

AD-A143 994

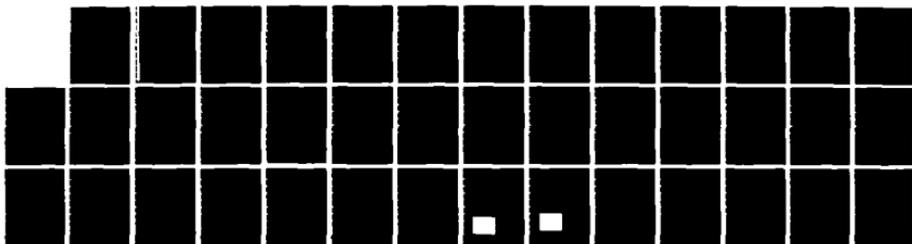
OPTICAL COMPUTING RESEARCH(U) STANFORD UNIV CA
INFORMATION SYSTEMS LAB J W GOODMAN ET AL JUN 84
ISL-L722-9 AFOSR-TR-84-0632 AFOSR-83-0166

1/1

UNCLASSIFIED

F/G 20/6

NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963-A

AFOSR-TR- 84-0632

INFORMATION SYSTEMS LABORATORY



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

AD-A143 994

OPTICAL COMPUTING RESEARCH

Joseph W. Goodman
Moshe Nazarathy
Qizhi Cao
Raymond Kostuk
Ellen Ochoa
Rae-Hong Park

June 1984

This manuscript is submitted for publication with the understanding that the United States Government is authorized to reproduce and distribute reprints for governmental purposes.

Annual Technical Report Number L722-9

Research supported by the Air Force Office of Scientific Research, Air Force Systems command, USAF, under Grant No. AFOSR 83-0146. The United States Government is authorized to reproduce and distribute reprints for governmental purposes, notwithstanding any copyright notation hereon

SEP 11 1984
AUG 3 1984
A

Approved for public release
distribution unlimited

OTIC FILE COPY

84 08 07 118

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION <i>Unclassified</i>		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) Annual Tech Report Number L722-9	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) Annual Tech Report Number L722-9		5. MONITORING ORGANIZATION REPORT NUMBER(S) AFOSR-TR. 84-0632	
6a. NAME OF PERFORMING ORGANIZATION Standard University	6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION AFOSR/NE	
6c. ADDRESS (City, State and ZIP Code) Dept. of Electrical Engineering Stanford Electronics Laboratories Stanford, CA 94305		7b. ADDRESS (City, State and ZIP Code) Bldg 410 Bolling AFB, DC 20332	
8a. NAME OF FUNDING/SPONSORING ORGANIZATION AFOSR	8b. OFFICE SYMBOL (If applicable) NE	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER <i>AFOSR-83-0166</i>	
8c. ADDRESS (City, State and ZIP Code) Bldg 410 Bolling AFB, DC 20332		10. SOURCE OF FUNDING NOS.	
11. TITLE (Include Security Classification) OPTICAL COMPUTING RESEARCH		PROGRAM ELEMENT NO. 61102F	PROJECT NO. 2305
		TASK NO. B1	WORK UNIT NO.
12. PERSONAL AUTHOR(S) Joseph W. Goodman, Moshe Nazarathy, Qizhi Cao, R. Kostuk, E. Ochoa, R. Park			
13a. TYPE OF REPORT Annual	13b. TIME COVERED FROM 18 MAR 83 TO 17 MAY 84	14. DATE OF REPORT (Yr., Mo., Day) JUNE 1984	15. PAGE COUNT 11
16. SUPPLEMENTARY NOTATION			
17. COSATI CODES		18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)	
FIELD	GROUP	<i>OPTICAL interconnections, non-linear optics, optical signal processing, photorefractive materials</i>	
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 March 1983 through 17 May 1984. The work covers several different areas of optical computing, as well as some work on digital processing of images. The primary emphasis of the work is on applications of optics to interconnections in the area of microelectronics. Other areas include diagonalization and inversion of circulant matrices, inversion of wavefronts using photorefractive crystals, suppression of speckle in coherently formed images, and data processing using dispersive anisotropic crystals. Publications during the last year arising out of the grant are also detailed.			
20. DISTRIBUTION/AVAILABILITY OF ABSTRACT UNCLASSIFIED/UNLIMITED <input type="checkbox"/> SAME AS RPT. <input type="checkbox"/> DTIC USERS <input type="checkbox"/>		21. ABSTRACT SECURITY CLASSIFICATION	
22a. NAME OF RESPONSIBLE INDIVIDUAL Lt Col Robert Carter	22b. TELEPHONE NUMBER 202-767-4934	22c. OFFICE SYMBOL NE	

OPTICAL COMPUTING RESEARCH

Joseph W. Goodman

Moshe Nazerathy

Qizhi Cao

Raymond Kostuk

Ellen Ochoa

Rae-Hong Park

June 1984

This manuscript is submitted for publication with the understanding that the United States Government is authorized to reproduce and distribute reprints for governmental purposes.

Annual Technical Report

Report Number L722-9

Research supported by the Air Force Office of Scientific Research, Air Force Systems Command, USAF, under Grant No. AFOSR ~~83-0000~~⁰¹⁶⁶. The United States Government is authorized to reproduce and distribute reprints for governmental purposes, notwithstanding any copyright notation hereon.

AIR FORCE
OFFICE OF SCIENTIFIC
RESEARCH
AFOSR
83-0166
L722-9
JUN 1984

Information Systems Laboratory
Stanford Electronics Laboratory
Stanford University, Stanford, California

ABSTRACT

This document contains information on the research accomplished under AFOSR Grant No. AFOSR 83-0166 during the time period 18 March 1983 through 17 May, 1984. The work ^{work on} covers several different areas of optical computing, as well as some work on digital processing of images. The primary emphasis of the work is on applications of optics to interconnections in the area of microelectronics. Other areas include diagonalization and inversion of circulant matrices, inversion of wavefronts using photorefractive crystals, suppression of speckle in coherently formed images, and data processing using dispersive anisotropic crystals. Publications during the last year arising out of the grant are also detailed.



Accession For	
NIS GRA&I	<input checked="" type="checkbox"/>
Dist TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	<input type="checkbox"/>
Distribution/	
Availability Codes	
Avail and/or	
Dist	Special
A-1	

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 March 18, 1983 through May 17, 1984. The research performed lies in five major areas: (1) Optical interconnections; (2) Nonlinear optical information processing using four-wave mixing; (3) Suppression of speckle in coherently formed images; (4) Diagonalization and inversion of circulant and Toeplitz matrices using coherent optics; and (5) Information processing in linear birefringent dispersive materials. We summarize the progress in each of these areas (Sections II through VI). In addition we present other information pertinent to the grant in the final section.

II. OPTICAL INTERCONNECTIONS

The primary and most important project under way is an investigation of the possible applications of optics to electronic interconnection problems. 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 buses 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. Work is aimed at both

the conceptual level and at the practical level. In the conceptual area, we are attempting to discover the most important scenarios in which optical interconnections make sense. A good summary of the level of our current conceptual understanding is contained in the preprint of the paper entitled "Optical interconnections in microelectronics", which is attached as Appendix 1. This paper will be published in the *Proceedings of the SPIE* volume entitled *Optical Computing: a Critical Review of Technology*.

At the practical level, we have embarked upon two different tasks. One is to evaluate the use of optical interconnections as a means for communication between chips. The other, just getting under way, is an investigation of clock distribution within a single chip using optics. Since the latter project is just starting, we concentrate our comments here on the progress in the former area.

Communication between distant locations (i.e. several mm to a few cm) on planar substrates is a limiting factor of VLSI system performance. The high density of input and output terminals of integrated subunits (CPU's, buffers, ROM's, etc.) present serious pad bonding and signal drive power difficulties. In addition, the point-to-point communications restrictions imposed by conventional electronic interconnection schemes encompers algorithm design.

The particular interconnect problem addressed in this work is described as follows. Two electronic integrated circuit units fabricated on a substrate with an array of input/output bonding pads are linked by deposited aluminum or a hybridized conducting material. The bonding pads are typically 125 micrometers square. The average power required to send a signal over a terminated line in this situation is about 40 milliwatts for a 2 volt signal with a 50 ohm termination. To date the highest modulation rates for arrangements of this type are

about 100 MHz. Projected required transmission modulation rates are on the order of tens of GHz. Our goal is to investigate the feasibility of an optical replacement for the interconnection just described.

The optical version of the interconnect geometry is described as follows. An information-bearing signal on one chip drives a light source; the emitted flux is collected by an imaging element and imaged onto a detector on the second chip, where it is converted to an electronic signal. Information transfer between the two chips thus takes place optically, rather than electronically.

An experimental system for testing the concept was constructed. It uses a Litronix R513 standard red LED with peak emission wavelength 660 nm in a 20 nm bandwidth. Approximately 70 mw of electrical power is required to produce an optical intensity of 60 microwatts per steradian. The detectors are Hewlett-Packard photodiodes from a high-speed opto-coupler. The active area of the photodiode is about 280 micrometers square, while the area of the emitter is about 240 micrometers square. Imaging is achieved with a reflection volume phase hologram formed on an Agfa-Gaevert 8E75 III emulsion. A pyrogallol-sodium carbonate developer (Agfa #GP-62) and a potassium bromide p-benzoquinone bleach (GP-432) were used. The holographic element was formed with on-axis converging and diverging spherical beams from a HeNe laser. The source and image points were about 4.5 cm from emulsion plane producing a beam overlap region of about 1.5 cm diameter.

When illuminated with the 660 nm LED, the peak hologram diffraction efficiency was about 5.5%. Image irradiance and resolution were at off-axis illumination angles were evaluated with a CCD line scanner. Image resolution at 1.0 cm varied from from 270 micrometers at 7 mm source-detector separation to

330 micrometers at 30 mm separation. Astigmatism was the predominant aberration at large angles of incidence. At an 18 degree angle of incidence, image intensity decreased to 50% of the value at a 9 degree angle of incidence. Thus 18 degrees is considered the useful field-of-view of this hologram.

Two source-detector mounting geometries were tested. In the first, an LED and detector diode were separated by 90 micrometers distance corresponding to typical bonding pad separations on an IC. The purpose of this geometry was to evaluate crosstalk between a source and a nearby detector on a single chip. A second test configuration with a source-detector separation of 1.1 mm was also used to simulate an interconnection between two different IC units. Unwanted scattered light was found to be sufficiently low in level to not pose a difficulty for the system. The limits to system performance posed by internal receiver noise are now being evaluated. A major effort in the future must be invested in the design and realization of low-noise receiver circuitry suitable for realization on integrated circuit chips.

The work at Stanford on optical interconnections and supported by AFOSR has received considerable recognition in the scientific community. As mentioned previously, an invited paper was presented at the SPIE critical review of optical computing technology. An invited paper will be published in the July 1984 issue of the *Proceedings of the IEEE*. The keynote address of the 13th General Meeting of the International Commission for Optics will be presented on this subject by Prof. Goodman in August 1984, and the first R.V. Pole Memorial Lecture will be presented on this topic by Prof. Goodman at the 1984 CLEO meeting in June 1984.

III. NONLINEAR INFORMATION PROCESSING USING PHOTOREFRACTIVE MATERIALS

Photorefractive materials are of interest in image processing because of their holographic analog to conventional Fourier optical processing systems. Other groups have reported using two-wave and four-wave mixing techniques in photorefractive crystals to perform optical phase conjugation, image convolution and correlation, edge enhancement, image amplification, and image division. We are interested in image inversion and have been investigating the characteristics of photorefractives which generate and limit the inversion process.

The origin of our interest in image inversion using photorefractive materials rests on the unusual signal processing operations that can be performed with such an effect. Image deblurring, code translation, detection of non-periodic structures in a periodic field, detection of periodic structures in a non-periodic field, and diagonalization and inversion of circulant matrices (see section V) are all examples of operations that can be performed if a method for image inversion, or more properly, wavefront inversion, is realized.

In a photorefractive medium, a space-charge field is generated when trapped charges are excited by an intensity-interference pattern and subsequently retrapped after transport by drift and diffusion. This field is proportional to the modulation depth, m , of the intensity grating, unless m approaches unity. If I_o represents the intensity of the object beam and I_r represents the intensity of the reference beam, then

$$m = \frac{2\sqrt{I_o I_r}}{I_o + I_r}$$

Through the electro-optic effect, the change in refractive index has the same

dependence as the space-charge field. A third beam performs read-out of the grating in real-time. When plane waves are used as the input beams, the resulting steady-state expression for diffraction efficiency is

$$\eta = \sin^2 \left(\frac{\pi \Delta n d}{\lambda \cos \theta} \right)$$

where

$$\Delta n = \frac{-m}{2} r_{eff} n_o^3 E_i \left[\frac{E_o^2 + E_j^2}{E_o^2 + (E_d + E_o)^2} \right]^{\frac{1}{2}}$$

In the above expressions,

$$E_o = \frac{q N_o}{\epsilon k} \quad ; \quad \text{the maximal space charge field}$$

$$E_d = \frac{k k_B T}{q} \quad ; \quad \text{the diffusion field}$$

E_o = applied electric field

N_o = concentration of traps

ϵ = dielectric constant

k = magnitude of grating wave vector

k_B = Boltzmann's constant

r_{eff} = effective electro-optic coefficient

n_o = appropriate index of refraction

The expression for diffraction efficiency can, for small argument, be approximated by the square of the argument; thus, diffraction efficiency is proportional to the square of the modulation depth. Because m depends only on relative intensities between the object beam and the reference beam, and not on absolute intensity of either beam, variation of the beam ratio away from unity causes diffraction efficiency to drop. At appropriate beam ratios, therefore, image inversion can occur. In certain cases, it is appropriate to use an effective modulation depth which includes effects of absorption of the crystal, α , and total light intensity incident on the crystal, I_T . Here,

$$m_{eff} = \frac{m}{1 + \frac{\beta}{\omega I_T}}$$

and β is dependent on crystal parameters.

Our experimental set-up uses Argon laser green light to write a grating in BSO or BGO, and a He-Ne laser is used to read it out. Minor adjustments need to be made to the formulas to account for a different read-out wavelength. We compare the experimental data with theoretical calculations. We have been exploring the parameters which affect the dynamic range of the beam ratio over which inversion occurs. In simulating the two-beam coupling process, we have shown that varying the electro-optic coefficient changes the diffraction efficiency at a constant beam ratio as well as the range of the beam ratio where inversion is seen. The concentration of donors and traps is another crystal parameter which affects diffraction efficiency and inversion. When keeping all crystal parameters fixed, inversion is influenced by features of the experimental set-up. In particular, it appears to be possible to "tune" the inversion dynamic range by varying an applied electric field across the crystal. Total light intensity reaching the crystal may be another tuning parameter. Once the two-beam process is well understood, we plan to use four-wave mixing techniques to invert images.

IV. SUPPRESSION OF SPECKLE IN COHERENTLY FORMED IMAGES

All coherently formed images, including those obtained from side-looking radar systems, laser-illuminated imaging systems, and medical ultrasound, contain a disturbing noise known as speckle. Speckle arises whenever the coherence length of the illuminating radiation is long compared to the surface roughness of the object to be imaged. For some time we have been involved in an evaluation

of a wide variety of techniques for suppressing speckle by means of digital image processing. Our work provides the first quantitative comparison of these diverse methods, and also introduces some new techniques that have never been tried before.

Over the past 20 years or so, many researchers have tried many different methods for suppressing speckle by post-detection filtering. Popular methods include (1) simple smoothing of the picture with linear filters; (2) homomorphic filtering consisting of a logarithmic transformation, followed by linear filtering, followed by an exponential transformation; and (3) two-dimensional median filtering. Our work has compared all of these methods for certain objects with mean-square error as a quality criterion. In addition, we have introduced new methods which include: (1) nonlinear one-dimensional filtering in a projection space; and a maximum entropy method.

The results of the work have identified what we believe to be the most promising approaches. Of the methods studied, most consistently good results were obtained using one dimensional square root filtering (i.e. a square root non-linearity, followed by linear filtering, followed by a square-law transformation) in a projection space.

Three papers have been prepared on the results of this work, and effective June 1984, the project has been terminated. The papers will be submitted to various journals in the near future.

V. DIAGONALIZATION AND INVERSION OF CIRCULANT MATRICES BY COHERENT OPTICS

For some time we have been studying, both theoretically and experimen-

tally, some methods for diagonalizing and inverting circulant matrices using coherent optics. Inversion of circulant approximations to Toeplitz matrices is a problem that arises frequently in optimal filtering problems. Using the methods we have been studying, it is possible to invert as many as several hundred circulant matrices, each of dimension several hundred by several hundred, simultaneously on one pass through a coherent optical system. Such a capability may have application to estimation problems involving antenna arrays with many elements, with many resolvable frequency bands in each antenna lobe.

The basic idea behind this work has recently been published in *Applied Optics*. Rather than repeat that material here, we have attached a reprint of the paper as Appendix 2. Since the publication of that paper, we have pushed the work further, successfully inverting circulant matrices, using photographic film as the nonlinear element needed to accomplish such inversion. Ultimately interest rests in the use of real-time nonlinear elements for accomplishing this inversion. Indeed the work on wavefront inversion using photorefractive crystals described in Section III is directly applicable to this problem.

A paper is now being written describing the further experiments that resulted in actual matrix inversions. This paper will complete our work on this subject, and the project will be terminated at the end of June 1984.

VI. INFORMATION PROCESSING IN LINEAR BIREFRINGENT DISPERSIVE MATERIALS

An area of recent interest to us has been the study of optical propagation in linear, birefringent, dispersive media, and the possibility of performing useful signal processing operations with devices based on such media. It has been shown through this work that in the process of propagation of a space-time wavefield in

such a medium, a cross-ambiguity function of the spatial and temporal parts of the input distribution is generated, provided the dispersion relations of the medium exhibit a certain coupling between temporal dispersion and spatial anisotropy. This new result opens new avenues for novel signal processing methods in the domain of ultrafast optics.

Essentially, the quality to be possessed by materials which generate the ambiguity function is the following: axial dispersion of temporal wavepackets varies with direction of propagation within the crystal. Thus the group velocity is anisotropic and the materials exhibit anisotropic dispersion. An alternative description is that the Poynting vector walkoff angle for the required materials exhibits dispersive anisotropy, in the sense that its spatial anisotropy characteristics change with temporal frequency.

To pursue the idea to the practical implementation of an optical processor, much work remains to be done in identifying suitable crystals with dispersive anisotropy (or anisotropic dispersion) compatible with the requisite time resolution. The significance of the current results for the domain of ultrafast optics remains to be evaluated. A particularly attractive prospect consists of converting the convolutional ambiguity function processor into a temporal time-invariant filter which could compute convolutions with an impulse response that is spatially preset by means of a spatial modulator.

VII. MISCELLANEOUS INFORMATION

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

1. Professor Joseph W. Goodman, Principal Investigator - overall supervision.
2. Dr. Moshe Nazarathy, Weizmann Foundation Post-Doctoral Fellow - Information processing in linear birefringent dispersive materials.
3. Ms. Qizhi Cao, Graduate Student Research Assistant - Diagonalization and inversion of circulant matrices.
4. Mr. Raymond Kostuk, IBM Doctoral Fellow - Optical interconnections.
5. Ms. Ellen Ochoa, IBM Doctoral Fellow - Wavefront inversion using photorefractive crystals.
6. Mr. Ray-Hong Park, Graduate Student Research Assistant - Suppression of speckle in coherently formed images.

Both Ms. Cao and Mr. Park are receiving their doctorates in June 1984, with their theses based on the work carried out under this grant.

We call special attention to the fact that, due to the support of fellowships from other sources, the actual number of personnel performing research under this grant is actually nearly twice the number of personnel actually supported with grant funds. Thus we have a high degree of leverage in this respect.

The publications on work supported in whole or in part by this grant during the last 14-month grant period are listed as follows:

Published Works

1. Q. Cao and J.W. Goodman, "Coherent optical techniques for diagonalization and inversion of circulant matrices and circulant approximations to Toeplitz matrices," *Applied Optics* **23**, 803-811 (1984).

2. J.W. Goodman, "Architectural development of optical data processing systems," *KINAM (Revista de Physica)*, Serie C, Vol. 5, 9-40 (1983).
3. E. Ochoa and J.W. Goodman, "Statistical distribution of ray directions in a fully developed speckle pattern," *J. Opt. Soc. Am.* 73, 943-949 (1983).
4. J.W. Goodman, "The optical data processing family tree," *Optics News* 10, 25-28 (May/June 1984).

Publications in Press

5. J.W. Goodman, "Optical interconnections in microelectronics," *Proc. SPIE*, Vol. 456.
6. J.W. Goodman, F.J. Leonberger, S.-Y. Kung, and R.A. Athale, "Optical interconnections for VLSI systems," *Proc. I.E.E.E.*, June 1984.
7. M. Nazarathy and J.W. Goodman, "Ambiguity function processors using linear birefringent dispersive media," *Proc. SPIE*, Vol. 465.

Under Submission

8. M. Nazarathy and J.W. Goodman, "Diffraction transforms in homogeneous birefringent media," submitted to *J. Opt. Soc. Am.*

Oral presentations on the items in numbers 4, 5 and 7 above were made at SPIE meetings and Optical Society of America meetings.

We are indebted to Prof. Lambertus Hesselink for the advice he has given on the wavefront inversion project.

APPENDIX I

OPTICAL INTERCONNECTIONS IN MICROELECTRONICS

Joseph W. Goodman

Department of Electrical Engineering
Stanford University
Stanford, California 94305

Abstract

As the complexity of microelectronic circuits increases, performance becomes more and more limited by interconnections. Continued scaling and packing lead to a dominance of interconnect delays over gate delays. This paper explores the potential of optical interconnections as a mean for alleviating such limitations. Various optical approaches to the problem are discussed, including the use of guided waves (integrated optics and fiber optics) and free space propagation (simple broadcast and imaging interconnections). The utility of optics is influenced by the nature of the algorithms that are being carried out in computations. Certain algorithms make far greater demands on interconnections than do others. Clock distribution is a specific application where optics can make an immediate contribution. Data interconnections are more demanding, and require the development of hybrid Si/GaAs devices and/or heteroepitaxial structures containing both Si and GaAs layers. The possibilities for future developments in this area are discussed.

I. Introduction

In the field of computation, optics is currently best known for its role in analog signal processing. Examples include acousto-optic spectrum analyzers, convolvers and correlators [1], [2], and systems for forming images from synthetic-aperture radar data [3], [4]. Analog approaches of this kind offer high processing speed, but low accuracy and limited flexibility in terms of the variety of operations that can be performed. These shortcomings have led to a search for applications of optics to digital [5], [6] and other types of high-accuracy numerical computation [7], [8]. However, regardless of the outcome of the new thrusts towards digital-optical computation, it is virtually certain that the vast majority of computation done in the world for many years to come will be performed by microelectronic chips. While current ideas in digital-optical computing have the potential to strongly impact special-purpose digital signal processing, they are very unlikely to invade a significant fraction of the total computing market in the foreseeable future. It is therefore natural to inquire as to whether there exists another role for optics in digital computing, one that has the potential for greater impact on the overall computing market.

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 new interest in 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. 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 needed nonlinear elements, nonetheless the horizon for electronics is not without clouds. It is generally realized and agreed that the exponential growth of semiconductor chip capabilities can not continue indefinitely, and 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 down and chip areas continue to increase [11], [12]. Indeed, interconnection difficulties extend beyond the chip level, and have major impact at the board level as well [13].

Given the above facts, it seems natural to inquire as to whether optics might offer important capabilities in overcoming the interconnect problems associated with microelectronic circuits and systems. Encouragement is offered by the fact that the very property of optics that makes it difficult to realize nonlinear elements (it is difficult to make two

GOODMAN

streams of photons interact) is precisely the property desired of an interconnect technology. However, at the chip level, optics is likely to be only one element of a hierarchy of interconnect technologies, with optical interconnection networks feeding more conventional interconnect lines constructed from polycrystalline silicon, metal silicides, or metals.

There exists an entire hierarchy of interconnect problems for which the implications of optics should be considered. At one extreme is the problem of machine-to-machine communication (local area networks). Such problems are excluded from our consideration here, due to the fact that so much work is already in progress by others. The next level is subsystem to subsystem communication within a single computer. Much less work has been done on optical approaches to this problem, but at least one example exists [14]. Again we exclude such problems from consideration here, preferring to concentrate on communication problems at lower levels of machine structure. Board-to-board communication with optics can be thought of as the problem of constructing an optical data bus within a machine. Some ideas pertinent to this level of interconnection have been published (see, for example, Ref. [15]), but again we exclude these problems from consideration. Interconnection within a single board (dimensions of approximately 1 ft x 1 ft) is in the realm where our discussions are pertinent, as is likewise the problem of interconnection within a wafer (typical dimension 7.5 cm diameter) or within a single chip (typical dimensions 10 mm x 10 mm). Most of our discussions will refer to the chip-level problem, but many of the ideas can be extrapolated to the wafer and board levels. In addition, most of the discussion will assume that we are dealing with common metal-oxide-semiconductor (MOS) integrated circuit technology.

The purpose of this paper is to stimulate further thought and research on optical interconnections. The ideas presented are rather preliminary and underdeveloped, but hopefully they have the potential to serve as a starting point for others in considering how optics might play a more widespread role in computing of the future. Section II reviews the origins and nature of the electronic interconnect problem from the technological point-of-view. Section III briefly discusses the interaction between interconnections and algorithmic considerations, thus examining the algorithmic side of the motivation for optical interconnects. Section VI discusses the potential benefits that optics might bring to a solution to these problems and outlines two fundamentally different optical approaches to interconnections, the index-guided approach and the free-space propagation approach. Section V specifically addresses the problem of clock distribution, while Section VI deals with the problem of data distribution. Finally, Section VII discusses future developments needed if these ideas are to have practical impact.

II. The Interconnect Problem - Technological Motivation

The growth of integrated circuit complexity and capabilities experienced since the birth of the industry has been achieved through a combination of scaling down 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 packing, i.e. the addition of circuitry in order to realize more complex chips in the same area of silicon that was used before scaling. Here we briefly discuss the good and bad effects of scaling. A more detailed discussion of the subject is found in Ref. [11].

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

Scaling also affects the interconnections between devices. Figure 1 shows the effect of scaling down a conductor by a factor d . Since the cross-sectional area of the conductor is decreased by a factor d^2 , the resistance per unit length will increase by a similar factor. If the length of the conductor is scaled down by d , as simple scaling implies, then the net increase of resistance is proportional to d . 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 d implies a decrease of capacitance by d^2 . However, scaling also implies a decrease by d of the thickness of the oxide insulating layer separating the plates of the capacitor. Hence the capacitance is inversely proportional to the first power of the scaling constant d . We see that the scaling up of resistance and the scaling down of capacitance exactly cancel, leaving the RC time constant unchanged.

Since gate delays scale down with d while interconnect delays remain independent of d ,

it is clear that eventually a point must be reached where interconnection delays dominate device delays. However, the situation is actually much worse than the above considerations imply, due to the fact that packing usually accompanies scaling, and the lengths of the interconnects required do not scale down with d . Rather, as the complexity of the circuit being realized increases, the distances over which the interconnections must be maintained on a chip of fixed area stay roughly constant. Statistical considerations show [13] that a good approximation to the maximum length L_{\max} of the interconnection required is given approximately by

$$L_{\max} = A^{1/2}/2 \quad (1)$$

where A represents the area of the chip. If the area of the chip stays roughly constant, then the maximum interconnect length stays roughly constant, interconnection resistance scales up as d^2 , and interconnection capacitance stays constant, yielding an overall scaling of interconnect delay that increases in proportion to d^2 . Note that if chip size is increased, rather than remaining constant, the interconnection problem becomes further exacerbated. As a consequence of these considerations, it has been estimated that by the late 1980's, MOS chip speeds will be limited primarily by interconnect delays [11].

Another different aspect of scaling is of considerable importance here. We refer to the effects of scaling and packing on the numbers of connections required from the outside world to chips. As the number of elements in a single chip grows, the number of interconnections required from that chip to other chips also increases. There is a well-known empirical relation, known as Rent's rule, which specifies that the number of interconnections M required for a chip consisting of N devices grows as approximately the 0.61 power of N , i.e.

$$M = N^{0.61} \quad (2)$$

However, the perimeter of a chip, to which the connections must be made, grows as only the square root of area, or equivalently as the 0.5 power of N . The disparity caused by the difference of these two exponents becomes more and more important as the number of devices within a chip grows due to scaling and packing. It should be noted that Rent's rule applies only to chips consisting of logic elements. Memory cells require fewer interconnections. In addition, it is required that each chip be a small "random" subset of the entire logic system [13].

The predictions of Rent's rule can also be applied to collections of chips on a board (providing the assumptions mentioned above are satisfied). At the chip level, some of the limitations implied by Rent's rule can be overcome by the 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.

One additional and final implication of scaling should be mentioned. While current scales down inversely with d , the cross-sectional area through which that current must flow scales down inversely with d^2 . The net result is that current density increases in proportion to d . Such an increase leads to increasing effects of electromigration, by which is meant the movement of conductor atoms under the influence of electron bombardment. The ultimate effect of electromigration is the breaking of conductor lines and the failure of the chip. The potential of optical interconnections as a means for alleviating electromigration problems is of considerable interest here.

III. The Interconnect Problem - Algorithmic Motivation

In Section II, some of the technological motivation for considering optical interconnections was presented. In this section we discuss motivations that arise from algorithmic considerations. We describe several classes of problems that place various degrees of burden on interconnections. Examples are drawn primarily from the fields of signal processing and image processing.

The first class of operations considered places the smallest burden on interconnections. We refer to this class as point processing operations. Given a two dimensional array $g(n,m)$ of data points to be processed and a two dimensional array of processors to perform these operations, the result of the processing is described by

$$g(n,m) = NL[f(n,m)] \quad (3)$$

where $NL[]$ is a (generally nonlinear) function that transforms each value $f(n,m)$ into a new value $g(n,m)$, independent of the values of $f(n,m)$ at indices other than (n,m) . Since each $g(n,m)$ depends only on one corresponding $f(n,m)$, no communications are required between processors in the array. Because the interconnections are relatively simple for such problems, it is likely that the only role for optics here might be in loading and unloading the data to and from the processor array. If the processor array is a large one, the number of parallel inputs and parallel outputs needed for maximum efficiency will be large, implying a large pin count for the processing chip. Relief can be found by integrating detectors with each processor, thus providing a parallel set of optical input channels. To achieve similar relief for output data, the more difficult problem of integrating a source with each processor must be solved. Potential paths to solutions of this problem are discussed in later sections.

As a second level of interconnect difficulties, consider the problem of matrix-matrix multiplication. Letting capital letters A, B , and C represent matrices, and lower-case letters with subscripts represent particular elements of those matrices, we wish to consider the computation described by

$$C = A B \quad (4)$$

or

$$c_{ij} = \sum_{k=0}^N a_{ik} b_{kj} \quad (5)$$

Note that a processor computing c_{ij} requires communication from one row of A and one column of B . Thus we might say that the communications required are semi-global (fully global communications would require that each c_{ij} receive data from all elements of A and all elements of B , which is not the case here). In spite of the fact that the communication requirements are semi-global, nonetheless it has been shown [16] that the computation can be performed by a two dimensional array of processors, with each processor connected only to its nearest neighbors (we refer to such processors as being "mesh connected"). If only a single matrix product is to be performed, then the time required by the mesh-connected set of processors is greater than would be the case if the processors were connected with semi-global communications. However, the mesh-connected array can readily be pipelined, and if a large number of matrix-matrix products are to be performed in succession, the pipelined mesh-connected array can do the job in the same time that a set of semi-globally connected processors could achieve. We conclude that for the matrix-matrix multiplication problem, the roles for optics are probably limited to providing data loading and unloading, and providing semi-global communications when the constraints of the problem prevent pipelining.

It is well recognized in the computer science field that there are classes of algorithms that are intensive in their requirements for interconnects [17], [18]. Such algorithms generally require periodic exchange of data between processors in an array, with the exchanges being most efficiently carried out with other than nearest-neighbor interconnections. Often the exchange pattern has a particular structure that allows the exchanges to be carried out with an interconnect network that is not fully general. Examples of useful exchange networks include the shuffle-exchange network (Fig. 2), and the closely related butterfly-exchange network (Fig. 3). These exchange networks play central roles in the FFT algorithm and in certain approaches to sorting problems. Because the interconnections are not nearest-neighbor, optical interconnect techniques may have an important role to play here. Note in particular that the butterfly exchanges could be carried out nicely with a dynamic optical interconnect device that changes its interconnect pattern at the end of each layer of exchanges.

Lastly, as an example of a class of problems with even greater demands on interconnects, we mention the problem of space variant linear filtering. In this case the output array $g(n,m)$ is computed from the input array $f(n,m)$ through the general linear filtering relation

$$g(n,m) = \sum_k \sum_p h(n,m;k,p) f(k,p) \quad (6)$$

where $h()$ is the kernel of the operation. In the most general case, the operation described above is fully global, in that each $g(n,m)$ receives data from every $f(n,m)$. The fact that the transformation is space invariant, as evidenced by the dependence of the kernel on both (n,m) and (k,p) , implies that the interconnect pattern required for the

computation of each output value $g(n,m)$ changes from output point to output point. Hence such operations, to the extent that they are fully global in their nature and to the extent that they require a multitude of different interconnect patterns, place an extremely intensive burden on interconnect technology, and may ultimately be prime targets for applications of optical interconnects.

IV. Optical Interconnections

This Section is devoted to discussing the general beneficial properties of optics as a technology for providing interconnections, and to describing some of the approaches that can be taken to using optics in this role. Properties of optics that are identified as being beneficial are done so with more conventional integrated circuit interconnections and their limitations in mind.

The first and perhaps most important advantage of optical interconnections over their electronic counterparts is 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 and with the closeness of the paths of electrons. In contrast, optical interconnections suffer no such effects, as long as care is exercised to assure that light scattering does not introduce an equivalent result (of different origin). Indeed, two streams of photons can pass directly through one another with no cross-coupling of their high-frequency modulations, provided the propagation medium is linear, as will nearly always be the case unless deliberate steps are taken to introduce nonlinear effects. The fundamental differences between electrons and photons in this regard can probably be traced to the fact that the latter are bosons, while the former are fermions. Two fermions can not occupy the same cell of phase space, whereas two bosons can.

A second advantage of optical interconnections is their freedom from capacitive loading effects. The speed of propagation of an electrical signal on a transmission line depends on the capacitance per unit length. Thus as more and more components having a capacitive component of admittance are attached to an interconnection line, the velocity of propagation decreases, and the time required to charge the line to a predetermined voltage 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.

A third potential advantage of optical interconnect technology lies in freedom from planar or quasi-planar constraints. Conventional integrated circuit design methodologies are constrained to the use of only a very few interconnecting layers, and within each layer interconnections can not cross. Optical waveguides can cross through other optical waveguides without significant cross-coupling (provided the angle of intersection is greater than about 10 degrees), and free-space light beams can cross through other 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 realization of reprogrammable interconnect patterns by means of dynamic optical interconnect components. In principle, the interconnection patterns associated with a chip can be changed at will by writing appropriate information to the interconnect element. While such a capability could be realized electronically with a suitable switching network, the cost in terms of wiring area is likely to be rather large.

Having reviewed the basic reasons for interest in optical interconnect technology, we now turn attention to various specific forms that optical interconnects can take. In this regard, it is useful to distinguish between two types of optical interconnections. The term index-guided interconnection is used to refer to an optical interconnection established when a light wave is guided along an interconnect path by means of a structure having a different refractive index than the surround. The usual examples are optical fibers and integrated-optic waveguides. The term free-space interconnection is used to refer to an optical interconnection established when a light wave travels from source to detector via a path that consists only of free space and possibly bulk optical components, such as lenses, mirrors, and holographic optical elements. Within this class it is possible to distinguish two subclasses, namely unfocused and focused interconnections. For unfocused interconnects, the optical signals are simply broadcast over a comparatively large region, whereas for focused interconnects, the signals are sent to specific desired locations via bulk focusing elements.

Index-guided interconnections are illustrated in Fig. 4, which shows sources connected to detectors via both fibers and integrated optic waveguides. Consider first the case of interconnections via optical fibers. The ends of the fiber must be carefully positioned over the source and detector, perhaps with the help of a wire-bonding machine, and a permanent attachment must be made, probably with UV-hardening epoxy. It should be noted that fibers

can not be allowed to bend too much, for bending induces radiation losses, which can be excessive if the radius of curvature of the bend is too small. Note that this interconnect technology requires a three-dimensional volume of space, for fibers must be crossed above the chip, and (for surface emitting sources as shown) bending limitations prevent keeping the fibers too close to the plane of the chip.

For integrated-optic waveguide interconnections, the waveguides might be formed by sputtering glass onto a silicon dioxide substrate. These guides again can not be bent too much, due to the problem of radiation losses. Light can be coupled out of a guide by using a taper at its end, causing radiation into and through the substrate. Alternatively, various kinds of integrated-optic junctions and couplers can be used to route signals from one source to a multitude of detectors. The chief difficulty with the integrated-optic approach lies in the problems of efficient coupling into and out of the waveguides.

Figure 5 shows two approaches to free-space unfocused interconnections. As mentioned, this approach involves broadcast of signals from one or more sources to one or more detectors, with no focusing of the light. Figure 5a shows a situation appropriate for the transmission of a single data stream to many sites on the same chip, as required, for example, for clock distribution. Detectors integrated into the silicon chip receive the common signal with the relative delays inherent in the path length differences between the source and the detectors. If a simple positive lens is inserted between the source and the chip such that the source lies at a focal point of the lens, then all paths from source to detectors are identical in length, and no relative delays are present. Such a configuration might be useful for clock distribution. Figure 5b shows a reflective configuration that achieves the same goal as the previous configuration [19]. A diffuser situated above the chip scatters any optical signal transmitted to it back towards the chip, where it can be intercepted by any number of detectors. The chief disadvantages of all unfocused broadcast schemes are (1) they are very inefficient in that only a small fraction of the total optical power falls on the detector sites where it is needed, (2) they provide basically a serial communication channel, with any parallelism achievable only through wavelength multiplexing, a complicating factor in any interconnect realization, and (3) they require a three dimensional volume rather than being strictly planar. It should also be mentioned that the broadcast of optical signals to the entire silicon chip has the potential to generate signals at unwanted locations, and therefore it is likely that an opaque dielectric blocking layer would be needed to prevent the optical signals from reaching undesired locations.

Figure 6 illustrates three approaches to free-space focused interconnects. In all of these cases, the light is sent from a source or several sources only to those locations where detectors are located. Figure 6a is appropriate, for example, for clock distribution, in which a single signal must be sent to many sites on the same chip. A transmissive holographic optical element is used for the routing of the signal. Rather high overall efficiencies can be achieved using dichromated gelatin holographic elements, with more than 99% of the light being sent to the desired locations. Note that any element that focuses the light to several different spots will also introduce relative time delays in those beams, but for typical geometries anticipated (e.g. chip, source and hologram confined to a 1 cm cube), the relative delays will be only a few picoseconds. Figure 6b shows a more general type of interconnection, in which several sources on the left are connected to several detectors on the right. The holographic optical element used in transmission connects each source to any number of detectors in whatever pattern might be desired. Note that a thin holographic element is restricted to providing the same interconnect pattern for all sources (i.e. it is a space-invariant element). However, a thick holographic element can provide different interconnect patterns for different sources, using the Bragg effect as the mechanism of selective interconnection. Finally, Fig. 6c shows a geometry in which a reflection hologram is used as the routing element. The free-spaced focused interconnections described above have two potential disadvantages. First, and most serious, they require very precise alignment of the focusing element with respect to the chip or chips. Second, as with the unfocused free-space interconnects, since the focusing element lies above the chip, they require a three-dimensional volume in order to establish the interconnect pattern.

We will close this section with a consideration of the inefficiencies inherent in optical interconnect schemes by virtue of the fact that they rely on the conversion of electrical signals to optical signals and back again to electrical signals. What price is paid for such conversion? An answer to this question can be obtained by considering a "canonical" example. It is reasonable to suppose that the sources used are LED's or lasing diodes operating in the near infrared with a wavelength near 8000 Å where the responsivity of p-i-n silicon detectors is near its peak value of about 0.6 amps/watt. It is not unreasonable to suppose an efficient lasing diode using 1.5mw of drive power (1 ma at 1.5 volts) and radiating 0.75mw of output optical power. If optical signals of 0.5mw power are available at the detector end, 0.3ma will be generated by a p-i-n photodiode. A transimpedance amplifier with perhaps two transistors could then utilize this current to generate voltages in the 100mv range. Further amplification (a few transistors) would be required to reach voltages compatible with logic devices. As the number of detectors to be connected to a single source

GOODMAN

6

increases, the power available at each detector decreases, thus requiring more amplification at each detector site (or alternatively more transmitted power from the source).

V. The Problem of Clock Distribution

A problem that appears amenable to immediate attack using optical technology is 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 common timing signal, called the clock. At the MOS chip level, for various reasons a two-phase clock is used. However, if a single phase clock is distributed, the two phases can be generated locally with a relatively simple circuit [20]. One of the chief difficulties encountered in designing circuits and systems for high-speed operation is "clock skew", a term which refers to the fact that different parts of the circuit or system receive the same state of the clock at different times. In this section we consider several different approaches to using optics for distribution of the clock, with the aim of minimizing or eliminating clock skew. Attention is focused primarily on clock distribution within a single chip, although many of the same concepts can be applied at the wafer level or even the board level.

The interconnections responsible for clock distribution are characterized by the fact that they must convey signals to all parts of the chip and to many 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 many parts of the chip. It is assumed that optics is used as only one part of a clock distribution hierarchy, in the sense that optical signals might carry the clock to various major sites on the chip, from which the signals would be distributed, on a local basis, by a more conventional electronic interconnection system.

If optical fibers are chosen as the interconnect technology, then the following approach might be used. A bundle of fibers is fused together at one end, yielding a single core into which light from the modulated source (probably a diode laser) must be coupled. Light coupled in at the fused end is split as the cores separate, and is transmitted to the ends of each of the fibers in the bundle. Each fiber is presumably located over an integrated optical detector which will convert the optical signal into an electrical one. Alignment of the fibers and the detectors can be accomplished with the help of a wire-bonding machine, and UV-hardening epoxy can be used to fix the fiber in its proper place permanently. Note that the fibers can not be bent too much, due to the problem of radiation loss, but some gradual bends will be inevitable and acceptable. The chief difficulty with this approach is associated with the alignment problem.

If integrated optical waveguides are chosen as the interconnect technology, then the following comments apply. Waveguides might be formed by sputtering glass onto a silicon dioxide substrate. Probably several straight waveguides would be run across the chip, thus avoiding problems associated with bending losses. The waveguides would all be fed by the same optical signal, probably generated by a single laser diode, with distribution to the individual waveguides accomplished by fibers. Alternatively, separate laser diodes could feed each waveguide, with synchronization of the sources provided by a single driving signal distributed electrically. Light must be coupled out of each waveguide at several sites along its length, with a detector converting the optical signal to electronic form at each such site. The primary difficulties associated with such an approach are the problem of efficient coupling of the light into the waveguides, and the problem of coupling light out with small coupling structures. Present optical waveguide technology requires couplers with sizes rather large compared with the feature sizes normally used in electronic integrated circuits.

Clock distribution can be accomplished with free-space unfocused interconnects if a single optical signal is simply broadcast to the entire chip, with detectors at selected sites generating the electrical version of the clock. Most desirable in terms of clock skew would be a single optical source located at the focal point of a positive lens. The collimated light falls normally on the surface of the chip, with the result that all detectors receive the optical clock signal with no skew whatsoever. However, this approach is extremely inefficient, in the sense that most of the light falls on portions of the chip where it is not needed or even wanted, and hence most of the optical energy is wasted. This waste has important implications when one considers that the amount of circuitry required on the chip to bring the detected clock signals up to a usable voltage level is dependent on the strength of the detected signals, which in turn depends on the amount of optical power delivered to each detector. In order to prevent the injection of unwanted electrical signals into the circuitry, an opaque dielectric overcoating should be used to block the light at all points except those where a detector resides.

Lastly we consider clock distribution by free-space focused interconnections. For such

interconnections, the optical source is imaged by an optical element onto a multitude of detection sites simultaneously. The required optical element can be realized by means of a hologram, which acts as a complex grating and lens to generate focused grating components at the desired locations. The optical efficiency of the scheme can far exceed that of the unfocused method, thereby reducing the electrical circuitry required to bring the detected signals up to logic-level voltages. Holographic optical elements with a single focal point can be made with optical efficiencies well in excess of 90%, using dichromated gelatin as a recording material. For elements with multiple focal points, it seems certain that overall efficiencies in excess of 50% can be achieved. The flexibility of the method is very great, for nearly any desired configuration of connections can be achieved. The chief difficulty of this method is the very high degree of alignment precision required in order 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. There exists a continuum of compromises between efficiency and alignment difficulty.

Figure 7 illustrates a possible configuration that retains high efficiency but minimizes alignment problems. The imaging operation is provided by a two element microlens, in the form of a block with a gap between the elements. A Fourier hologram can be inserted between the elements, and establishes the desired pattern of focused spots. The hologram itself consists of a series of simple sinusoidal gratings, and as such the positions of the spots generated on the chip are invariant under simple translations of the hologram. The source is permanently fixed to the top of the upper lens block after it has been aligned with respect to an alignment detector located at the edge of the chip, thereby establishing a fixed optical axis. The only alignment required for the hologram is rotation, which might be aided by the presence of a few more alignment detectors on the chip. The positions of the image spots are determined by the spatial frequencies present in the holographic grating. Such frequencies could be established very precisely if the hologram were written by an E-beam lithography machine.

Clock distribution at the wafer and board levels is also a strong candidate as an application of optical interconnections. The primary differences in the optical requirements for these cases lie in the different physical sizes of chips, wafers and boards. At the wafer level, the discussions of the previous paragraphs carry 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 optical fibers for conducting clock signals to remote parts of a single board, between which the greatest clock skew would otherwise be anticipated. Hierarchical schemes can also be envisioned, in which a network of fibers distributes an optical clock to a series of widely separated sites, from which either optical waveguides or free space interconnects are used to distribute the 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.

VI. Data Interconnects Using Optics

The use of optics for realizing data paths on a chip is inherently a more difficult problem than the clock distribution problem discussed above, due to the fact that there are many independent sources of information, each requiring an optical source. In addition, the interconnection patterns required by different sources may be quite different, thus requiring an extra flexibility of the interconnect technology. The chief practical difficulties arise from the fact that source technology is GaAs while the chip technology is predominantly Si. We consider first the problems of intrachip communications, and later make some comments on interchip communications.

One approach to the problem is hybrid in nature. GaAs chips containing sources are butted against a Si chip, with connections made by wire bonding. Signals to be sent to new locations on the Si chip are fed to the perimeter of that chip, where they are transferred to the GaAs chips. There they modulate optical sources, which send optical energy back to the interior of the Si chip, where it is detected and converted to electrical signals. Such a scheme is illustrated in Fig. 8, for the case of optical signal routing via a holographic optical element above the chip. A similar approach could use sputtered glass waveguides on an SiO₂ film to transport the optical signals, although with somewhat less flexibility in terms of the routing patterns that can be achieved. This latter scheme might be well suited to the realization of a perfect-shuffle network.

A second approach to the problem is a monolithic one that presupposes the successful development of heteroepitaxial growth of GaAs on Si, presumably with the help of a buffer layer (such as Ge) for lattice matching. The hope then is that devices can be fabricated in both the Si and the GaAs. Indeed such heteroepitaxial growth has already been carried out by several organizations. However, the critical question concerns the defect densities in the GaAs and the corresponding ability to realize high-quality devices in that material.

GOODMAN

8

Some success has been reported in this regard, with defect densities as low 10^4 per cm^2 [21]. However, as yet there remains a considerable distance to go before devices in both materials can be realized in monolithic form.

It has been tacitly assumed in the preceding discussions that the optical sources used must be on or close to the Si chip containing the computational circuitry. In fact the sources could be quite remote from the circuitry, provided modulators could be realized on the chip. A given source could be imaged onto the on-chip modulator, where the modulating signal would be imparted to the light beam. At present the only serious candidate for an on-chip modulator would be in integrated optic form [22]. Such devices can provide modulations in the Ghz range, but the difficulties associated with coupling the externally supplied light into an integrated optic waveguide on the chip will be non-trivial.

If holographic optical elements are to be used for data routing, the necessity to provide different geometrical interconnect patterns for different light beams dictates that the elements will have to be realized in the form of thick holograms. For such structures, the Bragg effect can be used to effectively provide a separate hologram for each source, and therefore to establish a different and unique interconnect pattern for each data path. Such holograms must be made interferometrically, since no fully synthetic way of generating the needed thick structures is known. One further point about holographic interconnect elements worth mentioning is that their use allows the possibility of "rewiring" the chip, simply by removing one holographic optical element and replacing it by another different element. Ideally one would like to have an electronically addressable light valve onto which a hologram can be written for the particular interconnect pattern desired at the moment. Such could be accomplished today if a thin hologram would suffice, i.e. if all interconnect patterns had the same geometrical structure. However, how to construct a light valve for realizing thick holographic optical elements is problematical, although some form of optical damage in electrooptic crystals might provide a suitable "real-time" medium for writing thick holograms.

For communication at the interchip level, one additional approach should be mentioned. Consider an array of silicon chips surrounding a GaAs chip, as shown in Fig. 9. Suppose that information is to be transferred from this cluster of chips to another similar cluster some distance away. It should be possible to realize high-speed multiplexers and demultiplexers in the GaAs chips, as well as optical sources. Very wide bandwidth optical signals can then be transferred between chip clusters by means of an optical fiber.

VII. Concluding Remarks

The discussions of the previous sections were by no means exhaustive, for the field of optical interconnections is so young that we have only scratched the surface in terms of ideas. However, some general conclusions are possible, and we list some in what follows.

1. The problem of optical clock distribution is one that is amenable to immediate attack by several different means. It provides a good vehicle for learning what the practical problems will be when optical detectors are to be used as a means for inputting information to chips. Furthermore, the problem of clock skew is a real and important one, and any contributions made by optics to solving this problem will be welcome.
2. Hybridization of Si and GaAs chips provides the short-term solution to making optical sources available to silicon circuitry. Investment in the development of such techniques will lead to ability to experiment with optical data interconnections, both at the interchip and intrachip levels.
3. Heteroepitaxial growth of structures that combine Si and GaAs layers is the most attractive long-term approach to making optical sources accessible by silicon chips. The development of such techniques to the point where Si and GaAs devices can be combined in monolithic form is probably a necessary condition before optical interconnections can have major impact at the intrachip level.
4. Integrated optoelectronics can play a major role in the optical interconnect field. Developments in this field are being pushed by other motivations, but optical interconnections can take direct advantage of any developments in this field.
5. The development of dynamic masks capable of changing optical interconnect patterns at high speeds would provide a unique capability, and could potentially speed up many important signal processing operations that depend on the fast Fourier transform algorithm, for example. To be maximally effective, such masks should be analogous to thick holograms, in that many different interconnect patterns can be multiplexed, one for each of many sources.

6. The perfect-shuffle exchange is a sufficiently ubiquitous operation, and requires sufficiently large wiring area in conventional integrated circuit form, that optical realizations are worth pursuing. Realizations could be by means of fibers, waveguides, or holographic optical elements.

It is hoped that the ideas presented above will serve to stimulate further thought and research in this interesting area of technology.

Acknowledgement

The author's thoughts have been strongly influenced by participation in an Army Research Office Palantir Meeting held October 31 through November 4, 1983. He would like to thank his co-participants in that meeting, Dr. Ravi Athale of the Naval Research Laboratory, Dr. S.Y. Kung (meeting chairman) of the University of Southern California, and Dr. Fred Leonberger of MIT Lincoln Laboratories for their many contributions to the ideas presented here, and Dr. Bobby Gunther for hosting the meeting. Partial support of the Air Force Office of Scientific Research is also gratefully acknowledged.

References

1. Special issue of the Proc. I.E.E.E. on acousto-optic signal processing, Vol. 69, January 1981.
2. N.J. Berg and J.N. Lee, Acousto-optic Signal Processing, Marcel Dekker, Inc., New York, N.Y. 1983.
3. A. Kozma, E.N. Leith, and N.G. Massey, "Tilted plane optical processor," Applied Optics, Vol. 11, pp. 1766-1777 (1972).
4. 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. I.E.E.E., Vol. 57, pp. 1026-1032 (1969).
5. For a review of this area, see H.J. Caulfield, J.A. Neff, 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. S.P.I.E., 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," Applied Optics, Vol. 18, pp. 149-162 (1979).
8. C.Y. Yen and S.A. Collins, Jr., "Operation of a numerical optical processor," Proc. S.P.I.E., Vol. 232, pp. 140-167 (1980).
9. A.A. Sawchuk and T.C. Strand, "Fourier optics in nonlinear signal processing," in Applications of Optical Fourier Transforms, (H. Stark, Editor), Chapter 9, Academic Press, New York, NY (1982).
10. C. Warde, A.M. Weiss, and A.D. Fisher, "Optical information processing characteristics of the microchannel spatial light modulator," Applied Optics, 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," Trans. I.E.E.E., Vol. ED-29, pp. 645-650 (1982).
12. R.W. Keyes, "Communication in computing," International J. Theoretical Physics, Vol. 21, pp. 243-273 (1982).
13. A.J. Blodgett, Jr., "Microelectronic packaging," Scientific American, Vol. 249, pp. 86-96, July 1983.
14. H. Tajima, Y. Okada, and K. Tamura, "A high speed optical common bus for a multiprocessor system," Trans. of the Inst. Electron. and Commun. Eng. Japan, Section E, Vol. E66, No. 1, pp. 47-48 (1983).
15. W.T. Cathy and B.J. Smith, "High concurrency data bus using arrays of optical emitters and detectors," Appl. Opt., Vol. 18, No. 10, pp. 1687-1691 (1979).

GOODMAN 10

15. C. Mead and L. Conway, Introduction to VLSI systems, Addison-Wesley, Reading MA, 1980, Chapter 9.
17. H.S. Stone, "Parallel processing with the perfect shuffle," I.E.E.E. Trans. on Computers, Vol. C-20, No. 2, pp. 153-161 (1971).
19. J.D. Ullman, Computational aspects of VLSI, Computer Science Press, Rockville, MD, 1984, Chapter 6.
19. The idea for using a diffuser above a chip for distributing optical signals was first suggested by Robert Kahn of DARPA. Extensions of this idea were pursued by Rockwell under DARPA funding. See P.D. Dapkus, "Optical communication for IC's," Report # MRDC 41072-IFR (Contract # N00014-80-C-0501), Rockwell International, Thousand Oaks, California, March 1982.
20. C. Mead and L. Conway, Introduction to VLSI systems, Addison-Wesley, Reading, MA 1980, p. 299.
21. 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).

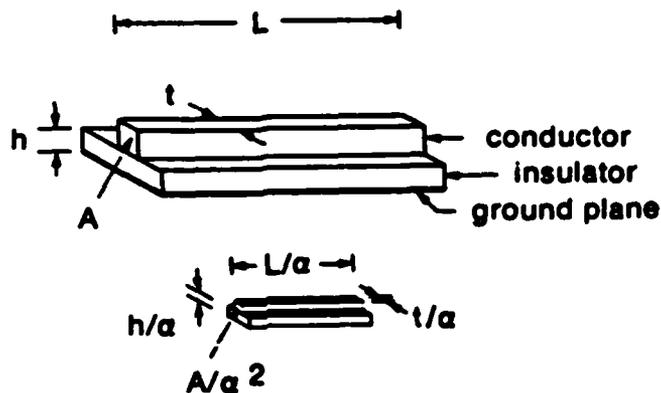


Figure 1. Scaling of interconnect lines.

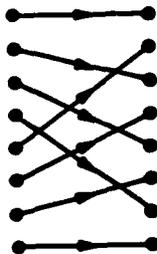


Figure 2. Shuffle-exchange network.

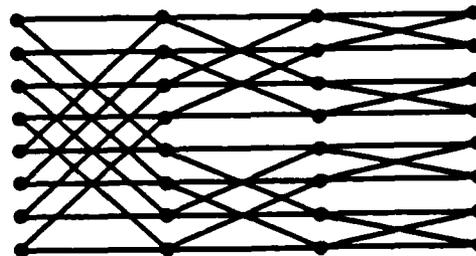


Figure 3. Butterfly network.

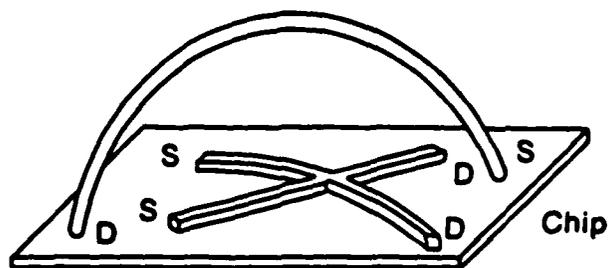
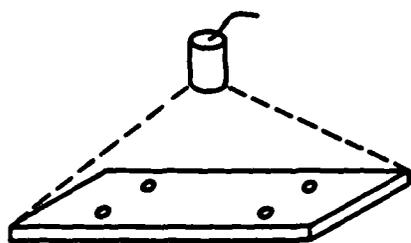
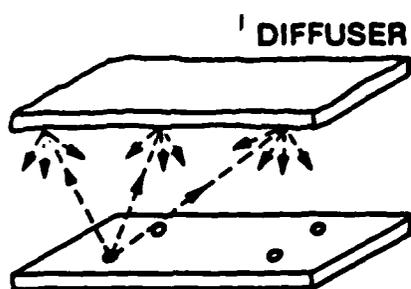


Figure 4. Index-guided interconnections.

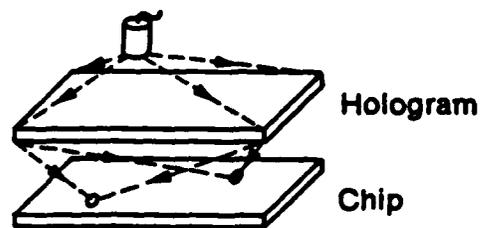


(a)

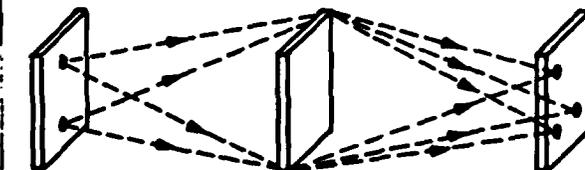


(b)

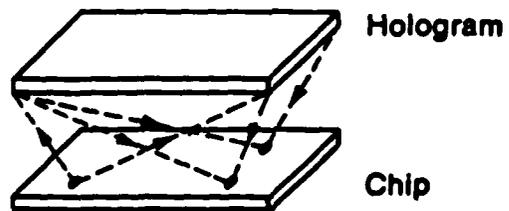
Figure 5. Free-space unfocused interconnections.



(a)



(b)



(c)

Figure 6. Free-space focused interconnections.

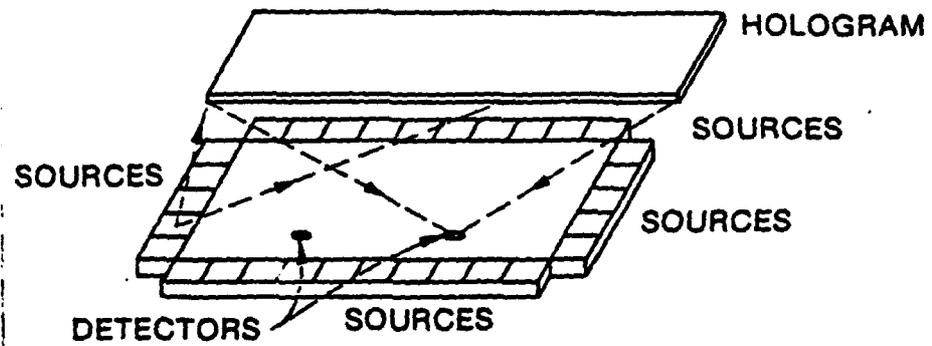


Figure 7. System for optical clock distribution with minimum alignment problems.

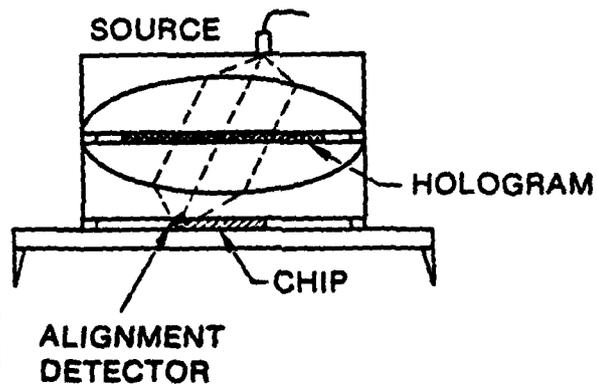


Figure 8. Hybrid approach to optical data interconnects.

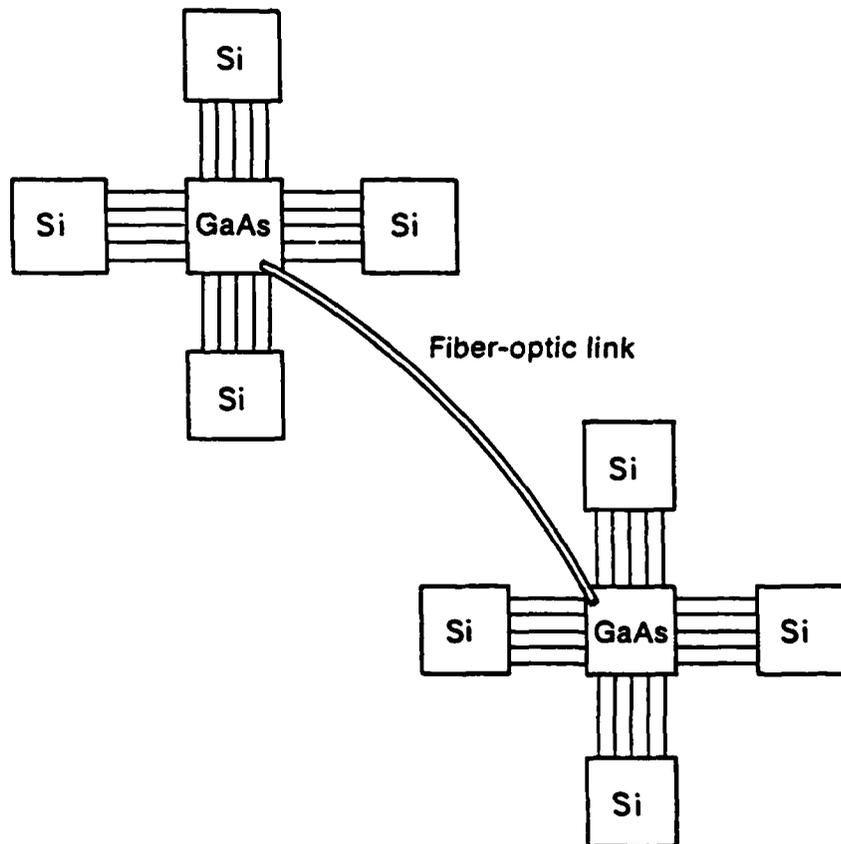


Figure 9. Silicon chip clusters with optical communication between GaAs chips.

APPENDIX II

**Coherent optical techniques for
diagonalization and inversion of
circulant matrices and circulant
approximations to Toeplitz
matrices**

Qizhi Cao and Joseph W. Goodman

a reprint from Applied Optics
volume 23, number 6, March 15, 1984

Coherent optical techniques for diagonalization and inversion of circulant matrices and circulant approximations to Toeplitz matrices

Qizhi Cao and Joseph W. Goodman

A coherent optical system for performing continuous Fourier transforms can be modified to perform discrete Fourier transforms. Such a system is capable of diagonalizing circulant matrices presented at its input. The diagonal elements of the new matrix are the eigenvalues of the original matrix. A suitable modification allows the eigenvalues of many different circulant matrices to be found simultaneously. Such a technique can be used for the initial portion of a coherent optical matrix inversion system, which can find the inverses of circulant matrices. The method can also be applied to the problem of inverting Toeplitz matrices in a hybrid digital and optical system.

1. Introduction

Matrix operations are receiving much attention from those working in the field of optical information processing. Primary attention has been focused on two problems, namely, matrix-vector multiplication and the solution of sets of simultaneous linear equations by means of iterative techniques. In this paper we discuss potential optical solutions to another class of problems, specifically the diagonalization and inversion of circulant matrices and circulant approximations to Toeplitz matrices. The basic ideas underlying the method were proposed some time ago but have just recently appeared in print.¹ Here we briefly review the basic ideas involved and focus attention on extensions of the method to matrices with complex elements, on some simple experiments verifying the most fundamental aspects of the ideas, and on the implications of the method for the problem of inverting Toeplitz matrices.

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 matrix *C* below, the numbers 1-4 stand for four distinct elements, and the organization of those elements in the circulant matrix is as follows:

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

A Toeplitz matrix is one for which all elements along the diagonal or any subdiagonal are constant. For example, again letting each number represent a distinct matrix element, a Toeplitz matrix has the form

$$\begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 1 & 2 & 3 \\ 6 & 5 & 1 & 2 \\ 7 & 6 & 5 & 1 \end{bmatrix}.$$

If only *M* of the $2N - 1$ possible diagonals are nonzero, the matrix is called a banded Toeplitz matrix with bandwidth *M*.

While any circulant matrix is also Toeplitz, the reverse is not true. Any covariance matrix of a wide-sense stationary random process is a Toeplitz matrix but in addition has Hermitian symmetry, implying that its eigenvalues are non-negative and real. The eigenvalues of a circulant matrix are in general complex-valued.

A very important problem in much of signal processing is the inversion of Toeplitz covariance matrices. Such inversions are required for the realization of optimum least-mean-square-error filters, and the inversion process must be repeated as knowledge of the covariance matrices is updated. While fast algorithms for such inversions have been devised, the issue of computational speed is still an important one, and any help offered by optical processing would be welcome. Inversion of Toeplitz matrices is, therefore, the ultimate motivation for this work.

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

Received 11 October 1983.

0003-6836/84/000803-08\$02.00/0.

© 1984 Optical Society of America.

In Sec. II we discuss the problem of diagonalizing circulant matrices using a coherent optical system and offer some simple experimental results verifying the basic idea. Section III discusses potential methods for inverting circulant matrices based on the diagonalization method of Sec. II. Such inversion techniques have not yet been experimentally verified, but there is strong evidence that they can be carried out in practice. Section IV provides a method to discover both moduli and phases of the output spots based on the intensity readings from a square-law detector. Section V discusses the application of the proposed technique to the inversion of a Toeplitz matrix or a multitude of Toeplitz matrices simultaneously. A combined optical and digital hybrid processor is required. Finally, Sec. VI summarizes the conclusions of the work at this stage.

II. Diagonalization of Circulant Matrices Using Coherent Optics

A remarkable property of circulant matrices (not shared by Toeplitz matrices) is that they are diagonalized by the discrete Fourier transform (DFT).² The resulting diagonal elements are the complex eigenvalues of the original matrix. Thus, if the DFT matrix W is defined by (again illustrating with a 4×4 example)

$$W = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & w & w^2 & w^3 \\ 1 & w^2 & w^4 & w^3 \\ 1 & w^3 & w^6 & w^9 \end{bmatrix} \quad (2)$$

where $w = \exp(-i2\pi/N)$, we have

$$\Lambda = W^{-1}CW = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \lambda_3 & \\ & & & \lambda_4 \end{bmatrix} \quad (3)$$

where the various λ are the eigenvalues of C .

A. Implementation with a 2-D Matrix Input

Two methods for diagonalizing a circulant matrix using coherent optics can be identified. One is esthetically pleasing in the sense that the entire circulant matrix appears at the input to the system and the entire diagonalized matrix appears at the output of the system. The second method is potentially more powerful. Only a single row of the circulant matrix is presented at the input, and all eigenvalues appear in a single row of the output plane. With the help of suitable astigmatic optics, the second method can be used to find the eigenvalues of many different circulant matrices in parallel, while the first method operates on one matrix at a time. We focus our attention first on the 2-D method and later discuss the 1-D method. Throughout this and the next sections we assume that the elements of the circulant matrices of interest are non-negative and real. Generalization to complex-valued matrices is deferred to a later section.

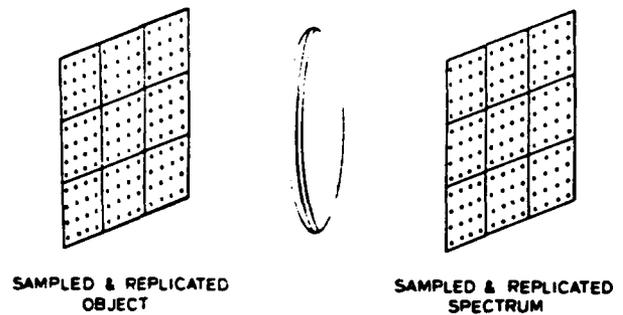


Fig. 1. System for performing the discrete Fourier transform.

The task of diagonalizing a circulant matrix using coherent optics can be performed if the normal continuous 2-D Fourier transform, so easily performed by such systems, can be changed to a discrete Fourier transform. Such a change can indeed be made. With reference to Fig. 1, the matrix to be diagonalized is entered into the coherent optical system as an array of transmitting cells in a mask, each cell representing one element of the circulant matrix (for the moment we assume that such elements are non-negative and real). The matrix is repeated at least 3×3 times in the horizontal and vertical directions causing the spectrum to form a series of discrete spots. The amplitude of each of the spots represents a different complex eigenvalue of the original matrix. Measurement of the intensities of these spots by discrete elements of a detector array is equivalent to measurement of the squared magnitudes of the eigenvalues of the matrix. If the full complex values are desired, interferometry or heterodyne detection must be used to extract both amplitude and phase information.

A mathematical description of the process can be given as follows. The coherent optical processor is assumed to perform a Fourier transform of the field $g(x, y)$ at its input. Thus,

$$G(u, v) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) \exp[-i2\pi(u_x x + v_y y)] dx dy \quad (4)$$

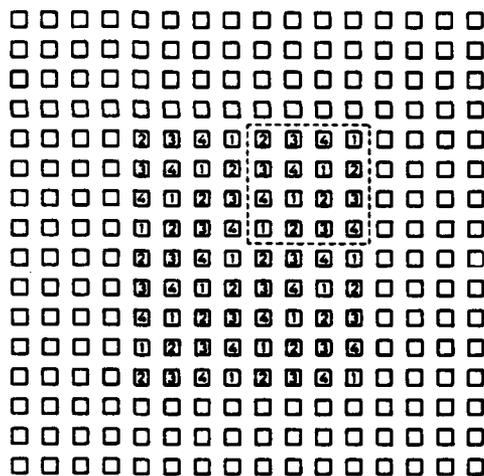
The input transparency representing the replicated matrix can be represented as an amplitude transmittance:

$$g(x, y) = \sum_{m=0}^{N-1} \sum_{n=0}^{N-1} \left[\sum_{p=0}^{N-1} \sum_{q=0}^{N-1} t_{pq} \cdot \delta(x - mL - pX - nL - qX) \right] \quad (5)$$

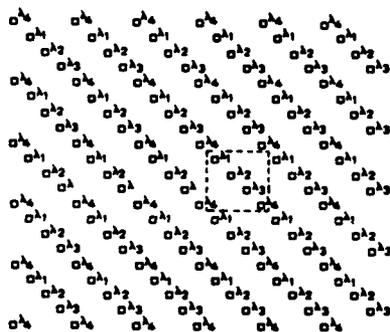
where X is the interval between any two cells along both x and y directions in the transparency, L is the period of replication, N is the number of elements in one row or column of the matrix, and t_{pq} are the amplitude transmittance of matrix elements. In Eq. (5), for simplicity we have neglected the finite size of the transparency and the finite size of the cells representing matrix elements. A more general formula will be given later.

Substitution of $g(x, y)$ into Eq. (4) yields a continuous spectrum

$$G(u_x, u_y) = \frac{1}{L^2} \sum_{k=0}^{N-1} \sum_{l=0}^{N-1} \delta \left(u_x - \frac{k}{L}, u_y - \frac{l}{L} \right) \left[\sum_{p=0}^{N-1} \sum_{q=0}^{N-1} t_{pq} \right. \\ \left. \times \exp[-j2\pi(u_x p X + u_y q X)] \right]. \quad (6)$$



(a)



(b)

Fig. 2. Matrix and its eigenvalues—2-D format: (a) replicated 4×4 circulant input matrix; (b) output for the matrix of (a). The λ are the eigenvalues of the original matrix.

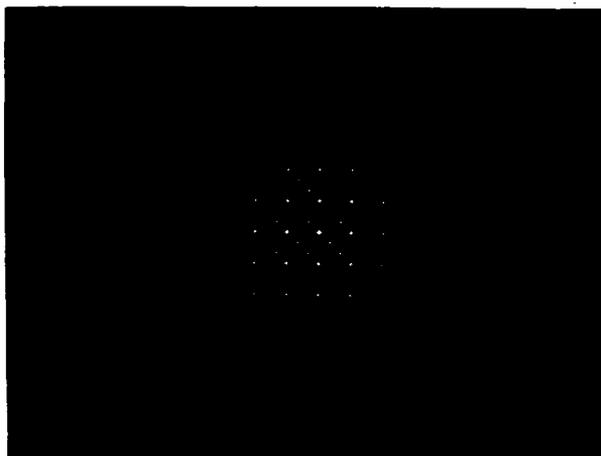


Fig. 3. Experimentally obtained output for a 3×3 circulant input matrix.

Finally, suppose that the continuous spectrum is sampled at particular points $u_x = k/L, u_y = l/L$. Then, with the definition $N = L/X$, the observed amplitudes can be expressed as

$$G_{kl} = \frac{1}{L^2} \sum_{p=0}^{N-1} \sum_{q=0}^{N-1} t_{pq} \exp \left[-j \frac{2\pi}{N} (kp + lq) \right].$$

The values of the complex spectrum at the chosen points are seen to represent the DFT coefficients of the original matrix (to within a known constant).

When the size a of a cell of the input matrix and the size w of the whole matrix transparency are taken into account, $g(x, y)$ and $G(u_x, u_y)$ become

$$g(x, y) = \left(\text{rect} \left(\frac{x}{a}, \frac{y}{a} \right) + \left\{ \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} \left[\sum_{p=0}^{N-1} \sum_{q=0}^{N-1} t_{pq} \right. \right. \right. \\ \left. \left. \left. \cdot \delta(x - mL - pX, y - nL - qX) \right] \right\} \right) \\ \times \text{rect} \left(\frac{x}{w}, \frac{y}{w} \right), \quad (7)$$

$$G(u_x, u_y) = \left(\frac{aw}{L} \right)^2 \text{sinc}(au_x, av_y) \left\{ \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} \delta \left(u_x - \frac{k}{L}, u_y - \frac{l}{L} \right) \right. \\ \left. \times \left[\sum_{p=0}^{N-1} \sum_{q=0}^{N-1} t_{pq} \exp \left[-j \frac{2\pi}{N} (kp + lq) \right] \right] \right\} \\ \cdot \text{sinc}(wu_x, wu_y). \quad (8)$$

Returning to the optics of the situation, if Fig. 2(a) represents the form of the matrix mask at the input to the system, Fig. 2(b) represents the form of the light amplitude distribution in the rear focal plane of the lens of Fig. 1. Each spot of light in the focal plane is represented by a small square, with a label indicating its complex value. The complex amplitudes of the spots are proportional to the eigenvalues of the original circulant matrix. In addition, as implied by Eqs. (6) and (8), there is a replication of these spots in the focal plane. The spots also are attenuated by a slowly falling envelope arising from the finite size of the cells of the input matrix.

There is one subtlety that should be mentioned. By using the 2-D forward DFT, we have in effect performed the matrix operation WCW, omitting the inverse sign on the first DFT matrix [compare Eq. (3)]. The result is only a 90° clockwise rotation of the diagonals in the eigenvalue plane, a change that is of no consequence provided its existence is recognized. As shown in Fig. 2(b), for example, the resulting eigenvalues are along the diagonal lines in the second and fourth quadrants instead of in the first and third quadrants.

A simple experiment has been performed to verify the basic ideas presented above. A circulant matrix of the form

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

was used. Figure 3 shows the distribution of light intensity present in the focal plane of the lens, each spot corresponding to an eigenvalue. The intensities of these spots were measured. The measured numbers

Table I. Measured Output Relative Values

	Ideal	Measured by power meter (μ W)	Values of sinc envelope	After compensation for sinc envelope	Normalized
$ \lambda_1 $	2	0.64	0.90	0.71	2.03
$ \lambda_2 $	1	0.33	0.94	0.35	1
$ \lambda_3 $	1	0.33	0.94	0.35	1

were compensated for the falloff induced by the finite size of the input matrix elements, and square roots were taken, yielding estimates of the magnitudes of the eigenvalues of the matrix. Table I compares the numbers deduced from measurement with the exact results obtained by analytical calculation. For this simple case the accuracy was found to be $\sim 1.5\%$. It is likely that the accuracy obtainable is a function of the size of the matrix, and greater care will be needed to obtain high accuracy from larger matrices.

B. Implementation with a 1-D Input

Since the successive rows of a circulant matrix are circularly shifted versions of the neighboring rows, such a matrix is obviously highly redundant. It may come as no surprise, therefore, to learn that the eigenvalues

of the entire circulant matrix can be obtained from operations on a single row. It can be shown that, if the top row of the matrix in Eq. (1) is subjected to a 1-D DFT, the complex DFT elements are, in fact, the eigenvalues of the original circulant matrix (see Appendix).

The above fact suggests that only one row of the circulant matrix need be entered into the optical system (with some replication of that row in the direction of its length) and only a 1-D Fourier transform need be performed. Thus, the spherical lens of Fig. 1 can be replaced by a cylindrical lens with power along the direction of the row. More important, by use of a suitable pair of spherical and cylindrical lenses, the optical system can be made to Fourier transform in one direction (along the rows) and image in the other direction (across the rows), allowing 1-D DFTs to be performed independently on all rows simultaneously. In this fashion the eigenvalues of many circulant matrices can be found in parallel. A 2-D detector array is then required at the output of the optical system if the squared magnitudes of all eigenvalues are to be measured. The optical system is illustrated in Fig. 4(a).

The idea of finding the eigenvalues of many circulant matrices in parallel as described above was tested (in a limited way) using the optical system of Fig. 4(a) and the same input mask that led to the results of Fig. 3. Originally the input mask was regarded as representing a full circulant matrix. For this particular mask, each row is a unit circular shift of the row above it, so all circulant matrices being represented by the rows of the mask are circularly shifted versions of the same matrix. The magnitudes of the eigenvalues of a circulant matrix are unaffected by circular shifts of the rows of the matrix. Therefore, all rows in the distribution of output intensity in the DFT plane should be identical for this particular mask. Figure 4(b), showing the experimental result, demonstrates that this is the case. With a more general set of circulant matrices represented by different rows of the input mask, the various rows in the output plane would differ from one another.

C. Diagonalizing Circulant Matrices with Complex-Valued Elements

Our previous discussions have all assumed implicitly that the elements of the circulant matrix or matrices to be diagonalized are non-negative and real; that is, it has been assumed that they can be represented physically by a non-negative and real amplitude transmittance of a mask. In practice, it is important to be able to deal with bipolar real inputs and in some cases complex-valued inputs. In this section we discuss several ways for generalizing the nature of the inputs.

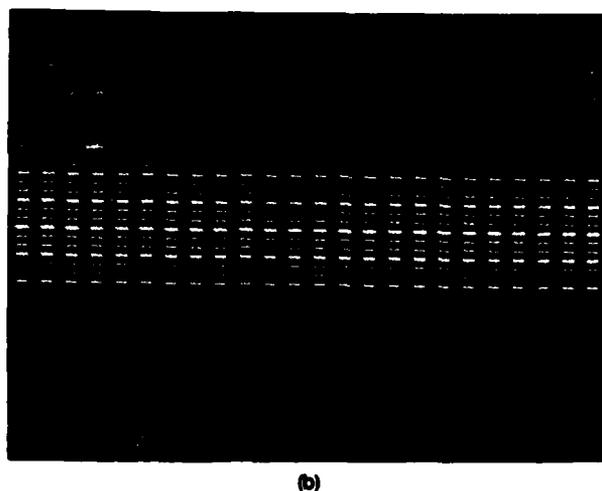
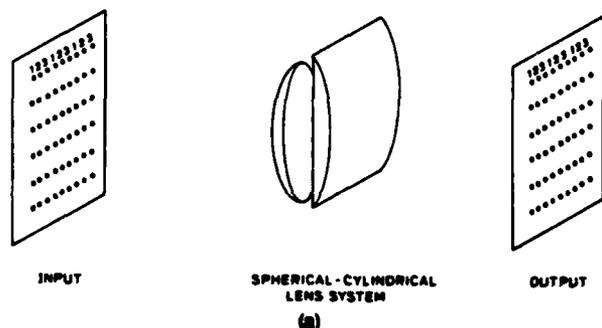


Fig. 4. Matrices and their eigenvalues—1-D format. (a) Optical system for finding eigenvalues of many circulant matrices. Each row of the output matrix contains the eigenvalues of one input matrix. (b) Experimentally obtained output for an input mask having many three-element sequences in parallel.

First, we suggest a way to input a 2-D complex matrix having positive real and imaginary parts for a 2-D optical Fourier transform system. From Fig. 2(b) we see that the output spots of a circulant matrix mask are all arranged along the diagonals. If we shift the entire input along the direction perpendicular to those diagonals, according to the shift theorem, we will have an output which is the same as the original one, except for a linear phase factor along the shift direction; that is, along each diagonal of the output, the complex field of each spot on this diagonal has an additional constant phase factor. Controlling the shift by various amounts, we can obtain various phase factors at will.

For example, as shown in Fig. 5(a), suppose we shift the matrix input along the diagonal toward the lower left with a displacement of $1/4$ of the diagonal interval between two adjacent cells of the input. The Fourier transform of this shifted matrix [the shaded one in Fig. 5(a)] has an additional phase factor $\phi = 2\pi(v_x + v_y) \cdot (X/4)$. In the output plane [see Fig. 5(b)] for the diagonals which satisfy

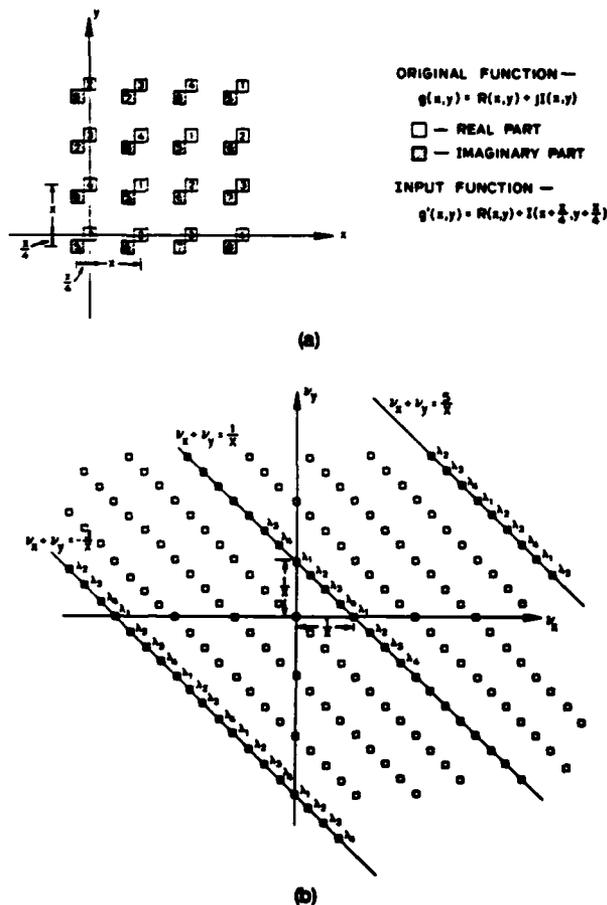


Fig. 5. Two-dimensional encoding of a matrix with complex elements. (a) An input mask. The original function is $g(x,y) = R(x,y) + jI(x,y)$. The input function is $g'(x,y) = R(x,y) + j(x + (X/4)y + (X/4))$. (b) Output for the input of (a). If $(v_x + v_y) = (4n + 1)/X$ with $n = 0, \pm 1, \pm 2, \dots$, $G^1(v_x, v_y) = R(v_x, v_y) + jI(v_x, v_y)$.

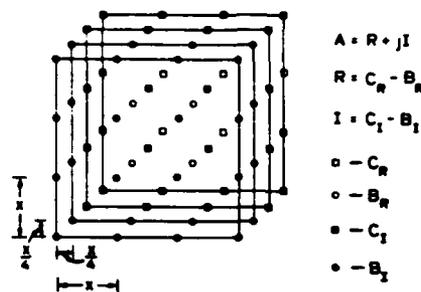


Fig. 6. Two-dimensional encoding of a circulant matrix with bipolar and complex elements. The four squares frame four matrices, each with relative shift of $X/4$ along the diagonal.

$$(v_x + v_y) = \dots - \frac{7}{X}, \frac{3}{X}, \frac{1}{X}, \frac{5}{X}, \dots, \frac{4n+1}{X}$$

with $n = 0, \pm 1, \pm 2, \dots$, the phase factors are

$$\phi = \dots - \frac{7\pi}{2}, \frac{-3\pi}{2}, \frac{\pi}{2}, \frac{5\pi}{2}, \dots, \frac{4n+1}{2}$$

with $n = 0, \pm 1, \pm 2, \dots$. Therefore, by making a mask with twin structures as shown in Fig. 5(a), with the shaded one representing the imaginary part and the unshaded the real part, we generate the effect of a complex input having positive real and imaginary parts.

In the same way, the effect of a complex matrix mask with negative real and imaginary parts can be generated.

Suppose the original matrix is $A = R + jI$, and both R and I have bipolar components. They can be decomposed as follows: $R = C_r - B_r$, and $I = C_i - B_i$, where both B_r and B_i are bias matrices with positive constant elements equal to the magnitude of the most negative elements from R and I , respectively. Then C_r and C_i have only real and positive elements.

Noticing that a negative sign is equivalent to a π phase factor, we can design an input mask for the matrix A as shown in Fig. 6. As before, the desired eigenvalues are located along the output diagonals as described earlier.

For the case of many 1-D inputs and a spherical-cylindrical lens system, we now suggest a way to generate a bipolar real-valued input. Assume that the three independent (real and bipolar) components of a 3×3 circulant matrix are $\{c_1, c_2, c_3\}$. By addition of this vector to a bias vector $\{b, b, b\}$, we have a new vector $\{c_1, c_2, c_3\}$ with all positive elements. Therefore, the 1-D Fourier transform for $\{c_1, c_2, c_3\}$ can be written as

$$\begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} = \mathcal{F} \left\{ \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} - \begin{bmatrix} b \\ b \\ b \end{bmatrix} \right\}.$$

But

$$\mathcal{F} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix},$$

and correspondingly

$$J \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} \lambda_1 + 3b \\ \lambda_2 \\ \lambda_3 \end{bmatrix}$$

$$J \begin{bmatrix} b \\ b \\ b \end{bmatrix} = \begin{bmatrix} 3b \\ 0 \\ 0 \end{bmatrix}$$

Now, as in the 2-D case, the original sequence can be represented by two sequences, one of which is shifted from the other appropriately. Knowing that the Fourier transform of the bias sequence has only one nonzero element, we can, for example, shift it from the other with a distance of $X/2$ [Fig. 7(a)] so that a π phase factor appears at the first spot of the second period of its Fourier transform, and so the sum of the two Fourier transforms within this period turns out to be the Fourier transform of the original bipolar sequence. This approach is illustrated in the diagrams of Fig. 7.

Furthermore, in Sec. III it will be seen that a second DFT is required to perform inversion of a circulant matrix. Therefore, we should have at least three consecutive periods of correct eigenvalues at this step to assume that, in the final output plane, the elements of the inverse matrix appear as separated spots. Another input geometry is designed for this purpose and is illustrated in Fig. 8.

III. Inversion of Circulant Matrices

While the problem of finding eigenvalues of a circulant matrix is of interest in its own right, there is much greater interest in inverting such matrices. Such a matrix inversion can be carried out if a means of inverting the complex eigenvalues of the diagonalized form of the matrix can be found. A discrete Fourier transform followed by inversion of the complex eigenvalues followed by an inverse DFT yields a new circulant matrix that is the inverse of the original matrix. Such an inversion method can be applied to a single circulant matrix or to many circulant matrices simultaneously using the astigmatic processor described earlier.

The key question to be addressed here is how to invert the complex eigenvalues of the matrix or matrices, those eigenvalues being represented by complex-valued fields in the DFT plane of the processor. An answer to this question is suggested by the following discussion.

Suppose we have a holographic recording device in the DFT plane that produces a diffracted field (into the -1 or conjugate diffraction order) having an amplitude that is proportional to the contrast of the fringes generated between an incident uniform plane reference beam and the object beam consisting of discrete spots of light with amplitudes proportional to the eigenvalues of the input circulant matrix. If U_r and U_o represent the amplitudes of the incident reference and object beams, respectively, we suppose that the amplitude U_i of the reconstructed image wave is given by

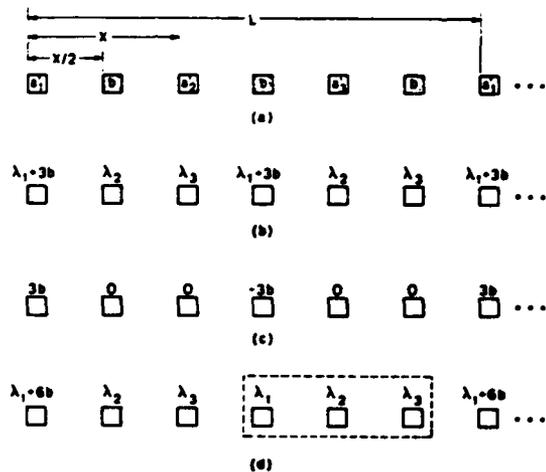


Fig. 7. One-dimensional encoding of a sequence with bipolar numbers, resulting in one correct period in the output: (a) $\{a_1, a_2, a_3\}$ is a positive sequence; $\{b, b, b\}$ is a bias sequence; (b) 1-D DFT of $\{a_1, a_2, a_3\}$; (c) 1-D DFT of $\{b, b, b\}$; (d) 1-D DFT of $\{a_1, a_2, a_3\}$.

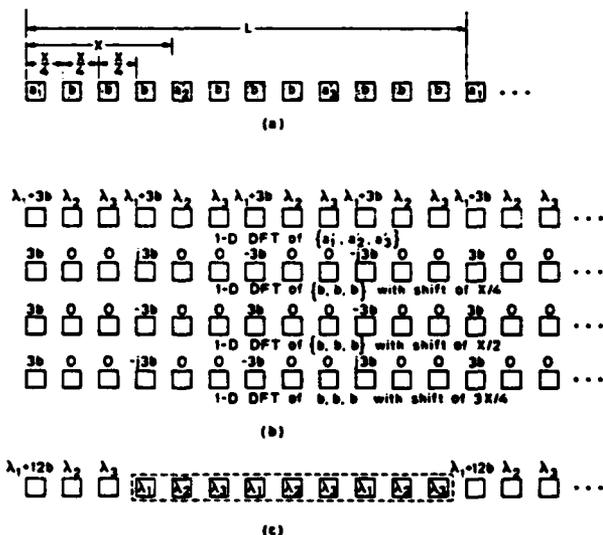


Fig. 8. One-dimensional encoding of a sequence with bipolar numbers resulting in three correct periods of the output: (a) positive sequence $\{a_1, a_2, a_3\}$ and three bias sequences, each of which has linear shifts of $X/4$ or $3X/4$ from the positive sequence, respectively; (b) 1-D DFT of four sequences; (c) 1-D DFT of $\{a_1, a_2, a_3\}$.

$$U_i = \frac{\alpha U_r U_o}{|U_r|^2 + |U_o|^2} \quad (9)$$

where α is a constant. If we now further assume that the reference beam is chosen to be much weaker than the object beam, we find that the reconstructed image wave is given approximately by

$$U_i = \alpha \frac{U_r}{U_o} \quad (10)$$

Since the reference wave amplitude is constant, the

reconstructed image wave has an amplitude proportional to the reciprocal of the complex object wave U_o . This is precisely the inversion process we desired.

That photographic film can be made to behave in the manner postulated in Eq. (9) is well established in optics literature. Such a dependence was exploited by Ragnerson³ and by Tichenor⁴ to realize coherent optical inverse filters using bleached photographic emulsions. More important, a similar dependence of the field diffracted in four-wave mixing experiments has been noted and exploited by White and Yariv for image edge enhancement.⁵ Most recently, Ja has demonstrated wave-front division using four-wave mixing.⁶ Thus, there seems to be ample reason to believe that methods for optical inversion of complex eigenvalues can be realized in the near term, particularly using real-time four-wave mixing devices.

The optical setup required for finding inverses of many circulant matrices simultaneously is illustrated in Fig. 9 under the assumption that a suitable nonlinear device is used for the reciprocation operation. While the reciprocation and inversion processes have not been experimentally demonstrated in this paper, the evidence that they can be carried out at least for some dynamic range of eigenvalues is extremely strong based on the above references.

IV. Detection of Output Data

In general it is possible for the elements of an inverse matrix to be bipolar (for real original matrices) or complex (for complex original matrices). However, the coherent optical system produces at its output a distribution of intensity representing the squared moduli of the elements of the inverse matrix. Limiting consideration to real-valued input matrices, some means for determining the sign of a real output must be found.

One possible solution is to resort to interferometric detection at the output, from which information concerning the signs associated with the various matrix elements can be found. Such an approach has the disadvantage of significantly complicating the optical system. We consider here an alternative approach which seems considerably simpler to implement. Our approach rests on a certain shift property of a circulant matrix as we now discuss in detail.

A. Shift Property of a Circulant Matrix

We first define a shift of a matrix to be the addition of a complex constant to each element of the matrix. The shift property of a circulant matrix can be stated as follows: if a circulant matrix C is shifted to C_s , the inverse matrix $[C_s]^{-1}$ is also shifted from C^{-1} with another shift constant. In other words, if $C_s = C + S$, then $[C_s]^{-1} = C^{-1} + E$, where S and E are matrices containing constant elements.

Furthermore, it can be shown that, if all the elements of S are s , the elements of E are all equal to $Ns/[\lambda_1(\lambda_1 + Ns)]$, where N and λ_1 are the dimension and the first eigenvalue of matrix C , respectively.

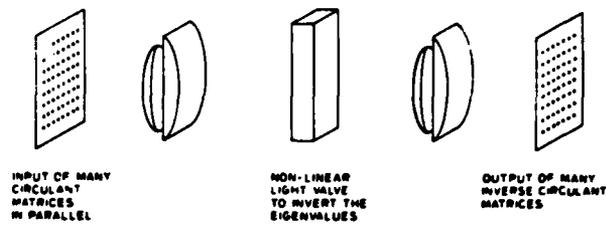


Fig. 9. System for performing the inversion of many circulant matrices.

In the following we will demonstrate this property for a 3×3 circulant matrix. The proof for an $N \times N$ circulant matrix is a simple extension of that for the 3×3 case. With c_1, c_2 , and c_3 as three independent components, the circulant matrix could be represented by vector C ,

$$C = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix},$$

consisting of the elements in the first row of C . Similarly, its inverse matrix could be represented by a vector C^{-1} , also consisting of the elements of the first row of C^{-1} .

As we know,

$$C = \mathcal{F} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix}, \quad (11)$$

$$C^{-1} = \mathcal{F}^{-1} \begin{bmatrix} \frac{1}{\lambda_1} \\ \frac{1}{\lambda_2} \\ \frac{1}{\lambda_3} \end{bmatrix}. \quad (12)$$

Now if C is shifted by vector S ,

$$S = \begin{bmatrix} s \\ s \\ s \end{bmatrix}, \quad (13)$$

with $C_s = C + S$, then, from Eqs. (11) and (13) and by recalling the DFT of S , a vector with constant elements, has only one nonzero element (the dc component) $3s$, we obtain

$$C_s = \begin{bmatrix} \lambda_1 + 3s \\ \lambda_2 \\ \lambda_3 \end{bmatrix}. \quad (14)$$

Finally, from Eqs. (12) and (14),

$$C_s^{-1} = \mathcal{F}^{-1} \begin{bmatrix} \frac{1}{\lambda_1 + 3s} \\ \frac{1}{\lambda_2} \\ \frac{1}{\lambda_3} \end{bmatrix} = \mathcal{F}^{-1} \begin{bmatrix} \frac{1}{\lambda_1} \\ \frac{1}{\lambda_2} \\ \frac{1}{\lambda_3} \end{bmatrix} - \mathcal{F}^{-1} \begin{bmatrix} \frac{3s}{\lambda_1(\lambda_1 + 3s)} \\ 0 \\ 0 \end{bmatrix}$$

that is, $C_s^{-1} = C^{-1} - E$, where the second shifting vector is

$$E = \begin{bmatrix} 3s \\ \lambda_1(\lambda_1 + 3s) \\ 3s \\ \lambda_1(\lambda_1 + 3s) \\ 3s \\ \lambda_1(\lambda_1 + 3s) \end{bmatrix}$$

B. Implementation of Sign Retrieval

With the help of the shift property of a circulant matrix we now can retrieve the sign information associated with the elements of the inverse matrix. When an input is designed, we interleave N additional rows between the N original rows of the input. Each of the additional rows represents a shifted version of one of the original matrices. Therefore, an input representing N matrices is now expanded to $2N$ rows; i.e., N pairs of rows. Similarly, at the output plane, we will also have N row-pairs. Each pair represents two respective inverse matrices, one of which is shifted from the other. For example, if the elements of one row are $\{a, b, c, \dots\}$, those of the other are $\{a - e, b - e, c - e, \dots\}$, where $e = Ns/[\lambda_1(\lambda_1 + Ns)]$ with all letters having the same meanings as in Sec. IV.A.

e can be easily calculated because λ_1 is simply the sum of all the elements of the original matrix and s is the shift constant, which can be any appropriate real number (for the 1-D input case). When $\lambda_1 = 0$, e will be infinite. However, this implies that the original circulant matrix is singular, which is a case we are excluding here.

With a square-law detector the values of $\{a^2, b^2, c^2, \dots\}$ and $\{(a - e)^2, (b - e)^2, (c - e)^2, \dots\}$ can be measured. Thus, the remaining problem of finding the values of $\{a, b, c, \dots\}$ is quite simple and can be solved rapidly with a digital computer.

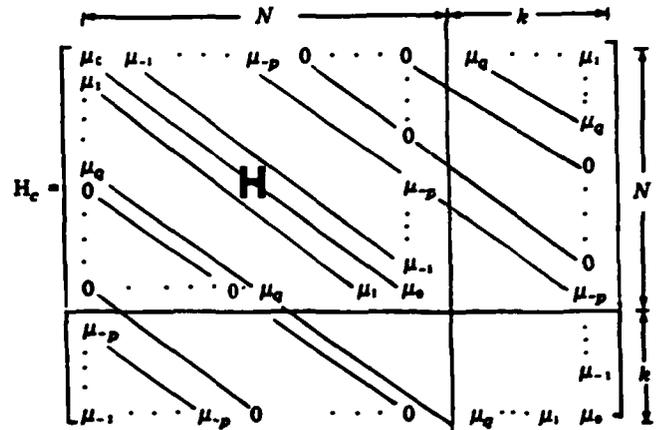
V. Inversion of Toeplitz Matrices using Circulant Approximations

Until this point we have considered the inversion of only circulant matrices. Using approximation methods developed by Jain,⁷ it is also possible to invert banded Toeplitz matrices using the same techniques supplemented by some digital processing. The resulting hybrid processor can be envisioned for use in inverting a multitude of Toeplitz matrices simultaneously. The digital processing required can be divided so that highly parallel digital hardware can be used with each separate digital processor devoted to inverting a small Toeplitz matrix, while the optical processor inverts a large circulant matrix. In what follows we describe the ideas that lie behind such an approach.

According to Jain,⁷ a Toeplitz matrix can be extended with terms so that it becomes a circulant matrix. This technique is called circulant decomposition. For example, the circulant matrix H_c is extended from a Toeplitz matrix H as shown below:

Table II. Comparison of Two Algorithms

Algorithm	Numbers of computations for H^{-1}
Trench	$2N^2 + 2Nk - k^2 + O(N)$
Circular decomposition	$2n_f + 3k^2 + O(N)$



where $k = \max(p, q)$

The inverse of the matrix H_c is

$$H_c^{-1} = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \begin{matrix} N \\ k \end{matrix}$$

where B_{11}, B_{12}, B_{21} , and B_{22} are four block matrices, each of which has the size shown.

It has been proved⁷ that the relation between the inverse of the original Toeplitz matrix T and the matrix H_c is

$$H^{-1} = B_{11} - B_{12}B_{22}^{-1}B_{21}. \quad (15)$$

Since B_{22} is still a Toeplitz matrix but with a smaller size of k , the advantage of decomposition now can be seen.

Apart from the algorithm for inverting a Toeplitz matrix without decomposition, this method includes four steps: first, extension of the banded Toeplitz matrix to a circulant one; second, inverting the circulant matrix by performing a forward FFT, inverting a diagonal matrix and performing an inverse FFT; third, inverting a Toeplitz matrix of small size, which is embedded in the inversion of the circulant matrix. This latter inversion is accomplished with a conventional fast algorithm. Finally, the matrix multiplication shown in Eq. (15) must be performed. Table II shows the numbers of computations required by two algorithms. One is the well-known Trench algorithm,⁸ the fastest conventional algorithm for inverting a Toeplitz matrix, and the other is the algorithm with decomposition as described above.⁷

In Table II, n_f is the number of operations involved in the FFTs for finding the inverse of the circulant matrix; i.e., $n_f = (N + k) \log(N + k)$. It can be seen that the computations required by the decomposition

methods are much less than those for the Trench algorithm when $k \ll N$ (a banded matrix case). In addition, when $k \approx \log_2 N$ so that $2n_f \gg 3k^2 + O(N)$ as in practice, the computations for inverting the circulant matrix become the greatest. For example, if $N = 512$ and $k = 10$, then $2n_f \approx 9000$ and $3k^2 + O(N) \approx 1000$. This result supports the idea of using an optical and digital hybrid system to accomplish the inversion of a banded Toeplitz matrix with the optical system taking care of the large circulant matrix and the electronic digital computer performing the remaining operations.

VI. Concluding Remarks

In this paper we have discussed methods by which coherent optical systems can be used to diagonalize and invert large circulant matrices. One configuration allows the simultaneous inversion of many such matrices in parallel. The methods can also be extended to the problem of inverting a multitude of banded Toeplitz matrices. The simultaneous inversion of many such matrices may be a useful capability in the processing of data from arrays of sensors. Often the data from such arrays are broken into many spectral bands with a separate matrix inversion operation required for the covariance matrices appropriate for each spectral band. Such operations are computationally very intensive, and the availability of a fast parallel, circulant matrix inversion system could speed the computations considerably.

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

Appendix

From $W^{-1}CW = \Lambda$, we have $[W^{-1}CW]^T = \Lambda$, where T denotes the transpose of a matrix. Therefore, $W^T C^T = \Lambda W^T$. But $W^T = W$, then $WC^T = \Lambda W$.

Writing out the first column of both sides of the above equality, we have $WC = \Lambda$, where

$$C = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix} \text{ and } \Lambda = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \end{bmatrix}.$$

References

1. J. W. Goodman, "Two Extensions of Fourier Optical Processors," *Transformations in Optical Signal Processing Proc. Soc. Photo-Opt. Instrum. Eng.* **373**, 89 (1983).
2. R. M. Gray, "Toeplitz and Circulant Matrices: II," in *Technical Report 6504-1*, Stanford U., Calif. (1977).
3. S. I. Ragnarson, *Phys. Scr.* **2**, 145 (1970).
4. D. Tichenor, "Extended Range Image Deblurring Filters," Ph.D. Thesis, Stanford U., Calif. (1974).
5. J. O. White and A. Yariv, *Opt. Eng.* **21**, 224 (1982).
6. Y. H. Ja, *Opt. Commun.* **44**, 24 (1982).
7. A. K. Jain, *IEEE Trans. Acoust. Speech Signal Process.* **ASSP-26**, 121 (1978).
8. S. Zohar, *J. Assoc. Comput. Mach.* **16**, 592 (1969).

FILMED