v	α (radian)	eff1·E_o
500	.041	1/500	.09	1.76	20.70
700	.072	1/700	.08	1.95	50.70
900	.156	1/900	.07	2.18	140.18
1100	.226	1/1100	.06	2.46	248.36
1300	.271	1/1300	.06	2.75	352.08
1400	.254	1/1400	.05	2.89	355.99
1500	.237	1/1500	.05	3.02	356.05
1600	.223	1/1600	.05	-3.14	356.11
1700	.210	1/1700	.05	-3.03	356.15
1800	.198	1/1800	.05	-2.92	356.19
1900	.187	1/1900	.05	-2.82	356.23
2000	.178	1/2000	.04	-2.72	356.27
2200	.162	1/2200	.04	-2.54	356.32
2400	.148	1/2400	.04	-2.37	356.37
2600	.137	1/2600	.04	-2.22	356.41
2800	.127	1/2800	.04	-2.08	356.45
2900	.123	1/2900	.04	-2.02	356.46
3000	.119	1/3000	.04	-1.95	356.48
3200	.050	1/3200	.04	-1.85	158.54
3400	.029	1/3400	.03	-1.78	99.75

(b)

$C = 1$ rad/den, $E_r = 10$ ergs/cm²					
E_o (ergs/cm²)	eff1(%)	k	v	α (radian)	eff1·E_o
500	.353	1/50	.28	1.78	176.41
700	.727	1/70	.24	1.97	508.92
900	1.439	1/90	.21	2.21	1295.13
1100	2.275	1/110	.19	2.47	2502.58
1300	2.554	1/130	.17	2.75	3320.53
1400	2.461	1/140	.17	2.89	3445.85
1500	2.323	1/150	.16	3.02	3484.60
1600	2.181	1/160	.16	3.14	3489.85
1700	2.056	1/170	.15	-3.03	3494.49
1800	1.944	1/180	.15	-2.92	3498.60
1900	1.843	1/190	.14	-2.82	3502.289
2000	1.753	1/200	.14	-2.72	3505.61
2200	1.596	1/220	.13	-2.54	3511.34
2400	1.465	1/240	.13	-2.38	3516.12
2600	1.354	1/260	.12	-2.23	3520.18
2800	1.258	1/280	.12	-2.09	3523.67
2900	1.139	1/290	.12	-2.03	3302.49
3000	.954	1/300	.12	-1.97	2862.33
3200	.597	1/320	.11	-1.88	1910.09
3400	.290	1/340	.11	-1.80	987.06

(c)

$C = 1$ rad/den, $E_r = 100$ ergs/cm²					
E_o (ergs/cm²)	eff1(%)	k	v	α (radian)	eff1·E_o
500	3.63	1/5	.75	1.96	1817.38
700	8.08	1/7	.66	2.17	5658.59
900	12.33	1/9	.60	2.39	11097.65
1100	15.12	1/11	.55	2.61	16632.14
1300	16.62	1/13	.52	2.81	21614.64
1400	16.92	1/14	.50	2.91	23696.16
1500	16.96	1/15	.48	3.02	25450.68
1600	16.77	1/16	.47	3.12	26835.01
1700	16.39	1/17	.46	-3.06	27863.24
1800	15.92	1/18	.45	-2.96	28602.44
1900	15.38	1/19	.44	-2.86	29233.93
2000	14.79	1/20	.43	-2.76	29592.41
2200	13.61	1/22	.41	-2.58	29937.41
2400	11.98	1/24	.39	-2.44	28762.42
2600	10.18	1/26	.38	-2.31	26458.07
2800	8.30	1/28	.36	-2.21	23233.68
2900	7.39	1/29	.36	-2.16	21445.99
3000	6.52	1/30	.35	-2.12	19563.65
3200	4.94	1/32	.34	-2.04	15810.37
3400	3.57	1/34	.33	-1.97	12147.03

(d)

$C = 1$ rad/den, $E_r = 1.3$ ergs/cm²					
E_o (ergs/cm²)	eff1(%)	k	v	α (radian)	eff1·E_o
500	.54	1/385	.10	1.76	26.91
700	.094	1/538	.09	1.95	65.98
900	.201	1/692	.08	2.18	180.76
1100	.296	1/846	.07	2.46	326.08
1300	.351	1/1000	.06	2.75	456.12
1400	.330	1/1077	.06	2.89	462.43
1500	.308	1/1152	.06	3.02	462.54
1600	.289	1/1230	.06	-3.14	462.63
1700	.272	1/1308	.06	-3.03	462.71
1800	.257	1/1385	.05	-2.92	462.78
1900	.244	1/1462	.05	-2.82	462.84
2000	.231	1/1538	.05	-2.72	462.89
2200	.210	1/1692	.05	-2.54	462.99
2400	.193	1/1846	.05	-2.37	463.07
2600	.178	1/2000	.04	-2.22	463.14
2800	.165	1/2154	.04	-2.08	463.20
2900	.160	1/2231	.04	-2.02	463.23
3000	.154	1/2308	.04	-1.95	463.26
3200	.067	1/2462	.04	-1.85	213.54
3400	.038	1/2615	.04	-1.78	129.66

$$\log E = \log E_{or} + \log(1 + v\cos\omega x),$$

where the first term is a logarithmic exposure bias for a fixed E_o and the second term is a log exposure excursion riding on the bias. When the excursion of the log exposure falls within the toe portion of the H-D curve, the resulting phase modulation can be expressed as

$$\phi = \phi(E_{or}) + \Delta\phi(v,\omega x), \tag{11}$$

where $\phi(E_{or})$, a function of E_{or}, is independent of spatial variables for a fixed E_{or}, and $\Delta\phi(v,\omega x)$ is a phase excursion with a high spatial frequency, Δ denoting the nonlinear mapping of exposure to phase in the toe portion. From Eq. (11) it follows that

$$t \sim \exp[j\phi(E_{or})] \cdot \exp[j\Delta\phi(v,\omega x)],$$

where $\exp[j\phi(E_{or})]$ is again a constant, but $\exp[i\text{-}\Delta\phi(v,\omega x)]$ is a function of v and x with a period of $(2\pi)/\omega$, its Fourier series representing a set of diffracted waves. The coefficients of this Fourier series may not be real in general, possibly generating some additional phase too. However, if the nonlinearity of the toe compensates for the logarithm of the exposure appropriately, $\Delta\phi$ could be approximately cosinusoidal, in which case the Fourier series of $\exp[j\Delta\phi(v,\omega x)]$ has real coefficients. In this case, this phase excursion term would not introduce a large dependence of phase on v and the extra phase involved in the transmittance is mainly due to $\phi(E_{or})$. Now if a complex object waveform consists of many object beams, $\phi(E_{or})$ becomes no longer constant but varies as E_o varies, causing a phase distortion of the reconstructed waveform. Nevertheless, for the case of a conventional phase hologram, because E_r is stronger than E_o, the variation of E_{or} is relatively small, finally leading to a small phase distortion.

From Table II(b) it can be seen that the total amount of change of $\alpha(\Delta\alpha)$ is indeed small (<0.4 rad for the entire dynamic range) and can be neglected.

B. $E_r \ll E_o$—*Inverse Reconstruction*

The simulations in the condition of $E_r \ll E_o$ (again E_r is fixed) were done with three different values of C, i.e., $C = 1$, 0.5, or 0.1 rad/den. Similar to the conventional case, the simulation is conducted with a fixed coefficient C and fixed $E_r(\ll E_o)$. E_o now varies over a large scale (e.g., 500–3400 ergs/cm²), and then a list of data with an associated graph is produced by means of the computer. Tables III(a)–(d) show the four lists of data obtained when $C = 1$ rad/den and $E_r = 1$, 10, 100, and 1.3 ergs/cm², respectively. Also the four graphs related to the four trials are shown in Figs. 5(a)–(c).

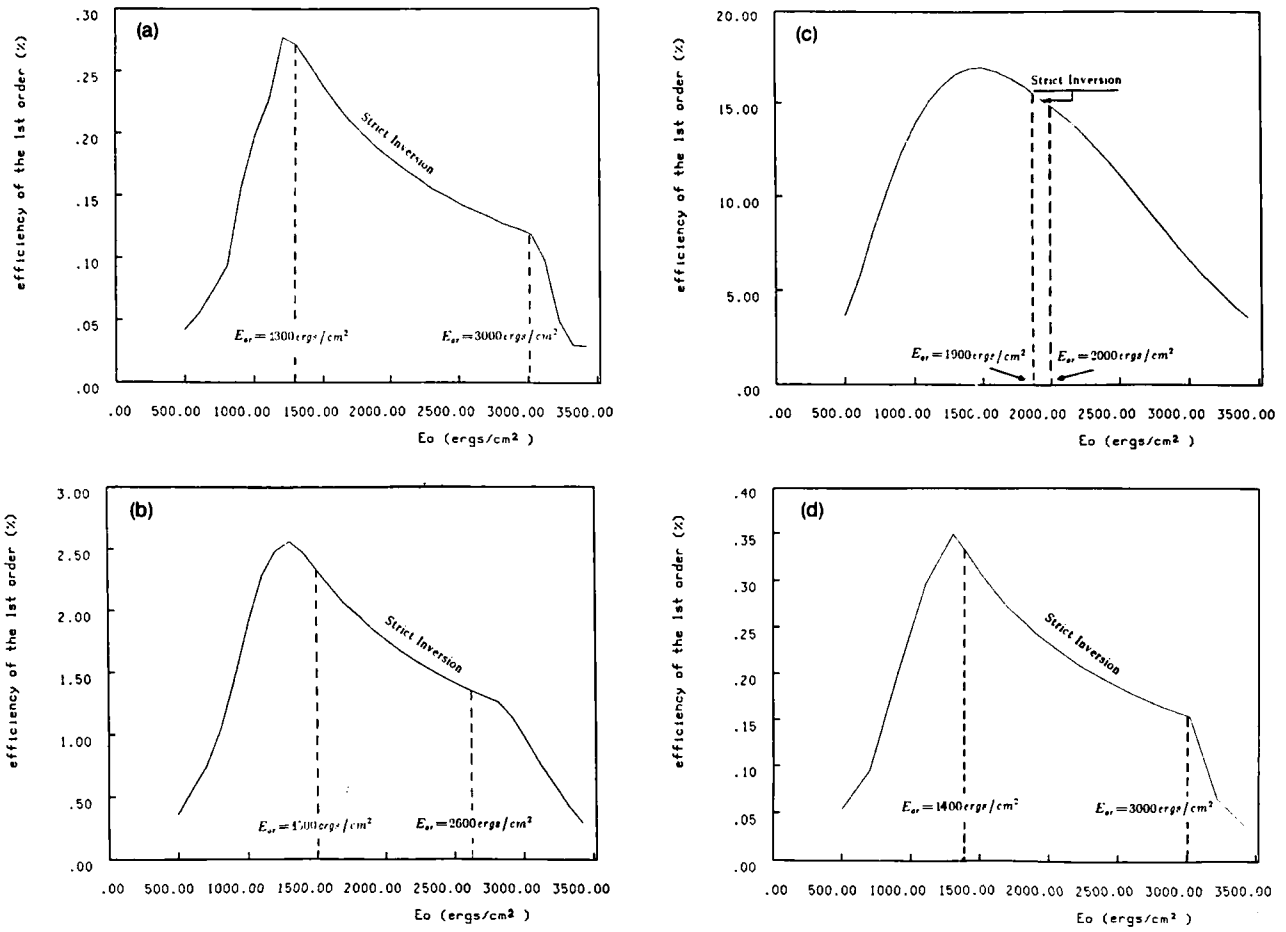


Fig. 5. eff1 vs E_o for $E_r \ll E_o$: (a) (E_r = 1 erg/cm^2,C = 1 rad/den); (b) (E_r = 10 ergs/cm^2,C = 1 rad/den); (c) (E_r = 100 ergs/cm^2,C = 1 rad/den); (d) (E_r = 1.3 ergs/cm^2,C = 1 rad/den).

The trial with E_r = 1.3 ergs/cm^2, in addition to that with E_r = 1 ergs/cm^2, was done to examine the changes of results due to small changes of E_r; particularly to study how the types of inversion depend on the beam ratio. This will become clearer later. For the cases of C = 0.5 or 0.1 rad/den, three trials were done for the three different combinations of C and E_r: E_r = 1 erg/cm^2 with C = 0.5 or 0.1 rad/den and E_r = 2 ergs/cm^2 with C = 0.1 rad/den. The results are shown in Tables IV(a)–(c) and Figs. 6(a)–(c). The columns in these lists are the same as in Table II, except that an extra column eff1 · E_o appears here to display values of the products of eff1 and E_o. A perfect inversion process would result in a constant value of this product.

From all the data presented, the following conclusions can be drawn:

1. The Regime for Inverse Reconstruction

Even though, as discussed previously, a thin phase hologram can peform an inverse reconstruction approximately, it is still useful to introduce criteria for strict and approximate inversion from a practical point of view. In the following context, we define strict inversion to mean that eff1 · E_o ~ const within an acceptable tolerance, e.g., ~1%. On the other hand, our definition for an approximate inversion is the regime where the tolerance of eff1 · E_o exceeds ~1% or more wherever the output brightness (eff1) drops as the input brightness (E_o) rises.

(1) Strict Inversion Occurs in the Straight Line Portion of the H-D Curve when the Beam Ratio k is Very Small.

Going through the data listed in Table III(a) and the corresponding Fig. 5(a), we see that the output [eff1(%)] first increases and then decreases as E_o goes up. The turning point for the start of inversion is at 1300 ergs/cm^2. From E_o = 1300–3000 ergs/cm^2, where the beam ratio k is from 1/1300 to 1/3000, eff1 · E_o remains constant with a tolerance of 1.1%; i.e., a strict inversion happens in this region. Again comparing with the H-D curve shown in Fig. 1, we find that in this trial strict inversion occurs when E_{or}(~E_o) is just within the linear portion of the H-D curve and its dynamic range covers the entire linear portion of the curve. Raising the reference beam to E_r = 1.3 ergs/cm^2 so that the beam ratio k varies from 1/1000 to 1/2308 while E_{or} is within the linear region of the H-D curve, as shown in Table III(d), we see that the tolerance of eff1 · E_o over this region rises to 1.5%(>1%). This implies that strict inversion

Table IV. Output Data from Simulation: (a) $E_r \ll E_o, C = 0.5$ rad/den, $E_r = 1$ erg/cm²; (b) $E_r \ll E_o, C = 0.1$ rad/den, $E_r = 1$ erg/cm²; (c) $E_r \ll E_o, C = 0.1$ rad/den, $E_r = 2$ ergs/cm²

$C = 0.5$ rad/den, $E_r = 1$ erg/cm²					
E_o (ergs/cm²)	eff1(%)	k	v	α (radian)	eff1·E_o
500	.010	1/500	.09	1.67	5.18
700	.018	1/700	.08	1.76	12.68
900	.039	1/900	.07	1.87	35.09
1100	.057	1/1100	.06	2.01	62.20
1300	.068	1/1300	.06	2.16	88.20
1400	.061	1/1400	.05	2.23	89.17
1500	.059	1/1500	.05	2.30	89.17
1600	.056	1/1600	.05	2.36	89.18
1700	.052	1/1700	.05	2.41	89.18
1800	.050	1/1800	.05	2.47	89.18
1900	.047	1/1900	.05	2.52	89.18
2000	.045	1/2000	.04	2.57	89.19
2200	.041	1/2200	.04	2.66	89.19
2400	.037	1/2400	.04	2.74	89.19
2600	.034	1/2600	.04	2.82	89.19
2800	.032	1/2800	.04	2.89	89.20
2900	.031	1/2900	.04	2.92	89.20
3000	.030	1/3000	.04	2.95	89.20
3200	.012	1/3200	.01	3.00	39.05
3400	.007	1/3400	.03	3.04	24.91

(b)

$C = 0.1$ rad/den, $E_r = 1$ erg/cm²					
E_o (ergs/cm²)	eff1(%)	k	v	α (radian)	eff1·E_o
500	.000	1/500	.09	1.59	.21
700	.001	1/700	.08	1.61	.51
900	.002	1/900	.07	1.63	1.40
1100	.002	1/1100	.06	1.66	2.49
1300	.003	1/1300	.06	1.69	3.53
1400	.003	1/1400	.05	1.70	3.57
1500	.002	1/1500	.05	1.72	3.57
1600	.002	1/1600	.05	1.73	3.57
1700	.002	1/1700	.05	1.74	3.57
1800	.002	1/1800	.05	1.75	3.57
1900	.002	1/1900	.05	1.76	3.57
2000	.002	1/2000	.04	1.77	3.57
2200	.002	1/2200	.04	1.79	3.57
2400	.001	1/2400	.04	1.80	3.57
2600	.001	1/2600	.04	1.82	3.57
2800	.001	1/2800	.04	1.83	3.57
2900	.001	1/2900	.04	1.84	3.57
3000	.001	1/3000	.04	1.85	3.57
3200	.000	1/3200	.01	1.86	1.59
3400	.001	1/3400	.03	1.86	1.00

(c)

$C = 0.1$ rad/den, $E_r = 2$ ergs/cm²					
E_o (ergs/cm²)	eff1(%)	k	v	α (radian)	eff1·E_o
500	.001	1/250	.13	1.59	.40
700	.001	1/350	.11	1.61	1.02
900	.003	1/450	.09	1.63	2.76
1100	.005	1/550	.09	1.66	5.10
1300	.005	1/650	.08	1.69	7.01
1400	.005	1/700	.08	1.70	7.14
1500	.005	1/750	.07	1.72	7.14
1600	.004	1/800	.07	1.73	7.14
1700	.004	1/850	.07	1.74	7.14
1800	.004	1/900	.07	1.75	7.14
1900	.004	1/950	.06	1.76	7.14
2000	.004	1/1000	.06	1.77	7.14
2200	.003	1/1100	.06	1.79	7.14
2400	.003	1/1200	.06	1.80	7.14
2600	.003	1/1300	.06	1.82	7.14
2800	.003	1/1400	.05	1.83	7.14
2900	.002	1/1450	.05	1.84	7.14
3000	.002	1/1500	.05	1.85	7.14
3200	.001	1/1600	.05	1.86	3.45
3400	.001	1/1700	.05	1.86	1.96

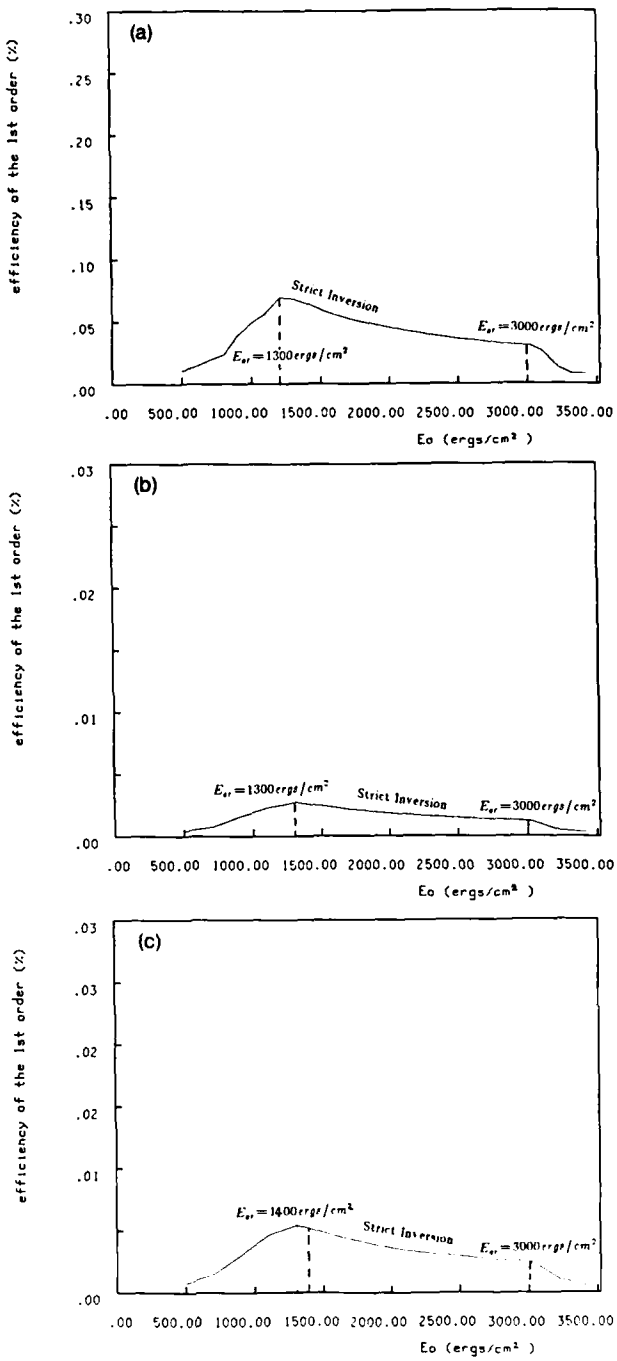


Fig. 6. eff1 vs E_o for $E_r \ll E_o$ and smaller C: (a) ($E_r = 1$ erg/cm², $C = 0.5$ rad/den); (b) ($E_r = 1$ erg/cm², $C = 0.1$ rad/den); (c) ($E_r = 2$ erg/cm², $C = 0.1$ rad/den).

does not hold over the entire linear region of the H-D curve but only over a part of it, i.e., from 1400 to 3000 ergs/cm², over which the tolerance is <1%. When $E_r = 10$ ergs/cm², the dynamic range over which strict inverse reconstruction holds drops to between 1500- and 2600 ergs/cm² with a beam ratio k from 1/150 to 1/260. This reduction continues until this range is just a small segment of the column, as in the case of $E_r = 100$ ergs/cm², where strict inversion remains at best only from $E_o = 1900$ to 2000 ergs/cm². With such a small dynamic range, this strict inversion is, in fact, useless. Therefore, it may be more realistic to claim that strict inversion occurs only when the total exposure is in the straight line portion of the H-D curve and the beam ratio is so small ($E_r \ll E_o$) that the maximum dynamic range can be achieved. Quantitatively, in this example the beam ratio is smaller than 1/1300 for a strict inversion. That is to say, with this number, the approximations [Eqs. (6) and (8)] introduced into the theoretical analysis in the foregoing section are valid in the sense of producing a strict inversion over a maximum dynamic range. This

number, as we will demonstrate later, has general significance, independent of the type of film or chemical process to a certain extent, and will be considered as an important reference for our experiments.

(2) Approximate Inversion Occurs in the Straight Line Portion Plus the Shoulder Portion of the H-D Curve.

According to the definition, an approximate inversion can be realized throughout the whole H-D curve, except for the toe portion. For example, in Fig. 5(a) approximate inversion for the case of E_r = 1 erg/cm^2 covers the whole range from 1300 to 3400 ergs/cm^2, but there is a steep drop prominent at about E_o = 3000 ergs/cm^2, which corresponds to the boundary between the straight line portion and the saturation portion of the H-D curve. As E_r increases, the dynamic range of the strict inversion is reduced, but the curve becomes smoother and the abrupt drop disappears when E_r = 100 ergs/cm^2. The dynamic range of the approximate inversion for E_r = 100 ergs/cm^2 is from ~1500 to 3400 ergs/cm^2.

(3) The Dynamic Range for Inverse Reconstruction is Determined Only by the Extent of the H-D Curve.

Investigating these two kinds of inversion, it becomes clear that the regime of inverse reconstruction is located in the linear region or the linear region plus the shoulder of a H-D curve. It follows that the dynamic range of the object energy is determined by the extent of the linear region and the shoulder of the H-D curve. In our case, the dynamic range is ~2:1 for a strict inversion, and ~2.6:1 for an approximate inversion. These numbers correspond to a total exposure region from 1300 to 3000 ergs/cm^2 or 1300 to 3400 ergs/cm^2, respectively.

Recall that a thin absorption hologram has linear reconstruction only when the total exposure is within the toe portion of the H-D curve. Since the region of the toe is much smaller than the linear region of the H-D curve, we see why a sandwich deblurring filter, which consists of a thin absorption hologram as mentioned at the beginning of this section, has a smaller dynamic range than Ragnarsson's filter, which is constructed by a thin phase hologram in the weak reference condition.

Inverse reconstruction will never occur when E_o is in the toe portion of a H-D curve. This can be explained qualitatively as follows. It can be seen that the efficiency (eff1) is proportional to the phase excursion, which is a function of either the excursion of the input E_o or the slope of the H-D curve. In the toe portion, the slope of the H-D curve continuously rises, causing the phase excursion to increase as E_o increases. Even though, at the same time, the reduced beam ratio and visibility due to the increase of E_o tend to suppress phase excursion, this effect cannot compete with the strength of the slope increase, so that finally the phase excursion goes up as E_o rises, and so does the diffraction efficiency.

2. Efficiency of the Inverse Reconstruction is Very Small for Conditions Yielding a Strict Inversion but Higher for Conditions Yielding an Approximate Inversion.

(1) In the straight line portion of the H-D curve, the slope of the curve remains constant and the efficiency rests solely on the square of the visibility, which is very small. Therefore, the efficiency is very small for strict inversion. It is shown from Table III(a) that the efficiency in this case is only 0.27–0.12%.

Table V. Efficiency of Two Types of Inversion

Linear Region of H-D Curve : 1300–3000 ergs/cm^2, C = 1 radians/unit density				
E_r	E_o	$k = \frac{E_r}{E_o}$	The Type of Inversion	eff1(%)
1	1300-3000	1/1300 - 1/3000	strict	0.12-0.27
10	1300-3400	1/130 - 1/340	approx.	0.29-2.6
100	1600-3400	1/16 - 1/34	approx.	3.57-16.8

(2) When E_r increases, the strict inversion changes to approximate inversion. The diffraction efficiency increases accordingly, e.g., ~3.57–16.8% when E_r = 100 ergs/cm^2 and E_o from 3400 to 1500 ergs/cm^2 [Table III(c)]. Thus, the efficiency and the degree of inversion are two contradictory parameters and some compromise must be made in practice. The comparison of efficiencies for the two types of inversion is listed in Table V.

3. Phase Distortion

Unlike the situation for a conventional phase hologram, the phase distortion becomes more severe for an inverse reconstruction phase hologram. For example, the difference of the phase angle, $\Delta\alpha$, for the whole dynamic range of the object is ~1.58 rad when C = 1 rad/den and E_r = 1 erg/cm^2 [Table III(a)] and remains around this number for the other cases. This is just the phase distortion ϕ_d we predicted in the previous section. It can be seen that this larger ϕ_d, as an intrinsic drawback, is due to the larger movement of the working point (total exposure) on the H-D curve when this type of hologram is formed. Yet, it can be much improved if the linear portion of the H-D curve has a very small slope (γ) and the constant C is smaller.

4. Effect of the Density–Phase Conversion Constant C

Tables IV(a)–(c) and the related Figs. 6(a) and (b) present the data of three trials in the conditions of E_r = 1 erg/cm^2 with C = 0.5 rad/den or 0.1 rad/den, and E_r = 2 ergs/cm^2 with C = 0.1 rad/den. From the data in the first two trials, we see that the smaller C values cause a reduction of the phase variation, which includes a slowly varying part and a high frequency excursion. As a result, the values of $\phi_d(\Delta\alpha)$ and eff1 are all reduced. For example, ϕ_d and eff1 become ~0.8 rad and 0.05% when C = 0.5 rad/den [Table IV(a)] and 0.15 rad and 0.002% for C = 0.1 rad/den [Table IV(b)]. In addition, to examine the regime of strict inversion for the case of smaller C values, Table VI is given, listing the tolerance of eff1 · E_0 for these three trials when E_o varies across the maximum dynamic range (linear region of the H-D curve. From the data in Table VI, we see that the variations of eff1 · E_o for the first two trials (C = 0.5 or 0.1 rad/den, and $k \leq$ 1/1300 for E_o, within the linear region of the H-D curve) are both ~1.1%, which is the same as when C = 1 rad/den [Table III(a)]. Yet,

Table VI. Effect of Beam Ratios on Types of Inversion

C (rad/den)	E_r (ergs/cm^2)	Dynamic Range E_r (ergs/cm^2)	Tolerance of eff'l E_r
0.5	1	1300-3000	1.1%
0.1	1	1300-3000	1.1%
0.1	2	1300-3000	2.0%

for the third trial when $C = 0.1$ rad/den and $E_r = 2$ ergs/cm^2, i.e., the beam ratio becomes 1/750 to 1/1500, this variation turns out to be 2%, which implies the inversion is no longer strict. This demonstrates again that the beam ratio k is indeed a more sensitive parameter than C and for the different C values [(e.g., 1–0.1 rad/den) the beam ratio of 1/1300 can still be taken as an upper bound to achieve a maximum range of strict inversion. Since the constants C and γ play the same roles in calculating t [Eqs. (5) and (7)], changing C in our computer simulation is the same as changing γ, which represents a change of film type or chemical process or both in the experiments. Therefore, the examination of the trials with smaller C values just demonstrates how the conclusion we drew previously about the beam ratio constraint ($\leq$1300) for inverse reconstruction is quite general, independent of chemical processing and film type.

IV. Experiment of Inverting a Circulant Matrix

4.1. Description of the Experiment

4.1.1. Input Matrix and Its Diagonal and Inverse Matrix

As we have presented in the pre-paper,[8] the 3×3 binary matrix used to demonstrate the idea of diagonalizing and inverting a circulant matrix is

$$C = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}.$$

Its diagonalized and inverse matrices can be calculated to be

$$\Lambda = \begin{bmatrix} 2 & 0 & 0 \\ 0 & \dfrac{1+j\sqrt{3}}{2} & 0 \\ 0 & 0 & \dfrac{1-j\sqrt{3}}{2} \end{bmatrix},$$

$$C^{-1} = \frac{1}{2}\begin{bmatrix} 1 & 1 & -1 \\ -1 & 1 & 1 \\ 1 & -1 & 1 \end{bmatrix},$$

respectively. Therefore, the moduli of the three eigenvalues are 2, 1, and 1 while those of the elements of the inverse matrix are all $\frac{1}{2}$. In other words, in the experiment we expect the intensity ratio of the three eigenvalue spots to be 4:1:1, which we have already reported in the pre-paper,[8] and all the elements of the inverse matrix to be equally bright.

The mask representing the input matrix is shown in Fig. 7, with the size of each cell and the intervals between cells labeled in the figure. While the primary matrix has a physical size of $\sim 480 \times 480$ μm, the entire mask, including 64×64 replications, is ~ 3 μm $\times$ 3 cm.

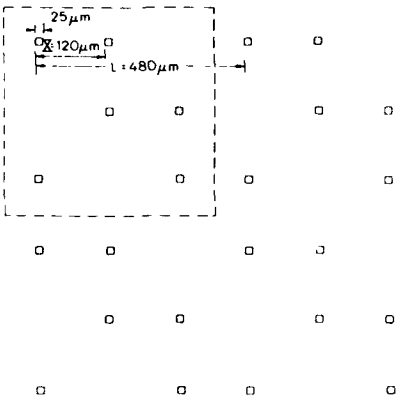


Fig. 7. Input matrix pattern.

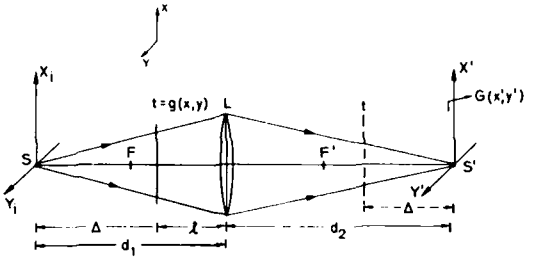


Fig. 8. Generalized optical Fourier transform geometry.

$$G(x',y') = -\frac{d_1}{\lambda d_2 \Delta} \cdot \exp\left\{j\frac{K}{2}\left[\frac{1}{d_2}\left(1 - \frac{ld_1}{\Delta d_2}\right)\right](x_3^2 + y_3^2)\right\}$$

$$\cdot \mathcal{F}[g(x_1,y_1)]|_{v_x = d_1/\lambda d_2 \Delta\, x_3, v_y = d_1/\lambda d_2 \Delta\, y_3}.$$

4.1.2 Optical System for Experiments

Up to now, for the convenience of demonstrating principles, the geometry of the coherent optical system we have dealt with for diagonalizing and inverting a circulant matrix is a typical one with two Fourier transform lenses, each of which has its input and spectrum at the front and the rear focal planes, respectively, as shown in Fig. 1 of the pre-paper.[8] With this geometry, each spot of the eigenvalue pattern we obtained is a sinc function with an infinite set of sidelobes, arising from the square window of the input matrix mask [see Eq. (8) and Fig. 3 of the pre-paper]. This fact is a potential flaw, which is going to show in the eigenvalue inversion process where the weaker sidelobe spots will be boosted and the whole inverse eigenvalue pattern degraded as a result. To solve this problem, a critical step must be taken in order to change the square window to a sinc window and a generalized optical Fourier transform geometry[15] has to be used (see Fig. 8).

Based on the generalized optical Fourier transform geometry, we arranged our system as shown in Fig. 9. A 200- $\times$ 200-μm square pinhole is now introduced in the system. With a parallel beam illumination, the pinhole

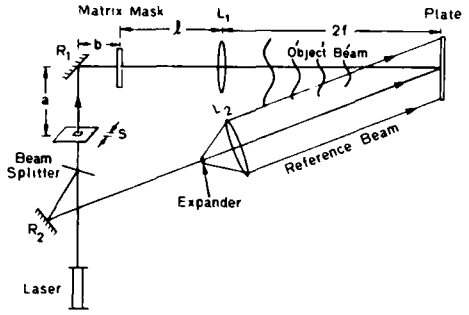


Fig. 9. Optical system for recording Λ^{-1}.

produces its Fourier transform, a 2-D sinc function, on the input matrix mask if the Fraunhofer diffraction approximation is valid. That is to say, the distance from the pinhole to the mask, Δ, and the pinhole size s are constrained[12] to

$$\frac{s^2}{\lambda\Delta} \ll 1,$$

where λ is the wavelength. Since the matrix mask has 64×64 replicas, it is sufficient to include almost eight sidelobes on both sides of this sinc function; the illumination can then be considered as a good approximation of a sinc window, ensuring that the amplitude of each eigenvalue spot is square instead of a sinc.

In addition, with this $2f$–$2f$ imaging configuration and the mask midway between the pinhole and the lens, the Fourier transform of the input mask pattern or the eigenvalue pattern will appear at the image plane of the pinhole.[15] This is just the configuration we use to generate an object field in forming the hologram.

Introducing an expanded off-axis parallel beam as a reference beam in the above system, we built our holographic geometry to record Λ^{-1} as shown in Fig. 9. In reconstruction, as we discussed in the previous section, a conjugate reference beam should be used as a reconstructing waveform. This is equivalent to using the original reference beam, but the hologram is turned over 180° with its glass substrate toward the reconstructing beam. With this reconstruction geometry, the waveform Λ^{-1} will appear immediately after the hologram and its Fourier transform is the inverse circulant matrix pattern, appearing at an infinite distance with an infinitely large size. To bring it back to a finite distance but with considerable magnification, a convergent reconstructing beam is finally used in our experiment instead of a parallel reconstructing beam.

4.2. Results

4.2.1. Pattern and Measurement of Eigenvalues

Although the measurement of one group of eigenvalue patterns has been reported in the pre-paper, we would like to show the measurements for nine groups for comparison with that for the inverse eigenvalue pattern. An eigenvalue pattern without sidelobes is shown in Fig. 10(a). The nine groups of eigenvalue about the center are measured by a silicon photodiode detector, and their relative moduli values are listed in Fig. 10(b). After

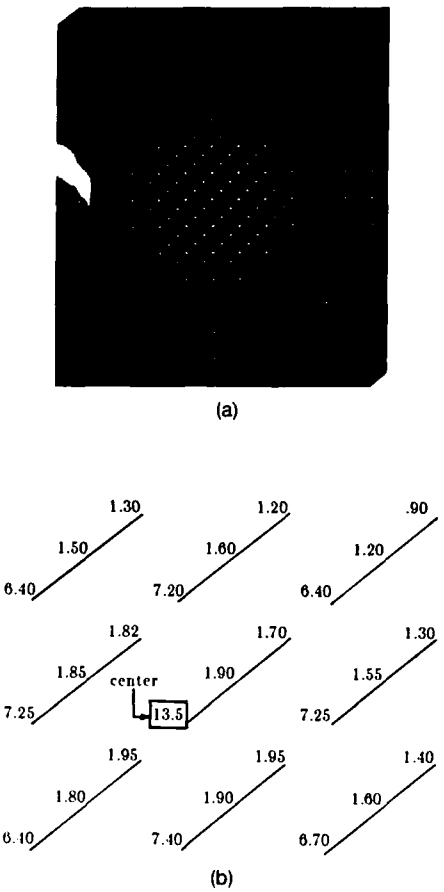


Fig. 10. Results of Eigenvalues: (a) eigenvalue pattern with a sinc window; (b) measurement of eigenvalues.

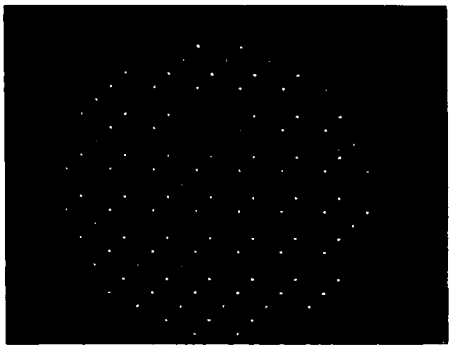
compensation for the sinc envelope for the whole eigenvalue pattern, the measured data are very close to the ideal values with an error of ~1.5% as reported in the pre-paper, except for the one at the center which is greater than the calculation, because the background of the input matrix mask is never zero, focusing and adding to the central eigenvalue spot and causing its error.

4.2.2. Holographic Process and Measurement of the Inverse Eigenvalues

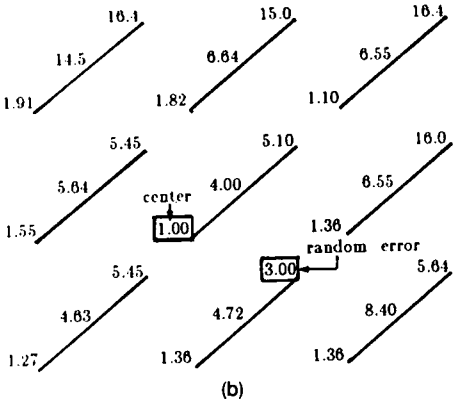
The holographic process for inversely reconstructing the eigenvalue pattern is described in Table VII. Some explanations need to be given. First, the linear region of the H-D curve given in the table does not have an upper limit because the exposures chosen were not large enough. In addition, the two limits of beam ratios and exposures listed in the table are for the darkest and brightest points of all the spots in the nine groups other than the group at the center. Because the strict inversion regime has its lower bound at a beam ratio of ~1/1300 according to the conclusions based on the computer simulations in the previous section, some spots around the edges of the nine groups are not in the strict inversion regime. This will introduce some errors on reconstruction. Moreover, the central eigenvalue,

Table VII. Holographic Process

Hologram forming	Beam Ratio k	1/500 - 1/4000
	Exposure	160- 1300 $ergs/cm^2$
Film Property	Holographic plate	Agfa 10E75
	Chemical Process	Development : 4', 20°C in D76[1:2]
	H-D Curve	Contrast $\gamma \sim 1$ Linear Region : 160 - over 1330 $ergs/cm^2$
Bleaching Process	Agent	Potassium Ferricyanide
	Const. C	~ 0.8



(a)

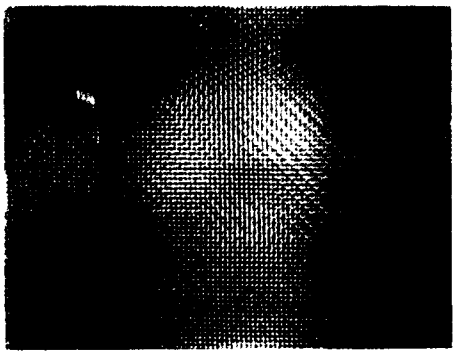


(b)

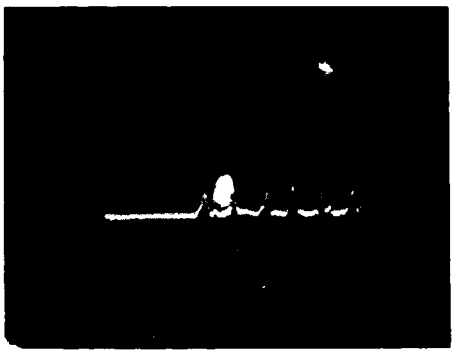
Fig. 11. Results of inverse eigenvalues: (a) inverse eigenvalue pattern; (b) measurement of inverted eigenvalues.

which is almost twice as bright as the brightest in the nine groups, may be beyond the linear region during the exposure. This may also introduce errors in the reconstructed pattern. Finally, since a small γ and C are used in the process, the phase distortion is suppressed to a rather low level. As an approximate estimate, the maximum phase distortion is about $\pi/2$ for the spots in the nine groups, corresponding to a dynamic range from 1/500 to 1/4000 (see Table VII).

Figures 11(a) and (b) show a reconstructed inverted eigenvalue pattern and the measured moduli of the inverted eigenvalues. From Fig. 11(a) it can be seen that



(a)



(b)

Fig. 12. Results of the inverse matrix: (a) inverse matrix pattern; (b) profiles of six elements of the inverse matrix pattern.

the brightness of the three spots in each group is inverted; i.e., the original distribution of bright–dim–dim becomes dim–dim–bright. From the values listed in Fig. 11(b) we can also see the errors involved in the three measured values. These errors exist not only at the edges and the center of the nine groups, due to the reasons we pointed out above, but also at the point indicated in Fig. 11(B). Most likely the latter arises from random error in the chemical process.

4.2.3. Pattern and Measurement of Inverse Matrix

As a result given by the generalized Fourier transform geometry,[15] when a convergent spherical wave is used as a reconstructing wave, the Fourier transform (up to a quadratic phase) of the inverse eigenvalue pattern will occur at the center of this convergent wave. This is just the inverse circulant matrix pattern we look for. A picture of this pattern (in 1:1 scale) is taken and shown in Fig. 12(a). Coming from the diagonalized eigenvalue pattern (Fig. 9), this pattern looks like a grid structure with a sinc-shaped envelope, which is actually the enlarged image of the sinc window of the input matrix. Even though the pattern is degraded due to the errors introduced by the foregoing steps, some portions are still sufficiently good to resolve and measure, e.g., the center of the pattern and some places in the sidelobes of the sinc. Because the spots of the pattern are crowded and dim, we have to use a detector array to take the intensity profile of the spots instead of measuring their integral intensities individually by a silicon photodetector. The

detector array we used is a Reticon RL-1024 array. The profile of six spots at the center of the pattern is recorded and shown in Fig. 12(b). As mentioned in the beginning of this section, we expect all spots of this pattern to be equally bright. From Fig. 12(b) we see they are indeed fairly close to this condition.

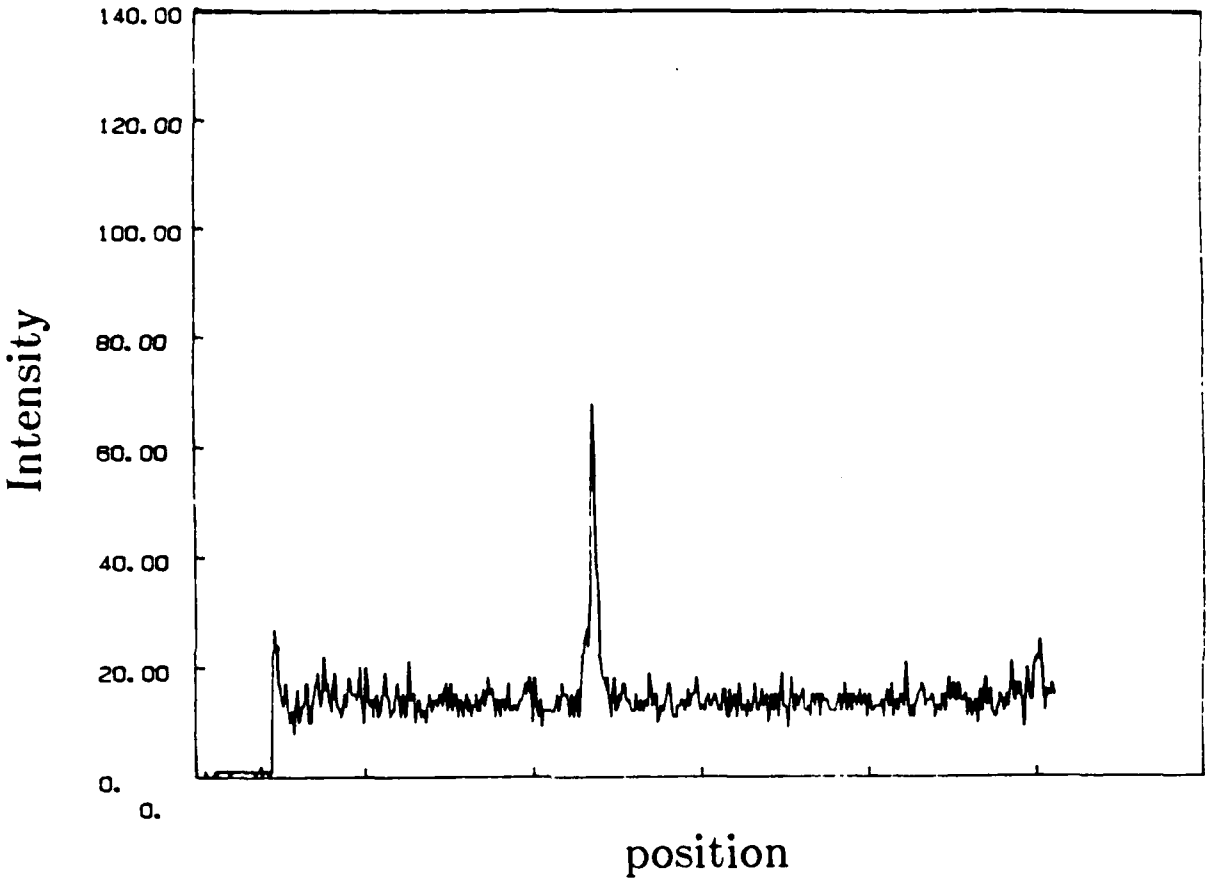
V. Concluding Remarks

In this paper, we pursue the study of wave-front inversion by means of a thin phase hologram. Reformulating the analysis of this type of hologram, we explicitly point out the inevitable phase distortion that accompanies the desired inversion. A computer simulation of this type of hologram, based on the theoretical analysis, demonstrates the principle and provides quantitative information and instructive conclusions. First, it is pointed out that the inversion regime strongly depends on the H-D curve of the film. Breaking the entire inversion regime into regions of strict and approximate inversion, we find that strict inversion takes place when the total exposure falls in the linear region of the H-D curve, while approximate inversion occurs in the shoulder portion. In addition, the beam ratio must be less than ~1/1300 for strict inversion with the largest dynamic region. Furthermore, the phase distortion can never be eliminated, which poses an intrinsic drawback for this type of hologram, but can be much improved by reducing the slope (γ) of the linear portion of the H-D curve and charging the bleaching chemicals. The results of the simulation were used for designing an experiment to invert a circulant matrix.

References

1. J. W. Goodman, "Coherent Optical Image Deblurring," in *Coherent Optical Engineering*, F. T. Arecchi and V. Degiorgio, Eds. (North-Holland, Amsterdam, 1977), pp. 263–280.
2. G. W. Stroke and R. G. Zech, "*A Posteriori* Image-Correcting 'Deconvolution' by Holographic Fourier-Transform Division," Phys. Lett. A **25,** 89 (1967).
3. A. Lohmann and H. W. Werlich, "Holographic Production of Spatial Filters for Code Translation and Image Restoration," Phys. Lett. A **25,** 570 (1967).
4. C. Zetzsche, "Simplified Realization of the Holographic Inverse Filter: A New Method," Appl. Opt. **21,** 1077 (1982).
5. S. I. Ragnarsson, "A New Holographic Method of Generating a High Efficiency, Extended-Range Spatial Filter with Application to Restoration of Defocused Images," Phys. Scr. **2,** 145 (1970).
6. D. Tichenor, "Extended Range Image Deblurring Filters," Ph.D. Thesis, Dept. of Electrical Engineering, Stanford U. (1974).
7. J. W. Goodman, "Architectural Development of Optical Data Procession Systems," Proc. Inst. Radio Electron. Eng. Aust. **2,** 139 (1982).
8. Q. Cao and J. W. Goodman, "Coherent Optical Techniques for Diagonalization and Inversion of Circulant Matrices and Circulant Approximations to Toeplitz Matrices," Appl. Opt. **23,** 803 (1984).
9. C. E. K. Mees, *The Theory of the Photographic Process* (Macmillan, New York, 1954).
10. "Kodak Plates and Films for Science and Industry," *Kodak Data Book* (Eastman Kodak Co., Rochester, N.Y., 1962).
11. R. L. van Renesee and F. A. J. Bouts, "Efficiency of Bleaching Agents for Holography," Optik **38,** 156 (1973).
12. J. W. Goodman, *Introduction to Fourier Optics* (McGraw-Hill, San Francisco, 1968).
13. J. Collier, C. B. Burckhardt, and L. H. Lin, *Optical Holography* (Academic, New York, 1971), p. 55.
14. A. Kozma, "Photographic Recording of Modulated Light," J. Opt. Soc. Am. **56,** 428 (1966).
15. Q. Cao, "Coherent Optical Techniques for Computing Eigenvalues and Inverse of Circulant Matrices," Ph.D Thesis, Dept. of Electrical Engineering, Stanford U. (1984).

We gratefully acknowledge the financial support of the Air Force Office of Scientific Research.


140.00
120.00
100.00
80.00
60.00
40.00
20.00
0.
0.
Intensity
position

APPENDIX V

OPTICAL INTERCONNECTIONS FOR VLSI SYSTEMS

Joseph W. Goodman, Frederick J. Leonberger,

Sun-Yuan Kung and Ravindra A. Athale

Reprinted from Proceedings of the IEEE, Vol. 72, No. 7, July 1984

Optical Interconnections for VLSI Systems

JOSEPH W. GOODMAN, FELLOW, IEEE, FREDERICK J. LEONBERGER, SENIOR MEMBER, IEEE, SUN-YUAN KUNG, SENIOR MEMBER, IEEE, AND RAVINDRA A. ATHALE

Invited Paper

The combination of decreasing feature sizes and increasing chip sizes is leading to a communication crisis in the area of VLSI circuits and systems. It is anticipated that the speeds of MOS circuits will soon be limited by interconnection delays, rather than gate delays. This paper investigates the possibility of applying optical and electrooptical technologies to such interconnection problems. The origins of the communication crisis are discussed. Those aspects of electrooptic technology that are applicable to the generation, routing, and detection of light at the level of chips and boards are reviewed. Algorithmic implications of interconnections are discussed, with emphasis on the definition of a hierarchy of interconnection problems from the signal-processing area having an increasing level of complexity. One potential application of optical interconnections is to the problem of clock distribution, for which a single signal must be routed to many parts of a chip or board. More complex is the problem of supplying data interconnections via optical technology. Areas in need of future research are identified.

I. Introduction

There are several roles that optics can play in the field of computation. Best known are the well-demonstrated applications of optics to analog computation. Examples include acousto-optic spectrum analyzers, convolvers, and correlators [1], [2], as well as systems for forming images from synthetic-aperture radar data [3], [4]. Such analog approaches offer very high processing speeds, but low accuracy and limited flexibility in terms of the types of operations that can be performed. These shortcomings have led to a search for applications of optics to digital [5], [6] or other types of numerical computation [7], [8].

A digital computer or computational unit consists primarily of nonlinear devices (logic gates) in which input signals interact to produce output signals, and interconnections between such devices or groups of devices of various sizes and complexity. The nonlinear interactions required of individual computational elements are realized in optics only with considerable difficulty. Various kinds of optical light valves have been utilized to realize a multitude of parallel nonlinear elements [9], [10], but the speeds at which such devices can operate are exceedingly slow by comparison with equivalent electronic elements. Recent discoveries in the area of optical bistability have generated renewed interest in the possibility of constructing optical logic gates that are even faster than their electronic counterparts, but currently the efficiency of such devices is low and the device concepts are too little explored to allow a full assessment of their potential. It seems safe to say that the construction of optical logic gates with speeds, densities, and efficiencies equaling or exceeding those of electronic gates remains problematical, although future progress is certainly possible.

While optics lags behind electronics in the realization of the needed nonlinear elements, nonetheless the horizon for electronics is not without clouds. It is generally realized that the exponential growth of semiconductor chip capabilities cannot continue indefinitely, and that indeed important limits are beginning to be felt already. These limits arise not from difficulties associated with the further reduction of gate areas and delays, but rather from the difficulties associated with interconnections as dimensions are further scaled downward and chip area continues to increase [11]–[13].

Given the above facts, it is natural to inquire as to whether optics might offer important capabilities in overcoming the interconnect problems associated with microelectronic circuits or systems. Encouragement is offered by the observation that the very property of optics that causes difficulty in realizing nonlinear elements (it is difficult to make two streams of photons interact) is precisely the property desired of an interconnect technology. Further encouragement is offered by the noted successes of fiber optics in satisfying modern communication needs on a more macroscopic scale.

The purpose of this paper is to explore the possible means by which optics might contribute to the solution of interconnect and communication problems in integrated circuits (ICs) and systems. Attention is by no means limited to fiber optics, but rather places emphasis on integrated optics and free-space interconnection techniques as well. The ideas are admittedly speculative to some degree, but an attempt is made to introduce realistic numbers wherever

Manuscript received February 3, 1984.

J. W. Goodman is with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA.

F. J. Leonberger is with Lincoln Laboratory, Massachusetts Institute of Technology, Lexington, MA 02173, USA.

S-Y. Kung is with the Department of Electrical Engineering-Systems, University of Southern California, University Park, Los Angeles, CA 90089, USA.

R. A. Athale is with the Naval Research Laboratory, Code 6530, Washington, DC 20375, USA.

0018-9219/84/0700-0850$01.00 ©1984 IEEE

possible. To our knowledge there has been no previous attempt to systematically explore the area of optical interconnections in a comprehensive way.

Section II provides an overview of the current state of IC technology, with emphasis on the limitations posed by interconnect requirements. Section III provides further motivation by considering the implications of signal/image processing algorithms with respect to interconnections. Some algorithms require only local interconnects, while others fundamentally benefit from global or more flexible interconnect capabilities. Section IV reviews the current state of the art of optical technology relevant to the interconnect problem. Section V introduces a number of specific optical approaches to interconnections, at various levels, including intrachip, interchip, and interboard (machine-to-machine communication is excluded from consideration, due to the large attention it has already received by others). Finally, Section VI identifies the outstanding problems that must be solved for optical interconnections to find real application in the microelectronics field, and suggests future directions for research.

II. Overview of Integrated Circuit Technology and Interconnect Limitations

The purpose of this section is to give a brief overview of the current state of metal–oxide–semiconductor (MOS) IC devices and systems, to outline the limitations posed by the interconnect problem, and to very briefly discuss some VLSI architecture issues. The treatment is far from exhaustive, but rather focuses on aspects that are relevant to the central question addressed in this paper; namely, what role might optics play in helping to solve the interconnect problem. Emphasis here is on silicon (Si) MOS circuits, since they are the base for the vast majority of VLSI-based computational power. Comments on the use of optical interconnections in hybrid GaAs–Si circuits as well as high-speed Si bipolar and GaAs ICs are found in Sections V and VI.

A. MOS Circuits [14]

MOS circuits are typically constructed from n-channel enhancement and depletion transistors (NMOS), or n-channel and p-channel enhancement transistors (CMOS). An n-channel transistor is made in a p-type substrate, whereas a p-channel transistor is made in an n-type substrate. In an n-channel transistor, the drain and source regions are created by n-type diffusions. The gate is made of a conductor (polysilicon) over a thin oxide covering the region between the drain and source diffusions. When the voltage of the gate is raised with respect to the drain, source, and substrate voltages, electrons are attracted to the surface of the substrate. Above a certain threshold, the number of electrons is so large that they form a conducting channel between the source and the drain.

The basic MOS module is the inverter circuit. Fig. 1 shows basic NMOS and CMOS inverter circuits. Usually, the transistor connected to ground is called the pulldown transistor, while the transistor connected to V_{dd} is called the pullup transistor. The pullup transistor of the NMOS inverter is a depletion-mode transistor, i.e., it is always on. The pullup transistor of the CMOS inverter is a p-channel enhancement-type; it will be on only when the voltage at

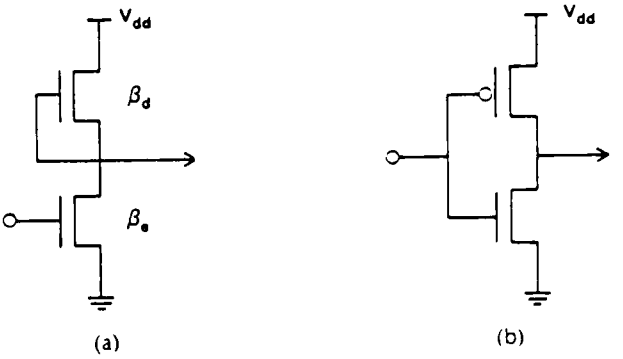


Fig. 1. Basic NMOS and CMOS inverter circuits. (a) NMOS-depletion load static inverter. (b) CMOS static inverter.

the gate is low with respect to the voltages at the drain and source. When the inverters are properly designed, their outputs can be used to drive the input of the next inverter stage.

A major difference between NMOS and CMOS is the power dissipated. A CMOS inverter draws power only in a transient condition, due to the fact that normally only one of the transistors is on. An NMOS inverter, on the other hand, draws power whenever the pulldown transistor is conducting. One way of alleviating this problem is to use dynamic logic, which is however more difficult to design and requires more area. Consequently, CMOS is in general preferable to NMOS.

B. Integrating Circuits on a Chip

A strong point of VLSI is the availability in the near future of a *hierarchical* and *multilevel* design method and the associated software packages. Such approaches are imperative due to the extremely large (> 100 000) numbers of MOS gates per chip in current technology. Usually four description levels are considered: 1) architectural, 2) register transfer, 3) logic/circuit, and 4) layout. An upper level description should be an elegant and powerful abstraction of the more detailed implementation at the lower level. For instance a control unit can be simply a box at the architectural level, one or more finite-state machines (FSM) at the register transfer level, an MOS programmable logic array (PLA) circuit description at the circuit level, and a collection of rectangles at the layout level. At each of the description levels the cells are hierarchically specified to decrease the complexity of the description. In order to reduce the design costs, a modular design approach is used; it is often less expensive to implement a general module that can be used in a number of different places than to implement a specific module that can be used only once.

The actual process of designing the layout of a VLSI chip is fairly well supported at the moment. Existing interactive layout editors and design rule checkers relieve the designer from most of the tedious work of specifying and checking the layout. An important aid in specifying the layout of a VLSI circuit is the so-called "stick diagram" [14]. A stick diagram specifies the topology of the circuit, i.e., the relative positions of the transistors and their interconnections. In a stick diagram the transistors are symbolically depicted as the crossing of polysilicon and diffusion lines. A stick diagram adequately models the functional behavior (i.e.

the logic gates and their interconnections) of the circuit. However, it does not allow the specification of certain capacitive effects, such as bootstrapping.

At the circuit level, we decompose the circuit into three major types of building blocks. The basic *memory* module is the one bit register cell; the basic *logic* module is the AND-OR-INVERT gate, and the basic *arithmetic* module is the full adder. Fundamentally, a VLSI circuit consists of these three types of modules. A somewhat special (but widely used) logic module is the PLA. A PLA can be used to implement any set of Boolean equations, and if combined with a state register, can even implement a complete FSM. A PLA can be generated directly from a register transfer level specification. In general, with the increasing use of high-level design aids, we see more and more programs that are able to synthesize large portions of a VLSI circuit from a high-level (register transfer) specification.

The building block approach, described above, when combined with high-level tools such as silicon compilers/assemblers, gives the VLSI designer the flexibility and modularity needed to cope with the ever increasing complexity of VLSI design.

C. Effects of Scaling on Device and Interconnection Delays

The exponential growth of IC complexity and capabilities experienced since the birth of the industry has been caused by a combination of scaling down of the minimum feature size achievable, and a scaling up of the maximum chip size, both subject to the constraint of reasonable yield. The scaling process has many beneficial effects, but also eventually causes difficulties if combined with "stuffing," i.e., the addition of circuitry in order to realize more complex circuits in the same area of silicon that was used before scaling [15]. Here we wish to briefly discuss the good and bad effects of scaling.

We will assume that all the dimensions, as well as the voltages and currents on the chip, are scaled down by a factor α (an α greater than one implies that sizes or levels are shrinking). Consider first the effects of device scaling. Obviously, when scaling down the linear dimensions of a transistor by α, the number of transistors that can be placed on a chip of given size scales up by α^2. In addition, the power dissipation per transistor decreases by a factor α [16], due to the fact that both the threshold voltage and the supply voltage are scaled down by α. Finally, we note that the switching delay of a transistor is scaled down by α, due to the fact that the channel length is decreased by a factor α.

Scaling also affects the interconnections between devices. Fig. 2 depicts the effect of scaling down a conductor by a factor α. Since the cross-sectional area of the conductor is decreased by a factor α^2, the resistance per unit length will increase by a similar factor. If the length of the conductor is scaled by α (as simple scaling implies), then the net increase of resistance is in proportion to α. At the same time, scaling implies changes of the capacitance of the interconnection. Regarding the conductor as one plate of a parallel-plate capacitor, scaling down of both linear dimensions of the plate by α implies a decrease of the capacitance by α^2. However, scaling down also implies a decrease by α of the thickness of the oxide insulating layer separating the plates of the capacitor. Hence the capacitance of a fixed interconnection scales down by α. We see the scaling up of resistance and down of capacitance exactly cancel, leaving the *RC* time constant and the interconnect delay unchanged.

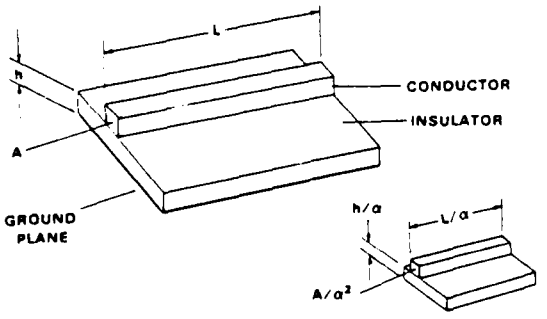


Fig. 2. Scaling of a conductor (scaling factor α).

It is well-recognized in the IC literature that the scaling laws outlined above will eventually pose serious problems for the VLSI industry [11]–[13], [17]–[19]. First, it is clear that since gate delays decrease with scaling while interconnect delays remain constant with scaling, eventually the speed at which a circuit can operate is dominated by interconnect delays rather than device delays. However, the situation is actually somewhat worse than the above considerations imply, due to the fact that as scaling and stuffing occur, the lengths of the interconnects required do not scale down with the inverse of α, as was assumed. Rather, as the complexity of the circuit being realized increases, the distances over which interconnections must be maintained on a chip of fixed area may stay roughly constant. It has been argued from statistical considerations [20] that a good approximation to the maximum length L_{max} of interconnection required is given by

$$L_{max} = \frac{A^{1/2}}{2} \tag{1}$$

where A represents the area of the chip. Note that if in addition to scaling, chip size is increased, the interconnect problem becomes further exacerbated. As a consequence of these considerations, it has been estimated that by the late 1980s, chip speeds will be limited primarily by interconnect delays [11]. The tantalizing possibilities for bringing optics to bear on this interconnection bottleneck are prime motivating factors for the considerations of this paper.

D. Effects of Scaling on Numbers of Interconnections Required

In accord with the hierarchical nature of design, a complicated VLSI chip can be regarded as consisting of a multitude of subunits of circuitry, called "blocks," connected to form larger circuit units, called "super-blocks" [11]. As scaling proceeds, the complexity of the blocks can be made greater, and the number of blocks that can be realized in a single super-block also grows. As the number of elements in a block increases, the number of interconnections required from that block to other blocks also increases. If the assumption is made that interconnections to or from a given block must be supplied around the perimeter of that block, then the limitation imposed by the amount of available perimeter implies that the number of interconnections that can be supplied grows as the square root of the area of the block, or equivalently as the square root of the number of devices contained within the block. However, there is a

well-known empirical relation, known as *Rent's rule*, which specifies that the number of interconnections M required for a block consisting of N devices grows as approximately the 2/3 power of N, i.e.,

$$M \approx N^{2/3}. \tag{2}$$

At the chip level, the disparity between the number of interconnections that can be realized and the number required becomes more and more severe as the number of devices within the chip grows larger through scaling and stuffing. For example, a circuit consisting of 100 000 gates requires about 2000 interconnections. For a 10 mm × 10 mm chip works out to a connection pad every 20 μm. It should be noted that Rent's rule applies only to circuits consisting of logic elements. Memory cells require fewer interconnections. In addition, it is required that each circuit be a small "random" subset of the entire logic system [13].

At the chip level, some of the limitations implied by Rent's rule can be overcome by use of metal bump technology for making interconnections possible from the interior of a chip, rather than just from the edges. Optical techniques may ultimately provide an alternate and more flexible means for providing interconnections directly to the interior of a chip.

As the number of devices realizable on a single chip grows, the assumptions underlying Rent's rule may become invalid, and the exponent associated with that rule may fall to less than 0.5. Nonetheless, from a practical point of view, connections to and from a chip are likely to remain a problem area, due to the bandwidths and driving powers required, as well as the continuing need to distribute signals to and from the interior of the chip.

E. Effects of Scaling on Electromigration

There is one further negative effect of feature size scaling that should be mentioned, namely *electromigration*. While current is scaled down inversely as α, the cross-sectional area through which that current must flow is scaled down inversely with α^2. The net result is that current density increases in proportion to α. Such an increase leads to greater electromigration effects, by which is meant the movement of conductor atoms under the influence of electron bombardments, resulting ultimately in the breaking of conductor lines. The potential of optical interconnections as a means for reducing the electromigration problem is of considerable interest here.

F. System Considerations

Arrays of identical VLSI processing elements can be organized to perform a variety of signal processing functions, as will be reviewed in the next section. Interconnections in such VLSI systems are customarily implemented in a two-dimensional circuit layout with a few crossover layers. With such technology, communication is ideally restricted to localized interconnections, since communication is very expensive in terms of area, power, and speed [14]. Dynamic interconnections are desirable for general-purpose computing and in certain signal processing applications.

The timing framework is a very critical issue in system design. There exist two different timing schemes, *globally synchronous* and *locally synchronous* approaches. In the globally synchronous scheme a global clock network must distribute the clocking signals throughout the entire system. Clock skew associated with global clock distribution has become an increasingly important factor in limiting achievable clock speeds. Under such circumstances the locally synchronous approach has some advantages, in that there need be no global clock, and information transfer is by mutual convenience and agreement between each processing element and its immediate neighbors. The performance of such a scheme is less affected by the scaling of technology than is the globally synchronous scheme, and it can be implemented with a simple handshaking protocol [21]. Optics may provide an alternative method for solving these synchronization problems, as discussed in greater detail in Section V.

III. Algorithmic Implications of Interconnects

Further motivation for considering optical interconnections is provided by algorithmic considerations. The capabilities and limitations of the interconnect technology utilized in realizing a computational or signal processing unit play a substantial role in determining the speed and flexibility of the operations that can be achieved by that unit. Different algorithms require different degrees of interconnect globality, and it is the implications of such considerations, particularly as they pertain to electronic and optical interconnections, that are the subject of this section.

To effectively exploit the special features of optical and electronic technologies, the mathematical operations needed in signal processing and computational operations must be cast in suitable algorithms. Optical signals can flow through three-dimensional space to achieve the required interconnect pattern between elements of a two-dimensional data array before executing the desired operation between them. Current VLSI-based electronic systems, on the other hand, are inherently two-dimensional in nature. While work on three-dimensional VLSI is in progress, success in this endeavor will primarily increase the density of computational elements, rather than alleviating the constraints imposed by interconnect limitations. For current VLSI technology, the interconnect paths as well as the processing elements have to share what is essentially a common plane. These topological considerations, as well as cross-talk and interconnect delay limitations, impose a restriction to nearest neighbor interconnections as being highly desirable in VLSI parallel-processing systems. The algorithmic mapping of the same mathematical operation with optical interconnects and electronic interconnects may be quite different in order to conform to the different constraints of the interconnect patterns. In subsequent discussion, we will consider four different classes of signal processing operations that are defined by the type of interconnect patterns that are needed to implement those operations on a parallel array of processors.

A. Point Operations

In this category of operations, each point in the one- or two-dimensional data array is processed completely independently. If the one-dimensional input is a time sequence, then these operations are referred to as "memoryless" operations. All the points in the input array may be processed in the same way or each could have its own independent instruction set. In either case, it is clear that once the input data array is loaded into the array of processing elements, then each element can carry out its own predetermined processing task completely independently of the rest of the

elements in the array. The interconnectivity required by these operations is therefore minimum and will be apparent only while loading the data into or unloading the data from the processor array. These operations can therefore be carried out in parallel, whether interconnections are provided by optical or electronic means.

Optical interconnections have the advantage of being able to input the entire two-dimensional data array in parallel using the third dimension for data propagation. On the other hand, in an electronic parallel processor, such as the Goodyear Massively Parallel Processor (MPP) [22], the data can be input and output only along the edges of the two-dimensional array, one row/column at a time.

In those cases where the operations needed are complex, the overhead associated with the data input and output may be small compared to the computational load of the main operation, thus minimizing the need for a completely parallel input/output link between the processing element and the outside world. Examples of such operations can be found especially in the field of image processing. The value of a picture element can be transformed according to a prescribed nonlinear function in order to modify its contrast. Another example would involve correcting for spatially varying sensitivity of a two-dimensional sensor array by suitable post-processing. Addition/subtraction of matrices is yet another example of this type of operation.

B. Matrix Operations

A large number of signal and image processing algorithms can be expressed in terms of matrix operations. The multiplication of two matrices is one of the most basic operations in matrix algebra (a vector–matrix multiplication can be considered as a special case of this more general operation). Such a multiplication is described mathematically as follows:

$$C_{ij} = \sum_{k=1}^{N} A_{ik} B_{kj}, \qquad i,j = 1,\cdots,N \tag{3}$$

where A, B, and C are assumed to be $N \times N$ square matrices, for the sake of convenience. Alternatively, the output matrix C can be defined as a sum of N outer product matrices formed by multiplying column vectors of A by corresponding row vectors of B

$$C = \sum_{k=0}^{N} C^{(k)}$$

$$C_{ij}^{(k)} = A_{ik} B_{kj}, \qquad i,j = 1,\cdots,N \tag{4}$$

where the second line defines the outer product between the kth column of A and the kth row of B. It is evident from the above equations that this operation involves a high degree of interconnectivity between the elements of the input matrices and the output matrix. Thus all the elements in a given row of A and a given column of B will contribute to one element of the output matrix C. Conversely, one element of matrix A (or B) contributes to all elements of the corresponding row (or column) of matrix C.

Taking advantage of these properties, the global interconnect capabilities of optics can be exploited to build high-speed, high-throughput parallel optical processors to perform matrix multiplication [23]. On the other hand, the regular nature of the interconnectivity suggested in (3) and (4) implies that these operations can also be carried out via recursive and locally interconnected algorithms implemented with systolic architectures in VLSI [21], [24]. In such algorithms, all processors perform nearly identical tasks and each processor repeats a fixed set of tasks on sequentially available data. A recursive algorithm is said to be locally interconnected if the space indices of the data elements input to the same processor in successive recursions are separated by no more than a given limit [21]. In matrix operations, these indices are found to differ by 1, thus indicating nearest neighbor type of interconnectivity.

It can be seen from (3) and (4) that the computation involved in matrix multiplication grows as $O(N^3)$, where N is the dimension of the matrices. The use of a two-dimensional array of processing elements that perform multiplication and addition of two numbers along with a suitable nearest neighbor interconnection network can be shown to carry out the matrix multiplication in $O(N)$ time. Thus the global interconnection capability offered by optics does not provide any significant computational advantage over systems using only nearest neighbor interconnections when dealing with simple matrix multiplications.

It is worth noting that some algorithms for matrix operations, more complicated than the simple product discussed above, have been proposed that require nearest neighbor connectivity on a three-dimensional surface (e.g., a torus) [25]. Clearly, the communication problem posed by cutting such a surface for compatibility with a planar processor geometry offers opportunities for contributions by optical interconnections. Another potential role for optical interconnections is in the problem of clock distribution in a globally synchronous systolic processor (see Section VI for more detailed consideration of the clock distribution problem).

C. Fourier Transforms and Sorting

Fourier transformation and sorting are two important signal processing operations that entail global interconnections between all the elements of the input array. In other words, every element of the output array is affected by all elements of the input array, and conversely, each element of the input array affects all elements of the output array. The computations involved in the Fourier transform are complex multiplication and addition, whereas in sorting it is the comparison operation.

The discrete Fourier transform (DFT) of a one-dimensional sequence is defined by

$$X(k) = \sum_{i=0}^{N-1} x(i) W_N^{ik}, \qquad k = 0,\cdots,(N-1) \tag{5}$$

where

$$W_N^{ik} = \exp\left[-j2\pi(ik)/N\right].$$

It can be seen that if implemented in a straightforward fashion, this operation involves computation that grows as $O(N^2)$. But the regular structure of the problem as well as the periodic nature of the W_N^{ik} suggest a more efficient algorithm, in which the computation grows as $O(N \log N)$ [26]. However, this computational savings comes at the expense of a global and more complicated interconnection pattern between the input elements than that implied by

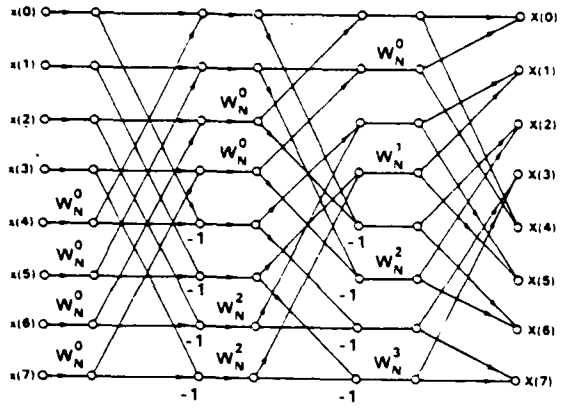


Fig. 3. Flowgraph for the FFT.

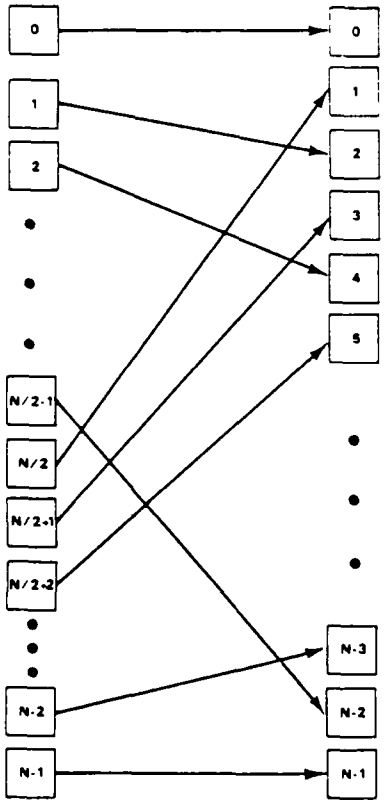


Fig. 4. Perfect shuffle interconnections.

(6). Fig. 3 shows the familiar butterfly diagram of a fast Fourier transform (FFT), which indicates that the interconnections change at different stages of the computation. This requirement for dynamic interconnections can be avoided by resorting to a fixed but global interconnect pattern known as the "perfect shuffle" and shown in Fig. 4. The perfect shuffle can be applied repeatedly at each stage of the FFT to produce the interconnect pattern required for that stage [27], presumably at a cost of extra time required to complete the interconnections.

The operation of sorting a sequence of numbers involves elementary operations of comparing two numbers and arranging them in descending order at the output. One algorithm for efficient sorting is Batcher's bitonic sort algorithm [28]. The basic principle behind this algorithm is the "divide-and-conquer" method of breaking a large problem

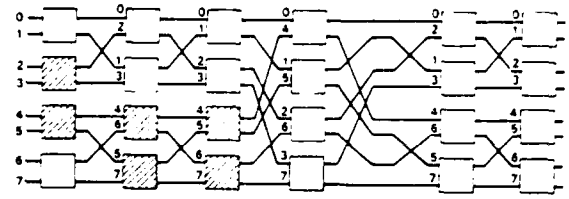


Fig. 5. Sorting network for eight items based on Batcher's bitonic sort algorithm.

into several smaller problems, and then combining their solutions at the output to generate a solution to the more general problem. The network for implementing Batcher's sort algorithm is shown in Fig. 5. Again it is seen that the interconnections between elements are global and dynamic. Since the FFT algorithm is also based on the divide-and-conquer principle, it should come as no surprise that the interconnections required for both operations are identical and hence can be generated by repeated application of the perfect shuffle [27].

Thus both of the operations discussed above belong to the class of operations requiring global and dynamic interconnection between the different elements of the input array. The regularity and structure of the different interconnect patterns leads to a simpler realization of the dynamic interconnects via repeated use of a global but fixed interconnect network (the perfect shuffle network in the case discussed). If either the fixed or the dynamic interconnect networks can be implemented easily and efficiently using optics, then a speedup of the throughput without a corresponding increase in the hardware can be achieved.

D. Space and Time Variant Operations

In image restoration and pattern recognition, one is interested in studying structures of various sizes in images, from a single point and its nearest neighbors to an image covering the entire available field. The interconnectivity between each input point and all object points is determined by the degree of globality of the operation under consideration. When the size and nature of the interconnectivity of the operation to be performed is invariant over the image space, as well as in time, then the problem is considerably simpler, and fixed optical or electronic (depending on the degree of globality required) interconnection networks can be used.

For problems encountered in restoration of images degraded by atmospheric turbulence and in processing signals obtained from a dynamic sensor array, the interconnectivity varies in space and time. Furthermore, the interconnect patterns could be data-dependent (as often is assumed in modeling the human visual system), making it impossible to foresee the interconnect requirements at different stages of processing without having foreknowledge of the input.

The computational throughput of a parallel processor implementing these types of operations will be critically affected by the availability of a dynamic and global interconnect network. Without such a capability, the processors could be idle for a significant number of cycles while the data are routed to the correct processors. A much higher degree of supervision on the part of the controller would be required, and there would be many more input/output operations to and from memory. All these

considerations decrease the computational efficiency of the parallel processors and point to the importance of flexible interconnects.

E. Overview

The four classes of signal/image processing operations discussed above require varying amount of complexity on the part of the interconnect network that routes signals to processors. As the complexity increases, so does the impact of a flexible interconnect technology on the throughput of a parallel processor, whether optical or electronic. In the case of point operations, the motivation for a flexible and global interconnect capability is weak; interconnects are most important in this case when many input/output operations to the processor array are required in practice. In the second category, namely, operations involving matrix algebra, the global but fixed interconnects inherent in the operations can be converted to nearest neighbor interconnects via the recursive formulation. The computational throughput can be high when pipelining is used. Therefore, in this type of application, global interconnections will be beneficial only when the nature of the application produces frequent breaks in the pipelining of the processor. The third category of operations involves global and dynamic interconnects, which can be reduced to repeated application of a global but fixed interconnect network. If these fixed interconnects are implemented in a fast and efficient way, perhaps with the help of optics, then a high computational throughput can be achieved without excessive hardware and without a dynamic interconnect structure. The last category considered contains operations with a minimum amount of regularity and structure, along with possible data and time dependency in the interconnect requirements. A global and programmable interconnect network will be vital to achieving high throughput and high efficiency with parallel processors implementing these operations.

IV. Overview of Electrooptic Technology as it Pertains to the Interconnect Problem

The rapid growth of the fiber telecommunications industry has been in part due to the development of near-infrared optical sources, modulators, and detectors. These components are particularly attractive for solving the interconnect problem. In this section the state of the art of these components, as well as of the transmission media for directing light to specific locations, will be reviewed.

A. Light Sources, Modulators, and Detectors

The alternatives for optical sources are diode lasers and light-emitting diodes (LEDs) [29]. While much of the telecommunications effort is now directed at InP-based devices whose emission wavelengths match the 1.3–1.5-μm optimum fiber band, the 0.85-μm range of GaAs/GaAlAs emitters is of most interest here because of the wavelength sensitivity compatibility with on-chip Si detectors. Typical efficient low-threshold GaAs lasers are of the index-guided type and emit 3–15 mW with threshold currents of approximately 30 mA and differential efficiencies in the 20–30-percent range. The spectral bandwidth is of the order of 20 Å (about 0.2 percent). Typical active laser area is 500×2 μm, but packaged chips are generally wider for bonding. Such lasers are edge emitters with cleaved mirrors and emission patterns typically 10×35 degrees in width. They have projected lifetimes in the $\geq 10^5$-h range. In the laboratory, a number of techniques have been proposed for noncleaved mirrors for more compact devices. Although surface-emitting GaAs lasers have seen limited development, both distributed feedback [30] and heterointerface-mirror [31] types have been reported. These devices are to date relatively inefficient and in the latter case require cryogenic cooling.

LEDs are well developed and can be designed for edge or surface emission. In contrast to diode lasers, they are inefficient (100-mA drive) with a low power output (of the order of 1 mW) that has a broad emission pattern (Lambertian for a surface emitter) and a wide spectral range ($\approx$ 450 Å). However, they are more stable and, in particular, less temperature sensitive than diode lasers.

To modulate these sources, either direct current modulation or external modulation can be used. Diode lasers can be large-signal modulated to rates of approximately 2 GHz with minimal pattern effects, and LEDs are generally limited to $<$ 100-MHz modulation but have recently been reported to operate at rates up to 500 MHz [32]. These sources present dynamic resistive loads of approximately 10 Ω. For situations where driving a capacitive load is preferred, one may use external modulators. Here full on–off modulation requires approximately 4 V in a waveguide modulator for a bandwidth to about 3 GHz with a capacitance of approximately 3 pF. [33] The use of $LiNbO_3$ waveguide modulators to monitor interchip signals in VHSIC circuits has recently been reported [34].

For modulating an optical wavefront propagating through free space, a large variety of electrically addressed light modulators can be used. Recently, several devices have been proposed with drive requirements compatible with Si integrated circuitry, and hence are particularly relevant to this study. These include a $LiNbO_3$ phase modulator [35], a cantilevered beam-deflector made in Si [36], and a deformable mirror device [37]. The last two approaches, being mechanical in nature, are somewhat limited in their speeds ($\approx$ 10 kHz) while the phase modulator needs a complicated optical system for conversion of phase modulation to intensity modulation. A recent proposal for using the electroabsorption effect in GaAs with guided or unguided optical waves has the advantage of potentially high-speed operation ($\approx$ 1 GHz) and of direct intensity modulation [38].

In the detector area, avalanche photodiodes, p-i-n photodiodes, and photoconductors are candidate devices [29]. Avalanche devices require large biases ($\approx$ 50 V) and are relatively temperature sensitive, and photoconductors have a relatively low impedance. Thus p-i-n photodiodes are the most attractive for interconnects. These diodes require only a few volts bias and have quantum efficiencies of approximately 70 percent. High-performance devices have been made in GaAs as well as Si. Self-scanned Si detector arrays are also well developed with scanning rates up to 40 MHz available commercially.

B. Circuits

For circuit interconnects, it is quite relevant to ask what overall transduction efficiency is obtainable using opto-

electronics. Laser efficiencies for devices with uncoated mirrors can approach approximately 30 percent, but minimum power dissipation will remain in the 1-mW range, due to the 1-V diode turn-on and 1-mA current fundamentally needed for lasing [39]. Optical signals in the 0.5-mW range could produce signals from a photodiode–transimpedance preamplifier of approximately 100 mV. To achieve this signal level (at frequencies typically $\leqslant$ 100 MHz) requires at least two transistors; subsequent amplification to volt level logic states will require a few additional transistors. Thus it is not unreasonable to consider integrating these detection circuits on silicon ICs if the photodiode fabrication process can be made compatible with logic circuit fabrication. This could be more of a difficulty with MOS than with bipolar technology.

An alternative method for changing logic states is by direct optical injection into a gate. Initial results utilizing this technique have been reported for both Si [40] and GaAs [41] circuits. At present the peak optical power levels required exceed those available from the low-threshold diode lasers discussed here, but more progress in this area is anticipated.

C. Interconnect Elements

Optics is attractive for interconnects because of the inherent noninteraction of multiple photon beams passing through or near one another. Here, media to be considered include free space, optical fibers, and integrated optical waveguides. An attractive means for exploiting optical beam noninteraction is to use free-space propagation with either focused (e.g., holographic) or unfocused techniques. Holographic optical elements are fairly well developed. They may be written with visible light in dichromated gelatin or silver halide emulsions. Reflective elements with efficiencies limited only by surface reflections (i.e., with efficiencies in the high 90 percentiles) can be realized in dichromated gelatin when imaging with visible light [42]. In bleached silver halide materials, high diffraction efficiencies are hard to achieve in reflective elements due to the limited spatial frequency response of even high-resolution materials, but efficiencies of the order of 70 percent are readily achieved on transmission [43]. Comparable performance should be possible in imaging diode laser or LED emission in the near infrared. Computer-generated holograms can also be constructed, in some cases with the help of electron-beam lithography for writing the hologram. In any of these cases, the equivalent of $F/1$ optical systems can be achieved.

Due to fiber-optic systems, multimode fibers and associated components such as microoptic lenses, star couplers, and wavelength multiplexing modules are under development with some commercial availability. These components could lend themselves to signal distribution. For example, an N-port fiber star coupler can distribute a signal at one of N input ports equally to N output channels. Recently, a coupler with $N = 100$ was reported with total channel loss of 5 ± 0.05 dB over that expected for equal power division [44].

Another candidate for signal distribution is integrated optics. While most work has been reported for guides in electrooptic materials such as $LiNbO_3$ and GaAs [45], numerous workers have fabricated guides in glass or in oxide films on Si substrates. These latter guides can have quite low propagation loss (< 0.1 dB/cm) [46] and passive components such as couplers and splitters can be formed. Here the work on multimode as well as single-mode devices is relevant.

The interconnection can be made changeable by inclusion of an active element in these schemes. With holographic optical elements, a tantalizing (though long-range) possibility is the incorporation of dynamic holographic materials, such as those now being studied for four-wave mixing applications. A shorter term solution might entail a bank of holographic mapping elements in conjunction with a real-time mask which selects the appropriate interconnection pattern. Some candidates for such a real-time mask are matrix-addressed liquid-crystal devices, and the matrix-addressed magnetooptic spatial light modulator [47]. Another approach for a dynamic interconnect could be the implementation of an optical crossbar switch as a special case of an optical matrix–vector multiplier [48]. Dynamic interconnects can also be obtained in the domain of guided-wave optics using integrated directional couplers in $LiNbO_3$ [49]. Cascading several such two-port switches can yield an arbitrary interconnection with high speed capabilities.

D. Hybrid and Monolithic Approaches

To discuss the use of optics in electronic interconnections raises immediately questions of materials technology, since optical sources cannot be formed from silicon. Hybrid and monolithic approaches are conceivable. In the former area, hybrid laser–amplifier circuits are available from a number of suppliers. For detection, hybrids containing a Si diode and preamplifiers are available. These hybridization techniques have to date followed conventional technological approaches to achieve several hundred-megahertz bandwidth. Although monolithic approaches can minimize parasitic capacitance for increased performance, there has been only limited work on monolithic Si photoreceivers, due to process compatibility issues and limited needs for bandwidth.

In the monolithic area, there have been limited reports of GaAs edge-emitting lasers monolithically integrated with electronics. In particular, FET drivers with lasers have been reported, including modulation up to 1 GHz [50]. Recently, a 4:1 multiplexer and laser were reported with speeds to 150 MHz [51]. Monolithic GaAs p-i-n diode–FET amplifiers have also been reported by several groups, and response times compatible with 500 Mbits/s have been achieved [52]. In each case, difficulties associated with process incompatibility between optoelectronic and electronic device requirements had to be overcome.

Similar efforts on both monolithic and hybrid InP-based circuits are being pursued because of the important telecommunications applications. However, these sources typically emit at 1.3 μm, a wavelength not readily detected by Si diodes. Thus this work will have most relevance to wideband fiber interconnects between machines which may be remotely located, a topic which, as previously mentioned, will not be discussed here.

An exciting topic in monolithic integration is the emerging area of heteroepitaxy. Here one attempts to grow crystalline films of one semiconductor on a dissimilar (and thus not lattice matched) crystalline substrate. The most recent

relevant example for the interconnect application is the growth of GaAs on Si and the subsequent fabrication of devices in both materials on the same wafer. Initial reported growth results utilized an intervening layer of Ge to relieve part of the lattice mismatch [53], but dislocation densities were too high (10^{7} cm^{-2}) to fabricate practical devices. Recently an overgrowth technique has been reported which has reduced the dislocation density to 10^{3}–10^{4} cm^{-2} [54], so that optoelectronic device fabrication in the GaAs layers looks promising.

V. Possible Applications of Optical Interconnects

A. Introduction

Some of the difficulties expected in the further extension of integrated electronic technology to circuits and systems with ever greater complexity have been mentioned in Section II. In what respects does optics offer a potential solution to some of these problems?

Propagation of signals on metallic interconnections is governed by the basic laws of circuits containing distributed elements. The exact character of the distributed elements depends on whether interconnections are considered at the chip level or the board level. At the chip level, interconnections can be viewed as distributed and unterminated *RC* transmission lines [11], [18]. At the board level, the transmission lines may contain significant distributed inductance and may be terminated or unterminated. The velocity of signal propagation on such transmission lines depends on the capacitance per unit length. Thus as more and more devices having capacitive components of admittance are attached to an interconnection, the velocity of propagation decreases, and the time required for charging the line to a predetermined voltage level increases.

By way of contrast, propagation of optical signals, whether confined to waveguides or through free space, takes place at a speed that is independent of the number of components that receive those signals, namely, at the speed of light in the medium of concern. Thus we identify the first of several advantages of optical interconnects, namely the *freedom from capacitive loading effects*. Such freedom is responsible for the greater flexibility of optical interconnections with respect to fan-in and fan-out, vis-a-vis electrical interconnections.

A second advantage of optical interconnections over their electronic counterparts is their superior *immunity to mutual interference effects*. The stray capacitances that exist between proximate electrical paths introduce cross-coupling of information to a degree that increases with the bandwidth of the signals of interest. In contrast, optical interconnects suffer no such effects, although care must be exercised to assure that light scattering does not introduce a similar result (of different origin). In any case, there is no electrical coupling between the high-frequency modulations of two proximate beams of light.

A third potential advantage of optical interconnect technology lies in *freedom from planar or quasi-planar constraints*. Waveguides can pass through waveguides without significant cross-coupling (provided the angle of intersection exceeds about 10°), and free-space light beams can pass through free-space light beams without significant interaction. Such flexibility can simplify the problems associated with routing signals on a complex chip or board.

A fourth advantage to be considered lies in the potential for certain types of optical interconnections (namely, free-space focused interconnects, to be discussed in more detail shortly) to achieve *reprogramming by means of a dynamic optical interconnect component*. In principle, the interconnection pattern associated with a chip can be changed at will, by writing appropriate information to the interconnect element. Such a capability would be rather difficult to realize using conventional electronic interconnect technology.

Finally, a fifth reason for considering optical interconnections lies in the possibility (discussed in Section IV-B) of *direct injection of optical signals into electronic logic devices*. Such intimate coupling of optical signals into electronic devices, with the bypassing of separate detectors for optical-to-electronic conversion, could greatly simplify the interface between the optical interconnect and electronic device technologies, as well as offer significant speed advantages vis-a-vis purely electronic connections to the logic, provided suitable low-power approaches can be found.

For optoelectronics to be useful for VLSI interconnections, the size, efficiency, and power requirements must be compatible with IC environments. This is in fact the case, based on the previously given review of component capabilities. For example, detector areas of about 100-μm diameter with about 1 mW of incident optical power should have acceptable power dissipation and signal levels. The overall efficiency of laser drive current to detector output current can approach 20 percent and only a modest number of transistors should be needed after the detector to produce logic-level signals.

B. Classification of Optical Interconnect Problems

In order to discuss the optical interconnect issue, it is useful to describe a hierarchy of connections with some perspective on size considerations. The fundamental chip we take for example is a 10 × 10-mm MOS chip, having $\geq$ 300 000 transistors. This level of complexity is less than that of recently reported NMOS circuits, but in general, at this level of complexity the chip can be wiring dominated and the "pin-out" problem can be severe. Such chips have typically 100-μm center-to-center spacing for bond pads at the periphery and are mounted on a chip carrier with about 50-mil spacing between contacts at its outer edge.

Levels of interconnects to be considered include within a single chip, at the wafer level between undiced chips, between packaged chips on a common board, and between boards. Here it is assumed that a board will be about 1 × 1 ft and will contain about 100 chips. As previously mentioned, the highest level interconnect, between machines, will not be considered here.

Within this hierarchy of interconnections it is appropriate to consider several types of signal distribution. The most straightforward case is clock distribution. Here an identical signal is distributed to a large number of nodes. Since only one-way fixed communication channels are involved, the source can be a separate discrete circuit. The more general case is the distribution of data and control signals. Here more specialized and two-way links would be desirable. In addition, since various degrees of flexibility in the intercon-

nects are desirable, both static and dynamic interconnects need to be explored. In the following sections, both clock and data distribution will be considered. For each case, the hierarchy of interconnections will be matched to the available optical technologies summarized previously.

C. The Problem of Clock Distribution

A problem that appears amenable to immediate attack using optical technology is that of clock distribution at the chip, wafer, or board level. Most (but not all) computing architectures require synchronous operation of a multitude of devices, circuits, and subsystems. Synchronism is maintained by distributing to all parts of the system a timing signal, called the clock. One of the chief difficulties encountered in designing circuits and systems for high-speed operation is the phenomenon known as "clock skew," a term which refers to the fact that different parts of the circuit or system receive the same state of the clock signal at different times. In this section we consider several possible approaches to using optics for distribution of the clock, with the aim of minimizing or eliminating clock skew. Attention is first focused on the problem of distributing the clock within a single chip. Consideration is then given to the problem at the wafer and board levels.

1. Intra-Chip Clock Distribution: The interconnections responsible for clock distribution are characterized by the facts that they must convey signals to all parts of the chip and to many different devices. These requirements imply long interconnect paths and high capacitive loading. Hence the propagation delays are large and depend on the particular configuration of devices on the chip. Here we consider methods for using optics to send the clock to various parts of the chip. It is assumed that optics is used in conjunction with electronic interconnects, in the sense that optical signals might be used to carry the clock to various major sites on the chip, from which the signals would be further distributed, on a local basis, by a conventional electronic interconnection system.

The clocks used in MOS technology are generally two-phase [14]. Presumably only one of these phases will be distributed optically, the other being generated on the chip after the detection of the optical timing signal.

A variety of optical techniques can be envisioned for accomplishing the task at hand, and therefore we devote some time to delineating these various approaches and specifying their strengths and weaknesses. The main distinction between these approaches occurs in the method used to convey light to the desired locations on the chip.

Index-guided optical interconnections. The first major category of optical interconnect techniques we refer to as "index guided." Light is assumed to be carried from some single source generating an optical signal modulated by the clock to many other sites by means of waveguides. The waveguides could be of either of two types. One type could use optical fibers for carrying the optical signals. The second type could use optical waveguides integrated on a suitable substrate.

If fibers are chosen as the interconnect technology, then the following approach, illustrated in Fig. 6, might be used. A bundle of fibers is fused together at one end, yielding a core into which light from the modulated optical source (probably a lasing diode) must be coupled. Light coupled in

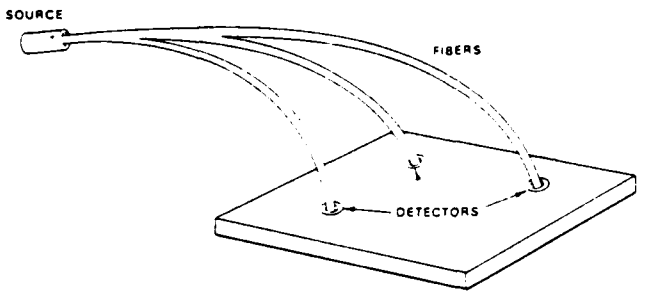


Fig. 6. Distribution of the clock by means of fibers.

at the fused end is split as the cores separate, and transmitted to the ends of each of the fibers in the bundle. Each fiber end must now be carefully located over an optical detector that will convert the optical clock to an electrical one. Alignment of the fiber and the detector might be accomplished with the help of micropositioners (analogous to a wire bonding machine), and UV-hardened epoxy could be used to hold the fiber in its proper place permanently. The difficulties associated with the fiber-optic approach stem from the alignment requirements for the fibers and detectors, and from the uniformity requirements for the fused-fiber splitter. It should also be noted that the fibers cannot be allowed to bend too much, for bends will cause radiation losses that may become severe. Lastly, we should mention that the use of fibers, and the requirements regarding allowable degrees of bending, imply that this interconnect technology will occupy a three-dimensional volume, rather than being purely planar, and this property could be a disadvantage in some applications.

If integrated optical waveguides are chosen as the interconnect technology, then the geometry might be that shown in Fig. 7. The waveguides might be formed by sputtering of

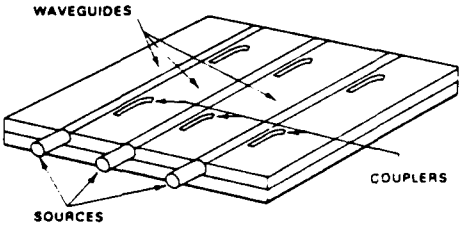


Fig. 7. Distribution of the clock by means of integrated optical waveguides.

glass onto a silicon dioxide film on the Si substrate. These guides are shown as straight in the figure, a configuration chosen again because of the large losses anticipated if this type of light guide is bent at a large angle. Optical signals must be coupled into each of the separate guides. Such signals might be generated by a single laser diode and carried to the waveguides by fibers, or separate sources might drive each of the guides, with the clock distributed to the different sources electrically. Presumably light must be coupled out of each of the straight waveguides at several sites along its length, with a detector converting the optical signal to electronic form at each such site. The difficulties associated with the waveguide approach to the problem, neglecting the bending problem which has been intentionally avoided, stem primarily from the requirement

to efficiently couple into and out of the guides. Careful alignment of the sources or fibers with the integrated waveguides is required, and couplers with short lengths are desired to remove the light from the guides and place it onto the appropriate detectors. Present waveguide technology requires distributed couplers with rather large dimensions (5 μm × 1 mm) compared with the feature sizes normally thought of in electronic IC technology. A major advantage of the integrated optics distribution system lies in its planar character and the small excess volume it requires. A disadvantage is the comparatively inflexible geometry dictated by the necessity to avoid large bends of the waveguides.

Free-space optical interconnections: A second major category of optical interconnects can be referred to as "free-space" techniques. For such interconnects, the light is not guided to its destination by refractive index discontinuities, but rather by the laws that govern the propagation of light in free space. It is helpful to distinguish between two types of free-space interconnect techniques, "unfocused" and "focused."

Unfocused interconnections are established simply by broadcasting the optical signals carrying the clock to the entire electronic chip. One such approach is shown in Fig. 8. A modulated optical source is situated at a focal point of

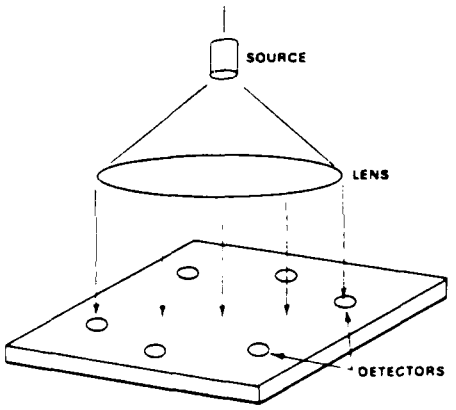


Fig. 8. Unfocused broadcast of the clock to the chip.

a lens that resides above the chip. The signal transmitted by that source is collimated by the lens, and illuminates the entire chip at normal incidence. Detectors integrated in the chip receive the optical signals with identical delays, due to the particular location of the source at the focal point of the lens. Hence in principle there is no clock skew whatever associated with such a broadcast system. However, the system is very inefficient, for only a small fraction of the optical energy falls on the photosensitive areas of the detectors, and the rest is wasted. Inefficient use of optical energy may result in requirements for the provision of extra amplification of the detected clock signals on the chip, and a concomitant loss of area for realizing the other electronic circuitry required for the functioning of the chip. Moreover, the optical energy falling on areas of the chip where it is not wanted may induce stray electronic signals that interfere with the proper operation of the chip. Therefore, it is likely that an opaque dielectric blocking layer would be needed on the chip to prevent coupling of optical signals at places where they are not wanted. Openings in this blocking layer would be provided to allow the optical signals to reach the detectors. Alternate unfocused interconnection techniques could be imagined that use diffusers rather than a lens. Note that all such techniques require a three-dimensional volume in order to transport the signals to the desired locations.

The last category of optical interconnections is free-space "focused" interconnections, which can also be called "imaging" interconnections. For such interconnections, the optical source is actually imaged by an optical element onto a multitude of detection sites simultaneously. As indicated in Fig. 9, the required optical element can be realized by

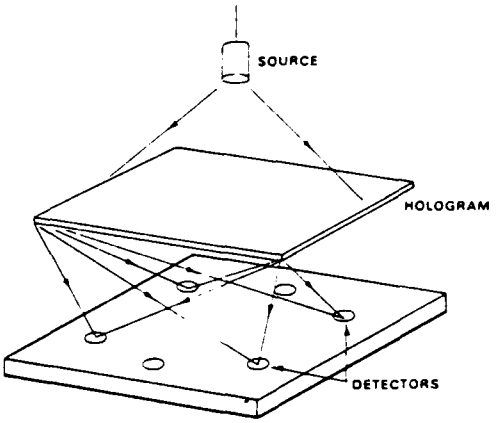


Fig. 9. Focused optical distribution of the clock using a holographic optical element.

means of a hologram, which acts as a complex grating and lens to generate focused grating components at the desired locations. The efficiency of such a scheme can obviously exceed that of the unfocused case, provided the holographic optical elements have suitable efficiency. Using dichromated gelatin as a recording material, efficiencies in excess of 99 percent can be achieved for a simple sine wave grating. When a multitude of focused spots are to be produced, the efficiency will presumably be lower, but should be well in excess of 50 percent. The flexibility of the method is great, for nearly any desired configuration of connections can be realized.

The chief disadvantage of the focused interconnect technique is the very high degree of alignment precision that must be established and maintained to assure that the focused spots are striking the appropriate places on the chip. Of course, the spots might be intentionally defocused, decreasing the efficiency of the system, but easing the alignment requirements. Thus there exists a continuum of compromises between efficiency and alignment difficulty. Fig. 10 illustrates a possible configuration that retains high efficiency but minimizes alignment problems. The imaging operation is provided by two two-element lenses, in the form of a block with a gap between the elements. A Fourier hologram can be inserted between the lenses, and it establishes the desired set of focused spots. The hologram itself consists of a series of simple sinusoidal gratings, and as such the position of the diffracted spots is invariant under simple translations of the hologram. The source is permanently fixed on the top of the upper lens block after it has been aligned with a detector at the edge of the chip.

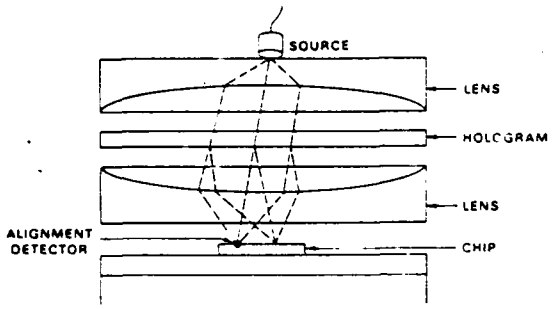


Fig. 10. Configuration for focused clock distribution that minimizes alignment problems

thereby establishing a fixed optical axis. The only alignment required for the hologram is rotation. The position of the image spots is determined by the spatial frequencies of the gratings in the hologram, which could be established very precisely if the hologram were written, for example, by electron-beam lithography.

Focused interconnect systems, like the unfocused ones, require a three-dimensional volume above the chip. If holographic elements are used, thought must be given to the effects of using a comparatively nonmonochromatic source such as an LED. A spread of the spectrum of the source results in a spread of the focused energy, so the primary effect is to reduce the efficiency with which light can be delivered to the desired detector locations.

2. Clock Distribution at the Wafer and Board Levels: Clock distribution at the wafer and board levels is also a strong candidate as an application for optical interconnections. The primary differences in the optical requirements in the two cases derive from the different physical sizes of chips, wafers, and boards, as discussed above. At the chip and wafer levels, the discussion of the previous sections carries over essentially without change. At the board level, the physical areas are sufficiently large that the free-space and integrated optics approaches could at best be used for coverage of only a portion of an entire board. The preferred approach seems to lie with the use of optical fibers for conducting signals to remote parts of a single board, between which the greatest clock skews would otherwise be anticipated. Hierarchical schemes can also be envisioned, in which a network of fibers distributes the clock to a series of widely separated sites, from which either integrated optical waveguides or free-space connects are used to distribute these signals more locally to detectors, where the optical signals are converted to electronic form and distributed on an even more local scale to the various devices that require the clock signal.

D. Data Interconnects

The clock distribution problem discussed above is a particular case where long propagation distances are encountered. The more general case is data exchange between different components of a system. It was seen earlier that, if the process of scaling down feature size and increasing chip size is continued, then at a certain point the speed of the chip will be limited by the delay time associated with the interconnections between different components of the circuit, rather than the switching times of the components themselves. Therefore implementing the longer interconnections optically could potentially enhance the performance of the chip. Communications between different chips are often limited by the number of pins available on the chips for communication with other chips. In addition, as interconnection lengths increase, the size and power requirements of the drivers on the chip also increase. Thus it becomes difficult to communicate with chips that are widely separated, and to have a large fan-out.

In addition to these hardware considerations, several important applications in signal and image processing demand much more flexible interconnections between processing elements of a VLSI-based parallel processor than electronic interconnect technology can provide. It was seen in Section III that some important classes of algorithms require global and dynamic interconnections between the elements of the processor array. Optical interconnections could make valuable contributions to these problems for all levels of integration, from one processor per chip to an entire array of processors on a single wafer.

The optical data interconnects utilize some components (the detectors and the interconnect elements) that are common with the problem of optical clock distribution. However, the conversion of signals from electronic to optical form poses much greater demands in the case of data interconnects. The different technologies available for direct modulation of sources, as well as external modulators (as reviewed in Section IV) will be particularly relevant here. As with the case of clock distribution, we discuss the problem of data interconnects at the two levels of intrachip and interchip communication.

1. Intrachip Data Communication: As a first scheme to be considered, a GaAs chip with optical sources is connected in a hybrid fashion (with conventional wire bond techniques) to a Si chip such that light is generated only along the edges of the Si chip (see Fig. 11). The sources

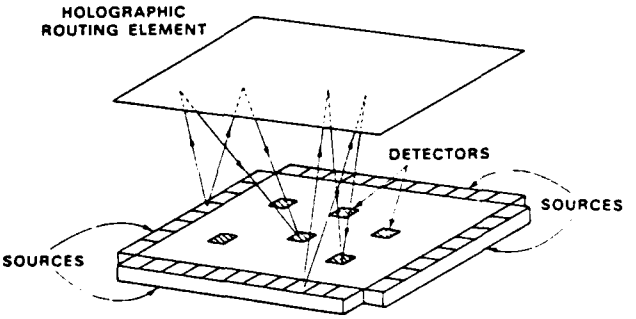


Fig. 11. Hybrid GaAs/Si approach to data communication

could be of the edge-emitting or surface-emitting type. The optical signals are routed to the appropriate locations on the Si chip using conventional and/or holographic optical elements. The Si chip will contain detectors to receive the optical data streams generated by the sources. Since the detector–amplifier combinations can be fabricated in Si, every computational component on the Si chip could be capable of receiving data. Free-space propagation of optical signals allows a large fan-out as well as the possibility of changing the routing via a programmable two-dimensional mask, such as a reflective spatial light modulator.

In the second scheme to be considered, index-guided

structures are used for routing signals instead of free-space propagation. This scheme can be well illustrated by a specific example involving a one-dimensional parallel processor array which is interconnected by a perfect shuffle network (see Section III). The schematic diagram of the system for the case of 4 processing elements is shown in Fig. 12. The Si chip contains 4 processing elements which

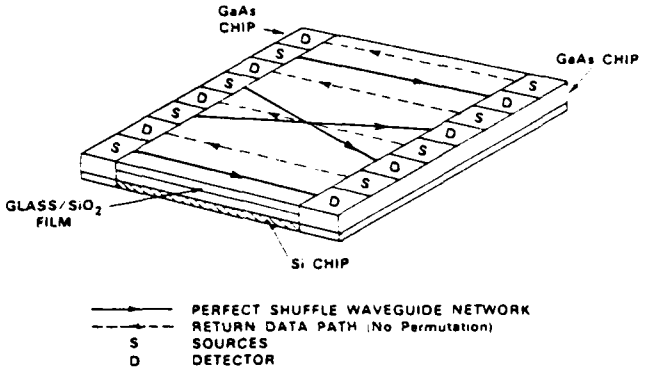


Fig. 12. Optical perfect shuffle network.

output data to the left and accept data from the right. Each processing element is associated with source–detector pairs in the GaAs chips that are mounted on both sides of the chip. Thus there exists a bidirectional optical data path for each processing element. The routing of optical data is performed by a crossing network of channel waveguides that can be formed in glass/SiO_2 layers grown on top of the Si chip. The waveguide network connecting the sources on the output side of the processor array (i.e., the left side) with the detectors on the input (right) side perform a perfect shuffle while the reverse optical paths correspond to a processing element connected to itself (no permutation). Using this scheme it is possible to shuffle data as many times as necessary to obtain the desired interconnect pattern before processing by the electronic part of the system. Since the sources and detectors can operate at several hundred megahertz rates (or more), and the optical interconnection delays are negligible, the permutation operation could be performed extremely rapidly. Such a processor design would be very efficient in implementing the FFT or Batcher's sorting algorithm.

An alternative approach to the same problem would use waveguides formed in a $LiNbO_3$ substrate, with active switches incorporated in the permutation network. Such a network would be programmable. A further modification might allow detection to be distributed throughout the Si chip, coupling light from $LiNbO_3$ waveguides to detectors via tapping mechanisms, such as grating couplers, evanescent wave coupling, etc.

In the third scheme to be discussed, electrical-to-optical conversion is carried out via external modulation of a uniform optical wavefront. In this scheme, the routing of optical signals is carried out via conventional and holographic optical elements and free-space propagation. The different modulation mechanisms were mentioned earlier along with their speed limitations. Fig. 13 shows the schematic diagram of this approach to optical interconnects, where the modulators are operated in a reflection mode.

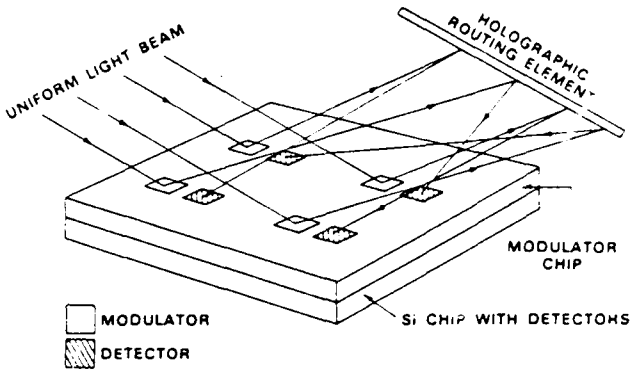


Fig. 13. Optical interconnections using external modulators operating in reflection mode.

The optical elements that perform the routing after the uniform light signals have been modulated could include a dynamic mask, thus introducing programmability in the interconnect patterns. The main advantage of using external modulators is their low power consumption and the relative ease with which components in the interior of the chip can be accessed. It should be noted that the problems encountered here in realizing the electrical-to-optical converters are somewhat similar to those encountered in the development of spatial light modulators. There are significant differences, however, in that the modulation mechanisms required for interconnect schemes need much faster response times (several megahertz versus tens of kilohertz), but need only binary (rather than analog) output (1 bit versus 10 bits).

An alternative approach to the same concept can achieve high-speed operation with low voltage levels via integrated optical modulators. As previously described, such modulators are highly asymmetric, being typically a few millimeters long and only tens of micrometers in width. Due to this shape, such an approach would make most sense when the communication distances involved are more than a few millimeters. The small widths can be exploited to fabricate a number of such modulators on a chip in order to provide one modulator for each source of data on the chip. Thus the two-dimensional arrangement of the processing elements in an array will be converted into a linear light distribution. This linear distribution can then be mapped into another linear distribution in an arbitrary fashion (including one-to-many mappings, if needed) via an optical crossbar switch, which can be realized using an optical matrix–vector multiplier, such as has been reported in the literature [48]. When the matrix mask that encodes the interconnect pattern is changeable in real time, a programmable interconnect network of the processing elements in an array is obtained. Fig. 14 shows the schematic diagram of such a system. The main limitation of this scheme will be the total number of components that can be connected this way within a chip. The theoretical maximum will be given by the number of integrated optical modulators that can be accommodated within the linear dimension of the chip. Since the routing is still performed by free-space propagation and focusing/imaging optics, the entire system will be three-dimensional, requiring large volume and a highly stable optical system. A hybrid version can also be envisioned, where the final delivery of the permuted optical data streams

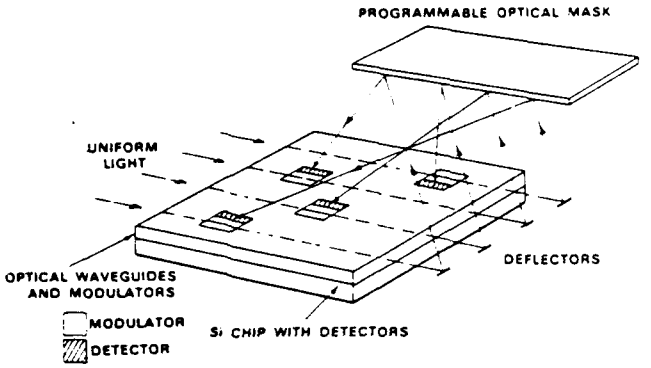


Fig. 14. Optical interconnections using integrated optical modulators.

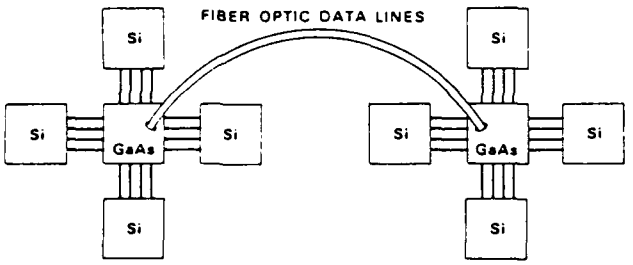


Fig. 15. Configuration with clusters of Si chips around GaAs chips, and with communication between GaAs chips via optical fibers.

to the detectors at appropriate locations on the chip can be performed with permanently bonded optical fibers.

A fourth and final approach to the data-routing problem is a monolithic one that presupposes the successful development of heteroepitaxial growth of GaAs on Si and subsequent construction of devices in both materials. The sources in the GaAs might be of the surface-emitting type. The optical routing could then be performed via free-space propagation and imaging components. The main advantage to such an approach is the optical access it provides to data sources that may be interior to the Si chip, rather than requiring that data be routed electrically to the edge of the chip before being converted into optical form.

2. Interchip Data Communication. Unlike the case of electrical interconnections, an increase in the length of optical data paths does not reduce the bandwidth of the link (all the paths of concern here are so short that optical attenuation is assumed to be insignificant). Therefore, the schemes discussed previously for intrachip communication using free-space propagation or fibers are equally applicable to interchip communication. The arguments are somewhat similar to those made for wafer- and board-level clock distribution with optics. Some advantages of this type of interchip data communication (above and beyond the potential for higher bandwidth and lower mutual interference afforded by optics) are the ability to access interior components of a chip directly and the potential for realizing programmable interconnects using real-time masks. On the other hand, optical data routing using imaging optics and free-space propagation requires careful alignment and uses the third spatial dimension, requiring a larger volume than would otherwise be the case. Therefore, optical fibers may be preferred as the communication medium at the board level.

A novel way of exploiting the large information carrying capacity of fiber optic systems would entail the use of time multiplexing. Thus several electrical connections could be replaced by a single high-speed fiber-optic line. Additional multiplexing and demultiplexing circuitry would be required on a GaAs chip. Conceptually, one can imagine a cluster of Si chips connected to a GaAs chip with short electrical connections. The GaAs chip performs the multiplexing and electrical-to-optical conversion before sending the data over a fiber-optic line to another cluster. The GaAs chip will also have detectors and demultiplexing circuits. The idea is described pictorially in Fig. 15.

VI. Future Directions

In this section our goal is to identify the most likely and profitable directions in which research on optical interconnections might move in the next few years. The view is based on the considerations discussed in previous sections, but inevitably has an element of subjectivity. These opinions should not be construed as ruling out the possibility of a "breakthrough," conceptual or practical, not foreseen in the previous discussions.

A. Clock Distribution

In many respects the problem of optical clock distribution seems to be the easiest to attack with current technology. The problem is a real and serious one. A single off-chip source is needed, and the distribution system can be as simple as global broadcast, with no focusing of light. However, presumably the extra efficiency of a focused distribution system will result in higher detected current and voltage levels, thus minimizing the need for extra amplification on the chip. The tradeoffs between optical efficiency and amplification needed on chip are interesting ones and worthy of future study.

Some experience with optically supplied clocks is needed before the benefits and drawbacks of the idea can be fully understood. Probably a combination of a macroscopic optical clock distribution feeding a collection of more local electrical distribution systems on the chip will be optimal. If unfocused broadcast is to be used, opaque overcoating layers will be needed to prevent introduction of unwanted signals at locations other than detectors. Experimentation with these ideas is badly needed.

If focused distribution systems are to be used, methods for relieving alignment problems, such as that illustrated in Fig. 10, must be developed. Further optical studies of the efficiencies of holographic optical elements in the infrared would be helpful.

B. Hybridization

Many of the proposed interconnect techniques rely on the mating of Si and GaAs chips. In some cases the GaAs provides only the optical source, but in others a more complex optoelectronic chip is needed. Initial efforts could focus on previously developed techniques for chip placement and bonding of the two dissimilar types of chip. However, to maximize the efficiency of relieving the interconnect and pin-out problems in large-scale Si circuits the development of techniques that minimize the length of

interchip interconnect lines, and in effect eliminate the bonding pads, are needed. The pads are considerably larger than the device lines, and pins on the conventional ceramic package are even larger, as described previously. Two examples of approaches that could be pursued are a metal-bump technique and a close-packed planar hybrid. In the former technique, the electrical and optoelectronic chips would be vertically mated by a series of metal posts formed on one of the circuits. This technique has been used, for example, to connect infrared detector arrays to Si chips [55]. In the latter approach, the chips could be butted up to one another and interchip metallization formed on an intervening planarizing dielectric [56]. Manufacturable versions of these technologies could allow both more interconnects to the Si and the efficient utilization of GaAs chips.

C. Monolithic Approaches

The integrating on GaAs of both optical and electrical devices is an exciting area currently in early development, and progress was summarized in Section IV-B. Progress in integrated optoelectronics will be paced by the development of GaAs ICs. If ICs with medium-scale integration or higher complexity can be fabricated with high yield, there may be strong motivation to use optics for interconnects in all-GaAs high-speed or radiation-tolerant processors.

A long-term and conceptually appealing technique for realizing optical interconnects is to grow single-crystal GaAs on Si and then form the appropriate devices in each material, thereby eliminating the need for hybridization. In addition to optoelectronic devices, GaAs electronic devices could also be formed (e.g., high-speed digital circuits for multiplexing signals to Si circuits). This technology could lend itself to many circuit configurations, including the formation of strategically located GaAs islands on a large Si wafer. To date, initial efforts have focused on materials issues, as discussed in Section IV-D. The demonstration of heteroepitaxial circuits, of course, requires the careful determination of process-compatible approaches to Si circuit formation, GaAs layer growth, and GaAs fabrication. For example, at present the GaAs growth occurs at temperatures of $\approx 680°C$, suggesting that higher temperature Si processing steps may have to occur prior to GaAs growth. This technique could, in the long run, realize in single-chip form many of the GaAs/Si hybrid approaches described earlier.

D. Perfect-Shuffle Exchange Network

The realization of a global but fixed interconnect network like the perfect shuffle with optoelectronic techniques could have a large impact on special-purpose, high-performance signal processing systems. The scheme discussed in Section V is just one possible approach. If the processing elements on the VLSI chip are not arranged in a simple linear form, as assumed in that scheme, then the use of $LiNbO_3$ guided-wave modulators to perform the electrical-to-optical signal transduction, as shown in Fig. 14, could be used. The optical permutation network chip consisting of linear arrays of sources and detectors, as well as the passive network of optical waveguides, can then be physically dissociated from the Si chip. In this way the processing elements could all be on one chip, or could be distributed among different chips. In any case, the components that need most development are the individual addressable arrays of laser diodes and $LiNbO_3$ guided-wave modulators that can be directly addressed by a Si chip at high speed.

E. Electrically Addressed Optical Modulators

One of the most important elements of an optical data interconnect system is the external modulator that can operate on guided or free-space optical wavefronts and can be directly addressable by a Si IC. Although several schemes have been proposed and demonstrated, as reviewed in Section IV, special attention is needed to make these modulators compatible with VLSI circuits with respect to size, power consumption, and speed. A matrix-addressed light modulator can also find application in dynamic interconnect schemes involving real-time masks. Depending on the applications, the update rate of these masks could be significantly lower than the clock rates of the VLSI circuits. On the other hand, these masks will be required to have a rather large space–bandwidth product in order to provide high interconnect flexibility. Thus in general future work will be required in developing high-speed, low-power, and small-size modulators.

F. High-Speed ICs

The focus of this paper has been on alleviating the interconnect problem for VLSI MOS circuits which at present operate at clock rates no greater than 30 MHz. A potentially equally important area for optical interconnects may be for smaller scale circuits that operate at very high speeds (100 to 2000 MHz). Si ECL circuits are commercially available with clock rates greater than 100 MHz, and gigahertz circuits in short-gate Si MOS and GaAs MESFET technology have been demonstrated. While the gate counts of such circuits are at least 1 to 2 orders of magnitude smaller than for VLSI chips, their higher speed places severe stress on the isolation and capacitance of the interconnects. In such cases, the inherent high speed and low crosstalk of optical approaches is very appealing. While there are cases where incorporating an optimum number of gates on a chip can minimize the number of high-speed interconnects [57], optics could still play a significant role. Many of the hybridization techniques described earlier could be applied to these high-speed situations. Hybridization of optical chips with fast complex Si circuits is especially attractive due to the recent report of integration of fast CMOS circuits with bipolar circuitry [58]. This achievement could eliminate the need for scaling up the MOS devices (and thus using wafer area and potentially slowing down the devices) in order to drive the optical sources. The most appealing long-term solution for high-speed applications may be the development of all-GaAs systems, since obviously the optical devices could then be integrated, as described above.

ACKNOWLEDGMENT

This paper was the result of an Army Research Office Palantir study. These studies address the physical founda-

tions of approaches to solutions of important technological problems with the aim of stimulating new avenues for progress toward their solution. The participants in the study consisted of the authors of this paper. S. Y. Kung served as chairman. The authors thank Dr. B. Guenther of ARO for his sponsorship and assistance. One of the authors (F. J. L.) wishes to acknowledge partial support by the Department of the Air Force.

REFERENCES

[1] *Proc. IEEE* (Special Section on Acoustooptic Signal Processing), vol. 69, no. 1, pp. 3–5, 48–118, Jan. 1981.

[2] N. J. Berg and J. N. Lee, *Acousto-Optic Signal Processing*. New York, NY: Marcel Dekker, 1983.

[3] L. J. Cutrona, E. N. Leith, L. J. Porcello, and W. E. Vivian, "On the application of coherent optical processing techniques to synthetic aperture radar," *Proc. IEEE*, vol. 54, no. 8, pp. 1026–1032, 1966.

[4] A. Kozma, E. N. Leith, and N. G. Massey, "Tilted plane optical processor," *Appl. Opt.*, vol. 11, no. 8, pp. 1766–1777, 1972.

[5] For a review of this area, see H. J. Caulfield, J. A. Neff, and W. T. Rhodes, "Optical computing: The coming revolution in optical processing," *Laser Focus*, vol. 19, no. 11, pp. 100–110, 1983.

[6] D. Psaltis, D. Casasent, D. Neft, and M. Carlotto, "Accurate numerical computation by optical convolution," *Proc. SPIE*, vol. 232, pp. 160–167, 1980.

[7] A. Huang, Y. Tsunoda, J. W. Goodman, and S. Ishihara, "Some new methods for performing residue arithmetic operations," *Appl. Opt.*, vol. 18, no. 2, pp. 160–167, 1980.

[8] C. Y. Yen and S. A. Collins, Jr., "Operation of a numerical optical processor," *Proc. SPIE*, vol. 232, pp. 160–167, 1980.

[9] A. A. Sawchuk and T. C. Strand, "Fourier optics in nonlinear signal processing," in *Applications of Optical Fourier Transforms*, H. Stark, Ed. New York, NY: Academic Press, 1982, ch. 9.

[10] C. Warde, A. M. Weiss, and A. D. Fisher, "Optical information processing characteristics of the microchannel spatial light modulator," *Appl. Opt.*, vol. 20, no. 12, pp. 2066–2074, 1981.

[11] K. C. Saraswat and F. Mohammadi, "Effect of scaling of interconnections on the time delay of VLSI circuits," *IEEE Trans. Electron Devices*, vol. ED-29, no. 4, pp. 645–650, 1982.

[12] R. W. Keyes, "Communication in computing," *Int. J. Theor. Phys.*, vol. 21, nos. 3/4, pp. 263–273, 1982.

[13] A. J. Blodgett, Jr., "Microelectronic packaging," *Sci. Amer.*, vol. 249, no. 7, pp. 86–96, 1983.

[14] C. Mead and L. Conway, *Introduction to VLSI Systems*. Reading, MA: Addison-Wesley, 1980.

[15] D. J. Kinniment, "VLSI and machine architecture," in *VLSI Architecture*, B. Randell and P. C. Treleaven, Eds. Englewood Cliffs, NJ: Prentice-Hall, 1983, pp. 24–33.

[16] F. Anceau and R. Ries, "Design strategy for VLSI," in *VLSI Architecture*, B. Randell and P. C. Treleaven, Eds. Englewood Cliffs, NJ: Prentice-Hall, 1983, pp. 128–137.

[17] R. W. Keyes, "Fundamental limits in digital information processing," *Proc. IEEE*, vol. 69, no. 2, pp. 267–278, 1981.

[18] P. M. Solomon, "A comparison of semiconductor devices for high-speed logic," *Proc. IEEE*, vol. 70, no. 5, pp. 489–509, 1982.

[19] G. W. Preston, "The very large scale integrated circuit," *Amer. Scientist*, vol. 71, no. 5, pp. 466–472, 1983.

[20] R. W. Keyes, "The evolution of digital electronics towards VLSI," *IEEE Trans. Electron Devices*, vol. ED-26, no. 4, pp. 271–278, 1979.

[21] S. Y. Kung, K. S. Arun, R. J. Gal-Ezer, and D. V. Bhaskar Rao, "Wavefront array processor: language, architecture, and applications," *IEEE Trans. Comput.* (Special Issue on Parallel and Distributed Computers), vol. C-31, no. 11, pp. 1054–1066, 1982.

[22] P. A. Gilmore, "The massively parallel processor (MPP) A large scale SIMD processor," *Proc. SPIE*, vol. 431, pp. 166–174, 1983.

[23] R. A. Athale, "Optical matrix algebraic processors," in *Proc. 10th Int. Optical Computing Conf.*, IEEE Cat. No. 83CH1880-4, Apr. 1983, pp. 24–31.

[24] H. T. Kung, "Why systolic architectures?" *IEEE Computer*, vol. 15, no. 1, pp. 37–46, 1982.

[25] R. S. Schreiber, "On the systolic arrays of Brent, Luk, and van Loan," *Proc. SPIE*, vol. 431, pp. 72–76, 1983.

[26] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing* Englewood Cliffs, NJ: Prentice-Hall, 1975, pp. 284–337.

[27] H. S. Stone, "Parallel processing with the perfect shuffle," *IEEE Trans. Comput.*, vol. C-20, no. 2, pp. 153–161, 1971.

[28] K. E. Batcher, "Sorting networks and their applications," in *1968 Spring Joint Comp. Conf., AFIPS Proc.*, vol. 32 (Washington, DC, 1968).

[29] See, for example, H. Kressel, *Semiconductor Devices for Optical Communications*. Berlin, Germany: Springer-Verlag, 1982.

[30] D. R. Scifres, R. D. Burnham, and W. Streifer, "Output coupling and distributed feedback utilizing substrate corrugations in double heterostructure GaAs lasers," *Appl. Phys. Lett.*, vol. 27, no. 5, 1975.

[31] K. Iga, H. Soda, T. Terakado, and S. Shimizu, "Lasing characteristics of improved GaInAsP/InP surface emitting injection lasers," in *Proc. 4th Int. Conf. on Integrated Optics and Optical Fiber Communications* (Tokyo, Japan, June 27–30, 1983), pp. 198–199.

[32] H. Grothe, G. Muller, and W. Harth, "560 Mb/s transmission experiments using 1.3 μm InGaAsp/InP LED," *Electron. Lett.*, vol. 19, no. 22, pp. 909–911, 1983.

[33] R. A. Becker, "Broad-band guided-wave electrooptic modulators," to be published in *IEEE J. Quantum Electron.*, vol. QE-20, pp. 723–727, July 1984.

[34] R. A. Boenning, V. B. Morris, and E. G. Vaerewyck, "PC board test signal extraction," presented at the 29th Int. Instrumentation Symp., Instrument Society of America, Albuquerque, NM, May 1983.

[35] R. V. Johnson, D. L. Hecht, R. A. Sprague, L. N. Flores, D. L. Steinmetz, and W. D. Turner, "Characteristics of the linear array total internal reflection (TIR) electrooptic spatial light modulator for optical information processing," *Opt. Eng.*, vol. 22, no. 6, pp. 665–674, 1983.

[36] K. E. Petersen, "Micromechanical light modulator array fabricated on silicon," *Appl. Phys. Lett.*, vol. 31, p. 521, 1977.

[37] D. R. Pape and L. J. Hornbeck, "Characteristics of the deformable mirror device for optical information processing," *Opt. Eng.*, vol. 22, no. 6, pp. 675–681, 1983.

[38] R. A. Kingston, B. E. Burke, K. B. Nichols, and F. J. Leonberger, "Spatial light modulation using electroabsorption in a GaAs charge-coupled device," *Appl. Phys. Lett.*, vol. 41, no. 5, pp. 413–415, 1982.

[39] J. N. Walpole, MIT Lincoln Laboratory.

[40] M. L. Levy, "An investigation of flaws in complex CMOS devices by a scanning photoexcitation technique," in *Proc. 15th Annu. IEEE Reliability Symp.*, pp. 44–53, 1977.

[41] R. K. Jain and D. E. Snyder, "Switching characteristics of logic gates addressed by picosecond light pulses," *IEEE J. Quantum Electron.*, vol. QE-19, no. 4, pp. 658–663, 1983.

[42] B. J. Chang, "Dichromated gelatin as a hologram storage medium," *Proc. SPIE*, vol. 177, pp. 71–81, 1979.

[43] K. Biedermann, "Silver halide photographic materials," in *Holographic Recording Materials*, H. M. Smith, Ed. Berlin, Germany: Springer-Verlag, 1977, p. 69.

[44] Y. Fujii, M. Suzuki, and J. Minowa, "A 100 input, output star coupler composed of low-loss slab waveguide," in *Proc. 4th Int. Conf. on Integrated Optics and Optical Fiber Communication*, (Tokyo, Japan, June 27–30, 1983), paper 29C2-4, pp. 342–343.

[45] See, for example, *Tech. Digest of the Topical Meeting on Guided Wave Optics* (Asilomar, CA), IEEE Cat. No. 82CH1719-4, 1982.

[46] S. Dutta, H. E. Jackson, and J. T. Boyd, "Scattering loss reduction in ZnO optical waveguides by laser annealing," *Appl. Phys. Lett.*, vol. 39, no. 3, pp. 206–208, 1981.

[47] W. E. Ross, D. Psaltis, and R. H. Anderson, "Two-dimensional Magneto-optic spatial light modulator for signal processing," *Proc. SPIE*, vol. 341, p. 191, 1982.
[48] J. W. Goodman, A.R. Dias, and L. M. Woody, "Fully parallel, high speed incoherent optical method for performing the discrete Fourier transform," *Opt. Lett.*, vol. 2, no. 1, pp. 1–3, 1978.
[49] M. Kondo, Y. Ohta, M. Fujiwara, and M. Sakaguchi, "Integrated optical switch matrix for single-mode fiber networks," *IEEE J. Quantum Electron.*, vol. QE-18, no. 10, pp. 1759–1765, 1982.
[50] M. Kim, C. Hong, D. Kasemet, and R. Milano, "GaAlAs/GaAs integrated optoelectronic transmitter using selective MOCVD epitaxy and planar ion implantation," in *Proc. GaAs IC Symp.*, pp. 44–47, Oct. 1983.
[51] J. Carnev, M. Helixal, and R. Kolby, "Gigabit optoelectronic transmitter," in *Proc. GaAs IC Symp.*, pp. 48–51, Oct. 1983.
[52] For a review, see N. Bar-Chaim, S. Margalit, A. Yariv, and I. Ury, "GaAs integrated optoelectronics," *IEEE Trans. Electron Devices*, vol. ED-29, no. 9, pp. 1372–1381, 1982.
[53] B. Y. Tsaur, M. W. Geis, J. C. C. Fan, and R. P. Gale, "Heteroepitaxy of vacuum-evaporated Ge films on single-crystal Si," *Appl. Phys. Lett.*, vol. 38, no. 10, pp. 779–781, 1981.
[54] B. Y. Tsaur, R. W. McClelland, J. C. C. Fan, R. P. Gale, J. P. Salerno, B. A. Vojak, and C. O. Bozler, "Low dislocation density GaAs epilayers grown on Ge-coated Si substrates by means of lateral epitaxial overgrowth," *Appl. Phys. Lett.*, vol. 41, no. 4, pp. 347–349, 1982.
[55] D. H. Pommerrenig, D. D. Enders, and T. E. Meinhardt, "Hybrid silicon focal plane development: an update," *Proc. SPIE*, vol. 267, pp. 23–30, 1981.
[56] A. Chu, MIT Lincoln Lab., private communication.
[57] B. Gilbert, T. Kinter, S. Hartley, and A. Firstenberg, "Exploitation of GaAs digital integrated circuits in wideband signal processing environments," in *Proc. GaAs IC Symp.*, pp. 58–61, Oct. 1983.
[58] J. Miyamoto, S. Saitoh, H. Momose, H. Shibata, K. Kanazaki, and S. Kohyama, "A 1.0 μm N-well CMOS/bipolar technology for VLSI circuits," in *Tech. Dig. Int. Electron Device Meet.*, pp. 63–66, Dec. 1983.

END

FILMED

1-86

DTIC