

AFOSR-TR-

USC OMDL 2001

2

**PHOTONIC MATERIALS AND DEVICES FOR OPTICAL
INFORMATION PROCESSING
AND COMPUTING APPLICATIONS**

DARPA 6015/AFOSR F49620-87-C-0056

**ANNUAL TECHNICAL REPORT
(RESEARCH PERIOD: 05/15/87 - 05/14/90)**

Submitted To:

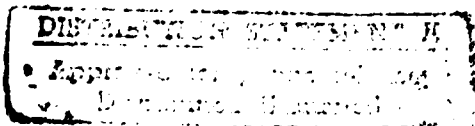
**Defense Sciences Office
Defense Advanced Research Projects Agency
1400 Wilson Boulevard
Arlington, Virginia 22209
Attn: Dr. Andrew Yang
Attn: Dr. Alan Craig (AFOSR/NE)**

Submitted By:

**Dr. Armand R. Tanguay, Jr.
Optical Materials and Devices Laboratory
University of Southern California
Los Angeles, California 90089**



OPTICAL MATERIALS AND DEVICES LABORATORY



91 3 06 132

**PHOTONIC MATERIALS AND DEVICES FOR OPTICAL
INFORMATION PROCESSING
AND COMPUTING APPLICATIONS**

**DARPA Order 6015
AFOSR F49620-87-C-0056**

**Combined Annual Technical Report
5/15/87 - 5/14/90**

**Armand R. Tanguay, Jr.
Optical Materials and Devices Laboratory,
Center for Photonic Technology, and
National Center for Integrated Photonic Technology
University of Southern California**

Approved by	
NTIS OF	
DAG	
U. S. Department of	
J. H. H. H.	
By	
D. H. H. H.	
Date	
Dist	
QUALITY INSPECTED 3	A-1

The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Defense Advanced Research Projects Agency or the U. S. government.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED			1b. RESTRICTIVE MARKINGS		
2a. SECURITY CLASSIFICATION AUTHORITY			3. DISTRIBUTION / AVAILABILITY OF REPORT APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED		
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE					
4. PERFORMING ORGANIZATION REPORT NUMBER(S) USC OMDL-2001			5. MONITORING ORGANIZATION REPORT NUMBER(S)		
6a. NAME OF PERFORMING ORGANIZATION UNIVERSITY OF SOUTHERN CALIFORNIA		6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION AIR FORCE OFFICE OF SCIENTIFIC RESEARCH		
6c. ADDRESS (City, State, and ZIP Code) UNIVERSITY OF SOUTHERN CALIFORNIA UNIVERSITY PARK, MC-0483 LOS ANGELES, CALIFORNIA 90089-0483			7b. ADDRESS (City, State, and ZIP Code) AFOSR/NE BUILDING 410 BOLLING AFB, DC 20332		
8a. NAME OF FUNDING / SPONSORING ORGANIZATION DEFENSE SCIENCES OFFICE		8b. OFFICE SYMBOL (If applicable) PKD	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER F49620-87-C-0056		
8c. ADDRESS (City, State, and ZIP Code) DEFENSE ADVANCED RESEARCH PROJECTS AGENCY 1400 WILSON BOULEVARD ARLINGTON, VIRGINIA 22209			10. SOURCE OF FUNDING NUMBERS		
			PROGRAM ELEMENT NO. 61101E	PROJECT NO. 6615	TASK NO. ARPA 6015
11. TITLE (Include Security Classification) PHOTONIC MATERIALS AND DEVICES FOR OPTICAL INFORMATION PROCESSING AND COMPUTING APPLICATIONS					
12. PERSONAL AUTHOR(S) DR. ARMAND R. TANGUAY, JR.					
13a. TYPE OF REPORT ANNUAL TECHNICAL		13b. TIME COVERED FROM 05/15/87 TO 05/14/90		14. DATE OF REPORT (Year, Month, Day) 1991, FEBRUARY	
15. PAGE COUNT 337					
16. SUPPLEMENTARY NOTATION COMBINED ANNUAL TECHNICAL REPORTS FOR 5/15/87 THROUGH 5/14/90					
17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number) OPTICAL INFORMATION PROCESSING, OPTICAL COMPUTING, SPATIAL LIGHT MODULATORS, VOLUME HOLOGRAPHIC OPTICAL ELEMENTS, PHOTO- REFRACTIVE MATERIALS, OPTICAL DISC, PHYSICS OF COMPUTATION		
FIELD	GROUP	SUB-GROUP			
19. ABSTRACT (Continue on reverse if necessary and identify by block number) See attached pages					
20. DISTRIBUTION / AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT ☐ DTIC USERS			21. ABSTRACT SECURITY CLASSIFICATION UNCLASSIFIED		
22a. NAME OF RESPONSIBLE INDIVIDUAL Chang			22b. TELEPHONE (Include Area Code) 202-767-4431		22c. OFFICE SYMBOL NE

PHOTONIC MATERIALS AND DEVICES FOR OPTICAL INFORMATION PROCESSING AND COMPUTING APPLICATIONS

Abstract

The research program described in this report addresses several generic avenues of opportunity in the advancement of a sophisticated component technology base for applications in optical signal processing, optical information processing, and optical computing. As such, the research program is multifaceted as well as highly interdisciplinary, spanning activities from materials growth, processing and characterization through device invention and evaluation to preliminary system level integration.

The primary program thrusts include:

- (1) Identification of the fundamental limitations inherent in optical information processing and computing systems that derive from physical laws, and discrimination of these limitations from those performance boundaries that result from device technological and materials parameter considerations;
- (2) An in-depth analysis of optical signal processing, optical information processing, and optical computing, with consideration of present levels of accomplishment and expected future development;
- (3) The invention and extensive characterization of a novel differential interferometric readout technique for high contrast ratio parallel readout of optical discs employed as two-dimensional spatial light modulators;

- (4) A critical assessment of the potential for fabrication of spatial light modulators and volume holographic optical elements in the III-V or II-VI compound semiconductor materials systems through the use of layered, multiple quantum well, and superlattice structures;
- (5) Analysis and implementation of real time volume holographic optical elements, including a novel multilayered structure (the so-called stratified volume holographic optical element, or SVHOE) with unique, potentially programmable diffraction characteristics;
- (6) The development of advanced techniques for the utilization of multi-dimensional dynamically programmable interconnections in hybrid optical/electronic multiprocessors based on VLSI and WSI technologies;
- (7) The invention, analysis, and development of a novel holographic recording and readout technique that allows highly multiplexed, weighted interconnections to be established in real time photorefractive materials, characterized by high throughput efficiency and very low interchannel crosstalk;
- (8) The practical development of photorefractive volume holographic optical elements, including the incorporation of novel two-layer antireflection coatings on very high index substrates to eliminate the reduction in diffraction efficiency attributed to the presence of multiple reflections, and the invention of methods for the elimination of a catastrophic collapse of the internal electric field in photorefractive materials under holographic recording conditions;
- (9) Optimization of the growth and processing of important electrooptic single crystal materials such as bismuth silicon oxide, lithium niobate, and strontium barium niobate;
- (10) Development of novel materials characterization techniques for dielectric single crystals and thin films, such as the electrooptic measurement of the volume resistivity of photorefractive materials, for integration with the growth and processing effort; and

- (11) Analysis and optimization of a number of candidate electrooptic spatial light modulator structures based on bulk single crystal materials, including the Photorefractive Incoherent-to-Coherent Optical Converter (PICOC), a Pockels Readout Optical Modulator (PROM) with no dielectric blocking layers and a $\langle 111 \rangle$ crystallographic orientation, single channel and linear array total internal reflection spatial light modulators, and the Optically Modulated Total Internal Reflection (OMTIR) spatial light modulator.

PHOTONIC MATERIALS AND DEVICES FOR OPTICAL INFORMATION PROCESSING AND COMPUTING APPLICATIONS

TABLE OF CONTENTS

<u>SECTION</u>	<u>PAGE</u>
ABSTRACT	3
1. PROGRAM SUMMARY	8
2. PROGRESS DURING THE CONTRACT PERIOD	13
3. PUBLICATIONS UNDER DARPA/AFOSR SPONSORSHIP	26
3.1 JOURNAL PUBLICATIONS	26
3.2 BOOK CHAPTERS	28
3.3 PATENTS	28
3.4 CONFERENCE PRESENTATIONS	28

APPENDIX 1:

D. A. Seery, M. H. Garrett, and A. R. Tanguay, Jr., "Electrooptic Measurement of the Volume Resistivity of Bismuth Silicon Oxide ($\text{Bi}_{12}\text{SiO}_{20}$)", Journal of Crystal Growth, **85**, 282-289, (1987).

APPENDIX 2:

A. Marrakchi, R. V. Johnson, and A. R. Tanguay, Jr., "Polarization Properties of Enhanced Self-Diffraction in Sillenite Crystals", IEEE Journal of Quantum Electronics, Special Issue on Electrooptic Materials and Devices, QE-23(12), 2142-2151, (1987).

APPENDIX 3:

R. V. Johnson and A. R. Tanguay, Jr., "Stratified Volume Holographic Optical Elements", Optics Letters, **13**(3), 189-191, (1988).

APPENDIX 4:

A. R. Tanguay, Jr., "Physical and Technological Limitations of Optical Information Processing and Computing", Materials Research Society Bulletin, Special Issue on Photonic Materials, XIII(8), 36-40, (1988); (Invited Paper).

APPENDIX 5:

R. V. Johnson and A. R. Tanguay, Jr., "Stratified Volume Holographic Optical Elements", Optics News, Special Issue on "Optics in 1988", **14**(12), 30-31, (1988); (Invited Paper).

APPENDIX 6:

J. Yu, D. Psaltis, A. Marrakchi, A. R. Tanguay, Jr., and R. V. Johnson, "Photorefractive Incoherent-to-Coherent Optical Conversion," in *Photorefractive Materials and Applications*, J. P. Huignard and P. Gunter, Eds., Springer-Verlag, New York, (1989).

APPENDIX 7:

R. V. Johnson and A. R. Tanguay, Jr., "Fundamental Physical Limitations of the Photorefractive Grating Recording Sensitivity", in *Optical Processing and Computing*, H. Arsenault, T. Szoplik, and B. Macukow, Eds., Academic Press, New York, (1989).

APPENDIX 8:

B. K. Jenkins and A. R. Tanguay, Jr., "Photonic Implementations of Neural Networks", Chapter 15 in *Neural Networks and Fuzzy Systems: A Dynamical Approach to Machine Intelligence*, B. Kosko, Ed., Prentice Hall, Englewood Cliffs, New Jersey, (1991).

APPENDIX 9:

A. R. Tanguay, Jr., "Fundamental Physical and Technological Considerations for Spatial Light Modulation", Proceedings of the International Conference on Optical Computing OC'90, Kobe, Japan, (1990); (Invited Paper).

APPENDIX 10:

C. Kyriakakis, Z. Karim, J. H. Rilum, J. J. Jung, A. R. Tanguay, Jr., and A. Madhukar, "Fundamental and Technological Limitations of Asymmetric Cavity MQW InGaAs/GaAs Spatial Light Modulators", OSA Topical Conference on Spatial Light Modulators and Applications, Incline Village, Nevada, Vol. 14 of the 1990 OSA Technical Digest Series, pp. 7-10, (1990).

APPENDIX 11:

J. H. Rilum and A. R. Tanguay, Jr., "Device Characteristics of Optical Disc Spatial Light Modulators", OSA Topical Conference on Spatial Light Modulators and Applications, Incline Village, Nevada, Vol. 14 of the 1990 OSA Technical Digest Series, pp. 200-203, (1990); (Invited Paper).

PHOTONIC MATERIALS AND DEVICES FOR OPTICAL INFORMATION PROCESSING AND COMPUTING APPLICATIONS

1. PROGRAM SUMMARY

The research program described in this report is focused on a critical evaluation of advanced photonic materials and device concepts for the implementation of optical information processing and computing systems. The effort ranges from a detailed investigation of the fundamental physical and technological limitations that impact the potential computational gain (*e.g.* increases in throughput, decreases in decision time subsequent to processing, or minimization of the energy expended during computation) of optical information processing and computing systems, through the invention and characterization of key enabling devices such as two-dimensional spatial light modulators and volume holographic optical elements, to the development of advanced techniques for materials growth, deposition, and processing that have a critical impact on potential device performance. As such, the research program is necessarily interdisciplinary, involving both faculty and students with physics, mathematics, electrical engineering, optics, materials science, and chemical engineering expertise. This multifaceted evaluation of novel materials, device, and system concepts has been directly responsible for the invention and characterization of a number of photonic devices and materials processing techniques that exhibit both high performance and capacity for practical manufacturing.

The research period covered by this report extends from 15 May, 1987 through 14 May, 1990. In addition to the set of publications attached as Appendices to this report, a significant number of manuscripts are in advanced stages of preparation, and will be

submitted for publication during the remaining program phase and included in the Final Technical Report.

The primary program thrusts can be organized into three principal categories: (1) fundamental and technological limitations of optical information processing and computing; (2) electrically and optically addressed spatial light modulators; and (3) volume holographic optical elements. The principal results of the research program in each category are outlined below. Further technical details and a guide to the various publications, as well as continuing directions of research, are provided in Section 2 (Progress During the Contract Period).

A detailed study of the fundamental as well as technological constraints that apply to an optically based information processing and computing technology has been undertaken. The purpose of the study is to delineate potential areas of opportunity that can be addressed by both optical and photonic techniques, in combination with electronic technology where appropriate. A major result of the study thus far has been an evaluation of the tradeoffs inherent in computation based on both analog and binary (digital) representations. Analog processing is favorable from an energy metric perspective whenever the computational complexity of the algorithm (or architecture) is large enough to overcome an inherently higher representation cost for a given dynamic range. A related result is that currently available analog optical devices operate much closer to the relevant quantum limits than *either* currently available binary optical or electronic devices. These results are in the process of being applied to analog *electronic* circuitry as well for direct comparison. In a parallel effort, the fundamental limits of the photorefractive grating recording sensitivity have been established. These limits clearly explain the apparent insensitivity of currently investigated photorefractive grating recording materials, and imply several promising opportunities for critically needed sensitivity enhancement.

Three different types of electrically and/or optically addressed spatial light modulators (SLMs) have been invented, developed, and characterized during the course of the research program to date. These devices include an optical disc spatial light modulator, a multiple quantum well (MQW) asymmetric Fabry-Perot spatial light modulator, and the Photorefractive Incoherent-to-Coherent Optical Converter (PICOC). The optical disc spatial light modulator leverages existing compact disc read-only memory (CD-ROM) technology by incorporating area encoding techniques and a differential interferometric readout configuration to yield a high resolution SLM with high contrast (100:1), *in situ* memory capacity, scrolling capability, and near term insertion prospects. The multiple quantum well asymmetric Fabry-Perot spatial light modulator offers potential hybrid integrability of III-V compound semiconductor modulation elements with silicon-based detection and control circuitry, and features a novel inverted cavity design made possible by the incorporation of InGaAs/GaAs strained layer multiple quantum wells on a transparent GaAs substrate. The Photorefractive Incoherent-to-Coherent Optical Converter is a two-dimensional spatial light modulator that exhibits remarkable fabrication simplicity by making use of the incoherent (image-bearing) erasure of a coherently recorded grating in a photorefractive material such as bismuth silicon oxide ($\text{Bi}_{12}\text{SiO}_{20}$).

Associated with the optical disc spatial light modulator characterization effort has been a parallel effort to develop an appropriate optical recording media test facility. To this end, we have participated in the modification of a commercially available optical media tester (Apex Systems OHMT-300) that is capable of recording two-dimensional image-formatted patterns on a wide range of optical disc media with an extremely high degree of track-to-track accuracy. The development and acquisition of this major piece of capital equipment represents a significant program achievement, and has allowed for preliminary tests that demonstrate the viability of optical disc based spatial light modulators.

In the area of volume holographic optical elements and interconnections, significant progress has been achieved on three separate issues: (1) the development of a viable approach to highly multiplexed weighted interconnections based on holographic techniques; (2) the characterization and implementation of Stratified Volume Holographic Optical Elements (SVHOEs); and (3) the development and characterization of photorefractive materials and devices for interconnection applications. During the research program, we have invented a novel approach for utilizing volume holographic optical elements in conjunction with arrays of individually coherent but mutually incoherent sources and two-dimensional spatial light modulators to provide highly multiplexed and weighted interconnections characterized by high throughput efficiency, low interchannel crosstalk, and capability for simultaneous (as opposed to sequential) initial recording and weight updates. We have also identified and demonstrated a number of novel features of SVHOE devices that have applications in programmable interconnections, tunable frequency filtering, and wavelength multiplexing/demultiplexing. In addition to the investigation of the fundamental limitations of photorefractive materials as used in interconnection applications described above, we have also attempted to improve the practical implementation of photorefractive devices. This effort has included the development of multilayer antireflection coatings that eliminate the source of multiple internal reflections, resulting in 100% increases in the saturation diffraction efficiency and two beam coupling gain; the identification and characterization of the dramatic electric field collapse within photorefractive crystals that occurs during grating recording, by means of a novel transverse electrooptic imaging method; the development of techniques to minimize or eliminate the electric field collapse, resulting in increases in diffraction efficiency and response time; the development of a novel electrooptic technique for the measurement of very high dark resistivities in electrooptic crystals (which determine the grating storage time); the analysis of the polarization properties of diffraction in optically active and

electrooptic photorefractive materials, including the effects of self-diffraction during grating recording; and the observation and characterization of the effects of microscopic charge screening on sub-hologram formation in photorefractive materials.

2. PROGRESS DURING THE CONTRACT PERIOD

As discussed in the Program Summary, the central theme of this research program is a critical assessment of the prospects for the implementation of optical information processing and computing systems, based on a parallel assessment of advanced photonic materials and device concepts. As such, the principal program elements have included studies of the fundamental physical and technological constraints that impact current and projected computational performance; invention, development, and characterization of critical optical processing and computing components, such as one- and two-dimensional spatial light modulators and volume holographic optical elements; and the development of advanced techniques for materials growth, deposition, and processing that have a critical impact on potential device performance. In this section, we describe the most significant results of the research program to date, provide a guide to the various publications that document these results, and indicate continuing directions of research where appropriate.

Fundamental Physical and Technological Constraints on Optical Information Processing and Computing

In order to examine the fundamental physical and technological limitations to optical information processing and computing, we have considered any computational process to comprise three separate functions: the representation of information, the implementation of computational complexity, and the detection of results. This separation provides a key for the analysis of any proposed algorithm, architecture, and implementation scheme from the perspective of a given metric, such as the total energy required to complete a particular calculation. In calculating the total energy dissipation inherent in a computational process, it is necessary to include the detection and storage of the inputs, the implementation of the computation, the costs of communication (interconnection) within the processor, and the

detection and communication of the results. The energy metric is of considerable importance, as many currently envisioned computational systems (both optical and electronic) are highly power consumptive, and are therefore limited in performance by the thermodynamics of cooling.

Our approach has been to examine the computational process for a number of important computational functions (*e.g.* two-dimensional Fourier transformation, image-based correlations, synthetic aperture radar image formation, and nonlinear dynamical systems such as neural networks) that have proven difficult to implement by electronic means, without resort to large scale systems that have size, weight, and power characteristics that make them inappropriate for a wide range of applications. We have examined both the fundamental boundaries implied by quantum statistics and thermodynamics, as well as the technological boundaries implied by the choice of particular components within a given implementation (such as spatial light modulators based on III-V compound semiconductor multiple quantum well modulators hybridized with silicon detection and control circuitry, for example) [Appendices 4, 7, 8, 9, and 10; Publs. 4, 16; Book Chs. 2, 3; Conf. Pres. 3, 5, 7, 13, 15, 19, 33, 43, 46, 51, 52, 56].

One of the most striking results to emerge from this study has been a realization of the dramatic difference in energy cost that exists between the digital (binary) and analog representations of a given number at the *same probability of error* (the equivalent of a bit error rate) due to quantum statistical fluctuations alone. The digital representation cost scales as the logarithm of the overall dynamic range, while by contrast the analog representation cost scales quadratically with increasing dynamic range. This result has direct implications for computation, in the sense that computational algorithms based on analog representations can be more energy efficient at the fundamental limits only if the computational complexity of the process implied by the analog-based algorithm is sufficient to make up for the increased representation cost. The implications of this result also extend

to the representation of information, for example, on optically addressed spatial light modulators, in which case a clear tradeoff exists among resolution (number of independent pixels), frame rate, dynamic range, and sensitivity for a given probability of representation error. In many cases, the stated performance characteristics quoted for certain spatial light modulators cannot be obtained simultaneously without provision for an unacceptably high input intensity.

Given the difference in representation cost, we have compared a number of implementations of such computational processes as the Fourier transform that are analog or digital, and optical or electronic in nature, at both the fundamental (quantum-limited) boundaries, and at the current limits of available device technologies at the device, circuit, and systems integration levels. The results are quite surprising. At the fundamental limit, analog (optical) and digital (electronic) approaches can exhibit comparable energy costs. At the technological level, however, electronic devices, circuits, and systems operate at progressively larger factors away from the fundamental limits, whereas currently available optical components are capable of operating much closer to the appropriate fundamental constraints. We are in the process of extending this line of inquiry to include *analog* electronic circuits and systems, as well as hybrid photonic components comprising both optical and electronic elements. [Appendices 4 and 8; Publ. 4; Book Ch. 3; Conf. Pres. 3, 5, 7, 13, 15, 19, 33, 43, 51, 52, 56]

A second major avenue of investigation has been to examine the fundamental and technological limitations of spatial light modulation, in particular focusing on the advantages and disadvantages inherent in phase and/or amplitude modulation, on methods for optimizing the utilization of available oscillator strength (particularly in multiple quantum well modulators), and on the optimization of asymmetric cavities to enhance device sensitivity, contrast ratio, and throughput [Appendices 4, 7, 8, 9, and 10; Publs. 4, 16; Book Chs. 2, 3; Conf. Pres. 3, 5, 7, 13, 15, 19, 33, 43, 46, 51, 52, 56]. In

particular, we have examined in detail a hybrid spatial light modulator with InGaAs/GaAs multiple quantum well modulators that utilize the quantum confined Stark effect in an asymmetric Fabry-Perot cavity configuration, in conjunction with silicon driver chips that contain appropriate detection and control circuitry [Appendices 8 and 10; Publ. 16; Book Ch. 3; Conf. Pres. 46, 51, 56].

Both pure phase and pure amplitude modulation cases have been examined from the perspective of minimizing signal-dependent amplitude modulation in the phase modulator case, and of minimizing signal-dependent phase modulation in the amplitude modulator case. The results of the study indicate that reflection amplitude modulators with contrast ratios in excess of 20:1 can be achieved with a dynamic range of about 50% (implying a 3 dB insertion loss), with acceptable signal-dependent phase modulation and optical bandwidth. Examination of the sensitivity of these results to process-induced variations (such as thickness nonuniformities characteristic of current state-of-the-art molecular beam epitaxy (MBE) techniques) revealed that such spatial light modulator designs must be detuned from optimized (theoretically achievable) performance parameters in order to obtain the requisite uniformity across a two-dimensional array. This study is continuing in conjunction with a device fabrication and characterization effort in collaboration with Prof. Anupam Madhukar's research group at USC.

During the contract period we have also undertaken and completed a study of the fundamental physical limitations of the photorefractive grating recording sensitivity, in order to evaluate the prospects for application of photorefractive materials to multiplexed interconnection applications [Appendices 4 and 7; Publ. 4; Book Chs. 2, 3; Conf. Pres. 3, 5, 7, 13, 15, 33]. In this study, we determined the quantum limited photorefractive sensitivity that can be achieved assuming a single photoexcited carrier for each absorbed photon, and compared the resulting optimum charge distribution with those that arise from sinusoidal grating exposure profiles in conjunction with realistic charge transport models.

These results clearly explain the fundamental origins of the observed relatively low quantum efficiencies of photorefractive recording processes in materials such as bismuth silicon oxide and barium titanate, and point out several potential directions for enhancement of the sensitivity. The importance of the grating recording sensitivity in optical processing and computing systems applications is that it establishes the maximum reconfiguration rate of a volume holographic optical element or interconnection device that can be achieved with a given (average) optical power.

Electrically and Optically Addressed Spatial Light Modulators

During the course of the research program, a number of different types of electrically and/or optically addressed spatial light modulators have been invented, developed, and characterized. These devices include an optical disc spatial light modulator, a multiple quantum well (MQW) asymmetric Fabry-Perot spatial light modulator, and the Photorefractive Incoherent-to-Coherent Optical Converter (PICOC). The optical disc spatial light modulator is electrically addressed (though optically written), while the latter two devices are optically addressed. These three devices span a considerable range of spatial light modulator functions and performance characteristics, as well as potential for inexpensive manufacturability and near-term insertion.

Optical disc spatial light modulators were first proposed by us [Conf. Publs. 15 and 24] in order to take advantage of the significant materials and device development leverage provided by commercially available CD-ROM technology. In this spatial light modulator approach, the format of an optical disc is altered to allow for the recording of a two-dimensional image in, for example, a 1000 x 1000 pixel array within a 1 cm² area. Grey scale is provided by area encoding techniques, such that within each "pixel" containing 100 binary bits, any number of bits can be written between 0 and 100, providing for a binary-

encoded analog representation. On readout, each pixel is under-resolved to allow the grey scale to be realized without imaging individual bits.

In order to accomplish such readout with a contrast ratio compatible with the desired grey scale, we developed a novel differential interferometric readout configuration to circumvent the inherently low contrast ratios (about 2:1) characteristic of commercially available CD-ROMs that are optimized for digital data recording applications. In this configuration, a shear plate is used to create two orthogonally polarized readout beams, displaced from each other along the track direction by one half of the inter-bit spacing. On readout, two independent images of the optical disc are created, which recombine within the shear plate (cancelling the displacement) and interfere with each other when passed through an appropriately oriented polarization analyzer. A π phase shift is introduced between the beams by the shear plate, such that the beams exactly cancel everywhere except where bits have been written. In this manner, the background reflections from the tracks and grooves are eliminated, and written bits appear as a pair of bright dots against an essentially black background. In experimental tests of the technique using a mirror in place of the optical disc, we have achieved contrast ratios in excess of 600:1 as limited by the AR coatings of the various optical elements. On double sided ablative media without AR coatings, contrast ratios of 5:1 have been measured, whereas on single sided bump forming media the experimentally achieved contrast ratios exceed 20:1 [Appendix 11; Publs. 11, 12; Conf. Pres. 15, 24, 33, 36, 40, 47, 52, 55].

In order to write two-dimensional images directly onto an optical disc in a format that allows for parallel readout, the traditional sequential bit recording process must be modified. In collaboration with Apex Systems in Boulder, Colorado, we have developed an optical media tester that is capable of very high (sub-micron) track-to-track accuracy, and of recording on a wide range of optical disc media including ablative, bump-forming, and magneto-optic. In addition, the design is optimized to minimize the time required to

record each two-dimensional image. Using this optical media tester, which was recently acquired during the contract period, we have successfully written a wide range of both binary and analog two-dimensional images on various optical disc media. The largest images written to date have 512 x 512 pixels with 101 binary-encoded grey scale levels. Parallel readout of these images using the differential interferometric readout configuration has been accomplished with striking results. Further investigations are under way to optimize the recording parameters utilized, as well as the readout configuration.

One of the most useful features of the optical disc spatial light modulator is its capability for operation in a "scrolling" readout mode, as is desirable for example in synthetic aperture radar image formation. In the scrolling mode, it is necessary to read out the entire image for subsequent processing following the arrival of *each* additional line of data, with the consequent deletion from the field of view of the oldest line of data. Operating in this readout mode, the optical disc is currently capable of parallel readout rates exceeding 5 *terabits* of information per second.

A second type of spatial light modulator that we have actively investigated during the program period is a hybrid SLM in which the active modulator elements are based on InGaAs/GaAs multiple quantum wells in an asymmetric Fabry-Perot cavity configuration, and the optical address function is carried out by detection and control circuitry integrated on a silicon chip by standard VLSI foundry processing. We have chosen the strained layer InGaAs/GaAs system for epitaxial growth on GaAs substrates in order to achieve substrate transparency at the operational wavelength of about 950 nm. This substrate transparency can be used to advantage by allowing for an inverted asymmetric Fabry-Perot cavity geometry in which the low reflectivity Bragg mirror is grown by MBE techniques (in Prof. Anupam Madhukar's laboratory) *below* the MQW structure, and the high reflectivity mirror is an externally deposited dielectric multilayer coating [Appendix 10; Publ. 16; Book Ch. 3; Conf. Publs. 46, 48, 49, 50, 51, 52, 54, 56]. This configuration relieves the MBE growth

process of including the (several micron thick) high reflectivity Bragg mirror below the MQW structure, at once minimizing the growth time and complexity as well as the accumulation of strain and lattice defects. In addition, it allows for face-to-face contact by means of flip-chip bonding techniques between the GaAs substrate containing the modulator elements, and the Si chip containing the detectors and control electronics. This in turn eliminates the need for vias through both substrates, as is characteristic of traditional hybridization.

To date we have performed a thorough analysis of the potential performance characteristics that can be expected from such a hybrid SLM based on InGaAs/GaAs MQWs (as well as on AlGaAs/GaAs MQWs) in order to determine both optimized mirror and quantum well designs, as well as the sensitivity of such designs to anticipated process-induced variations. On the basis of this analysis, we have grown a number of candidate modulator structures, for which extensive optical and electrical characterization is in progress at the present time. In every case examined thus far, the predictions based on our design analysis have been borne out. In parallel with this effort, we have also designed and fabricated (through the Metal-Oxide-Semiconductor Implementation Service (MOSIS) operated for DARPA by USC's Information Sciences Institute) a number of silicon chips that contain various types of photodetectors and control circuitry. These chips are fully functional, and are in process of being characterized to evaluate their maximum operational bandwidth, the accuracy of the control circuitry functional transformation, and the optical sensitivity of the photodetectors. Preliminary studies of the requisite flip-chip bonding requirements are under way.

The final type of spatial light modulator that has been developed and characterized during this research program is the Photorefractive Incoherent-to-Coherent Optical Converter (PICOC), which has the unusual feature of combining photorefractive volume holographic grating recording techniques with incoherent-to-coherent conversion capability

within the same device [Appendix 6; Book Ch. 1]. This feature at once allows for simple and inexpensive device fabrication, as any of a number of bulk photorefractive crystals can be used to configure this device without the need for additional device processing. In addition, it allows for a number of unique optical information processing and computing applications that involve both spatial light modulation and volume holographic recording techniques. During the contract period, an intensive study of the factors that affect PICOC operational mode, readout configuration, sensitivity, resolution, and contrast ratio was completed. The results of this study are described in detail in Appendix 6.

Volume Holographic Optical Elements

One of the most important potential advantages of optical techniques for application to information processing and computing systems is the capacity for parallel, densely packed crosstalk-free interconnections among multiple processing planes provided by volume holographic recording and reconstruction techniques. Should real time implementations of such optical interconnections not prove viable, this advantage may prove to be minimal, if it exists at all. As such, a critical feature of this research program has been the evaluation of the prospects for high performance volume holographic interconnections, as well as the development of practical photorefractive materials and devices.

During the contract period, significant progress has been achieved in the development of a novel approach to highly multiplexed weighted interconnections based on holographic techniques, on the characterization and implementation of Stratified Volume Holographic Optical Elements (SVHOEs), and on the development and characterization of photorefractive materials and devices for interconnection applications. Progress in each of these three areas is summarized below.

We have recently discovered that the traditional methods for recording volume holographic interconnections lead to either high interchannel crosstalk in the case of fully

coherent, simultaneous recording schemes, or to interchannel crosstalk, high throughput losses, and complex recording schedules in the case of sequential (coherent or incoherent) recording schemes. By utilizing the optical beam propagation method, we have simulated a number of such point-to-point weighted interconnections with multiple holographically recorded gratings, and have been able to quantify the degree of crosstalk and throughput loss in each case [Appendix 8; Publs. 9 and 10; Book Ch. 3; Patent 1; Conf. Pres. 37, 38, 39, 41, 42, 45, 49, 50, 52, 53, 54, 57]. In addition, we have identified a new form of crosstalk, so-called *beam degeneracy* crosstalk, that produces substantial interchannel crosstalk even in the incoherent recording case, and accounts for a dramatic throughput loss as well. For an N input point to N output point holographic interconnection recorded by traditional sequential methods, for example, the optical throughput efficiency is reduced by of order $1/N$. This loss is potentially catastrophic for large scale interconnection networks such as those envisioned for optical implementations of neural networks.

Subsequent to this analysis, we have invented an architecture based on two-dimensional parallel source arrays, spatial light modulators, and volume holographic optical elements that is capable of circumventing the crosstalk and throughput problems, as well as providing for simultaneous rather than sequential weight updates [Appendix 8; Publs. 9 and 10; Book Ch. 3; Patent 1; Conf. Pres. 37, 38, 39, 41, 42, 45, 49, 50, 52, 53, 54, 57]. In this approach, a two-dimensional source array (*e.g.* the surface emitting semiconductor diode laser arrays recently announced by Bellcore and ATT) is employed, in which each source is individually coherent, but all sources are mutually incoherent over the response time of the holographic medium. Two optically addressed spatial light modulators are employed to represent the input and output planes. Parallel illumination of the spatial light modulators by the source array produces (with appropriate optics) sets of mutually incoherent holograms in the volume holographic recording medium, all of which are doubly angularly multiplexed to avoid the effects of beam degeneracy.

Simulations of this architecture using the optical beam propagation method have shown a remarkable reduction in interchannel crosstalk, accompanied by a corresponding substantial increase in optical throughput efficiency. Laboratory demonstrations of 8×8 and 2×64 interconnections in real time photorefractive crystals have confirmed the theoretical and numerical predictions. Extension of both the numerical simulations and the laboratory demonstrations to larger array sizes is under way at the present time.

The Stratified Volume Holographic Optical Element or SVHOE is a unique optical element capable of emulating volume holographic diffraction characteristics in devices consisting of only thin layered photosensitive materials, and at the same time exhibits a large number of novel properties that are useful in array interconnection, wavelength multiplexing and demultiplexing, and spatial frequency filtering applications [Appendices 3 and 5; Publs. 3, 5, 6, 7; Conf. Pres. 1, 2, 4, 6, 9, 11, 12, 14, 15, 16, 17, 21, 29, 32, 33, 49, 50, 52].

An SVHOE consists of a number of thin photosensitive layers separated by substrate (or buffer) layers that are optically insensitive. This type of construction is particularly advantageous for materials that exhibit large photoinduced refractive index variations, but that are difficult to fabricate in the thicknesses required for highly multiplexed volume holographic device operation (of order 1 mm to 1 cm). Examples of such materials include the commercially available DuPont Holopolymer material, nonlinear organic materials, and III-V compound semiconductor multiple quantum well structures. In operation, the device is exposed to two coherent recording beams that interfere in the photosensitive layers to record the hologram, and readout is performed in direct analogy with other volume holographic optical elements.

A number of novel diffraction characteristics of SVHOE structures have been identified, numerically modeled, and in most cases experimentally demonstrated during the

contract period. These include a unique periodic angular tuning characteristic, the emulation of Bragg limited angular response characteristics, the generation of regularly spaced arrays of output angles (or positions) when illuminated by a strongly focused beam, spatial frequency notch filtering, wavelength notch filtering, and wavelength multiplexing and demultiplexing. In addition, SVHOE structures fabricated with active photorefractive materials, such as III-V compound semiconductor multiple quantum well structures that are voltage or field enabled, are capable of exhibiting tunable diffraction characteristics that can be externally modulated by altering the distribution of applied voltages on a layer by layer basis. Preliminary studies of such active SVHOE structures in quantum well devices have been completed, and fabrication and evaluation of test structures is under way. Finally, demonstration of key SVHOE characteristics with the DuPont Holopolymer material is underway in a cooperative effort with DuPont, in order to take advantage of the *in situ* fixing capabilities of this class of organic materials.

In addition to the investigation of the fundamental limitations of photorefractive materials (as used in interconnection applications) described above [Appendices 4 and 7; Publ. 4; Book Chs. 2, 3; Conf. Pres. 3, 5, 7, 13, 15, 33], we have also undertaken a multifaceted effort to improve the prospects for practical implementation of photorefractive devices. This effort has included the development of multilayer antireflection coatings on $\text{Bi}_{12}\text{SiO}_{20}$, LiNbO_3 , BaTiO_3 , SBN, GaAs, and CdTe that eliminate the source of multiple internal reflections within the photorefractive material [Publs. 13, 14, 15; Conf. Pres. 25, 33, 34, 35, 49, 50, 52], resulting in 100% increases in the saturation diffraction efficiency and two beam coupling gain; the identification and characterization of the electric field collapse within photorefractive crystals that occurs during grating recording, by means of a novel transverse electrooptic imaging method [Publ. 17; Conf. Pres. 27, 32, 33]; the development of a novel electrooptic technique for the measurement of very high dark resistivities in electrooptic crystals (which determine the grating storage time) [Appendix 1;

Publ. 1; Conf. Pres. 8, 18, 22]; the application of the electrooptic measurement technique to the determination of the dark resistivities of undoped, doped, and nonstoichiometric single crystals of bismuth silicon oxide [Appendix 1; Publ. 1; Conf. Pres. 8, 18, 22]; the analysis of the polarization properties of diffraction in optically active and electrooptic photorefractive materials, including the effects of self-diffraction during grating recording [Appendix 2; Publ. 2; Conf. Pres. 26]; and the observation and characterization of the effects of microscopic charge screening on sub-hologram formation in photorefractive materials [Publ. 8; Conf. Pres. 30], which has important implications for both the implementation of spatially segmented volume holograms in real time materials, as well as for the analysis of the effects of macroscopic space-variant illumination effects in the recording of full aperture volume holograms.

3. PUBLICATIONS UNDER DARPA/AFOSR SPONSORSHIP

The following technical publications and conference presentations describe research supported in part by this contract.

3.1 Journal Publications

1. D. A. Seery, M. H. Garrett, and A. R. Tanguay, Jr., "Electrooptic Measurement of the Volume Resistivity of Bismuth Silicon Oxide ($\text{Bi}_{12}\text{SiO}_{20}$)", *Journal of Crystal Growth*, **85**, 282-289, (1987).
2. A. Marrakchi, R. V. Johnson, and A. R. Tanguay, Jr., "Polarization Properties of Enhanced Self-Diffraction in Sillenite Crystals", *IEEE Journal of Quantum Electronics*, Special Issue on Electrooptic Materials and Devices, **QE-23**(12), 2142-2151, (1987).
3. R. V. Johnson and A. R. Tanguay, Jr., "Stratified Volume Holographic Optical Elements", *Optics Letters*, **13**(3), 189-191, (1988).
4. A. R. Tanguay, Jr., "Physical and Technological Limitations of Optical Information Processing and Computing", *Materials Research Society Bulletin*, Special Issue on Photonic Materials, **XIII**(8), 36-40, (1988); (Invited Paper).
5. R. V. Johnson and A. R. Tanguay, Jr., "Stratified Volume Holographic Optical Elements", *Optics News*, Special Issue on "Optics in 1988", **14**(12), 30-31, (1988); (Invited Paper).
6. G. P. Nordin, R. V. Johnson, and A. R. Tanguay, Jr., "Physical Characterization of Stratified Volume Holographic Optical Elements", in preparation for *Optics Letters*.
7. G. P. Nordin, R. V. Johnson, and A. R. Tanguay, Jr., "Diffraction Properties of Stratified Volume Holographic Optical Elements", in preparation for *Journal of the Optical Society of America*.

8. P. Asthana and A. R. Tanguay, Jr., "Charge-Screening-Induced Switching in Spatially Multiplexed Sub-Holograms in $\text{Bi}_{12}\text{SiO}_{20}$ ", in preparation for Applied Physics Letters.
9. P. Asthana, G. P. Nordin, A. R. Tanguay, Jr., and B. K. Jenkins, "Beam Degeneracy Crosstalk in Fan-Out/Fan-In Volume Holographic Interconnections", in preparation for Optics Letters.
10. P. Asthana, G. P. Nordin, A. R. Tanguay, Jr., and B. K. Jenkins, "Analysis and Minimization of Inter-channel Crosstalk in Volume Holographic Interconnections for Optical Neural Networks", in preparation for Journal of the Optical Society of America.
11. J. H. Rilum and A. R. Tanguay, Jr., "Characterization of Optical Memory Discs for Optical Information Processing Applications", in preparation for Applied Optics.
12. J. H. Rilum and A. R. Tanguay, Jr., "Differential Interferometric Readout Optical Memory Disc Spatial Light Modulators", in preparation for Optics Letters.
13. Z. Karim and A. R. Tanguay, Jr., "A Bandpass AR Coating Design for Bismuth Silicon Oxide", in preparation for Applied Physics Letters.
14. Z. Karim and A. R. Tanguay, Jr., "A Bandpass AR Coating for the Photorefractive Materials LiNbO_3 , BaTiO_3 , CdTe , and PLZT ", in preparation for Applied Optics.
15. Z. Karim, C. Kyriakakis, and A. R. Tanguay, Jr., "Improved Two-Beam Coupling Gain and Diffraction Efficiency in Bismuth Silicon Oxide Crystals Using a Bandpass AR Coating", in preparation for Applied Physics Letters.
16. C. Kyriakakis, Z. Karim, J. H. Rilum, J. J. Jung, A. R. Tanguay, Jr., and A. Madhukar, "Fundamental and Technological Limitations of Asymmetric Cavity MQW InGaAs/GaAs Spatial Light Modulators", in preparation for IEEE Journal of Quantum Electronics.
17. E. J. Herbulock and A. R. Tanguay, Jr., "Electric Field Profile Effects on Photorefractive Grating Formation in Bismuth Silicon Oxide", in preparation for Applied Physics Letters.

3.2 Book Chapters

1. J. Yu, D. Psaltis, A. Marrakchi, A. R. Tanguay, Jr., and R. V. Johnson, "Photorefractive Incoherent-to-Coherent Optical Conversion," in *Photorefractive Materials and Applications*, J. P. Huignard and P. Gunter, Eds., Springer-Verlag, New York, (1989).
2. R. V. Johnson and A. R. Tanguay, Jr., "Fundamental Physical Limitations of the Photorefractive Grating Recording Sensitivity", in *Optical Processing and Computing*, H. Arsenault, T. Szoplik, and B. Macukow, Eds., Academic Press, New York, (1989).
3. B. K. Jenkins and A. R. Tanguay, Jr., "Photonic Implementations of Neural Networks", Chapter 15 in *Neural Networks and Fuzzy Systems: A Dynamical Approach to Machine Intelligence*, B. Kosko, Ed., Prentice Hall, Englewood Cliffs, New Jersey, (1991).

3.3 Patents

1. "Incoherent/Coherent Multiplexed Holographic Recording for Photonic Interconnections and Holographic Optical Elements", B. K. Jenkins and A. R. Tanguay, Jr., patent pending.

3.4 Conference Presentations

1. A. R. Tanguay, Jr., "Optical Information Processing Components", Georgia Institute of Technology, Atlanta, Georgia, (1987); (Invited Colloquium).
2. A. R. Tanguay, Jr., and R. V. Johnson, "Stratified Volume Holographic Optical Elements", Conference on Lasers and Electro-Optics, Baltimore, Maryland, (1987).
3. A. R. Tanguay, Jr., "Fundamental Physical Limitations of Optical Information Processing and Computing", NATO Collaborative Award Colloquium, University College London, London, England, (1987); (Invited Colloquium).
4. A. R. Tanguay, Jr., "Optical Information Processing Components", GEC Research, Marconi Research Centre, Chelmsford, England, (1987); (Invited Colloquium).

5. A. R. Tanguay, Jr., "Fundamental Physical Limitations of Optical Information Processing and Computing", British Telecom Research Laboratories, Ipswich, England, (1987); (Invited Colloquium).
6. A. R. Tanguay, Jr., "Optical Information Processing Components", University of California at Santa Barbara, Santa Barbara, California, (1987); (Invited Colloquium).
7. A. R. Tanguay, Jr., "Fundamental Physical Limitations of Optical Information Processing and Computing", University of Rochester, Institute of Optics, Rochester, New York, (1987); (Invited Colloquium).
8. M. H. Garrett, D. A. Seery, and A. R. Tanguay, Jr., "Electrooptic Measurement of the Volume Resistivity of Bismuth Silicon Oxide ($\text{Bi}_{12}\text{SiO}_{20}$)", American Conference on Crystal Growth-7, Monterey, California, (1987); (Invited Paper).
9. A. R. Tanguay, Jr., "Optical Processing and Computing Devices: The Materials Perspective", American Conference on Crystal Growth-7, Monterey, California, (1987); (Invited Paper).
10. A. R. Tanguay, Jr., "The Professor Profession", 1987 Annual Meeting of the Optical Society of America, Rochester, New York, (1987); (Invited Presentation).
11. A. R. Tanguay, Jr., "Optical Information Processing Components", Bell Communications Research, Morristown, New Jersey, (1987), (Invited Colloquium).
12. A. R. Tanguay, Jr., "Optical Information Processing Components", IBM Almaden Research Center, San Jose, California, (1987); (Invited Colloquium).
13. A. R. Tanguay, Jr., "Fundamental Physical Limitations of Optical Information Processing and Computing", DARPA Panel on Neural Networks, California Institute of Technology, Pasadena, California, (1987); (Invited Presentation).
14. A. R. Tanguay, Jr., and R. V. Johnson, "Stratified Volume Holographic Optical Elements", IEEE-LEOS Southern California Section Winter Regional Meeting, University of Southern California, Los Angeles, California, (1988).

15. A. R. Tanguay, Jr., "Photonic Materials and Devices for Optical Information Processing and Computing Applications", DARPA Annual Conference on Optical Processing and Computing, Leesburg, Virginia, (1988); (Invited Paper).
16. A. R. Tanguay, Jr., "Electro-Optical Information Processing and Computing Components", Conference on Lasers and Electro-Optics, Anaheim, California, (1988); (Invited Paper).
17. A. R. Tanguay, Jr., and R. V. Johnson, "Spatial Light Modulators and Volume Holographic Optical Elements", Symposium of the Center for the Integration of Optical Computing, University of Southern California, Los Angeles, California, (1988); (Invited Paper).
18. M. H. Garrett and A. R. Tanguay, Jr., "Crystal Growth and Characterization of Nonstoichiometric Bismuth Silicon Oxide ($\text{Bi}_x\text{SiO}_{1.5x+2}$)", AACG/West Tenth Conference on Crystal Growth, Fallen Leaf Lake, California, (1988).
19. C. Kyriakakis, P. Asthana, R. V. Johnson, and A. R. Tanguay, Jr., "Spatial Light Modulators: Fundamental and Technological Issues", Optical Society of America Topical Meeting on Spatial Light Modulators, Lake Tahoe, Nevada, (1988); (Invited Paper).
20. S. Mroczkowski and A. R. Tanguay, Jr., "Impurity Induced Photochromic Behavior in Bismuth Silicon Oxide ($\text{Bi}_{12}\text{SiO}_{20}$)", American Conference on Crystal Growth/East-2, Atlantic City, New Jersey, (1988).
21. G. P. Nordin, R. V. Johnson, and A. R. Tanguay, Jr., "Physical Characterization of Stratified Volume Holographic Optical Elements", 1988 Annual Meeting of the Optical Society of America, Santa Clara, California, Vol. 11 of the 1988 OSA Technical Digest Series, p. 106, (1988).
22. M. H. Garrett, S. W. McCahon, and A. R. Tanguay, Jr., "Crystal Growth and Characterization of Nonstoichiometric Bismuth Silicon Oxide, $\text{Bi}_{12}\text{SiO}_{1.5x+2}$ ", 1988 Annual Meeting of the Optical Society of America, Santa Clara, California, Vol. 11 of the 1988 OSA Technical Digest Series, p. 106, (1988).

23. P. Asthana and A. R. Tanguay, Jr., "Charge-Screening-Induced Switching in Spatially Multiplexed Sub-Holograms in $\text{Bi}_{12}\text{SiO}_{20}$ ", 1988 Annual Meeting of the Optical Society of America, Santa Clara, California, Vol. 11 of the 1988 OSA Technical Digest Series, p. 151, (1988).
24. J. H. Rilum and A. R. Tanguay, Jr., "Utilization of Optical Memory Discs for Optical Information Processing Applications", 1988 Annual Meeting of the Optical Society of America, Santa Clara, California, Vol. 11 of the 1988 OSA Technical Digest Series, p. 43, (1988).
25. Z. Karim, M. H. Garrett, and A. R. Tanguay, Jr., "A Bandpass AR Coating Design for Bismuth Silicon Oxide", 1988 Annual Meeting of the Optical Society of America, Santa Clara, California, Vol. 11 of the 1988 OSA Technical Digest Series, p. 125, (1988).
26. A. Marrakchi, R. V. Johnson, and A. R. Tanguay, Jr., "Polarization Properties of Enhanced Self-Diffraction in Sillenite Crystals", 1988 Annual Meeting of the Optical Society of America, Santa Clara, California, Vol. 11 of the 1988 OSA Technical Digest Series, p. 107, (1988).
27. E. J. Herbulock, M. H. Garrett, and A. R. Tanguay, Jr., "Electric Field Profile Effects on Photorefractive Grating Formation in Bismuth Silicon Oxide", 1988 Annual Meeting of the Optical Society of America, Santa Clara, California, Vol. 11 of the 1988 OSA Technical Digest Series, p. 143, (1988).
28. S. Mroczkowski and A. R. Tanguay, Jr., "Crystal Growth and Impurity Induced Photochromic Behavior in Bismuth Silicon Oxide ($\text{Bi}_{12}\text{SiO}_{20}$)", The Eighth National Conference on Crystal Growth and Materials, Guilin, China, (1988).
29. G. P. Nordin, R. V. Johnson, and A. R. Tanguay, Jr., "Physical Characterization of Stratified Volume Holographic Optical Elements", Signal and Image Processing Institute Annual Research Review, University of Southern California, Los Angeles, California, (1989).

30. P. Asthana and A. R. Tanguay, Jr., "Charge-Screening-Induced Switching in Spatially Multiplexed Sub-Holograms in $\text{Bi}_{12}\text{SiO}_{20}$ ", Signal and Image Processing Institute Annual Research Review, University of Southern California, Los Angeles, California, (1989).
31. M. Hibbs-Brenner, S. D. Mukherjee, M. P. Bendett, and A. R. Tanguay, Jr., "Integrated Optoelectronic Cellular Array for Fine-Grained Parallel Processing Systems", Proc. OSA Topical Meeting on Optical Computing, Salt Lake City, Utah, (1989).
32. A. R. Tanguay, Jr., "Device Development for Optical Computing", Conference on Lasers and Electro-Optics (CLEO '89), Baltimore, Maryland, (1989); (Invited Paper).
33. A. R. Tanguay, Jr., "Photonic Materials and Devices for Optical Information Processing and Computing Applications", DARPA Annual Conference on Optical Processing and Computing, Airlie, Virginia, (1989).
34. Z. Karim, C. Kyriakakis, and A. R. Tanguay, Jr., "Improved Two-Beam Coupling Gain and Diffraction Efficiency in Bismuth Silicon Oxide Crystals Using a Bandpass AR Coating", 1989 Annual Meeting of the Optical Society of America, Orlando, Florida, Vol. 18 of the 1989 OSA Technical Digest Series, p. 29, (1989).
35. Z. Karim and A. R. Tanguay, Jr., "Bandpass AR Coating for the Photorefractive Materials LiNbO_3 , BaTiO_3 , CdTe , and PLZT ", 1989 Annual Meeting of the Optical Society of America, Orlando, Florida, Vol. 18 of the 1989 OSA Technical Digest Series, p. 78, (1989).
36. J. H. Rilum and A. R. Tanguay, Jr., "Performance Characteristics of Optical Memory Disc Spatial Light Modulators", 1989 Annual Meeting of the Optical Society of America, Orlando, Florida, Vol. 18 of the 1989 OSA Technical Digest Series, pp. 171-172, (1989).
37. B. K. Jenkins, G. C. Petrisor, S. Piazzolla, P. Asthana, and A. R. Tanguay, Jr., "Photonic Architecture for Neural Network Implementation", Signal and Image Processing Institute Annual Research Review, University of Southern California, Los Angeles, California, (1990).

38. P. Asthana, H. Chin, G. Nordin, A. R. Tanguay, Jr., S. Piazzolla, B. K. Jenkins, and A. Madhukar, "Component Technology Development for Optical Implementations of Neural Networks", Signal and Image Processing Institute Annual Research Review, University of Southern California, Los Angeles, California, (1990).
39. P. Asthana, G. Nordin, H. Chin, and A. R. Tanguay, Jr., "Incoherent/Coherent Holographic Interconnections and Optoelectronic Components for Application to Optical Neural Networks", Signal and Image Processing Institute Annual Research Review, University of Southern California, Los Angeles, California, (1990).
40. J. H. Rilum and A. R. Tanguay, Jr., "Optical Memory Disc Based Neural Networks", Signal and Image Processing Institute Annual Research Review, University of Southern California, Los Angeles, California, (1990).
41. B. K. Jenkins, G. C. Petrisor, S. Piazzolla, P. Asthana, and A. R. Tanguay, Jr., "Photonic Architecture for Neural Nets Using Incoherent/Coherent Holographic Interconnections", Proceedings of the International Conference on Optical Computing OC'90, Kobe, Japan, (1990).
42. P. Asthana, H. Chin, G. Nordin, A. R. Tanguay, Jr., S. Piazzolla, B. K. Jenkins, and A. Madhukar, "Photonic Components for Neural Net Implementations Using Incoherent/Coherent Holographic Interconnections", Proceedings of the International Conference on Optical Computing OC'90, Kobe, Japan, (1990).
43. A. R. Tanguay, Jr., "Fundamental Physical and Technological Considerations for Spatial Light Modulation", Proceedings of the International Conference on Optical Computing OC'90, Kobe, Japan, (1990); (Invited Paper).
44. A. R. Tanguay, Jr., "Comments on Photonic Information Systems", ATT Workshop on Free-Space Digital Optics, Naperville, Illinois, (1990).
45. B. K. Jenkins and A. R. Tanguay, Jr., "Photonic Neural Networks with Incoherent/Coherent Holographic Interconnections", Joint USA (NSF)/Korea (KOSEF) Workshop on Optical Neural Networks, Seoul, South Korea, (1990).

46. C. Kyriakakis, Z. Karim, J. H. Rilum, J. J. Jung, A. R. Tanguay, Jr., and A. Madhukar, "Fundamental and Technological Limitations of Asymmetric Cavity MQW InGaAs/GaAs Spatial Light Modulators", OSA Topical Conference on Spatial Light Modulators and Applications, Incline Village, Nevada, Vol. 14 of the 1990 OSA Technical Digest Series, pp. 7-10, (1990).
47. J. H. Rilum and A. R. Tanguay, Jr., "Device Characteristics of Optical Disc Spatial Light Modulators", OSA Topical Conference on Spatial Light Modulators and Applications, Incline Village, Nevada, Vol. 14 of the 1990 OSA Technical Digest Series, pp. 200-203, (1990).
48. A. R. Tanguay, Jr., "Advances in Spatial Light Modulator Technology", OSA Topical Conference on Spatial Light Modulators and Applications, Incline Village, Nevada, Vol. 14 of the 1990 OSA Technical Digest Series, (1990).
49. P. Asthana, H. Chin, S. de Mars, E. Herbulock, Z. Karim, C. Kyriakakis, G. Nordin, J. H. Rilum, and A. R. Tanguay, Jr., "The Critical Role of Dielectric and Optoelectronic Materials in Optical Information Processing and Computing Devices", Gordon Research Conference on Dielectric Materials, Plymouth, New Hampshire, (1990); (Invited Paper).
50. P. Asthana, H. Chin, S. de Mars, E. Herbulock, Z. Karim, C. Kyriakakis, G. Nordin, J. H. Rilum, and A. R. Tanguay, Jr., "The Critical Role of Dielectric and Optoelectronic Materials in Optical Information Processing and Computing Devices", DARPA Materials Research Council Meeting on Optical Computing, La Jolla, California, (1990); (Invited Paper).
51. C. Kyriakakis, P. Asthana, Z. Karim, and A. R. Tanguay, Jr., "Fundamental and Technological Limitations of Optical Information Processing and Computing", DARPA Materials Research Council Meeting on Optical Computing, La Jolla, California, (1990); (Invited Paper).
52. A. R. Tanguay, Jr., "Photonic Materials and Devices for Optical Information Processing and Computing Applications", DARPA Annual Conference on Optical Processing and Computing, Reston, Virginia, (1990).

53. B. K. Jenkins, A. R. Tanguay, Jr., S. Piazzolla, G. C. Petrisor, and P. Asthana, "Photonic Neural Network Architecture Based on Incoherent-Coherent Holographic Interconnections", 1990 Annual Meeting of the Optical Society of America, Boston, Massachusetts, Vol. 15 of the 1990 OSA Technical Digest Series, p. 56, (1990).
54. P. Asthana, H. Chin, G. Nordin, A. R. Tanguay, Jr., G. C. Petrisor, B. K. Jenkins, and A. Madhukar, "Photonic Components for Neural Network Implementations Using Incoherent-Coherent Holographic Interconnections", 1990 Annual Meeting of the Optical Society of America, Boston, Massachusetts, Vol. 15 of the 1990 OSA Technical Digest Series, p. 57, (1990).
55. J. H. Rilum and A. R. Tanguay, Jr., "Optical Memory Disc Spatial Light Modulators", 1990 Annual Meeting of the Optical Society of America, Boston, Massachusetts, Vol. 15 of the 1990 OSA Technical Digest Series, p. 72, (1990); (Invited Paper).
56. C. Kyriakakis, P. Asthana, Z. Karim, G. Nordin, J. Rilum, and A. R. Tanguay, Jr., "Fundamental Physical and Technological Constraints on Optical Information Processing and Computing", 1990 Annual Meeting of the Optical Society of America, Boston, Massachusetts, Vol. 15 of the 1990 OSA Technical Digest Series, p. 241, (1990); (Invited Paper).
57. P. Asthana, G. Nordin, S. Piazzolla, A. R. Tanguay, Jr., and B. K. Jenkins, "Analysis of Interchannel Crosstalk and Throughput Efficiency in Highly Multiplexed Fan-out/Fan-in Holographic Interconnections", 1990 Annual Meeting of the Optical Society of America, Boston, Massachusetts, Vol. 15 of the 1990 OSA Technical Digest Series, p. 242, (1990).

ELECTROOPTIC MEASUREMENT OF THE VOLUME RESISTIVITY OF BISMUTH SILICON OXIDE ($\text{Bi}_{12}\text{SiO}_{20}$)

David A. SEERY, Mark H. GARRETT and Armand R. TANGUAY, Jr.

Optical Materials and Devices Laboratory, University of Southern California, 523 Sower Science Center, University Park, MC-0483, Los Angeles, California 90089-0483, U.S.A.

Single crystals of bismuth silicon oxide ($\text{Bi}_{12}\text{SiO}_{20}$) and its isomorphs (including, for example, bismuth germanium oxide ($\text{Bi}_{12}\text{GeO}_{20}$)) have been utilized in a wide range of active electrooptic and acoustooptic devices, including the Pockels Readout Optical Modulator (PROM), the PRIZ, the Photorefractive Incoherent-to-Coherent Optical Converter (PICOC), volume holographic storage devices, and surface acoustic wave devices. A key material parameter that influences device performance characteristics is the volume resistivity, which is difficult to measure accurately using standard techniques in refractory oxides like $\text{Bi}_{12}\text{SiO}_{20}$ due to its large magnitude (typically $> 10^{13} \Omega \text{ cm}$). We present here a technique for the measurement of such very high resistivities in electrooptic materials; this method utilizes the electrooptic modulation induced by a voltage placed across the (crystallographically oriented) sample as a probe of temporal voltage transients that are in turn directly related to the sample volume resistivity. In our experiments, a very weak optical probe is frequency modulated, phase detected, and employed at low duty cycle to avoid ambiguities due to photoconductive voltage decay. The technique is described in detail, and experimental results are presented on a number of undoped and doped samples of bismuth silicon oxide grown by the Czochralski technique.

1. Introduction

Bismuth silicon oxide is a wide bandgap, high resistivity semi-insulator that is also photoconductive, electrooptic, acoustooptic, magneto-optic, and optically active. As such, it has found widespread application in optical information processing and computing components such as spatial light modulators and volume holographic optical elements [1]. Examples of such devices include the Pockels Readout Optical Modulator (PROM) [2,3], the Photorefractive Incoherent-to-Coherent Optical Converter (PICOC) [4], the Optically Modulated Total Internal Reflection Spatial Light Modulator (OM-TIR SLM) [5], and photorefractive volume holographic optical elements (VHOEs) [1,6].

Several important operational parameters of both electrooptic spatial light modulators and volume holographic optical elements depend directly on the volume resistivity (ρ) of the active electrooptic material. For example, the dielectric relaxation time ($\tau = \rho \epsilon \epsilon_0$, in which ϵ is the relative dielectric permittivity of the material and ϵ_0 is the permittivity of free space) establishes the frame storage time in the PROM, PICOC, and OM-TIR

spatial light modulators, as well as the maximum frame integration time. In addition, the dielectric relaxation time determines the holographic grating storage time in $\text{Bi}_{12}\text{SiO}_{20}$ volume holographic optical elements. Furthermore, spatial nonuniformities in the resistivity will produce undesirable space-variant erasure and image decay characteristics.

Determination of the volume resistivity also has significant implications for materials characterization. Just as in the case of semiconductor materials, the as-grown resistivity provides an indication of both overall crystal purity and crystallographic perfection, as well as potential for correlation with impurity and dopant analysis techniques. This latter point is especially significant, since such techniques are at present not well developed for low impurity/dopant levels in dielectric matrices. In fact, resistivity variations can yield indirect information regarding the nature of states (shallow level, deep level, recombination center) induced by specific selected incorporants. Finally, the dark resistivity provides a critical baseline against which the photoconductive sensitivity can be evaluated.

Traditional techniques for measuring resistivity

ties of order $10^{13} \Omega \text{ cm}$ and greater, such as direct measurement of the current-voltage relationship, use of the four-point probe [7], and Van der Pauw's method [8], are impractical due to the resultant high voltages and extremely low currents implied by the total sample resistance. Several other difficulties bear on the measurement of very high resistivities, including the surface conductivity of the specimen crystal and crystal mount, the possible influence of humidity-enhanced surface conduction, and the dependence of the empirically-derived dark resistivity on the optical and thermal exposure history of the sample.

We have utilized an electrooptic measurement technique as described herein that largely circumvents these difficulties, in order to evaluate the dark volume resistivities of both undoped and doped samples of single crystal bismuth silicon oxide. The technique is applicable to a wide range of electrooptic refractory oxides, such as bismuth germanium oxide ($\text{Bi}_{12}\text{GeO}_{20}$), bismuth titanium oxide ($\text{Bi}_{12}\text{TiO}_{20}$), barium titanate (BaTiO_3), lithium niobate (LiNbO_3), and strontium barium niobate ($\text{Sr}_{x}\text{Ba}_{1-x}\text{Nb}_2\text{O}_6$).

The theoretical basis of the experimental technique is presented in section 2, and the details of the method are described in section 3. Results of resistivity measurements on representative undoped and doped single crystals of $\text{Bi}_{12}\text{SiO}_{20}$ are given in section 4. Discussion and conclusions are provided in section 5.

2. Theoretical considerations

The essence of the measurement technique is the formation of a parallel plate capacitor with the specimen electrooptic single crystal acting as the dielectric, by depositing semi-transparent counter-electrodes on two optically polished, parallel faces of the sample. The electrooptic effect is then utilized to provide a direct measurement of the instantaneous voltage applied across the sample. A high voltage is first applied to the sample in order to charge the capacitor to an initial voltage V_0 , and is then disconnected to allow the resultant transient voltage decay of the sample capacitor in parallel with its internal resistance to be moni-

tored. The time constant of the voltage decay provides a direct measurement of the resistivity of the sample independent of geometric factors, as outlined below.

Consider a single crystal sample of $\text{Bi}_{12}\text{SiO}_{20}$ oriented along the $[001]$ direction as shown schematically in fig. 1. Bismuth silicon oxide is a cubic (I23) non-centrosymmetric crystal, characterized by a single value of the electrooptic coefficient $r_{41} = r_{52} = r_{63}$. Hence, a voltage $V = E_{[001]} d$ applied across the transparent conductive electrodes (in which d is the sample thickness) will induce principal axes along the $[110]$ and $[\bar{1}10]$ directions, with indices of refraction given by [9]:

$$n_{[110]} = n_0 - \frac{1}{2} n_0^3 r_{41} E_{[001]} \quad (1)$$

and

$$n_{[\bar{1}10]} = n_0 + \frac{1}{2} n_0^3 r_{41} E_{[001]}. \quad (2)$$

In eqs. (1) and (2) above, the natural optical activity of bismuth silicon oxide is neglected (as will be discussed further below). Consider further propagation of an optical probe beam in the $[001]$ direction, polarized along the $[010]$ direction. The two resultant principal components will experi-

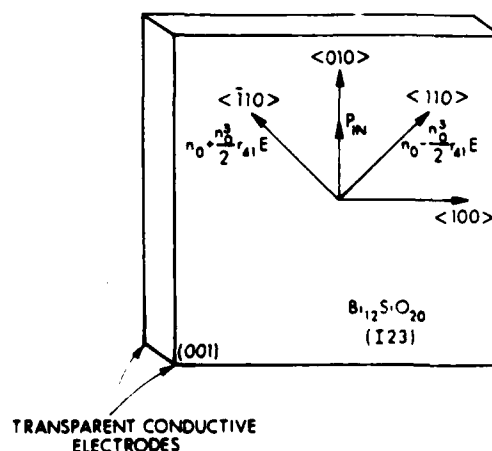


Fig. 1. Typical electrooptic configuration for measurement of the volume resistivity of bismuth silicon oxide ($\text{Bi}_{12}\text{SiO}_{20}$). Both crystallographic and electric-field-induced principal axes are shown for a sample oriented along the $[001]$ direction.

ence a net phase shift given by:

$$\Gamma = \frac{2\pi}{\lambda} \Delta n d = \frac{2\pi}{\lambda} n_0^3 r_{41} V. \quad (3)$$

Note that in this longitudinal electrooptic configuration, the phase difference depends only on the applied voltage, and not on the internal electric field. For an analyzer oriented along the [100] direction, the resultant transmitted intensity will be:

$$I = I_0 \sin^2(\Gamma/2) = \sin^2\left(\frac{\pi}{\lambda} n_0^3 r_{41} V\right). \quad (4)$$

This expression illustrates the one-to-one correspondence between the transmitted intensity through the system and the instantaneous voltage across the crystal (provided that the applied voltage is always less than the half wave voltage $V_{\pi} = \lambda/2n_0^3 r_{41}$).

The transient voltage decay of a resistor in parallel with a capacitor is given by:

$$V(t) = V_0 \exp(-t/\tau), \quad (5)$$

in which τ is the characteristic RC time constant. For a parallel plate capacitor of sufficiently high aspect ratio to avoid significant field fringing effects, the RC product is approximately equal to the dielectric relaxation time $\rho\epsilon\epsilon_0$, which is thus independent of geometric parameters. Since the relative dielectric permittivity of $\text{Bi}_{12}\text{SiO}_{20}$ is known ($\epsilon = 56$ [10]), measurement of the voltage decay time constant directly yields the desired volume resistivity.

Two issues are particularly worthy of note at this point in the discussion. First and foremost, the measurement technique utilized herein (as do all resistivity measurement techniques) assumes a linear, or at least piecewise linear, current-voltage relationship. As will be shown in section 4, bismuth silicon oxide samples exhibit significant non-linear current-voltage behavior in the high field regime ($> 10^5$ V/cm). Hence the measurement described herein is more accurately a measure of the differential volume (dark) resistivity at the specified initial value of applied voltage V_0 . We have chosen a value of V_0 (1600 V) which is of particular relevance to the device applications discussed in section 1. The implications of this point

will be discussed further in section 4.

The second issue concerns the absolute orientation of the single crystal sample, the effects of finite absorption coefficients, and the natural optical activity of $\text{Bi}_{12}\text{SiO}_{20}$. As described in section 3, an essentially arbitrarily oriented sample with any combination of linear and nonlinear optical properties can be easily accommodated by the simple device of obtaining a calibration curve of transmitted intensity as a function of voltage during the initial charging of the capacitor. The only requirement is that a measurable fraction of the transmitted light experience an electrooptic (or electrorefractive) effect such that a monotonic calibration curve can be established. This further eliminates any experimental dependence on the measurement wavelength, the electrooptic coefficient, the index of refraction, the crystal thickness, the angle of incidence, and the optical rotatory power.

Finally, a similar resistivity measurement technique has been described previously [11] in which the sample to be measured is connected in parallel with an electrooptic crystal, which in turn is utilized to measure the voltage decay. This technique is applicable for the measurement of resistivities small compared with that of the calibrated sample.

3. Experimental procedure

The basic configuration utilized for the measurement of high volume-resistivities in electrooptic crystals is as shown in fig. 2. The sample to be measured is typically oriented as shown in fig. 1 and mounted in an environmental chamber to allow control of the ambient atmosphere. The chamber is flushed with filtered ultra high purity dry N_2 gas prior to and during the measurement procedure in order to minimize adsorption of surface moisture and contaminants. Incident probe light derived from a polarized helium neon laser (chosen to minimize the photoconductive decay of the bismuth silicon oxide samples) is spatially filtered, collimated, apertured, and passed through a final polarizer before transmittal through the sample at near normal incidence and a polariza-

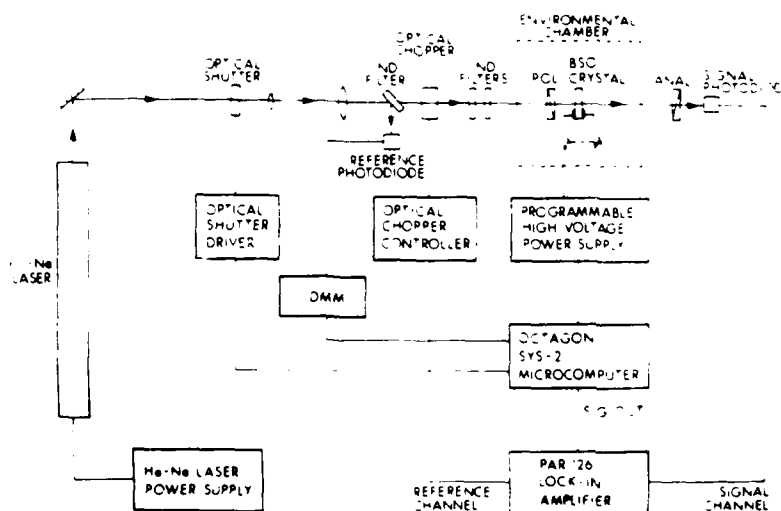


Fig. 2. The experimental configuration for measurement of volume resistivity in electrooptic materials, as described in detail in section 3.

tion analyzer. Light passing the analyzer is detected by a silicon photodetector, which provides the signal channel input to a phase sensitive detector (Princeton Applied Research Model 126 Lock-In Amplifier). The reference output of the lock-in is used to drive an optical chopper through a controller to complete a phase-locked loop. This allows operation at very low signal amplitudes (incident intensities), again minimizing photoconductive decay of the sample voltage. The average incident power during the course of the measurement is further reduced by utilization of neutral density filters and a computer controlled optical shutter that remains closed at all times except during brief measurement intervals.

Voltage is applied to the crystal by means of a programmable high voltage power supply through a single pole, single throw knife switch. Both a mercury relay and a high voltage contactor were employed in trial runs in the hopes of implementing full computer control of the voltage disconnect sequence, but proved to have intolerably high parasitic capacitances. The knife switch was automated for computer control by incorporation of a long throw solenoid interfaced to the microcomputer.

In operation, the microcomputer increases the

voltage across the sample in increments, recording the transmitted intensity at each incremental voltage. This produces the requisite calibration curve for direct interpretation of the voltage transient decay, as described in section 2. Examples of such calibration curves are shown in fig. 3 for the four samples employed in this study. Following completion of the calibration sequence, the optical shutter is automatically closed, and the knife switch automatically opened by means of the electromechanical solenoid. The microcomputer controls subsequent data acquisition during the decay transient, controlling the shutter, recording the transmitted intensity, and recording the instantaneous laser power by means of a reference photodiode for purposes of normalization. Voltage decay measurements were typically recorded every 15 min, with five independent measurements of the intensity at a given time averaged to form each plotted data point. Dependent on the decay time constant to be measured, voltage decay runs varied in duration from 15 min to over one hundred hours.

A typical series of voltage decay curves are plotted in fig. 4 for one of the undoped $\text{Bi}_{12}\text{SiO}_{20}$ samples. Four curves are shown, for average power levels during measurement of 2 μW , 200 nW, 20

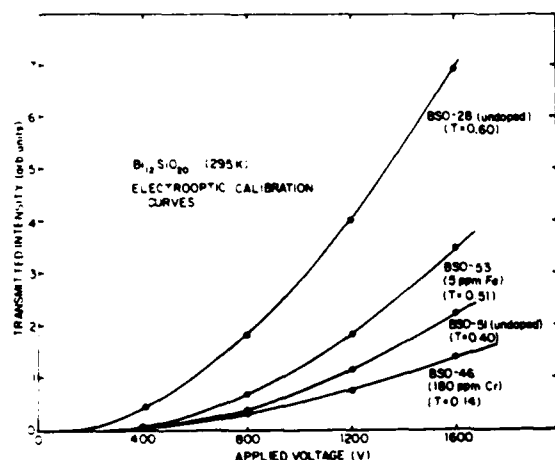


Fig. 3. Electrooptic calibration curves for the four samples utilized in this study. The quantity T is the intensity transmission coefficient of each sample at the probe wavelength, uncorrected for reflection losses.

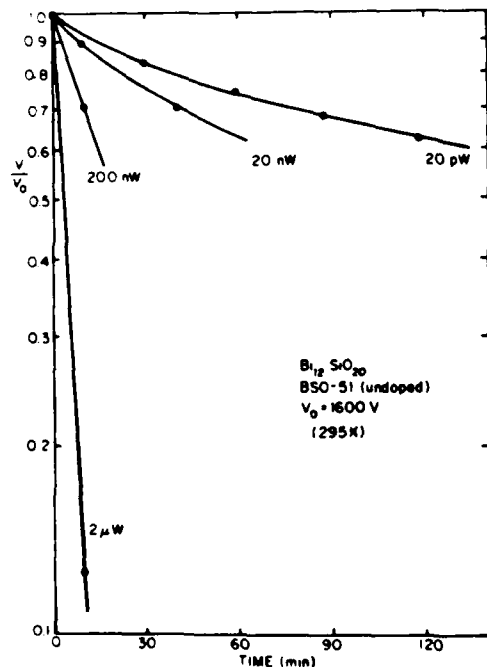


Fig. 4. Characteristic voltage decay curves for undoped sample BSO-51, showing the dependence of the decay time constant on the average incident power level over the time span of the measurement.

nW, and 20 pW. The first three curves show the effects of enhanced photoconductive decay, while the 20 pW decay curve showed little sensitivity to continued decreases in average power level, and hence can be considered representative of the dark resistivity.

It should be noted that bismuth silion oxide exhibits the effects of multiple trap levels [10], and as such the occupancy state of these traps prior to measurement will affect the obtained values of the dark resistivity. It has further been observed, for example, that pre-illumination with IR can significantly alter the trap occupancy and dark conductivity [12]. In order to standardize the initial conditions in a manner representative of the operating conditions of most spatial light modulators and volume holographic optical elements, the samples were pre-illuminated with intense broad band radiation while the counterelectrodes were shorted together. This serves the further purpose of uniformly distributing any accumulated space charge prior to measurement.

Careful sample preparation was found to be critical in obtaining consistently repeatable results. Single crystal samples were oriented by Laue back reflection X-ray diffraction techniques, cut, and polished to better than $\lambda/4$. The samples were then ultrasonically cleaned in semiconductor grade TCE, acetone, and methanol, followed by a thorough rinse in 18 M Ω cm deionized water. Indium tin oxide (ITO) transparent counterelectrodes were deposited by RF magnetron sputtering through pattern-defining masks. Contact to the samples was made through opposing conductive o-ring gaskets, and in some cases by small wires attached to the ITO electrodes with silver paint. Between measurements, samples were kept in a desiccated environment to avoid adsorbed moisture.

4. Resistivity measurements

Four single crystal samples of $\text{Bi}_{12}\text{SiO}_{20}$ were chosen for measurement, representative of both undoped and doped crystals grown from the melt by the Czochralski technique [9]. The four crystal samples and their characteristics are listed in table

Table 1
Summary of $\text{Bi}_{12}\text{SiO}_{20}$ sample characteristics

Sample		Description
BSO-28	Undoped	Nominally pure
BSO-51	Undoped	Nominally pure
BSO-53	Fe doped (5 ppm)	Very low level dopant
BSO-46	Cr doped (180 ppm)	Intermediate level dopant

1. The first two samples, BSO-28 and BSO-51, are representative of undoped single crystals grown from different starting batches of five 9's purity Bi_2O_3 and SiO_2 under similar but not identical growth conditions. Sample BSO-53 was grown from the same starting melt as BSO-51, with the addition of 5 ppm (atomic) iron (from iron oxide) to the melt. Sample BSO-46 was intentionally doped with 180 ppm (atomic) chromium in the melt (from chromium oxide). Using reported values of the segregation coefficients in $\text{Bi}_{12}\text{SiO}_{20}$ (Fe: < 0.05; Cr: 1.8 [13]), this implies crystalline doping densities of approximately 250 ppb or less Fe in BSO-53 and 320 ppm Cr in BSO-46. The optical quality of the samples was determined by careful inspection for strain induced birefringence using polarization microscopy. No growth induced strain was evident in any of the samples.

Characteristic decay curves for each of the four samples are shown in fig. 5. Sample BSO-28 exhibited the shortest time constant to the $1/e$ point (approximately 40 min), and Cr-doped BSO-46 exhibited the longest (in excess of 100 h). As discussed in section 2, note that none of the voltage decay curves shown in fig. 5 can be adequately fit by a straight line as would be expected for a purely exponential decay. This is a clear indication of an inherent nonlinearity in the current-voltage characteristic for this voltage range. However, the calculated variation in differential resistivity along each of the curves amounts to of order a factor of two.

Resistivity values for each of the four representative samples are plotted for comparison in fig. 6 as a function of the average incident power level over the time span of the measurement. These values were calculated from the first two reliable data points following initiation of the voltage decay measurement in each case, and as

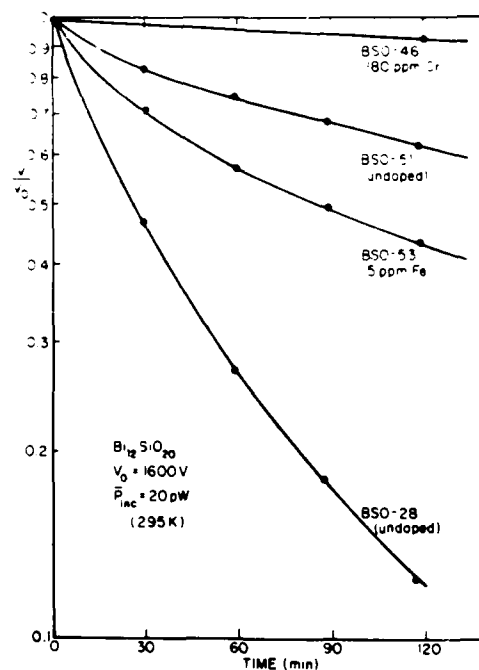


Fig. 5. Characteristic voltage decay curves for all four $\text{Bi}_{12}\text{SiO}_{20}$ samples, all at an average incident power level of 20 pW over the time span of the measurement.

such closely approximate the differential resistivity at or near 1600 V for each sample. Note that each curve is seen to saturate with decreasing average incident power, indicating a valid assignment of the dark resistivity at the lower limiting power level. Furthermore, note the variation in slopes for higher incident power levels, which implies a corresponding variation in sample photoconductivity at the probe wavelength of 6328 Å.

The lowest resistivity sample measured is BSO-28, with a value near $4 \times 10^{14} \Omega \text{ cm}$. The second undoped sample (BSO-51) is considerably more resistive, at $1.5 \times 10^{15} \Omega \text{ cm}$. This variation can be attributed either to differences in starting batch impurity levels or to distinct growth conditions. The addition of only 5 ppm Fe to the melt is seen to lower the resistivity from that of BSO-51 to that of BSO-53 ($9 \times 10^{14} \Omega \text{ cm}$), while the addition of 180 ppm Cr to the melt produced a sample with the highest resistivity observed among our samples to date (BSO-46: $1.9 \times 10^{16} \Omega \text{ cm}$).

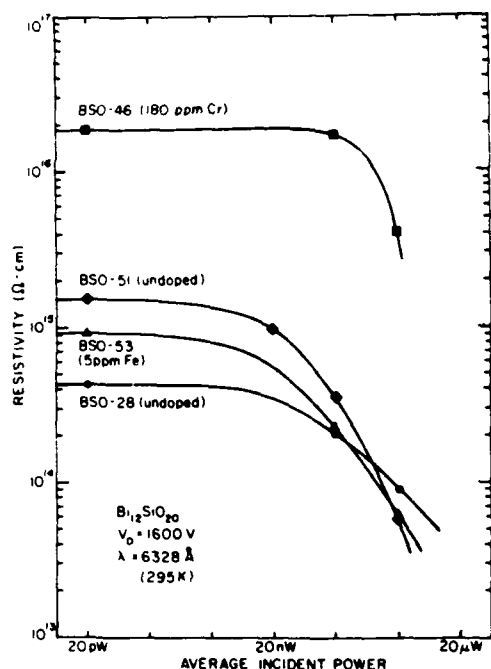


Fig. 6. Measured values of the volume resistivity for all four $\text{Bi}_{12}\text{SiO}_{20}$ samples as a function of average incident power over the time span of the measurement.

It should be noted that the Cr doped sample was quite unusual in that it proved to be quite insensitive to considerable levels of illumination. For example, with 1600 V applied, minutes of exposure to bright, broad band illumination caused negligible voltage decay, as did continual exposure to nominal ambient room illumination levels. The combination of an increased resistivity with a significantly diminished photosensitivity suggests that Cr may form an active recombination center with a large capture cross section in bismuth silicon oxide. Other dopants in $\text{Bi}_{12}\text{SiO}_{20}$ have also been observed to both decrease the photoconductivity in the visible spectrum and change the volume resistivity [14].

5. Discussion and conclusions

Several device implications accrue to the electrooptic measurement of the volume resistivity of

bismuth silicon oxide, particularly with regard to its potential correlation with growth and doping conditions. The attainment of very large resistivity values in certain electro-optic single crystals through either growth modification or dopant incorporation is of considerable importance, as such values imply very large charge storage (frame) times in most device configurations. In addition, the achievement of long dielectric relaxation times implies the added flexibility of decay-free temporal integration processing modes in both optically addressed electrooptic spatial light modulators and photorefractive volume holographic optical elements. In some cases, device structures utilize dielectric blocking layers to prevent longitudinal charge transport through the device. An important example is the Pockels Readout Optical Modulator, which employs vapor deposited parylene organic thin films as high resistivity, high dielectric breakdown strength blocking layers. In such cases, pre-determination of the volume resistivity allows the effects of sub-optimum dielectric blocking layers to be accurately assessed [15–17].

Effects of growth modification and/or dopant incorporation on the photoconductive response at particular wavelengths can also have considerable impact for potential enhancements of device performance. For example, increased photoconductivity at specified (writing) wavelengths optimizes the exposure sensitivity of optically addressed spatial light modulators. On the other hand, decreased photoconductivity at specified (readout) wavelengths reduces the potential for optical damage in single and multiple channel electrooptic modulators, and can also increase the available readout gain for several classes of spatial light modulators.

We have presented herein a technique for accurate and repeatable determination of the volume resistivity (and photoconductivity) of electrooptic single crystals. The technique is quite straightforward to implement, is nondestructive, and allows for in situ variation of parameters such as external illumination, ambient atmosphere, and ambient temperature. In addition, the technique does not require accurate crystallographic orientation or sample alignment for successful implementation.

Acknowledgements

This research was supported in part by the Defense Advanced Research Projects Agency through the Office of Naval Research, the Joint Services Electronics Program, and ITT Corporation. The authors are pleased to acknowledge the crystal growth contributions of Leroy Fisher, and technical assistance from Frank Lum.

References

- [1] A.R. Tanguay, Jr., *Opt. Eng.* 24 (1985) 002.
- [2] Y. Owechko and A.R. Tanguay, Jr., *J. Opt. Soc. Am. A1* (1984) 635.
- [3] Y. Owechko and A.R. Tanguay, Jr., *J. Opt. Soc. Am. A1* (1984) 644.
- [4] A. Marrakchi, A.R. Tanguay, Jr., J. Yu and D. Psaltis, *Opt. Eng.* 24 (1985) 124.
- [5] S. McCahon, S. Kim and A.R. Tanguay, Jr., *J. Opt. Soc. Am. A1* (1984) 1314.
- [6] P. Gunter, *Phys. Rept.* 93 (1982) 199.
- [7] F.M. Smits, *Bell Syst. Tech. J.* 37 (1958) 711.
- [8] L.J. van der Pauw, *Philips Res. Rept.* 13 (1958) 1.
- [9] A.R. Tanguay, Jr., *The Czochralski Growth and Optical Properties of Bismuth Silicon Oxide*, PhD Dissertation, Yale University, New Haven, CT (1977).
- [10] S.L. Hou, R.B. Lauer and R.E. Aldrich, *J. Appl. Phys.* 44 (1973) 2652.
- [11] G.A. Massey, N.G. Eror and G.W. Nelson, *Appl. Opt.* 19 (1980) 1282.
- [12] A.A. Kamshilin and M.G. Miteva, *Opt. Commun.* 36 (1981) 429.
- [13] B.C. Grabmaier and R. Oberschmid, *Phys. Status Solidi* 96 (1986) 199.
- [14] T. Mori, T. Okamoto, and M. Saito, *J. Electron. Mater.* 8 (1979) 261.
- [15] Y. Owechko and A.R. Tanguay, Jr., *J. Opt. Soc. Am.* 71 (1981) 1630.
- [16] Y. Owechko and A.R. Tanguay, Jr., *J. Opt. Soc. Am.* 72 (1982) 1832.
- [17] A.R. Tanguay, Jr. and Y. Owechko, *J. Opt. Soc. Am.* 72 (1982) 1832.

Polarization Properties of Enhanced Self-Diffraction in Sillenite Crystals

ABDELLATIF MARRAKCHI, RICHARD V. JOHNSON, MEMBER, IEEE, AND
ARMAND R. TANGUAY, JR., MEMBER, IEEE

Abstract—Doppler-enhanced self-diffraction in two-beam coupling experiments with sillenite crystals exhibits pronounced optical polarization effects due to the concomitant presence of natural optical activity and electric-field-induced linear birefringence. Coupled wave equations that describe the polarization properties, exclusive of self-diffraction effects, have previously been derived for the two principal crystal orientations most commonly used in photorefractive recording. In this paper, the coupled wave equations are combined with a linearized model of the photorefractive recording process (single trap level, single mobile charge species) to analyze the impact of self-diffraction effects on the polarization state evolution. Numerical solutions of these equations yield optimum configurations for enhanced gain and improved image contrast in two-wave mixing with such materials. In addition, inclusion of optical activity in the model emphasizes the contribution of this effect to the apparent reduction of the effective electrooptic coefficient of bismuth silicon oxide crystals.

I. INTRODUCTION

A MOST interesting class of photorefractive media is that of the sillenite crystals, which includes bismuth silicon oxide ($\text{Bi}_{12}\text{SiO}_{20}$, or BSO), bismuth germanium oxide ($\text{Bi}_{12}\text{GeO}_{20}$, or BGO), and bismuth titanium oxide ($\text{Bi}_{12}\text{TiO}_{20}$, or BTO) [1]. These crystals exhibit high photorefractive sensitivity for volume holographic grating formation [2], fast response, long storage times under dark conditions, and essentially unlimited recyclability. Such crystals are potentially useful for dynamic real-time and time average interferometry [3], nonlinear optical signal processing [4], [5], phase-corrected image propagation through aberrating media [6], optical interconnections [7], spatial optical switching [8], self-pumped laser resonators [9], and image amplification [10].

Light diffraction from volume holograms in sillenite crystals exhibits pronounced polarization properties, encompassing the effects of both natural optical activity and electric-field-induced linear birefringence. A detailed model of these polarization properties is critical for obtaining maximum performance in photorefractive appli-

cations. For example, Herriau *et al.* [11], [12] have demonstrated a technique for significantly improving the holographic image quality by suppressing spurious scattered light. This technique relies on the fact that each of the two holographic writing beams generally evolves into a distinctly different polarization state, so that the two beams can be distinguished by a polarization analyzer.

Numerous studies of the polarization properties have been published for each of the two principal crystallographic orientations (Figs. 1 and 2) in which the presence of optical activity is neglected [13]–[15]. The effects of optical activity have been incorporated by coupled wave equation techniques, as derived by the authors [16], and independently by Vachss and Hesselink [17], which describe transmissive holographic recording geometries in the absence of self-diffraction effects. Mallick *et al.* have derived analytic solutions for these coupled wave equations in the limit in which one of the two recording beams remains undepleted as it propagates, and in the limit of perfect Bragg matching [18]. Kukhtarev *et al.* have derived a set of coupled wave equations which includes both optical activity and self-diffraction effects, and have observed that self-diffraction can significantly modify the polarization states exhibited by the light beams [19]. The detailed analysis presented in [19] emphasized a reflection holographic recording geometry, but the coupled wave equations can easily be recast for a transmissive geometry. The scope of the present paper is to explore the impact of self-diffraction effects on the polarization properties of volume holograms in sillenite crystals recorded in a transmissive geometry (as shown schematically in Fig. 3). Operation deep within the Bragg regime is assumed throughout the following analysis.

Self-diffraction refers to the process whereby the two writing laser beams, which interfere to form the photorefractive grating, diffract from the forming grating, thereby modifying the interference fringe profile deeper within the crystal [20]–[22]. This effect modifies both the modulation depth of the grating and the phase of the fringe system. The result is a temporal and spatial evolution of the grating strength and phase, which eventually stabilizes in the steady-state limit to a spatially-varying grating strength and phase. Representative steady-state grating profiles induced by self-diffraction are shown in Figs. 4 and 5, based upon the model described in the next section.

Manuscript received February 9, 1987; revised September 2, 1987. This work was supported in part by the Defense Advanced Research Projects Agency (through the U.S. Office of Naval Research and the U.S. Air Force Office of Scientific Research) and by the Joint Services Electronics Program.

A. Marrakchi is with Bell Communications Research, Red Bank, NJ 07701.

R. V. Johnson and A. R. Tanguay, Jr. are with the Departments of Electrical Engineering and Materials Science, University of Southern California, Los Angeles, CA 90089.

IEEE Log Number 8717506.

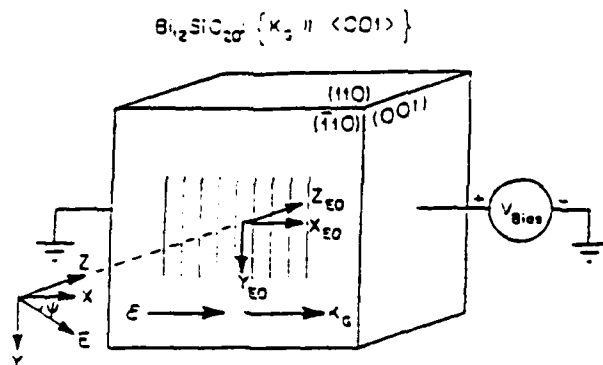


Fig. 1. The $\{K_G \parallel \langle 001 \rangle\}$ crystal orientation of bismuth silicon oxide ($\text{Bi}_{12}\text{SiO}_{20}$, or BSO) for volume holography. The dashed line represents a normal to the entrance face. The x , y , and z coordinate system shown to the left of the crystal is assumed in the coupled wave analysis. The x_{EO} , y_{EO} , and z_{EO} axes shown on the crystal face refer to the principal electrooptic axes induced by the applied bias electric field.

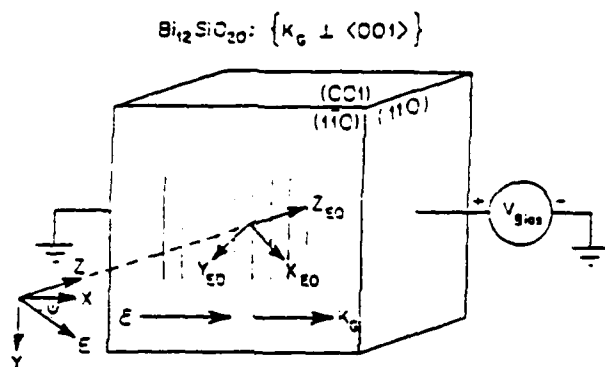


Fig. 2. The $\{K_G \perp \langle 001 \rangle\}$ crystal orientation of $\text{Bi}_{12}\text{SiO}_{20}$ for volume holography.

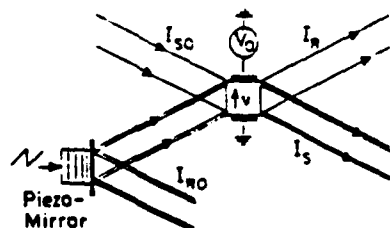


Fig. 3. Optical arrangement for Doppler-enhanced two-beam coupling, in which the temporal frequency of one of the two writing laser beams is slightly shifted by reflection from a piezoelectrically driven mirror, creating a moving grating within the photorefractive medium.

Note the curious structural features of the grating profiles, especially for the $\{K_G \perp \langle 001 \rangle\}$ orientation (defined in Fig. 2). In one case, the grating strength stays essentially uniform for the conditions considered, but the grating phase fronts are corrugated rather than flat. In another case, the grating phase fronts remain flat but the grating strength exhibits a layered structure.

One manifestation of self-diffraction effects can be energy coupling between the two recording laser beams, in which the intensity of one of the beams can be amplified at the expense of the intensity in the second beam. This phenomenon has been extensively investigated in the literature [10], [23]–[25]. Maximum energy coupling oc-

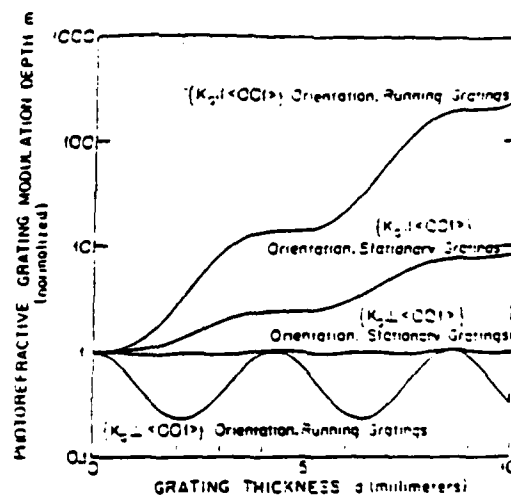


Fig. 4. Self-diffraction induced growth of the space charge field modulation depth as a function of crystal depth for various recording configurations. A bias field of 6 kV/cm and the material parameters for BSO given in Table I are assumed.

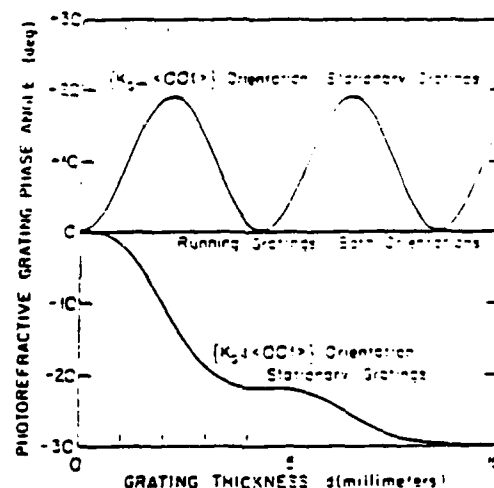


Fig. 5. Variation of the space charge field phase with depth in the crystal resulting from self-diffraction effects, for the same recording conditions assumed in the previous figure. The phase angle plotted here is relative to the phase angle at the entrance face of the crystal.

curs when a 90° phase shift exists between the optical interference pattern and the resulting space charge field. This optimum phase shift occurs naturally when recording stationary interference patterns in the diffusion regime, but typical space charge fields and hence energy coupling effects are low in this regime. The photosensitivity can be significantly enhanced by applying a bias electric field (as shown in Figs. 1–3), but the phase relationship between a stationary light interference pattern and the resulting space charge field is generally not optimum for energy coupling.

At least two techniques have been demonstrated for enhancing the self-diffraction process for energy coupling applications. One technique is to Doppler shift one of the recording laser beams, for example by reflecting this beam from a piezoelectrically-driven constant velocity mirror, as shown in Fig. 3 [10], [23]–[25]. A Doppler shift causes

the light interference pattern to translate with speed v , which modifies the phase between the optical interference and space charge field profiles. One particular grating speed v_{90} exists which asserts the 90° phase shift needed for maximum energy coupling. An additional benefit of the Doppler technique is a significant increase in the magnitude of the space charge field, potentially by as much as an order of magnitude (see Fig. 6, which is described more fully in Section II-B). A second technique for enhancing energy coupling is to record a stationary interference pattern with a rapidly reversing applied bias field [26].

Self-diffraction effects can also be manifested as an alteration of the polarization states of the two writing beams, as shown in Figs. 7-18. These figures compare representative evolutions of polarization angle and ellipticity for the undepleted pump beam and for the signal beams associated with three alternative recording configurations. One recording configuration involves no self-diffraction, but rather presumes a photorefractive grating which is uniform in both amplitude and phase throughout the volume of the hologram. This is the model assumed in our previous polarization properties analysis [16], as well as those of Vachss and Hesselink [17], and Mallick *et al* [18]. The second recording configuration incorporates self-diffraction effects associated with a stationary interference pattern, and the third configuration incorporates self-diffraction effects associated with a moving interference pattern which can enhance energy-coupling in sillenite crystals [10], [23]-[25]. These numerical solutions are reviewed more fully in Section III-A.

Figs. 7-18 demonstrate graphically that self-diffraction effects have a profound impact on the polarization properties of light diffraction from volume holograms in sillenite crystals, especially when techniques such as Doppler-shifting are used to enhance the self-diffraction. A model of light diffraction in sillenite crystals which incorporates self-diffraction effects is detailed in Section II, and the sample solutions are studied in detail in Section III-A. To understand the key implications of the numerical solutions, fundamental features of the diffraction anisotropy are reviewed in Section III-B. Finally, conclusions are drawn from the analysis in Section IV.

II. DESCRIPTION OF THE MODEL AND NUMERICAL METHOD

The complete model of light diffraction in the presence of self-diffraction effects synthesizes two principal components. One component specifies the light diffraction, coupling, and polarization evolution in an electrooptic medium in response to a prescribed quasi-static electric field pattern $E(x, z)$. The second component specifies the photorefractive recording of a space charge field $E_{SC}(x, z)$ in response to a given optical intensity profile formed by the coherent interference of two intersecting laser beams. Each component is explored separately below, and the numerical algorithm for combining them is then described.

A. Coupled Wave Equations for Light Diffraction

The diffraction of light, energy coupling, and evolution of the polarization states for two intersecting laser beams in a volume hologram are all governed by a set of coupled wave equations, one set for each of the two principal crystallographic orientations. These two sets of equations have been derived and discussed in detail previously [16] and are therefore not reproduced herein. The parameters appearing in these equations include natural circular birefringence and electric-field-induced linear birefringence. The electric field profile $E(x, z)$ in steady state is assumed to have the form

$$E(x, z) = E_0 + \frac{1}{2} [E_{SC}(z) \exp(iK_0 x) + \text{complex conjugate}] \quad (1)$$

in which E_0 is a bias electric field applied to the crystal to enhance the photosensitivity (shown in Figs. 1-3), and $E_{SC}(z)$ is the space charge field induced by the photorefractive effect. Higher order spatial harmonics of the space charge field may exist, but are neglected in this analysis because these harmonics are not Bragg-matched to the incident light beams.

Given a particular form for the space charge field $E_{SC}(z)$, the coupled wave equations can be readily integrated to give the light intensities and polarization states at the exit window of the volume hologram. In our previous studies, the space charge field E_{SC} was presumed to be uniform in amplitude and phase throughout the hologram. However, self-diffraction effects are known to induce spatial variations in the amplitude and/or phase, as shown in Figs. 4 and 5, and so we must study the effects of self-diffraction on the photorefractive recording process, considered next.

B. Photorefractive Recording Model

In the photorefractive recording process, a space charge field $E_{SC}(z)$ is induced in response to an optical interference pattern formed between two writing beams with amplitudes R and S (see Fig. 3). A useful parameter for analyzing the photorefractive recording process is the modulation depth m of the optical interference pattern, which is related to the light amplitudes R and S by

$$m = 2R^* \cdot S / (|R|^2 + |S|^2) \quad (2)$$

in which the asterisk denotes complex conjugation, and the dot denotes an inner product. The modulation depth m as defined above is a complex number with phase determined by the relative phase of the optical interference pattern with respect to the coordinate system.

A linearized recording model applies whenever the intensity of one of the recording beams is significantly smaller than that of the second beam, such that $|m| \ll 1$. The stronger beam is generally called the pump beam, with amplitude R , and the weaker beam is called the signal beam, with amplitude S . In the linear recording and

steady-state limits, the space charge field E_{sc} can be written as

$$E_{sc} = mE_{sat} \quad (3)$$

at least to first order in m . The electric field factor E_{sat} remains constant throughout the volume of the hologram, whereas the modulation depth m varies slowly with depth coordinate z due to energy coupling and also due to the nonidentical polarization state evolution of the two writing laser beams.

Detailed expressions for the field term E_{sat} in the linear approximation have been published by Kukhtarev *et al.* for recording stationary gratings using a single mobile charge species and single trap species model [22], and alternatively using two types of photoexcited carriers and two trap species in the charge transport model [19]. In the context of the present analysis, the only effect induced by the additional charge and trap species is to reduce the coupling between the two light beams, and perhaps to modify its phase. For simplicity, we restrict our attention to a single mobile species/single trap species model for the remainder of this article.

Analytic expressions for the field factor E_{sat} for Doppler-enhanced recording, assuming a single mobile charge species and a single trap site, have been derived by Valley [25] and Refregier *et al.* [10]. Fig. 6 shows typical magnitudes of the field factor E_{sat} for Doppler-enhanced recording at optimum speed to maximize the energy coupling, based upon the published analytic solutions, and assuming the material parameters listed in Table I. The grating speed has been reoptimized for each combination of applied bias field and grating spatial frequency.

C. Numerical Algorithm

A typical numerical analysis of self-diffraction effects proceeds as follows. First, one specifies two incident light beams, including their intensities and polarization states, at the entrance window $z = 0$ of the photorefractive crystal. From this, one can determine the modulation depth $m(z = 0)$ of the intensity profile, and hence the space charge field $E_{sc}(z = 0)$ and the birefringent coupling factor (discussed previously in [16]). Next, the coupled wave equations can be integrated directly to determine the light amplitudes and polarization states at a deeper level Δz within the crystal. This defines the space charge field recorded at that depth through the modulation index m . The process is then repeated as needed until the exit window of the photorefractive crystal is reached. The numerical integration can be performed either by a Runge-Kutta algorithm [27], [28] or by the anisotropic optical beam propagation method [29]. It is interesting to note that the computation time with self-diffraction effects included is only minimally longer than the time for the original set of coupled wave equations, at least for a linear model of the photorefractive recording process in steady state. Also, it must be emphasized that the numerical solutions presented in this article apply only to linear recording ($|m| \ll 1$) in the steady-state regime. This concludes the de-

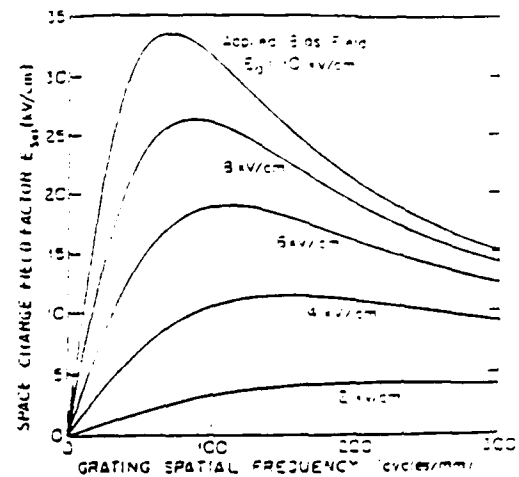


Fig. 6. Magnitude of the space charge field factor E_{sat} as a function of the grating spatial frequency for varying values of the applied bias field E_0 , achieved by the Doppler-enhanced recording technique. The grating velocity has been reoptimized for each point on the curves. See also Valley [25] for similar curves.

TABLE I
PARAMETERS ASSUMED IN THE CALCULATIONS*

Parameter	$\text{Bi}_{12}\text{SiO}_{20}$ at $\lambda = 515 \text{ nm}$ Value	Dimension
refractive index (n_0)	2.615	dimensionless
absorption (α)	—	cm^{-1}
optical rotatory power (ρ)	35.0	degrees/mm
electrooptic coefficient (r_{33})	4.32	pm/V
intensity ratio (I_{50}/I_{90})	1.0×10^{-2}	dimensionless
trap density (N_t)	1.0×10^{16}	cm^{-3}
dielectric constant (ϵ)	56	dimensionless
electron mobility (μ)	0.03	$\text{cm}^2/\text{V s}$
electron recombination rate (γ_R)	2.0×10^{-4}	cm^3/s

*[32], [33].

scription of the model and the corresponding numerical methods.

III. DISCUSSION OF NUMERICAL SOLUTIONS

Having defined the model, let us now turn to a study of representative solutions. Numerical solutions given in Figs. 7–16 are explored in Section III-A. To understand aspects of these solutions, the diffraction anisotropies inherent in sillenite crystals in the absence of optical activity are reviewed in Section III-B. Finally, a heuristic argument is proposed in Section III-C to explain some of the more striking aspects of the numerical solutions.

A. Numerical Solutions

Diffraction and polarization properties of a number of recording configurations have been studied numerically, with results as shown in Figs. 7–16. Configuration parameters include the crystal orientation, the applied bias field, the grating spatial frequency, the assumption of either a stationary or a Doppler-shifted optical interference pattern, and the incorporation of either a self-diffraction or a uniform grating recording model. Two alternative com-

binations of bias field and grating spatial frequency have been considered as representative of different solution limits. One combination consists of a bias electric field of 6 kV/cm and a grating spatial frequency of 300 cycles/mm. This bias field is high enough to enhance significantly the photosensitivity, and the grating spatial frequency is high enough to allow for excellent resolution in any reconstructed volume hologram. In the second combination, the bias electric field is increased to 10 kV/cm, which is about the upper limit for BSO in practical applications. The grating spatial frequency has been reduced from 300 to 70 cycles/mm, even though this reduces the holographic resolution capacity, because at this spatial frequency the resulting space charge field for the case of an optimized velocity moving grating is at a maximum for a 10 kV/cm bias field, as shown in Fig. 6. The second combination of parameters gives almost three times the coupling strength as the first configuration, for the case of Doppler-enhanced recording. The space charge field factors E_{sc} resulting from these two combinations of parameters, assuming the material parameters given in Table I and the analytic solutions published by Refregier *et al.* [10], are shown in Table II. The phase angles quoted in Table II give the phase of the space charge field relative to the optical interference pattern in each case. In addition, an incident intensity ratio (signal to pump beam) of 10^{-3} has been assumed throughout to assure applicability of the undepleted pump approximation. It can be expected that similar effects will occur for higher incident intensity ratios, but modification of the photorefractive charge transport model to incorporate the effects of a depleted pump beam may then be necessary.

Consider first the inhomogeneities induced in the modulation depth m by self-diffraction, which have been calculated for the 6 kV/cm bias field, 300 cycles/mm grating spatial frequency recording combination, with results as shown in Figs. 4 and 5. Fig. 4 shows the photorefractive grating modulation depths as a function of grating thickness calculated for both orientations, as well as for both running and stationary gratings. For the case of stationary gratings, self-diffraction yields significant oscillatory behavior in the $\{K_G \parallel \langle 001 \rangle\}$ orientation, but only minor perturbations in the $\{K_G \perp \langle 001 \rangle\}$ orientation. On the other hand, in the case of running gratings, both orientations yield striking modulation effects. In particular, the $\{K_G \perp \langle 001 \rangle\}$ configuration exhibits a periodic layered structure in the modulation depth, with regions at particular grating thicknesses that approach zero modulation. The photorefractive grating phase angle is plotted for the same set of cases in Fig. 5. For both orientations, the phase is fixed throughout in the case of running gratings at the optimum velocity. For stationary gratings, the $\{K_G \perp \langle 001 \rangle\}$ orientation exhibits an oscillatory phase behavior throughout the depth, while the $\{K_G \parallel \langle 001 \rangle\}$ orientation exhibits a nonlinear warping of the phase fronts with an additional component of periodic behavior. Note then that the $\{K_G \perp \langle 001 \rangle\}$ orientation in the case of stationary gratings has essentially constant modulation but

TABLE II
SPACE CHARGE FIELD FACTORS E_{sc} ASSUMED IN THE CALCULATIONS

Bias field of 6 kV/cm and grating spatial frequency of 300 cycles/mm	
Stationary gratings:	$E_{sc} = 5.5 \text{ kV/cm} \pm 23^\circ$
Doppler-enhanced:	$E_{sc} = 12.3 \text{ kV/cm} \pm 90^\circ$
Bias field of 10 kV/cm and grating spatial frequency of 70 cycles/mm	
Stationary gratings:	$E_{sc} = 3.9 \text{ kV/cm} \pm 3^\circ$
Doppler-enhanced:	$E_{sc} = 33.5 \text{ kV/cm} \pm 90^\circ$

oscillatory phase, while the same orientation in the case of running gratings has constant phase but oscillatory modulation.

Similar profiles have been calculated for the 10 kV/cm, 70 cycle/mm combination, with results almost exactly like those shown in Figs. 4 and 5. The phase modulation is almost identical to that shown in Fig. 5, with only a very slight increase for the larger coupling case. The amplitude growth has the same general form as shown in Fig. 4, but spans a larger range of values.

Next consider the polarization state evolution results shown in Figs. 7-14. The advance of the polarization angle for the $\{K_G \parallel \langle 001 \rangle\}$ orientation, shown in Figs. 7 and 9, predominantly follows a linear trend defined by the magnitude of the optical activity, with a spatially periodic oscillation about this linear progression. The periodic oscillation is induced by coupling of the signal beam to the pump beam, which undergoes its own spatially periodic evolution in polarization state. In the absence of an applied bias field, this spatial period is defined by $d = 180^\circ/\rho$, in which ρ is the optical rotatory power. In the presence of a bias-field-induced linear birefringence C_{11} , the spatial period is somewhat smaller, and is defined by $d = 180^\circ/(\rho^2 + C_{11}^2)^{1/2}$. The periodic perturbation is strongest for Doppler-enhanced self-diffraction, and weakest for the uniform space charge grating model. The evolution of the ellipticity for the $\{K_G \parallel \langle 001 \rangle\}$ orientation, shown in Figs. 8 and 10, also exhibits the same periodicity. (Ellipticity is defined as the ratio of the minor to the major axes of the polarization ellipse, with opposite signs for right-handed and left-handed electric vector rotation.) The ellipticity for the Doppler-enhanced self-diffraction model of the signal beam deviates most strongly from that of the pump beam.

The $\{K_G \perp \langle 001 \rangle\}$ orientation exhibits even more striking modulation of the signal beam polarization state by self-diffraction effects, at least for Doppler-enhanced coupling, as shown in Figs. 11-14. Note the strong variations in the polarization angle for the Doppler-enhanced self-diffracted signal, as shown in Figs. 11 and 13. The stationary grating model with self-diffraction exhibits a polarization response which is much closer to that shown by the uniform grating model. The ellipticity evolution for the Doppler-enhanced signal beam also differs remarkably from that of the stationary grating and uniform grating models, as seen in Figs. 12 and 14.

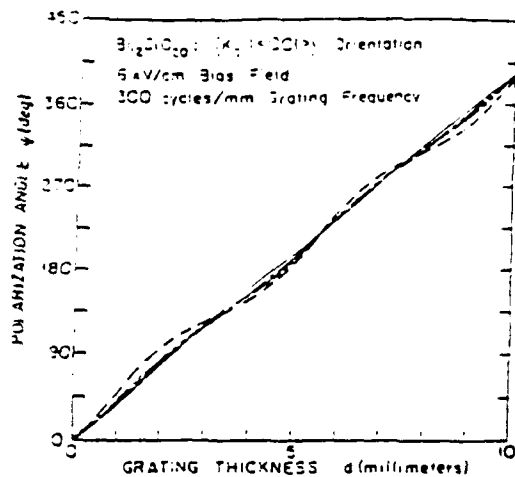


Fig. 7. Evolution of the polarization angle (major axis orientation) for diffraction in the $\{K_C \parallel \langle 001 \rangle\}$ orientation for the undepleted pump beam (solid line), signal beam with Doppler-enhanced self-diffraction at v_m (dashed line), signal beam with a stationary grating and self-diffraction (single dashed line), and signal beam without self-diffraction assuming a phase grating uniform in magnitude and phase throughout the holographic volume (double dashed line). The pump and signal beams are assumed to enter the crystal with identical linear polarization states, with polarization angle parallel to the grating wave vector and the applied bias field.

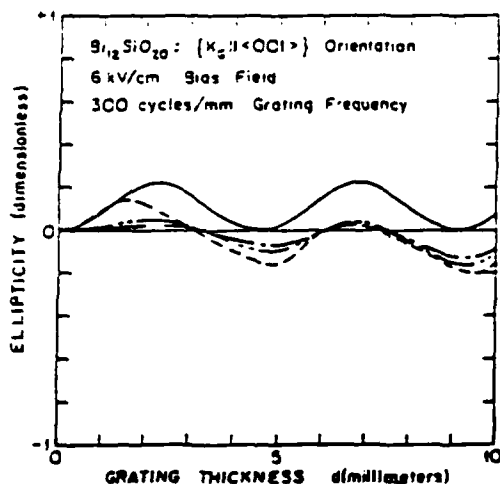


Fig. 8. Evolution of the ellipticity for diffraction in the $\{K_C \parallel \langle 001 \rangle\}$ orientation for the same recording conditions considered in Fig. 7.

Figs. 15 and 16 show the degradation in energy-coupling gain with increasing amounts of optical rotation ρd in the diffusion and drift regimes, respectively, for the two principal crystallographic orientations. The results shown in Fig. 15 are derived from analytic solutions presented by Foote and Hall [30], while the results shown in Fig. 16 are numerically derived from the coupled wave equations. In both figures, the pump and signal beams are assumed to enter the photorefractive medium with identical linear polarization states. In each case, the incident polarization angle is varied from 0 to 180° to find the minimum and the maximum coupling gain; the optimization of the polarization angle is repeated for each value of ρd . These two figures were derived by keeping the op-

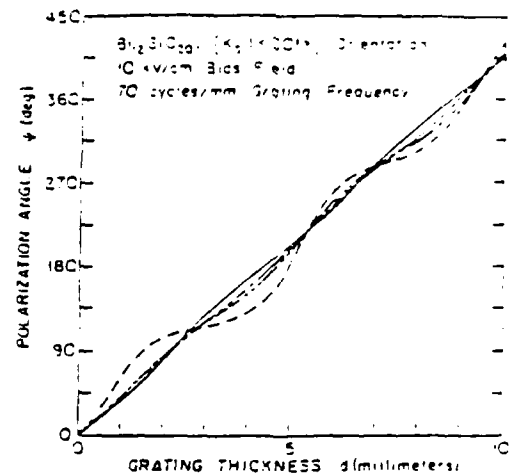


Fig. 9. Evolution of the polarization angle for diffraction in the $\{K_C \parallel \langle 001 \rangle\}$ orientation, as shown in Fig. 7, but for a higher bias field of 10 kV/cm and a lower grating spatial frequency of 70 cycles/mm, which combine to give almost a threshold increase in the coupling strength.

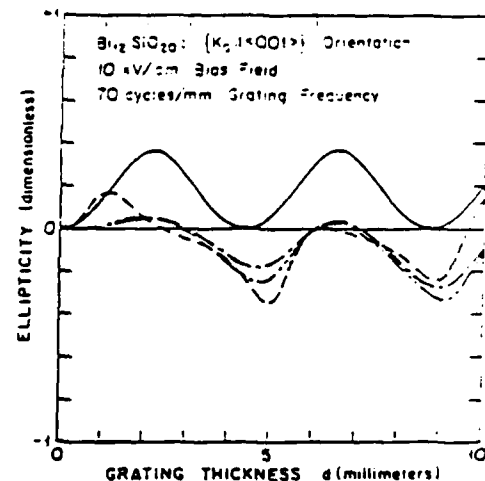


Fig. 10. Evolution of the ellipticity for diffraction in the $\{K_C \parallel \langle 001 \rangle\}$ orientation for the same recording conditions considered in Fig. 9.

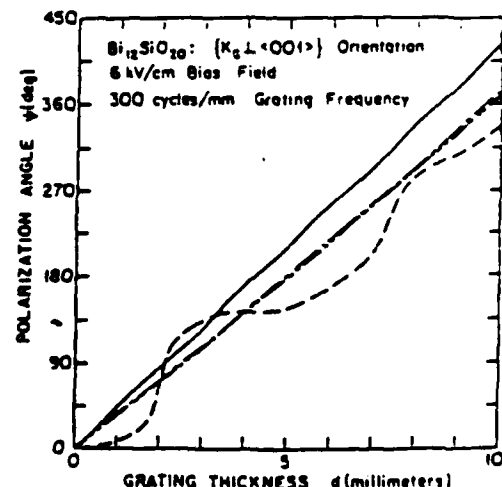


Fig. 11. Evolution of the polarization angle for recording conditions considered in Fig. 7, but for the orthogonal $\{K_C \perp \langle 001 \rangle\}$ recording orientation.

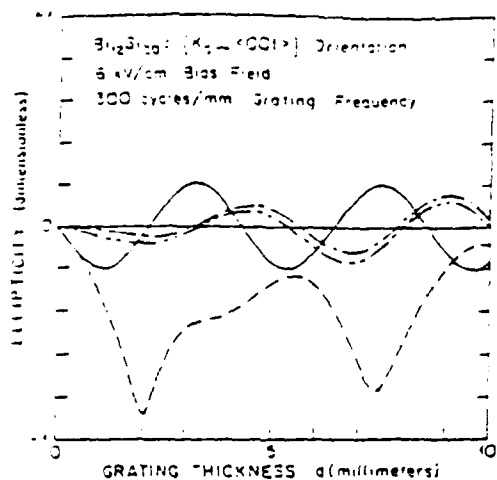


Fig. 12. Evolution of the ellipticity for recording conditions considered in Fig. 11.

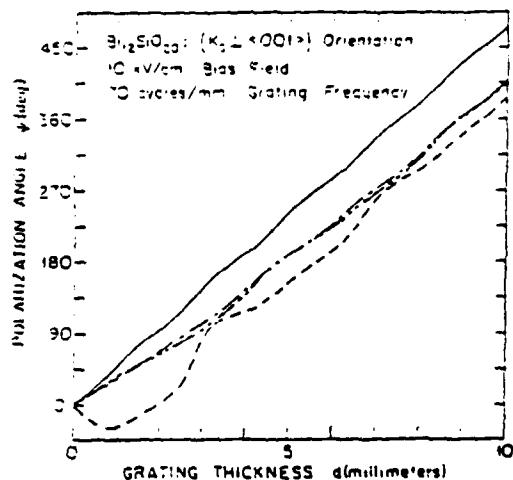


Fig. 13. Evolution of the polarization angle for recording conditions considered in Fig. 9, but for the $\{K_2 \perp \langle 001 \rangle\}$ orientation.

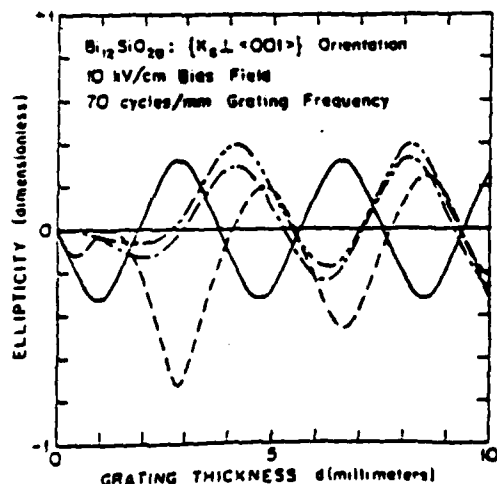


Fig. 14. Evolution of the ellipticity for recording conditions considered in Fig. 13.

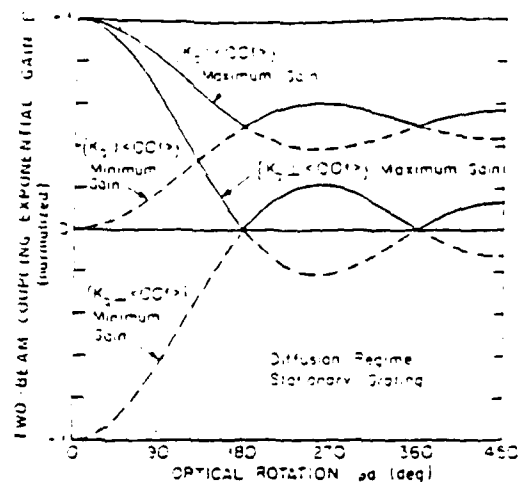


Fig. 15. Degradation of energy coupling gain Γ as defined in (30) with increasing levels of optical rotation ρd , in which ρ is the optical rotatory power and d is the crystal thickness, for the two principal recording orientations. These curves are based upon analytic solutions derived by Foote and Hail for diffusion regime recording [30].

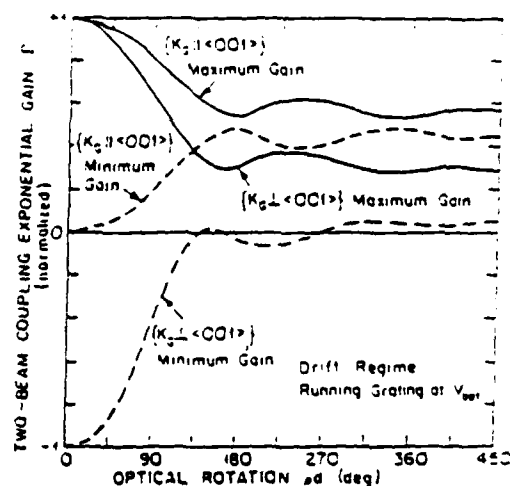


Fig. 16. Degradation of energy coupling with increasing levels of optical rotation ρd for Doppler-enhanced drift-regime recording. Recording conditions are identical to those assumed in Figs. 7 and 11.

tical rotatory power ρ fixed at $38.6^\circ/\text{mm}$ and varying the crystal thickness d , but very similar curves have been obtained by keeping the thickness d constant and varying the rotatory power ρ . These curves have important implications for holographic experiments which measure the effective electrooptic coefficient. These measurements are difficult to interpret correctly for a number of reasons, one being the intricate polarization properties of volume gratings in sillenite crystals. Figs. 15 and 16 indicate that a 50 percent reduction in coupling gain can easily be accounted for because of degradation induced by optical activity.

We have also explored the impact on energy coupling gain of using nonidentically polarized incident pump and signal beams, and of using incident beams with other than linearly polarized states. A modest increase in the gain of

order 10 percent has been obtained for particular configurations.

A heuristic argument is presented in Section III-2 to explain some of the behavior shown in Figs. 7-16. First, however, we need to review the fundamental anisotropies in the diffraction process for the two principal orientations of sillenite crystals typically employed for volume holography.

3. Inherent Diffraction Anisotropies

Much of the interesting polarization evolution behavior exhibited in Figs. 7-16 derives from the anisotropies inherent in the light diffraction process, independent of optical activity. When the optical activity terms are set equal to zero, the sets of coupled wave equations with self-diffraction effects included can be completely decoupled into orthogonal polarization eigenmodes, and analytic solutions can be obtained for each eigenmode. The eigenaxes are different for the two principal recording configurations, and are shown by the (X_{EO}, Y_{EO}, Z_{EO}) axes in Figs. 1 and 2.

For example, the $\{K_G \parallel \langle 001 \rangle\}$ orientation shown in Fig. 1 has eigenaxes X_{EO} and Y_{EO} parallel and perpendicular to the grating wave vector and the applied electric field, respectively. No energy coupling is exhibited for light polarized along the X_{EO} eigenaxis. Instead, maximum coupling occurs for light polarized along the orthogonal axis, Y_{EO} in Fig. 1. This coupling can be either positive or negative, depending upon the direction of the applied electric field. We assume throughout this paper that the field direction is oriented for positive coupling.

The $\{K_G \perp \langle 001 \rangle\}$ orientation has its two eigenaxes X_{EO} and Y_{EO} oriented at 45° with respect to the grating wave vector and the applied bias field, as shown in Fig. 2. Positive energy coupling gain is exhibited along one of these eigenaxes, and negative energy coupling gain is exhibited along the orthogonal axis, although the identification of which axis exhibits positive gain can be altered by reversing the direction of the applied bias field. This can be understood with reference to the distortion of the index ellipsoid imposed by a uniform bias electric field. In the absence of any electric field, the index is isotropic, so that the constant index surface is a sphere. A uniform bias electric field distorts this index surface into an ellipsoid with principal axes aligned along the (X_{EO}, Y_{EO}, Z_{EO}) eigenaxes shown in Fig. 2. The index of refraction is increased by a small amount Δn along one of the axes, say the X_{EO} axis, while it is decreased along the Y_{EO} axis. Now consider a sinusoidal space charge field, which induces a corresponding sinusoidal phase grating for light polarized along the X_{EO} eigenaxis, and a similar phase grating for light polarized along the Y_{EO} eigenaxis. Note, however, that the grating induced along the Y_{EO} eigenaxis is 180° out of phase with respect to the grating induced along the X_{EO} eigenaxis, because the index shift along the one axis is equal and opposite to the index shift along the orthogonal axis. Therefore, if one of the two gratings is optimally phased for positive energy coupling, then the

second grating must be phased for negative energy coupling.

The $\{K_G \perp \langle 001 \rangle\}$ exhibits even more interesting polarization behavior for light linearly polarized along directions bisecting the eigenaxes, i.e., either parallel or perpendicular to the grating wavevector and the applied bias field. For light so polarized, the volume grating exhibits birefringent diffraction properties, also known in the literature as anisotropic Bragg diffraction, wherein the diffracted light has polarization orthogonal to the incident light [29]. This birefringent mode is exhibited only by the $\{K_G \perp \langle 001 \rangle\}$ orientation, not by the $\{K_G \parallel \langle 001 \rangle\}$ orientation, and has important implications in the polarization properties in the presence of self-diffraction, as explained next.

C. Comparison of the Two Principal Orientations

A critical insight into the light diffraction process in the presence of self-diffraction is that the space charge grating can only be recorded where the pump and signal beams share the same polarization state, not when they have orthogonal states. Therefore, an interesting aspect to study in the numerical solutions is the comparative content of signal light intensity with polarization state parallel to and orthogonal to that of the local pump beam. Note the use of the term "local"—the polarization state of the pump beam continuously evolves as it propagates through the crystal, so that the signal beam polarization decomposition of interest similarly evolves.

Representative numerical solutions for these two polarization components are shown in Fig. 17 for the $\{K_G \parallel \langle 001 \rangle\}$ orientation and in Fig. 18 for the $\{K_G \perp \langle 001 \rangle\}$ orientation. The two polarization components in these figures are labeled "coherent" and "incoherent," referring to the components with polarization states identical to and orthogonal to the local pump beam, respectively. Note that the coherent component dominates throughout the propagation path for the $\{K_G \parallel \langle 001 \rangle\}$ orientation, as shown in Fig. 17. Recall that this orientation exhibited no birefringent diffraction modes in the absence of optical activity. Addition of optical activity introduces only a small amount of birefringent diffraction. Compare this with the response for the $\{K_G \perp \langle 001 \rangle\}$ orientation shown in Fig. 18. Assuming that the incident beams are polarized to maximize the energy coupling (which implies different incident polarizations for the two configurations), the growth of the coherent component is identical for both recording orientations, at least for the first half millimeter or so of propagation. For the $\{K_G \perp \langle 001 \rangle\}$ orientation, the optical activity rotates the pump polarization angle so that after the first millimeter or so of propagation significant birefringent diffraction is encountered. At this point, further growth of the coherent signal light is halted, while the incoherent component continues to grow. One consequence is that further growth of the space charge field which supports the diffraction is similarly halted. After more propagation, say at about the

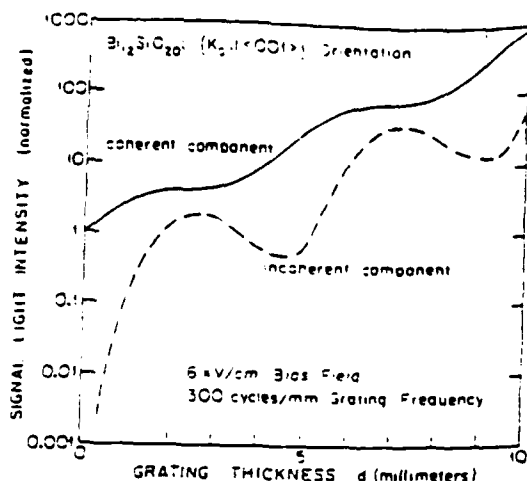


Fig. 17. Growth of the incoherent and coherent signal beam components for self-diffraction in the $\{K_G \parallel \langle 001 \rangle\}$ orientation. The coherent component is defined as that which has a polarization state identical to the local pump beam, and hence can interfere coherently with the pump beam to form a fringe pattern and photorefractively induce a space charge field. The incoherent component has an orthogonal polarization state, and hence can only erase the space charge field. Recording conditions are identical to those assumed in Figs. 7 and 8, except that the incident polarization angle is 90° with respect to the grating wave vector to maximize the energy coupling.

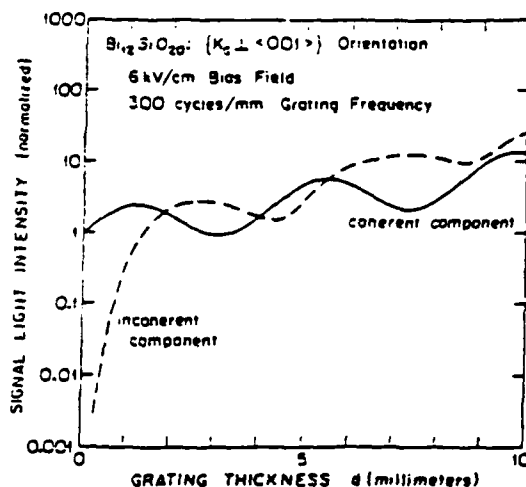


Fig. 18. Growth of the incoherent and coherent signal beam components for self-diffraction in the $\{K_G \perp \langle 001 \rangle\}$ orientation. Recording conditions are identical to those assumed in Figs. 11 and 12, except that the incident polarization angle is -45° with respect to the grating wave vector to maximize the energy coupling.

2 mm depth, the optical activity has rotated the pump beam polarization enough so that now significant negative energy coupling is experienced, and the coherent signal component loses intensity rather than gains in this region. No comparable birefringent modes or negative gain axes exist for the $\{K_G \parallel \langle 001 \rangle\}$ orientation. This explains why the $\{K_G \parallel \langle 001 \rangle\}$ orientation exhibits superior energy-coupling gain compared with the $\{K_G \perp \langle 001 \rangle\}$ orientation, as has been reported in the literature [31].

IV. CONCLUSIONS

A method for calculating the polarization state evolution in photorefractive sillenite crystals in the presence of

self-diffraction effects has been described, and representative solutions have been studied. Self-diffraction effects are shown to alter the diffraction process remarkably. For the $\{K_G \parallel \langle 001 \rangle\}$ orientation, self-diffraction induces significant energy coupling gain; for the $\{K_G \perp \langle 001 \rangle\}$ orientation, self-diffraction significantly modifies the polarization states, especially for Doppler-enhanced recording. Striking spatial inhomogeneities are induced in the space charge field, especially for the $\{K_G \perp \langle 001 \rangle\}$ orientation, leading to striations in the space charge field along the direction of propagation for the case of Doppler-enhanced recording, and essentially constant amplitude but corrugated phase fronts for stationary recording. It is interesting to note that the complexity and time to compute for the numerical model are only modestly increased with the addition of self-diffraction, at least for the linear photorefractive recording model in steady state used herein. The diffraction properties of the two principal recording configurations prove to be distinctly different, and in part this difference is attributed to fundamental anisotropies inherent in the diffraction process. Finally, the degradation in energy coupling gain with increasing amounts of optical rotation ρd has been calculated for a few representative cases. This degradation can amount to of order 50 percent and must be considered in any holographically-based measurements of effective electrooptic coefficients.

REFERENCES

- [1] B. Aurivillius and L. G. Sillen, "Polymorphism of bismuth trioxide," *Nature*, vol. 155, pp. 305-306, 1945.
- [2] J. P. Huignard and F. Micheron, "High-sensitivity read-write volume holographic storage in $\text{Bi}_{12}\text{SiO}_{20}$ and $\text{Bi}_{12}\text{GeO}_{20}$ crystals," *Appl. Phys. Lett.*, vol. 29, pp. 591-593, 1976.
- [3] J. P. Huignard, J. P. Hernau, and T. Valentin, "Time average holographic interferometry with photoconductive electrooptic $\text{Bi}_{12}\text{SiO}_{20}$ crystals," *Appl. Opt.*, vol. 16, no. 11, pp. 2796-2798, 1977.
- [4] V. Markov, S. Odulov, and M. Soskin, "Dynamic holography and optical image processing," *Opt. Laser Technol.*, vol. 11, pp. 95-99, 1979.
- [5] J. O. White and A. Yariv, "Real-time image processing via four-wave mixing in a photorefractive medium," *Appl. Phys. Lett.*, vol. 37, pp. 5-7, 1980.
- [6] J. P. Huignard, J. P. Hernau, P. Aubourg, and E. Spitz, "Phase-conjugate wavefront generation via real-time holography in $\text{Bi}_{12}\text{SiO}_{20}$ crystals," *Opt. Lett.*, vol. 4, pp. 21-23, 1979.
- [7] J. W. Goodman, F. J. Leiberger, S.-Y. Kung, and R. A. Athale, "Optical interconnections for VLSI systems," *Proc. IEEE*, vol. 72, pp. 850-866, 1984.
- [8] J. P. Hernau, A. Delboulbe, B. Loiseau, and J. P. Huignard, "Commutateur optique bidimensionnel par réseaux holographiques photoinduits," *J. Opt. (Paris, France)*, vol. 15, pp. 314-318, 1984.
- [9] H. Rajbenbach and J. P. Huignard, "Self-induced coherent oscillations with photorefractive $\text{Bi}_{12}\text{SiO}_{20}$ amplifier," *Opt. Lett.*, vol. 10, pp. 137-139, 1985.
- [10] Ph. Refregier, L. Solymar, H. Rajbenbach, and J. P. Huignard, "Two-beam coupling in photorefractive $\text{Bi}_{12}\text{SiO}_{20}$ crystals with moving grating: Theory and experiments," *J. Appl. Phys.*, vol. 58, pp. 45-57, 1985.
- [11] J. P. Hernau, J. P. Huignard, and P. Aubourg, "Some polarization properties of volume holograms in $\text{Bi}_{12}\text{SiO}_{20}$ crystals and applications," *Appl. Opt.*, vol. 17, no. 12, pp. 1851-1852, 1978.
- [12] J. P. Hernau, J. P. Huignard, A. G. Apostolidis, and S. Mallick, "Polarization properties in two-wave mixing with moving grating in photorefractive BSO crystals. Application to dynamic interferometry," *Opt. Commun.*, vol. 56, no. 3, pp. 141-144, 1985.
- [13] M. P. Petrov, S. V. Mindonov, S. I. Stepanov, and V. V. Kulikov,

- "Light diffraction and nonlinear image processing in electrooptic $\text{Bi}_{12}\text{SiO}_{20}$ crystals," *Opt. Commun.*, vol. 31, no. 3, pp. 301-305, 1979.
- [14] M. P. Petrov, T. G. Penecheva, and S. I. Stepanov, "Light diffraction from volume phase holograms in electrooptic photorefractive crystals," *J. Opt. Paris, France*, vol. 12, no. 5, pp. 187-192, 1981.
- [15] S. I. Stepanov and M. P. Petrov, "Photorefractive crystals of the $\text{Bi}_{12}\text{SiO}_{20}$ type for interferometry, wavefront conjugation and processing of non-stationary images," *Opt. Acta*, vol. 31, no. 12, pp. 1335-1343, 1984.
- [16] A. Marrakchi, R. V. Johnson, and A. R. Tanguay, Jr., "Polarization properties of photorefractive diffraction in electrooptic and optically active sillenite crystals: Bragg regime," *J. Opt. Soc. Amer. B*, vol. 3, no. 2, pp. 321-336, 1986.
- [17] F. Vachss and L. Hesselink, "Holographic beam coupling in anisotropic photorefractive media," *J. Opt. Soc. Amer. A*, vol. 4, no. 2, pp. 325-339, 1987.
- [18] S. Maitlick, D. Rouede, and A. G. Apostolidis, "Efficiency and polarization characteristics of photorefractive diffraction in a $\text{Bi}_{12}\text{SiO}_{20}$ crystal," *J. Opt. Soc. Amer. B*, vol. 4, no. 3, pp. 1247-1259, 1987.
- [19] N. V. Kukhtarev, G. E. Dovgaleenko, and V. N. Starkov, "Influence of the optical activity on hologram formation in photorefractive crystals," *Appl. Phys.*, vol. A35, pp. 227-230, 1984.
- [20] N. V. Kukhtarev, V. B. Markov, S. G. Odulov, M. S. Soskin, and V. L. Vinetski, "Holography storage in electrooptic crystals," *Ferroelectr.*, vol. 22, pp. 949-964, 1979.
- [21] V. Kondilenko, V. Markov, S. Odulov, and M. Soskin, "Diffraction of coupled waves and determination of phase mismatch between holographic grating and fringe pattern," *Opt. Acta*, vol. 26, pp. 238-251, 1979.
- [22] N. V. Kukhtarev, V. B. Markov, and S. G. Odulov, "Nonstationary energy exchange during interaction between two light beams in electro-optical crystals," *Sov. Phys. Tech. Phys.*, vol. 25, pp. 1109-1114, 1980.
- [23] J. P. Huignard and A. Marrakchi, "Coherent signal beam amplification in two-wave mixing experiments with photorefractive $\text{Bi}_{12}\text{SiO}_{20}$ crystals," *Opt. Commun.*, vol. 38, no. 4, pp. 249-254, 1981.
- [24] H. Rajbenbach, J. P. Huignard, and B. Loiseaux, "Spatial frequency dependence of the energy transfer in two-wave mixing experiments with BSO crystals," *Opt. Commun.*, vol. 48, pp. 247-252, 1983.
- [25] G. C. Valley, "Two-wave mixing with an applied field and a moving grating," *J. Opt. Soc. Amer. B*, vol. 1, pp. 868-873, 1984.
- [26] S. I. Stepanov and M. P. Petrov, "Efficient unstationary holographic recording in photorefractive crystals under an external alternating electric field," *Opt. Commun.*, vol. 53, pp. 292-295, 1985.
- [27] B. Carnahan, H. A. Luther, and J. O. Wilkes, *Applied Numerical Methods*. New York: Wiley, 1969, ch. 6, sect. 6.5.
- [28] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes: The Art of Scientific Computing*. New York: Cambridge, 1986, ch. 15, sect. 15.1.
- [29] R. V. Johnson and A. R. Tanguay, Jr., "Optical beam propagation method for birefringent phase grating diffraction," *Opt. Eng.*, vol. 25, pp. 235-249, 1986.
- [30] P. D. Foote and T. J. Hall, "Influence of optical activity on two beam coupling constants in photorefractive $\text{Bi}_{12}\text{SiO}_{20}$," *Opt. Commun.*, vol. 57, pp. 201-206, 1986.
- [31] A. Marrakchi, J. P. Huignard, and P. Gester, "Diffraction efficiency and energy transfer in two-wave mixing experiments with $\text{Bi}_{12}\text{SiO}_{20}$ crystals," *Appl. Phys.*, vol. 24, pp. 131-138, 1981.
- [32] A. R. Tanguay, Jr., "The Czochralski growth and optical properties of bismuth silicon oxide," Yale Univ., New Haven, CT, Ph.D. dissertation, 1977.
- [33] G. C. Valley and M. B. Klein, "Optimal properties of photorefractive materials for optical data processing," *Opt. Eng.*, vol. 22, pp. 704-711, 1983.



Abdellatif Marrakchi was born in Tlemcen, Morocco, on June 25, 1955. He received the M.S. and Engineer's degrees and a Doctorate (Ph.D.) Cycle in electrical engineering with majors in optics and instrumentation, all from the University Pierre and Marie Curie, Paris, France, in 1973, 1979, and 1981, respectively. His thesis work at Thomson-CSF Research Laboratories, Orsay, France, and the Swiss Federal Institute of Technology, Zurich, Switzerland, involved holographic optical elements for head-up displays,

real-time holography, phase conjugation, and two-beam coupling in photorefractive materials for nondestructive testing. He also received the M.S. and Ph.D. degrees in electrical engineering, with a major in electrophysics and optics, both from the University of Southern California, Los Angeles, CA, in 1983 and 1986, respectively. His dissertation work involved the study of the polarization properties of beam coupling in sillenite crystals and the application of real-time holography to spatial light modulation.

In 1986, he joined Bell Communications Research, Red Bank, NJ, where his primary interests are the analysis of current technologies with potential application in the field of photonic switching, and the optical implementation of neural networks. He has published numerous papers in the field of photorefractive materials, and holds two patents.

Dr. Marrakchi is a member of the Optical Society of America and SPIE.



Richard V. Johnson (M'77) was born in Los Angeles, CA, on January 3, 1945. He received the B.S., M.A., and Ph.D. degrees from the University of Southern California, Los Angeles, in 1965, 1967, and 1973, respectively, with majors in physics and minors in mathematics. His Ph.D. dissertation was on computer studies of relaxation oscillations in stimulated Raman and Brillouin scattering.

In 1973, he joined the Xerox Corporation, where he worked on laser scanning components and systems for electronic printing applications, optical data storage technologies, the characterization of photoresist for printed circuit board manufacturing, and magnetography. In 1984, he joined the Optical Materials and Devices Laboratory at the University of Southern California, where he has been active in modeling the polarization properties of light diffraction from volume holograms in sillenite crystals, computer studies of photorefractive recording processes, and characterizing novel grating structures. He has published numerous articles and holds seven patents.

Dr. Johnson is a member of the Optical Society of America and SPIE.

Armand R. Tanguay, Jr. (S'66-M'67), for a photograph and biography, see this issue, p. 2103.

Stratified volume holographic optical elements

R. V. Johnson and A. R. Tanguay, Jr.

Optical Materials and Devices Laboratory, and Center for Photonic Technology, University of Southern California, University Park, MC-0483, Los Angeles, California 90089-0483

Received October 20, 1987; accepted December 21, 1987

A computational algorithm for analyzing diffraction properties of optical devices, the optical beam propagation method, has suggested a new class of devices by which Bragg regime (thick grating) response can be obtained from a spaced sequence of thin grating layers. Such stratified volume holographic optical elements (SVHOE's) can emulate distributed volume gratings in terms of diffraction efficiency and angular selectivity and in addition possess periodic diffraction properties that might serve, for example, as interconnections for optical cellular logic arrays. SVHOE's also offer a unique capability for altering the device diffraction response on a layer-by-layer basis, allowing for control of both the diffraction peak width and the angular separation of adjacent peaks.

Holographic optical elements are of fundamental importance to a number of applications in optical information processing and computing, including optical interconnections, content-addressable memories, and various linear and nonlinear signal processing configurations. Numerical analyses of typical volume holographic structures are critical for characterizing and optimizing such optical elements. One of the most flexible and broadly applicable numerical tools for studying the diffraction behavior of volume holograms is the optical beam propagation method (BPM).^{1,2} In this method, the distributed optical inhomogeneities that characterize a typical hologram are approximated by a discrete sequence of physically and mathematically simplified elements: infinitesimally thin phase and/or polarization modulation layers, interleaved with optically homogeneous layers of finite thickness. By approximating the distributed grating with a sufficiently large number of these discrete elements, the resulting numerical model of the grating can be brought arbitrarily close in its response to that of the desired distributed bulk grating.

The BPM concept of separating the modulation process from the diffraction process in turn suggests a new class of optical devices: the stratified volume holographic optical element, or SVHOE,³ as shown in Fig. 1. The SVHOE device structure consists of a sequence of thin photosensitive holographic recording layers that perform the optical modulation function, interleaved with optically passive buffer layers, i.e., layers that impress no modulation on the light beam but rather allow the diffraction processes necessary for thick grating response to occur. A grating recorded in any individual modulation layer would necessarily exhibit Raman-Nath characteristics because each recording layer is quite thin. However, a surprisingly small number of recording layers, each spaced from its neighbors by a passive layer of appropriate thickness, can exhibit pronounced Bragg regime (thick grating) response. In addition, SVHOE structures exhibit striking diffraction characteristics not observed in bulk gratings, as demonstrated below.

The SVHOE structure is of particular interest for materials that can be produced in only finite thickness layers but that nonetheless exhibit either strong mod-

ulation effects or fast response times. Examples of such strongly modulating but limited-thickness media include III-V and II-VI compound semiconductor multiple quantum wells and superlattices, and ordered noncentrosymmetric polymer layers characterized by significant electro-optic coefficients.

For simplicity, the following discussion emphasizes the angular alignment sensitivity of the grating structure; alternative performance measures, such as sensitivity to the wavelength of the readout light, correspond closely to this angular alignment sensitivity metric. To measure the angular alignment response, a grating structure is illuminated with a well-collimated single-wavelength laser beam. That fraction of the incident light intensity diffracted into an adjacent diffraction order, such as the +1 order, is measured with a photodetector as a function of the tilt angle of the grating.

The angular response for a representative distributed bulk grating is shown in the top half of Fig. 2 to serve as a reference for comparison with sample response curves for typical SVHOE structures. Such a uniform volume (or distributed bulk) grating has an alignment sensitivity curve characterized by essentially a sinc² profile. The angular width of this response

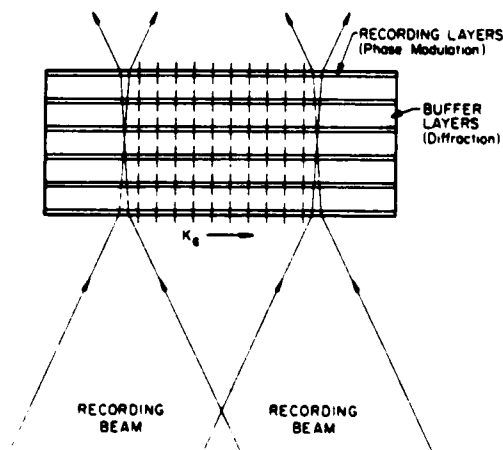


Fig. 1. The SVHOE structure, showing the recording of a grating with wave vector K_G by means of two coherent, collimated recording beams.

curve is inversely proportional to the grating thickness: a thin grating has a broad response curve indicative of Raman-Nath regime behavior, whereas a thick grating has a quite narrow response curve indicative of Bragg regime behavior.⁴

Comparative angular response curves for several SVHOE structures are shown in Figs. 2 and 3. In these figures, the recording layers in each thin grating stack are assumed for simplicity to have infinitesimal thickness and to impose only phase (as opposed to amplitude) modulation. When the structure comprises several thin phase modulation layers, the sum of the phase modulation for all layers has been set equal to that of the reference case of the uniform volume grating (Fig. 2, top) and is divided equally among all the thin recording layers. The thickness of each optically passive buffer layer is adjusted such that the total device thickness is equal to that of the thick uniform volume grating. Finally, the index of refraction of the buffer layers has been set equal to the average index of the modulation layers.

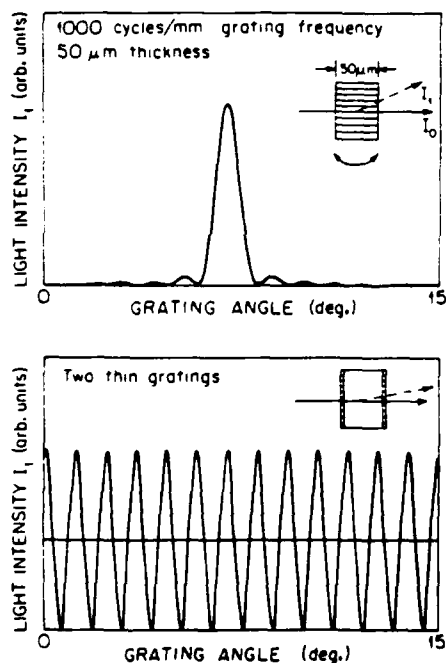


Fig. 2. Angular response for a distributed thick grating (top), for a single thin grating (horizontal line, bottom), and for a pair of thin gratings (bottom).

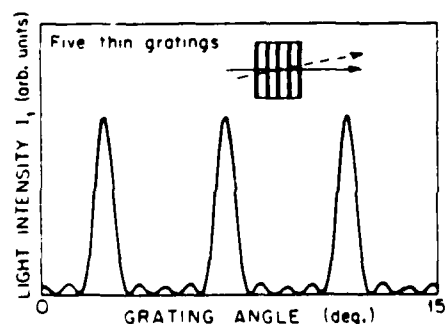


Fig. 3. Angular response for five thin gratings.

Consider first a single thin grating with modulation strength equal to that of the bulk grating. Such a thin grating has negligible angular alignment sensitivity, as shown by the horizontal line in the bottom half of Fig. 2. The angular alignment sensitivity of a two-grating structure, however, exhibits essentially a sinusoidal oscillation, as shown in the bottom half of Fig. 2, with an angular period defined by the ratio of the grating period to the thickness separating the two gratings. (Note that this structure has potentially interesting applications as a high-resolution angle encoder.)

A stack of five thin gratings, as shown in Fig. 3, suppresses three of every four peaks that appear when only the two outer gratings are present. In general, a stack of N thin gratings exhibits an angular alignment sensitivity that is periodic, with a period of $N - 1$ two-grating peaks, in which $N - 2$ of the two-grating peaks have been suppressed by the interior gratings.

For a given angular interval, the incorporation of a sufficiently large number of thin gratings leads to a structure with an angular alignment sensitivity that is indistinguishable from that of the bulk grating. Thus a SVHOE structure can emulate closely the Bragg response of a bulk grating, using surprisingly few thin grating layers. In addition, this structure exhibits features not found in bulk gratings that may prove useful for certain system applications. For example, illumination of a SVHOE structure with focused monochromatic light produces uniformly spaced angular response peaks, which can be transformed with a lens to produce an array of equally spaced points. This function is useful for interconnection of optical cellular logic arrays. For example, illumination with 850-nm light of a two-grating SVHOE with a grating separation of 1 cm and a grating spatial frequency of 350 cycles/mm at $F/3.3$ will produce in excess of 1000 regularly spaced diffracted beams.

Further, for real-time material implementations of SVHOE structures in which the photoresponse for grating formation can be either optically or electrically switched on a layer-by-layer basis, a structure with a given number of layers can interconnect either every element or every p th element with the source, in a programmable manner. For example, such layer-by-layer programmability may prove feasible by either optical bleaching⁵ or electrical tuning⁶ of the exciton resonance in a multiple quantum well structure or by modulation of a doping superlattice.⁷ SVHOE structures can also be conceived that exhibit entirely different diffraction behavior in response to grating formation by distinct writing wavelengths, by actively altering the layer photosensitivity spectrum on a layer-by-layer basis. These features collectively allow for additional degrees of freedom in the formulation of both passive and active holographic elements.

Several simple experiments have been performed to demonstrate and verify the SVHOE concept. Phase gratings have been recorded in positive photoresist (Shipley 1450J) deposited upon microscope cover glass pieces (22 mm $\times$ 22 mm, approximately 200 μ m thick). The grating frequency was adjusted to be 74 cycles/mm, as recorded by a 442-nm laser interferometer. Readout was performed by a 633-nm laser beam, which probed either individual gratings or stacks of

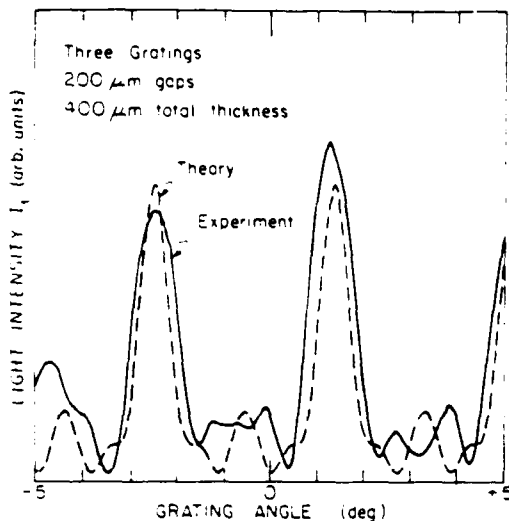


Fig. 4. Angular response for three gratings, each spaced from the next by a single glass plate thickness: experiment (solid curve) and computer model (dashed curve).

gratings. Gratings were exposed individually, processed, and then assembled using moiré imaging techniques to achieve proper alignment both in angle and in linear translation with respect to each other.

The angular response has been measured and compared with theoretical calculations for the following cases: (1) a single grating, (2) two gratings separated by 200 μm (the thickness of a single glass plate), (3) two gratings separated by 400 μm , (4) two gratings separated by 800 μm , and (5) three gratings, each separated from the next by 200 μm , for a total device thickness of 400 μm . In all cases, the agreement between experiment and theory is excellent. Representative experimental and theoretical curves for the three-grating case are shown in Fig. 4.

The peak diffraction efficiency of a SVHOE structure is a function not only of the modulation strength of each grating layer but also of the spacing between the layers. A single thin phase grating can diffract no more than 33.9% of the incident readout light power into the +1st diffraction order, assuming a sinusoidal grating profile. A two-grating structure can diffract as much as 67.7%, as discussed by Zel'dovich *et al.*,⁸ or as little as 22.2%, assuming the same grating strength for each layer but changing only the buffer layer thickness. A three-grating structure can diffract as much as 87.0%, a four-grating structure in excess of 90%, and a five-grating structure more than 95%, if the buffer-layer thickness is chosen to be optimum in each case. Maximum diffraction efficiency (at least for grating strengths ranging from zero to the first maximum in diffraction efficiency) occurs when the thickness of each buffer layer is given by

$$Q_{\text{buffer}} = \lambda D_{\text{buffer}} / n \Lambda^2 = 2\pi(m + 1/2), \quad (1)$$

in which λ is the light wavelength, D_{buffer} is the buffer-layer thickness, n is the index of refraction of the buffer layers, Λ is the period of the grating, and m is an integer. Thus, neglecting the thickness of individual

grating layers and counting only the buffer layers, the optimum total thickness D_{total} for a structure with N gratings is given by

$$Q_{\text{total}} = \lambda D_{\text{total}} / n \Lambda^2 = 2\pi(m + 1/2)(N - 1). \quad (2)$$

Note that the broadest spatial-frequency response occurs when the smallest optimal thickness is employed, i.e., when $Q_{\text{total}} = (N - 1)\pi$.

The angular response analysis shown in Figs. 2 and 3 has assumed infinitesimally thin recording layers. The effect of finite recording-layer thicknesses is to introduce an angular response rolloff associated with any single layer, which in turn serves as an envelope function to modulate the angular response of more intricate structures constructed from multiple layers. In fact, the angular response characteristic is just the Fourier transform of the depth distribution of the grating strength.⁹ Thus for certain applications one design constraint might be an upper bound on the thickness of each recording layer to keep the overall response rolloff envelope sufficiently broad.

We have demonstrated that SVHOE's can satisfactorily emulate distributed bulk gratings over a prescribed angular interval. However, even more intriguing are those attributes unique to SVHOE's that suggest new device applications. Specifically, three features have been identified to date: (1) the ability to control the recording sensitivity of individual recording layers, independent of adjacent layers, either optically or electrically, thus altering the diffraction characteristics of the device; (2) the ability to record holograms with distinct diffraction characteristics at two or more wavelengths, also through the use of electrical or optical control on a layer-by-layer basis, and (3) the existence of periodic multiple response peaks in the angular response profiles.

This research was supported in part by the Defense Advanced Research Projects Agency (Office of Naval Research), the U.S. Air Force Office of Scientific Research (University Research Initiative Program), and the Joint Services Electronics Program.

References

1. R. V. Johnson and A. R. Tanguay, Jr., *Opt. Eng.* **25**, 235 (1986).
2. J. A. Fleck, Jr., J. R. Morris, and M. D. Feit, *Appl. Phys.* **10**, 129 (1976).
3. A. R. Tanguay, Jr., and R. V. Johnson, *J. Opt. Soc. Am. A* **3**(13), P53 (1986).
4. W. R. Klein and B. D. Cook, *IEEE Trans. Sonics Ultrason.* **SU-14**, 123 (1967).
5. D. A. B. Miller, D. S. Chemla, P. W. Smith, A. C. Gossard, and W. Weigmann, *Opt. Lett.* **8**, 477 (1983).
6. D. A. B. Miller, D. S. Chemla, T. C. Damen, A. C. Gossard, W. Weigmann, T. H. Wood, and C. A. Burrus, *Phys. Rev. Lett.* **53**, 2173 (1984).
7. G. H. Dohler, *Opt. Eng.* **25**, 211 (1986).
8. B. Ya. Zel'dovich, D. I. Mirovitskii, N. V. Rostovtseva, and O. B. Serov, *Sov. J. Quantum Electron.* **14**, 364 (1984).
9. J. R. Rogers, Institute of Optics, University of Rochester, Rochester, New York 14627 (personal communication, 1986).

Physical and Technological Limitations of Optical Information Processing and Computing

Armand R. Tanguay, Jr.

Introduction

Over the past four decades, the growth of information processing and computational capacity has been truly remarkable, paced to a large extent by equally remarkable progress in the integration and ultra-miniaturization of semiconductor devices. And yet it is becoming increasingly apparent that currently envisioned electronic processors and computers are rapidly approaching technological barriers that delimit processing speed, computational sophistication, and throughput per unit dissipated power. This realization has in turn led to intensive efforts to circumvent such bottlenecks through appropriate advances in processor architecture, multiprocessor distributed tasking, and software-defined algorithms.

An alternative strategy that may yield significant computational enhancements for certain broad classes of problems involves the utilization of multi-dimensional optical components capable of modulating and/or redirecting information-carrying light wavefronts. Such an optical processing or computing approach relies for its competitive advantage principally on massive parallelism in conjunction with relative ease of implementation of complex (weighted) interconnections among many (perhaps simple) process-

ing elements. A wide range of computational problems exist that lend themselves quite naturally to optical processing architectures, including pattern recognition, earth resources data acquisition and analysis, texture discrimination, synthetic aperture radar (SAR) image formation, radar ambiguity function generation, spread spectrum identification and analysis, systolic array processing, phased array beam steering, and artificial (robotic) vision. In addition, many neural network processes that inherently rely on intricate interconnection patterns have been or can be implemented optically. These and other applications are treated in more detail in accompanying articles in this special issue of the *MRS Bulletin*^{1,2} and in special issues of *IEEE Proceedings*³ and *Optical Engineering*⁴ on optical computing.

A generalized optical processor or computer can be depicted schematically as shown in Figure 1. The physical constituent elements of such a system include a central processing unit (CPU) that performs the essential implementable function, a data management processor that orchestrates the flow of data and sequence of operations (usually considered part of the CPU in a traditional electronic computer), several types of memory elements for both short-term and long-term data storage

and buffering, format devices to spatially organize input data fields, input devices to convert data input types to a form amenable to subsequent processing, output devices to convert processed results to detectable and interpretable forms, and detectors to produce externally addressable results. In Figure 1, feedback interconnects are explicitly shown as separate components to emphasize their crucial role in implementing parallel iterative algorithms and complex weighting functions.

A wide variety of optical components are required to implement processors based on the generalized architecture shown in Figure 1.⁵ These include one- and two-dimensional spatial light modulators, volume holographic optical elements, threshold arrays, optical memory elements, sources, source arrays, detectors, and detector arrays. The state of the technology is such that while demonstration devices and prototypes are proliferating, with the exception of sources, detectors and detector arrays, few (if any) such components have as yet achieved significant commercial success or even demonstrated technological viability. In large part, this is due to the fact that each of the candidate technologies has placed rather severe demands on the state-of-the-art of the materials employed regardless of the nature of the optical effect utilized (e.g., electrooptic, magneto-optic, photorefractive, or electroabsorptive). In other words, the magnitudes of observable optical perturbations per unit excitation are just not large enough with readily available materials to allow flexible device engineering. As such, the answer to whether optical processing and computing will come of age may ultimately rest on the capabilities and fortunate discoveries of materials scientists and process engineers.

It is of considerable interest, nonetheless, to examine the physical as well as technological limitations that apply to optical information processing and computing systems in order to assess their potential performance advantages (if any) over comparable electronic counterparts, and to provide much-needed guidance for continued research efforts in optical materials, devices, algorithms, and system architectures. Although the study of fundamental physical limitations in the context of digital (binary) electronic systems (particularly VLSI) is well established,^{6,7} it should be noted that comparable studies of optical processing and computing systems are rela-

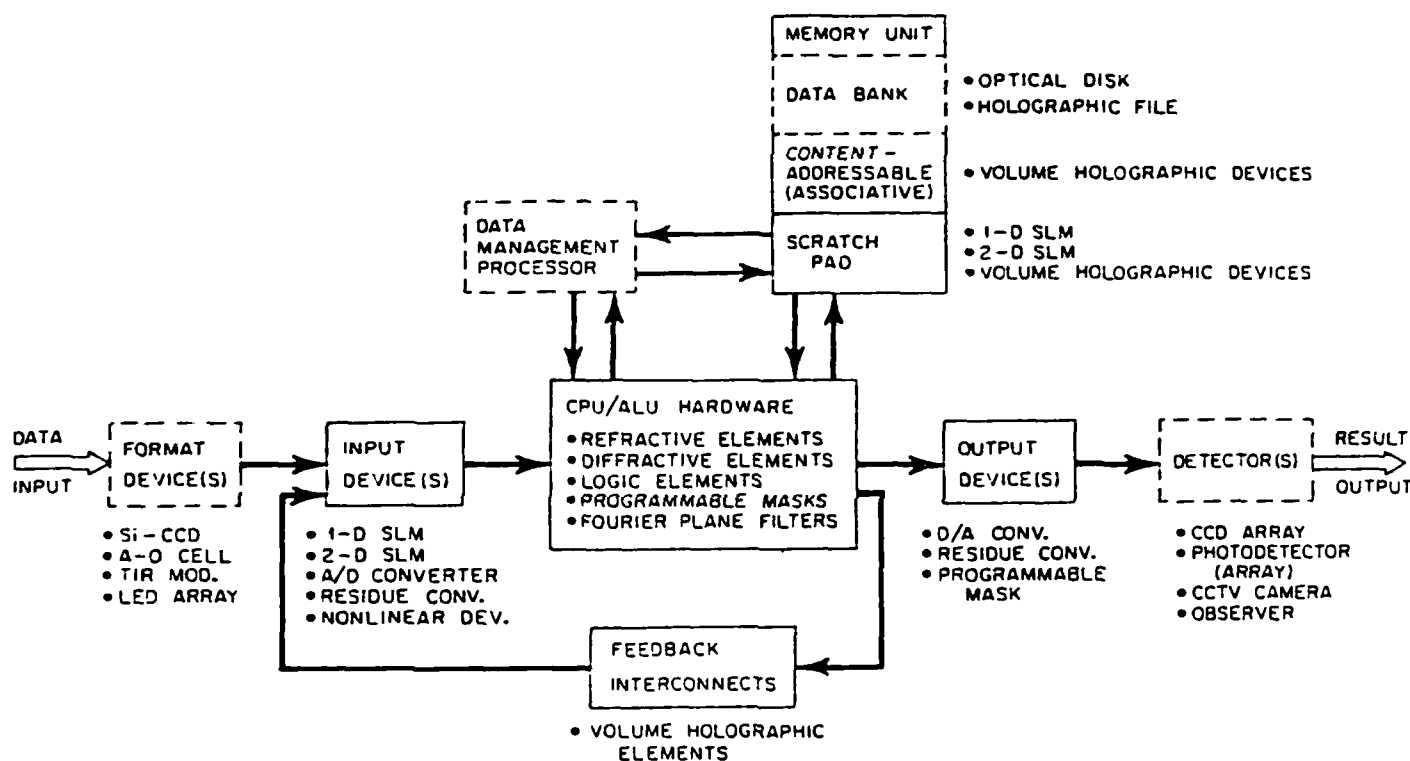


Figure 1. Schematic diagram of the elements of a generalized optical processor or computer (after Ref. 5).

tively recent, and have to date focused primarily on digital optical computing as opposed to analog optical processing.⁶⁻¹¹

The remainder of this article will describe several such physical and technological limitations of both optical information processing and computing. The following section addresses the nature of computation from the perspective of identifying physical (*technology-independent*) limits. The next section will then provide an example of a *technology-dependent* limit in the context of photorefractive volume holographic optical interconnections. Conclusions are drawn in a final section.

The Nature of Computational Constraints

In order to implement any computational algorithm on a machine of given architecture and component hardware, we must first choose a *representation* for inputting data and executing computational steps. Such a representation might, for example, be digital, analog, or even symbolic. As we shall see, such a choice of representation has profound implications on the physical limits that apply to subsequent computation.

For a given representation, we can then assess the *computational complexity* of the chosen algorithms. We define the computational complexity of an operation as the equivalent number of irreducible binary switching operations required to perform the same operation in the most efficient manner possible. For example, the relatively simple operation of transferring 1,000 ten-bit words from the CPU to memory will require at the very least 10^4 binary switching operations (assuming that all of the data is transferred in parallel), and most likely considerably more if shift registers are employed to multiplex the data transfer. This definition of computational complexity provides a convenient means for comparing the overall efficiency or total minimum energy cost of a given computation performed by different algorithms on a machine of given architecture and technology, or by machines based on vastly different architectures and perhaps disparate technological hardware. It does not, however, take into account the interconnection complexity required to implement communications pathways at the device, circuit, and subsystem levels.

Finally, we must end each computation with a *detection* of the desired results, in order to allow the result to be used (for example, to implement a desired action). This separation of every computational process into the implementation of representation, computational complexity, and detection functions allows us to establish sets of interrelated limits that apply to various combinations of choices.

For purposes of discussion in this article, let us examine each of these functions from the point of view of energy cost. In so doing, we seek to identify physical limits on the maximum computational throughput achievable per watt of dissipated power. It should perhaps be noted here that the "requirement" of energy dissipation is not in fact a fundamental limit, and derives instead from a system design demand to *assert* each stored value as rapidly as possible in order to complete the entire computation deterministically and in minimum time. Thus adiabatic computational processes which do not require a minimum energy dissipation¹² are not applicable to the present discussion.

In the digital (binary) realm, the lowest possible energy cost of representing

a number requiring a given number of bits is equal to the number of bits times the minimum energy to store a single bit. For semiconductor electronic switches, the minimum energy in turn is equal to a few tens of $k_B T$ per electron^{6,7} times the number of electrons required to guarantee detection with a given bit error rate (BER) [a quantum rather than a thermal limit]. For a bit error rate of 10^{-3} or so, about ten electrons are required (assuming an "ideal" detector [or following switch] that can unambiguously differentiate between the presence and absence of a single electron). Therefore each switching event (or representation of each bit) requires the dissipation of approximately $200 k_B T$. This places an immediate upper boundary on the maximum computational throughput of such an electronic digital (binary) computing engine of approximately 10^{18} transitions (irreducible binary switching operations) per second per watt at room temperature. Note that this bound does not include the assessment of any cost for internal communication of information, as would be required to execute an actual computation.

To place this number in proper perspective, we need only look at the switching energies representative of current semiconductor technology. For electronic devices, we might examine a typical CMOS capacitor with a $100 \mu\text{m}^2$ cross-sectional area and a $1,000 \text{ \AA}$ oxide thickness, for which the charge/discharge cycle consumes CV^2 worth of energy. If we operate at the minimum switching voltage of a few tens of $k_B T/q$,^{6,7} the switching energy is about $10^6 k_B T$ at room temperature, about four orders of magnitude above the fundamental limit. Current digital logic circuits operate at roughly $2.5 \times 10^{-12} \text{ J/transition}$, which is about $10^6 k_B T$.⁷ This is also roughly the minimum energy required to operate even a local interconnection at GHz rates,^{7,8} an unavoidable cost of communication between devices at the circuit level. At the system level, a vast amount of additional overhead comes into play. For example, the DEC VAX 11-750 dissipates approximately 3 kW running fully loaded at its maximum throughput rate of $750,000$ instructions/second ($200,000$ operations/second). Thus the energy required to perform a single instruction is about 3 mJ or $10^{15} k_B T$, which corresponds to a throughput rate of 250 instructions/second/watt.

By comparison, analog representations (as used extensively, for example,

in optical processors) require far more energy than the binary equivalent. This is due to the necessity of utilizing a much higher particle count (electrons or photons) in order to minimize the effects of quantum statistical fluctuations on the BER. For example, if we wish an analog representation of the number $1,000$, then we require a dynamic range of at least $1,000:1$. For incoherent illumination, quantum fluctuations in the emission/detection process produce a photon number distribution with a relative standard deviation of $\sigma \approx \sqrt{N}/N$. The equivalent of a 10^{-9} BER for the digital case corresponds to roughly 12 standard deviations. Therefore, the number of photons required must be greater than 1.5×10^8 from statistical considerations alone. For a GaAs semiconductor laser characterized by a photon energy of $\sim 1.5 \text{ eV}$, this corresponds to about $10^{10} k_B T$. To represent $1,000$ optically in binary requires approximately 14 bits (10 bits for the number plus 4 bits of overhead) at 15 eV each (10 photons at 1.5 eV each, assuming direct detection and an ideal detector), or about $10^6 k_B T$.

Given the remarkably higher cost of analog representation as compared with digital, any competitive advantages for analog systems from the perspective of an energy dissipation metric due to physical limits must come from enhanced computational complexity. The various tradeoffs involved can perhaps best be described in terms of an example. Consider, then, a highly interconnected, nonlocal problem such as the Fourier transformation of a two-dimensional function. Assume that the function is sampled on a $1,000 \times 1,000$ element array at 10 bits.

If we operate in the binary regime, we can utilize a highly efficient discrete Fourier transform (DFT) routine such as the Cooley-Tukey algorithm. The number of complex operations (multiplies and adds) scales as $1.5N^2 \log_2 N^2$ for an $N \times N$ input. If we assume an electronic machine implementation that requires 20 switching events/bit/complex operation, then approximately 10^{10} switching events are required to perform the computation, or 10 nJ for operation near the physical limits described above.

If, on the other hand, the input data field (two-dimensional function) is represented in analog form by means of a spatial light modulator, illumination by a coherent wavefront will produce the required transform in the back focal plane of a (following) lens. Here we see illustrated the notion of a computation

as the transformation of information. The energy cost of this particular operation accrues only to the initial representation of the function (essentially, the optical or electronic addressing of the spatial light modulator) and to the subsequent detection of the final result. This is then equivalent to 2×10^6 individual pixel detection operations for the optically addressed case, or $70 \mu\text{J}$ at the physical limits for 1.5 eV photons, still about four orders of magnitude larger than the binary equivalent. The computational complexity implemented by the analog optical processor is thus seen to partially offset the large initial difference in representation energy costs. For other classes of problems with even higher inherent computational complexity, the physical boundary on energy cost may thus favor an analog approach.

The situation looks quite a bit different if we examine current technological (rather than more fundamental) constraints. For current digital circuits that switch at around 2.5 pJ/bit , the DFT just described dissipates 25 mJ for the required number of switching operations alone. Current digital systems such as the VAX 11-750 consume about 10^{-4} J/bit , or 1 MJ for the DFT. With regard to optical systems, spatial light modulators are available with input sensitivities of about 200 pJ/pixel at high signal-to-noise ratio, and CCD detector arrays dissipate approximately 1 mJ/Mpixel on readout. Hence the analog optical Fourier transform can be performed for an energy cost of about 1.2 mJ . This apparent capability of analog optical systems to generate significantly enhanced computational throughput per unit input power is thus due largely to the fact that currently available analog optical components operate far closer to the relevant thermal and quantum limits than current digital electronic (VLSI) components. It should be noted that such a comparison begs the question of overall computational accuracy, which clearly favors the digital implementations due to nonlinearities and nonuniformities in currently available analog optical components.

The third principal component of the computational process, that of detection of the results (whether electronic or optical), is subject to the same thermal and quantum statistical limitations as the process of representation. This statement follows directly from an assumption inherent in assessing the minimum energy required for representation—that each number must be capable of detection at a given BER.

assuming an ideal detector. Technological constraints can then be added to the fundamental limit problem as before to assess realistic current capabilities. For example, available optical detectors typically require a mean photon number of 1,000 for a 10^{-9} BER in the direct detection mode, rather than the value of 10 assumed above.

Volume Holographic Optical Interconnections

As pointed out in the Introduction (and discussed in detail in Ref. 1), the ability to provide complex, multidimensional, programmable, weighted interconnections by means of volume holograms is an attractive feature of optical processing and computing systems. In addition, this ability perhaps more so than any other contributes significant computational complexity to optical architectures. To this end, it is of considerable importance to assess the fundamental limitations that apply to such volume holographic optical interconnections.

Although many types of photosensitive media can be utilized for the recording and readout of volume holograms, the dynamic reprogrammability offered by photorefractive crystals such as lithium niobate, barium titanate, bismuth silicon oxide, and gallium arsenide has made such materials the objects of intensive study. In these materials, the interference between two coherent optical wavefronts (the signal and reference beams) generates a space-variant photoexcitation distribution that produces in turn a related space charge redistribution and associated electric field pattern by carrier drift and diffusion. Many interesting limits apply to such a process, including the maximum number of independent interconnections that can be sequentially or simultaneously recorded at a given value of allowable crosstalk, the highest achievable diffraction efficiency of each independent interconnection at the maximum interconnection density, the absolute minimum number of photoevents per unit volume required to record an interconnection of given analog weight within the quantum fluctuation limits for statistical accuracy of recording and reconstruction (readout), and the maximum asymmetry possible between the recording and erasure processes.¹⁵ An additional limit of considerable importance is that of the photorefractive sensitivity,¹⁶⁻¹⁸ or the refractive index modulation obtained in recording a uniform grating of fixed spatial fre-

quency per unit (incident or absorbed) energy density, as this parameter places an upper bound on the maximum rate of interconnection reprogrammings that can be accomplished per unit average power. In what follows, the fundamental physical limitations on the photorefractive sensitivity are examined in a bit more detail.

A number of factors contribute to the photorefractive sensitivities characteristic of photoconductive, electrooptic materials. One such factor is the photogeneration quantum efficiency, which represents the number of photogenerated mobile charge carriers per photon absorbed from the recording beam(s). A second factor is the charge transport efficiency, which is a measure of the degree to which the average photogenerated mobile charge carrier contributes to the forming space charge grating after separation from its original site by means of drift and/or diffusion and subsequent trapping. The magnitude of the space charge field generated by a given space charge grating is inversely proportional to the dielectric permittivity ϵ of the photorefractive material, which thus contributes a third factor to the grating recording sensitivity. A fourth factor describes the perturbation of the local index ellipsoid (dielectric tensor at optical frequencies) that results from a given space charge field through the electrooptic frequencies) that results given space charge field through the electrooptic (Pockels or Kerr) effect.

In addition, several other physical quantities factor into an evaluation of the photorefractive sensitivity, including the wavelength of the recording illumination (to convert the number of absorbed photons into an equivalent energy), the absorption coefficients of the material at the wavelengths of both the recording and readout beams (to correct for the fractional absorbance of the recording beams and the fractional transmittance of the readout beam), and the magnitude of the applied voltage (which significantly alters the sensitivity characteristics for certain materials by changing the nature of the dominant charge transport mechanism from the diffusion regime to the drift regime).

In order to provide a quantitative metric that does in fact have a fundamental physical limitation, we define the *grating recording efficiency*¹⁹ of a photorefractive recording model or configuration as the magnitude of the space charge field produced by a fixed number of photogenerated mobile charge carriers at a given spatial frequency,

normalized by the maximum quantum limited space charge field that can be produced by the same number of photogenerated carriers. This metric thus effectively combines the notions of a photogeneration efficiency and a charge transport efficiency, and provides an estimate of the fundamental quantum efficiency of the photorefractive grating recording process. For simplicity, we confine our attention here to a photorefractive model characterized by a single mobile charge species, and a single type of donor site with associated un-ionized donor and ionized donor (trap) states.

Let us compare the grating recording efficiencies of four idealized grating recording models, generated by combining two types of photogeneration profiles (a comb function and a sinusoid) with two types of charge transport processes (a translation of each photogenerated charge carrier by exactly half of a grating wavelength; and uniform redistribution, or randomization, of all photogenerated charge). The resulting four combinations are shown schematically in Figure 2.

The maximum quantum limited space charge field that can result from a fixed number of photogenerated mobile charges is given by the bipolar comb distribution, which generates a square wave electric field profile from alternating positive and negative sheets of charge spaced by half of a grating wavelength. It is in fact the first harmonic component of this electric field profile that we are interested in, as we assume a volume grating operated deep within the Bragg diffraction regime. The bipolar comb distribution can thus be assigned a grating recording efficiency of unity.

By comparison, the monopolar comb, transport-efficient sinusoid, and baseline sinusoid cases yield grating recording efficiencies of $1/2$, $1/2$, and $1/4$, respectively. These results are summarized in Figure 3, which shows the evolution of the first harmonic component of the space charge field for all four cases as a function of grating recording time.

The saturation behavior of each case derives from assuming a fixed finite trap density prior to grating recording. Since the resultant diffraction efficiencies of such gratings scale as the square of the space charge field (for small modulation), these results imply that the baseline sinusoid distribution will exhibit a diffraction efficiency lower than the bipolar comb distribution by a factor of 16. This conclusion is of considerable

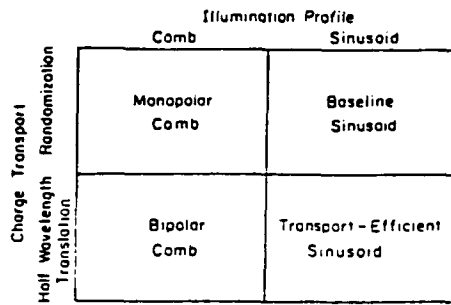


Figure 2. Schematic diagram depicting the illumination profile and charge transport assumptions underlying each of the four idealized photorefractive recording models discussed in the text.

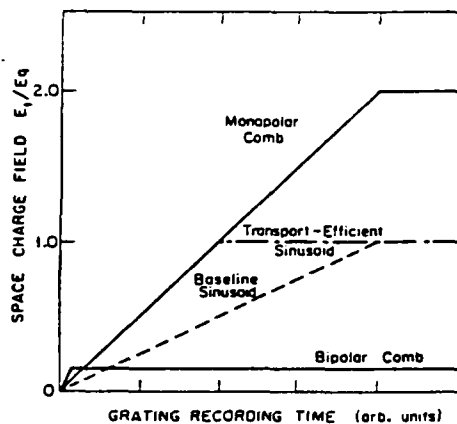


Figure 3. The evolution of the space charge field as a function of grating recording time for each of the four idealized photorefractive recording models. The normalization parameter $E_q = eN_A/\epsilon\epsilon_0K_G$ in which N_A is the maximum available trap density and K_G is the grating wavevector.

interest because the baseline sinusoid case can be shown¹⁹ to represent an asymptotic upper bound on the grating recording efficiency predicted by more realistic (rather than idealized) photorefractive grating recording models that assume a sinusoidal illumination profile. In fact, several photorefractive materials have been demonstrated to exhibit photorefractive sensitivities that approach the upper bound represented by the baseline sinusoid case.¹⁸

Nonetheless, the photorefractive grating recording process is from this

perspective somewhat quantum inefficient, in that it does not make the best possible use of either input photons or photogenerated carriers. In view of the analysis presented above, therefore, it is of considerable interest to imagine novel engineered photorefractive materials with donor planes spatially separated from trap planes, illuminated by a comb-like intensity pattern (which can be generated, for example, by utilizing stratified volume holographic optical elements [SVHOEs]²⁰). Such materials could find applications in devices based on generation of a carrier grating of given spatial frequency, such as the Photorefractive Incoherent-to-Coherent Optical Converter (PICOC).²¹

Furthermore, it should be noted that the limits derived for interconnections based on photorefractive materials pertain to a particular physical mechanism involving photoexcitation, charge transport, and electrooptically induced index changes. The importance of multiplexed interconnections to overall schemes for optical processing and computing suggests continued emphasis on the search for alternative physical effects that can provide either enhanced sensitivity or more conveniently implementable desirable features such as selective erasure.²²

Conclusions

An examination of both the fundamental physical and current technological limitations to computational performance shows that in terms of throughput per unit power, digital (binary) representations (whether electronic or optical) presently operate many orders of magnitude away from the relevant boundaries. Analog representations and detections are inherently much more power consumptive than digital representations and detections, and processing schemes based on analog algorithms must demonstrate considerably enhanced computational complexity prior to each intermediate detection plane in order to be energy competitive at the appropriate thermal and quantum limits.

Hybrid architectures and algorithms (such as those employed by numerous neural network models) that effectively combine analog representations as weights in highly multiplexed parallel interconnections with binary representations in the form of threshold arrays (decision planes) may ultimately prove to be a nearly optimum compromise. This is particularly applicable to optical processing and computing architec-

tures, for which available analog components (spatial light modulators and photorefractive volume holographic optical elements) operate much closer to the relevant quantum statistical limits than the corresponding digital components (threshold arrays). In such hybrid architectures, the threshold array performs the equivalent of a parallel level restore function (saturated nonlinearity) that is essential to the proper convergence of many processing and computing algorithms.

Further study of the fundamental physical limitations that apply to optical processing and computing will provide important guidance for the continued development of active optical materials, primarily by differentiating between avenues of opportunity with large potential gain that depend on key materials parameters, and those for which current materials characteristics are sufficient to approach relevant performance boundaries.

Acknowledgments

Research on the fundamental limitations of optical information processing and computing at the University of Southern California is supported in part by the Air Force Office of Scientific Research, the Defense Advanced Research Projects Agency, and the Army Research Office.

References

1. D.Z. Anderson, *MRS Bull.* (this issue).
2. A.M. Glass, *MRS Bull.* (this issue).
3. J.A. Neff, ed., *Opt. Eng.*, Special Issue on Optical Computing 24 (1) (1985).
4. H.J. Caulfield, S. Horvitz, G.P. Tricoles, and W.A. Von Winkle, eds., *Proc. IEEE*, Special Issue on Optical Computing 72 (7) (1984).
5. A.R. Tanguay Jr., *Opt. Eng.* 24 (1) (1985) p. 2-18.
6. C. Mead and L. Conway, *Introduction to VLSI Systems*, (Addison-Wesley, Reading, MA, 1980) p. 333-372.
7. R.W. Keyes, *The Physics of VLSI Systems*, (Addison-Wesley, Reading, MA, 1987).
8. R.W. Keyes, *Proc. IEEE*, 63 (5) (1985) p. 740-767.
9. P.W. Smith, *Bell Systems Tech. J.* 61 (1982) p. 1975-1993.
10. R.L. Fork, *Phys. Rev. A* 26 (4) (1982) p. 2049-2064.
11. R.W. Keyes, *Opt. Acta* 32 (5) (1985) p. 525-535.
12. R. Landauer, in *Optical Information Processing*, edited by Yu. E. Nesterikhin, G.W. Stroke, and W.E. Kock (Plenum Press, New York, 1976) p. 219-253.
13. C.H. Bennett, *Int. J. Theor. Phys.* 21 (12) (1982) p. 905-941.
14. C. Kyriakakis, P. Asthana, R.V. Johnson, and A.R. Tanguay Jr., *Proc. Opt. Soc. Am. Topical Meeting on Spatial Light Modulators*.

Physical and Technological Limitations of Optical Information Processing and Computing

Lake Tahoe, Nevada, (1988).

15. See Reference 1 for a discussion of the usefulness of this asymmetry.

16. A.M. Glass and D. von der Linde, *Ferroelectrics* 10 (1976) p. 163-166.

17. F. Micheron, *Ferroelectrics* 18 (1978) p. 153-159.

18. A.M. Glass, *Opt. Eng.* 17 (5) (1978) p. 470-479.

19. R.V. Johnson and A.R. Tanguay Jr., in *Optical Processing and Computing*, edited by H. Arsenault and T. Szoplik (Academic Press, New York, 1988), (in press).

20. R.V. Johnson and A.R. Tanguay Jr., *Opt.*

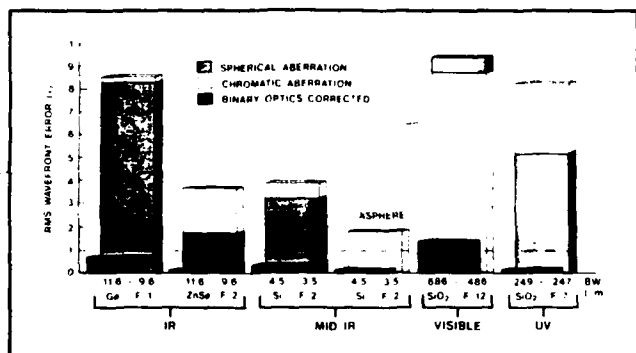
Lett. 13 (3) (1988) p. 189-191.

21. A. Marrakchi, A.R. Tanguay Jr., J. Yu, and D. Psaltis, *Opt. Eng.* 24 (1) (1985) p. 124-131.

22. D. von der Linde and A.M. Glass, *Appl. Phys.* 8 (1975) p. 95-100.



OPTICS IN 1988



Binary optics correction of typical lenses.

ly add the power of laser diode arrays (modular laser power), 2) beam steering or wavefront manipulation, and 3) opto-electronic integration of imager focal planes.

Thus far, work has concentrated on the technology involved in designing and producing individual elements for systems. The vast potential of binary optics to spawn new products and industries will materialize only as industrial expertise develops, system designers become familiar with its cost, weight, and design benefits, and manufacturers put it in production through replication, embossing, and forging or molding from a single master element.

STRATIFIED VOLUME HOLOGRAPHIC OPTICAL ELEMENTS

R.V. JOHNSON AND A.R. TANGUAY JR.
OPTICAL MATERIALS AND DEVICES LABORATORY, AND
CENTER FOR PHOTONIC TECHNOLOGY
UNIVERSITY OF SOUTHERN CALIFORNIA
LOS ANGELES, CALIF.

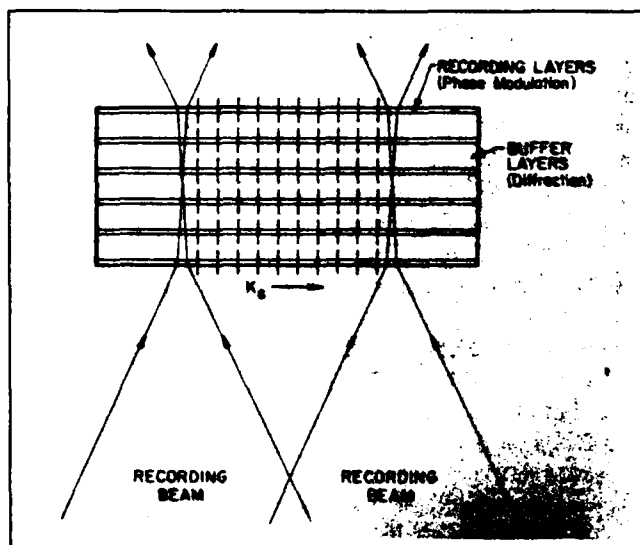
Holographic optical elements are of fundamental importance to a number of applications in optical information processing and computing, including optical interconnections, content-addressable memories, and various linear and nonlinear signal processing configurations. Numerical analyses of typical volume holographic structures are critical for characterizing and optimizing such optical elements. One of the most flexible and broadly applicable tools for studying the diffraction behavior of volume holograms is the optical beam propagation method (BPM).^{1,2} In this method, the distributed optical inhomogeneities that characterize a typical hologram are approximated by a discrete sequence of physically and mathematically simplified elements: infinitesimally thin phase and/or polarization modulation layers, interleaved with optically

homogeneous layers of finite thickness. By approximating the distributed grating with a sufficiently large number of these discrete elements, the resulting numerical model of the grating can be brought arbitrarily close to that of the desired distributed bulk grating.

The BPM concept of separating the modulation process from the diffraction process in turn suggests a new class of optical devices: the stratified volume holographic optical element, or SVHOE^{3,4} (see figure). The SVHOE device structure consists of a sequence of thin photosensitive holographic recording layers that perform the optical modulation function, interleaved with optically passive buffer layers, i.e., layers that impress no modulation on the light beam but rather allow the diffraction processes necessary for thick grating response to occur. A grating recorded in any individual modulation layer would necessarily exhibit Raman-Nath characteristics because each recording layer is quite thin. But for a given angular interval, the incorporation of a sufficiently large number of thin gratings leads to a structure with an angular alignment sensitivity that is indistinguishable from that of the bulk grating. Thus, an SVHOE structure can closely emulate the Bragg response of a bulk grating, using surprisingly few thin grating layers.

In addition, this structure exhibits features not found in bulk gratings that may prove useful for certain system applications. For example, illumination of an SVHOE structure with focused monochromatic light produces uniformly spaced angular response peaks, which can be transformed with a lens to produce an array of equally spaced points. This function is useful for interconnection of optical cellular logic arrays. For example, illumination with 850 nm light of a two grating SVHOE with a grating separation of 1 cm and a grating spatial frequency of 350 cycles/mm at F/3.3 will produce in excess of 1000 regularly spaced diffracted beams. The occurrence of multiple (periodic) peaks in the angular response profile can be applied to angular encoding applications, in which a two grating SVHOE mounted on the part to be tracked can be probed by a single beam to allow coarse angular measurement by peak counting, and fine angular measurement by interpolation. Highly selective wavelength notch filtering can also be accomplished in SVHOE structures by appropriate choice of buffer layer thickness and grating spatial frequency.⁵

Further, for real-time material implementations of SVHOE structures in which the photoresponse for grating formation can be either optically or electrically switched on a layer-by-layer basis, a structure with a given number of layers can interconnect either every element or every *p*th element with the source, in a programmable manner. For example, such layer-by-layer programmability may



The SVHOE structure, showing the recording of a grating with wave vector K_G by means of two coherent, collimated recording beams.

prove feasible by either optical bleaching or electrical tuning of the exciton resonance in a multiple quantum well structure, or by modulation of a doping superlattice. SVHOE structures can also be conceived that exhibit entirely different diffraction behavior in response to grating formation by distinct writing wavelengths, by actively altering the layer photosensitivity spectrum on a layer-by-layer basis.

These novel features of SVHOEs collectively allow for numerous additional degrees of freedom in the formulation of both passive and active holographic elements.

REFERENCES

1. R.V. Johnson and A.R. Tanguay Jr., *Opt. Eng.* 25, 235, 1986.
2. J.A. Fleck Jr., J.R. Morris, and M.D. Feit, *Appl. Phys.* 10, 129, 1976.
3. A.R. Tanguay Jr., and R.V. Johnson, *J. Opt. Soc. Am. A* 3(13), P53, 1986.
4. R.V. Johnson and A.R. Tanguay Jr., *Opt. Lett.* 13(3), 189, 1988.
5. A.R. Tanguay Jr., *Digest of the Conference on Lasers and Electro-Optics* (Optical Society of America, Washington, D.C., 1988) paper WY4.

OPTICAL COMMUNICATIONS

16 GB/S LIGHTWAVE TRANSMISSION BY OPTICAL TIME-DIVISION MULTIPLEXING

RODNEY S. TUCKER, GADI EISENSTEIN, AND STEVEN K. KOROTKY
AT&T BELL LABORATORIES
CRAWFORD HILL LABORATORY
HOLMDEL, N.J.

Optical data at a bit rate of 16 Gb/s were recently transmitted over optical fiber in an experiment at AT&T Bell Laboratories. This is the highest transmission bit-rate reported for any single-channel lightwave system—an achievement made possible using optical technology to time-division multiplex and demultiplex the data.

Commercial lightwave transmission systems use electronic time-division multiplexing (TDM) to combine a number of electronic data channels into a single higher capacity channel for transmission. In essence, time-division multiplexing is a digital technique in which data bits are interleaved in time. As bit-rates move into the multi-gigabit per second range, the speed of available electronics is being pushed to the limit and the electronic multiplexing

circuitry causes a speed bottleneck. The optical time-division multiplexing techniques used in the Bell Laboratories experiments retain the digital format of the signal, but overcome the speed bottleneck. This is achieved by capitalizing on the inherent wide-band capabilities of optical components such as pulsed semiconductor lasers and lithium niobate (LiNbO_3) directional coupler switches.

In the 16 Gb/s optical time-division multiplexed system, four optical data streams at 4 Gb/s are time-multiplexed directly in the optical domain to give a 16 Gb/s data stream for transmission. At the receiver end of the system, the 16 Gb/s data stream is demultiplexed optically to four lower bit-rate optical streams before detection and conversion to the electrical domain.

The 4 Gb/s input data streams originate from short optical pulses generated by four mode-locked semiconductor lasers. Data are encoded on the pulses using LiNbO_3 waveguide optical modulators. Optical multiplexing is carried out by simply delaying the optical pulses streams and combining them in a passive fiber directional coupler array.

The demultiplexer, the most important component in the system, separates the pulses on the incoming 16 Gb/s stream into four streams each at 4 Gb/s. A block diagram of the demultiplexer and its associated electronics is

Reprint from

Topics in Applied Physics, Volume 62:

Photorefractive Materials and Their Applications II

Survey of Applications

Editors: P. Günter and J.-P. Huignard

© Springer-Verlag Berlin Heidelberg 1989

Printed in Germany. Not for Sale.

Reprint only allowed with permission from Springer-Verlag



Springer-Verlag
Berlin Heidelberg New York
London Paris Tokyo

Photorefractive Materials and Their Applications II

Survey of Applications

Editors: P. Günter and J.-P. Huignard

1. **Introduction.** By J.-P. Huignard and P. Günter
 2. **Amplification, Oscillation, and Light-Induced Scattering in Photorefractive Crystals.** By S. G. Odoulov and M. S. Soskin
(With 25 Figures)
 3. **Photorefractive Effects in Waveguides.** By V. E. Wood, P. J. Cressman, R. L. Holman, and C. M. Verber (With 15 Figures)
 4. **Wave Propagation in Photorefractive Media.** By J. O. White, Sze-Keung Kwong, M. Cronin-Golomb, B. Fischer, and A. Yariv
(With 34 Figures)
 5. **Phase-Conjugate Mirrors and Resonators with Photorefractive Materials.** By J. Feinberg and K. R. MacDonald (With 35 Figures)
 6. **Optical Processing Using Wave Mixing in Photorefractive Crystals.** By J.-P. Huignard and P. Günter (With 56 Figures)
 7. **The Photorefractive Incoherent-To-Coherent Optical Converter.** By J. W. Yu, D. Psaltis, A. Marrakchi, A. R. Tanguay, Jr., and R. V. Johnson (With 32 Figures)
 8. **Photorefractive Crystals in PRIZ Spatial Light Modulators.** By M. P. Petrov and A. V. Khomenko (With 12 Figures)
-

7. The Photorefractive Incoherent-To-Coherent Optical Converter

Jeffrey W. Yu, Demetri Psaltis, Abdellatif Marrakchi, Armand R. Tanguay, Jr., and Richard V. Johnson

With 32 Figures

7.1 Overview

High performance spatial light modulators (SLMs) are essential in many optical information processing and computing applications for converting incoherent images to coherent replicas suitable for subsequent processing [7.1, 2]. A typical spatial light modulator consists of a photosensitive element to capture the incoherent light image and an optical modulator element to impress the incoherent input image content onto a coherent readout beam. A particularly important class of spatial light modulators employs photorefractive crystals that combine both photosensitive and electrooptic modulation functions within the same medium. Examples of electrooptic spatial light modulators that utilize photorefractive crystals include the Pockels Readout Optical Modulator (PROM) [7.3] and the PRIZ (a Soviet acronym for a crystallographically modified PROM) [7.4].

During operation of the Pockels Readout Optical Modulator, the input image-bearing beam creates photoinduced carriers which are longitudinally separated by an applied bias electric field. This charge separation produces a space-variant division of the applied field across the active electrooptic crystal and one or more dielectric blocking layers, as shown in the upper left quadrant of Fig. 7.1. The local birefringence of the medium depends on the longitudinal component of the local electric field, and hence can be sensed by a polarized readout beam observed through an exit analyzer. In the PRIZ, the same charge generation and separation process is utilized, with the difference that the crystallographic orientation is chosen to emphasize readout sensitivity to the transverse components of the induced electric field distribution, as shown in the upper right quadrant of Fig. 7.1.

The PROM and the PRIZ are typically limited in spatial frequency response to of order 10 cycles/mm at optimum optical exposure [7.5], and hence are limited to relatively modest bandwidth optical processing and computing applications. The physical configurations of the PROM and PRIZ devices do not lend themselves readily to exploitation of the remarkably high spatial bandwidths available in holographic configurations, in which the input image is encoded on a spatial carrier (as shown in the lower left quadrant of Fig. 7.1 and discussed extensively elsewhere in this book). Even when utilizing the same electrooptic crystal as in the PROM and PRIZ (typically bismuth silicon oxide), holographic recording configurations employing transverse applied elec-

ELECTROOPTIC SPATIAL LIGHT MODULATORS
($\text{Bi}_{12}\text{SiO}_{20}$)

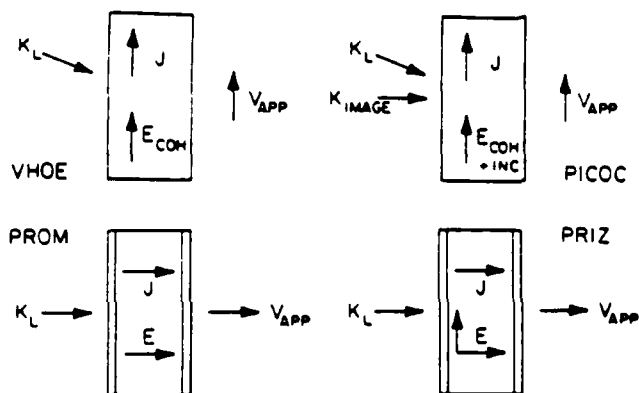


Fig. 7.1. Schematic diagram of the principal components of the applied voltage, charge transport, sensed electric field components, and input light wave vectors for four types of spatial light modulators that utilize single crystal bismuth silicon oxide ($\text{Bi}_{12}\text{SiO}_{20}$). The acronym VHOE stands for volume holographic optical element; likewise PROM stands for the Pockels Readout Optical Modulator, PRIZ is a Soviet acronym for a crystallographically modified PROM, and PICOC stands for the photorefractive incoherent-to-coherent optical converter, the subject of this chapter

tric fields and no blocking layers have been shown to exhibit spatial frequency bandwidths in excess of 2000 cycles/mm. However, such purely holographic recording requires input of image-dependent information on one of two coherent input beams, and as such cannot be directly utilized for performing the incoherent-to-coherent conversion function.

A fourth distinct type of photorefractive spatial light modulator has been independently proposed by *Kamshilin* and *Petrov* [7.6] and by the present authors [7.7, 8] which combines the incoherent-to-coherent conversion function with an essentially holographic recording process, and thereby exhibits several advantages of each. As in the holographic recording case, a transverse applied electric field is used in conjunction with two uniform coherent writing beams to produce a volume grating that is then selectively modified by a third beam encoded with the information content to be stored or converted. This configuration is shown schematically in the lower right quadrant of Fig. 7.1. The Photorefractive Incoherent-to-Coherent Optical Converter (PICOC) [7.7, 8] device is capable of recyclable real time operation, is characterized by an enhanced spatial frequency response, and is even simpler to construct than the PROM or PRIZ. In addition, the PICOC device configuration allows its use in many quasi-holographic techniques (such as optical phase conjugation), which in turn lead to potentially novel optical information processing and computing architectures [7.9–11].

The photorefractive incoherent-to-coherent optical conversion process is described in Sect. 7.2, as are alternative sequencing schemes and optical implementations. In Sect. 7.3, the limiting assumptions needed to derive a tractable analytical model of PICOC performance are specified, in preparation for detailed discussion of the recording stage in Sect. 7.4, and of the readout stage in Sect. 7.5. Conclusions and future research directions are offered in Sect. 7.6.

7.2 Physical Principles and Modes of Operation

The photorefractive incoherent-to-coherent optical conversion (PICOC) process is perhaps best understood as an extension of the more familiar holographic recording process in a photorefractive medium. The physical principles governing such recording are briefly reviewed in this section, and in much greater detail in Sect. 7.4. The extension of this recording process to include PICOC allows for at least three different temporal modes for sequencing the coherent grating with respect to the incoherent image. These modes are identified and compared in this section. In addition, two alternative optical architectures are defined, and converted images generated by one representative configuration are presented.

The high sensitivity of photoconductive and electrooptic crystals such as bismuth silicon oxide ($\text{Bi}_{12}\text{SiO}_{20}$, or BSO) in the visible portion of the spectrum has allowed the simultaneous recording and reading of volume holograms to be achieved with time constants amenable to real-time operation [7.12]. The holographic recording process in photorefractive materials involves photoexcitation, charge transport, and trapping mechanisms [7.13]. When two coherent beams are allowed to interfere within the volume of such a crystal, free carriers are nonuniformly generated by absorption and are redistributed by diffusion and/or drift under the influence of an externally applied electric field. Subsequent trapping of these charges at relatively immobile trapping sites generates a stored space-charge field, which in turn modulates the refractive index through the linear electrooptic (Pockels) effect and thus records a volume phase hologram. If both coherent writing beams are plane waves, the induced hologram consists of a uniform grating.

In the photorefractive incoherent-to-coherent optical conversion (PICOC) process, an incoherent image is focused in the volume of the photorefractive material in addition to the coherent grating beams, creating an additional spatial modulation of the charge distribution stored in the crystal. This spatial modulation can be transferred onto a coherent readout beam by reconstructing the holographic grating. The spatial modulation of the coherent reconstructed beam will then be a negative replica of the input incoherent image, as shown in Fig. 7.2. It should be noted here that a related image encoding process can be implemented nonholographically by premultiplication of the image with a grating [7.14].

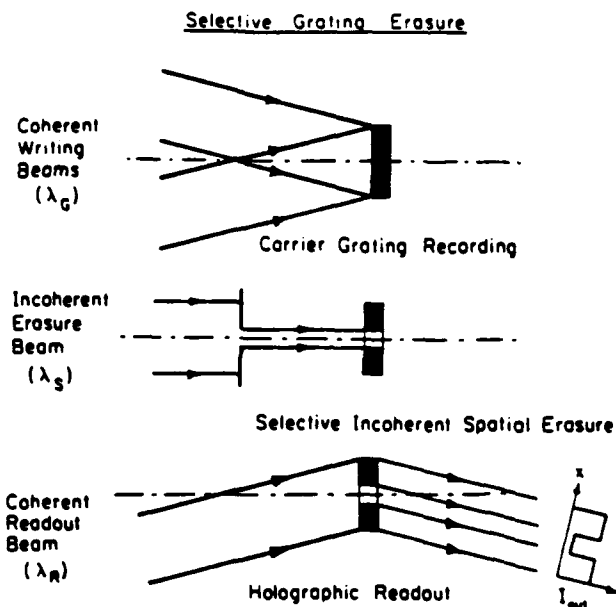


Fig. 7.2. Principle of operation of the photorefractive incoherent-to-coherent optical converter (hereinafter referred to as PICOC)

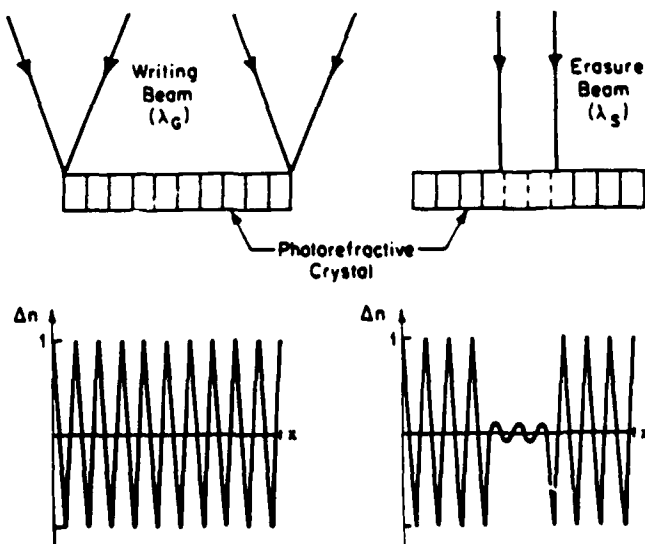


Fig. 7.3. PICOC in the grating erasure mode (GEM), in which the carrier grating is recorded before the incoherent image-bearing signal

In PICOC, the holographic grating can be recorded before, during, or after the crystal is exposed to the incoherent image. Therefore, several distinct operating modes are possible. These include the grating erasure mode (GEM; Fig. 7.3); the grating inhibition mode (GIM; Fig. 7.4), and the simultaneous erasure/writing mode (SEWM; Fig. 7.5).

In the *grating erasure mode* (GEM), shown schematically in Fig. 7.3, a uniform grating is first recorded by interfering two coherent writing beams in the photorefractive crystal. The writing beams are turned off, and this grating is then selectively erased by incoherent illumination of the crystal with an image-bearing beam. The incoherent image may be incident either on the same face of the crystal as the writing beams, or on the opposite face. When the absorption coefficients at the writing and image-bearing beam wavelengths give rise to significant depth nonuniformity within the crystal, these two cases will have distinct wavelength-matching conditions for response optimization [7.8].

In the *grating inhibition mode* (GIM), shown schematically in Fig. 7.4, the crystal is pre-illuminated with the incoherent image-bearing beam prior to grating formation. This serves to selectively decay (enhance) the applied transverse electric field in exposed (unexposed) regions of the crystal. After this pre-exposure, the coherent writing beams are then allowed to interfere within the crystal, causing grating formation with spatially varying efficiency due to significant differences in the local effective applied field.

In the *simultaneous erasure/writing mode* (SEWM), shown schematically in Fig. 7.5, the incoherent image modulation, the coherent grating formation

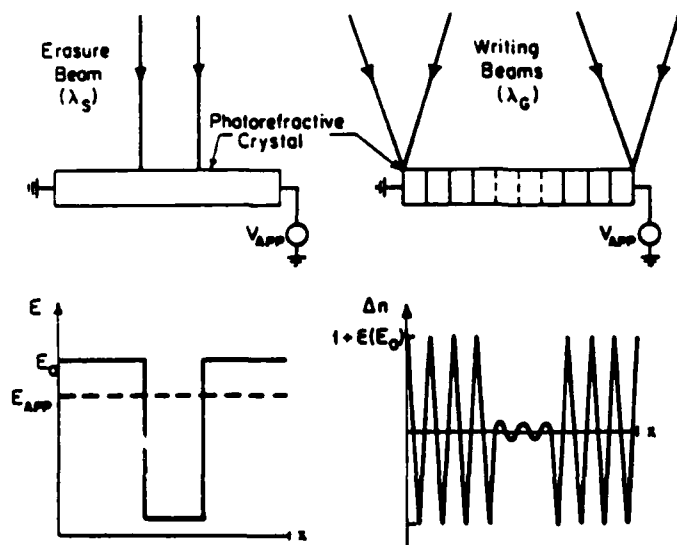


Fig. 7.4. PICOC in the grating inhibition mode (GIM), in which the carrier grating is recorded after the incoherent image-bearing signal

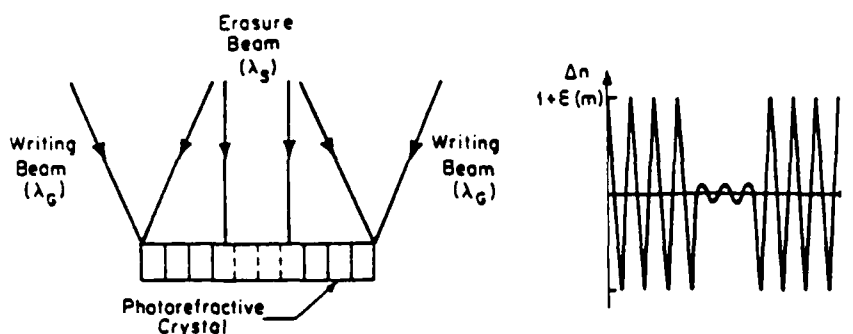


Fig. 7.5. PICOC in the simultaneous erasure/writing mode (SEWM), in which the carrier grating and the incoherent image-bearing signal are recorded simultaneously

process, and the readout function are performed simultaneously. A stable image is transcribed after the space-charge field has reached steady state.

A further distinction can be made in the operating modes between strictly cyclic exposure/readout, with an upper limit on the readout time interval, and operation in which prolonged readout times are required. Cyclic readout can be achieved by any one of the three sequencing modes introduced above, and the sensitometry requirement for achieving good quality images can be expressed in terms of optical exposure (i.e., optical energy per unit area). When prolonged readout is required, degradation of the stored space-charge profile can occur, due either to dark current or erasure induced by the optical readout beam. The most appropriate sequencing mode for prolonged readout is the simultaneous erasure/writing mode (SEWM) because it constantly regenerates the space-charge field profile. However, the sensitometry requirement for this latter mode is better expressed in terms of optical power rather than optical exposure, assuming readout time intervals long compared with the time required to achieve a stable steady state readout image. In exchange for this optical energy penalty, SEWM offers a considerably simplified experimental configuration with no need for temporal sequencing, a much greater tolerance for photorefractive crystals with increased dark conductivity, and readout of essentially unlimited duration. Readout light beams of much shorter wavelength and/or much higher intensities can be accommodated in SEWM without incurring unacceptable erasure of the charge pattern. Because of its experimental convenience and analytical simplicity, SEWM is emphasized in this chapter, with more detailed discussion of GEM and GIM given later in Sect. 7.4.4.

Optical Implementations. The original implementation of PICOC described by Kamshilin and Petrov [7.6] is a modification of the nondegenerate four-wave mixing geometry to include simultaneous exposure by an incoherent image-bearing beam. This configuration requires a readout wavelength separate and distinct from the coherent grating writing wavelength, which then allows the readout wavelength to be selected for significantly reduced grating erasure

rates. Thus, the grating inhibition mode (GIM) and the grating erasure mode (GEM) are best implemented in this configuration. However, the optical alignment is more intricate with this architecture than it is with the degenerate four-wave mixing geometry introduced next, since the Bragg angle of the read-out beam will not be the same as the Bragg angle of the coherent writing beams (shown in Fig. 7.6).

An alternative optical implementation is a modification of the conventional degenerate four-wave mixing geometry to include simultaneous exposure by an incoherent image-bearing beam, as shown in Fig. 7.7. This implementation has the advantage of extremely easy optical alignment, as the readout beam is readily Bragg aligned by retroreflecting one of the two coherent grating writing beams. It has the disadvantage that the readout beam, being at the same wavelength as the coherent grating writing beams, must erase the grating structure being probed at rates comparable to the writing process, assuming a readout light intensity comparable to the writing light intensities. Thus this implementation can be utilized for the simultaneous erasure/writing mode (SEWM), and can be adopted for the grating erasure mode (GEM) and the grating inhibition mode (GIM) only by significantly reducing the probe beam intensity, with a correspondingly reduced readout signal intensity.

As a specific example of the PICOC process, consider a degenerate four-wave mixing configuration in which the coherent writing beams and the incoherent image-bearing beam were made to illuminate the same face of a 1.3 mm thick crystal of bismuth silicon oxide, obtained from Crystal Technology, Inc. An electric field of 4 kV/cm was applied along the $\langle 110 \rangle$ axis, as shown in

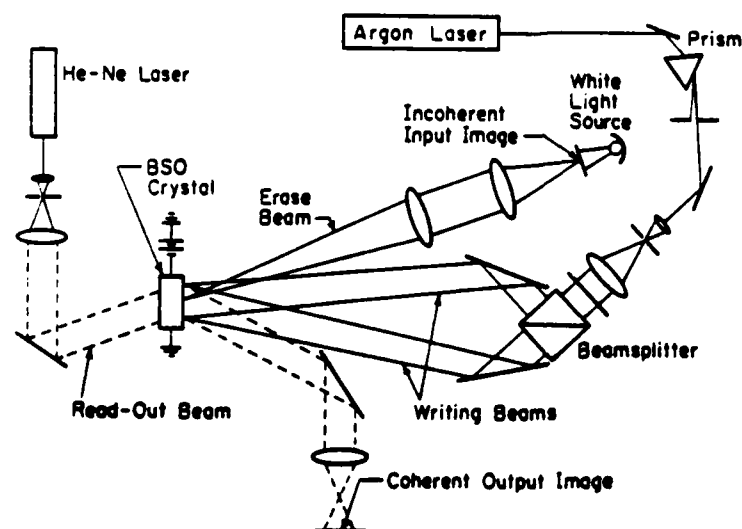


Fig. 7.6. Nondegenerate four-wave mixing architecture to perform the photorefractive incoherent-to-coherent optical conversion

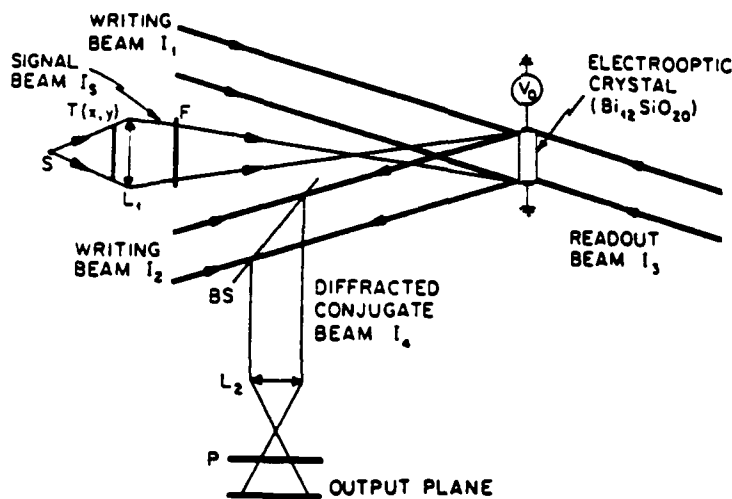


Fig. 7.7. Degenerate four-wave mixing architecture to perform the photorefractive incoherent-to-coherent optical conversion

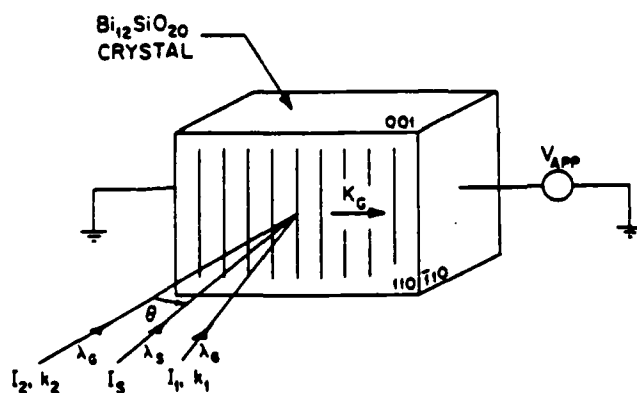


Fig. 7.8. Bismuth silicon oxide crystal orientation for the PICOC transverse electrooptic configuration and recording geometry. The volume holographic grating with wave vector K_G is formed by the coherent writing beams I_1 and I_2 , and the incoherent image information is encoded on beam I_3

Fig. 7.8. A 300 cycles/mm grating was written by the 515 nm line of an argon ion laser with the grating wave vector oriented parallel to the applied bias field to maximize the diffraction efficiency. The image-bearing light source was either a xenon arc lamp, a tungsten lamp, or the 488 nm line of the argon ion laser. The average coherent grating intensity was 0.4 mW/cm^2 and the image-bearing light intensity was typically 8.0 mW/cm^2 . The coherent grating writing beams were polarized orthogonal to the applied electric field. A polarizer was inserted at the output to minimize coherent optical scatter from the crystal [7.15].

Sample converted images obtained from two binary transparencies (a spoke target and an U.S. Air Force resolution target) and from two black-and-white slides with continuous tone gray scale are shown in Fig. 7.9. The original transparency and its converted image have reversed contrast, as shown in this figure and as explained by Fig. 7.2. An approximate resolution of 15 line pairs/mm (determined from the resolution target image) was achieved without optimizing factors such as the Bragg readout condition. Such optimization results in striking enhancements of the resolution to of order 50 line pairs/mm, as discussed in Sect. 7.5 below. Similar images of comparable quality have also been recorded in a bismuth silicon oxide crystal in which the $\langle 001 \rangle$ axis is aligned parallel to the coherent grating wave vector and to the applied bias electric field.

Having reviewed in broadest terms the physical principles and modes of operation of the PICOC process, let us now delineate the scope and limitations



Fig. 7.9a-d. Examples of the conversion of binary and gray-level transparencies: (a) spoke target, (b) U.S. Air Force resolution target, (c) airplane, and (d) an incoherent student

of the analytical model, as given in the next section, in preparation for more detailed studies of the recording process in Sect. 7.4 and the readout process in Sect. 7.5.

7.3 Delineation of the Analytical Model

A reasonably accurate and complete model of the photorefractive incoherent-to-coherent optical conversion process is based upon a set of equations that govern trap and electron balance, electron transport, and the buildup of a space-charge field, as detailed in Sect. 7.4.1. This model exhibits both nonlocal response due to charge transport and striking nonlinearities. No analytical solution has yet been identified that is broadly applicable to the full range of important experiments (e.g., experiments involving large modulation depths and photo-induced variations in the recombination time). Numerical solutions are certainly feasible, but have not yet been fully explored. To help refine physical insight into the conversion process, an approximate model capable of analytic solution needs to be defined, but its interpretation must be tempered with careful attention to its limitations.

Two such approximate solution models are identified herein. The first approach, the "constant recombination time approximation," is based upon the analytical studies of *Moharam et al.* [7.16], and was presented previously by the authors [7.8]. This approach is discussed in sufficient detail in Sect. 7.3.1 to enable comparison with the second approach, the "perturbation series approximation," as introduced in Sect. 7.3.2. These two approaches are compared and contrasted in this section. The perturbation series approximation is then used as the basis for continued discussion of PICOC performance characteristics throughout the remainder of the chapter, and is detailed more fully in Sect. 7.4.1.

7.3.1 Constant Recombination Time Approximation

One such approximate model of the PICOC process [7.8] evolves from analytic solutions of single spatial frequency grating recording as derived by *Moharam et al.* [7.16]. This model approximates quite well the nonlinearity of the conversion process, but it is limited in scope to steady state behavior. Hence the analytical solutions of this model (derived in [7.8]) are suitable only for studying the simultaneous erasure/writing mode (SEWM) in the steady state regime, and cannot cope with the grating erasure mode (GEM) and the grating inhibition mode (GIM) response. A further limitation of this approximation is its assumption of a constant recombination time (which is equivalent to assuming an infinite trap-limited saturation field strength), and so this will hereinafter be referred to as the "constant recombination time" model. A more accurate study of the photorefractive effect, as defined by a set of photoconductivity equations, allows for photoinduced variations of the recombination time from its nominal value. Although these variations may be quite small in amplitude compared with the nominal value, they nevertheless provide a very important mecha-

nism that can significantly modify the space-charge field for certain recording configurations.

A key assumption of the constant recombination time model is the absence of self-diffraction. Self-diffraction is the process by which the grating written at the entrance of the photorefractive crystal diffracts a portion of the coherent writing beams, modifying the interference pattern and hence the grating that is recorded deeper in the crystal [7.13, 17, 37]. This process can be neglected for sufficiently thin crystals and low diffraction efficiencies (e.g., 2 mm of bismuth silicon oxide) and allows a considerable simplification of the analysis. If alternative crystals with significantly higher electrooptic coefficients and/or thicker crystals are used, then the diffraction efficiency would increase, self-diffraction effects would be far more pronounced, and thus the mathematical framework described in this chapter would need to be modified.

7.3.2 Perturbation Series Approximation

By recasting the chosen set of photorefractive equations in Fourier transform space, additional physical insights into the transcription process can be achieved which lead naturally to a perturbation series formulation of the conversion process. Analytic solutions can be derived for the first few terms of this series without restricting such solutions to the steady state regime, and so define an additional approximate model of the PICOC process. Limiting the analysis to the first few terms means that the strong nonlinearities that characterize the incoherent-to-coherent optical conversion process are estimated at best, but the most compelling advantage of this approach is its ability to model the temporal evolution of the space-charge field. Study of the temporal behavior is crucial to the analysis of the grating erasure mode (GEM) and the grating inhibition mode (GIM). In this chapter, we use a perturbation model that predicts the various spatial frequency components of the space-charge field.

Consider for example a coherent grating beam $I_G(x)$ of the form

$$I_G(x) = I_0(1 + m_G \cos K_G x) \quad , \quad (7.1)$$

in which m_G is the modulation depth and K_G is the wave vector associated with the coherent grating, and an incoherent signal beam $I_S(x)$ of similar form

$$I_S(x) = I_1(1 + m_S \cos K_S x) \quad , \quad (7.2)$$

in which m_S is the modulation depth and K_S is the wave vector associated with the image profile. The coordinate system is defined such that x is parallel to the applied bias electric field E_0 , which is also parallel to the coherent grating wave vector K_G ; z is orthogonal to the entrance face of the photorefractive crystal; y is defined to complete a right-handed coordinate system. Because of the nonlinearities in the recording process, the space-charge field $E(x)$ transcribed by these light beams contains spatial frequencies that do not exist in the original light profiles. These intermodulation terms have the form

$$E(x) = \sum_{m=-\infty}^{+\infty} \sum_{n=-\infty}^{+\infty} E_{mn} \exp[i(mK_G + nK_S)x] \quad . \quad (7.3)$$

If the recording process were perfectly linear, then only terms such as E_{10} (i.e., $m = 1$ and $n = 0$) and E_{01} ($m = 0$ and $n = 1$) would appear in which one of the two subscripts is zero, assuming light profiles as defined by (7.1.2). The nonlinearity of the recording process induces new spatial harmonics such as E_{11} which describe the modulation of the coherent grating by the incoherent image beam and hence are central to the photorefractive incoherent-to-coherent optical conversion process.

The intermodulation decomposition (7.3) provides a natural framework for a perturbation series analysis of the photorefractive incoherent-to-coherent optical conversion process in powers of the modulation depths m_G and m_S , a technique first demonstrated by *Kukhtarev et al.* as applied to the analysis of single grating transcription [7.18, 19], and extended by *Ochoa et al.* [7.20] and by *Refregier et al.* [7.21]. In practice, a typical image consists of a multiplicity of spatial frequency components, not just the single frequency signal term postulated in (7.2). Such a spectrally rich image will necessarily introduce an additional summation over the spectral harmonics in (7.3). The resultant series, if limited to terms of the form E_{11} , predicts a linear transcription of the image profile for which the response to each spatial frequency component can be evaluated separately, such that the total response is determined by summing up all harmonics. Higher order terms, such as E_{12} , describe nonlinear distortion of the image spectrum in which new image frequencies are generated that do not exist in the original incoherent image profile. These higher order terms prove to be extremely tedious to calculate. If these higher order terms contribute significantly to the conversion process, then alternative methods of analysis such as numerical modeling are recommended. The analysis presented in this chapter concentrates on the E_{11} term.

Analytical expressions defining the temporal evolution of the first few harmonic terms of (7.3) are readily derived. As shown in Appendix 7.A and in [7.22], specific perturbation terms for the simultaneous erasure/writing mode (SEWM) in the saturation limit can be approximated by

$$E_{01} = -\frac{1}{2}m_S^{\text{eff}}E_0 \quad (7.4)$$

$$E_{10} = -\frac{1}{2}m_G^{\text{eff}}E_0 \quad \text{and} \quad (7.5)$$

$$E_{11} = \frac{1}{2}m_S^{\text{eff}}m_G^{\text{eff}}E_0 \quad (7.6)$$

in which E_0 is the applied bias field shown in Fig. 7.8, and m_G^{eff} and m_S^{eff} are effective modulation depths to be defined in the next section.

If bismuth silicon oxide is chosen as the photorefractive medium of interest, then the refractive index modulation $\Delta n(x)$ induced along a principal electrooptic axis of the crystal by the linear electrooptic effect is proportional to the induced space-charge field $E(x)$ [7.23]. Each of the spatially periodic terms in the space-charge field (7.3) thus induces a distinct volume phase grating through the linear electrooptic effect. Such a superposition of phase gratings

diffraction an incident collimated readout laser beam into a multiplicity of discrete beams, as shown in Fig. 7.10.

Another fundamental insight into the PICOC process that evolves naturally from the perturbation series approach is the existence of a one-to-one mapping of E_{mn} , the spatial frequency components of the space-charge field, into I_{mn} , the diffraction orders and suborders of the spatially modulated readout light. This identification presumes weak diffraction efficiencies, as have thus far been typical of most PICOC implementations. In the far field diffraction limit, shown in Fig. 7.10, the spatially modulated readout light profile $I_R(x)$ decomposes into the usual set of discrete diffraction orders associated with the coherent grating, labeled by subscript m , while each of these orders further decomposes into subharmonics associated with the incoherent image signal, labeled by subscript n . This mapping offers a most convenient method to test the behavior of distinct spatial frequency components of the space-charge field E_{mn} during the conversion process.

For typical coherent grating spatial frequencies and typical crystal thicknesses, the volume phase gratings in the photorefractive crystal are recorded deep within the Bragg regime. Therefore the optical readout process can respond at most to one diffraction order such as I_{10} and its immediate subharmonics such as I_{11} , while excluding other diffraction orders and associated sidebands such as I_{01} . Thus for the illumination profiles $I_G(x)$ and $I_S(x)$ of the form (7.2,3), and assuming selective diffraction into only the +1st diffraction order and its immediate sidebands, a typical form of the modulated readout light profile $I_R(x)$ would be

$$I_R(x) = I_{10} + I_{11} \cos(K_S x) + \text{higher image harmonics} \quad (7.7)$$

Furthermore, assuming perfect Bragg alignment for the I_{10} term, and assuming image frequencies K_S small enough to avoid Bragg misalignment of the immediately adjacent subharmonics, the diffracted intensity terms I_{10} and I_{11} are proportional to the squares of the corresponding space-charge field components E_{10} and E_{11} for sufficiently low diffraction efficiencies. Combining (7.4-7) and approximating for small modulation depths m_S^{eff} gives

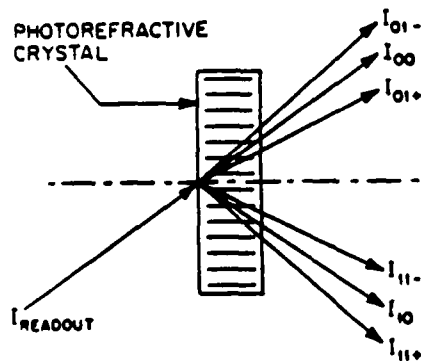


Fig. 7.10. Far-field diffraction pattern created by the conversion of an incoherent grating into its coherent replicas, showing the mapping of spatial harmonics of the space-charge field into diffraction orders of the modulated readout light beam

$$I_R(x) \propto (m_G^{\text{eff}})^2 [1 - 4m_S^{\text{eff}} \cos(K_S x)] \quad (7.8)$$

in which higher order terms have been neglected. Thus we can see from (7.8) that the incoherent image has been transcribed onto the coherent readout beam with a reversal of image contrast. The mapping of E_{mn} into I_{mn} also enables the possibility of spatial filtering of the image beam to change the negative image into a positive image (by a schlieren technique), and to improve the contrast ratio without suffering an associated reduction in image intensity.

The perturbation series approach that leads to (7.8) is most effective whenever the higher spatial harmonics in (7.7) can be neglected in favor of the lowest harmonics, and this applies whenever the modulation depths m_G^{eff} and m_S^{eff} are sufficiently small (or the spatial frequencies are sufficiently large). Such small modulation depths occur only for very restrictive conditions which seldom match the experimental conditions. Therefore the scope of the perturbation solution is limited, which reflects as much a fundamental difficulty with the PICOC process as it does a difficulty with the analytic technique. The PICOC process requires a nonlinear response to generate the modulation of the coherent grating by the incoherent image field, and hence strong levels of nonlinear distortion inextricably occur with significant levels of modulated light intensity. Restricting the attention to the lowest order spatial harmonics underestimates the nonlinearities, but leads to simple and powerful analytical expressions for the temporal response that contain considerable physical insight into the PICOC process and are unobtainable by any other analytical technique.

With these comments as backdrop, we can now proceed with a more detailed study of the recording process in Sect. 7.4 and of the readout process in Sect. 7.5.

7.4 The Recording Process

During the recording process, a space-charge electric field $E(x)$ is formed in the photorefractive crystal in response to the combined illumination by a coherent grating light beam $I_G(x)$ and an incoherent image beam $I_S(x)$. A physical model describing this transcription process is reviewed in Sect. 7.4.1, and particular numerical and analytical solutions are listed for the simultaneous erasure/writing mode (SEWM). The implications of these results on the nonlinear transfer function of the conversion process are explored in Sect. 7.4.2. Issues affecting the resolution of the recording process are discussed in Sect. 7.4.3. Finally, the temporal evolution characteristics are studied in Sect. 7.4.4.

7.4.1 Physical Model and Sample Solutions

The model of photoconductivity that is most frequently selected to describe holographic grating formation in photorefractive crystals assumes a single mobile charge species (electrons) and a single trapping level [7.13], although more

intricate models involving hole transport [7.24, 25] and multiple trapping levels [7.26] have occasionally been proposed to achieve a better fit with particular experiments. The simple single trapping level/single mobile charge species model which has been chosen to describe the PICOC recording process consists of the following set of equations:

$$\frac{\partial N_D^+}{\partial t} = [S_G I_G(x, t) + S_S I_S(x, t) + \beta](N_D - N_D^+) - \gamma_R N_D^+ n \quad (7.9)$$

$$\frac{\partial n}{\partial t} = \frac{\partial N_D^+}{\partial t} + \frac{1}{e} \frac{\partial j}{\partial x} \quad (7.10)$$

$$\frac{\partial E}{\partial x} = \frac{e}{\epsilon \epsilon_0} (N_D^+ - n - N_A^-) \quad (7.11)$$

$$j = en\mu E + kT\mu \frac{\partial n}{\partial x} \quad (7.12)$$

in which

N_D	is the total concentration of donor-like trapping centers,
$N_D^+_{eq}$	is the concentration of ionized donor-like trapping centers in quasi-equilibrium under dark conditions,
$N_D^+(x, t)$	is the concentration of ionized donor-like trapping centers,
N_A^-	is the concentration of negatively charged acceptor-like centers that compensate for the charge $N_D^+_{eq}$ under dark thermal quasi-equilibrium conditions (N_A^- is a constant of the crystal),
$n(x, t)$	is the concentration of electrons in the conduction band,
$E(x, t)$	is the internal space-charge electric field,
e	is a positive number with magnitude equal to the electronic charge,
S_G	is the cross section of photo-ionization for the coherent grating beams with wavelength λ_G divided by the photon energy, hereinafter referred to as a photo-ionization cross section,
S_S	is the cross section of photo-ionization for the incoherent image beam with wavelength λ_S divided by the photon energy, hereinafter referred to as a photo-ionization cross section,
β	is the thermal generation rate of electrons into the conduction band,
γ_R	is the carrier recombination constant,
$I_G(x, t)$	is the optical intensity profile for the coherent grating,
$I_S(x, t)$	is the optical intensity profile for the incoherent image,
$j(x, t)$	is the current density in the crystal,

μ	is a positive number with magnitude equal to the charge carrier mobility,
k	is Boltzmann's constant,
T	is the absolute temperature of the crystal,
ϵ	is the static dielectric constant of the crystal, and
ϵ_0	is the free space electric permeability.

Rationalized MKS units are assumed. The x coordinate is transverse to the nominal light propagation direction, and parallel to the applied bias field direction; t denotes time. Equation (7.9) is the rate equation describing the excitation of electrons into the conduction band and subsequent recombination into traps. Equation (7.10) states the conservation of electric charge. Equation (7.11) is Maxwell's first equation for the electric field. Equation (7.12) defines the current density in terms of drift and diffusion components. The material parameters for bismuth silicon oxide assumed in the numerical calculations are listed in Table 7.1 taken from the works of *Tanguay* [7.27] and *Valley and Klein* [7.28]. We wish to solve the equations for the space-charge electric field $E(x)$ which is induced by exposure to the input optical intensities $I_G(x)$ and $I_S(x)$.

Single Grating Response. Consider first the case of a single spatial frequency grating $I_G(x)$, as defined by (7.1) of the previous section, recorded in the absence of an incoherent image beam $I_S(x)$. Because of the nonlinearity of the recording process, the key variables of the photoconductivity model consist of a superposition of harmonics of the incident light beam, i.e.,

$$n(x, t) = \sum_{m=-\infty}^{+\infty} n_m(t) \exp(imK_G x) \quad , \quad (7.13)$$

$$N_D^+(x, t) = \sum_{m=-\infty}^{+\infty} N_{Dm}^+(t) \exp(imK_G x) \quad , \quad (7.14)$$

$$E(x, t) = \sum_{m=-\infty}^{+\infty} E_m(t) \exp(imK_G x) \quad , \quad \text{and} \quad (7.15)$$

$$j(x, t) = \sum_{m=-\infty}^{+\infty} j_m(t) \exp(imK_G x) \quad . \quad (7.16)$$

These harmonic decompositions can be substituted into (7.9–12), resulting in a set of coupled differential equations defining the temporal evolution of each harmonic component. This coupled set of equations can either be integrated numerically, as demonstrated by *Moharam et al.* [7.16] on a reduced subset of the equations, or else be solved approximately by perturbation series methods.

Consider first a perturbation series expansion in powers of the modulation depth m_G . Analytical expressions for the first order terms have been derived

Table 7.1. Material parameters of bismuth silicon oxide ($\text{Bi}_{12}\text{SiO}_{20}$)

Parameter	Symbol	Value	Reference
Mobility	μ	$0.03 \text{ cm}^2/\text{Vs}$	[7.28]
Carrier lifetime	τ	$5 \times 10^{-6} \text{ s}$	[7.28]
Donor-like trap density	N_D	10^{19} cm^{-3}	[7.28]
Dark ionized trap density	$N_{D_{eq}}^+$	10^{16} cm^{-3}	[7.28]
Recombination coefficient	γ_R	$2 \times 10^{11} \text{ cm}^3/\text{s}$	[7.28]
Dielectric constant	ϵ	56	[7.28]

	Symbol	488 nm	515 nm	633 nm	Units	Reference
Index of refraction	n_0	2.650	2.615	2.530		[7.27]
Optical absorption	α	7.0	2.8	0.6	cm^{-1}	[7.27]
Electrooptic coefficient	r_{41}	4.52	4.51	4.41	pm/V	[7.27]
Photo-ionization cross section (divided by photon energy)	S_G	—	0.42	—	cm^2/Joule	[7.28]

Note: A photo-ionization cross section S_3 at 488 nm has been estimated to be of order $1 \text{ cm}^2/\text{Joule}$ by $S_3 = (\alpha_3 \lambda_3 / \alpha_G \lambda_G) S_G$, assuming identical quantum efficiency at 488 and 515 nm. However see the discussion in Sect. 7.4.2 and Sprague [7.30].

by Kukhtarev et al. [7.18,19], such that in the steady state regime the first spatial harmonic component E_1 of the space-charge field is given by

$$E_1 = -\frac{1}{2} m_G \frac{E_0 + iE_D(K_G)}{D(K_G)} \quad (7.17)$$

to first order in the modulation depth m_G , in which E_0 is the applied bias electric field strength. In (7.17), the denominator function $D(K_G)$ is defined by

$$D(K_G) = 1 - i \frac{E_0 + iE_D(K_G)}{E_q(K_G)} \quad (7.18)$$

and the diffusion field $E_D(K_G)$ and the trap-limited saturation field strength $E_q(K_G)$ are defined by

$$E_D = \frac{kTK_G}{e} \quad (7.19)$$

$$E_q = \frac{eN_{D_{eq}}^+}{\epsilon\epsilon_0 K_G} \quad (7.20)$$

The trap-limited saturation field strength E_q defines the highest electric field that can be generated by a sinusoidal charge distribution with maximum charge density of $N_{D_{eq}}^+ = N_A^-$. For reduced values of N_A^- (and hence of $N_{D_{eq}}^+$), the amount of space charge and resulting space-charge field strength is limited as described by (7.17, 18). Considerable variation has been reported in the literature in estimates of the equilibrium donor-like trap density $N_{D_{eq}}^+$ for bismuth silicon oxide [7.29], which directly affects the estimate of the saturation field strength E_q , and hence the upper limit on the space-charge field strength.

The first order perturbation analysis is accurate only in the limit of low modulation depths m_G , whereas many gratings are written with the highest possible modulation depths. Solutions accurate at higher modulation depths can be obtained by numerical methods, with sample solutions presented in Fig. 7.11 showing the space-charge field profiles induced by a single grating frequency in the steady state regime for various applied bias fields and various grating frequencies. A modulation depth of $m_G = 0.99$ and the material parameters listed in Table 7.1 are assumed for all curves. Note that even when the writing light profile is cosinusoidal, the resulting space-charge field profiles exhibit significant distortion because of the nonlinearity of the recording process.

In practice, only one of the multiple spatial harmonics of the coherent grating wave vector K_G can dominate the optical readout process because the grating is typically recorded very deeply into the Bragg regime. Figure 7.12 therefore shows numerical solutions for the strength of just the first spatial harmonic component of the space-charge field E_1 as a function of the modulation depth m_G (labeled in the figure as m_G^{eff} in anticipation of the two grating transcription discussion that follows, but to be interpreted for now as m_G). The family of curves shown in Fig. 7.12 is parametrized by the ratio E_q/E_0 , which scales the relative magnitudes of the trap-limited saturation field to the applied bias field. Those portions of the response curves in Fig. 7.12 that can be approximated by a straight line, namely for low levels of the modulation depth m_G up to about 0.5 or so, are the regions in which the linear approximation of Kukhtarev et al. [7.18, 19] most accurately describes the transcription process. In anticipation of the discussion of nonlinear PICOC response given in Sect. 7.4.2, we find that high effective modulation depths m_G^{eff} (defined by (7.23) below) for which the linear approximation breaks down are associated with weak levels of average incoherent image intensity I_1 compared with the

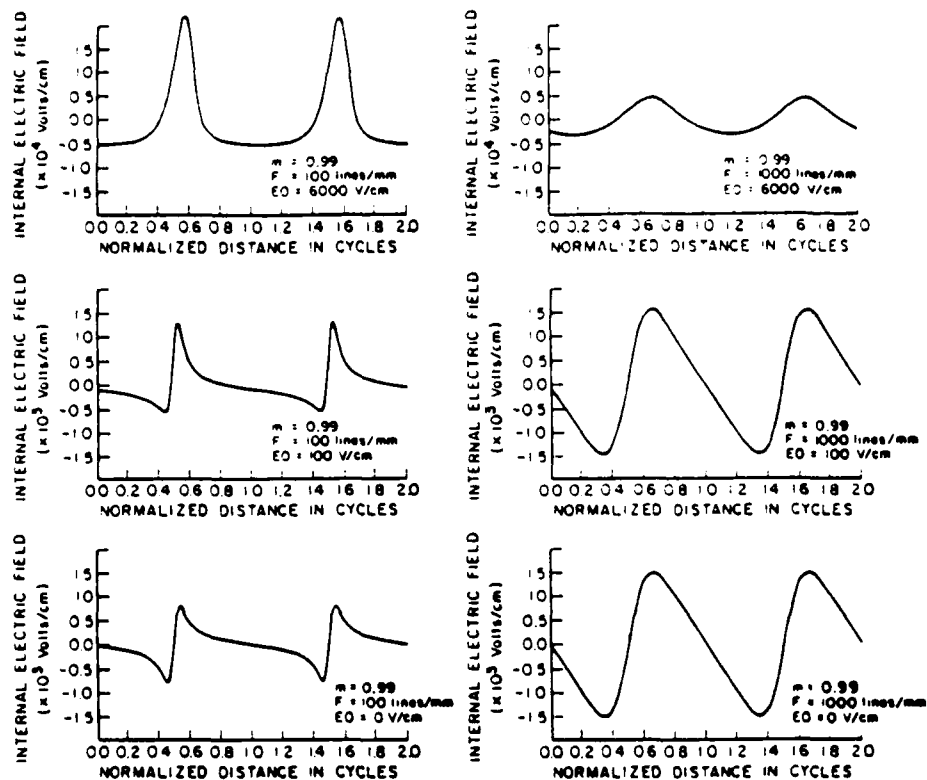


Fig. 7.11. Sample profiles of the space-charge field generated by a single frequency (unmodulated) grating, as determined by numerical solution of the photoconductive equations

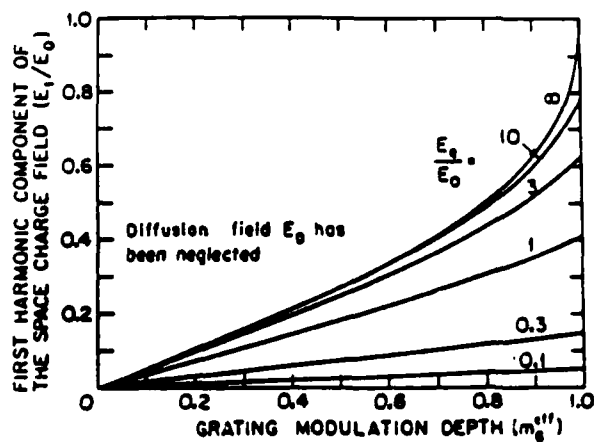


Fig. 7.12. First spatial harmonic components of the space-charge field profile as a function of the modulation depth of the coherent carrier grating for various levels of the equilibrium donor-like trap density $N_{D,eq}^+$

average coherent grating light intensity I_0 , while low modulation depths for which the linear approximation is most accurate occur for strong levels of average incoherent light illumination I_1 .

Two Grating Transcription. When an incoherent image beam is present simultaneously with the coherent grating, as in the simultaneous erasure/writing mode (SEWM), then additional terms corresponding to harmonics of the image frequencies must be included in the space-charge field. Consider for example an image consisting of just one sinusoidal component, as given by (7.2), with a resulting space-charge field decomposition of the form of (7.3). Double harmonic decompositions analogous to (7.3) can be assumed for each of the variables in the photoconductive equations (i.e., the ionized trap density N_D^+ , the electron density n , and the current density j).

The resulting set of coupled differential equations can be solved either numerically or by a perturbation series analysis with respect to this harmonic decomposition, with the latter being the approach adopted herein [7.22]. The perturbation series approach leads to analytical expressions for the linear terms E_{10} and E_{01} of the form

$$E_{10} = -\frac{1}{2}m_G \frac{S_G I_0}{S_G I_0 + S_S I_1} \frac{E_0 + iE_D(K_G)}{D(K_G)} \quad (7.21)$$

$$E_{01} = -\frac{1}{2}m_S \frac{S_S I_1}{S_G I_0 + S_S I_1} \frac{E_0 + iE_D(K_S)}{D(K_S)}, \quad (7.22)$$

as given in Appendix 7.A and detailed in [7.22]. These expressions correspond very closely to (7.17), the first order field expression derived by Kukhtarev et al. [7.18, 19]. The effective modulation depths m_G^{eff} and m_S^{eff} in (7.21, 22) which account for the reduction of the original modulation depths m_G and m_S by the presence of both the incoherent image and the coherent grating beams are defined by

$$m_G^{\text{eff}} = m_G \frac{S_G I_0}{S_G I_0 + S_S I_1} \quad (7.23)$$

$$m_S^{\text{eff}} = m_S \frac{S_S I_1}{S_G I_0 + S_S I_1} \quad (7.24)$$

The expressions (7.21, 22) can be simplified over broad operating regions as follows. For spatial frequencies less than of order 200 cycles/mm and bias fields over 2 kV/cm, and assuming the equilibrium donor-like trap density $N_{D\text{eq}}^+$ given in Table 7.1, the denominator terms $D(K_G)$ and $D(K_S)$ can be approximated by unity, and the diffusion field E_D can be neglected in favor of the applied bias field. For higher spatial frequencies, the denominator factors $D(K_G)$ and $D(K_S)$ and the diffusion field E_D primarily contribute a phase shift to the coherent grating's charge distribution, with negligible degradation of its magnitude. (This assertion is justified in the discussion on material limitations in

Sect. 7.4.3.) Thus the lowest harmonics of the space-charge field can be approximated by

$$E_{10} = -\frac{1}{2}m_G^{\text{eff}}E_0 \quad (7.25)$$

$$E_{01} = -\frac{1}{2}m_S^{\text{eff}}E_0 \quad (7.26)$$

By invoking similar approximations, the magnitude of the intermodulation term E_{11} , which is crucial to the incoherent-to-coherent optical conversion process, is well approximated by

$$E_{11} = \frac{1}{2}m_G^{\text{eff}}m_S^{\text{eff}}E_0 \quad (7.27)$$

as discussed in Sect. 7.4.3 and in [7.22]. An identical expression for E_{11} was derived by Marrakchi et al. [7.8], starting from the constant recombination time approximation discussed in Sect. 7.3.1. This derivation involves an additional linearization of the response in the limit of low modulation depths m_G^{eff} and m_S^{eff} .

In addition to the magnitude expressed by (7.27), the perturbation series analysis predicts that the E_{11} field is phase shifted with respect to the incident coherent grating. For steady state response in SEWM, this phase shift does not significantly impact the performance of the conversion process, assuming that it remains reasonably constant over the recording bandwidth; for temporal response it can significantly degrade the usefulness of PICOC for particular optical processing architectures.

Equations (7.25–27) indicate that the SEWM response in the steady state limit is predominantly governed by the reduced modulation depths m_G^{eff} and m_S^{eff} . The consequent impact on the overall readout image light intensity and on the modulation transfer characteristic is discussed in the nonlinear transfer response analysis, presented next.

7.4.2 Nonlinear Transfer Response

The image transfer for PICOC involves a performance trade-off between two competing mechanisms: the contrast ratio (which involves the ratio of I_{11} and I_{10}) improves steadily with increasing intensities of incoherent image-bearing light (in the absence of spatial filtering), but at the same time the average intensity (which involves I_{10} alone) steadily declines with increasing incoherent intensity because the uniform background in the incoherent light erases the carrier grating pattern in the photorefractive crystal.

In many optical information processing applications, optimization of the image contrast ratio is desirable within the image intensity constraints implied by the performance trade-off described above. In other types of signal processing applications such as correlation with a Vander Lugt filter, the dc image content contained in the I_{10} diffraction component does not contribute to the processing, or can be readily modified by spatial filtering, whereas maximizing

the intensity of nonzero spatial frequency components such as I_{11} is critical to good conversion performance. For these cases, an optimum level of incoherent image-bearing beam intensity exists for maximizing the I_{11} component, beyond which the I_{11} component decays because of space-charge erasure. Therefore the following study of the nonlinear transfer response includes explicit consideration of the I_{11} diffraction order term.

To obtain a quantitative estimate of these effects, consider as a representative example the simultaneous erasure/writing mode (SEWM) in the steady state regime, with a single spatial frequency coherent grating given by (7.1), and a single spatial frequency incoherent image profile given by (7.2). One possible method of assessing the image transfer response is to determine the modulation depth of the readout image as a function of the input (incoherent and coherent) light characteristics. A convenient parameter that describes these characteristics is the product BR , in which $R = I_1/I_0$ is the ratio of the average incoherent image light to the average coherent grating light intensity levels, and $B = S_S/S_G$ is the ratio of the photoconductive sensitivity of the incoherent image beam with respect to the coherent grating beam. With these definitions, the effective modulation depths m_G^{eff} and m_S^{eff} defined by (7.25, 26) can be expressed as

$$m_G^{\text{eff}} = \frac{m_G}{1 + BR} \quad \text{and} \quad (7.28)$$

$$m_S^{\text{eff}} = \frac{m_S BR}{1 + BR}$$

The diffracted intensity I_{10} is directly proportional to the square of m_G^{eff} , while the modulation depth of the readout image is similarly proportional to m_S^{eff} . The image transfer response is plotted in Fig. 7.13, which shows the improve-

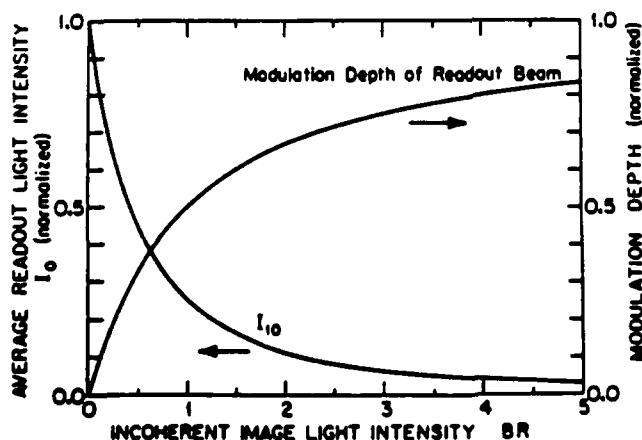


Fig. 7.13. Normalized diffraction efficiency of the I_{10} beam (uniform erasure), and the modulation transfer function for modulated erasure, as a function of the intensity ratio R

ment in modulation depth of the readout light beam with increasing levels of incoherent image illumination, but with concomitant decay of the average transferred image intensity I_{10} .

An unusual but intuitive feature of the photorefractive incoherent-to-coherent conversion process (as opposed to more typical linear spatial light modulation techniques) is that the image quality is primarily determined by the *ratio* of the incoherent image intensity to the coherent grating intensity, not by the *sum* of these intensities (assuming negligible dark conductivity β in the photogeneration rate equation (7.9)). Remember, however, that this analysis assumes a recording process that has already reached steady state, and hence does not address the question of the time required to reach saturation, which is indeed a function of the total light intensity. Temporal response issues are considered in Sect. 7.4.4 below.

To test the predictions of this nonlinear transfer model, the erasure of the readout beam's first diffraction order I_{10} in response to spatially uniform incoherent illumination has been measured by the nondegenerate four wave mixing configuration discussed in Sect. 7.2, with results as shown in Fig. 7.14. The grating was written with the 515 nm line of an argon ion laser, while the 488 nm line was used to simulate the spatially uniform image beam. For comparison, predictions of two different models of the conversion process are included in Fig. 7.14. The curve labeled "linear approximation" corresponds to the perturbation analysis term E_{10} presented in this chapter, while the second curve labeled "constant recombination time approximation" derives from (7.9) of the previously published model [7.8], which in turn is based upon the saturation regime model of *Moharam et al.* [7.16]. The linear approximation model has been scaled in intensity to match the experimental points at $R = 0.6$, and the constant recombination time approximation (Moharam's analytical solution) has been scaled to converge with the linear approximation for very large beam ratios R .

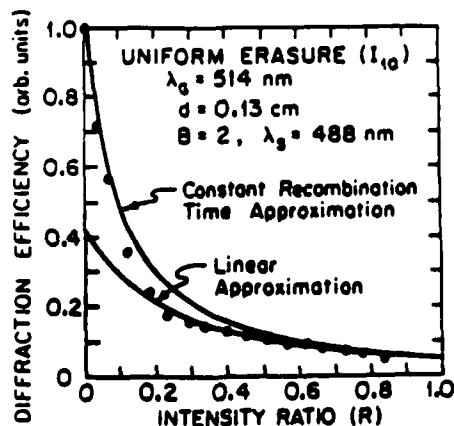


Fig. 7.14. Experimental diffraction efficiency of the I_{10} beam (uniform erasure) as a function of the intensity ratio R . Corresponding theoretical predictions for the linearized model and the constant recombination time model are also shown for comparison

Note that the experimental data points and both models converge reasonably well for high levels of incoherent image illumination, which correspond to large intensity ratios R that by (7.28) imply small effective modulation depths m_G^{eff} . Recall from Fig. 7.12 that small modulation depths correspond to the most accurate region of the linearized models of the recording process, as assumed in the lowest order terms of the perturbation analysis (as well as in the linearized constant recombination time approximation discussed following (7.27), which was shown to yield expressions identical to the perturbation analysis). In contrast, for small intensity ratios R with concomitantly large effective modulation depths m_G^{eff} , the full nonlinear constant recombination time model as developed in [7.8] offers a significantly better recording approximation than the linearized models for the simultaneous erasure/writing mode (SEWM) in saturation, as shown clearly in Fig. 7.14 by the curve marked "constant recombination time approximation". Note in addition that the choice of scaling of the constant recombination time approximation does not automatically guarantee good agreement with the data near $R = 0$.

A more challenging test of the theory is to predict accurately the conversion response to a *sinusoidally* modulated image beam. Such a test can be performed with the nondegenerate four wave mixing geometry described previously in which a 488 nm argon ion laser line passes through a Michelson interferometer to generate a sinusoidal spatial modulation, as shown in Fig. 7.15. The sinusoidal image modulation introduces an additional diffraction order I_{11} not observed in the uniform erasure case. The diffracted light sideband intensity I_{11} has been measured as a function of the intensity ratio R , with results as shown in Fig. 7.16, and as compared against the theoretical prediction of the perturbation expansion method.

The general features of Fig. 7.16 can be understood with reference to Fig. 7.13. For low levels of image intensity, corresponding to low intensity ratios R , the coherent grating's charge pattern is not significantly erased by the

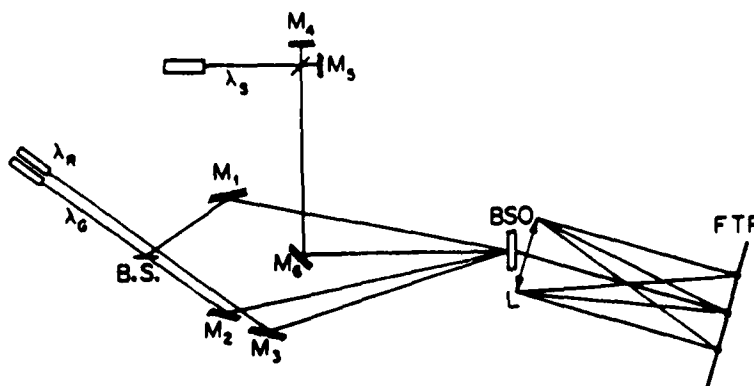


Fig. 7.15. Experimental arrangement for sensitivity and transfer function measurements, as described in detail in the text

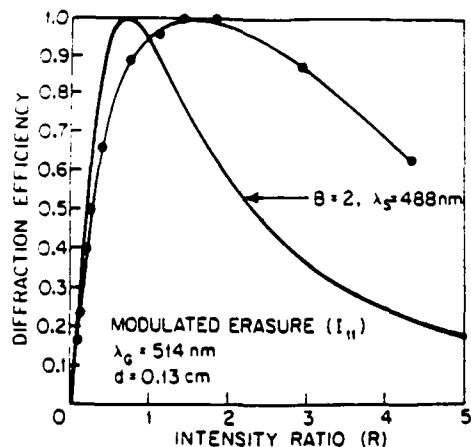


Fig. 7.16. Normalized experimental diffraction efficiency of the I_{11} beam (strongly modulated erasure) as a function of the intensity ratio R . The corresponding theoretical prediction is also shown for comparison

image beam, as indicated by the high level of I_{10} in Fig. 7.13. In this regime, increasing the incoherent image-bearing light intensity increases the transfer of image modulation onto the space-charge grating profile without destroying that profile. For high levels of incoherent light intensity, corresponding to larger intensity ratios R , the transfer of image modulation onto the coherent grating's space-charge field profile is very high as shown by the readout beam's modulation depth curve in Fig. 7.13, but the high intensity of the uniform incoherent image light I_1 strongly erases the coherent grating's space-charge profile, as shown by the I_{10} curve in Fig. 7.13. Hence the I_{11} diffracted light component, which is derived from the combination of the modulation depth and the average grating intensity I_{10} , exhibits the peaking behavior shown in Fig. 7.16.

The match between the theoretical curve in Fig. 7.16 and the experiment is not expected to be perfect because the perturbation theory requires small modulation depths m_G and m_S for reasonable accuracy, whereas the experiment is configured with the highest possible modulation depths. To impose small modulation depths in the experiment would make the diffracted I_{11} light intensity too weak to measure reliably, so a more exacting test of the theory must await numerical modeling of high modulation depth transcriptions.

An additional difficulty in achieving agreement between the theoretical models and the experiment concerns the appropriate value for the photoconductive sensitivity ratio B . For the experiment described in Fig. 7.14, in which the coherent writing wavelength is 515 nm, and the incoherent image wavelength is 488 nm, an estimate of $B = 2$ can be derived from a single photon absorption model that assumes the quantum efficiencies of the photoconductive processes for both the coherent grating beam and the incoherent image beam to be identical [7.8]. However, measurements of the photoconductive quantum efficiency reported by Sprague [7.30] show a strong dependence on the light wavelength, such that an increase in the sensitivity ratio B by a factor of 1.5 can reasonably be argued due to this dispersion of quantum efficiency. Further-

more, this factor can be expected to vary from one crystal sample to another, depending upon the detailed growth conditions. On the other hand, a reduction of the effective sensitivity ratio B of as much as 1.7 can also be argued for the crystal thickness used in this experiment because of the dispersion of the optical absorption, i.e., the ratio of coherent grating to incoherent image beam intensities changes continuously as both beams propagate through the crystal [7.8]. Because of these conflicting arguments, the nominal ratio of $B = 2$ in Figs. 7.14 and 7.16 has been assumed for the wavelengths considered. This gives a good fit for the uniform erasure I_{10} experimental points, especially for the constant recombination time approximation, but this ratio causes the predicted modulated erasure I_{11} curve to peak at a significantly lower intensity ratio R than is indicated by experiment.

In summary, the broad features of the recording nonlinearities are well understood and successfully modeled by the perturbation series approach, e.g., the decay of the coherent grating pattern with increasing amounts of image intensity, and the dependence of the I_{11} diffracted light intensity component on exposure parameters. Detailed agreement will require both further analysis to model the nonlinearities more accurately, and better information about the wavelength dispersion of the photoconductive sensitivities of the coherent grating and the image-bearing beams.

7.4.3 Spatial Resolution Issues for the Recording Process

A number of distinct factors influence the ultimate resolution achievable with the PICOC spatial light modulator. These factors can be classified as geometric, configurational, and materials related in nature. The geometric and materials related factors influence the recording of the image, and hence are reviewed in this section. The configurational factors influence the readout of the volume gratings, and as such are reviewed in the subsequent section.

Geometrical Limitations. Geometric resolution limitations derive principally from the incorporation of an incoherent imaging system in the four-wave mixing geometry, and from the finite crystal thickness d required to create a volume holographic grating. These effects are illustrated in Figs. 7.17 and 7.18. Distinctly different resolution performance is expected, depending upon the optical absorption coefficients α_G of the coherent grating light and α_S of the incoherent image light.

Figure 7.17 describes the case for low optical absorption ($\alpha_G d \ll 1$ and $\alpha_S d \ll 1$), such that the induced holographic grating has essentially uniform amplitude throughout the volume of the crystal. As can be expected from physical considerations, the optimum focal point occurs in the center of the crystal, and is not localized on the front surface of the crystal. The spatial frequency response will then be proportional to $(W/2)^{-1}$, which in turn is equal to $4n_0 F\# / d$ for the case of 1:1 imaging, in which W is the diameter of the incoherent image beam at the front surface of the crystal (as shown in Fig. 7.17), n_0 is the refractive index of the electrooptic crystal, $F\#$ is the F -

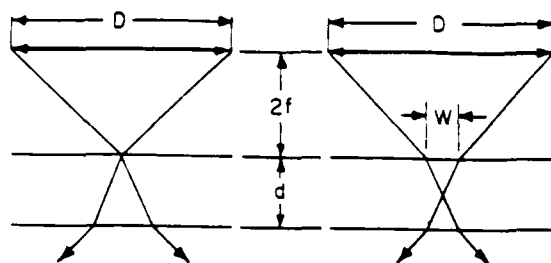
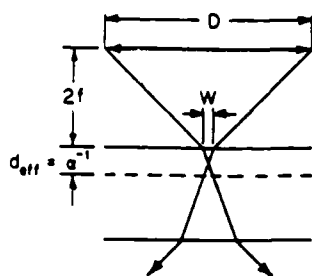


Fig. 7.17. Geometrical constraints imposed on PICOC resolution by the finite F -number of the image input optics for the case $\alpha d < 1$. A distance of $2f$ between the imaging lens and the photorefractive crystal has been assumed in this and in the next figure, corresponding to a presumed 1:1 lens magnification

CASE I: $\alpha d < 1$ $R \sim \frac{1}{(W/2)} = \frac{4nF\#}{d}$

EXAMPLE: FOR $n=2.5$, $d=1\text{mm}$, $F\#=5$,
 $R=50$ line pairs/mm



CASE II: $\alpha d \gg 1$ $R \sim \frac{1}{(W/2)} = 4nF\#\alpha$

EXAMPLE: FOR $n=2.5$, $d=1\text{mm}$, $F\#=5$, $\alpha=100\text{cm}^{-1}$
 $R=500$ line pairs/mm

Fig. 7.18. Geometrical constraints imposed on PICOC resolution by the finite F -number of the image input optics for the case $\alpha d \gg 1$

number of the incoherent imaging system, and d is the crystal thickness. For example, for $n_0 = 2.5$, $d = 1$ mm, and an F -number of 5, the resolution limit is approximately 50 cycles/mm.

In contrast, Fig. 7.18 describes the case for which $\alpha_G d \gg 1$, such that the induced holographic grating has significant amplitude only within a thin layer of thickness $d_{\text{eff}} = \alpha_G^{-1}$. In this case, the resolution is given by $4n_0 F\# \alpha_G$. For $n_0 = 2.5$, $d = 1$ mm, $\alpha_G = 100\text{cm}^{-1}$, and an F -number of 5, the resolution limit is approximately 500 cycles/mm, a factor of 10 improvement in resolution. However, the diffraction efficiency is reduced by a factor of order 100 because of the reduction in effective thickness of the grating. If the absorption coefficient α_S is chosen to be significantly larger than α_G , then the resolution will be constrained by α_S instead of α_G .

Material Limitations. An additional resolution limitation stems from material-dependent parameter constraints which influence the physics of grating formation, in particular the finite supply of compensating traps N_A^- , which is equal to the equilibrium donor-like trap density $N_{D\text{eq}}^+$. If the trap density is limited,

then the space-charge field that can be recorded is similarly limited because sufficient space charge cannot be generated to establish any higher field strengths. This limitation becomes progressively more severe at higher spatial frequencies as expressed by the $D(K_G)$ factor in (7.17), in which the trap-limited saturation space-charge field E_q is a function of the spatial frequency and the equilibrium donor-like trap density N_{Deq}^+ , as shown in (7.20). The reduction in space-charge field with an increase in spatial frequency (as exhibited by a corresponding decrease in E_q) is shown in Fig. 7.12. If an unlimited supply of compensating traps were available for establishing the space-charge field, then the saturation field E_q would be infinite and the factor $D(K_G)$ would converge to unity, indicating negligible degradation in the space-charge field. In practice, the finite level of compensating traps leads both to a reduction of the space-charge field and to a phase shift as consequences of the finite saturation field. These two effects are discussed in more detail below.

Consider for example an equilibrium donor-like trap density N_{Deq}^+ of 10^{16} cm^{-3} and a coherent grating frequency of 300 cycles/mm, which implies a saturation field E_q of order 19 kV/cm and a diffusion field E_D of order 0.5 kV/cm. For an applied bias field of 6 kV/cm, the effect of the factor $D(K_G)$ in (7.17) is only a 5% reduction in the magnitude of the induced space-charge field in the linear approximation. As these parameters are typical of PICOC operation, the resultant effect induced by this mechanism on the spatial frequency response is therefore negligible compared with alternative response degradation mechanisms such as Bragg detuning on readout, as discussed in the next section.

The addition of harmonic components to the image will in general introduce an additional amplitude variation in the space-charge field through the denominator factor $D(K)$ in (7.17) (as shown explicitly in (7.A4) of Appendix 7.A). This variation is negligible compared with the effect induced by the coherent grating spatial frequency described above.

To verify the predicted high bandwidth of the recording process and to eliminate depth of focus issues as discussed previously, the Michelson interferometer configuration shown in Fig. 7.15 was used to record sinusoidal image patterns onto a bismuth silicon oxide crystal. The ratio R of the image-bearing light intensity I_1 to the coherent grating light intensity I_0 was adjusted experimentally to maximize the intensity of the I_{11} diffraction order. As the image grating frequency was varied, the angular alignment of the readout beam was also varied to maintain optimum Bragg angle alignment, thereby removing Bragg detuning effects on the readout resolution. Thus the measured diffracted intensity I_{11} corresponds to the strength of the space-charge field stored in the crystal, with results as shown in Fig. 7.19. A slight decrease in diffraction efficiency with increasing spatial frequency is observed, but the effective bandwidth for grating writing far exceeds the expected bandwidths for Bragg detuning on readout, as discussed in the readout section below.

A far more serious implication of (7.17) for the space-charge field component E_1 is a shift in the phase of the recorded space-charge field profile with

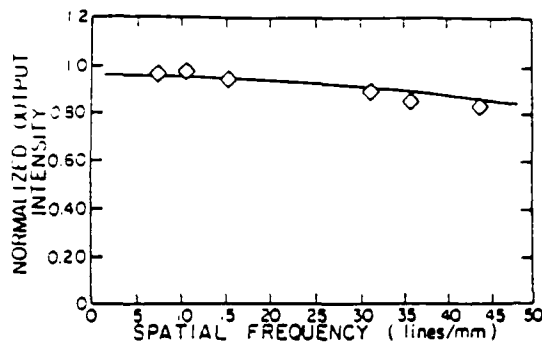


Fig. 7.19. Measured resolution response of the photorefractive recording process, limited by material dependent constraints. These measurements were performed by continuously Bragg matching the I_{11} order, thereby removing any dependence of bandwidth on Bragg detuning, as discussed in Sect 7.5

respect to the incident coherent grating illumination profile. Considering again an equilibrium donor-like trap density $N_{D\text{eq}}^+$ of 10^{16} cm^{-3} , a coherent grating frequency of 300 cycles/mm, and an applied bias field of 6 kV/cm, the resultant phase shift is of order 18° . This phase shift poses no problems for PICOC performance so long as it remains uniform over the full aperture of the photorefractive crystal and the full bandwidth, as it in fact varies weakly with K_S and m_S . On the other hand, an image-induced differential phase shift can prove to be problematic for particular optical processing architectures.

Resolution Anisotropy. A final resolution issue that has been predicted by the perturbation series is a moderate anisotropy in the recording of image structures parallel as opposed to perpendicular to the applied bias field for the simultaneous erasure/writing mode (SEWM), and a severe resolution anisotropy for the grating inhibition mode (GIM). Features of this analysis are quite intriguing and hence are briefly reviewed here.

The perturbation series analysis is conducive to a physically intuitive interpretation of the perturbation terms as a sequence of discrete recording events. For the simultaneous erasure/writing mode (SEWM), two recording paths contribute to the I_{11} diffraction order. In one recording path, the incoherent image writes a space-charge field E_{01} , independent of the coherent grating profile. This space-charge field E_{01} then modulates the recording of the coherent grating light beam. This transcription path is analogous to the grating inhibition mode (GIM). In the second SEWM transcription path, the coherent grating profile writes a space-charge field E_{10} , which then modulates the recording of the incoherent image-bearing beam. This second path is analogous to the grating erasure mode (GEM). The GEM-like path exhibits nearly perfect isotropy of response to an arbitrary image, but the GIM-like path exhibits a very strong anisotropy, with image structures oriented perpendicular to the applied bias field generating much weaker space-charge fields than structures oriented parallel to the bias field.

To elaborate, consider an image profile consisting of a single spatial frequency, similar to (7.2), but oriented such that the wave vector K_S is orthogonal

to the applied bias field E_0 . Thus the incident image intensity $I_S(y)$ is given by

$$I_S(y) = I_1 (1 + m_S \cos K_S y) , \quad (7.30)$$

in which the y axis is orthogonal to the applied bias field E_0 . This image profile, in combination with the coherent grating profile as given by (7.1), induces a space-charge field $E(x, y)$ of the form

$$\begin{aligned} E(x, y) = & \hat{x} E_{10} \cos K_G x + \hat{y} E_{01} \cos K_S y \\ & + (\hat{x} E_{11X} + \hat{y} E_{11Y}) \cos K_G x \cos K_S y \\ & + \text{higher order terms} \end{aligned} \quad (7.31)$$

in which $\hat{x}$ and $\hat{y}$ are unit vectors parallel and perpendicular to the applied bias field respectively.

Terms such as E_{01} and E_{11Y} involve recording a charge pattern along a direction orthogonal to the applied bias field, and hence do not benefit from the enhancement of the photoconductivity induced by this applied field. In practice, these terms are very much smaller than terms such as E_{10} and E_{11X} which involve recording a charge pattern along a direction parallel to the applied bias field. Analysis using the perturbation method shows that the E_{11X} term is reduced by a factor of two compared with the E_{11} term given by (7.5) that results when the image wave vector K_S is parallel to the applied bias field E_0 .

Thus we find that the material limitations on recording resolution are negligible compared with the geometrical limitations, which in turn are of comparable magnitude to the Bragg sensitivities on readout to be discussed in the Sect. 7.5.

7.4.4 Temporal Response

A temporal response analysis is necessary for the study of the grating inhibition mode (GIM) and the grating erasure mode (GEM) since these involve temporal sequencing, and the issue of timing is crucial to the optimization of their performance. The temporal analysis is also of interest for the simultaneous erasure/writing mode (SEWM) because it clarifies the duration and nature of the transient writing period before a stable response is achieved, and also because it leads to a very important reciprocity law between the incident light power and the response time of the converter (assuming that the dark conductivity of the photorefractive crystal can be neglected). A final and very important issue that arises from the temporal response is an image-induced phase modulation of the coherent grating's charge profile that can degrade PICOC performance in some coherent processing architectures.

The analysis reported here is restricted to small modulation regimes for simplicity. The results are modified substantially when operating with large modulation depths. In particular, the transient period increases significantly in such a regime.

To introduce some of the key concepts, consider the temporal evolution of the space-charge field in response to coherent grating illumination in the absence of an incoherent image. *Kukhtarev's* study of this problem using first order perturbation analysis [7.19] (in the absence of self-diffraction effects) has shown that if the photorefractive crystal is illuminated with an intensity

$$I(x, t) = \begin{cases} 0 & \text{for } t < 0 \\ I_0(1 + m_G \cos K_G x) & \text{for } t > 0 \end{cases} \quad (7.32)$$

then the space-charge field component $E_1(t)$ has the form

$$E_1(t) = -\frac{1}{2} m_G [E_0 + iE_D(K_G)] D(K_G)^{-1} \{1 - \exp[-t/T(K_G)]\} \quad (7.33)$$

in which $D(K_G)$, $E_D(K_G)$, and $E_q(K_G)$ have been defined by (7.18–20). The time constant $T(K)$ is defined by

$$T(K) = T_0 \frac{\{1 - i[E_0 + iE_D(K)]/E_M(K)\}}{\{1 - i[E_0 + iE_D(K)]/E_q(K)\}} \quad (7.34)$$

in which T_0 is the dielectric relaxation time, defined by

$$T_0 = \frac{\epsilon \epsilon_0}{\mu e n_0} \quad (7.35)$$

and n_0 is the zeroth order estimate of the electron density given by

$$n_0 = S_G I_0 N_D \tau \quad (7.36)$$

In (7.36), τ is the free carrier lifetime, as given by

$$\tau = (\gamma_R N_{D \text{ eq}}^+)^{-1} \quad (7.37)$$

A group of parameters occurs in (7.34) having the dimensions of an electric field. This field parameter is assigned the symbol $E_M(K_G)$, and is defined by

$$E_M(K_G) = (\mu K_G \tau)^{-1} \quad (7.38)$$

The physical interpretation of this field parameter can be best understood by considering the average transport length of the mobile charges. When the drift contribution to the current density in (7.12) dominates over the diffusion contribution, the mobile charges during their limited lifetime τ travel an average distance $L = \mu \tau E$ while under influence of an electric field E . The drift transport length is equal to the grating period when the field strength $E = 2\pi E_M$.

The space-charge field in (7.33) exponentially approaches its saturation limit, generally with a complex time constant $T(K_G)$ which denotes oscillatory as well as decaying behavior. At low spatial frequencies, the drift-induced transport length is very short compared with the grating period (equivalent to having a field E_M large compared to the applied bias field E_0). In this case, the decay predicted by (7.33) is governed essentially by the dielectric relaxation time T_0 and exhibits negligible oscillatory behavior. This gives the fastest possible response time. For higher spatial frequencies and high applied bias fields, for which the drift-driven charge transport length greatly exceeds the grating pe-

riod, the response time increases substantially beyond the dielectric relaxation time and in addition the response exhibits oscillatory behavior. An intuitive interpretation of this phenomenon is that the finite transport length blurs the charge pattern being transcribed, forcing a longer recording time to achieve a given level of charge profile modulation. Furthermore, the blur pattern is one-sided because the applied bias field forces the mobile charges always in one direction, inducing a phase shift of the charge pattern being transcribed. In all cases, the response time is inversely proportional to the incident light intensity, so that doubling the incident intensity reduces the response time by a factor of two.

By similar analysis, one finds that the erasure of the resulting space-charge field by a spatially uniform incident light beam also exhibits an exponential response with a time constant that is inversely proportional to the erasure light intensity. In the absence of an applied bias field, this response is described by a simple exponential function with no shift in the phase of the coherent grating's charge profile. In the presence of an applied bias field, the uniform erasure light beam induces a drift of the charge pattern that was originally recorded by the coherent grating beams, resulting in a phase modulation as well as an amplitude modulation.

Let us now review the analysis of one particular version of the simultaneous erasure/writing mode (SEWM) for the photorefractive incoherent-to-coherent optical conversion (PICOC) process. Consider a crystal in which a coherent grating of wave vector K_G has been written and has reached steady state. At time $t = 0$, an incoherent image grating with wave vector K_S is turned on. The intensity incident upon the crystal is then described by

$$I(x, t) = \begin{cases} I_0(1 + m_G \cos K_G x) & \text{for } t < 0 \\ I_0(1 + m_G \cos K_G x) + I_1(1 + m_S \cos K_S x) & \text{for } t > 0 \end{cases} \quad (7.39)$$

The temporal evolution of the various components of the space-charge field can be solved by perturbation techniques, valid for small levels of the modulation depth m_G , with explicit expressions for the lowest order terms as presented in Appendix 7.A. In particular, the temporal response of the E_{11} component has the general form

$$E_{11}(t) = M_0 + M_1 \exp[-t/T_1(K_G)] + M_2 \exp[-t/T_2(K_S)] \\ + M_3 \exp[-t/T_3(K_G + K_S)] + M_4 \exp[-t/T_4(K_G, K_S)] \quad (7.40)$$

in which the M 's are complex coefficients that depend upon the applied field, material parameters, and the incident light intensities, with explicit expressions as given in Appendix 7.A and in [7.22]. Time constants T_1 , T_2 , and T_3 are given by (7.34) for spatial frequencies corresponding to K_G , K_S , and $(K_G + K_S)$, respectively. The fourth time constant T_4 is given by

$$T_4(K_G, K_S) = [1/T(K_G) + 1/T(K_S)]^{-1} \quad (7.41)$$

The first term M_1 corresponds to the recording of the coherent grating wave vector K_G in the absence of the image profile. The second term M_2 corresponds to the recording of the image profile wave vector K_S in the absence of the coherent grating. The third term M_3 corresponds to the recording of the intermodulation frequency $K_G + K_S$, and represents a resonant response of the system of photorefractive equations. The fourth term M_4 corresponds to the nonresonant response driven by the product of the coherent grating and the incoherent recording processes, which arises through the nonlinearity of the recording process. The steady state value of the E_{11} field component is governed by the term M_0 .

The longest response time to reach steady state is comparable in value to the pure coherent grating response as discussed above in (7.34–38) and in [7.19]. This overall response time t_{sys} obeys the following reciprocity law

$$t_{sys} = \frac{G}{S_G I_0 + S_S I_1} \quad (7.42)$$

in the absence of appreciable dark conductivity β in the photorefractive crystal, in which the proportionality constant G involves only material parameters and the applied bias field E_0 . As a result, the rate at which new information can be recorded is determined by the total available intensity incident on the crystal.

Similar temporal response analyses have been performed for both the grating inhibition mode (GIM) and the grating erasure mode (GEM) [7.22]. In the GEM mode, the response time constant of the system is inversely proportional to the incoherent erasing intensity I_1 , rather than the sum of the incoherent and coherent intensities. Conversely, in the GIM mode, the response time constant of the system is inversely proportional to the coherent grating intensity I_0 .

Sample perturbation analysis solutions for GEM, GIM, and SEWM are shown in Figs. 7.20, 7.21, and 7.22 respectively for a 1 mW/cm^2 average intensity coherent grating beam at 515 nm wavelength with a grating spatial frequency of 300 cycles/mm and a small modulation depth m_G , a 1 mW/cm^2 average intensity incoherent image-bearing beam at 488 nm wavelength with an image spatial frequency of 10 cycles/mm and also with a small modulation depth m_S , an applied bias field of 6 kV/cm , and with the material parameters for bismuth silicon oxide as given in Table 7.1.

Figure 7.20 shows the GEM response for three diffracted light components: the direct recording of the unmodulated coherent grating component I_{10} , the direct recording of the incoherent image-bearing beam I_{01} , and the intermodulation component I_{11} (the image-modulated grating component). The coherent grating has been recorded to saturation and then turned off before the time interval shown in the figure. The time $t = 0$ is defined when the image-bearing beam is turned on. Thus the coherent grating frequency component I_{10} starts at its saturated level and decays for times $t > 0$ because of erasure by the image-bearing beam. The direct recording of the image-bearing beam I_{01} grows from an initial value of zero to its saturation level. Note that the response time for

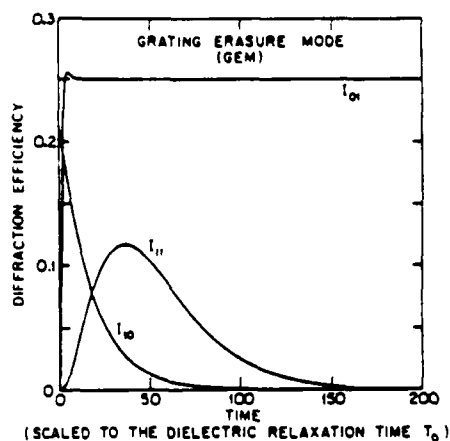


Fig. 7.20

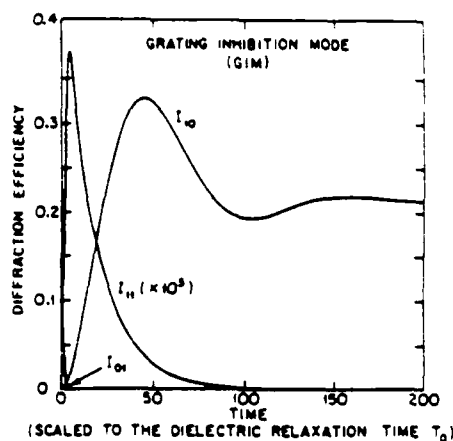


Fig. 7.21

Fig. 7.20. Temporal evolution for the grating erasure mode (GEM) as predicted by the perturbation series analysis. Time $t = 0$ starts after the coherent grating has been recorded and the incoherent image-bearing beam has just been turned on. In general the I_{11} diffraction component would be much smaller than either the I_{10} or the I_{01} components in this and in the next two figures, corresponding to small modulation depths m_G^{eff} and m_S^{eff} , but for simplicity all three curves are shown as if these modulation depths were unity

Fig. 7.21. Temporal evolution for the grating inhibition mode (GIM) as predicted by the perturbation series analysis. Time $t = 0$ starts after the incoherent image-bearing beam has been recorded and the coherent grating beams have just been turned on

the I_{01} image component, with its much lower spatial frequency, is significantly faster than that for the I_{10} coherent grating component, with its order of magnitude higher spatial frequency. The image-modulated grating component I_{11} exhibits a temporal response which is derived from a combination of the I_{10} and I_{01} response, eventually evolving into a slow decay in time when the incoherent image-bearing light beam erases the coherent grating. The I_{11} intensity is eventually erased by the image-bearing beam for very long recording times, so that the image-bearing light exposure time must be truncated.

Figure 7.21 shows a similar set of temporal response curves for the grating inhibition mode (GIM). The image-bearing light has been recorded to saturation, then turned off before the time interval shown in this figure. Time $t = 0$ is defined when the coherent grating light is turned on. In this mode, the pure coherent grating component I_{10} starts with zero intensity and gradually grows to saturation, whereas the directly recorded incoherent image-bearing beam I_{01} starts from its saturation level and is quite rapidly erased by the uniform component of the coherent grating light. The image-modulated grating component I_{11} shows initially a very rapid rise, followed by a much slower erasure by the coherent grating beam, eventually decaying to zero for very long recording times. In GIM, the image-modulated grating I_{11} component generally falls far short of its levels for GEM and SEWM, at least as predicted by the pertur-

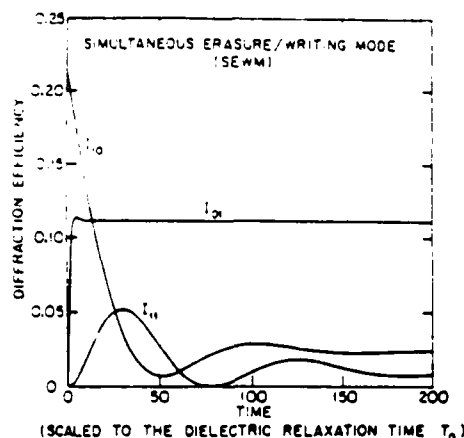


Fig. 7.22

Fig. 7.22. Temporal evolution for one version of the simultaneous erasure/writing mode (SEWM) as predicted by the perturbation series analysis. Time $t = 0$ starts after the coherent grating has been recorded and the incoherent image-bearing beam has just been turned on.

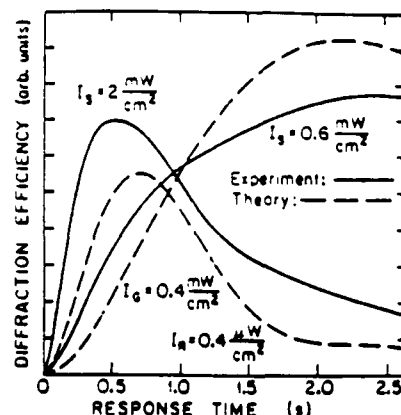


Fig. 7.23

Fig. 7.23. Measured diffraction efficiency of the I_{11} beam (strongly modulated erasure) as a function of time for two values of the signal intensity for the simultaneous erasure/writing mode, shown for comparison with the corresponding theoretically predicted response curves from the perturbation series analysis. In this figure, I_3 is the intensity of the signal beam(s), I_G is the intensity of the grating recording beams, and I_R is the intensity of the readout beam.

bation series analysis, because the direct image recording is erased before the coherent grating recording has a chance to grow.

Figure 7.22 presents the simultaneous erasure/writing mode response, assuming the temporal sequencing given by (7.39). The pure coherent grating recording I_{10} starts with its initial saturation level at time $t = 0$, and slowly drops to a reduced saturation level. The direct recording of the incoherent image-bearing beam I_{01} rapidly builds from zero intensity to saturation. The image-modulated grating component I_{11} exhibits strong oscillations, eventually settling at its saturation level, with a response time much longer than that of the I_{01} component.

Measurements of the temporal response of the I_{11} diffraction order are compared in Fig. 7.23 with the temporal response solutions generated by the perturbation series analysis for the cases of intensity ratios $R = 1.5$ and $R = 5.0$ and for a grating written in the nondegenerate simultaneous erasure/writing mode (SEWM). The experiments are shown as solid lines, the theoretical predictions as dashed lines. In this figure, the coherent grating is established in the saturation regime at time $t = 0$, at which point the incoherent erasure beam is allowed to expose the crystal, as given by (7.39). For a small R ratio ($I_1 = I_3 = 0.6 \text{ mW/cm}^2$ in Fig. 7.23), the experimental diffraction efficiency increases monotonically, at least within the time interval of this figure. For a strong incoherent beam ($I_1 = I_3 = 2 \text{ mW/cm}^2$ in Fig. 7.23), a transient

effect appears in the experimental response within this same time interval. Initially, the incident beam diffracts from the composite grating at wave vector ($K_G + K_S$) to generate a rapid rise in the amplitude of I_{11} , but the strong incoherent illumination eventually erases the coherent grating and hence the diffraction efficiency decreases to a small steady-state value.

The theoretical curves are scaled in peak intensity and in dielectric relaxation time to achieve a reasonable match with the experiments, but once the scale is defined for one curve, it fixes the scale for both theoretical curves. The dielectric relaxation time needed to achieve agreement between the theoretical curves and the experimental data is a factor 3 slower than that derived from the bismuth silicon oxide parameters given in Table 7.1 per (7.35–37). When comparing the theoretical with the experimental curves in Fig. 7.23, it should be noted that the theory is most accurate for low modulation depths, whereas the experiment is performed with the highest possible modulation depths for both the coherent grating and the incoherent image beams. Even so, the match between the theoretical response and the measured response is quite striking.

The time constant for this particular set of experimental parameters is in the range 0.5 to 1.5 s. To achieve conversion rates of 30 frames per second in bismuth silicon oxide, a total light intensity of order 35 to 45 mW/cm² is extrapolated, based upon the reciprocity law given in (7.42), and assuming that the ratio of incoherent image-bearing light to coherent grating light is kept constant.

7.5 The Readout Process

The readout process consists of the optical modulation of the coherent readout beam by the space-charge field. This modulation occurs in PICOC through the linear electrooptic effect which modifies the refractive index within the photorefractive crystal, thus establishing a volume phase grating. The phase hologram is then read out by a coherent auxiliary beam to achieve the conversion.

The diffraction characteristics of such volume phase gratings have been studied using many different analytical techniques, including coupled wave analysis by *Kogelnik* [7.31] with extensive numerical studies by *Klein and Cook* [7.32], a Born approximation to a scattering integral by *Gordon* [7.33], and the optical beam propagation method by *Yevick and Thylen* [7.34] and *Johnson and Tanguay* [7.35]. Methods for studying the polarization properties of light diffraction in electrooptic crystals such as bismuth silicon oxide include anisotropic versions of the coupled wave formalism [7.23, 36] and the optical beam propagation method [7.35].

This section examines the readout of the phase gratings and its consequences for the performance of PICOC as a spatial light modulator. Because of the high spatial frequencies typically used in PICOC, the grating exhibits pronounced Bragg diffraction characteristics with rapid degradation of the readout quality whenever the Bragg condition is detuned, whether by increasing

the spatial frequency of the incoherent image, by slight angular misalignment, or by sub-optimum alignment of the incoherent image beam. The Bragg sensitivity is discussed in the section on the isotropic phase grating properties. In addition, because of the optical activity exhibited by bismuth silicon oxide, the polarization states of both the transmitted probe light and the diffracted signal light will in general be elliptical, with implications for the optimum readout conditions, as discussed in the section on polarization properties.

7.5.1 Isotropic Phase Grating Model

The space-charge field induces a small perturbation in the index of refraction through the linear electrooptic effect, such that

$$\Delta n(x) = \frac{1}{2} n_0^3 r_{41} E_1 \cos(K_G x) \quad (7.43)$$

in which n_0 is the nominal index of refraction and r_{41} is the electrooptic coefficient appropriate for bismuth silicon oxide. The index perturbation in turn modulates the optical phase fronts of an incident light beam. Thus the sinusoidal space-charge field induces a volume phase grating in the crystal. To gain some feeling for the readout performance issues involved in PICOC, consider a 2 mm thick piece of bismuth silicon oxide with a simple unmodulated sinusoidal grating with spatial frequency of 300 cycles/mm, space-charge field of 5 kV/cm, and probed by a 633 nm laser beam. A space-charge field E_1 of order 5 kV/cm induces an index perturbation Δn of order 2×10^{-5} , assuming the material parameters for bismuth silicon oxide listed in Table 7.1.

Dimensionless parameters that characterize the diffraction characteristics of such a phase grating include the grating thickness parameter Q and the grating strength v , defined by

$$Q = \frac{\lambda_R d K_G^2}{2\pi n_0} \quad \text{and} \quad (7.44)$$

$$v = \frac{2\pi \Delta n d}{\lambda_R} \quad (7.45)$$

in which λ_R is the wavelength of the readout light in *vacuo* and d is the thickness of the grating [7.32]. For a 300 cycles/mm grating in a 2 mm thick crystal, read out by a 633 nm laser beam, the associated thickness parameter Q is almost 300, which is considered to be a very thick grating [7.30]. A space-charge field of the order of 5 kV/cm induces a grating strength of the order of 0.35 radians, which gives fairly weak diffraction efficiency (on the order of a few per cent). The combination of large thickness parameter Q and weak grating strength v places the grating deep in the Bragg regime with an optical diffraction efficiency η given approximately by

$$\eta = [(v/2\sigma) \sin \sigma]^2, \quad \text{with} \quad (7.46)$$

$$\sigma = \left(\xi^2 + \frac{v^2}{4} \right)^{\frac{1}{2}} \quad \text{and} \quad (7.47)$$

$$\xi = \frac{1}{2} K_G d \sin(\theta_{in} - \theta_B) \quad (7.48)$$

(according to [Ref. 7.31, Eqs. 17, 42, and 43] or [Ref. 7.32, Eqs. 6-8, 35, and 36]), in which θ_{in} is the entrance angle of the incident probe light measured with respect to the constant phase lines of the coherent grating, and $\theta_B(K_G)$ is the Bragg angle associated with wave vector K_G , defined by

$$\sin \theta_B = \frac{\lambda_R K_G}{4\pi n_0} \quad (7.49)$$

The parameter ξ in (7.47, 48) is a measure of the Bragg misalignment; it is equal to zero when the readout beam is perfectly Bragg aligned with respect to the volume hologram.

For perfect Bragg alignment in which $\theta_{in} = \theta_B$, and assuming a grating strength $v = 0.35$ radians, the peak diffraction efficiency is estimated to be of order 3%. Doubling the thickness of the grating would increase the diffraction efficiency to 12% (ignoring polarization issues), but also increases the Bragg detuning sensitivity, as described in the following paragraphs.

Bragg detuning impacts PICOC performance in two ways: the angular alignment sensitivity of the hologram to the coherent readout beam, and the spatial frequency response of the hologram readout with its concomitant angular alignment dependence on the incoherent image-bearing beam. Consider first the alignment sensitivity to the coherent readout beam. An angular misalignment $\Delta\theta$ from the optimum Bragg alignment introduces a Bragg mismatch ξ of

$$\xi = \frac{1}{2} K_G d \Delta\theta \quad (7.50)$$

as seen from (7.48). Assuming a grating strength $v = 0.35$ radians, one finds from (7.46, 47) that the angular alignment sensitivity of the diffraction efficiency has essentially a sinc^2 profile. The angular misalignment $\Delta\theta$ needed to reach the first null of this profile occurs when $\xi = \pi$, i.e., for

$$\Delta\theta = \frac{2\pi}{K_G d} \quad (7.51)$$

Hence the angular alignment needed to achieve optimum diffraction efficiency is extremely sensitive, on the order of 0.1° for a 2 mm thick photorefractive crystal with a 300 cycles/mm grating frequency.

One possible alignment of the PICOC system is to orient the readout beam to be Bragg matched precisely to the coherent grating, thereby maximizing the intensity of the I_{10} diffracted order. However, when an incoherent grating is also incident upon the crystal, the nonlinear recording process creates a new grating with wave vector $(K_G + K_S)$ that in general does not satisfy the same Bragg condition as the coherent grating wave vector K_G . The resulting I_{11} intensity component is then attenuated by an amount dependent upon the image wave vector $K_S = 2\pi f_S$, in which f_S is the spatial frequency of the image profile. The dependence of the attenuation on the image frequency f_S is a function of the orientation of the image-bearing beam with respect to the coherent grating.

The importance of the image-bearing beam orientation on the spatial frequency response of the converter is strikingly illustrated by Fig. 7.24, in which are shown converted images of two orthogonal orientations of a 5 line pair/mm Ronchi ruling and the associated coherent Fourier transforms. As can be seen from the figure, a significant difference in resolution exists between cases in which the wave vector of the ruling (incoherent grating) is parallel or perpendicular to the coherently written (holographic) grating. This difference derives principally from the fact that a different wave vector matching condition exists for these two cases.

Consider the alternative wave vector matching conditions shown in Figs. 7.25 and 7.26. In Fig. 7.25, the incoherent grating wave vector K_S is parallel to the coherent grating wave vector K_G , a condition achieved by symmetrically disposing the incident coherent beams about the normal to the crystal while simultaneously arranging the incoherent imaging system such that its optical axis is parallel to the crystal normal. In this case, significant Bragg detuning occurs for even small incoherent grating wave vectors.

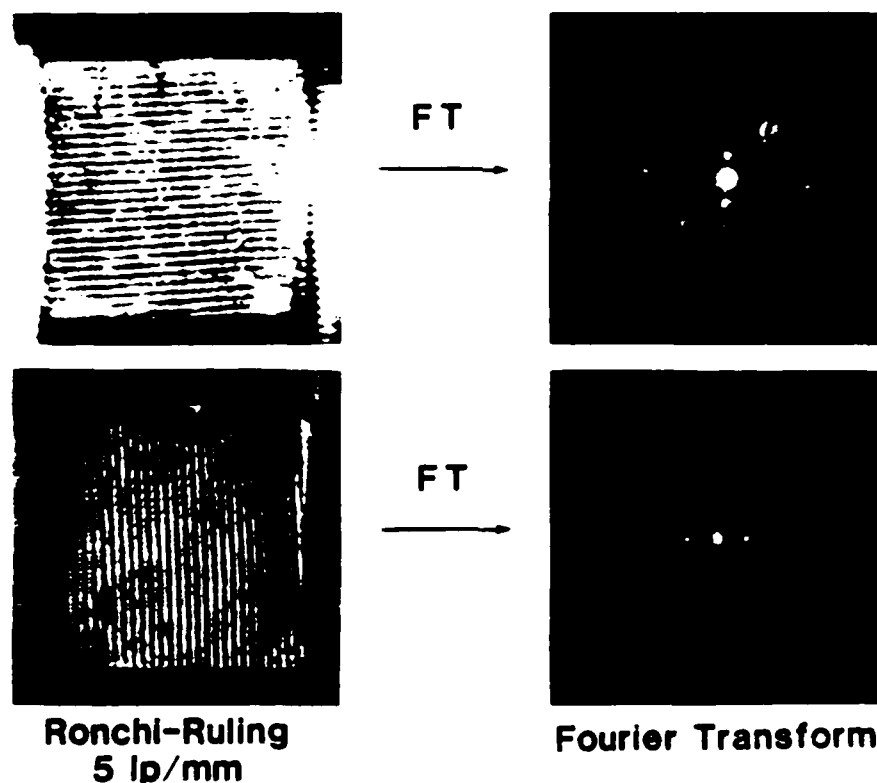


Fig. 7.24. Photograph of the Fourier transform of a Ronchi ruling recorded in the PICOC configuration, showing a strong anisotropy in the resolution performance when the image-bearing light beam bisects the two coherent grating writing beams

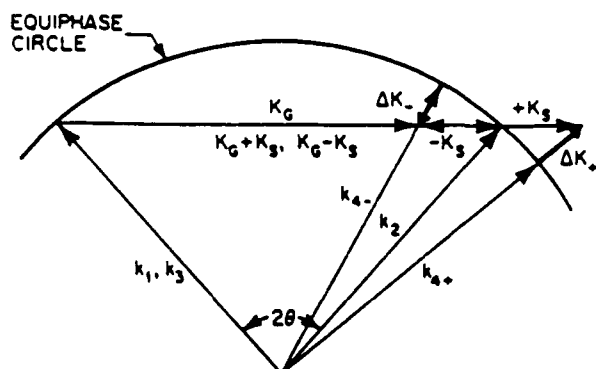


Fig. 7.25. Wave vector mismatch diagram for the case in which the image-bearing beam bisects the two coherent grating writing beams

Let the spatial bandwidth of the readout process be defined as that spatial frequency f_S for which the magnitude of the I_{11} diffracted intensity component degrades to 25% of its peak value. This occurs when $\xi = 1.9$, as determined from (7.46, 47), assuming a small grating strength v on the order of 0.35 radians. The Bragg mismatch parameter ξ is a function of the image spatial frequency f_S , and can be approximated by the first term or two in a Taylor series expansion of the form

$$\xi(f_S) = \xi_0 + f_S \frac{d\xi}{df_S} + \frac{1}{2} f_S^2 \frac{d^2\xi}{df_S^2} + \text{higher order terms} \quad (7.52)$$

in which ξ_0 represents the Bragg mismatch associated with the coherent grating. Standard alignment procedure is to set this zeroth order mismatch term ξ_0 to zero, i.e., to Bragg match the incident readout beam to the coherent grating.

For the case of the image-bearing beam bisecting the two coherent grating beams, as shown in Fig. 7.25, the first order term in (7.52) is a sufficiently accurate estimate of the Bragg mismatch parameter ξ , and this term can be derived from (7.48) to give

$$\frac{d\xi}{df_S} = -\frac{1}{2} K_G d \frac{d\theta_B}{df_S} \quad (7.53)$$

By (7.49), the term $d\theta_B/df_S$ is equivalent to

$$\frac{d\theta_B}{df_S} = \frac{\lambda_R}{2 n_0 \cos \theta_B} \quad (7.54)$$

Hence the spatial bandwidth f_S is

$$f_S = \frac{7.6 n_0 \cos \theta_B}{\lambda_R K_G d} \quad (7.55)$$

which can be expressed in terms of the fringe spacing Λ_G of the coherent grating wavelength as

$$f_S = \frac{1.2 n_0 \Lambda_G \cos \theta_B}{\lambda_R d} \quad (7.56)$$

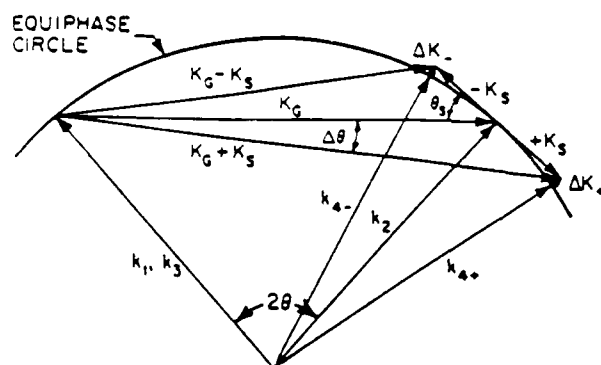


Fig. 7.26. Optimum wave vector alignment diagram in which the image-bearing light is tilted with respect to the bisector of the two coherent grating writing beams by the Bragg angle associated with the probe beam

For the parameters used in our experiments, this bandwidth is estimated to be on the order of 8 cycles/mm. Note that doubling the crystal thickness d to increase the diffraction efficiency by almost a factor of four has the adverse effect of reducing the spatial bandwidth f_s by a factor of two for this particular alignment configuration.

Compare now the resolution performance associated with the alignment of Fig. 7.25 with that for Fig. 7.26. In Fig. 7.26, the incoherent grating wave vector is arranged to lie tangentially to the circle defined by the readout beam wave vector, such that a significantly increased angular deviation of the diffracted beam is allowed before serious Bragg detuning effects occur. Such a wave vector tangency condition is automatically satisfied when the incoherent image wave vector is normal to the coherent grating wave vector (as it is in the y orientation normal to the plane of incidence). Alternatively, the wave vector tangency condition is satisfied when the central ray of the incoherent image beam is parallel to the diffracted probe light, for which the horizontal and vertical resolutions become degenerate. This is not the case for the situation depicted in Fig. 7.25, which explains the observations apparent in Fig. 7.24. An equivalent condition has been described by Kamshilin and Petrov [7.6] for the nondegenerate four-wave mixing optical architecture.

In the geometry of Fig. 7.26, the addition of an image frequency f_s shifts the pointing direction of the combined wave vector ($K_G + K_S$) such that the incident readout light remains almost perfectly Bragg matched over a much broader range of image frequencies. Mathematically, this condition is equivalent to setting the first order term $d\xi/df_s$ of the Taylor series expansion in (7.52) for the Bragg mismatch parameter ξ equal to zero. To demonstrate this, consider the more general case in which the image profile wave vector K_S is oriented at a small angle θ_s with respect to the coherent grating wave vector K_G . The combination ($K_G + K_S$) rotates through a small angle $\Delta\theta$ compared with wave vector K_G alone, with $\Delta\theta$ given approximately by

$$\Delta\theta = \frac{K_S \sin \theta_s}{K_G} = \frac{2\pi f_s \sin \theta_s}{K_G} \quad (7.57)$$

The Bragg mismatch parameter ξ induced by the image wave vector K_S is to first order given by

$$\xi = \frac{d\xi}{d\Delta\theta} \Delta\theta + \frac{d\xi}{df_S} f_S \quad (7.58)$$

and these two terms cancel when the skew angle $\theta_S = \theta_B$, the Bragg angle for the readout light beam associated with the coherent grating frequency. For a 633 nm readout beam wavelength and a 300 cycles/mm grating frequency, the Bragg angle θ_B is of order 5° , implying that a mere 5° separates the optimum resolution alignment of Fig. 7.26 from the alignment of Fig. 7.25. In addition, the angular alignment tolerance on the incoherent image-bearing beam is much tighter than the nominal alignment angle of 5° , typically of order 0.7° .

With the alignment of Fig. 7.26, the degradation in diffraction efficiency then becomes a second order function of the image spatial frequency f_S . The second derivative term $d^2\xi/df_S^2$ in (7.52) is

$$\frac{d^2\xi}{df_S^2} = \frac{\pi d \lambda_R}{n_0} \quad (7.59)$$

and the resulting spatial bandwidth of the tangential configuration is

$$f_S = \left(\frac{1.2 n_0}{\lambda_R d} \right)^{1/2} \quad (7.60)$$

This expression was first published by *Kamshilin* and *Petrov* [7.6]. For our experimental parameters, a tangential geometry increases the converter's bandwidth from 8 to 48 cycles/mm. It is interesting to note that, in addition to the increased bandwidth of the tangential geometry, the spatial resolution is independent of the coherent grating frequency, and that doubling the thickness of the crystal d does not halve the converter's bandwidth, as it would for the alignment configuration shown in Fig. 7.25, but only reduces it by a factor of $\sqrt{2}$.

Further experimental tests of the Bragg detuning hypotheses are presented in Figs. 7.27–29, in which the image source was a Michelson interferometer (shown in Fig. 7.15) to alleviate the depth of focus issues discussed previously. In these experiments, the intensity of the diffracted component I_{11} was measured as a function of the spatial frequency of the image source grating. Figure 7.27 shows the response for the geometry in which the incoherent image beam bisects the coherent grating writing beams to achieve the wave vector mismatch condition diagrammed in Fig. 7.25. The predicted frequency response rolloff with bandwidth of 8 cycles/mm, shown by the solid line, compares reasonably well with the experimental data points which indicate a bandwidth of 6 cycles/mm. Figure 7.28 shows the rolloff when the signal beam wave vector is tangent to the equiphase circle as shown in Fig. 7.26, giving much improved frequency response of 45 cycles/mm which is in excellent agreement with the theoretically predicted 48 cycles/mm. Figure 7.29 shows the rolloff when the signal light wave vector is nominally aligned to be perpendicular to the applied

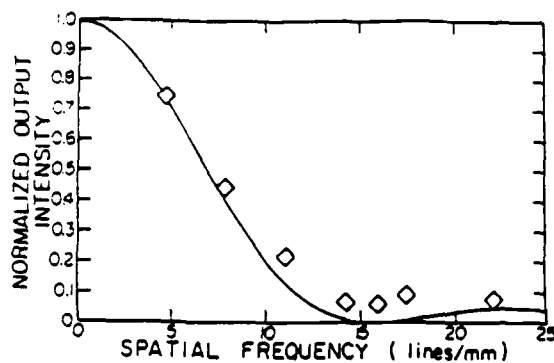


Fig. 7.27. Experimental measurement of the diffraction efficiency as a function of image spatial frequency associated with the alignment shown in Fig. 7.25

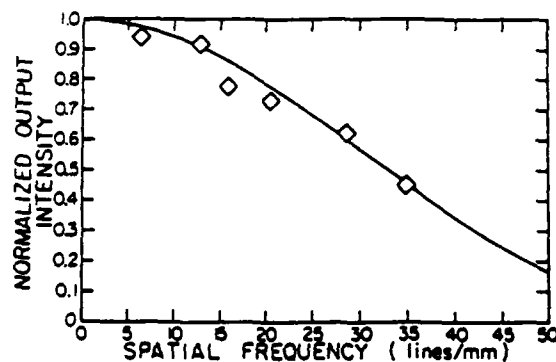


Fig. 7.28. Experimental measurement of the diffraction efficiency as a function of image spatial frequency associated with the optimum alignment shown in Fig. 7.26

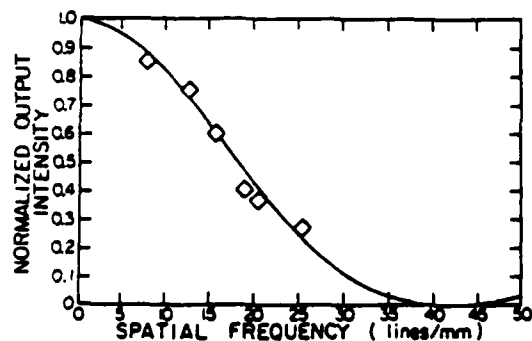


Fig. 7.29. Experimental measurement of the diffraction efficiency as a function of image spatial frequency for the alignment in which the image-bearing beam's wave vector is normal to the applied bias electric field

bias field and to the coherent grating wave vector. The theoretical bandwidth for this configuration should be identical to that shown in Fig. 7.28, but the measurements indicate a bandwidth of only 25 cycles/mm. This discrepancy may be attributable to the very high alignment sensitivity of the incoherent image beam relative to the plane defined by the two coherent writing beams, e.g., a deviation of approximately 0.7° would cause a deterioration of the bandwidth from 48 cycles/mm to 25 cycles/mm.

To summarize, the presence of the coherent grating in PICOC defines a volume hologram that is typically operated quite deep into the Bragg regime, with very strong misalignment sensitivities. As a consequence, the readout beam must be aligned typically within 0.1° of optimum, the image-bearing beam must be aligned typically within 0.7° of its optimum, and the optimum alignment for the image-bearing beam is not to bisect the two coherent grating beams (Fig. 7.25), but rather offset from this by a small angle typically of order 5° (as shown in Fig. 7.26).

7.5.2 Polarization Issues

Bismuth silicon oxide has quite remarkable optical polarization properties that strongly influence the optimization of the readout process [7.23]. These properties include significant levels of natural optical activity (as high as $46^\circ/\text{mm}$ for the 488 nm argon ion laser wavelength – see Fig. 7.30), a uniform linear birefringence induced by the applied bias field through the electrooptic effect, and a spatially varying linear birefringence induced by the image-defined space-charge field. In almost all readout configurations, these properties will cause both the readout and the diffracted signal beams to exhibit elliptical polarizations.

Consider Fig. 7.31, in which is shown the evolution of the polarization states for both the readout and the diffracted signal beams as a function of depth into the bismuth silicon oxide crystal for a 633 nm wavelength readout beam and an applied bias field of 6 kV/cm in the absence of self-diffraction effects. A hologram induced by a single coherent grating unmodified by any image profile is assumed. In this figure, we have chosen to plot the polarization states that result from an input polarization set at 45° with respect to the grating wave vector, which is therefore along one of the two electrooptically in-

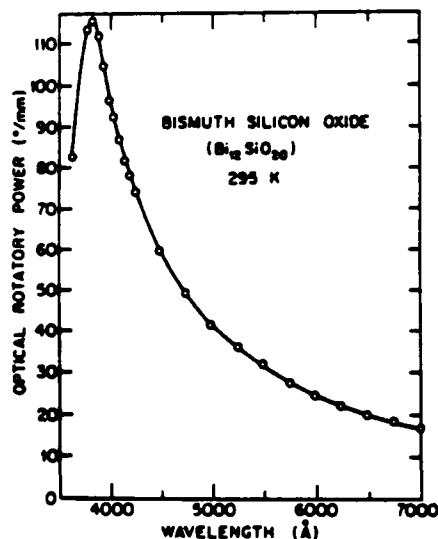


Fig. 7.30. Optical rotatory power of bismuth silicon oxide as a function of the wavelength of the readout light (after [7.27])

duced principal axes of the crystal's index ellipsoid. As such, the polarizations of the readout and signal beams are nearly parallel for very shallow depths but quickly evolve toward a 90° major axis separation with increasing depth because of the influence of the optical activity. This separation of polarization states enables the use of polarization analyzer techniques to suppress the scattered readout beam light in favor of the diffracted signal light. The intricacy of the polarization state evolution can be appreciated from Fig. 7.31.

One potential application of the polarization properties in bismuth silicon oxide has been demonstrated by *Herriau et al.* [7.15] for obtaining optimum holographic readout. They attained excellent suppression of scattered readout light noise for a nearly on-axis recording configuration by placing a polarization analyzer into the diffracted signal beam path. The analyzer is then adjusted to eliminate the scattered readout beam, and since the diffracted signal beam generally has a different polarization state from that of the transmitted readout beam, a significant fraction of the signal beam will pass through the analyzer. This technique greatly improves the signal-to-noise ratio of the holographic reconstruction process.

The presence of spatial modulation further complicates the polarization properties of the diffracted light, especially when the incoherent image beam is misaligned from the optimum wave vector matching configuration shown in Fig. 7.26, which introduces a strong polarization state dispersion. That is, the

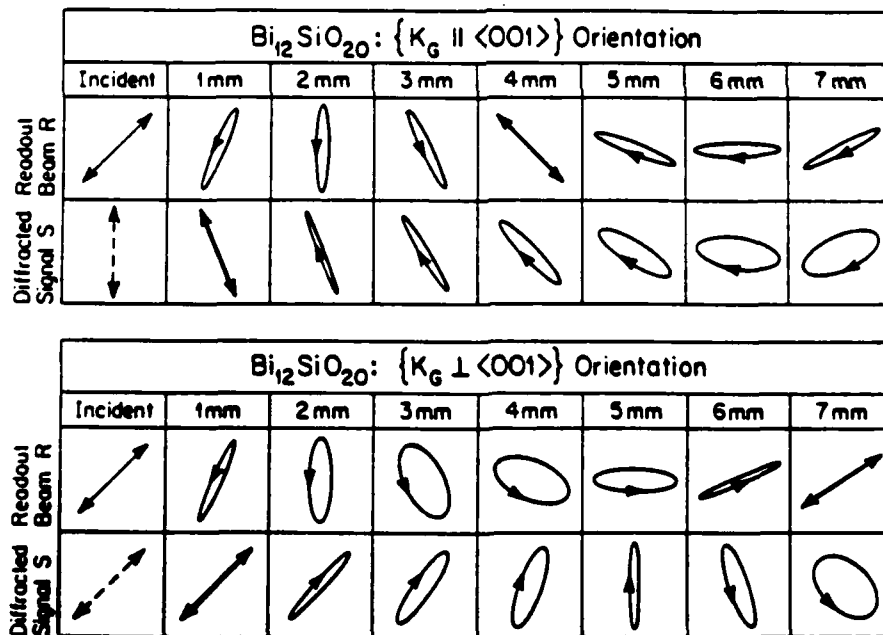


Fig. 7.31. Sample evolution of the polarization states of the incident readout beam and the diffracted signal beam for a simple sinusoidal grating

polarization states of the modulated readout beam's diffraction orders (I_{mn} in Fig. 7.10) are strongly dependent on the image spatial frequency when the optimum alignment of Fig. 7.26 is not achieved. Such dispersion further degrades the resolution if a polarization analyzer is used to separate the scattered readout light from the diffracted signal light. In contrast, when the optimum alignment of Fig. 7.26 is met, the polarization dispersion is negligible. This is one more reason why the alignment of Fig. 7.26 is crucial to achieving the best performance from the PICOC device.

The most serious issue concerning the polarization properties of volume holograms in bismuth silicon oxide is the degradation of the light diffraction efficiency that is imposed by the optical rotatory power [7.23]. However, one readout configuration has been identified in extensive polarization analyses that does not degrade the diffraction efficiency, namely the crystal geometry of Fig. 7.8 with no applied bias field and with a circularly polarized readout light beam. The diffraction efficiency for a linearly polarized readout beam is compared with that of a circularly polarized readout beam in Fig. 7.32, showing the marked improvement in diffraction efficiency that can be achieved by using circularly polarized light at the correct alignment.

In conclusion, the polarization properties of light diffraction in bismuth silicon oxide significantly affect the performance of the PICOC modulator, and can be exploited to improve this performance in terms of signal-to-noise ratio and to attain the highest possible diffraction efficiency.

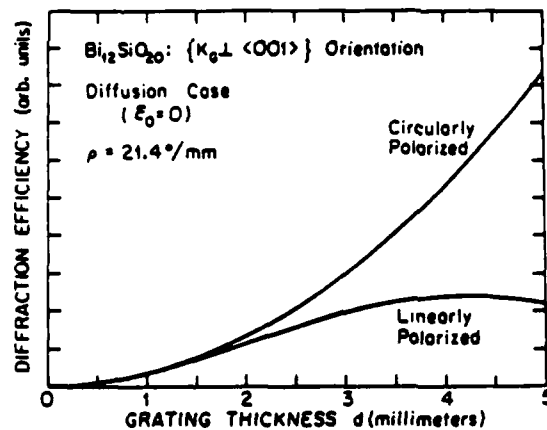


Fig. 7.32. Comparison of the diffraction efficiency achieved by linearly polarized and circularly polarized probe light when the applied bias field is set to zero during the readout process

7.6 Conclusion

Through the use of the photorefractive incoherent-to-coherent optical conversion (PICOC) process, we have successfully converted incoherent images into their negative coherent replicas. The PICOC system is inexpensive, easily implementable, and compares favorably in its performance with other photorefrac-

tive spatial light modulators. In addition, we have developed a mathematical framework within which to analyze the performance of the converter. Using this formalism, we have derived the conditions necessary to achieve high linearity, good contrast ratio, and fast temporal response. Experiments have been conducted which verified the essential features of these predictions, and which have underscored important materials issues. It is of considerable importance to extend the model to account more accurately for these materials issues, to predict the response of the converter at higher modulation depths, and to explore the optical phase modulation induced in the output image and its impact on various image processing architectures. In addition, testing of the photorefractive incoherent-to-coherent optical converter in a representative optical image processing system will undoubtedly highlight significant features worthy of further study and advanced development.

7.A Appendix. Steady State and Temporal Behavior of the Space-Charge Field Components in PICOC (Simultaneous Erasure/Writing Mode)

In this Appendix, the expressions for the lowest order components of the space-charge field E_{mn} are presented, both in the steady state and with full temporal evolution. Details of the derivation of these expressions can be found in [7.22].

7.A.1 Steady State Behavior

If the light incident on the photorefractive crystal is given by

$$I(x) = I_0(1 + m_G \cos K_G x) + I_1(1 + m_S \cos K_S x) \quad (7.A1)$$

then to first order the steady state response of the E_{01} , E_{10} , and E_{11} components are

$$E_{10} = -\frac{1}{2}m_G^{\text{eff}} \frac{E_0 + iE_D(K_G)}{D(K_G)} \quad (7.A2)$$

$$E_{01} = -\frac{1}{2}m_S^{\text{eff}} \frac{E_0 + iE_D(K_S)}{D(K_S)} \quad (7.A3)$$

$$E_{11} = \frac{1}{4}m_S^{\text{eff}}m_G^{\text{eff}}[D(K_G)D(K_S)D(K_G + K_S)]^{-1} \times \{F_1[E_0 + iE_D(K_G)] + F_2[E_0 + iE_D(K_S)]\} \quad (7.A4)$$

in which

$$m_G^{\text{eff}} = m_G \frac{S_G I_0}{S_G I_0 + S_S I_1} \quad (7.A5)$$

$$m_S^{\text{eff}} = m_S \frac{S_S I_1}{S_G I_0 + S_S I_1} \quad (7.A6)$$

$$D(K) = 1 - i \frac{E_0 + iE_D(K)}{E_q(K)} \quad (7.A7)$$

$$F_1 = D(K_G) + \frac{E_D(K_S)}{E_q(K_G)} \quad (7.A8)$$

$$F_2 = D(K_S) + \frac{E_D(K_G)}{E_q(K_S)} \quad \text{and} \quad (7.A9)$$

$$E_q(K) = \frac{eN_D^{+}eq}{\epsilon\epsilon_0 K} \quad (7.A10)$$

$$E_D(K) = \frac{kTK}{e} \quad (7.A11)$$

for which K can assume the values K_G , K_S , and $(K_G + K_S)$.

7.A.2 Temporal Response

The temporal behavior of the space-charge field components for the simultaneous erasure/writing mode (SEWM) is discussed in this section. Consider the following sequencing of light intensity profiles:

$$I(x, t) = \begin{cases} I_0(1 + m_G \cos K_G x) & \text{for } t < 0 \\ I_0(1 + m_G \cos K_G x) + I_1(1 + m_S \cos K_S x) & \text{for } t > 0 \end{cases} \quad (7.A12)$$

The temporal evolution of the various components of the space-charge field is given by

$$E_{10} = -\{(m_G - m_G^{\text{eff}}) \exp[-t/T(K_G)] + m_G^{\text{eff}}\} \times [E_0 + iE_D(K_G)]D^{-1}(K_G) \quad (7.A13)$$

$$E_{01} = -m_S\{1 - \exp[-t/T(K_S)]\}[E_0 + iE_D(K_S)]D^{-1}(K_S) \quad (7.A14)$$

$$E_{11} = M_0 + M_1 e^{-t/T_1} + M_2 e^{-t/T_2} + M_3 e^{-t/T_3} + M_4 e^{-t/T_4} \quad (7.A15)$$

in which

$$M_0 = \frac{1}{4} m_G^{\text{eff}} m_S^{\text{eff}} [D(K_G)D(K_S)D(K_G + K_S)]^{-1} \times \{F_1[E_0 + iE_D(K_G)] + F_2[E_0 + iE_D(K_S)]\} \quad (7.A16)$$

$$M_1 = \frac{1}{4} (m_G - m_G^{\text{eff}}) m_S^{\text{eff}} [D(K_G)D(K_S)D(K_G + K_S)]^{-1} \frac{T_1}{T_1 - T_3} \times \left\{ F_1[E_0 + iE_D(K_G)] + F_2[E_0 + iE_D(K_S)] \left(1 - \frac{T_0}{T_1}\right) \right\} \quad (7.A17)$$

$$M_2 = -\frac{1}{4}m_G^{\text{eff}}m_S^{\text{eff}}[D(K_G)D(K_S)D(K_G + K_S)]^{-1}\frac{T_2}{T_2 - T_3} \\ \times \left\{ F_1[E_0 + iE_D(K_G)]\left(1 - \frac{T_0}{T_2}\right) + F_2[E_0 + iE_D(K_S)] \right\} \quad (7.A18)$$

$$M_4 = -\frac{1}{4}(m_G - m_G^{\text{eff}})m_S^{\text{eff}}[D(K_G)D(K_S)D(K_G + K_S)]^{-1}\frac{T_4}{T_4 - T_3} \\ \times \left\{ F_1[E_0 + iE_D(K_G)]\left(1 - \frac{T_0}{T_2}\right) + F_2[E_0 + iE_D(K_S)]\left(1 - \frac{T_0}{T_1}\right) \right\} \quad (7.A19)$$

$$M_3 = -(M_0 + M_1 + M_2 + M_4) \quad (7.A20)$$

The dielectric relaxation time constant T_0 is defined by

$$T_0 = \frac{\epsilon\epsilon_0}{\mu\epsilon n_0} \quad (7.A21)$$

the time constants T_1 , T_2 , and T_3 are defined by

$$T(K) = \frac{T_0 C(K)}{D(K)} \quad (7.A22)$$

for $K = K_G$, K_S , and $(K_G + K_S)$ respectively; and the time constant T_4 is defined by

$$T_4 = \left(\frac{1}{T_1} + \frac{1}{T_2} \right)^{-1} \quad (7.A23)$$

In (7.A22), the factor $C(K)$ is defined by

$$C(K) = 1 - i \left(\frac{E_0 + iE_D(K)}{E_M(K)} \right) \quad (7.A24)$$

Note that the GEM temporal response can be derived from the SEWM expression by setting the average coherent grating intensity $I_0 = 0$ for $t > 0$. Thus the M_0 and M_2 terms disappear for GEM, leaving the M_1 , M_4 , and $M_3 = -(M_1 + M_4)$ terms. The GIM response can in turn be derived from the GEM response.

Acknowledgements. The authors thank F. Lum, D. Seery, M. Garrett, and Y. Shi for their technical assistance. This research was supported in part at the University of Southern California by the Air Force Systems Command (RADC) under Contract No. F19628-83-C-0031, the Defense Advanced Research Projects Agency (Office of Naval Research), the Joint Services Electronics Program, and the Army Research Office; and at the California Institute of Technology by the Air Force Office of Scientific Research and the Army Research Office.

References

- 7 1 D. Casasent: *Proc. IEEE* **65**, 143-157 (1977)
- 7 2 A.R. Tanguay, Jr.: *Opt. Eng.* **24**, 2-18 (1985)
- 7 3 B.A. Horwitz, F.J. Corbett: *Opt. Eng.* **17**, 353-364 (1978)
- 7 4 M.P. Petrov, A.V. Khomenko, M.V. Krasin'kova, V.I. Marakhonov, M.G. Shlyagin: *Sov. Phys. Tech. Phys.* **26**, 816-821 (1981)
- 7 5 Y. Owechko, A.R. Tanguay, Jr.: *J. Opt. Soc. Am.* **A1**, 635-652 (1984)
- 7 6 A.A. Kamshilin, M.P. Petrov: *Sov. Tech. Phys. Lett.* **6**, 144-145 (1980)
- 7 7 Y. Shi, D. Psaltis, A. Marrakchi, A.R. Tanguay, Jr.: *Appl. Opt.* **22**, 3665-3667 (1983)
- 7 8 A. Marrakchi, A.R. Tanguay, Jr., J. Yu, D. Psaltis: *Opt. Eng.* **24**, 124-131 (1985)
- 7 9 M.W. McCall, C.R. Petts: *Opt. Commun.* **53**, 7-12 (1985)
- 7 10 L.M. Bernardo, O.D.D. Soares: *Appl. Opt.* **25**, 592-593 (1986)
- 7 11 M.B. Klein, G.J. Dunning, G.C. Valley, R.C. Lind, T.R. O'Meara: *Opt. Lett.* **11**, 575-577 (1986)
- 7 12 M. Peltier, F. Micheron: *J. Appl. Phys.* **48**, 3683-3690 (1977)
- 7 13 D.L. Staebler, J.J. Amodei: *J. Appl. Phys.* **43**, 1042-1049 (1972)
- 7 14 R. Grousson, S. Mallick: *Appl. Opt.* **19**, 1762-1767 (1980)
- 7 15 J.P. Herriau, J.P. Huignard, P. Aubourg: *Appl. Opt.* **17**, 1851-1852 (1978)
- 7 16 M.G. Moharam, T.K. Gaylord, R. Magnusson, L. Young: *J. Appl. Phys.* **50**, 5642-5651 (1979)
- 7 17 V. Kondilenko, V. Markov, S. Odulov, M. Soskin: *Opt. Acta* **26**, 238-251 (1979)
- 7 18 N.V. Kukhtarev, V.B. Markov, S.G. Odulov, M.S. Soskin, V.L. Vinetskii: *Ferroelectrics* **22**, 949-960 (1979)
- 7 19 N.V. Kukhtarev: *Sov. Tech. Phys. Lett.* **2**, 438-460 (1976)
- 7 20 E. Ochoa, F. Vachas, L. Hesselink: *J. Opt. Soc. Am.* **A3**, 181-187 (1986)
- 7 21 Ph. Refregier, L. Solymar, H. Rajbenbach, J.P. Huignard: *J. Appl. Phys.* **58**, 45-57 (1985)
- 7 22 J. Yu, D. Psaltis, R.V. Johnson, A.R. Tanguay, Jr.: "Temporal evolution of multiple gratings in photorefractive crystals", to be published
- 7 23 A. Marrakchi, R.V. Johnson, A.R. Tanguay, Jr.: *J. Opt. Soc. Am.* **B3**, 321-336 (1986)
- 7 24 R. Orlovski, E. Kratzig: *Solid State Commun.* **27**, 1351-1354 (1978)
- 7 25 M.B. Klein, G.C. Valley: *J. Appl. Phys.* **57**, 4901-4905 (1985)
- 7 26 G.C. Valley: *Appl. Opt.* **22**, 3160-3164 (1983)
- 7 27 A.R. Tanguay, Jr.: "The Czochralski growth and optical properties of bismuth silicon oxide," Ph. D. dissertation (Yale University, New Haven, Conn., 1977)
- 7 28 G.C. Valley, M.B. Klein: *Opt. Eng.* **22**, 704-711 (1983)
- 7 29 G. Lesaux, J.C. Launay, A. Brun: *Opt. Commun.* **57**, 166-170 (1986)
- 7 30 R.A. Sprague: *J. Appl. Phys.* **46**, 1673-1678 (1975)
- 7 31 H. Kogelnik: *Bell Syst. Tech. J.* **48**, 2909-2947 (1969)
- 7 32 W.R. Klein, B.D. Cook: *IEEE Trans. Sonics and Ultrason.* **SU14**, 123-134 (1967)
- 7 33 E.I. Gordon: *Appl. Opt.* **5**, 1629-1639 (1966)
- 7 34 D. Yevick, L. Thylen: *J. Opt. Soc. Am.* **72**, 1084-1089 (1982)
- 7 35 R.V. Johnson, A.R. Tanguay, Jr.: *Opt. Eng.* **25**, 235-249 (1986)
- 7 36 F. Vachas, L. Hesselink: *J. Opt. Soc. Am.* **A1**, 1221 (1984)
- 7 37 A. Marrakchi, R.V. Johnson, A.R. Tanguay, Jr.: *IEEE J. Quant. Electron.* **QE-23**, 2142-2151 (1987)

3. ---

Fundamental Physical Limitations of the Photorefractive Grating Recording Sensitivity

R.V. Johnson* and A.R. Tanguay, Jr.

Optical Materials and Devices Laboratory
Departments of Electrical Engineering and Materials Science
and Center for Photonic Technology
University of Southern California
University Park
Los Angeles, California

Contents ---

1. Introduction	60
2. Factors Contributing to the Photorefractive Sensitivity	63
3. The Grating Recording Efficiency	68
4. Representative Grating Recording Efficiency Calculations	95
5. Conclusions	98
Acknowledgments	99
References	99

* Current affiliation: Crystal Technology, Inc., 1060 E. Meadow Circle, Palo Alto, CA 94303

1.

Introduction

A number of inherent inefficiencies exist in the photorefractive recording of grating structures due to the nature of the photoexcitation and charge transport processes. These inefficiencies can be quantified by postulating highly idealized photoexcitation and charge transport models that yield optimum (quantum limited) space charge field distributions, assuming a photoexcitation constraint of no more than one mobile charge per incident photon. By comparing such highly idealized photorefractive recording models with more realistic models, the fundamental origins of several such inherent inefficiency factors can be identified and their magnitudes estimated. In this manner, the grating recording efficiencies of photorefractive materials can be directly compared with the fundamental physical limitations imposed by quantum constraints.

In this chapter, we will establish the absolute quantum efficiency of the photorefractive grating recording process by deriving the optimum idealized photorefractive recording model subject to such quantum constraints. A more realistic photoexcitation and charge transport model applicable to numerous currently investigated real time photorefractive materials will then be examined in depth, with emphasis on a comparison with the characteristics of the optimum quantum limited model. This realistic charge transport model, based on the extensive previous work of numerous authors, is presented in such a manner as to illustrate its statistical nature and to provide a physically intuitive interpretation of its principal attributes.

Of all of the parameters that seek to quantify the absolute or relative performance of photorefractive materials, one of the most important is the photorefractive sensitivity (von der Linde and Glass, 1975; Micheron, 1978; Glass, 1978; Gunter, 1982; Yeh, 1987a and 1987b; Glass et al., 1987; Valley and Klein, 1983). This key parameter is typically defined in theoretical analyses either as the refractive index modulation obtained in writing a uniform grating of fixed spatial frequency per unit absorbed recording energy density (energy per unit volume) (von der Linde and Glass, 1975; Micheron, 1978; Glass, 1978; Gunter, 1982; Glass et al., 1987; Valley and Klein, 1983), or as the inverse of the recording energy density required to achieve a specified value of the diffraction efficiency for a uniform grating of fixed spatial frequency in a material of given thickness (Valley and Klein, 1983; Huignard and Micheron, 1976). Alternatively, for purposes of experimental measurement, the photorefractive sensitivity may be specified as the inverse of

the recording energy density required to reach a given fraction of the saturation diffraction efficiency of a particular material (Gunter, 1982; Amodi and Staebler, 1972). The photorefractive sensitivity, and the fundamental physical limitations that apply to it, are of current significant interest because they establish the maximum reconfiguration rate of volume holographic optical elements (VHOEs) (von der Linde and Glass, 1975; Tanguay, 1985) at constant average optical input power. The maximum reconfiguration rate of VHOEs is in turn important for applications ranging from massively parallel interconnections in optical processing and computing systems (Tanguay, 1985) to the development of the photorefractive incoherent-to-coherent optical converter (PICOC) (Kamshilin and Petrov, 1980; Shi et al., 1983; Marrakchi et al., 1985).

A number of factors contribute to the various photorefractive sensitivities characteristic of photoconductive, electrooptic materials. One such factor is the photogeneration quantum efficiency, which represents the number of photogenerated mobile charge carriers per photon absorbed from the recording beam(s). A second factor is the charge transport efficiency, which is a measure of the degree to which the average photogenerated mobile charge carrier contributes to the forming space charge grating after separation from its original site by means of drift and/or diffusion and subsequent trapping. The magnitude of the space charge field generated by a given space charge grating is inversely proportional to the dielectric permittivity ϵ of the photorefractive material, which thus contributes a third factor to the grating recording sensitivity. A fourth factor describes the perturbation of the local index ellipsoid (dielectric tensor at optical frequencies) that results from a given space charge field through the electrooptic (Pockels or Kerr) effect. And finally, a fifth factor pertains to the physical optics inherent in the readout process, whereby the diffraction efficiency and polarization properties of the readout beam are derived directly from the index ellipsoid modulation.

In addition, several other physical quantities factor into an evaluation of the photorefractive sensitivity, including the wavelength of the recording illumination (to convert the number of absorbed photons into an equivalent energy), the absorption coefficients of the material at the wavelengths of both the recording and readout beams (to correct for the fractional absorbance of the recording beams and the fractional transmittance of the readout beam), and the magnitude of the applied voltage (which significantly alters the sensitivity characteristics for certain materials by changing the nature of the dominant charge transport mechanism from the diffusion regime to the drift regime).

In previous treatments of the photorefractive sensitivity and its associated limits, all of the abovementioned factors have been addressed to some degree, though not necessarily within a single unified treatment, or even within a consistent set of constraints, assumptions, and approximations. Such a unification is also beyond the scope of the present work. In this chapter, we present the results of a study in which we have approached the photorefractive sensitivity issue from a somewhat different perspective: that of the *absolute quantum efficiency* of the photorefractive grating recording process. As such, we attempt to answer an oft-stated but as yet unanswered question as to the origin of the apparent insensitivity of photorefractive grating recording, particularly in comparison with the relatively high sensitivities characteristic of electrooptic spatial light modulators that utilize a similar combination of photoconductive charge separation and electrooptic modulation in the same single crystal materials (Tanguay, 1985).

In order to provide a quantitative metric that does in fact have a fundamental physical limitation, we define herein the *grating recording efficiency* of a photorefractive recording model or configuration as the magnitude of the space charge field produced by a fixed number of photogenerated mobile charge carriers at a given spatial frequency, *normalized by the maximum quantum limited space charge field that can be produced by the same number of photogenerated carriers*. This metric thus effectively combines the notions of a photogeneration efficiency and a charge transport efficiency, and it provides an estimate of the fundamental quantum efficiency of the photorefractive grating recording process as determined by the particular photoexcitation and charge transport mechanism invoked.

In order to fully utilize the concept of the grating recording efficiency, we first calculate the maximum quantum limited space charge field that can be produced by a fixed number of photogenerated mobile charge carriers, assuming optimum photoexcitation and redistribution (transport) functions. We then calculate the grating recording efficiencies that describe several important limiting cases with selected idealized photoexcitation and redistribution functions. This in turn allows for the definition of a baseline case against which more realistic charge transport models can be compared. Examination of a particular charge transport model as a test case then indicates several other generically applicable factors that act to further decrease the grating recording efficiency in various recording configurations. These results allow the above analysis to be conveniently utilized for a wide range of applications of current technological interest.

The organization of the remainder of this chapter is as follows. A brief

summary of the origin, material dependence, and implications of the several parameters that affect the photorefractive grating recording sensitivity is provided in the following section (Section 2). Alternative models of the photorefractive recording process are defined in Section 3, and the appropriate grating recording efficiencies are presented therein for each case. Assumptions and limitations common to all of the models are discussed in Section 3.1. Idealized photogeneration and charge transport models are defined and analyzed in Section 3.2, and more realistic models based upon the analyses of Young et al. (1974) and Moharam et al. (1979) in the initial stages of recording before significant space charge field amplitudes evolve, and of Kukhtarev (1976) for temporal evolution with small modulation depths of the illumination profile, are discussed and analyzed in Section 3.3. Representative grating recording efficiency calculations are presented in Section 4 for several common materials and applications in order to illustrate the effect of the quantum inefficiency factors on the overall quantum efficiency of photorefractive recording. Finally, conclusions drawn from the above analyses are discussed in Section 5.

2. --- Factors Contributing to the Photorefractive Sensitivity

In this section, each of the five factors outlined in the introduction that collectively determine the photorefractive sensitivity is briefly discussed, in order to provide a suitable context for the derivation of the quantum limited grating recording efficiency as presented in the following section. It should perhaps be emphasized at the outset that although each of the primary factors considered herein affects at least one of the aforementioned alternative photorefractive sensitivity parameters, not all of the factors enter into each defined parameter.

In the context of photorefractive grating recording, the photogeneration quantum efficiency is related to the fraction of the incident photon flux that generates mobile charge carriers free to participate in subsequent charge transport and trapping processes. For a given recording wavelength, several distinct photoexcitation processes can contribute to the total absorption coefficient. In most commonly considered models of the photorefractive effect, the dominant process is the photogeneration of free carriers from deep donor

or acceptor states, such that only one sign mobile carrier (either an electron or a hole) is liberated for each photoevent. In wavelength regions of significant photoconductivity, a second important process is the creation of electron-hole pairs (as well as excitons in certain materials and material structures) by means of band-to-band transitions. Examples of photoinduced transitions that are not likely to contribute to the photorefractive effect, and hence tend to reduce the photogeneration quantum efficiency, are intersub-band absorptions, intraionic level promotions, quantum well interlevel excitations, and photochromic charge transfer exchanges. Since each contributing process will in general be characterized by its own charge transport efficiency (discussed below), it is most appropriate to assign separate photogeneration quantum efficiencies not only to each charge carrier type, but also to each distinct photoproduction origin (or photoexcitation channel).

The inherent absorptive inefficiency implied by a finite thickness photorefractive medium also affects the overall photogeneration quantum efficiency. In the case of a thin crystal (such that $\alpha d \ll 1$, in which α is the absorption coefficient at the recording wavelength and d is the crystal thickness), only a small fraction ($= \alpha d$) of the recording beam intensity is absorbed and hence has an opportunity to participate in the photorefractive process. In a thicker crystal, for which the thickness may be optimized for maximum saturation diffraction efficiency, the absorbed fraction is $[1 - e^{-\alpha d}]$ if only the entrance surface of the photorefractive medium is allowed to achieve saturation. In this case the recorded grating will exhibit an exponential nonuniformity throughout the crystal thickness that will decrease the maximum achievable diffraction efficiency. If the entire crystal is exposed to saturation, the photogeneration quantum efficiency will be further reduced by a factor of order $e^{-\alpha d}$. Finally, the effects of reflection at both front and rear crystal surfaces reduce the fraction of incident photons that contribute to the formation of a given grating component and, hence, also reduce the effective quantum efficiency. For a crystal thickness optimized for maximum saturation diffraction efficiency with equal write and read wavelengths, and for indices of refraction typical of common photorefractive materials, the combined effects of absorption of the recording beams, absorption of the readout beam, and reflection losses (assuming uncoated surfaces) on the photorefractive sensitivity is approximately an order of magnitude.

An additional effect that acts to reduce the photogeneration quantum efficiency is the constraint imposed indirectly by the nature of the photoexcitation distribution. As we shall demonstrate in the next section, the sinusoidal intensity interference pattern generated by two coherent recording

beams is not the optimum (quantum limited) photoexcitation distribution function.

Perhaps the most critical factor that determines the photorefractive sensitivity is the charge transport efficiency, in that this quantity, above all others, exhibits the greatest degree of variation among commonly investigated photorefractive materials. The charge transport efficiency quantifies the degree to which the average photoproduced charge carrier contributes to the forming space charge grating following photoexcitation, charge transport, and subsequent recombination or trapping. Charge transport in the refractory oxides, as well as in the compound semiconductor family, is a statistical process in which the net result of a given photoinduced event may be to increase, decrease, or leave unchanged the magnitude of the space charge modulation at the fundamental grating frequency.

The statistical nature of the charge transport process can be subsumed in the standard band transport treatment (Young et al., 1974; Moharam et al., 1979; Kukhtarev et al., 1976, 1979), or made explicit as in the hopping conduction model (Feinberg et al., 1980); both approaches lead to essentially equivalent results (Feinberg et al., 1980; Jaura et al., 1986). The charge transport efficiency is strongly affected in photorefractive materials by the dominant conduction mechanism (diffusion and/or drift in an externally applied field), and by the ratio of the average displacement of a photoexcited carrier (before recombination or trapping) to the grating spacing. In both the drift and diffusion regimes, the average displacement depends primarily on the mobility-lifetime product of the photoexcited species. As such, separate charge transport efficiencies should be assigned to each carrier type. A significant net charge transport efficiency will be realized only if there is a net differential in the carrier displacement and trapping process.

It should be noted that there are several situations that can yield vanishing charge transport efficiencies, even with large photogeneration quantum efficiencies. For example, in a single donor/single trap model, if the initial Fermi level is more than $10 k_B T$ or so above the un-ionized donor level, no substantial charge rearrangement is possible due to the unavailability of ionized donors (traps) outside the regions of significant photoexcitation. Likewise, if the initial Fermi level is more than $10 k_B T$ or so below the un-ionized donor level, only band-to-band photoexcitations can contribute with nonvanishing photogeneration quantum efficiencies, which again is likely to yield negligible charge transport efficiency unless the mobilities and/or lifetimes are significantly different, or unless operation in the drift regime is engendered by employing an externally applied electric field.

The space charge grating that results from the combination of photoexcitation and charge transport processes in turn gives rise to a modulation of the local electric field at the same spatial frequency through the first Maxwell equation:

$$\nabla \cdot [\epsilon \mathbf{E}(x)] = \frac{\rho(x)}{\epsilon_0}, \quad (3.1)$$

in which $\mathbf{E}$ is the total electric field at each point in space x , ϵ is the dielectric permittivity tensor, ρ is the local space charge amplitude, and ϵ_0 is the dielectric permittivity of free space. Note that the magnitude of the space charge field derived from a given space charge grating amplitude is inversely proportional to the grating wave vector, as implied by the differential relationship expressed in Eq. (3.1). The tensor character of ϵ is important to note, as many photorefractive materials (particularly the ferroelectric oxides) exhibit marked dielectric anisotropy. Hence, space charge gratings oriented in different directions within the same crystal can give rise to quite large variations in the resultant space charge field. Note further that the magnitude of the space charge field scales inversely with a diagonal component of the dielectric permittivity for a given space charge grating oriented along a principal dielectric axis of the crystal. Thus, materials with large dielectric constants (such as BaTiO_3 and SBN) require correspondingly large space charge amplitudes in order to produce an internal electric field modulation of given amplitude.

In the types of photorefractive materials considered herein, the index of refraction is a function of the local electric field. This dependence can arise from a number of electrorefractive effects, including among others the linear electrooptic (Pockels) effect, the quadratic electrooptic (Kerr) effect, the Franz-Keldysh effect, and the quantum confined Stark effect (Chemla et al., 1985). In some materials, notably multiple quantum well structures in compound semiconductors, more than one such electrorefractive effect can contribute simultaneously to the establishment of the resultant index perturbation. For our purposes herein, we consider only the linear electrooptic effect, in which the change in the dielectric impermeability tensor $\mathbf{B}$ (the inverse of the dielectric tensor ϵ) is linear in the electric field:

$$\Delta B_{ij} = \Delta(\epsilon^{-1})_{ij} = r_{ijk} E_k, \quad (3.2)$$

in which the Einstein summation rule is implied, and in which r_{ijk} is the third rank tensor representing the electrooptic coefficient (Kaminow, 1974). As shown in Eq. (3.2), the tensor nature of the electrooptic effect implies

a dependence of the effective index of refraction on the orientation of the grating within the crystal, as well as on the direction of propagation of the readout beam and its polarization. Since in general one can derive an effective electrooptic coefficient for a given experimental configuration, Eq. (3.2) may be rewritten in the form

$$\Delta n_{\text{eff}}(x) = -\frac{1}{2} n_0^3 r_{\text{eff}} E(x) \quad (3.3)$$

in which the x coordinate is taken parallel to the grating wave vector, $\Delta n_{\text{eff}}(x)$ is the effective index modulation resulting from the combination of the space charge field and the readout configuration, and n_0 is the corresponding unperturbed index at the wavelength of the readout beam.

In discussions of the photorefractive sensitivity, it is of considerable value to combine the effects of the previous two factors, since it has been shown that for a wide range of common photorefractive materials, the ratio $n_0^3 r_{\text{eff}}/\epsilon$ exhibits considerably reduced variation compared with that of each parameter separately (Glass et al., 1984; Glass, 1984). This is indicative of the general observation that materials with large static polarizabilities typically also exhibit concomitantly large perturbations of the dielectric tensor at optical frequencies in response to low frequency applied (or internal) electric fields.

Once the magnitude of the index perturbation is established, the resultant diffraction efficiency can be directly determined from the thickness (and uniformity) of the grating, the grating wave vector, the readout wavelength, and the corresponding absorption coefficient. Provided that the grating structure is sufficiently thick to assure diffraction in the Bragg regime and that the phase and amplitude distortions associated with self-diffraction effects can be neglected (Kukhtarev et al., 1979; Marrakchi et al., 1987), the diffraction efficiency is given by (Kogelnik, 1969)

$$\eta = \exp\left(\frac{-\alpha d}{\cos \theta_B}\right) \sin^2\left(\frac{\pi \Delta n d}{\lambda \cos \theta_B}\right), \quad (3.4)$$

in which θ_B is the Bragg angle. The first term in this expression denotes the inherent inefficiency associated with finite absorption at the readout wavelength, while the second term derives from the grating-modulation-induced diffraction process. Since for suitably small values of the argument $\pi \Delta n d / \lambda \cos \theta_B$ the diffraction efficiency scales as the square of both the index modulation and the grating thickness, this photorefractive sensitivity factor is inherently nonlinear and must be utilized with considerable caution.

Until this point in the discussion, we have assumed implicitly that the grating recording exposures and spatial frequencies employed have been large enough and small enough, respectively, to keep the photorefractive recording process outside the additional limitations imposed by quantum statistical fluctuations. For the recording of photorefractive gratings at very high spatial frequencies and at very low grating recording exposures, several additional factors will come into play, including statistical fluctuations in the photogeneration process, corresponding fluctuations in the charge transport and trapping processes that yield a locally inhomogeneous charge distribution with spatial frequency components near that of the grating wave vector, and concomitant variations in the direction and magnitude of the local space charge field. These additional factors may act to further reduce the overall photorefractive sensitivity.

3.

The Grating Recording Efficiency

As defined in the introduction (Section 1), the grating recording efficiency of a given photorefractive recording model or configuration is the ratio between the magnitude of the space charge field at a given spatial frequency produced by a fixed number of photogenerated mobile charge carriers, and the maximum quantum limited space charge field that can be produced by the same number of photogenerated carriers. The grating recording efficiency is introduced as a useful metric that effectively compares the photogeneration and charge transport efficiencies of any given model with the optimum quantum limited case, and as such provides an estimate of the overall fundamental quantum efficiency of the photorefractive grating recording process. Note that since the grating recording efficiency is normalized, any effect of the dielectric permittivity tensor in establishing the magnitudes of the space charge fields cancels out. Hence the grating recording efficiency may be equivalently defined directly in terms of the space charge grating amplitudes or in terms of the space charge fields for a dielectric constant of unity.

In this section, idealized photogeneration and charge transport models are postulated and analyzed in order to determine the maximum quantum limited space charge amplitude that can be produced by a fixed number of photogenerated mobile charge carriers at a given spatial frequency. The grating recording efficiencies of several such idealized models are then calculated

to form a set of baseline cases against which the corresponding efficiencies of more realistic photorefractive recording models can be directly compared. For one such model, that of a single mobile charge species transported between a single type of donor site and its associated (ionized) trap sites, several factors are identified that contribute to the grating recording efficiency and that are particularly illustrative of the fundamental limitations on the photorefractive sensitivity inherent in the model.

3.1. Constraints Common to Alternative Models of Photorefractive Recording

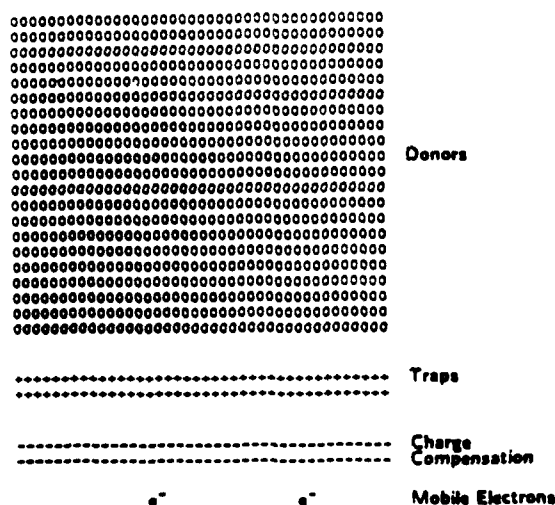
The idealized photogeneration and charge transport models defined herein are abstractions of a more realistic model that has been studied by many authors and has been analyzed in considerable detail by Kukhtarev et al. (1976, 1979), and as such these models share key assumptions concerning quantum recording limitations and implications of the readout process. For simplicity, we confine our attention herein to a version of the model characterized by a single mobile charge species, and a single type of donor site with associated un-ionized donor and ionized donor (trap) states, although more intricate models have been proposed for particular materials systems [e.g., the existence of mobile holes as well as electrons in lithium niobate (Orlowski and Kratzig, 1978) and barium titanate (Strohkendl et al., 1986), or the existence of multiple trap levels in bismuth silicon oxide (Attard and Brown, 1986; Valley, 1986)].

Four principal material species are considered in the single mobile charge/single trap level model, as diagrammed in Fig. 3.1. The mobile charge species (usually electrons) has number density $n(x, t)$ and is represented by the symbol e^- in Fig. 3.1. It is assumed that negligible densities of these mobile charges exist under dark conditions; essentially all mobile charges are created by photoionization of donors. In this model, *donors* and *traps* are assumed to be different valence states of the same impurity atom (e.g., iron in lithium niobate) or lattice defects, as diagrammed in Fig. 3.2. The total number of such impurity ions or defects is distributed uniformly throughout the crystal at potential donor sites. A donor is converted into a trap (ionized donor), simultaneously with the creation of a mobile charge, by photogeneration; conversely, a mobile charge is removed from the conduction band, and a trap is converted into a donor, by recombination (Fig. 3.2). The sum of the number densities of donors and ionized donors is denoted by N_D , which is therefore the total density of potential donor sites and is assumed to be constant in space and time. If the number density of ionized donors

is taken to be $N_D^+(x, t)$, then the number density of (un-ionized) donors is $[N_D - N_D^+(x, t)]$. In Fig. 3.1, the donors are represented by the symbol O and the traps are represented by the symbol $\cdot$. A fourth material species is needed in this model to ensure charge conservation, as the density of ionized donors $N_D^+(x, t)$ frequently exceeds the density of mobile charges $n(x, t)$ by several orders of magnitude. This fourth species, with number density N_A , has traditionally been called an *acceptor*, but is also called a *charge compensation site* herein and is represented in Fig. 3.1 by the symbol $-$. The charge compensation sites are electrically negative with respect to the donors and are presumed to be negatively charged impurity ions or lattice defects incorporated during the crystal growth process. The density N_A is assumed to be constant in space and time, and is further assumed to be numerically equal to $N_D^+(x, t = 0)$, the concentration of ionized donors in equilibrium. In point of fact, the only requirement for charge compensation is that the product of the charge compensation site density and the effective number of negative charges on each be equal to $N_D^+(x, t = 0)$.

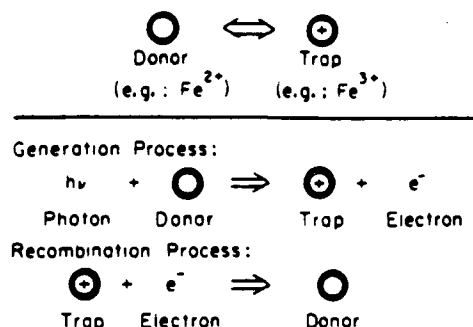
It is instructive to consider the relative densities of these various species. Consider, for example, a crystal of bismuth silicon oxide ($\text{Bi}_{12}\text{SiO}_{20}$). Typ-

Fig. 3.1.



Schematic representation of the single mobile charge/single trap level photo-refractive recording model. Indicating both the uniform distributions and the relative densities of the four principal material species involved prior to grating recording.

Fig. 3.2.



Schematic representation of the single active species photorefractive recording model, indicating the electron photogeneration (and recombination) processes involved in the transformation of donors into traps (and traps into donors).

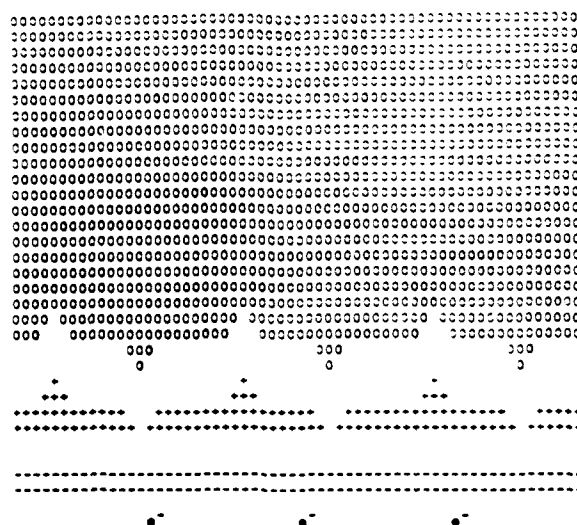
ical concentrations of donors, charge compensation sites, and mobile charges under reasonable optical intensities are of order 10^{19} cm^{-3} , 10^{16} cm^{-3} , and no more than 10^{12} cm^{-3} , respectively (Peltier and Micheron, 1977; see also Hou et al., 1973). (The number density of electrons under dark conditions for typical crystals of bismuth silicon oxide is entirely negligible.) Thus, $N_D \gg N_D^+ = N_A \gg n$. A number of other photorefractive media exhibit similar proportionalities. For these materials, the total space charge density $\rho(x, t)$ is given by

$$\rho(x, t) = e[N_D^+(x, t) - n(x, t) - N_A] \approx e[N_D^+(x, t) - N_A]. \quad (3.5)$$

The local density of ionized donors N_D^+ can be spatially redistributed under the influence of inhomogeneous photogeneration, as shown by comparing Figs. 3.1 and 3.3. This redistribution occurs as follows. A photon is absorbed by a donor, converting the donor into a trap (ionized donor) and generating a mobile charge carrier. The mobile charge carrier is transported some distance through the photorefractive medium due to drift and/or diffusion, and it is subsequently captured by a trap thus generating another donor. If we choose to follow a particular electron, the entire process can be viewed as a simple exchange between equivalent donor sites of a donor state with an ionized donor (or trap) state.

As a final note, the space charge profile $\rho(x, t)$ might evolve into a highly distorted profile with respect to the distribution of incident illumination because of nonlinearities inherent in the recording process. However, since the grating readout is typically performed deep within the Bragg regime, at most

Fig. 3.3.



Schematic representation of the single mobile charge/single trap level photorefractive recording model, indicating the spatial redistribution of the four principal material species following grating recording. Note that the sum of the donor and trap densities is space-invariant. The symbols identifying each species are given in Fig. 3.1.

a very limited range of spatial frequencies is effective in diffracting the read-out light. Thus only one spatial harmonic of the space charge field is of interest, which is assumed herein to be the fundamental harmonic.

In summary, the photorefractive recording model studied by Kukhtarev and all idealized photogeneration and charge transport abstractions considered in the following analysis share these basic assumptions: a) only a finite amount of space charge [$N_D^+(x, t = 0) = N_A$] exists for generating the space charge electric field, b) this space charge can be spatially redistributed under the influence of an illumination pattern, c) no more than one mobile charge is generated for each absorbed photon of the illumination beam, and d) only the fundamental spatial harmonic of the space charge is effective in the holographic readout process. The photorefractive recording model and the idealized models differ, however, in the details of the photogeneration and charge transport processes, as discussed next.

3.2. Idealized Photogeneration and Charge Transport Models

The grating recording efficiency comprises three successive physical processes: photogeneration, charge transport, and trapping. In order to deter-

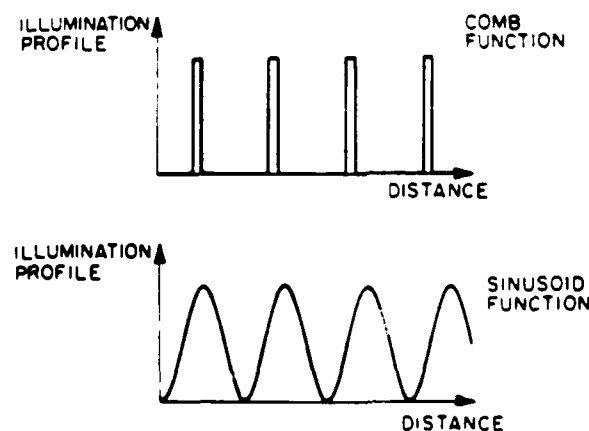
mine the maximum quantum limited space charge field (at unity dielectric constant) that can result from a fixed number of photogenerated mobile charges, we investigate four idealized photogeneration and charge transport models, as defined below.

The photogeneration process is controlled by the recording illumination profile. Two alternative profiles are considered herein, a periodic comb function and a sinusoidal function. The comb illumination profile $I_C(x)$, as shown in Fig. 3.4, is defined by

$$I_C(x) = I_{C0} \text{comb}\left(\frac{x}{\Lambda_G}\right) \Rightarrow I_{C0} \sum \delta\left(p + \frac{x}{\Lambda_G}\right), \quad (3.6)$$

in which the arrow $\Rightarrow$ indicates that the desired function asymptotically approaches a sequence of Dirac delta functions, i.e., a series of intensity peaks with spatial extent small compared with the spatial period Λ_G . Unlike a mathematical delta function, however, the intended comb peaks are assumed to be large enough to overlap a reasonable number of donor sites. Note that the periodic comb function does not correspond to any normal recording configuration. This is acceptable for purposes of this analysis because the maximum quantum limited space charge field at unity dielectric constant is intended to define an upper limit against which more realistic models might

Fig. 3.4.



Schematic representation of the two principal illumination profiles (comb and sinusoid functions) utilized in the idealized photogeneration and charge transport models. The comb function is illustrated with finite width to incorporate a given number of donors at a predetermined donor density.

be compared. The sinusoid illumination profile $I_S(x)$ of unity modulation depth is defined by

$$I_S(x) = I_{S0} \left[1 + \cos\left(\frac{2\pi x}{\Lambda_G}\right) \right], \quad (3.7)$$

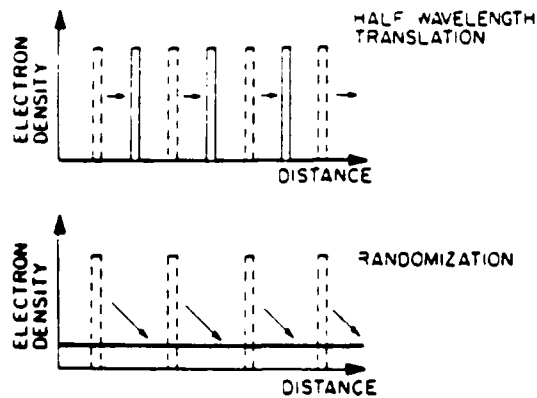
in which Λ_G is the spatial period of the illumination profile. To compare these two illumination profiles, the same photon flux is assumed; that is, the scaling parameters I_{S0} and I_{C0} are adjusted such that the following normalization integral is satisfied:

$$\int I_S(x) dx = \int I_C(x) dx, \quad (3.8)$$

from which we derive the relation that $I_{S0} = I_{C0} \equiv I_0$. In the idealized photogeneration models, all photons in the illumination profile are assumed to be absorbed and to generate mobile charge carriers.

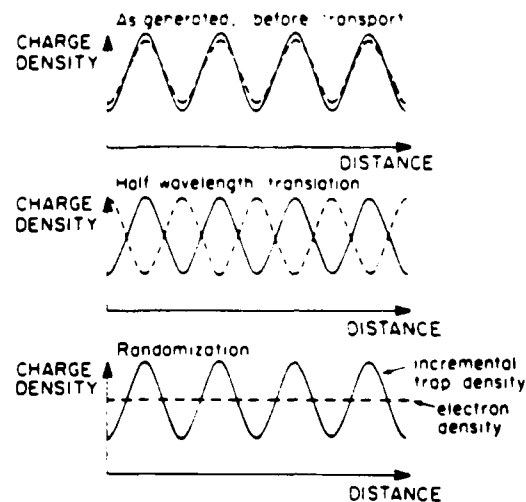
Two alternative models of charge transport and trapping are also considered: half wavelength translation, and randomization (Figs. 3.5 and 3.6). In the half wavelength translation process shown schematically in Fig. 3.5, each electron is assumed to translate precisely half the grating period of the illumination profile before capture, without diffusive blooming of the elec-

Fig. 3.5.



Schematic representation of the two principal charge redistribution mechanisms (half wavelength translation and randomization) utilized in the idealized photogeneration and charge transport models, for the case of comb illumination. A single mobile charge species (electrons) is assumed for purposes of illustration.

Fig. 3.6.



Schematic representation of the two principal charge redistribution mechanisms (half wavelength translation and randomization) utilized in the idealized photogeneration and charge transport models, for the case of sinusoidal illumination.

tron cloud. In the analysis that follows, it will be shown that half wavelength translation corresponds to the quantum limited charge transport process, giving rise to the maximum grating recording efficiency. The second idealized transport and capture process to be considered herein is complete randomization, as shown schematically in Fig. 3.6, in which the photogenerated electrons are assumed to redistribute randomly until they exhibit no spatial variation in density, and only then are recaptured by local traps. The randomization model is intended to represent a more realistic charge transport model; even so, this model corresponds to an *upper* bound on realizable charge transport efficiencies, as shown in the next section.

Combination of the two photogeneration models and the two charge transport and trapping models produces four alternative idealized recording models for comparison (Fig. 3.7), which will hereinafter be identified as the bipolar comb, the monopolar comb, the transport-efficient sinusoid, and the baseline sinusoid. Each of the four combinations will be considered in turn. Grating recording efficiencies are calculated for these four combinations in Section 3.2.1, and the corresponding space-charge-field saturation behavior is considered in Section 3.2.2.

Fig. 3.7.

		Illumination Profile	
		Comb	Sinusoid
Charge Transport	Randomization	Monopolar Comb	Baseline Sinusoid
	Half Wavelength Translation	Bipolar Comb	Transport-Efficient Sinusoid

Matrix representation of the four idealized photogeneration and charge transport models, as derived from the principal illumination profiles and charge transport mechanisms.

3.2.1. Idealized Grating Recording Efficiencies

The bipolar comb results from a periodic comb illumination function combined with a half wavelength translation. The resulting space charge density $\rho(x)$ induced by this illumination and transport combination is a periodic comb function superimposed on a periodic comb function of opposite sign shifted by half of a grating wavelength, as shown in Fig. 3.8. For a given peak photogenerated charge density ρ_0 (proportional to $I_0 t$, in which t is the exposure time), the resultant space charge distribution $\rho(x)$ is given by

$$\rho(x) = \rho_0 \left[\text{comb}\left(\frac{x}{\Lambda_G}\right) - \text{comb}\left(\frac{1}{2} + \frac{x}{\Lambda_G}\right) \right]. \quad (3.9)$$

This space charge profile $\rho(x)$ can be readily integrated [see Eq. (3.1)] to yield a space charge field $E(x)$ that is a square pulse train (Fig. 3.8). The first spatial harmonic component, E_1 , defined by

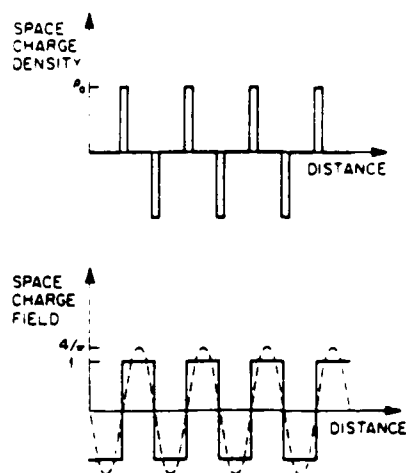
$$E_1 = 2\Lambda_G^{-1} \int E(x) \sin\left(\frac{2\pi x}{\Lambda_G}\right) dx, \quad (3.10)$$

has a magnitude of

$$E_1 = 4e \frac{\rho_0}{\epsilon \epsilon_0 K_G}, \quad (3.11)$$

in which $K_G = 2\pi/\Lambda_G$ is the wave vector of the illumination profile.

Fig. 3.8.



Space charge density and field profiles for the bipolar comb distribution. The first spatial harmonic of the space charge field is represented by the dashed curve and is scaled to the magnitude of the total space charge field.

The monopolar comb commands interest because it can support the highest space charge field before saturation due to limited ionized trap density, as discussed in the next section. The monopolar comb results from a periodic comb illumination function combined with electron randomization. The resulting space charge profile $\rho(x)$ induced by this illumination and transport combination is a periodic comb function superimposed on a uniform background of opposite charge, as shown in Fig. 3.9. The resulting charge distribution is represented by

$$\rho(x) = \rho_0 \left[\text{comb} \left(\frac{x}{\Lambda_G} \right) - 1 \right]. \quad (3.12)$$

The corresponding space charge field exhibits a sawtooth profile (Fig. 3.9), and the magnitude E_1 of its first spatial harmonic is

$$E_1 = 2e \frac{\rho_0}{\epsilon \epsilon_0 K_G}. \quad (3.13)$$

The transport-efficient sinusoid results from a sinusoidal illumination profile combined with a half wavelength translation, as shown schematically in Fig. 3.10, leading to a charge density $\rho(x)$ given by

$$\rho(x) = 2\rho_0 \cos\left(\frac{2\pi x}{\lambda_G}\right) \quad (3.14)$$

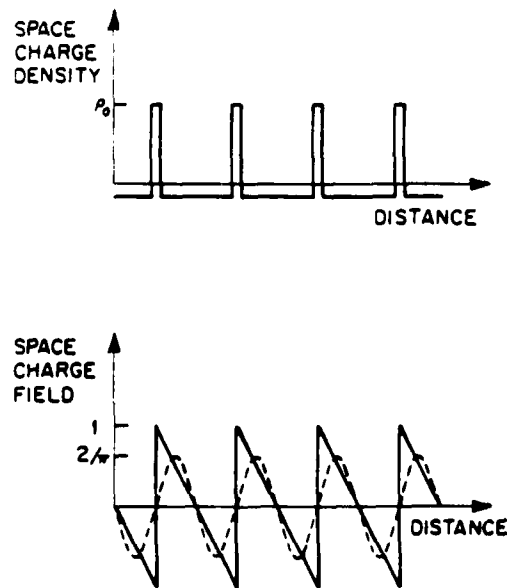
and a first spatial harmonic field component E_1 of

$$E_1 = 2e \frac{\rho_0}{\epsilon\epsilon_0 K_G} \quad (3.15)$$

The transport-efficient sinusoid combination is included for completeness, but is not emphasized because it is neither realistic nor does it correspond to any upper bound of grating recording efficiency or saturation performance.

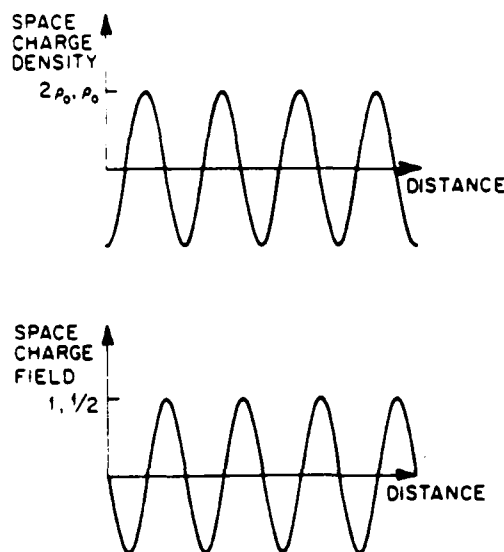
The baseline sinusoid, although seemingly artificial, is of very pronounced interest because it represents an asymptotic upper bound on the grating recording efficiency predicted by more realistic recording models, as discussed in Section 3.3 below. The baseline sinusoid results from a sinusoidal illumination profile combined with randomization of the electron distribution. The resulting space charge distribution exhibits a sinusoidal

Fig. 3.9.



Space charge density and field profiles for the monopolar comb distribution. The first spatial harmonic of the space charge field is represented by the dashed curve and is scaled to the magnitude of the total space charge field.

Fig. 3.10.



Space charge density and field profiles for the transport-efficient (first scale values) and baseline sinusoid (second scale values) distributions. The space charge fields are scaled to the magnitude of the transport-efficient case.

profile, as does the corresponding space charge field, as shown in Fig. 3.10. The charge density is given by:

$$\rho(x) = \rho_0 \cos\left(\frac{2\pi x}{\Lambda_G}\right), \quad (3.16)$$

corresponding to a first spatial harmonic field component E_1 of

$$E_1 = e \frac{\rho_0}{\epsilon \epsilon_0 K_G}. \quad (3.17)$$

As can be seen from the analysis above, the bipolar comb photogeneration/charge transport combination generates the maximum quantum limited space charge field at unity dielectric constant, and hence provides the normalization constant required for evaluation of the grating recording efficiencies of both idealized and realistic photorefractive recording models. Physically, this optimum combination of photogeneration and charge transport results from the maximum possible average separation of the positive and negative charge distributions, as well as from the fact that the amplitude

of the first harmonic of a square wave exceeds the amplitude of the square wave itself.

The grating recording efficiencies for the four idealized photorefractive grating recording combinations are shown in Table 3.1. The most quantum-efficient recording occurs for the bipolar comb, which therefore is assigned a grating recording efficiency of unity. The next most efficient configurations are the monopolar comb and the transport-efficient sinusoid, which both yield a grating recording efficiency of 0.5. The least efficient configuration is the baseline sinusoid, with a grating recording efficiency of 0.25. As we shall show later, the efficiency of the baseline sinusoid is an asymptotic upper limit of more realistic recording models. Table 3.2 gives the relative photorefractive grating recording sensitivities for these four cases in terms of diffraction efficiency per unit incident photon flux, assuming low diffraction efficiencies such that the efficiency is proportional to the square of the space charge field component E_1 . Note that by this measure the sensitivity of the baseline sinusoid is degraded by a factor of 16, over an order of magnitude, compared with the quantum limit represented by the bipolar comb charge distribution function.

3.2.2. Space Charge Saturation for the Idealized Models

Not only is the photorefractive recording sensitivity of concern, but also saturation limitations of the space charge field occurring because of limited ionized trap density. Consider, for example, the bipolar comb example shown in Fig. 3.8. The regions of positive space charge grow by the photoexcitation of neutral donors, which converts them into positively ionized donors (traps) and generates mobile electrons. The electrons are removed from this region by the various transport processes, leaving behind the positively ionized donors. As will be shown later, it is these regions of net positive space charge that primarily contribute to the buildup of the space charge field. The regions of negative space charge grow by the reverse process, i.e., by capturing mobile electrons at local ionized donor sites to form neutral donors. In pho-

TABLE 3.1

Grating Recording Efficiencies of the Various Idealized Models

Bipolar comb	1.0
Monopolar comb	0.5
Transport-efficient sinusoid	0.5
Baseline sinusoid	0.25

TABLE 3.2

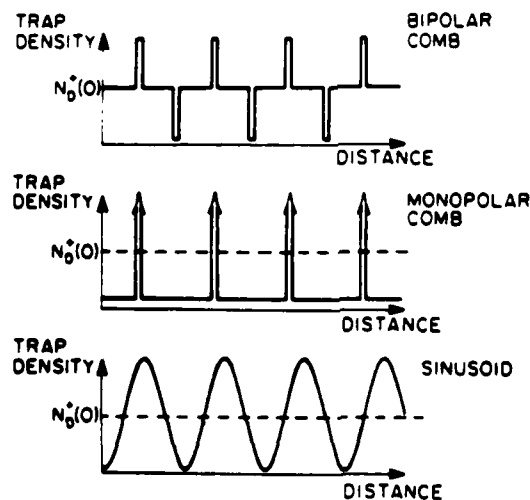
Relative Diffraction Efficiencies of the Various Idealized Models

Bipolar comb	1.0
Monopolar comb	0.25
Transport-efficient sinusoid	0.25
Baseline sinusoid	0.0625

torefractive crystals that initially are in quasi-thermodynamic equilibrium, the density of donors typically far exceeds the density of ionized donors, which implies that the regions of negative space charge will saturate first. This corresponds to the complete conversion of all ionized donors into neutral donors at the peaks of the negative space charge distribution, resulting in the local complete cancellation of $N_D^-(x, t = 0) = N_A$, as shown schematically in Fig. 3.11.

The bipolar comb combination exhibits by far the most rapid charge saturation at the lowest space charge field of the four combinations considered herein, because for this combination the electrons after transport are concentrated into a very small volume, with a correspondingly small number

Fig. 3.11.



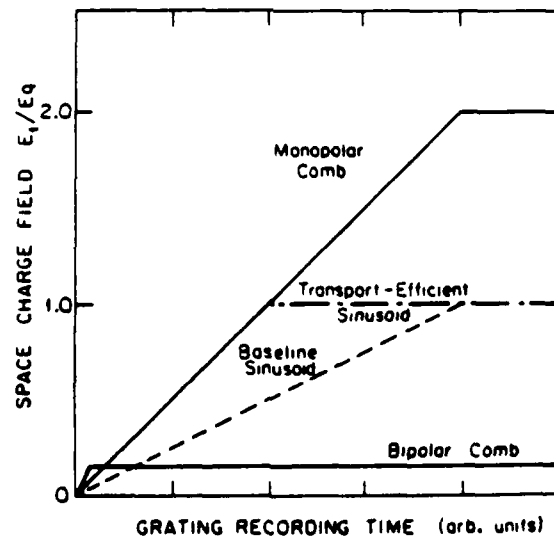
Saturation of the space charge modulation for the bipolar comb, monopolar comb, and sinusoid charge distributions as a result of the finite available trap density. An initially uniform trap density of $N_D^-(0)$ is assumed prior to photo-generation and charge redistribution.

of ionized donors available to capture these electrons. Conversely, the monopolar comb combination exhibits the slowest saturation at the highest space charge field because the electrons are uniformly distributed throughout the volume of the photorefractive medium and, hence, can be captured by all of the ionized donors. The saturation characteristics of the sinusoidal combinations are intermediate between the bipolar and the monopolar combinations.

The relative photorefractive grating recording sensitivities (grating recording efficiencies) and saturation characteristics of these four photogeneration/charge transport combinations are schematically diagrammed in Fig. 3.12. The photosensitivities are indicated by the initial linear slopes and the saturation by the final space charge field levels. The space charge field is plotted here in units of E_q , which is defined as $eN_A/\epsilon\epsilon_0K_G$ (Amodei, 1971). (This expression is valid when the density of charge compensation sites N_A is much smaller than the total density of potential donor sites N_D .) Note that the saturation level for the bipolar comb should really be much closer to the horizontal axis; it has been overstated for clarity of illustration.

Having calculated the grating recording efficiencies and saturation fields for these four highly idealized cases, we now proceed to compare these ideal

Fig. 3.12.



Space charge field as a function of grating recording time, showing the initial linear growth (photorefractive sensitivity) and subsequent saturation regimes.

results with more realistic photogeneration and charge transport models in the next two sections.

3.3. Realistic Recording Models

The idealized photorefractive recording models discussed above allow the effects of the photogeneration profile and charge transport process on the grating recording efficiency to be assessed relative to the fundamental quantum limits. We now examine a more realistic model applicable to a wide range of commonly investigated photorefractive media, which exhibits an overall efficiency degraded from that of the baseline sinusoid case presented above by several additional factors. In this model, the photogeneration profile is assumed to be sinusoidal, and the effects of the photogeneration quantum efficiency, absorption coefficients, and reflection losses are assumed to be space-invariant efficiency factors and hence can be directly incorporated in any estimate of the photorefractive grating recording sensitivity. Another major factor is an inherent inefficiency in the charge transport process; this inefficiency is studied in Section 3.3.1 using analytical solutions derived by Young et al., (1974; see also Moharam et al., 1979), which are valid in the initial recording interval before significant space charge fields have evolved. A related factor derives from the reduced recording sensitivity exhibited as the space charge field approaches its steady state limit. This is studied in Section 3.3.2 using analytical solutions derived by Kukhtarev (1976) that describe the temporal evolution of the space charge field in the limit of very low illumination profile modulation depths. Recording configurations that generate low modulation depths are inherently inefficient, as most of the photons in the illumination contribute a uniform background photocurrent, and only a fraction of the incident intensity contributes to the spatial structure of the image. The modulation depth, therefore, is also a factor that reduces the photorefractive grating recording sensitivity. However, low modulation depths are necessary in certain recording techniques for enhancing the space charge field in the steady state limit, such as the running grating process discussed in Section 3.3.3.

3.3.1. Initial Recording Sensitivity

To determine the existence of degraded charge transport efficiency, the idealized recording models discussed in Section 3.2 must be compared with more realistic charge transport solutions, such as those given by Young et al. (1974; see also Moharam et al., 1979). Under normal recording condi-

tions, the coupled photorefractive recording equations are nonlinear, making analytical solutions difficult or impossible to derive. However, a significantly simplified analysis can be utilized during the initial recording period, which enables analytical solutions to be derived at least for certain illumination profiles. These analytical solutions are well worth studying for the physical insight they furnish into the charge transport process.

The analytical simplification described above derives from a linearization of the recording equations, in which two recording parameters, the total electric field and the ionized donor (trap) density, remain essentially constant throughout space during the initial recording interval (Young et al., 1974; Moharam et al., 1979). This assumes that the photorefractive crystal is initially in quasi-thermodynamic equilibrium, i.e., with a spatially uniform distribution of ionized donors $N_D^+(x, t = 0) = N_A$, and that the trap density remains essentially constant throughout this initial recording interval.

The analytical solutions that exist in this regime have typically emphasized recording with sinusoidal illumination profiles. While such a profile corresponds closely with typical experimental situations and simplifies the mathematics, the physics of the transport process is somewhat obscured in comparison with an alternative illumination profile, that of a very narrow slit, which can be viewed as an approximation of a Dirac delta function. Typical trapped electron density profiles obtained in response to a narrow slit illumination profile are shown in Fig. 3.13 for the cases of diffusion-only transport (top illustration), drift-only transport (middle illustration), and one particular combination of drift and diffusion processes (bottom illustration); the derivation of these figures is described below. These figures emphasize several important features of a more realistic transport analysis. The transport mechanism is inherently a random process, with a spread in characteristic transport lengths associated with a corresponding spread in charge carrier lifetimes. Useful parameters for characterizing the transport processes are the average transport lengths L_E for drift-induced transport and L_D for diffusive transport, defined as (Young et al., 1974; Moharam et al., 1979)

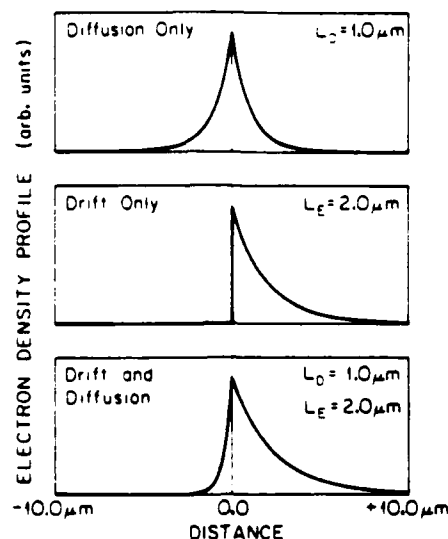
$$L_E = \mu\tau E_0, \quad (3.18)$$

and

$$L_D = (D\tau)^{1/2} = \left(\frac{k_B T}{e}\right)^{1/2} (\mu\tau)^{1/2}, \quad (3.19)$$

in which μ is the mobility of the charge carriers, E_0 is the applied bias electric field, τ is the charge carrier lifetime, D is the diffusion coefficient

Fig. 3.13.



Typical trapped electron density profiles obtained in response to a narrow slit illumination profile. The average transport lengths L_D for diffusive transport and L_E for drift-induced transport are defined in the text.

for the charge carriers, k_B is Boltzmann's constant, T is the crystal temperature, and e is the charge of an electron. Einstein's relation between the diffusion coefficient and the mobility has been used in Eq. (3.19). Note that in both cases the transport lengths are functions of the $\mu\tau$ product.

For the diffusion-only case, the electron spread is symmetrical, which introduces no net phase shift in the photorefractive response to any arbitrary illumination profile. For the drift case, as well as for the combined drift/diffusion case, the electron distribution is skewed to one side by the presence of an applied bias field, which does in fact introduce a phase shift when recording particular illumination profiles, as shown in Fig. 3.13.

Now consider an illumination profile $I(x)$ that is sinusoidal and of the form

$$I(x) = I_0[1 + m \cos(K_G x)], \quad (3.20)$$

in which m is the modulation depth of the light profile and K_G is the wave vector associated with the interference pattern. For such an illumination profile, Young et al. (1974; see also Moharam et al., 1979) have derived expressions for the growth of the space charge field during the initial recording interval. When transport is dominated by diffusion, the initial growth

of the first harmonic component E_1 of the space charge field is expressed by

$$E_1 = m \left[\frac{teq_0}{\epsilon\epsilon_0 K_G} \right] \left[\frac{K_G^2 L_D^2}{1 + K_G^2 L_D^2} \right], \quad (3.21)$$

in which q_0 is the photogeneration rate and t is the time relative to the initiation of grating recording. The initial growth when drift transport dominates is expressed by

$$E_1 = m \left[\frac{teq_0}{\epsilon\epsilon_0 K_G} \right] \left[K_G L_E (1 + K_G^2 L_E^2)^{-1/2} e^{i\phi} \right], \quad (3.22)$$

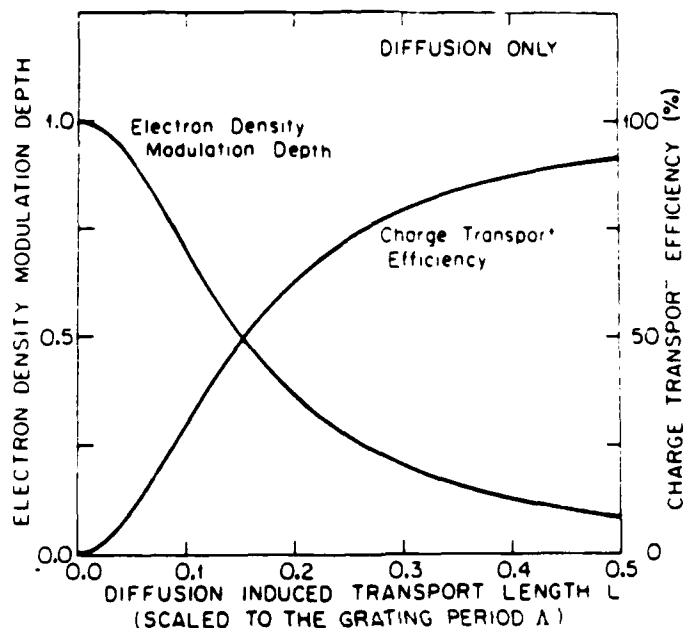
in which the phase shift ϕ is defined by

$$\tan \phi = K_G L_E. \quad (3.23)$$

The first bracketed term $[teq_0/\epsilon\epsilon_0 K_G]$ in Eqs. (3.21) and (3.22) corresponds to the grating recording efficiency predicted by the baseline sinusoid model (with $\rho_0 = teq_0$), as discussed in Section 3.2. This represents the upper bound on achievable recording sensitivity. The second bracketed terms in Eqs. (3.21) and (3.22) correspond to an additional charge transport inefficiency inherent in more realistic models of photorefractive recording, the subject of this section. This transport inefficiency factor is plotted as a function of increasing transport length in Fig. 3.14 for diffusion-only transport and in Fig. 3.15 for drift-only transport. Recall that these curves apply only during the initial recording interval, before significant space charge has accrued. Later recording will be characterized by a lower transport efficiency because of the presence of the space charge field. Note in Figs. 3.14 and 3.15 that the charge transport efficiency asymptotically approaches its maximum value in the limit of very long transport lengths, as intuitively expected, although even in this limit the maximum recording sensitivity is that of the baseline sinusoid, not that of the transport-efficient sinusoid.

The reason that the recording sensitivity only reaches the baseline sinusoid level in this limit is best understood by considering the spatial modulation profile of the mobile charge density $n(x)$. Analytical expressions for the modulation depth of the mobile charge density can be readily derived from the same analysis that led to Eqs. (3.21) and (3.22) (Young et al., 1974; Moharam et al., 1979), for times sufficiently long compared with the mobile carrier lifetime so that the mobile charge density represents a quasi-steady-state distribution, and also sufficiently short so as to remain in the initial recording regime. In the diffusion-only case, the mobile charge density is

Fig. 3.14.



Electron density modulation depth and charge transport efficiency as a function of the diffusion-induced transport length, in the diffusion-only regime.

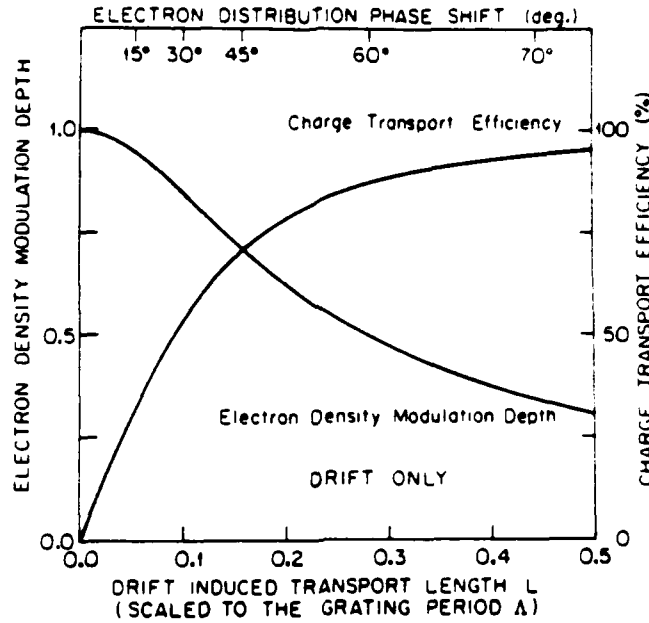
$$n(x) = \tau g_0 \left[1 + m_e(K_G L_D) \cos\left(\frac{2\pi x}{\Lambda_G}\right) \right], \quad (3.24)$$

in which τ is the mobile carrier lifetime and in which the modulation depth $m_e(K_G L_D)$ is given by

$$m_e(K_G L_D) = \frac{m}{1 + K_G^2 L_D^2}. \quad (3.25)$$

Note that this charge distribution has the least efficient possible phase for building up a space charge field. Recall from the discussion in Section 3.2.1 that the optimum phase of the mobile charge profile is a 180° phase shift with respect to the incident illumination profile, allowing the space charge field contribution of the mobile charge after subsequent trapping to add to the contribution of the excess positively charged trap profile produced by the photogeneration process. Here, however, the mobile charge profile is aligned coincident with the illumination and with the excess positively charged

Fig. 3.15.



Electron density modulation depth and charge transport efficiency as a function of the drift-induced transport length, in the drift-only regime. The resultant phase shift of the mobile electron distribution is shown at the top of the figure.

trap profiles. As a consequence, when a mobile charge is captured, it removes one of the incremental photogenerated traps, thereby reducing the overall space charge profile. A net space charge field can then accrue in the diffusion-dominated transport case only by partially randomizing the mobile charge distribution, corresponding to a reduction in the modulation depth $m_e(K_G L_D)$. Note in Fig. 3.14 that the rise in charge transport efficiency with increasing transport length $K_G L_D$ is coincident with a reduction of the mobile charge modulation depth, as expected. Optimum charge transport efficiency occurs when the mobile charge profile has been completely randomized, which, according to Eq. (3.25), occurs in the limit $L_D \gg \Lambda_G$.

Similarly, for drift-dominated transport, the mobile charge profile has a modulation depth of

$$m_e(K_G L_E) = \frac{m}{(1 + K_G^2 L_E^2)^{1/2}} \quad (3.26)$$

and a phase shift of ϕ , as given by Eq. (3.23). The charge transport efficiency, mobile charge modulation depth, and phase shift of the mobile charge profile with respect to the incident illumination are plotted in Fig. 3.15 as a function of the transport length $K_G L_E$ for drift-dominated transport, which is in turn proportional to the applied bias field E_0 . For short transport lengths, the phase shift is close to 0° , which as pointed out above is the least efficient phase for building a space charge field, and the modulation depth m_c exhibits its maximum value of m , the modulation depth of the illumination profile. For longer transport lengths, the mobile charge profile phase shifts away from 0° , but is always less than 90° , and hence always degrades the net space charge profile. The modulation depth m_c similarly decreases with increasing charge transport length. The most efficient charge transport occurs for very large transport lengths, $K_G L_E$, for which the mobile charge profile has become completely randomized.

The reason that the phase of the mobile charge profile never exceeds a 90° phase shift for drift-dominated transport (and that the phase is always 0° for diffusion-dominated transport) can perhaps best be appreciated from Fig. 3.13, which shows the mobile charge profile that is generated in response to a narrow slit (i.e., very tightly focused) illumination profile. The profiles shown in Fig. 3.13 can be derived by Fourier decomposing the Dirac delta function of the illumination profile into an equivalent set of spatial harmonics, applying Eqs. (3.25) and (3.26) to find the mobile charge profile in response to each spatial frequency and then performing an inverse Fourier transform. But the profiles shown in Fig. 3.13 are also intuitively reasonable. Diffusion tends to broaden the mobile charge density symmetrically about the location of an illumination region, whereas drift tends to pull the mobile charges to one side. In addition, note that the mobile charge distribution resulting from any arbitrary illumination profile can be derived by convolving the illumination profile with the appropriate distribution shown in Fig. 3.13 (which can be considered to be a blur function). Because of the monotonically decreasing shape of these blur functions, no phase shift in excess of 90° is feasible.

Thus we find that the baseline sinusoid model represents an upper bound on the grating recording efficiency predicted by the single mobile charge species/single donor/single trap photorefractive recording model.

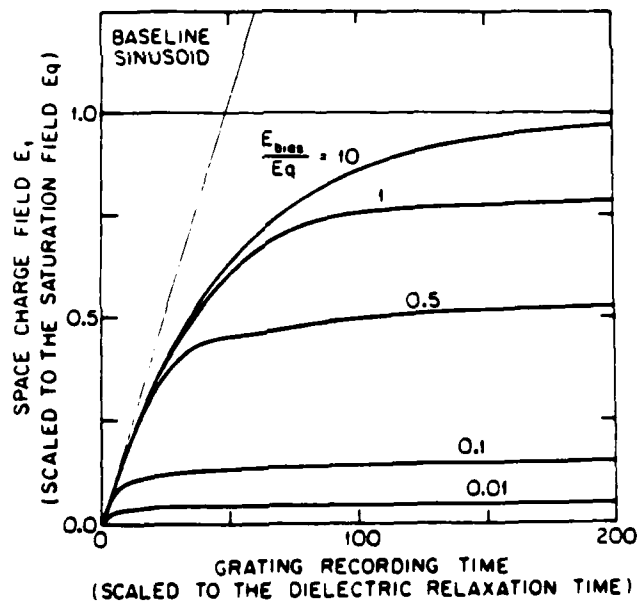
3.3.2. Temporal Approach to Steady State

In addition to the charge transport efficiency factor just discussed, two additional factors must be considered when evaluating any realistic recording

situation. One factor pertains to the saturation in temporal growth of the space charge field, resulting in a reduced growth rate as the field approaches its steady state limit. The second factor pertains to the modulation depth m of the illumination profile, which is often chosen to be small to enable enhanced recording techniques, as discussed in the next section.

The reduction in sensitivity of the recording process as the space charge field approaches its steady state limit can be assessed from analytic solutions that have been derived by Kukhtarev (1976). Typical solutions are shown in Fig. 3.16 for a variety of bias fields, E_0 , scaled to the maximum possible field, E_q , due to limited ionized trap density (discussed in Section 3.2.2). These curves were generated based upon typical charge mobility-lifetime product ($\mu\tau$) parameters for bismuth silicon oxide ($\text{Bi}_{12}\text{SiO}_{20}$; BSO), as given in Table 3.3. (A bias field to saturation field ratio of 10:1 is unphysical for several photorefractive materials such as bismuth silicon oxide, requiring exceptionally high bias fields and/or spatial frequencies, but is nonetheless included for generality.) Note that in all cases shown the time needed to

Fig. 3.16.



Temporal approach of the first harmonic of the space charge field E_1 to its steady state limit with applied field as a parameter for $\text{Bi}_{12}\text{SiO}_{20}$, assuming a stationary illumination profile with small modulation depth m .

TABLE 3.3
Material Parameters Assumed in the Calculations

	Bi ₁₂ SiO ₂₀	BaTiO ₃	Units
Mobility-lifetime product $\mu\tau$	15	0.5	$\mu\text{m}^2/\text{V}$
Trap density N_t	1×10^{17}	2×10^{17}	cm^{-3}
Dielectric constant ϵ	56	168 ($\epsilon_{\text{BaTiO}_3}$)	
Index of refraction n	2.5	2.4	
Electrooptic coefficient r_{33}	5 (r_{33})	80 (r_{33})	pm/V
n/r_{33} , ϵ	1.4	6.8	pm/V

(After Valley and Klein, 1983, and references cited therein)

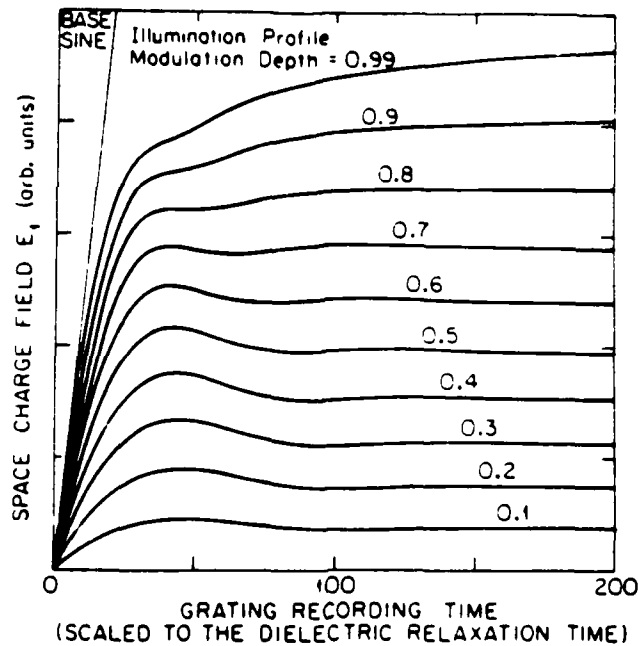
reach saturation is approximately a factor of two longer than that predicted by the baseline sinusoid idealized transport model.

Due to the fact that the analytical solutions for temporal evolution derived by Kukhtarev (1976) are valid only for small modulation depths m of the illumination profile, such solutions describe low recording efficiency situations in which the majority of the photons in the illumination beam contribute a uniform photocurrent and only a small fraction of the photons convey the spatial structure in the image profile. This point is emphasized in Fig. 3.17, in which numerical solutions of the photorefractive equations are presented, showing the temporal evolution of the first spatial harmonic component of the space charge field for various modulation depths m . These solutions have been produced by the authors using numerical techniques discussed by Moharam et al. (1979) but applied to the full set of photorefractive equations proposed by Kukhtarev (1976).

In Fig. 3.17, a crystal of bismuth silicon oxide illuminated by a 300 cycle/mm sinusoid has been assumed. Note in this figure the slight oscillations that can be observed in the temporal evolution. The strength of these oscillations is directly dependent on the charge mobility-lifetime product, $\mu\tau$; the curves in Fig. 3.17 correspond to the material and configuration parameters as listed in Table 3.3. It should be noted that a wide variation in mobility-lifetime products has been reported, even for crystals with nominally the same composition (Lesaux et al., 1986).

Note also in Fig. 3.17 that the highest space charge fields are associated with the highest illumination profile modulation depths, a regime for which the time-dependent analytical solutions derived by Kukhtarev are no longer applicable. Furthermore, to first order the resultant space charge field E_1 , both in the initial recording regime as well as in saturation, is directly proportional to the modulation depth parameter m .

Fig. 3.17.



Increase in saturation space charge field with increasing modulation depth for $\text{Bl}_{12}\text{SiO}_{20}$, showing transition from linear to nonlinear recording regimes.

3.3.3. Enhanced Recording Techniques

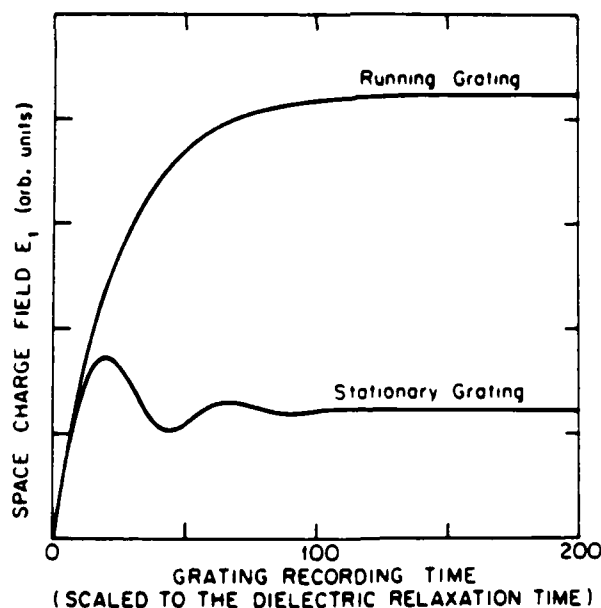
The recording configuration just considered assumes a stationary illumination profile and a constant applied bias field (for brevity, hereinafter called the *stationary illumination technique*). Space charge fields with much higher steady state limits can be obtained by either of two alternative nonstationary recording configurations: One technique is to translate the illumination profile with respect to the photorefractive crystal (hereinafter referred to as the *running grating technique*) (Huignard and Marrakchi, 1981; Stepanov et al., 1982; Valley, 1984; Refregier et al., 1985), and the second technique is to periodically reverse the direction of the applied bias field (the *alternating field technique*) (Stepanov and Petrov, 1985). These techniques do not improve the rate at which space charge builds up with time under constant illumination intensity, compared with the stationary illumination/constant field recording configuration.

The running grating technique, in particular, was chosen for study herein.

both to illustrate enhanced photorefractive recording concepts and because it is commonly employed to provide significant amplification of weak images. This technique can be studied by a combination of analytical solutions that are valid in the linearized regime of very small modulation depths and by numerical solutions for larger image modulations.

The relative advantage of the running grating technique is summarized in Fig. 3.18, which has been derived from the analytical solutions of Valley (1984) and Refregier et al. (1985) in the limit of small illumination profile modulation depths. A grating spatial frequency of 110 cycles/mm and parameters typical of bismuth silicon oxide have been assumed in the solutions shown in Fig. 3.18. The spatial frequency is chosen to be 110 cycles/mm, rather than 300 cycles/mm, because the enhancement of the steady state space charge field is maximized for this grating frequency, assuming material parameters for $\text{Bi}_{12}\text{SiO}_{20}$ as given in Table 3.3. Note that the steady state limit of the space charge field is in fact increased by the running grating technique relative to that obtained with a stationary grating, but that the

Fig. 3.18.

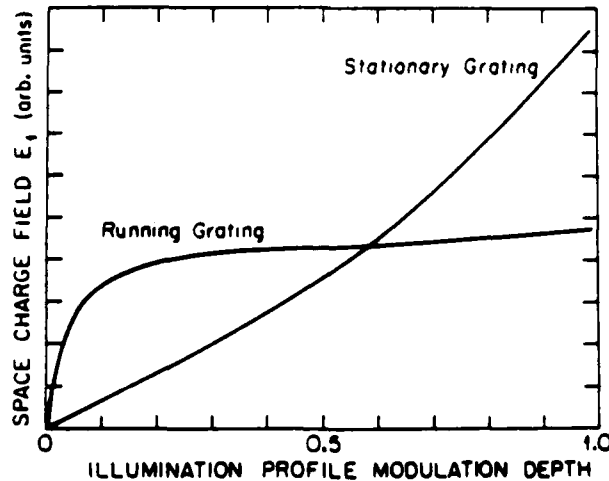


Temporal evolution of the space charge field, comparing the cases of stationary and running gratings for $\text{Bi}_{12}\text{SiO}_{20}$ in the low modulation limit.

initial recording sensitivities are asymptotically equal. The initial growth of the space charge field continues to be bounded by a combination of quantum limitations and charge transport inefficiencies, as described above. Furthermore, the additional factor of two reduction in sensitivity observed on approach to saturation obtains for both solutions.

The magnitude of the steady state space charge field is necessarily quite small in the linearized regime for which the analytical solutions apply, as the field is proportional to the modulation depth m of the illumination profile, which must be kept small to ensure accurate analytical solutions. Larger space charge fields require larger illumination modulation depths, which in turn result in eventual nonlinear saturation of the space charge field itself. Numerical solutions have been generated by the authors to explore the onset of this field saturation, with typical results as shown in Fig. 3.19, again assuming the nominal $\text{Bi}_{12}\text{SiO}_{20}$ material parameters listed in Table 3.3. Note that the saturation of the steady state field occurs at quite modest values of the modulation depth, which is consistent with the experimental observations reported by Refregier et al. (1985). Thus the running grating technique proves to be most effective for amplifying weak images, but not for recording large space charge fields. For recording the largest fields, stationary illumination is preferred.

Fig. 3.19.



Steady state space charge field as a function of modulation depth for $\text{Bi}_{12}\text{SiO}_{20}$, comparing the cases of stationary and running gratings.

4. Representative Grating Recording Efficiency Calculations

In order to illustrate the above concepts, we proceed in this section to consider the several factors that contribute to the overall grating recording efficiency for two different types of materials and for two different types of applications. The materials considered are bismuth silicon oxide ($\text{Bi}_{12}\text{SiO}_{20}$, or BSO) and barium titanate (BaTiO_3); the principal material parameters assumed in the estimates are listed in Table 3.3, based upon a set of values utilized in previous related analyses by Valley and Klein (1983). The two applications considered are those of reconfigurable holographic interconnections and the amplification of weak images.

Let us first consider a reconfigurable holographic interconnection implemented in bismuth silicon oxide. For simplicity, we assume that only one grating with a spatial frequency of 300 cycles/mm is recorded in the crystal. To achieve maximum diffraction efficiency, the modulation depth of the recording beams should be as large as possible; ideally, $m = 1$. Also, a bias electric field is typically applied to crystals of $\text{Bi}_{12}\text{SiO}_{20}$ to enhance the photosensitivity; a field of 6 kV/cm is typical, implying a drift transport length L_E from Eq. (3.18) of 9 μm , assuming the mobility-lifetime product given in Table 3.3. For a 300 cycle/mm grating frequency, this implies an (L_E/Λ_G) ratio of about 2.7, and from Fig. 3.15 we see that this corresponds to essentially 100% charge transport efficiency in the initial stages of recording.

The recording efficiency for a $\text{Bi}_{12}\text{SiO}_{20}$ interconnect, compared with the ideal (i.e., bipolar comb) quantum efficiency limit, is listed in Table 3.4. Three factors are considered in this and subsequent tables. The first is the

TABLE 3.4

Representative Grating Recording Efficiency Calculation: Reconfigurable Interconnection
Drift Recording in Bismuth Silicon Oxide

Modulation depth = 1.0	
300 cycles/mm grating frequency	
Space charge efficiency factors:	
Baseline sinusoid/bipolar comb	0.25
Charge transport (to saturation)	0.5
Modulation depth factor	1.0
Grating recording efficiency	0.125
Diffraction efficiency derating factor	0.016

25% efficiency factor that applies between the baseline sinusoid and bipolar comb cases, a factor that is common to all recording configurations considered in this section. The second factor is the charge transport efficiency, comparing actual recording performance to that of the baseline sinusoid case. The $\text{Bi}_{12}\text{SiO}_{20}$ interconnect is assigned a 50% charge transport efficiency to account for the factor of two increase in recording energy (photon flux) estimated to reach saturation, as shown in Fig. 3.16 and as discussed in Section 3.3.2. The final factor is the modulation depth, which we have assumed to be unity for a reconfigurable interconnect. Thus the total grating recording efficiency, in terms of the magnitude of the space charge field generated per unit photon flux, compared with ideal quantum efficient recording, is only about 12.5% for a $\text{Bi}_{12}\text{SiO}_{20}$ interconnection. This gives a diffraction efficiency derating factor of only 1.6% (assuming a diffraction efficiency that is proportional to the square of the space charge field).

As a second example, consider image amplification in $\text{Bi}_{12}\text{SiO}_{20}$ using a running grating enhanced recording technique. We will again assume a 110 cycle/mm grating spatial frequency (see Section 3.3.3) and a bias field of 6 kV/cm applied to the $\text{Bi}_{12}\text{SiO}_{20}$ crystal. However, the modulation depth must be reduced from 100% to approximately 10% to achieve the peak space charge enhancement in saturation provided by the running grating recording technique, as shown in Fig. 3.19 and as discussed in Section 3.3.3. For our calculations, we chose a modulation depth equal to 10%, implying a corresponding reduction in the quantum efficiency of the recording process; i.e., most of the photons must supply the pump beam, with comparatively few photons in the signal beam containing signal information. Thus the total quantum efficiency for image amplification, shown in Table 3.5, is an order of magnitude lower than that given in Table 3.4 for the reconfigurable in-

TABLE 3.5

Representative Grating Recording Efficiency Calculation: Two-Wave Image Amplification
Running Gratings in Bismuth Silicon Oxide

Modulation depth = 0.1	
110 cycles/mm grating frequency	
Space charge efficiency factors:	
Baseline sinusoid/bipolar comb	0.25
Charge transport (to saturation)	0.5
Modulation depth factor	0.1
Grating recording efficiency	0.0125
Diffraction efficiency derating factor	0.00016

terconnection. Correspondingly, the diffraction efficiency derating factor is two orders of magnitude lower for this case than for the case of the reconfigurable interconnection and nearly four orders of magnitude less efficient than the absolute quantum limitation.

As a final example, consider a reconfigurable interconnection in barium titanate, with grating recording efficiency as shown in Table 3.6 for the case of recording by diffusion transport only. Because of the large electrooptic coefficient in barium titanate, only very modest space charge fields, typically a fraction of the diffusion field for a 300 cycle/mm grating, are needed to achieve peak diffraction efficiencies in crystals of reasonable size. Hence diffusion transport often proves to be sufficient in this material in order to generate experimentally useful diffraction efficiencies. Unfortunately, two factors work against the high electrooptic coefficient in barium titanate. One is the concomitantly high dielectric constant, which from Maxwell's first equation implies that a considerable amount of space charge must be moved to achieve a modest space charge field. As mentioned in the introduction, the combined material parameter ($n_0^3 r_{eff}/\epsilon$), which is a measure of the amount of optical index modulation per unit space charge, proves to be surprisingly constant from material to material (Glass et al., 1984; Glass, 1984). Based upon this measure alone, barium titanate proves to be modestly superior to bismuth silicon oxide, as shown in Table 3.3.

The second, and far more serious, factor degrading grating recording efficiency for diffusion recording in barium titanate is its low mobility-lifetime product $\mu\tau$, which is almost two orders of magnitude lower than that for bismuth silicon oxide, implying a significantly degraded charge transport efficiency. The diffusion transport length L_D predicted by Eq. (3.19) is of order 35 nm, based upon the mobility-lifetime product $\mu\tau$ given in Table

TABLE 3.6
Representative Grating Recording Efficiency Calculation: Reconfigurable Interconnect
Diffusion Recording in Barium Titanate

Modulation depth = 1.0	
300 cycles/mm grating frequency	
Space charge efficiency factors:	
Baseline sinusoid/bipolar comb	0.25
Charge transport (to saturation)	0.002
Modulation depth factor	1.0
Grating recording efficiency	0.0005
Diffraction efficiency derating factor	2.5×10^{-4}

3.3. Assuming a 300 cycle/mm grating, this implies by Eq. (3.21) a charge transport efficiency factor of order 0.004. We include an additional factor of 0.5 in Table 3.6 to account for the additional photon flux required to reach saturation, as indicated in Fig. 3.16 and by the discussion in Section 3.3.2. Thus we find that the grating recording efficiency for diffusion recording in barium titanate is some 250 times less than that for drift recording in bismuth silicon oxide, in terms of space charge generated per unit incident photon.

Thus we have considered representative examples of several different materials, transport mechanisms and efficiencies, and recording applications. We find that, even in the most efficient grating recording configurations, a significant inefficiency still exists between the absolute quantum limits and actual performance.

5. --- Conclusions

In this chapter, we have considered a number of the fundamental physical limitations that constrain the potential performance of photorefractive materials. In particular, we have described several idealized photogeneration and charge transport models in terms of the grating recording efficiency, and we have identified one such model (the bipolar comb) as the absolute quantum limit against which other such models may be compared. The bipolar comb model generates the maximum possible fundamental harmonic of the space charge field at unity dielectric constant for a given number of photoexcited mobile charge carriers. A second idealized model, the baseline sinusoid, provides an upper bound for the grating recording efficiency of a more realistic photorefractive grating recording model involving a single mobile charge carrier and a single donor/single trap photorefractive center. This upper bound was shown to be a factor of four less efficient in generating a given space charge field than the quantum limitations imply for an optimum photogeneration distribution and perfectly efficient charge transport. An additional factor of two accrues from the nonlinearity of the grating recording process observed as the space charge field nears saturation. The combination of absorption and reflection losses in typical photorefractive recording configurations (without antireflection coatings) contributes approximately one order of magnitude to the inefficiency of grating recording and readout relative to the incident photon flux. Hence, the usual photorefractive recording

configuration exhibits an overall sensitivity that is approximately two orders of magnitude less than that achievable in the quantum limit. This results in roughly four orders of magnitude reduction in the corresponding diffraction efficiency per incident photon. These considerations explain to a certain degree why photorefractive recording has proven to be relatively insensitive as compared with distinct but related mechanisms of spatial light modulation.

Perhaps far more important, however, are the implications of the above analysis for the conceptual design and technological implementation of optimized grating recording media that operate far closer to the quantum limits. For example, the bipolar comb illumination profile is not necessarily as unphysical as it may at first seem, and it can be closely approximated by the utilization of stratified volume holographic optical elements (Johnson and Tanguay, 1988) for beam formation. This approach is perhaps most appropriate for the generation of a grating of given spatial frequency, as in the photorefractive incoherent-to-coherent optical converter (Marrakchi et al., 1985). As a second example, the factor of two inherent in the approach to saturation can be avoided by seeking photorefractive materials of near unity charge transport efficiency and high electrooptic figures of merit, such that experimentally suitable diffraction efficiencies can be obtained well within the linear recording regime. As a final example, the inefficiency implied by surface reflection losses can be dramatically reduced, even over the broad spectrum of angles and wavelengths characteristic of photorefractive recording, by the design and incorporation of appropriate antireflection coatings (Karim et al., 1988).

Acknowledgments

This research was supported in part by the Defense Advanced Research Projects Agency (through the Office of Naval Research and the Air Force Office of Scientific Research), the Air Force Office of Scientific Research, the Army Research Office, and the Joint Services Electronics Program.

References

- Amodei, J.J. (1971). "Analysis of transport processes during holographic recording in insulators," *RCA Review* **32**, 185-198.
- Amodei, J.J., and Staebler, D.L. (1972). "Holographic recording in lithium niobate," *RCA Review* **33**, 71-93.

- Attard, A.E., and Brown, T.X. (1986). "Experimental observations of trapping levels in BSO," *Appl. Opt.* **25**(18), 3253-3259.
- Chemla, D.S., Miller, D.A.B., and Smith, P.W. (1985). "Nonlinear optical properties of GaAs-GaAlAs multiple quantum well material: Phenomena and applications," *Opt. Eng.* **24**(4), 556-564.
- Feinberg, J., Heiman, D., Tanguay, A.R. Jr., and Hellwarth, R.W. (1980). "Photorefractive effects and light-induced charge migration in barium titanate," *J. Appl. Phys.* **51**(3), 1297-1305.
- Glass, A.M. (1978). "The photorefractive effect," *Opt. Eng.* **17**(5), 470-479.
- Glass, A.M. (1984). "Materials for optical information processing," *Science* **226**, 657-662.
- Glass, A.M., Johnson, A.M., Olson, D.H., Simpson, W., and Ballman, A.A. (1984). "Four-wave mixing in semi-insulating InP and GaAs using the photorefractive effect," *Appl. Phys. Lett.* **44**(10), 948-950.
- Glass, A.M., Klein, M.B., and Valley, G.C. (1987). "Fundamental limit of the speed of photorefractive effect and its impact on device applications and material research: Comment," *Appl. Opt.* **26**(16), 3189-3190.
- Gunter, P. (1982). "Holography, coherent light amplification, and optical phase conjugation with photorefractive materials," *Phys. Rep.* **93**, 199-299.
- Hou, S.L., Lauer, R.B., and Aldrich, R.E. (1973). "Transport processes of photoinduced carriers in $\text{Bi}_{12}\text{SiO}_{20}$," *J. Appl. Phys.* **44**(6), 2652-2658.
- Huignard, J.P., and Marrakchi, A. (1981). "Coherent signal beam amplification in two-wave mixing experiments with photorefractive $\text{Bi}_{12}\text{SiO}_{20}$ crystals," *Opt. Commun.* **38**(4), 249-254.
- Huignard, J.P., and Micheron, F. (1976). "High-sensitivity read-write volume holographic storage in $\text{Bi}_{12}\text{SiO}_{20}$ and $\text{Bi}_{12}\text{GeO}_{20}$ crystals," *Appl. Phys. Lett.* **29**(9), 591-593.
- Jaura, R., Hall, T.J., and Foote, P.D. (1986). "Simplified band transport model of the photorefractive effect," *Opt. Eng.* **25**(9), 1068-1074.
- Johnson, R.V., and Tanguay, A.R. Jr. (1988). "Stratified volume holographic optical elements," *Opt. Lett.* **13**(3), 189-191.
- Kaminow, I.P. (1974). "An Introduction to Electrooptic Devices." Academic Press, New York.
- Kamshilin, A.A., and Petrov, M.P. (1980). "Holographic image conversion in a $\text{Bi}_{12}\text{SiO}_{20}$ crystal," *Sov. Tech. Phys. Lett.* **6**(3), 144-145.
- Karim, Z., Garrett, M.H., and Tanguay, A.R. Jr. (1988). "A bandpass AR coating design for bismuth silicon oxide," *Tech. Digest 1988 Annual Meeting of the Optical Society of America*, Santa Clara, California.

- Kogelnik, H. (1969). "Coupled wave theory for thick hologram gratings." *Bell Syst. Tech. J.* **48**(9), 2909-2947.
- Kukhtarev, N.V. (1976). "Kinetics of hologram recording and erasure in electrooptic crystals." *Sov. Tech. Phys. Lett.* **2**(12), 438-440.
- Kukhtarev, N.V., Markov, V.B., Odulov, S.G., Soskin, M.S., and Vinetskii, V.L. (1979). "Holographic storage in electrooptic crystals." *Ferroelectrics* **22**, 949-964.
- Lesaux, G., Launay, J.C., and Brun, A. (1986). "Transient photocurrent induced by nanosecond light pulses in BSO and BGO." *Opt. Commun.* **57**(3), 166-170.
- von der Linde, D., and Glass, A.M. (1975). "Photorefractive effects for reversible holographic storage of information." *Appl. Phys.* **8**, 85-100.
- Marrakchi, A., Johnson, R.V., and Tanguay, A.R. Jr. (1987). "Polarization properties of enhanced self-diffraction in sillenite crystals." *IEEE J. Quantum Electron.* **QE-23**(12), 2142-2151.
- Marrakchi, A., Tanguay, A.R. Jr., Yu, J., and Psaltis, D. (1985). "Physical characterization of the photorefractive incoherent-to-coherent optical converter." *Opt. Eng.* **24**(1), 124-131.
- Micheron, F. (1978). "Sensitivity of the photorefractive process." *Ferroelectrics* **18**, 153-159.
- Moharam, M.G., Gaylord, T.K., Magnusson, R., and Young, L. (1979). "Holographic grating formation in photorefractive crystals with arbitrary electron transport lengths." *J. Appl. Phys.* **50**(9), 5642-5651.
- Orlowski, R., and Kratzig, E. (1978). "Holographic method for the determination of photo-induced electron and hole transport in electro-optic crystals." *Solid State Comm.* **27**(12), 1351-1354.
- Peltier, M., and Micheron, F. (1977). "Volume hologram recording and charge transfer process in $\text{Bi}_{12}\text{SiO}_{20}$ and $\text{Bi}_{12}\text{GeO}_{20}$." *J. Appl. Phys.* **48**(9), 3683-3690.
- Refregier, Ph., Solymar, L., Rajbenbach, H., and Huignard, J.P. (1985). "Two-beam coupling in photorefractive $\text{Bi}_{12}\text{SiO}_{20}$ crystals with moving grating: Theory and experiments." *J. Appl. Phys.* **58**(1), 45-57.
- Shi, Y., Psaltis, D., Marrakchi, A., and Tanguay, A.R. Jr. (1983). "Photorefractive incoherent-to-coherent optical converter." *Appl. Opt.* **22**(23), 3665-3667.
- Stepanov, S.I. and Petrov, M.P. (1985). "Efficient unstationary holographic recording in photorefractive crystals under an external alternating electric field." *Opt. Commun.* **53**(5), 292-295.
- Stepanov, S.I., Kulikov, V.V., and Petrov, M.P. (1982). "Running holograms in photorefractive $\text{Bi}_{12}\text{SiO}_{20}$ crystals." *Opt. Commun.* **44**(1), 19-23.

- Strohkendl, F.P., Jonathan, J.M.C., and Hellwarth, R.W. (1986). "Hole-electron competition in photorefractive gratings." *Opt. Lett.* **11**(5), 312-314.
- Tanguay, A.R. Jr. (1985). "Materials requirements for optical processing and computing devices." *Opt. Eng.* **24**(1), 2-18.
- Valley, G.C. (1984). "Two-wave mixing with an applied field and a moving grating." *J. Opt. Soc. Am. B* **1**(6), 868-873.
- Valley, G.C. (1986). "Simultaneous electron/hole transport in photorefractive materials." *J. Appl. Phys.* **59**(10), 3363-3366.
- Valley, G.C., and Klein, M.B. (1983). "Optimal properties of photorefractive materials for optical data processing." *Opt. Eng.* **22**(6), 704-711.
- Yeh, P. (1987). "Fundamental limit of the speed of photorefractive effect and its impact on device applications and material research." *Appl. Opt.* **26**(4), 602-604.
- Yeh, P. (1987). "Fundamental limit of the speed of photorefractive effect and its impact on device applications and material research: Author's reply to comment." *Appl. Opt.* **26**(16), 3190-3191.
- Young, L., Wong, W.K.Y., Thewalt, M.L.W., and Cornish, W.D. (1974). "Theory of formation of phase holograms in lithium niobate." *Appl. Phys. Lett.* **24**(6), 264-265.

CHAPTER 15

PHOTONIC IMPLEMENTATIONS OF NEURAL NETWORKS

B. Keith Jenkins and Armand R. Tanguay, Jr.

TOWARDS THE DEVELOPMENT OF A NEURAL NETWORK IMPLEMENTATION TECHNOLOGY

As described in other chapters of this book, neural networks provide a different approach to solving problems as compared with more conventional algorithmic techniques, and can be applied to a wide range of applications. In some of these application domains the simulation performance of neural networks has been comparable to that of more conventional algorithmic approaches. In many application domains, however, a realistic (and therefore large scale) problem may overwhelm the conventional approach, in that it may be too computation intensive to be implemented on a sequential digital computer, and may not parallelize sufficiently well (if at all) for efficient computation on a parallel digital machine. On the other hand, because a neural network algorithm is inherently parallel, it immediately suggests a parallel architecture, which may in turn be implemented using either analog or digital hardware. And for the case of large-scale problems, analog hardware will typically provide a much more efficient neural implementation than digital hardware.

In the previous chapter, fully electronic (primarily VLSI-based) neural networks were described in which the primary functionality of both the neuron units and the weighted

(synaptic) interconnection matrix is incorporated on a planar microelectronic chip. An important advantage of the integrated circuit approach to neural network implementation is the capability for near-term technology insertion, with leverage provided by a well-established technology base characterized by a fully developed computer aided design and computer aided manufacturing (CAD/CAM) device and circuit repertoire. An equally important limitation is the difficulty in scaling up neural chips to incorporate large numbers of neuron units in fully (or near fully) interconnected architectures. This limitation derives from the limited pin-out, off-chip communication bandwidth, and on-chip interconnection density available in both current generation and projected chip designs.

In this chapter, we consider the utilization of optical (free-space) interconnection techniques in conjunction with photonic switching and modulating devices to expand the number of neuron units and complexity of interconnection, by using the off-chip (3^{rd}) dimension for synaptic communication. As we shall see, the merging of optical and photonic devices with appropriately matched electronic circuitry can provide novel features such as fully parallel weight updates and modular scalability, as well as both short and long term synaptic plasticity.

Many approaches to the incorporation of photonic and optical technology in the implementation of neural networks are currently being pursued in the research community. The intent of this chapter is not to present a review of these various approaches, the details of which can be obtained from several of the references given in the Suggested Further Reading section at the end of this chapter. Instead, our focus herein is directed toward a description of key photonic devices and techniques based on fundamental optical phenomena, as well as toward a unique and generalizable approach to their potential use in the implementation of large-scale, highly parallel neural network architectures. The unusual nature of some of these techniques has interesting implications for the design of naturally mapped architectures and associated learning/computing algorithms. We will address a number of these unique features in the context of a description of the basic optical phenomena that can be used to advantage, and of the array of photonic devices that comprise the system designer's palette. Because the incorporation of optical and photonic hardware casts the subject of neural network implementations in a somewhat unfamiliar

light, we first discuss a set of desirable and requisite characteristics for neural network implementation technologies.

An important feature of any implementation technology is that of generality: a "building block" approach. The growth and synthesis of material structures, and their incorporation into devices, must be well characterized, understood, and repeatable, for a *small number* of specific material combinations and device structures. These devices are then assembled into circuits or architectures for the implementation of specific computational models. This provides leverage in two ways: (1) the small number of useful and well understood components are used repeatedly in different structures for different applications, and (2) architecture and system level designers need not be experts in the properties of the material and device structures used to configure the components, saving many man-hours in the design of computational systems over a completely custom approach. Such a purely custom approach could preclude the widespread use of these architectures and systems, as has been characteristic, for example, of optical information processing and optical signal processing systems over the past two decades.

Not only is it important for the implementation *technology* to be of a building block nature, but it is also important for the *models* underlying the computational architectures to support such an approach, and in fact to be of a building block nature themselves. Ideally the models would comprise a set of common components and operations at the functional level, such as specific types of neuron units, weighted interconnections, weight updates, and comparisons with desired target values. Then the mapping from model and functional architecture to hardware architecture, layout, and implementation can proceed efficiently as well. This building block approach at both the model and hardware levels has certainly been characteristic of the development of digital electronics, and has been largely responsible for its success.

Assuming that appropriate neural network models and a corresponding technology base can be merged within a compatible building block approach, neural network systems potentially provide a unique capability for large-scale *analog, nonlinear* computation. As such, the neural network paradigm potentially alleviates two critical bottlenecks that have impeded the widespread implementation of large-scale analog, nonlinear computing

systems based on non-neural architectures: the lack of appropriate generic hardware components and of the sufficiently leveraged manpower required for their economical design and manufacture, as well as difficulty in establishing efficient techniques for mapping from the application and model domain onto compatible hardware. With regard to the former bottleneck, neural network architectures are generally forgiving with respect to device nonuniformities and imperfections, creating much needed latitude for the device designer. With regard to the latter bottleneck, the neural network paradigm inherently provides a mapping from the problem domain onto a highly parallel architecture, which immediately yields a starting point for its layout in analog hardware.

In the case of photonics for neural network applications, the hardware technology is being developed *simultaneously* with the neural computation model(s). This implies at least two things. First, it is crucial to retain *flexibility* in the functionality of each component, so that as the neural computation models evolve, the hardware can evolve along with it. The development of an entirely new technology base typically takes at least a decade; such a delay between model development and hardware realization is generally unacceptable. Thus, the generic technology base *must* provide sufficient flexibility. Second, not only should the neural computation model steer the technology development, but the reverse can, and indeed *must*, also occur. This assures a mutual compatibility in outcome.

The basic requisite functions for a neural network technology base appear to be: neuron unit response, weighted interconnections (fixed and variable), input/output, learning computation and weight update, and duplication capability (*i.e.*, the capability of making a copy of a network structure). In addition, other features are desirable, such as higher order connection capability. The neuron unit response, at the most common and basic level, is a sum-of-inputs followed by a monotonic nonlinearity. The nonlinearity should have the flexibility of providing different amounts of gain; for example, it is useful to have a high gain to implement a binary threshold, and a low gain to implement a nearly linear response. The neuron unit should be bipolar in that it permits both positive and negative inputs, so that inhibitory and excitatory connections can be realized. This we consider to be the minimal requisite functionality of a neuron unit. In addition, a very desirable feature is the capability for bipolar outputs. Note that biologically this is not necessarily

the case, but useful neural computation models will likely deviate substantially from biological reality and may require this capability. Other capabilities are also desirable, such as leaky integrator effects [Mead, 1989] and more complex behavior such as that required in shunting networks [Carpenter, 1987].

Each weighted interconnection must store a learned or initialized value, and perform a multiplication operation on the signals passing through. An analog multiplication is generally much more efficient than a digital one for reasons of speed and device area or volume. For this reason, it is worth some effort to provide analog storage for the weights. Note that the very large number of weights used in many networks implies that minimization of the incorporated hardware complexity of the requisite storage and multiplication operations is crucial for physical realizability. Input/output is often ignored at the higher levels, but can critically affect the physical architecture and can be a major factor in determining the overall throughput. The input/output function includes the input of signals, the output of results, and the input of weights if necessary.

It is important at the outset to distinguish among systems that have fixed weights; systems that have programmable weights (that are externally computed but loaded into the network), and systems that have full learning capability, in which the learning algorithm is implemented as part of the parallel system. The associated hardware complexity can be quite different in each of these cases. Finally, duplication capability is useful, for example, for replicating a pre-learned network when multiple copies of the network are to be produced and subsequently used with fixed connections. Probing the weight values in a hardware implementation may at first sound straightforward, but implementing very large numbers of weights in a small volume can in some cases preclude such capability.

Applications of hardware implementations of neural networks could include sensor signal processing and fusion, pattern recognition, associative memory, and robotic control. These applications imply a wide range of hardware requirements. For example, most vision processing is characterized by moderately large numbers of neuron units, with small to moderate connectivity and primarily local interconnections. Associative memory, on the other hand, typically requires a very high connectivity.

Semiconductor-based VLSI technology has proven to be very capable for the implemen-

tation of most, if not all, of the above functions. It also has the capability for integration of control circuitry and/or arbitrary digital or logical operations on the same chip as the neural processing circuitry. However, an important issue in neural implementations is that of *scalability*, since many neural network applications are likely to require very large numbers of neuron units and connections. As pointed out above, it is primarily the consideration of scalability that leads us to the conclusion that purely electronic VLSI will work well for certain applications, but can benefit greatly from the incorporation of photonics for other applications.

Two important considerations in VLSI implementations are area complexity and pinout requirements. A fully connected network of N neuron units requires area $O(N^2)$. One can think of this as having only a single linear dimension available for the neuron units themselves, in order to leave room for the connections. This of course limits the size of fully connected networks that can be accommodated on a single chip. For example, chips have been fabricated with 54 neuron units and 2916 ternary (3-level) synapses, and it is estimated that approximately 700 neurons, fully connected with $(10^5 - 10^6)$ similar synapses, could be implemented on a CMOS chip using $0.5 \mu m$ design rules. Learning capability with analog synapses may require substantially more area per synapse [Jackel, 1988]. On the other hand, networks with low connectivity and only local connections between neuron units permit a much larger number of neuron units to be implemented on a chip. For example, the silicon retina of Mead *et al.* comprises a 48×48 array of neuron-like units, each connected to its 6 nearest neighbors on a hexagonal grid [Mead, 1988]. A much larger number of neuron units, on the order of 10^5 , could be implemented with such a locally-connected array, since the neuron units and the connections each require area only $O(N)$. So we see that a critical limiting factor in the VLSI implementation of neural networks is the area required for on-chip interconnections, and that the area required for neuron units is relatively inconsequential.

The number of pinouts that can be provided on a chip is proportional to its linear dimension, not to its area. This degree of pinout capacity is well matched to fully connected networks in which the weights do not need to be input or output frequently, and to locally connected networks of low to moderate bandwidth. An example of the latter is a vision

network that operates at video frame rates; in this case the signal can be easily time multiplexed onto a relatively small number of lines for communication onto and off of the chip.

On the other hand, other application areas will require a higher input/output (I/O) bandwidth and/or a large number of neuron units with high connectivity. For example, if the weights are to be fed onto and off of a chip frequently, substantial multiplexing would be required for reasonably large networks (*e.g.*, up to 200 pinouts can generally be accommodated, but as described above, as many as 10^5 - 10^6 synaptic weights can be incorporated on the chip). Wafer scale integration with bump bonding techniques can help by providing large, multi-wafer structures, but the number of I/O lines is still likely to be modest due to practical and physical constraints. So we see that a second critical limiting factor in the VLSI (and wafer scale integration) implementation of neural networks is the number of I/O lines that can be practically incorporated. Neural applications utilizing large fully connected networks such as associative memory would benefit greatly from implementations of 10^5 - 10^6 fully connected neuron units, implying the need for 10^{10} - 10^{12} synaptic weights. In addition, the intermediate realm of large networks with partial but moderate-to-large connectivity will likely also prove beneficial to a wide range of applications.

In this chapter, we discuss a variety of issues that impact the development of photonic technology as applied to hardware implementations of neural networks with enhanced capabilities. Photonics has the potential for the implementation of networks with large numbers of neurons (10^5 - 10^6) and high connectivity (approximately 10^{10} analog-weighted interconnections) in one "module". The approach taken here is to use electronics to implement the internal function of each neuron unit, and to use optics to implement the connections, weights, and I/O. With this technique most of the area of a two-dimensional (2-D) "chip" can be used to implement the neuron units themselves, and optical free-space propagation and volume holograms can be used to implement the interconnections. Thus the interconnections actually occupy a three-dimensional (3-D) *volume*, which improves scalability dramatically.

The next section of this chapter describes the fundamental optical principles and key

photonic technology concepts that are needed for neural network implementations, and covers photonic analog arithmetic, switching, interconnections, sources and detectors. Architectural considerations are discussed in the subsequent section, including the use of volume interconnections, signal representation, and desired architectural features. The next section then presents a photonic implementation strategy that satisfies most of the desired criteria. In the two concluding sections we investigate the ultimate limitations of photonic implementations of neural networks, and consider the future of such implementations.

FUNDAMENTAL PRINCIPLES OF PHOTONIC TECHNOLOGY

In order to effectively appreciate the potential advantages as well as the limitations of extending the VLSI (electronic) repertoire to include photonic components and optically inspired functionality, we will first identify and then explain a few truly fundamental principles of the optical and photonic technologies on which such hybrid neural network implementations are based. In this section, therefore, we discuss the basic features of optical analog computation with both coherent and incoherent illumination sources, photonic switching devices and their various neuron-like functions, the characteristics of photonic interconnections that are essential to the implementation of synapse-like interneuron wiring, and the principal features of sources (photonic power supplies) and detectors (photonic-to-electronic power converters).

Optical Analog Computation

At the present time, most proposed photonic implementations of neural networks are based on analog operations, both in the representation of neuron outputs as well as in the incorporation of interconnection weights. This emphasis of *analog* computation over

perhaps more familiar *digital* computation derives principally from several distinct advantages that accrue to optical systems designed to handle the switching and interconnection of very large numbers of inputs and outputs at each circuit node (neuron unit). These advantages include a significant reduction in the number of switching components required to sum multiple inputs [Abu-Mostafa, 1989], an increase in the degree of fan-in and fan-out allowable from each circuit node, the elimination of analog-to-digital and digital-to-analog converters, a significant decrease in signal routing and interconnection complexity, the potential utilization of natural physical phenomena within certain photonic devices to accomplish difficult computational functions directly, and the possibility of higher computational throughput per unit dissipated energy in operations characterized by high computational complexity. Additionally, two notable disadvantages of analog systems, error accumulation and lack of precision in representation, may prove to be relatively unimportant in the neural network environment, due in part to the self-organizing and error correcting nature of many neural learning algorithms [von der Malsburg, 1987]. We will return to a number of these issues throughout the remainder of this chapter.

The computational operations necessary for the implementation of a wide range of neural and neural-like networks are surprisingly simple, consisting primarily of addition, subtraction, multiplication, and nonlinear thresholding. The operation of addition usually must be performed over a very large number of inputs at a given neuron unit, representing weighted excitatory signals from other interconnected neuron units; subtraction is utilized to differentiate excitatory inputs from inhibitory inputs to a given neuron unit, as is common, for example, in models of the visual process and of associative memories. Multiplication is necessary for the provision of linear interconnections with signal-independent weights, which in turn store learned (or pre-programmed) information, and hence form an important constituent of the neural paradigm. Finally, nonlinear thresholding operations are performed in order to provide the appropriate transfer function between the neuron activation potential (sum of all inputs, both positive and negative, to a given neuron unit), and the output each neuron unit generates in response. It is this nonlinearity in particular that gives multilayer neural networks their computational power, and allows recurrent networks to iteratively approach one of several stable states of the system. The detailed

nature of the nonlinearity itself can affect a number of critical system properties, including the number of iterations (or equivalently the time) required to achieve steady state, and the stability of the network in the presence of noise, inaccuracy, and nonuniformities among both the neuron units and the interconnections. In some envisioned neural network implementations, it is important to also be able to alter the nature of the nonlinear thresholding function dynamically during the computational phase of operation.

In addition to these elementary operations (and combinations thereof), it is necessary to accommodate operationally for both the learning function (which includes input-dependent interconnection weight updates) and for the computational function (which in general requires iterative feedback and the implementation of nonlinear thresholding, usually with fixed interconnection weights). *The only additional fundamental operation required by these features, that of input-dependent interconnection weight updates, can usually be reduced to at most a combination or sequence of the previously described fundamental operations of addition, subtraction, multiplication, and nonlinear thresholding.* It is important to note, however, that in this case the operations pertain directly to the implementation of interconnection weights, rather than to functions performed by the neuron units; hence, the physical processes involved may in fact be considerably different in nature, and thus subject to a quite different set of constraints. This differentiation between the two different "types" of basic operations (those pertaining to the neuron units and those pertaining to the interconnection pathways and weights) is discussed in more detail below, as well as in the section describing the fundamental physical and technological limitations of neuro-optical computing.

At the outset, we must further differentiate between those fundamental photonic operations that are intended for "optical" implementation, and those that are envisioned for "optoelectronic" implementation. In the first category, we place those types of physical processes in which one or more beams of light interact either *directly* (as in coherent interference) or through an intermediate physical medium (as in the summation of two incoherent beams of light on a single detector). In the second category, we place operations that involve one or more photon-to-electron or electron-to-photon conversion processes prior to the actual implementation of the desired function, which is then assumed to be

accomplished using intermediate (primarily analog, but perhaps digital) circuitry. An example of this latter category might be the subtraction of two optical signals by independent but simultaneous photodetection of each signal, followed by the use of an analog electronic differential amplifier to execute the functional subtraction. Most proposed neuro-optical processors rely to some degree on both types of physical implementation mechanisms (*c.f.* the sections "Architectural Considerations for Photonic Neural Network Implementations" and "An Implementation Strategy", below). In fact, the eventual degree of success achievable by neuro-optical computing techniques rests heavily on an appropriate balance of these mechanisms within a given system, optimized to yield the greatest computational advantage within the allowable physical and system constraints.

Optoelectronic implementations of neural functions for the most part rely on optical signals as inputs and outputs to and from the neuron units in order to allow for both high bandwidth and high interconnection multiplexing capacity, and on electronic signal combination and processing locally within each neuron unit. As such, optoelectronic functions are characterized primarily by the *optical* characteristics of interconnection, and by the *electronic* characteristics of computation. The former will be described below, while the latter has been discussed both by Bang Lee and Bing Sheu elsewhere in this volume [Lee, 1991] as well as by Carver Mead in a recent elegant monograph [Mead, 1989]. On the other hand, optical implementations of these functions rely on communication by, as well as the interaction of, two or more optical signals in order to accomplish the relevant computation, which is often (but not always) followed by a photon-to-electron conversion process in some form of single channel or array detector. It is to the fundamental principles of such optical computational interactions that we next turn our attention.

In order to adequately consider even so basic a process as optical addition, it is essential to differentiate between two basic types of optical interactions (as determined by the nature of the optical signals involved): *incoherent* and *coherent*. Incoherent interactions occur whenever the light wavefronts representing the input signals temporally dephase (do not oscillate in unison) over the relevant time of observation (detector temporal integration window), in that they are either both temporally incoherent at the outset, or are *each* temporally coherent but separated in optical frequency by more than the inverse of the

observation time. Interactions in which the input optical signals spatially dephase over the aperture of the relevant detector wherever the output is utilized (detector spatial integration window) are also incoherent for all practical purposes, and will obey incoherent summation rules as given below. Coherent interactions occur, on the other hand, whenever the light wavefronts representing the input signals simultaneously maintain a constant phase relationship over the detector spatial and temporal integration windows.

From these remarks, it can be seen that it is quite important to understand the distinction between coherent (or incoherent) *light* and coherent (or incoherent) *interactions* as defined by the eventual detector configuration and operational parameters. For example, it is perfectly acceptable to consider a situation in which two mutually coherent optical beams interact to produce an interference pattern with a spatial scale small compared with the relevant detector aperture. In such cases, the interaction will in fact follow *incoherent* summation rules, as the detector effectively *integrates* the space-variant interference pattern over the full detector aperture to produce exactly the same result as the interaction of two mutually incoherent (temporally) optical beams.

Given these preconditions, then, the actual rules for the basic operations are quite straightforward. Consider first the case of addition of two incoherent optical signal beams in a *collinear* geometry, in which the two distinct input beams with intensities I_1 and I_2 are assumed to emerge from a beam-combining optical system (as yet undetermined) such that the two output beams are collinear. If the beams are combined, for example, by a nondispersive (wavelength-insensitive) 50/50 (50% transmission, 50% reflection) beam-splitter as shown in Figure 15.1(a), two possible output beams I_{out} and I'_{out} are created, each with an intensity given by:

$$I_{out} = I'_{out} = \frac{1}{2}(I_1 + I_2). \quad (1)$$

The output intensity is thus linearly proportional to the sum of the input intensities. Note that this operation cannot be accomplished without an inherent loss, in the case shown above equal to 0.5 or about 3 *dB*. In fact, if we wish to combine N beams collinearly by this technique (using a linear chain of non-dispersive beamsplitters), $N - 1$ beamsplitters

are required with transmissivities given by $1/2, 2/3, 3/4, \dots, (N-1)/N$, representing a total loss of $(N-1)/N$ with an overall throughput of $1/N$:

$$I_{out} = \frac{1}{N}(I_1 + I_2 + I_3 + \dots + I_N). \quad (2)$$

Instead of using a linear chain of beamsplitters with different transmissivities, we could alternatively construct a binary tree structure by pairing the inputs that again requires $N-1$ beamsplitters, but in this case with *equal* transmissivities of $\frac{1}{2}$. This system of beamsplitters also exhibits an overall throughput of $\frac{1}{N}$ (for even values of N). Although beam combination of a large number of inputs by the multiple beamsplitter method is impractical, we will utilize this result a little later in order to understand the essential features of holographic beam combiners, which are subject to many of the same constraints.

FIGURE 15.1 Illustration of optical addition utilizing a 50/50 beamsplitter: (a) collinear *incoherent* beam geometry; (b) collinear *coherent* beam geometry, showing input and output *amplitudes*; (c) collinear *coherent* beam geometry, showing input and output *intensities*.

It should be noted in passing that the overall throughput loss implied by Equation (2) can be circumvented *if* the beams to be summed incoherently are sufficiently distinct in wavelength that a *dichroic mirror* can be used to combine them. A dichroic mirror reflects light within a given wavelength range, and transmits light outside of that range. Multiple dichroic mirrors can be used to collinearly sum multiple beams through appropriate choice of the input wavelengths in each arm, and of the characteristic wavelengths of each succeeding dichroic mirror.

Consider next the case of addition of two *coherent* optical signal beams in a collinear geometry. An example, again using a beamsplitter, is shown schematically in Figure 15.1(b). The presence of the beamsplitter generates a $\pi/2$ phase shift for each transmitted

beam, and a π phase shift for each reflected beam [Haus, 1984]. In this case, due to the coherent nature of the two input signal beams, the output intensity is no longer given simply by the sum of the input intensities. In fact, the output *amplitude* is proportional to the sum of the input signal *amplitudes* (provided that the two beams have identical polarizations):

$$a_{out} = \frac{1}{\sqrt{2}}(ia_1 - a_2e^{i\phi}) \quad (3)$$

in which ϕ is the relative phase between the two wavefronts (here assumed constant over the detector aperture). It should be noted parenthetically that optical beams with orthogonal polarizations do not interfere, and hence follow the incoherent addition rule. We assume throughout this chapter that all beams are polarized identically.

From this simple equation, several important principles can be seen to emerge. First, the representation we must choose for simple addition to occur with coherent light is different than in the case of incoherent light: in the coherent case, we must use the *amplitudes* (containing phase information), whereas in the previous (incoherent) case addition is linear in the *intensities*. Second, the input-output transformation represented by Equation (3) reveals an easy method for implementing both addition and subtraction: we merely set the phase difference to $-\pi/2$ for addition, and to $\pi/2$ for subtraction. This can be accomplished either by adjusting the relative path lengths of the two input beams, or by inserting an appropriately oriented wave plate in one of the two beams. In the incoherent case treated above, no such algorithm exists since we are adding intensities (which are positive definite quantities), and direct subtraction is not possible without intermediate intervention by an active optical or optoelectronic device. Third, note that the second output beam is now not symmetric with the first:

$$a'_{out} = \frac{1}{\sqrt{2}}(-a_1 + ia_2e^{i\phi}). \quad (4)$$

This asymmetry in output amplitudes directly results from the asymmetry between the phases of the reflected and transmitted components in a partially transmitting mirror [Haus, 1984].

Since the intensity of an optical beam is related to its amplitude by the relation:

$$I_m = (\mathbf{a}_m^*)^T \cdot \mathbf{a}_m \quad (5)$$

in which $\mathbf{a}_m$ is the vector amplitude of the wave representing its polarization, $*$ represents the complex conjugation operation, and T represents the transpose of the vector, the output intensities in the two coherently summed channels are given by:

$$I_{out} = \frac{1}{2}[a_1^2 + a_2^2 + 2a_1a_2 \sin\phi], \quad (6)$$

and

$$I'_{out} = \frac{1}{2}[a_1^2 + a_2^2 - 2a_1a_2 \sin\phi], \quad (7)$$

as shown in Figure 15.1(c). From these two equations, it can be seen that for arbitrary values of the phase shift ϕ , the output intensity is not simply related to the sum of the input intensities, but instead has a seemingly undesirable cross term. We can use this cross term to advantage by noting that for phase shifts of 0 or any integer multiple of π , both output intensities are equal, and reduce to the expression previously noted for the incoherent case (Equation (1)). Thus for coherent illumination, we can either perform addition directly with the amplitudes, or with the intensities if we are careful about proper phasing of the input signals. One difficulty with the former approach is that most detectors are linear in intensity but not in amplitude, as will be discussed in further detail below.

In direct analogy with the analysis presented for the case of incoherent multiple signal beam summation, we can extend the above equations to include the case of multiple collinear coherent inputs using appropriate combinations of beamsplitters. For N coherent input beams summed optimally, the output amplitude is given by:

$$a_{out} = \frac{i^{N-1}}{\sqrt{N}}[a_1 + a_2 + a_3 + \cdots + a_N]. \quad (8)$$

In order to achieve this result, one must again use $N - 1$ beamsplitters with (intensity) transmissivities identical to those employed in the incoherent case, and in addition the

relative phases must be arranged such that the phase difference between a_2 and a_1 is $-\pi/2$, and each successive beam is *increased* in phase by $\pi/2$.

In all of the cases described above, we have constrained the problem by requiring that the output beams all be *collinear*, and in fact many proposed neuro-optical architectures implicitly demand such a constraint. We will show in later sections, however, that this is perhaps an unnecessary and in many cases undesirable constraint. Hence we consider here also the case of optical addition with *noncollinear* output beams, requiring instead only that the summed beams fill the same detector aperture. There are in fact a number of interesting variants of these two constraints, but we will limit the discussion to the two principal cases only. One possible such configuration is shown in Figure 15.2(a), in which two incoherent signal beams are summed within the detector aperture by using two 100% reflecting mirrors, producing an output intensity given by:

$$I_{out} = I_1 + I_2. \quad (9)$$

This output intensity is uniform across the detector aperture, as shown schematically in the figure. In addition, relaxation of the requirement for collinearity can be seen to now allow for the use of mirrors instead of beamsplitters, eliminating the loss we found in the previous case at the expense of increased *angular* multiplexing.

FIGURE 15.2 Illustration of optical addition utilizing mirrors: (a) angularly multiplexed *incoherent* beam geometry; (b) angularly multiplexed *coherent* beam geometry.

Up to this point in the discussion, we have had to consider the *phase* of the optical wavefronts only for the case of collinear, coherent addition. In that case, we only needed to use the *relative* phase shift between the two beams to derive Equations (6) and (7), since the phase shift is constant in both space and time over the detector spatial and

temporal integration windows. In order to consider the case of *noncollinear, coherent* addition, however, we must allow for the space-variant phase shifts that naturally result when two coherent wavefronts cross at a non-zero angle. These effects are automatically taken into account if we express each wave (beam) amplitude in a form that incorporates both its *magnitude* and its *phase* everywhere in space at a given instant of time. For a plane wave (in which the planes of constant phase are oriented normal to the direction of propagation), it proves convenient to use the form $\mathbf{a}\exp(i\mathbf{k}\cdot\mathbf{r})$, in which $\mathbf{a}$ is a vector representing the polarization of the optical wave (its amplitude in each of the principal coordinate directions), $\mathbf{k}$ is the *wave vector* of the optical wave (defined as a vector with direction normal to the planes of constant phase, and with magnitude $|\mathbf{k}| = 2\pi n/\lambda$, in which n is the refractive index of the propagation medium and λ is the wavelength of the light wave in vacuum), and $\mathbf{r}$ is a position vector defined from an arbitrary origin in space ($\mathbf{r} \equiv x\hat{x} + y\hat{y} + z\hat{z}$, in which $\hat{x}$, $\hat{y}$, and $\hat{z}$ are unit vectors along the Cartesian coordinate axes).

For the case of coherent illumination, then, the result of noncollinear summation is as shown schematically in Figure 15.2(b). In this case, the phase of each input wave varies across the detector aperture (assumed to lie in the plane $z = 0$) at a rate that is a function of the angular separation of the incident beams, as well as of the angular deviation of the bisector from normal incidence. This condition can be represented by writing the wave amplitudes with a space-variant phase, which is in turn dependent on the x -component of the wave vector k_x in the form $a_1\exp(ik_x x)$. The two waves will thus interfere in the plane of the detector, forming essentially a new wave with a local amplitude given by the sum of the incident amplitudes. The resulting intensity pattern has both a space-invariant (uniform) and a space-variant (sinusoidal) component:

$$I_{out} = a_1^2 + a_2^2 + 2a_1a_2 \cos 2k_x x. \quad (10)$$

If we assume that the detector is linear in intensity over the dynamic range represented by this equation, and furthermore that the detector is uniform in responsivity over its aperture, then the output from the detector will be the *spatial average* of this interference

pattern, resulting in an output intensity that is in fact a sum of the input intensities, as represented by Equation (9). In this case, the result is independent of the relative phase shift (difference) between the two beams at their points of entry into the beam combiner system, since such a phase shift will merely result in a translation of the interference pattern without altering its integrated value. This result also can be extended to the case of multiple input signal beams, with the stipulation that mirrors cannot be allowed to occlude each other; hence, for a given beam width, only a certain number of beams can be combined without loss by means of this method without overcrowding the available angular spectrum.

The operation of optical multiplication is fundamentally different in a number of ways from those of addition and subtraction. Perhaps the most important difference is that the multiplication of either beam amplitudes or intensities cannot be accomplished directly, but must instead utilize a nonlinear medium of some form within which the beams can interact. There are two principal types of interactions to consider: those in which the two beams must be present *simultaneously* in order to form the desired product, and those in which the two beams are utilized *sequentially* in time. In general, the former interactions tend to operate on the amplitudes and hence require mutual coherency, whereas the latter interactions typically form products of (incoherent or coherent) intensities, which are therefore more straightforward to detect with currently available intensity-sensitive detectors.

Simultaneous multiplication of two optical beams is suggested by Figure 15.2(b), in which two coherent signal beams are angularly multiplexed to form the interference pattern given by Equation (10). Note that the space-variant part of the output intensity in the plane of the "detector" is proportional to the product of the amplitudes, *i.e.*, is of the form $2a_1a_2 \cos 2k_x x$. If instead of employing a uniform (spatially averaging) detector as before, we were now to employ a space-variant detector sensitive to the local intensity, it is possible to record this modulation term along with the unmodulated (uniform) bias represented by the squares of the two amplitudes. If in addition the "detector", for example, is assumed to generate a change in either its absorption coefficient or its refractive index as a function of the recorded intensity pattern (for a given exposure), a diffraction grating

will be formed. The resulting diffraction grating can be characterized by an amplitude that is proportional to the product of the input signal beam amplitudes, and that can be probed by a third so-called "readout" beam. This is at once the basic principle of holographic recording (as explained in more detail below in the subsection on "Photonic Interconnections"), and at the same time allows the implementation of the multiplicative operation for coherent inputs. It should be noted that although this process produces a useful result for the case of two inputs, extension to larger numbers of inputs is not trivial, and requires the utilization of higher order terms in the susceptibility tensor (representing the complex dielectric constant) for implementation. The one exception to this rule is the use of the probe (readout) beam intensity I_p as a third effective input, in which case output intensities proportional to either $I_p a_1 a_2$ or $I_p I_1 I_2$ can be detected depending on the operational parameters of the recording medium and readout configuration.

If the simultaneity requirement is relaxed to allow for sequential interactions in an intervening photosensitive medium, then it is possible to multiply two incoherent input signals by means of the simple generic scheme shown in Figure 15.3. The medium again acts as an effective detector for beam 1, generating a transmittance (in its range of linearity) proportional to the intensity of beam 1. This transmittance can be generated either directly, or through the exposure given by the product of the intensity and the exposure time as in the familiar case of photographic film. Beam 2 is effectively employed as a probe beam, such that the output intensity is given by:

$$I_{out} = c I_1 I_2, \quad (11)$$

as desired, in which c is a proportionality constant subject to the constraint that:

$$c I_1 \leq 1. \quad (12)$$

This process, as in the previous case, is extendable to accommodate an arbitrary number of inputs by iteration, unfortunately resulting both in a lengthy generation sequence for a large number of inputs, and in the potential for significant nonlinear effects with a heavily constrained overall dynamic range. For the case of N input beams, we can utilize $N - 1$

exposure steps in combination with $N - 1$ intermediate readout steps and a final readout step with beam N to generate an output intensity of the form:

$$I_{out} = c^{N-1} I_1 I_2 I_3 \cdots I_{N-1} I_N, \text{ with } c^{N-1} I_1 I_2 I_3 \cdots I_{N-1} \leq 1. \quad (13)$$

In direct analogy to the case of summation, we could instead utilize a binary tree structure, which requires only $\log_2 N$ time steps but uses the same *number* of devices.

Finally, it should be noted that this latter process of incoherent beam multiplication through an intervening medium by sequential illumination is suggestive of the process of *spatial light modulation*, in which the same basic concept is extended to cover a two-dimensional array of multiplication elements. In fact, this process is an essential component of the general area of photonic switching, to which we will turn our attention below.

FIGURE 15.3 Illustration of optical multiplication utilizing a medium with variable transparency.

Before turning to the topics of photonic switching and photonic interconnections, we conclude this section with a discussion of the fourth principal computational process that can be performed optically (as opposed to optoelectronically, as discussed below): that of the incorporation of functional nonlinearity. Although many types of functional nonlinearities are of interest in a generalized analog computational system, those of primary utility in the neural network environment are for the most part threshold-like in nature. A threshold function $f_T(x)$ of some input variable x (such as the input intensity, for example) can be described in general by:

$$\begin{aligned} f_T(x) &\cong T_{min}; & -\infty \leq x < x_1 \\ f_T(x) &= m(x); & x_1 \leq x < x_2 \\ f_T(x) &\cong T_{max}; & x_2 \leq x \leq \infty \end{aligned} \quad (14)$$

in which the function $m(x)$ is a monotonic function with a minimum value of T_{min} and a maximum value of T_{max} . For a step function response, the function $m(x)$ can be eliminated by setting $x_1 = x_2$. In many cases, a smoother transition between the two extreme states has been found to generate enhanced network stability and faster settling times. In such cases, the function $m(x)$ may be taken, for example, as a sigmoid with an exponential onset and an asymptotic approach to the saturation level.

The incorporation of such nonlinear functionality by direct optical means can be achieved through the use of a number of different types of nonlinear materials; such materials typically exhibit a change in their refractive index or absorption coefficient proportional to the first and higher order powers of the local optical intensity. One example of such a material is photographic film, which after development exhibits a (negative) sigmoid-like exposure characteristic, with a saturation value determined by the maximum optical density achievable within a given film thickness. (The optical density (OD) of a medium is given by the negative of the decadic logarithm of its transmittance; for example, a film that transmits 1% of the incident illumination has an optical density of two (OD2)). Another common example of an optical nonlinearity is the photoconductive saturation behavior of certain semiconductor materials such as cadmium sulfide, zinc selenide, and silicon. In this latter case, the distinction between an "all-optical" nonlinearity and an optoelectronic nonlinearity becomes somewhat blurred, as the photoconductor can be thought of as a light-sensitive electronic device.

Such optical techniques for the generation of functional nonlinearities at present suffer several inherent disadvantages, in that they often require either an off-line post-exposure development step (which is unsuitable for real time operation at high frame rates), long response times, or very high optical intensities to achieve saturation. In addition, such materials as of yet have not proven to be readily programmable, which is often a desirable feature from the systems perspective in order to accommodate variable threshold functions, gain, saturation values, and offsets. As we will see in the next section, the incorporation of electronic circuitry with optical detectors and modulators to achieve *optoelectronic* nonlinearities can in fact greatly increase the threshold sensitivity and operational bandwidth of nonlinear switching elements, while simultaneously providing flexible programming ca-

pabilities.

Photonic Switching

The switching function, that of providing an output that is (perhaps nonlinearly) dependent on one or more inputs, is a principal distinguishing characteristic of neuron units. Electronic circuit elements (particularly as configured by very large scale integration techniques) are quite well suited to the switching task, as long as the number of inputs (representing the fan-in) and the number of outputs (representing the fan-out) are both kept relatively small (less than a few hundred or so for the case of analog fan-in and fan-out). However, for neural network implementations that demand a high degree of connectivity (with a concomitantly large number of neuron units), the required gate count as well as the *area* required for interconnection routing in purely electronic implementations rapidly gets out of hand.

The fundamental aspects of the fan-in and fan-out components of the switching function are quite distinct, and lead to different types of demands on the chosen implementation technology. The *fan-in* of a number of inputs requires that a particular functional relationship be established between the generated output, on the one hand, and the set of inputs, on the other. In the case of a neural network, the output typically depends on both sums and differences of various combinations of the inputs. Therefore, a given implementation technology must properly generate the requisite logical or functional relationship, as well as provide for an appropriate physical input mechanism (*e.g.* the input leads in the case of an electronic implementation). For electronic circuits, the network area required for the provision of input leads and functional circuitry typically scales directly with the number of inputs, which is an unfortunate dependence when the number of inputs is large. *Fan-out*, on the other hand, usually implies the broadcast of a single output value to a number of (input) locations or nodes. In electronics, the output power required to drive the inputs to a large number of nodes scales directly with the number of these nodes, which again does not scale favorably (but turns out to be an unavoidable penalty in any case). Significant

(N -fold) fan-out often involves the incorporation of high power driver circuitry, which may have to be duplicated M times ($M < N$) in order to avoid unacceptable loading of the output stages.

The combination of both fan-in and fan-out components of the switching function reveals a further demand on the real estate required for the establishment of weighted *interconnections*. In a fully connected neural network with N neurons, for example, area must be provided for the incorporation (storage and programming) of N^2 independent weights as well as N^2 independent signal pathways. Hence, the chip area required in a VLSI circuit implementation of such a fully connected neural network will scale at least as the square of the number of neuron units ($O(N^2)$). Network segmentation into a number of interconnected chips can help somewhat to expand the network size beyond the limitation imposed by applying this constraint to a single chip. However, the limiting factor in the multiple-chip case rapidly becomes interchip communication (I/O), as pinouts from VLSI chips of greater than two hundred or so are not technologically feasible at present.

Photonic implementations of neural networks take advantage of the simple beam-combining mechanisms outlined above to multiplex inputs and outputs, and as such exhibit much higher capacity for fan-in and fan-out than do typical electronic implementations. The utilization of *optical* rather than electronic interconnections for the fan-in and fan-out functions provides for completely different scaling laws at large numbers of inputs and outputs to a given neuron unit, as described in more detail in the following subsection ("Photonic Interconnections").

Even given photonic interconnections with a high degree of fan-in and fan-out capability, the nonlinear functional (switching) relationship between the output and combinations of the inputs must still be provided for. For purposes of neural network implementation, the primary photonic switching component is the *spatial light modulator*, a device that alters either the amplitude or phase across an expanded probe beam in response to the local intensity (or exposure) across an input (writing or recording) beam.

The simplest example of a spatial light modulator, albeit one that cannot operate at real time frame rates, is photographic film. Following exposure to an information-bearing optical field, in which an image of a given scene is brought into focus on the two-dimensional

plane of the film, a "latent" (undeveloped) image is formed within the photographic emulsion on the surface of the film. Chemical development is used to transform the latent image into a measurable change in the optical transparency (transmissivity) of the film, which can then be "read out" or probed by secondary illumination to reveal features of the recorded scene. In this context, slide projection is in fact the equivalent of *amplified* readout with a probe beam, in the sense that the reconstruction of the image is accomplished at a much higher level of intensity (for a longer period of time) than the original exposure.

As can be seen from this example, the basic functions performed by a spatial light modulator are those of *detection*, *functional transformation*, and *optical modulation*, as shown schematically in Figure 15.4(a). In the case of photographic film, the detection process occurs at photosensitive centers during exposure, the functional transformation (the transfer function that relates the output transparency to the input exposure) is incorporated during development and fixing, and the optical modulation process is accessed during readout. This division of the spatial light modulation function into three key elements is particularly useful in the discussion of optoelectronic spatial light modulators, which typically consist of separately identifiable detectors, control circuitry, and modulation elements, as shown schematically in Figure 15.4(b-d). This functional division also allows extensive use to be made of sophisticated electronic circuitry deployed locally within each pixel, both to generate programmable nonlinear control functions and to compensate to a certain degree for the nonidealities inherent in the optical detection and optical modulation elements.

FIGURE 15.4 Fundamental principles of spatial light modulator function: (a) block diagram of the principal functions of an optically-addressed spatial light modulator, including the detection, functional implementation, and modulation functions; (b) schematic diagram of an $N \times N$ array of spatial light modulator pixels, in which three pixels are shown in different transmission states; (c) expanded view of the pixel array, showing an incomplete fill factor within each pixel; (d) expanded view of a single pixel within the array, illustrating

ing one possible pixel configuration that incorporates two detector elements D_1 and D_2 , control electronics for impedance matching and functional implementation, and two modulator elements, shown here in different transmittance states.

Up to this point in the discussion, we have focused on *optically-addressed* spatial light modulators (OASLMs) that respond locally to the incident light intensity, as this light detection function is common to most envisioned photonic neural network architectures. Another way of controlling the modulation within an array on a pixel-by-pixel basis is to configure the spatial light modulator such that it can accept a serial or parallel electronic input signal, which can be decoded (or demultiplexed) to drive each individual modulator element. Such *electrically-addressed* spatial light modulators (EASLMs) can be driven, for example, by the output of a television camera to again combine the functions of detection, functional transformation (which may be accomplished in an external circuit), and modulation. One advantage of such a combination is the current advanced state of the art in closed circuit television cameras (CCTVs), which exhibit exceedingly high performance at relatively low cost. One notable disadvantage, however, is the implied limitation on the frame rate of the combined device, since most high resolution TV cameras are designed to operate at less than one hundred frames per second.

Over the past two decades, a wide range of physical modulation mechanisms have been investigated for use in various types of spatial light modulators. Such mechanisms include the modulation of the index of refraction or birefringence in single crystal materials by means of an applied electric field (the *electrooptic* effect), the reorientation of liquid crystal molecules (producing in turn a change in the index of refraction or birefringence) by either an applied electric field or by local optically-induced heating, changes in coloration produced by optical absorption (the *photochromic* effect), modulation of the polarization of reflected light by application of local magnetic fields (the *magnetooptic* effect), surface deformations in a thin film or membrane induced by either applied electric fields or local optically-induced heating, changes in the local refractive index induced by the application of pressure or by the transmission of an acoustic wave (the *acoustooptic* effect), and

electric field modulation of the absorption or dispersion properties of semiconductor device structures. The utilization of these physical modulation mechanisms in various spatial light modulator configurations has been addressed in a number of review articles [Tanguay, 1985; Warde, 1987], journal special issues [e.g., *Spatial Light Modulators for Optical Information Processing*, 1989], and topical conference proceedings [e.g., *Spatial Light Modulators and Applications*, 1990].

The principal configurational and operational characteristics of spatial light modulators that are of interest for application to neural networks include optical sensitivity, write (input) wavelength, read (output) wavelength, input-output transfer function, functional programmability, operational bandwidth, degree of integration, pixel size, total number of pixels per chip, output modulation contrast ratio, dynamic range, and dissipated power density. In many cases, these characteristics are interdependent, and thus impose at times contradictory design constraints that must be optimized in the overall systems context. The fundamental and technological limitations that affect device design and performance are discussed further below and in a succeeding section.

As is the case for electronic circuitry, both monolithic and hybrid approaches to the development of optoelectronic spatial light modulators with suitable functionality have been employed. In the *monolithic* approach, the detectors, control circuitry, and modulation elements within each individual picture element (pixel) are integrated within a single class of materials on a supporting substrate, as shown schematically in Figure 15.5. An example of such an approach is the integration of $p-n$ or $p-i-n$ junction photodiodes with metal-semiconductor field effect transistors (MESFETs) [Sze, 1981a] to drive multiple quantum well (MQW) optical modulators based on the quantum confined Stark effect (QCSE) [Miller, 1990], all fabricated by means of photolithographic processing with multiple mask levels on gallium arsenide ($GaAs$) substrates. In Figure 15.5, two distinct approaches to the monolithic integration of spatial light modulators are illustrated, differentiated primarily by the method employed to physically or electrically isolate (pixelate) the modulator elements.

Two particularly critical parameters of spatial light modulators used in neural network implementations are the contrast ratio and dynamic range of the modulator. Their values

can in certain cases be increased by incorporating the active modulation layer (for example, a multiple quantum well device) within a symmetric or asymmetric optical (Fabry-Perot) cavity [Whitehead, 1989a; Whitehead, 1989b; Whitehead, 1989c; Yan, 1989]. The asymmetric case is shown schematically in Figure 15.5(a), in which two multilayer Bragg mirrors are used to form a reflective cavity with a high reflectivity (R) on the substrate side, and a lower reflectivity on the air-incident side. One of several advantages of monolithic integration is the potential for utilizing common components for multiple purposes. For example, the basic MQW modulator structure can also be used as a $p-i-n$ photodetector by application of appropriate bias voltages, as shown in Figure 15.5(b). To date, significant progress in such monolithically integrated optical modulators has been achieved, although spatial light modulators with large numbers ($> 10^4$) of pixels have not yet been fabricated that exhibit the relatively high degree of integration described above.

FIGURE 15.5 Examples of monolithically-integrated spatial light modulators. The chosen examples incorporate photodetectors, control circuitry, and multiple quantum well modulators within each pixel on a single gallium arsenide ($GaAs$) substrate. In (a), the control electronics and photodetector elements are fabricated following the photolithographic definition and physical isolation of the modulator elements, while in (b) a buffer (isolation) layer is used to allow fabrication and interconnection of all of the elements without chemical or ion beam etching.

In the *hybrid* approach, on the other hand, certain of the device functions may be integrated on a substrate within one materials system (with its associated process technology) in order to optimize either their performance characteristics or manufacturability, while others are integrated on a separate substrate within a different materials system (with a necessarily distinct process technology). Following separate processing sequences for each individual component, the two substrates are then interconnected (bonded together) such

that the mating pixels on each substrate are in pairwise electrical contact. For example, several currently investigated types of spatial light modulators (SLMs) incorporate the detection elements and control circuitry on a silicon (Si) substrate utilizing standard VLSI design rules, while the modulation elements are based in a separate technology (such as multiple quantum well structures integrated on a $GaAs$ substrate). Alternatively, hybrid spatial light modulators can be fabricated on a single common substrate, with additional functionality provided by the growth, deposition, or coating of a second active material onto the substrate. Examples of this type of hybrid SLM include silicon VLSI/ferroelectric liquid crystal devices [Drabik, 1990] and silicon/PLZT devices [Lin, 1990]. Such hybrid SLMs are also in the early stages of advanced development, and are the subject of current intensive research and development efforts [*Spatial Light Modulators for Optical Information Processing*, 1989; *Spatial Light Modulators and Applications*, 1990].

Using either of these two approaches to spatial light modulator fabrication, devices based on both transmissive and reflective readout can be constructed, with different implications on the overall systems design in each case. In particular, the reflective mode can be used to advantage in configuring a hybrid-integrated SLM to mate the detection and control circuitry functions of the device with the optical modulation function. Use of the reflective readout configuration allows the detection and control circuitry to be integrated on a substrate that is opaque to the readout illumination wavelength [Kyriakakis, 1990], as shown schematically in Figure 15.6.

FIGURE 15.6 Example of a hybrid spatial light modulator, in which the photodetectors and control electronics are fabricated on a silicon substrate, and the multiple quantum well modulator elements are fabricated on a gallium arsenide ($GaAs$) substrate. The two sets of devices are bump contacted on a pixel-by-pixel basis to provide parallel electrical continuity.

As an example of the degree of functional integration currently envisioned for spatial

light modulators that are specifically designed for photonic implementations of neural networks, a silicon-based CMOS chip has recently been designed and fabricated [Asthana, 1990a; Asthana, 1990b] that incorporates two input detectors, control circuitry, and two (optical modulator) output drivers within each $100 \times 100 \mu m$ pixel as shown schematically in Figure 15.7(a). It should be noted that these current dimensions do not in any sense represent a lower limit, but rather a practical size for laboratory demonstrations and experiments, as well as a useful size from the perspective of neural network applications. The pixel layout allows for two $30 \times 50 \mu m$ detectors, followed by a 15 transistor dual input, dual output differential amplifier that implements a sigmoid-like transfer function, with externally programmable saturation characteristics. Output pads are also provided for hybrid bonding (by bump contact techniques [Shirouzu, 1986]) to an *InGaAs/GaAs* multiple quantum well modulator structure fabricated on a *GaAs* substrate [Kyriakakis, 1990]. Utilizing $2 \mu m$ CMOS design rules, the control circuitry easily fits within $25 \times 100 \mu m$, leaving adequate space for the modulator output pads as shown in Figure 15.7(b, c). This currently allows for the integration of 10^4 pixels per cm^2 , or 6×10^4 pixels per in^2 . The functional operation of this circuit will be discussed in the section, "An Implementation Strategy", below.

FIGURE 15.7 VLSI layout of a generalizable silicon-based spatial light modulator structure: (a) neuron pixel layout; (b) photograph of a single neuron unit in VLSI implementation, with probe pads substituted for the two detectors (bottom) and for contact to the two modulation elements (top); (c) photograph of a 6×6 array of neuron units on a VLSI chip that incorporates additional test circuitry.

Before leaving the subject of photonic switching, it should be noted that the general principles outlined above can be used to design a wide variety of mutually compatible devices with different functionalities as well as different tradeoffs among the set of con-

figurational and operational parameters. For example, it is relatively straightforward to design time-integrating and time-differentiating circuits; sharp (step-like) thresholds; level slices; sigmoid-like functions, their complements, and their derivatives; inverters; and logarithmic amplifiers. Many such functions can be implemented with only a few integrated components, such as capacitors, diodes, transistors employed as current amplifiers, and biased transistors employed as resistors [Mead, 1989]. Therefore, these functions can easily be incorporated within each pixel (neuron unit) of a two-dimensional spatial light modulator, as well as in some cases between pixels for the implementation of non-local (other than pointwise) operations such as automatic gain control and nearest-neighbor inhibition.

Photonic Interconnections

Given that the neuron units are to be represented by individual pixels within a two-dimensional spatial light modulator, interconnections must now be established between each individual neuron unit (pixel) and many (if not all) other neuron units. As such, the chosen interconnection scheme must be capable of the appropriate degree of fan-in and fan-out, be characterized by sufficient transmission bandwidth in each channel, and be scalable to relatively large numbers of neuron units. In addition, the neural network paradigm presents the additional requirement that the interconnections be *weighted*, such that the output from a given point-to-point interconnection is proportional to the product of the input and a stored constant or weight. It is in fact this requirement that eliminates from consideration a large number of possible switching networks that provide full reconfigurability in a non-blocking manner (such as a crossbar or shuffle-exchange network), but without the capacity to incorporate weights within each interconnection pathway. In adaptive networks (those that incorporate learning algorithms), these interconnection weights must have the capability of being updated in a manner determined by the particular learning algorithm employed. A nontrivial consequence of these last two requirements is that the interconnection weights must be *stored* for at least as long as the average iterative computation, if not *much* longer; yet, they must simultaneously exhibit dynamic

programmability if the network is to exhibit either short-term or long-term plasticity.

For very small numbers of neurons with a low degree of connectivity, one possible way of forming the interconnection network would be to use fiber optic transmission lines with modulated semiconductor laser diodes as sources and optical receivers as detectors, much like a fiber optic local area network. The weights could be incorporated by means of a variable gain amplifier at either end of each fiber optic link, with weight storage in local dynamic random access memory (RAM) or static read only memory (ROM) circuits. Unfortunately, the sheer bulk of each transmitter, receiver, and fiber optic link precludes scalability to large neural network systems. For example, a fully connected twenty neuron network would involve four hundred sets of sources, transmission lines, and detectors, which would currently represent a prohibitive requirement. The same would be true of a fifty neuron network with a fan-out and fan-in of eight, representing a relatively low degree of connectivity.

In order to be able to satisfy the interconnection requirements for a large number of neuron units that are fully or nearly fully interconnected, the appropriate photonic technology to employ is that of *holographic* interconnections, in which the weights as well as the interconnection patterns themselves are stored as holograms in either a fixed (static) or real time (dynamic) holographic recording material. In this section, we first discuss the basic principles that apply to the utilization of holographic recording for point-to-point interconnections. Next, we describe the physical origins of a number of complexities with holographic interconnection schemes that lead to both interchannel crosstalk and throughput losses. An architecture that lends itself to the minimization of such complications will be described in detail in the section, "An Implementation Strategy", below. Finally, the potential for incorporation of real time volume holographic recording media such as photorefractive materials in holographic interconnection networks is addressed.

The essential principle of holographic recording, that of the space-variant interference of two mutually coherent wavefronts, was discussed briefly in reference to Equation (10) and is illustrated in Figure 15.2(b). In this figure, two angularly separated (noncollinear) collimated beams are incident on a photosensitive material, such that their mutual interference locally exposes the material to the intensity distribution given by Equation (10).

In Figure 15.2(b), the photosensitive material was assumed to spatially integrate across the interference pattern, producing an output that depends on only the spatial *average* of the intensity distribution. Suppose now that we use instead a photosensitive material with the property that its *local* index of refraction or absorption coefficient depends on the *local* incident intensity (exposure), which allows the complete interference pattern to be *recorded*. The resulting change in the local optical properties of the medium may either be immediate (as in the case of a photochromic transformation, for example), or may require development following exposure (as in the case of bleached photographic negatives or dichromated gelatin thin films). Figure 15.8(a) shows such a detection or recording geometry in which a thin semi-transparent layer of photosensitive material acts as a quasi-planar holographic recording medium. The interference pattern produced by the mutually coherent signal and reference beams within the holographic recording medium is recorded to form a diffraction grating within the volume accessed by both beams simultaneously, as shown in the figure. For simplicity in Figure 15.8 (as well as in subsequent figures), we have not shown the refraction of the incident and transmitted beams at the input and output faces of the holographic recording medium that occurs due to a difference between the refractive indices of the medium and its surround. The amplitude (and intensity) of the reflected beam shown in Figure 15.8(b) depends directly on the index difference, and represents a throughput loss on readout.

FIGURE 15.8 A simplified holographic recording configuration: case of plane wave signal and reference beams, and a *thin* holographic recording medium; (a) recording, and (b) reconstruction with a plane wave readout beam.

Consider first the case of an exposure-dependent refractive index variation. Illumination of such a space-variant modulation of the refractive index by a coherent collimated beam of the same wavelength λ as the exposure (writing) beams will result in a diffraction pattern consisting of several collimated beams, each emanating in a characteristic direction as

shown schematically in Figure 15.8(b), and as given by the following equation:

$$\mathbf{k}_{mx} = \mathbf{k}_{rx} + m\mathbf{K}_G; \quad |\mathbf{K}_G| = \frac{2\pi}{\Lambda_G} = |\mathbf{k}_2 - \mathbf{k}_1|; \quad m = 0, \pm 1, \pm 2, \dots \quad (15)$$

In this equation, $\mathbf{k}_{mx}$ is the x -component of the wave vector of the m^{th} diffracted beam (diffraction order), $\mathbf{K}_G$ is the wave vector (assumed oriented along the x -axis) of the interference pattern (diffraction grating) formed by the two writing beams (with wave vectors $\mathbf{k}_1$ and $\mathbf{k}_2$), $\mathbf{k}_{rx}$ is the x -component of the wave vector of the incident readout beam propagating in the x - z plane, and Λ_G is the spatial wavelength of the diffraction grating. The multiple diffracted orders result from the phase modulation of the readout (probe) beam by the refractive index modulation $n(x)$ of the thin holographic grating; the magnitude and phase of the readout beam amplitude immediately after passing through the hologram (located at the position z_0) can be written in the form:

$$A_{diff} = a_r e^{i(k_{rx}x + k_{rz}z_0)} e^{i\phi_G(x)}, \quad (16)$$

in which A_{diff} is the amplitude of the diffracted wavefront, $\phi_G(x) = 2\pi n(x)d/\lambda$ is the local phase shift induced by the diffraction grating (assumed to be of thickness d), $a_r e^{i(k_{rx}x + k_{rz}z_0)}$ is the incident readout beam amplitude at the exit plane of the hologram (z_0), and k_{rx} and k_{rz} are the x and z components of the wave vector $\mathbf{k}_r$, respectively. Each of the diffracted orders can then be directly associated with a corresponding Fourier component of the modulated amplitude [Goodman, 1968], which can be expanded in terms of the form:

$$A_m e^{imK_G x}. \quad (17)$$

In order to assess the effectiveness with which the holographic grating diffracts the incident beam into a particular diffraction order, it is convenient to define the *diffraction efficiency* η of each order as:

$$\eta \equiv \frac{|A_m|^2}{|a_r|^2}. \quad (18)$$

The essential diffraction properties of thin absorption gratings (in which the modulation

occurs in the local absorption coefficient) are the same as for the case of thin pure phase gratings, with two principal exceptions: (1) for sinusoidal absorption gratings, the diffracted orders are limited to $m = -1, 0$, and 1 ; and (2) the presence of absorption significantly decreases the maximum first order diffraction efficiency that can be achieved.

In order to illustrate how such holographic gratings can be employed to generate weighted point-to-point interconnections, we need to introduce two additional concepts: the lens as an angle-to-position encoder, and the superposition of holographic gratings recorded with different diffraction efficiencies. The first concept can be understood with reference to Figure 15.9, in which a simple lens is placed one focal length away from a point source in the input plane of a photonic interconnection, and a second simple lens is placed one focal length away from the output plane. What is normally thought of as the focal property of a lens results in the generation of a collimated beam (a beam comprising both parallel rays and planar wavefronts) following the first lens, with an *angle* (both in and out of the plane of the page) that depends on the *position* of the point source in the focal (input) plane. In this sense, the first lens acts as a position-to-angle encoder, providing a one-to-one correspondence between the input location and the output collimated beam angle. Depending on the nature of the grating stored within the holographic optical element, the collimated beam will be diffracted into a new direction characterized by a *different* angle. The second lens will then focus the diffracted beam to a point in the output plane that depends on this angle, thus acting as an angle-to-position encoder. The utilization of different orientations of gratings within the holographic optical element allows for the interconnection of any arbitrary point in the input plane to any other point in the output plane.

FIGURE 15.9 A point-to-point interconnection system, using a holographic optical element (HOE) for interconnection routing, and lenses as position-to-angle and angle-to-position encoders. In this example, the holographic optical element effectively performs an input angle to output angle transformation, such that light emitted (or transmitted) at point p_1 in the input plane (P_1) is detected at point p_2 in the output plane (P_2).

Suppose now that we do in fact choose to superimpose a number of planar gratings within the holographic medium, each with a different wave vector (orientation and grating period) and grating modulation (variation of the refractive index or the absorption coefficient). Assuming for the moment that the diffraction process is linear, each input point will be interconnected with a number of output points as determined by the set of recorded gratings. Likewise, each output point will be interconnected with a specified number of input points. Each interconnection will be weighted by its diffraction efficiency as determined by Equation (18), which is in turn dependent on the index of refraction (or absorption coefficient) variation recorded for each grating. As such, the holographic optical element acts as a multi-port variable beamsplitter, redirecting (diffracting) a given fraction of each input beam to a specified set of output beams. By employing lenses as described above, this feature allows the construction of a point-to-point interconnection with weights and arbitrary fan-out/fan-in (delimited only by the number of gratings recorded).

There is at least one obvious problem with the interconnection scheme outlined above, however, in that any *given* grating will connect *any* of the input points to specific output points pairwise, as shown by Equation (15). This particular feature occurs because each input point generates a collimated beam with a distinct wave vector $\mathbf{k}_i$ corresponding to a particular direction (angle) of propagation, each of which satisfies Equation (15) with a different diffracted wave vector (for each diffracted order) $\mathbf{k}_m$. The result of this degeneracy is that any recorded hologram that is designed to connect a single input point to one or more output points will in fact also connect *every* other input point to corresponding sets of output points, using the same relative interconnection pattern for each input point. This effect can be utilized to advantage, for example, in parallel digital optical computing systems with interconnection symmetry or regularity, since one simple hologram can in effect implement a very large number of point-to-point interconnections (the equivalent of wires in the case of an electronic implementations) [Jenkins, 1984]. For neural networks, however, the common requirement of nearly arbitrary (highly irregular) interconnections makes this feature undesirable.

A second problem with the proposed interconnection scheme is the presence of a multiplicity of diffracted *orders* for each diffraction grating, as shown in Figure 15.8, which occasions the connection of each input point to a number of geometrically related output points even for the case of a single stored grating.

The solution to this seeming dilemma is to extend the holographic medium into the third dimension (the direction of light propagation), creating a *volume* holographic optical element (VHOE) to take the place of the thin planar element discussed above. There are two essential properties of VHOEs that bear directly on the utilization of such elements in photonic interconnections. The first is that diffraction is limited to the first order only and all higher diffracted orders are suppressed if the holographic medium is thick enough, as defined below and as shown schematically in Figure 15.10. This occurs because each additional "layer" in the thickness direction of the holographic medium provides an additional constraint on the diffraction phenomenon; these constraints act collectively to enhance the amplitude diffracted into the first order by means of constructive interference, at the expense of the other diffracted orders.

FIGURE 15.10 Volume holographic recording with plane wave signal and reference beams; (a) recording, and (b) reconstruction, showing the elimination of the higher diffracted orders.

The second important property of a volume holographic optical element is that of *angular selectivity*; specifically, the range of input angles that can diffract from a given grating decreases as the thickness of the grating is increased. The central angles that are allowed in the case of a thick grating are the same angles that define the two beams that initially *created* the holographic grating. This property therefore eliminates the inadvertent connection of all input points pairwise to a matching set of output points, and allows for the generation of *independent, weighted* interconnections as are desired for neural network applications.

In order to differentiate “thin” grating diffraction behavior (the so-called Raman-Nath diffraction regime) from “thick” grating behavior (the so-called Bragg diffraction regime), it is convenient to define a dimensionless “thickness” parameter Q such that:

$$Q = \frac{2\pi\lambda d}{n\Lambda_G^2} \quad (19)$$

in which n is the average refractive index of the holographic recording medium, and the remaining parameters are as specified previously. In general, gratings for which $Q \geq 10$ operate well within the Bragg regime, while gratings with Q parameters less than unity exhibit unacceptable degrees of Raman-Nath character for truly independent multiplexed interconnection applications. The angular response characteristics of both planar and volume diffraction gratings are shown as a function of the Q parameter in Figure 15.11, in which the transition from pure Raman-Nath to pure Bragg behavior for increasing values of Q can be seen. Note that the number of input points that can be independently connected to an equally-sized array of output points is a decreasing function of the width of the angular response.

FIGURE 15.11 The angular alignment sensitivity of a volume holographic optical element, as a function of the dimensionless Q -parameter defined in the text. The grating strength for all of the curves (3.14 radians) is optimized to produce 100% diffraction efficiency in the limit of large Q (Bragg diffraction regime), and is not optimized for low Q gratings. Note that the diffraction efficiency is essentially independent of angle for low Q gratings, and is very strongly peaked at the Bragg angle (7.5 degrees in this case) for high Q gratings.

The throughput efficiency of a volume holographic optical element as used in an interconnection application is determined to first order by the diffraction efficiency of each

individual interconnection grating, in direct analogy to the definition of the diffraction efficiency for the planar hologram case in Equation (18). For example, for the case of an unslanted pure phase grating with equiphase fronts (*i.e.* planes of constant phase) parallel to the bisector of the recording beams with wave vector k and perpendicular to the entrance face of the volume holographic recording medium, the diffraction efficiency at the Bragg (optimum readout) angle is given by [Kogelnik, 1969]:

$$\eta = e^{-\alpha d / \cos \theta_B} \sin^2 \left(\frac{\pi \Delta n d}{\lambda \cos \theta_B} \right), \quad (20)$$

in which α is the absorption coefficient of the holographic recording medium of thickness d at the optical readout wavelength λ , Δn is the amplitude of the refractive index modulation, and θ_B is the Bragg angle defined by $2k \sin \theta_B = K_G$. As can be seen from Equation (20), the diffraction efficiency of the first order for a single grating can approach 100% if the absorption coefficient satisfies the requirement $\alpha d \ll 1$, provided sufficient index modulation Δn can be produced by the exposure process. The dependence of the diffraction efficiency on the grating strength is shown in Figure 15.12 for both thin (Raman-Nath) and thick (Bragg) pure phase diffraction gratings. The grating strength v is defined as the integrated peak phase modulation of the grating in each case, and is given by:

$$v = \frac{2\pi \Delta n d}{\lambda \cos \theta_B}. \quad (21)$$

The maximum diffraction efficiency of the thin diffraction grating is about 34%, which occurs at a grating strength of 1.8 radians. Thick diffraction gratings achieve 100% diffraction efficiency at a grating strength of π radians, at which point the diffraction efficiency of the thin grating has peaked and is nearly at its first node, as shown in Figure 15.12.

FIGURE 15.12 The diffraction efficiency of thin (Raman-Nath diffraction regime) and thick (Bragg diffraction regime) holographic gratings as a function of the grating strength.

The extremely narrow angular alignment characteristics of volume diffraction gratings in principle allow the simultaneous multiplexing of large numbers of independent, weighted interconnections to be recorded between the input plane and the output plane (c.f. Figure 15.9). In addition, the use of angular multiplexing allows for both fan-out from a given input point to a number of output points, as well as fan-in from a number of input points to a single output point.

The holographic implementation of the fan-out from a single input point to a number of output points uses several multiplexed (superimposed) holographic gratings to achieve the desired weighted fan-out, one for each output point. Consider a 4 input, 4 output interconnection as shown in Figure 15.13. For each input point x_j that we wish to interconnect to an output point y'_i , the recording process requires the pairwise coherent interference within the holographic recording medium of x_j with a second beam y_i corresponding to y'_i . The interconnection of x_1 to y'_1, y'_2, y'_3 , and y'_4 therefore requires the pairwise coherent interference of x_1 with y_1, x_1 with y_2 , and so on. This process results in the fourfold fan-out of x_1 to all of the outputs.

The fan-out from a single reference beam to a number of output beams is directly analogous to the readout of a traditional hologram (of, for example, a two-dimensional or three-dimensional image), provided that the full set of beams $\{y_i\}$ is coherently recorded with the given reference beam x_j . Although up to this point we have formulated the point-to-point holographic interconnection problem in terms of collimated (plane wave) input and output beams that record individual diffraction gratings (characterized by a single grating wave vector) within the holographic recording medium, many alternative recording and reconstruction geometries can be envisioned that produce equivalent results. In the case of traditional holography, for example, the input transparency bearing the image to be recorded is illuminated with a collimated beam, resulting in a complex diffraction pattern at the front entrance plane of the holographic recording medium. Collimated, converging, or diverging reference beams can be utilized to produce reconstructed images with a wide variety of optical imaging characteristics. Likewise, various input and output beam geome-

tries can be used in a point-to-point interconnection system to optimize the overall system characteristics, such as freedom from interchannel crosstalk, optimum use of the spatial frequency recording characteristics of the holographic recording medium, optical system complexity, and convenience of the optical layout (particularly when viewed in conjunction with associated optical subsystems).

FIGURE 15.13 Schematic representation of a 4 input, 4 output holographic interconnection, showing 4 coherent input beams x_1 - x_4 and 4 coherent recording beams y_1 - y_4 , each of which corresponds to a desired output y'_1 - y'_4 . In (a), the sets $\{x_j\}$ and $\{y_i\}$ interfere within the volume holographic medium, recording the desired interconnection diffraction gratings. In (b), a new set of input beams $\{x_j\}$ illuminates the volume holographic medium, reading out the weighted interconnection pattern and forming appropriately weighted sums at each of the outputs $\{y'_i\}$.

The fourfold fan-in of inputs x_1, x_2, x_3 , and x_4 to y_1 can likewise be accomplished by recording each of the necessary interconnections pairwise, as before for the fan-out case. The recording process for the fully implemented 4 to 4 interconnection therefore involves the generation of 16 individually weighted diffraction gratings that connect the full set of inputs $\{x_j\}$ to the full set of outputs $\{y'_i\}$. The multiplexed hologram that accomplishes this function can be recorded in a number of ways, each characterized by certain advantages and disadvantages [Psaltis, 1988].

In the fully coherent approach, the requisite gratings can be recorded by illuminating the holographic recording medium with $\{x_j\}$ and with $\{y_i\}$ simultaneously. This can be accomplished, for example, by using a spatial light modulator to store each of the sets of values, and a pair of mutually coherent readout beams to encode these values and interfere them within the holographic element. In this manner, all of the required gratings are recorded in a single exposure; however, there are two difficulties inherent in this single

exposure, fully coherent approach. The first problem is that a fully independent N to M interconnection requires NM stored interconnection weights, whereas the single exposure described above supplies only $N + M$ input values that can be used to generate the weights. The resulting interconnection matrix can in fact connect all of the input points to all of the output points, but the relative fan-out weights from each input point will be degenerate. One way to avoid this degeneracy is to illuminate the holographic recording medium with each input x_j and a full set of corresponding outputs $\{y_i\}$, sequencing through all N of the inputs (and changing the set of corresponding outputs) one at a time. This procedure generates an independent fan-out from each input point. The second problem with the single exposure, fully coherent approach is that undesirable gratings will be recorded among the $\{x_j\}$ and among the $\{y_i\}$ that can lead to considerable coherent crosstalk among the *desired* interconnection pairs. This coherent interference process diminishes the degree of independence of the interconnections.

This coherent-recording-induced crosstalk can also be avoided by sequencing the recording, but in this case each desired grating pair is recorded separately such that only one input beam x_j interferes with one output beam generator y_i (recording beam for the desired output beam y'_i) at a time. This scheme effectively eliminates the coherent crosstalk, but does not eliminate another form of crosstalk (called *beam degeneracy* crosstalk [Jenkins, 1990a; Jenkins, 1990b; Asthana, 1990c], the origin of which is described below) that can be equally severe; in addition, the complication imposed by the incorporation of such a sequential recording schedule can be a serious constraint for large $N \times M$ (N input points to M output points) interconnections, as NM independent recording steps are required for full programming of the interconnection. This proves to be particularly problematic for the rapid generation of weight updates in a large scale neural interconnection network that incorporates synaptic plasticity. Furthermore, sequential recording of holographic exposures can cause partial erasure of previously recording interconnection weights in certain types of holographic recording materials, necessitating the use of recording schedules that attempt to balance the weights recorded at the beginning of the sequence (and hence partially erased by all subsequent exposures) with the weights recorded at the end of the sequence [Psaltis, 1988]. The use of such recording schedules usually implies an overall

decrease in both the exposure efficiency and throughput efficiency of the resulting holographically recorded interconnection matrix, as well as the buildup of noise resulting from the series of space-variant erasures.

One potential scheme for reducing coherent-recording-induced crosstalk, beam degeneracy crosstalk, and sequential recording schedules involves the use of an array of coherent but mutually incoherent sources to simultaneously expose the holographic recording medium to only the desired sets of gratings [Jenkins, 1990a; Jenkins, 1990b; Asthana, 1990c]. This scheme will be discussed in detail in a later section.

The fan-out process is illustrated in Figure 15.14, in which implementations using both beamsplitters and volume holographic optical elements are shown. The case of fan-out utilizing beamsplitters is shown schematically in Figure 15.14(a). As can be seen in the figure, the input beam can be divided among the output beams with arbitrary weights set by the transmissivities of the beamsplitting elements BS_i . If the final beamsplitter is a mirror, the fan-out process can be accomplished with essentially zero throughput loss. By analogy to the beamsplitter case, as well as by direct analysis, it can be proven that the holographic fan-out process shown in Figure 15.14(b) can also be accomplished with essentially arbitrary weights, with no optical throughput loss inherent in the fan-out process itself. It is interesting to note that these two implementations differ in at least one essential feature, in that the beams fanned out from the holographic implementation originate within the same volume, while the beams fanned out from the beamsplitter implementation originate from vertically displaced beamsplitters. If we were to extend the fanned out beams in the latter case backwards toward the left hand side of Fig. 15.14(a), we could imagine replacing the three discrete, vertically displaced beamsplitters with a single, multiplexed "virtual" beamsplitter that generates the same set of output beams. One physical realization of such a "virtual" beamsplitter component is in fact the multiplexed volume hologram shown in Figure 15.14(b).

FIGURE 15.14 Schematic representation of the fan-out process for optical beams, for the case of one input and three outputs: (a) with beamsplitters

($BS_1 - BS_3$): (b) with a single holographic optical element containing three multiplexed (spatially superimposed) diffraction gratings.

The collinear fan-in process is illustrated in Figure 15.15 for both types of implementations. As was discussed above, for the beamsplitter implementation an intrinsic fan-in loss is encountered for the case of collinear fan-in, while the intrinsic loss can be circumvented by resorting to mirrors and employing angular multiplexing. For the case of volume holographic optical elements, the situation is identical, such that collinear fan-in is grossly inefficient for large numbers of fan-in interconnections to the same node. On the other hand, appropriate use of angular multiplexing can eliminate this seemingly inherent fan-in loss, giving rise to a highly multiplexed, efficient interconnection element [Jenkins, 1990a; Jenkins, 1990b; Asthana, 1990c] as described in a later section.

FIGURE 15.15 Schematic representation of the fan-in process for optical beams, for the case of three angularly distinct inputs and one combined collinear output beam: (a) with beamsplitters, showing the unavoidability of a throughput loss associated with the set of transmitted (and multiply reflected) beams; (b) with a single holographic optical element containing three multiplexed (spatially superimposed) diffraction gratings, showing an analogous throughput loss.

The physical origin of this intrinsic optical throughput loss in the case of collinear fan-in is directly related to the mechanism that gives rise to beam degeneracy crosstalk. In Figure 15.16 we show a 4 to 4 holographic interconnection that is assumed to have been recorded by the sequential exposure technique described above in reference to Figure 15.13, in order to include all 16 individually weighted interconnection gratings but none of the undesirable gratings that can give rise to coherent-recording-induced crosstalk. In

this case, readout by the input beam x_1 generates the four output beams y'_1 through y'_4 , with values given by the stored interconnection weights w_{ij} :

$$y'_i = w_{i1}x_1. \quad (22)$$

Within the holographic medium, however, each of the four output beams can in turn act as an *input* beam, generating undesired output beams in the directions x'_2, x'_3 , and x'_4 . These undesired output beams are a result of diffraction from the gratings recorded between each output generating beam y_i and the full set of input beams $\{x_j\}$. Each output beam is automatically Bragg matched (at the correct Bragg angle) to the full set of input beams due to the collinear recording geometry employed. We refer to the fan-in as *collinear* in this case because each input beam x_j that is fanned in to a given output y'_i produces an output beam in the *same* direction. The generation of diffracted intensity in the directions x'_2 - x'_4 from readout with x_1 results in a throughput loss for the interconnections between x_1 and the set of output beams $\{y'_i\}$. In addition, the throughput losses of the individual output beams $\{y'_i\}$ will not be equal in general. Furthermore, the undesired diffracted beams x'_2 - x'_4 can *also* act as input beams, generating additional output beams in the directions $\{y'_i\}$ that coherently interfere with the beams directly diffracted in those directions by the input beam x_1 . The combination of interconnection-dependent losses from the output beams $\{y'_i\}$ into the "cross-coupled" beams $\{x'_i\}$, and of interconnection-dependent coupling from $\{x'_i\}$ into $\{y'_i\}$ gives rise to an undesired redistribution of the intensities of the output beams. This phenomenon is referred to as *beam degeneracy crosstalk*, as it arises from the beam direction degeneracy (collinearity) of the output beams fanned into a single output point.

FIGURE 15.16 Illustration of the generation of crosstalk in holographic optical interconnections due to beam degeneracy: recording/readout configuration. The input beams $\{x_j\}$ are assumed to have interfered within the volume holographic medium with the set of recording beams $\{y_i\}$, producing

the desired set of interconnection gratings with weights w_{ij} . Illumination of the volume holographic medium with beam x_1 produces a 1 to 4 fanout into the output beams $\{y'_i\}$, as well as the zeroth order beam x'_1 . Due to the effects of beam degeneracy, power is also coupled into the zeroth order beams x'_2 - x'_4 , and crosstalk terms $\{c_i\}$ are introduced into the outputs.

Both the throughput loss and the beam degeneracy crosstalk that characterize holographic interconnection geometries with collinear fan-in can be estimated by numerical simulation of the diffraction process from a multiplexed grating [Asthana, 1990c]. By using the optical beam propagation method [Johnson, 1986] to simulate the diffraction process, we can analyze the 4 to 4 interconnection described above for the case of a single beam readout, as shown in Figure 15.16. The results of such an analysis are presented in Figure 15.17, which shows the diffraction efficiency of each of the four beams fanned out from the single input point, as well as the three cross-coupled beams in the directions $\{x'_j\}$ and the zero order (undiffracted) beam. For this illustration, all 16 interconnection weights are equal in magnitude. As the grating strength is increased, a significant amount of intensity is coupled into the cross-coupled components, robbing the desired fan-out beams of the desired diffraction efficiency. In addition, the *relative* diffraction efficiencies observed in the designated fan-out beams are no longer independent of the grating strength, as desired in a fully independent weighted interconnection. Extensive modeling of N -to- N holographic interconnections with collinear fan-in suggests that the throughput loss increases approximately as $1/N$, which is potentially catastrophic for large interconnection networks. In a later section, we will describe an alternative holographic recording approach that obviates this $1/N$ loss.

FIGURE 15.17 Illustration of the generation of crosstalk in holographic optical interconnections due to beam degeneracy: diffraction efficiency as a function of grating strength for the readout configuration of Figure 15.16. Shown

are the depletion of the zero order beam x'_1 and the rise of the desired output beams y'_i , accompanied by a strong buildup of the cross-coupled beams x'_2 - x'_4 .

The development of a viable photonic interconnection technology is based in no small part on the availability of appropriate photosensitive recording materials [Psaltis, 1988; Smith, 1977; Gunter, 1988; Gunter, 1989]. Many interconnection demonstration experiments have been performed in the laboratory on bleached photographic emulsions and dichromated gelatin films, both of which are thick enough (10 - 30 μm) to exhibit sufficient Bragg-like diffraction behavior to allow a limited degree of multiplexing to be incorporated. Neither material, however, exhibits capacity for real time operation, which is essential for the implementation of photonic neural networks with at least some degree of synaptic plasticity. On the other hand, one principal advantage of photographic film and dichromated gelatin is their essentially infinite read-write asymmetry, which is highly desirable in many applications as described below.

By *real time operation*, we mean that the holographic interconnections can be programmed (exposed) and used (read out) on roughly the same time scale (perhaps at $k\text{Hz}$ frame rates), without the necessity of chemical development processes or the like. By *read-write asymmetry*, we mean that the readout of a programmed interconnection should not erase the stored weights at an accumulated readout exposure equal to that of the recording exposure. Ideally, we would like to have the capability of exposing the holographic interconnection to the recording beams with essentially instantaneous "development" of the stored gratings, with the recording process characterized by very high sensitivity during the "learning" process. At the same time, we would like to be able to initiate readout of the stored interconnection pattern without altering the stored weights for a length of time equal to the desired "computation" time. Although in many applications the learning and computation times may differ by only an order of magnitude, in other cases it is desirable to compute for very long times compared with the learning phase, and yet still maintain the capacity for (slowly varying) weight updates.

The class of photosensitive recording materials that has been most extensively investi-

gated for photonic interconnection applications does in fact have the capacity for sensitive holographic recording, is available in "thick" samples that allow for the formation of Bragg-regime diffraction gratings, exhibits a high multiplexing capacity, and allows for the inclusion of modest read-write asymmetries. This class is that of the so-called "photorefractive" materials [Gunter, 1988; Gunter, 1989], which includes single crystals of semi-insulating optical materials such as bismuth silicon oxide ($Bi_{12}SiO_{20}$), bismuth germanium oxide ($Bi_{12}GeO_{20}$), lithium niobate ($LiNbO_3$), strontium barium niobate ($Sr_{1-x}Ba_xNb_2O_6$), potassium niobate ($KNbO_3$) and barium titanate ($BaTiO_3$), as well as semi-insulating semiconductors such as gallium arsenide ($GaAs$), indium phosphide (InP) and cadmium telluride ($CdTe$). The use of the term "photorefractive" to describe these materials exclusively is somewhat misleading, in that many other classes of materials are known to undergo a refractive index change following illumination as well as those traditionally included in the class described above. But at least the term is descriptive of the basic phenomenon involved, as outlined below.

In photorefractive materials such as bismuth silicon oxide, exposure to an interference pattern at an appropriate wavelength (characterized by significant photosensitivity) generates free charge carriers (electrons or holes) liberated from deep traps. The number of photogenerated carriers is in general proportional to the local intensity absorbed by the crystal; as such, the photogenerated carrier population mimics the exposure pattern in both amplitude and phase. The photogenerated carriers are free to diffuse to regions of lower intensity, or they can be assisted out of the brightest regions by application of a bias electric field to produce carrier drift. In either case, they tend to be retrapped, in turn creating a space charge distribution that has the same spatial frequency as the interference pattern. This space-variant space charge distribution produces a locally modulated electric field with the same spatial frequency (as determined by the grating spacing or grating wavelength), which in turn induces a local change in the refractive index of the photorefractive material through the linear (Pockels) or quadratic (Kerr) electrooptic effect [Kaminow, 1974]. The refractive index grating can then be probed by a readout beam to generate a diffracted beam, just as in the case of the pure phase gratings described previously.

An excellent set of review articles on the physical properties and applications of photorefractive materials has been assembled by Gunter and Huignard [Gunter, 1988; Gunter, 1989]. The state of the art is such that 1 cm^3 crystals of many of these materials have been grown, and shown to exhibit a very high degree of optical quality. Exposure sensitivities vary widely, but several crystals require of order $500\text{ }\mu\text{J}/\text{cm}^3$ for full exposure to saturation (the highest grating strength that can be achieved in that particular crystal). This corresponds to the absorption of about $50\text{ mW}/\text{cm}^2$ of optical intensity throughout 1 cm^3 of material for an exposure period of 10 msec . The range of spatial frequencies that can be supported in these materials ranges from a few lines/mm to over 2000 lines/mm. Diffraction efficiencies close to 100% have been observed in several types of crystals, while others saturate nearer to 10% for thicknesses of order 1 cm .

Optimization of photorefractive materials for interconnection device applications is under way, including the development of growth processes for large photorefractive crystal boules with a high degree of optical uniformity; the characterization of both unintentionally incorporated impurities and intentionally incorporated dopants that alter the holographic recording, readout, and storage characteristics; the use of applied d.c. and a.c. bias electric fields to enhance the holographic recording sensitivity; the use of polarization effects to enhance the reconstructed image signal-to-noise ratio; and the antireflection coating of the (typically high index) front and rear surfaces to increase the diffraction efficiency and avoid the presence of unwanted gratings due to multiple reflections [Karim, 1988; Karim, 1989a; Karim, 1989b]. In addition, the origin of electric field nonuniformities that occur within photorefractive crystals during grating recording is under active investigation, and several methods of eliminating the field collapse have been discovered [Herbulock, 1988]. Use of these methods increases both the saturation diffraction efficiency and the grating response time, resulting in more efficient interconnection devices that operate at higher recording sensitivities.

Sources and Source Arrays

In reviewing a large fraction of the journal articles published over the past decade on optical information processing and computing, including the most recent coverage of photonic implementations of neural networks, you will be inspired perhaps by the cleverness of a particular proposed architecture, or intrigued by the novel features of a particular device structure. But you will also be amazed at the apparent lack of emphasis on certainly one of the most fundamental components in any proposed photonic computational system: the source of the light! This oversight may be caused in part by direct analogy to the situation in VLSI electronics, in which it is a bit unglamorous (and probably also to a certain extent unnecessary) to concentrate on the battery or the power supply. After all, electrical power is relatively inexpensive, widely available, well characterized, and reasonably abundant. At peak usage, your home probably uses about 10 kW, most of which is dissipated in the air conditioner.

However, the situation in photonic technology is quite different. Sources of coherent optical radiation that can produce average output powers in the 10 kW range exist in only a few laboratories, are very large (about 15 m³), usually emit in the far infrared (10.6 μ m), and are far from inexpensive. Incoherent sources in the range of 100 - 1000 W are available (xenon-mercury (*Xe-Hg*) gas discharge lamps, for example), but this type of source is typically noisy (exhibits large intensity fluctuations), difficult to collimate, and characterized by a very short lifetime (from the systems perspective). In addition, gas discharge lamps are broadband sources, and as such usually require wavelength filtering in order to provide compatibility with wavelength sensitive devices such as volume holographic optical elements and spatial light modulators. A broadband source that has been suitably filtered to allow readout of a typical volume holographic optical element (within the allowable spectral bandwidth of the stored diffraction gratings) might generate only about 10^{-5} - 10^{-6} of its total rated power in the wavelength region of interest. For the 1000 W *Xe-Hg* lamp, this results in only about 1 - 10 mW of quasi-monochromatic optical power.

Coherent, monochromatic optical power can be provided by an array of different types of laser sources [Milonni, 1988], including the argon-ion (Ar^+) laser, the neodymium-YAG ($Nd-YAG$) laser, the helium neon ($He-Ne$) laser, the helium cadmium ($He-Cd$) laser, dye lasers, excimer lasers, and semiconductor laser diodes. Typical monochromatic (single laser line) power outputs from the first two types range from about 500 mW to 25 W . Helium neon and helium cadmium lasers are readily available as well as relatively inexpensive, but have output powers that are typically in the range 1 - 5 mW , peaking out at about 50 mW . Dye lasers are often optically pumped by argon-ion lasers, and hence exhibit power outputs slightly lower than that of the pump laser. Excimer lasers are typically operated in the pulsed mode of operation at repetition rates of 10 - 1000 pulses per second, and emit average powers in the 10 - 100 W range. Finally, semiconductor laser diodes are available with very long lifetimes at output powers of 1 - 20 mW , and much shorter lifetimes in the 100 mW - 1 W range.

Of these six different types of coherent sources, the first five are still relatively bulky (about 0.1 m^3), consume considerable electrical power, generate significant amounts of heat (many must be water cooled to ensure stable operation and practical lifetimes), and are very expensive (especially when compared with a comparable electronic power supply!). Although these sources can be (and indeed are) employed in current systems-level demonstrations, their collective liabilities do not augur well for their eventual incorporation in commercially viable computational systems in general, and perhaps neural network applications in particular. This leaves the last category, that of semiconductor diode lasers (including, possibly, miniaturized diode-pumped $Nd-YAG$ lasers), for further consideration.

Before discussing the properties of semiconductor diode lasers as optical power sources any further, we should at least note that the range of output powers available from these sources (1 - 100 mW for single element devices) is rather limited. Taking an upper bound (with continued research and development) of about a watt per device gives us a realistic estimate of the amount of average coherent source power available for at least circuit level implementation of photonic neural networks, though certainly at the systems level phased arrays of stripe laser diodes and/or multiple sources could conceivably be employed.

Semiconductor diode lasers [Kressel, 1977; Casey, 1978a; Casey, 1978b] have been ex-

tensively investigated and developed over the past two decades for a broad range of commercial applications, including compact disk player recording and readout, fiber optical communications systems [Jones, 1988], merchandise optical scanners, and laser printers. The physical size of these lasers is small enough (about $0.3 \times 1 \times 5 \text{ mm}$) to fit in a standard transistor (or IC) package, as long as external cooling is not required. Lasers with power outputs of 1 - 10 mW are relatively inexpensive, costing a few tens of dollars in quantity on the average. Higher output power lasers are considerably more expensive, however, as are lasers with very narrow spectral linewidths (so-called *single longitudinal mode* lasers). For the higher power lasers (as well as for the intermediate power lasers that are required to maintain a high degree of center wavelength accuracy), external cooling (*e.g.* by means of a thermoelectric cooler) must be provided in order to maintain thermal stability in both wavelength and output power.

The wavelength ranges spanned by semiconductor diode lasers are dictated by the direct bandgap materials used to fabricate the coherent light-emitting diode (semiconductor *p-n* junction). Aluminum gallium arsenide/gallium arsenide (*AlGaAs/GaAs*) lasers grown on single crystal gallium arsenide substrates emit at wavelengths in the range 780 to 900 nm, while lasers based in the quaternary indium gallium arsenide phosphide (*In-GaAsP*) compound semiconductor system (and grown on indium phosphide substrates) emit at wavelengths further into the infrared (1.2 to 1.6 μm). The aluminum gallium arsenide/gallium arsenide lasers in particular are nearly wavelength matched to the peak sensitivity of both silicon and gallium arsenide photodetectors, as might be employed for photonic switching in spatial light modulator arrays, or for detection of computed results in a system diagnostic or output plane.

Within these ranges, a typical multimode semiconductor diode laser has a spectral bandwidth of 0.5 - 2 nm; a single longitudinal mode laser has a much narrower spectral bandwidth of order 10^{-4} nm (about 50 MHz centered at an optical frequency of $3.5 \times 10^{14} \text{ Hz}$). Both multimode and single longitudinal mode diode lasers can be used to write and read holographic optical interconnection elements, as long as the coherence length of the laser is larger than the thickness of the holographic recording medium. The coherence length of a laser is essentially the maximum path difference over which two

beams derived from the same laser can maintain the stable phase relationship necessary to exhibit an interference pattern. In applications requiring high multiplexing capacity within the holographic interconnection medium (or significant path differences among beams that must coherently interfere), the narrower linewidths of the single longitudinal lasers are often preferable since their coherence lengths are several orders of magnitude longer. For example, typical multimode semiconductor diode lasers operated above threshold exhibit coherence lengths in the range 0.1 - 10 mm, while stabilized single longitudinal mode diode lasers can have coherence lengths exceeding 1 m.

Employing a single, high intensity optical power source in a typical neural network application carries with it a potential penalty: an inherent tradeoff between energy efficiency on the one hand, and the need for array generation optics on the other. This tradeoff arises from the fact that most optoelectronic implementations of neuron unit arrays have either photodetectors or modulation windows (in some cases both) that are smaller in size than each individual pixel, as was shown schematically in Figures 15.4 and 15.7. The ratio of the area of a given photosensitive element to the entire pixel area is referred to as the *fill factor* of the pixel (with respect to that particular element). Typical fill factors for the photodetectors and modulation windows may range from less than 0.1 in the case of monolithic integration to about 0.5 for hybrid integrated devices. Light that falls outside the appropriate areas within a given pixel will at best contribute to the overall system throughput loss, and at worst may adversely affect the function of adjacent devices that exhibit photosensitivity.

In order to efficiently channel the optical illumination to the correct photosensitive regions, we need to (a) *expand* the source illumination uniformly to fill the entire aperture of the device in question (a spatial light modulator or volume holographic optical element, for example), (b) in many cases *collimate* (or re-collimate) the light source to produce a planar wavefront with a beam of constant width, (c) *spatially filter* the beam to enhance its uniformity by eliminating significant fixed-pattern noise, (d) *focus* the light within each individual pixel to a size compatible with the relevant photosensitive area (in effect thereby generating a two-dimensional array of focused beamlets), and (e) *align* the resulting array of focused beamlets with each succeeding device in the optical path.

The procedures and optical elements required for beam expansion, collimation, and spatial filtering are well understood among the optical community for the case in which the source beam is initially *axially symmetric*, as is typical of gas and excimer laser systems. In typical semiconductor laser diodes, however, the planar nature of the light-emitting heterojunction region often gives rise to a diffraction-induced beam divergence *parallel* to the junction of 3 - 10 degrees, and a corresponding beam divergence *perpendicular* to the junction of 20 - 60 degrees. Comparable procedures and optical elements for such *anamorphic* (non-axially symmetric) beams are more complex, and are currently under development. Also under development are a number of types of semiconductor diode lasers that emit approximately axially symmetric beams suitable for standard collimation and filtering systems.

The optical source array generation problem has received considerable attention recently, due primarily to significant interest in optical interconnection systems. In one promising approach, a two-dimensional array of computer generated and photographically reduced amplitude-encoded Fresnel zone plates has been used to form an 8×8 grid of microlenses that function by means of *diffraction* (from what is, practically speaking, a computer generated hologram (CGH)) rather than *refraction* [Marrakchi, 1990]. In another well-developed approach, computer generated binary phase holograms (so-called *Dammann* gratings [Dammann, 1971]) have been configured to form large grid patterns of regularly spaced illuminated spots with predetermined locations and fill factors [Morrison, 1989]. Using this latter technique, 32×32 arrays have been generated with both high throughput efficiencies and low scattered light by crossing two fabricated 1×32 grating arrays. In addition, an 81×81 array has been experimentally demonstrated by using two pairs of crossed 1×9 grating arrays in an optical arrangement that generates multiple images by means of a convolution operation [McCormick, 1989]. In both of these techniques, all of the resulting light beamlets are mutually coherent, as they derive from the same source. This mutual coherence has an impact on the utilization of such source arrays for the generation of independent holographic interconnection networks, as described in the subsection on "Photonic Interconnections" above.

An interesting alternative to the generation of pixelated optical sources by modification

of the properties of a *single* source is that of direct fabrication of *multiple source arrays*. One striking example is the recent successful fabrication of over one million independent surface-emitting semiconductor diode lasers on a single gallium arsenide chip [Jewell, 1990]. Both cylindrical and square cross-section microlasers have been fabricated with diameters and edge dimensions in the range $1 - 5 \mu m$, with heights above the surface of the wafer of about $5.5 \mu m$ as shown schematically in Figure 15.18. In the fabrication process employed, the laser mirrors are arranged to generate laser emission *through* the $500 \mu m$ thick gallium arsenide substrate, as shown in the Figure. In order to accomplish this without significant absorption in the substrate, the active (lasing) medium is composed of *InGaAs* quantum wells with *GaAs* barriers, giving rise to an emitted infrared wavelength ($\approx 950 nm$) that lies in a region of substrate transparency.

FIGURE 15.18 Illustration of a surface-emitting laser diode source array [after Jewell, 1990]. In this example, the individual semiconductor laser diodes are isolated by chemically assisted ion beam etching techniques, must be individually contacted, and emit *through* the *GaAs* substrate.

In the present configuration, the lasers are essentially optically isolated, and hence are not designed to be mutually coherent (phase locked). In fact, over time constants typical of holographic recording in currently available photorefractive crystals (milliseconds), it is likely that such arrays are for all practical purposes *mutually incoherent*, due both to the optical isolation as well as to process-induced variations in device parameters that alter the wavelength emitted from each individual laser. Arrays of surface-emitting semiconductor lasers that have been specifically *designed* to have uniformly separated wavelengths have also been demonstrated [Chang-Hasnain, 1990]. We shall return to this characteristic in a succeeding section that addresses a particular strategy for photonic neural network implementation.

At present, each laser within the array operates at a threshold voltage of about 10

volts at a threshold current of a few milliamperes, resulting in a power dissipation of 10 - 50 milliwatts per device at threshold, and higher for power outputs significantly above threshold. In order to keep the overall power density within established limits ($1 - 10 \text{ W/cm}^2$) and thus to keep from overheating the substrate (resulting in potentially deleterious effects on wavelength stability and/or catastrophic failure), the lasers must either be spaced appropriately, operated in a pulsed (on/off) mode at less than unity duty cycle, or temporally multiplexed (turning on only a few lasers at a time) by resorting to individual rather than parallel addressing. Given the current rate of progress in the development of these and other types of surface-emitting laser arrays, it is reasonable to expect demonstration of continuous operation of up to 10^4 microlasers per square centimeter within the near future.

It should be noted that the array shown in Figure 15.18 is not currently configured for parallel operation of all of the sources simultaneously, which would require electrical contact to the tops of each selectively etched microcavity. This feature could likely be provided by an additional surface passivation and metallization step. Matrix-addressable surface-emitting laser arrays have recently been fabricated by forming columns of lasers separated by etched isolation grooves, and interconnected across the grooves by striped row contacts [Orenstein, 1990a]. Application of an appropriate bias voltage across a given pair of electrodes (column and row) activates the laser diode at the intersection, allowing for raster-scanned operational modes as well as fully parallel operation [Von Lehmen, 1990].

Other currently investigated approaches to surface-emitting laser array fabrication use various techniques to form the microlaser cavities *within* the planar substrate without the need for deep etched isolation grooves, such as by the use of ion implantation to form electrically insulating isolation layers between the laser cavities [Tai, 1989a; Orenstein, 1990b] or by the current confinement that results from photolithographic definition of one of the two laser mirrors and its associated electrical contact [Tai, 1989b]. Fabrication processes that yield planar or quasi-planar device structures allow for direct parallel contact if desired without the complications of depositing contacts on vertical sidewalls.

Before leaving the subject of semiconductor laser diodes and surface-emitting laser arrays, it is worthwhile to note a very useful feature of such devices: their capacity for *high*

bandwidth direct modulation. By this we mean that the output intensity of the semiconductor laser source can be modulated (at full modulation depth, *i.e.* from well below the threshold for lasing to peak output power) at frequencies up to a few gigahertz by direct variation of the voltage applied across the device. This attribute can be used to advantage in many neuro-optical implementation architectures by eliminating the need for mechanical or electrooptical shutters, as well as by offering temporal multiplexing as an additional degree of freedom for the systems designer.

One additional type of solid state device that is capable of both single source and source array fabrication is the light emitting diode (LED). Closely related to the semiconductor laser diode, the LED is also a *p-n* junction device that can be fabricated with considerably less processing complexity by elimination of the high reflectivity mirrors that form the semiconductor laser cavity. An additional advantage is the lack of a threshold for operation, allowing the LED to emit over a much wider dynamic range of applied voltages. One drawback of light emitting diodes is that they are relatively broadband (incoherent) sources, and as such are not usable as sources for holographic recording applications (and in many cases for readout of multiplexed holographic optical elements as well). In addition, they are relatively inefficient emitters with typical electrical-to-optical conversion efficiencies of a few percent. This feature tends to make LEDs rather power consumptive for a given amount of usable output intensity.

Detectors and Detector Arrays

Detectors are optoelectronic components that act as photon-to-electron converters, in that they transform incident optical intensity into electronic form, usually a voltage or a current. Detectors therefore allow the optical representation of neuron unit outputs, for example, to be converted into an electronic representation for further processing. As such, they are important components for the photonic implementation of neural networks in at least two functional areas: (a) as input transducers for the necessary optical detection function of optically addressed spatial light modulators, and (b) as output transducers

for the translation of optically generated intermediate and final results to an appropriate electronic format. After all, once you've gone to all of the trouble of learning and computing with a neural network, it might prove worthwhile occasionally to actually get the answer out and use it to initiate some other useful process!

In both of these functional areas, we can further categorize detectors as (a) single pixel detectors, and (b) interconnected detector arrays. In the first category, we include both single element detectors that have one optical input aperture and one output channel, as well as the single pixel detectors employed as part of an array in two-dimensional spatial light modulators. This latter assignment is made because even though detectors used in spatial light modulators are perhaps *configured* in an array, their outputs are used only within one or at most a few local pixels. In the second category, we include arrays of detector elements that are interconnected in such a way that the *entire* parallel (one- or two- dimensional) array can be read out electronically through one or more output channels. An example of a detector array in this category might be the light sensitive element in the CCD (charge-coupled-device) camera, now commonplace in many solid state cameras and video cassette recorders.

This distinction between single pixel detectors and detector arrays is important because the technologies that are commonly employed in these two cases differ in a number of respects, and as a result can exhibit wide differences in performance characteristics such as bandwidth, sensitivity, linearity, and dynamic range. In the case of single pixel detectors, for example, it proves easier to jointly optimize performance parameters because of the larger number of degrees of freedom available to the device designer in a single input, single output system. The detector array designer, on the other hand, often must make additional tradeoffs dictated by the nature of the charge storage and readout process employed over the full set of integrated pixels.

In the context of photonic neural network implementations, single pixel detectors have two primary functions. The first is to act as optical signal to electronic signal converters within optically addressed spatial light modulators, to translate a pixel's worth of incident light intensity (representing, for example, the weighted sum of signals from the output of a plane of neuron units) into a voltage or current. The resultant electronic signal can then

be processed by local intrapixel circuitry to produce the desired neural threshold function for subsequent optical encoding (modulation). This process could be accomplished either onboard a monolithic or hybrid integrated optically addressed spatial light modulator (OASLM), or on a separate detector chip that interfaces with an electrically addressed spatial light modulator (EASLM). In this latter case, the detector will most likely fall under the *detector array* category discussed further below, since a parallel-to-serial conversion is typically required to extract the array of data (e.g., an image) from the detector chip (followed by a serial-to-parallel conversion to load the signal into the EASLM). It should be noted that even in the case of monolithic spatial light modulators that do not feature discrete detectors, electronic control circuitry, and modulators, converting a two-dimensional optical input distribution into a modified two-dimensional output distribution *necessarily* involves a local detection function, even if it is not particularly easy to separate the detection process from the modulation process.

The second important single pixel detector function is to provide for single point monitoring functions within the system, such as the output power from a given laser source, the average power emitted from a laser source array, or a particular system output that activates a desired process or function (for example, the identification of a specific defect pattern on a manufactured part within the input image field of a neural image processor, that in turn results in rejection of the part).

Perhaps the simplest type of detection element that can be incorporated in a single pixel is the *photoconductor*, which typically consists of a thin film of material that alters its resistance to electrical current in response to the intensity of incident illumination. The most commonly used single pixel photodetectors, however, are based in some way or other on the *semiconductor p-n junction diode*. Under reverse bias in a *p-n* junction diode, photocarriers created by light absorbed within the region of the junction between *n*-type and *p*-type semiconductor layers are swept *out* of the junction region by the internal electric field across the junction, and collected in the external circuit. If the internal electric field is high enough, each photocarrier can acquire enough energy during sweepout to generate an avalanche of additional carriers, leading to significant gain in the class of so-called *avalanche photodiodes*.

The inclusion of an intrinsic (undoped or compensated) layer of semiconductor material between the n -type and p -type layers allows for a significant reduction in the junction capacitance of the device, with a corresponding improvement in signal bandwidth. Such devices are commonly referred to as $p-i-n$ photodiodes, packaged versions of which are commercially available for a wide variety of photosensor functions. Typical $p-i-n$ photodiodes exhibit risetimes of a few nanoseconds, are linear in output over seven orders of magnitude of input intensity, and are sensitive to very low light intensity levels. For silicon $p-i-n$ photodiodes, sensitivities of about 0.4 milliamperes of output current per milliwatt of optical input power at a wavelength of 830 nm are common, which represents a conversion efficiency from photons to electrons of approximately 60%.

Phototransistors are light sensitive devices that exhibit current gain in exactly the same manner as a transistor, with the exception that the controlling base current is injected *optically* rather than through the base lead. In fact, most VLSI transistors (both bipolar and MOS) are photosensitive (though perhaps not optimized for the photodetector role), and must be protected from stray light in order not to compromise their performance characteristics. The principal advantage of a phototransistor is its inherent current gain of order 100 to 1000, which often makes the interface of the photodetector to following circuitry more straightforward. In cases requiring exceptionally high gain in the front (photodetection) end, two transistors can be paired as shown in Figure 15.19 so that one acts as a phototransistor, and the other as a current amplifier. Such a two transistor combination has achieved widespread use, and is referred to as a *photo-Darlington pair* [Sze, 1981b]. The tradeoffs for increased gain in both of these cases are risetime (which translates directly into signal bandwidth) and area required for integration. Typical risetimes for phototransistors are almost three orders of magnitude higher (a few microseconds) than those characteristic of $p-i-n$ photodiodes. Photo-Darlingtons are yet another factor of ten or so slower in response time. Optimized phototransistors and photo-Darlingtons require relatively large collector-base junctions in order to provide an appropriately sized photosensitive region that can be accessed by optical imaging techniques.

FIGURE 15.19 Schematic diagram of a photodarlington pair utilized as a high gain detector/amplifier combination.

In many if not most cases, the type of photodetector chosen for use as a single pixel detector in a spatial light modulator application depends on its integrability with associated control electronics and modulation elements. This, in turn, depends on whether the particular spatial light modulator in question is monolithically or hybrid integrated, as discussed in the section on photonic switching above, and on which semiconductor substrate the photodetection element itself is to be fabricated. In some cases, the desire for integration of a high density of neuron units may place strict bounds on the area allocated to each separate function in general, and on the photodetection and requisite amplification function in particular.

In traditional applications of photodetector technology, for example in spectroscopy and optical metrology, linearity of response (output voltage or current as a function of the input intensity) is prized, as is a wide dynamic range over which linearity is assured. In neural network applications, however, linearity is typically less of an issue. In fact, it is often convenient to use the inherent nonlinearity of the input-output characteristic of a particular photodetector device to generate part or all of the nonlinearity required of the overall neural unit function. This can result in a lower overall expenditure of real estate for each neuron unit, increasing the neuron array density, as well as in a reduction of circuit complexity within each pixel. One such example is the output current saturation characteristic of phototransistors at high input intensities, which can be used to emulate the upper saturation regime of the sigmoidal neuron response function.

Detector arrays are employed whenever the intensity distribution of a one- or two-dimensional image field requires conversion to electronic form for interface with succeeding computational or output stages of the system. In a very real sense, a two-dimensional detector array is nothing more than the business end of an optoelectronic camera that can be positioned anywhere within the optical system that the local intensity distribution represents a desired result. In fact, low reflectivity beamsplitters can be used to merely

"sample" the local intensity distribution of a given beam of light, allowing most of the incident light to propagate in a further computational arm of the optical train for use elsewhere.

Detector arrays are inherently different in at least one key respect from the single pixel photodetectors (as well as arrays of photodetectors used in optically addressed spatial light modulators) discussed previously: the need to provide for some form of output channel multiplexing, in order to avoid the requirement for a one-to-one correspondence between pixels in the array and output pins. For example, in a 1000×1000 element detector array, fully parallel readout requires one *million* output channels or pinouts. As a result, detector arrays are usually configured to perform some form of parallel-to-serial conversion of the data into a single high bandwidth serial channel prior to the readout of each frame (though multiple output channels can also be used). This can either add significant circuit complexity to the area surrounding each pixel in order to accommodate for the parallel-to-serial conversion and interpixel communication function, or be directly incorporated into the design of the photodetection elements themselves, as in the case of the CCD arrays discussed below.

The state of the art of detector arrays has advanced tremendously even over the past decade, to the point where solid state detector arrays with quite spectacular performance are used everywhere from astronomical applications (as detectors for even the largest telescopes), to earth observation satellites (infrared focal plane arrays), to photomicroscopy (in place of the traditional film-based photographic camera), to consumer products (electronic still photography and video cameras).

One of the most successful and generally available types of solid state detector array is the *charge-coupled-device* (or CCD) array [Sze, 1981c; *Optical Engineering*, 1987a; *Optical Engineering*, 1987b]. In this technology, usually based on MOS fabrication techniques in silicon (but adaptable to compound semiconductor substrates as well), incident illumination within a given pixel causes the accumulation of photogenerated charge in an electrostatic potential well formed by the application of bias voltages on a set of electrodes, as shown schematically in Figure 15.20. In operation, the CCD array is illuminated for a given exposure time (slightly less than one full frame interval), during which time the

charge generated by the incident illumination is integrated within each primary well. Subsequently, appropriate voltages are applied by means of multi-phased electrode structures to *spill* the accumulated charge packet into the neighboring well, while simultaneously moving the charges in the neighboring well to *its* neighboring well, and so on throughout the array.

FIGURE 15.20 Schematic diagram of a charge coupled device (CCD) photodetector array fabricated on a silicon substrate. Electrostatic potential wells are created by application of appropriate voltages to the three phase bias electrode structure, with electrical isolation provided by the gate oxide layer. Light incident through the transparent electrodes creates stored charge that can be transferred to an output signal terminal by proper sequential phasing of the bias voltages ($P_1 - P_3$).

The overall operation resembles the function of an array of one-dimensional shift registers. At one edge of the structure, the charge packets from each row are collected into a single column that is read out by a very high speed shift register (a linear, usually buried channel CCD array) to form the single output channel. Full readout of the array must occur before the next frame is exposed (except in specifically designed cases such as the time-delay-and-integrate or TDI mode of operation, in which only *one* shift is interposed between successive exposures).

One-dimensional arrays of CCD elements have been successfully fabricated in sizes of 1×2048 , while special purpose two-dimensional CCD imaging arrays 2048×2048 in size are commercially available [Blouke, 1987]. This represents a parallel detector with 4,194,304 individual pixels! In one particular 2048×2048 CCD array, the imaging area is 5.5×5.5 centimeters, with a pixel size of 27×27 microns. This array exhibited a dark (unilluminated) noise buildup in each pixel of only 6 to 12 electrons when read out at a rate of 50,000 pixels per second, which allows for detection of extremely low level signals

with excellent signal-to-noise ratio. Given a well capacity of about 700,000 electrons, this very low noise figure suggests a dynamic range in excess of 70,000, nearly five orders of magnitude! For well charge densities less than 200,000 per pixel, the linearity is better than 0.5% over this portion of the full dynamic range. Finally, this array exhibited an extraordinarily high charge transfer efficiency of 0.999992, representing the fraction of charge within a given pixel that is routinely transferred to an adjacent pixel without loss.

The integration of large scale detector arrays by means of VLSI techniques provides the prospect of special purpose arrays that perform part of the computational function *within* the confines of the array. One example of such special purpose chips is the incorporation in a CCD array of charge-coupled analog circuitry to perform arithmetic operations such as addition, subtraction, and magnitude comparison [Fossum, 1987]. Such an array could allow for detection of parallel differential outputs, with both positive (excitatory) and negative (inhibitory) weighted sums as dual optical inputs in a (positive definite) intensity representation.

ARCHITECTURAL CONSIDERATIONS FOR PHOTONIC NEURAL NETWORK IMPLEMENTATIONS

We now turn our attention to the *use* of the photonic components and fundamental principles described above in the implementation of highly parallel neural network *architectures*. The focus in this section is on a general framework that emphasizes characteristics common to different approaches to photonic and optical neural network implementations, as well as on illuminating some of the key fundamental differences among the various implementation approaches. A review of recent and ongoing research in photonic and optical neural network implementations is beyond the scope of this chapter; sources of such information can be found in the Suggested Further Reading section at the end of this chapter.

Photonic neural network implementations can be adaptive or non-adaptive, can represent the signal using different physical quantities, and can be built using one-dimensional

(1-D) or two-dimensional (2-D) arrays of neuron units with two-dimensional or three-dimensional (3-D) interconnection elements. These issues, in addition to other features that are desirable in any photonic implementation of a neural network, are discussed in this section. Throughout, one should keep in mind the distinctions that exist among systems with fixed interconnections, programmable systems, and truly adaptive systems. We will initially concentrate on the implementation of a single layer of a network, and subsequently show how this generalizes to multiple layers.

The computation process of any one layer of a neural network can be represented by:

$$y_i = f \left[\sum_j w_{ij} x_j \right] \quad (23)$$

in which neuron unit j is situated at the input to the layer of interconnections, neuron unit i is situated at the output of the layer of interconnections, y_i is the output of neuron unit i , x_j is the output of neuron unit j , w_{ij} is the weight associated with the interconnection between them, and the function f represents the neuron unit nonlinearity. The term inside the brackets, the activation potential, will be denoted by ρ_i . Note that the term in brackets is a matrix-vector product between an interconnection weight matrix and an input vector. The function f then operates independently on each element of the resulting vector; this is called a *point nonlinearity*, and as such lends itself to implementation with a spatial light modulator (SLM).

Most current learning algorithms fall into one of a small number of classes. For example, one such class can be specified by:

$$\Delta w_{ij} = \alpha \delta_i x_j - \beta w_{ij} \quad (24)$$

in which $\Delta w_{ij} = w_{ij}(k+1) - w_{ij}(k)$ is the weight update, k represents the iteration index, α is the learning gain constant, and β is a decay constant that is included primarily for hardware convenience; β can be set to 0 when so desired. Suitable choices of δ_i give different learning algorithms, such as Hebbian, Widrow-Hoff, single-layer least minimum squares (LMS), and for the case of multilayer networks, backward error propagation. (For

example, an optical architecture that potentially implements backward error propagation in a multilayer neural network is described by Wagner and Psaltis [Wagner, 1987]). In this chapter we will restrict our attention to this particular class of algorithms for illustrative purposes. Although other classes of learning algorithms can likely also be implemented using photonic hardware, research to date has focused primarily on the class represented by Equation (24). An important aspect of Equation (24) for implementation is the outer product between the training vector δ and the input vector $\mathbf{x}$ for the weight matrix update.

An example of a photonic neural system is shown in block diagram form in Figure 15.21. This system utilizes a 1-D array of neuron units at the input and output, and a 2-D interconnection mask. Each pixel in the input is expanded optically (using cylindrical lenses) and illuminates the corresponding row of the interconnection mask. The mask stores the analog weights, and provides a pointwise multiplication before the beam is contracted so that one column from the mask is incident onto one corresponding output pixel. The optical system in effect provides a fully parallel analog optical matrix-vector multiplication as represented by the bracketed term in Equation (23), performed over all i . Threshold functions and feedback connections are provided by means of either photonics or electronics. The first experimental demonstration of such a system applied to neural network implementations used an array of light emitting diodes (LEDs) as inputs to, and a linear detector array as the output from, the optical interconnection [Psaltis, 1985; Farhat, 1985]. This particular system utilized electronics to provide the threshold functions and feedback connections.

FIGURE 15.21 Block diagram of a 1-D to 1-D photonic neural network, in which a one-dimensional neuron array is fully interconnected to a one-dimensional detector array by means of a two-dimensional interconnection mask.

It should be noted that many variants of Figure 15.21 are possible; some of them are

more compact than others, though all of them share essentially the same basic characteristics. The interconnection mask can be fixed (*e.g.*, photographic film) or variable (*e.g.*, an SLM). In the latter case the SLM can be electronically or optically addressed. Electronic addressing is appropriate for straightforward interfacing to an electronic machine that supplies the (updated) interconnection weights, whereas for a maximum adaptation rate an optical addressing technique would ultimately be optimal. Currently available SLMs with large numbers of pixels tend to be slow (500×500 analog pixels with 1 - 100 ms frame times) [Tanguay, 1985]; much faster technologies are being developed for future use [see, for example, Lentine, 1988; Lentine, 1991; McCormick, 1989b]. Such a system, with 1-D inputs, 1-D outputs, and 2-D interconnections, will likely scale up to 100 - 1000 fully connected neuron units.

A photonic system that can implement larger numbers of neuron units and interconnections is shown in Figure 15.22. All neuron unit planes are now 2-D arrays, and the interconnection medium is a 3-D structure, implemented in a volume holographic material. In effect, there is a separate volume grating connecting each input neuron unit j to each output neuron unit i . The diffraction efficiency of each grating is proportional to the weight, w_{ij} , of the corresponding interconnection. Note that each such grating is analogous to a beamsplitter, as discussed in the previous section, with the primary difference that the volume gratings are direction (and wavelength) selective. Thus, beams incident on such a "beamsplitter" at other than the correct angle are not affected by the presence of the holographic beamsplitter. Properly recorded, then, the grating w_{ij} is situated in angular orientation and grating period so that it affects only the inputs at the angle corresponding to x_j , and will direct the corresponding output $w_{ij}x_j$ to the correct summation node ρ_i . The achievable numbers of neuron units and interconnections are currently subjects of considerable debate, but would likely be 10^4 - 10^6 neuron units per plane and on the order of 10^{10} independent interconnections with weights, assuming continued research unveils no impassable boundaries.

FIGURE 15.22 Block diagram of a 2-D to 2-D photonic neural network.

in which a two-dimensional neuron array is fully interconnected to a two-dimensional output array by means of a three-dimensional volume holographic optical interconnection mask. The input plane, output plane, and optional training plane are shown. Many variants of this geometry with similar properties are possible.

For the case of an adaptive network, we use a variable (typically photorefractive) holographic material for recording and implementing the interconnections. To incorporate learning, a training plane comprising a 2-D array of nodes generates the δ_i terms (Figure 15.22). During a weight update, an exposure is made of the interference pattern between beams emanating from the two left hand planes in the figure. Each of the two left hand planes could be implemented using, for example, a 2-D spatial light modulator illuminated by an expanded beam. This results in a change in the refractive index modulation representing the current weight that is dependent on the product $\delta_i x_j$, so that with appropriate choices of parameters, the increment in diffraction efficiency can be made proportional to $\delta_i x_j$. Ideally, this records changes (updates) in the interconnection weights within the hologram given by Equation (24), above, in the form of gratings situated with appropriate angular orientation and grating period. It should be noted that generating and recording these weight updates is not a simple matter, and care must be taken to insure that the appropriate interference terms are recorded and that not too much crosstalk is inadvertently created. Recording and recall of the correct values is primarily a number representation issue and is discussed below; undesirable crosstalk depends on the recording and reconstruction technique as previously discussed in the subsection on "Photonic Interconnections".

An example of one source of holographically-induced interconnection crosstalk is an inadvertent degeneracy of gratings. Even though each volume grating affects only the beams incident at a particular angle with respect to the grating, it affects *all* of the beams at that particular angle. Because of this, an entire cone of beams (with its axis of symmetry aligned with the grating wave vector) can be affected by a single diffraction grating. This

degeneracy creates an undesired coupling between different interconnections in a fully connected network. For neuron unit sources on an ideal, rectangular grid, this coupling can be eliminated by removing neuron units from certain locations in the array, leaving sparsely distributed neuron units arranged in a degeneracy breaking pattern. This eliminates the undesired coupling, at the expense of a reduction in the number of neuron units from N^2 (for an $N \times N$ array) to $N^{1.5}$ [Psaltis, 1989].

The case of a non-adaptive network is likely to be an important one as well. In this case the interconnection hologram does not have to be recorded in accordance with a specific learning algorithm. If the weights are known *a priori*, then any applicable recording technique will suffice. In many cases, however, the weights may not be known. A common scenario may involve the training of a "master" network; once it has been trained, copies of the network could be produced in a production environment. If the network is large, and particularly if it utilizes volume holographic optical interconnections, then probing the values of all of the weights could be impractical. The most efficient production means in this case would be to make direct copies of the volume hologram. Thus, the capability of rapidly copying a multiplexed volume interconnection hologram is important.

The physical representation of the signal directly impacts the operation of a photonic neural network. The physical quantities available for optical representation of a signal level are field amplitude, phase, intensity, polarization, spatial position or frequency, and wavelength. We will consider only the most likely candidates: field *amplitude* (with phase) and *intensity*. For the case of an amplitude (with phase) representation, the signals may in general be complex valued; bipolar signals, of course, represent a subset of these numbers, and thus can be represented. Given that x and y are represented as (electric or magnetic) field amplitudes, the resulting detected activation potential of neuron unit i , ρ_i , is given by

$$\rho_i^{(\text{coh})} = \left| \sum_j w_{ij} x_j \right|^2 \quad (25)$$

for the case of a coherent sum, and by

$$\rho_i^{(\text{incoh})} = \sum_j |w_{ij}x_j|^2 \quad (26)$$

for the case of an incoherent sum (*c.f.* the preceding section on "Fundamental Principles of Photonic Technology"). In both Equations (25) and (26), the weight w_{ij} is represented physically by the *amplitude* diffraction efficiency. The coherent sum given by Equation (25) has the advantage of allowing for the addition of both positive and negative numbers in computation of the neuron unit potential, as desired for the incorporation of both excitatory and inhibitory neuron unit inputs. Clearly, Equations (25) and (26) deviate from conventional neural network models. The effects on different neural network models of such deviations in the summation before thresholding are not currently well understood.

If we instead encode the signal levels as intensities, the activation potential becomes

$$\rho_i^{(\text{int})} = \sum_j w_{ij}x_j, \quad (27)$$

which is the desired activation potential, but at the expense of all terms in the summation being nonnegative. In this case the weight w_{ij} is represented physically by the *intensity* diffraction efficiency. A technique for effectively achieving bipolar signals in this case will be discussed in the section describing "An Implementation Strategy".

The signal representation used also impacts the nature of the weight updates. The physical weight updates can be derived using common models of photorefractive (or other) recording materials. Such a derivation requires a number of approximations and assumptions to be made regarding the chosen operational mode. By appropriate choice of the operational mode, the ideal weight update rule given by Equation (24) can be approximately obtained for both intensity representation and amplitude representation cases. The operational mode may not prove to be the same in each case, and may differ in such parameters as the size of the weight updates, the size of the existing weights before the update, and the exact characteristics of the holographic material used. The "second order" terms that deviate from the precise form of Equation (24) are also different in the two cases; the effect of such terms on learning algorithm performance is not well characterized or

understood, and is currently an active area of research.

So far we have discussed only a single interconnection layer with neuron units for inputs and outputs. If such a physical network includes feedback, it can be generalized to functionally implement an arbitrary multilayer feedforward or recurrent network. Figure 15.23 illustrates this principle, showing one *physical* layer of neuron units, one layer of interconnections from the neuron units to a set of fan-in nodes, and feedback from each fan-in node to the corresponding neuron unit. These neuron units can be conceptually divided into groups corresponding to different *functional* layers. Some of the physical interconnections then represent functionally feedforward connections (represented by solid lines and boxes in Figure 15.23), and some represent functionally lateral connections within a layer (represented by broken lines and boxes in Figure 15.23). Feedback connections to previous layers, and feedforward connections that bypass the next subsequent layer, can also be incorporated in a similar manner, but are not shown in the figure. This technique for implementing multilayer networks using a single physical layer has been discussed by Farhat for the case of 1-D neuron unit arrays interconnected by a 2-D mask, and used in the implementation of parallel optoelectronic simulated annealing [Farhat, 1987]. Thus any photonic (single physical layer) architectures discussed herein generalize to multilayer networks, provided that they have capability for arbitrary connections and feedback.

FIGURE 15.23 A single layer physical neural network with feedback, used to implement a multilayer recurrent functional network. The solid boxes indicate feedforward connections, and the broken boxes indicate lateral connections.

In summary, the desirable characteristics of a photonic implementation of neural networks include: (1) modularity, so that multiple "modules" can be cascaded; (2) capability for lateral, feedforward, and feedback interconnections, which can be achieved physically by use of a single layer network with feedback and arbitrary interconnection capability;

(3) analog, weighted connections with analog signals; (4) bipolar signals and weights; (5) scalability to large numbers of neuron units with high connectivity; (6) generality, so that different neuron models, network models, and learning algorithms can be implemented within the same basic technology; (7) compatibility of different components within a given architecture; and (8) overall feasibility of the proposed combination of algorithm, architecture, devices, and materials. In addition, the optical/photonic hardware would ideally incorporate the following features: (1) simultaneous, parallel updates of all interconnection weights at each iteration; (2) high optical throughput; (3) low interconnection crosstalk; and (4) flexible functionality for neuron unit response, so that different neuron models and learning algorithms can be accommodated.

AN IMPLEMENTATION STRATEGY

In this section a photonic technique for the implementation of neural networks is described that potentially satisfies the aforementioned desirable characteristics and features [Jenkins, 1990a; Asthana, 1990a; Jenkins, 1990b; Asthana, 1990b; Jenkins, 1990c]. This photonic neural network implementation technique utilizes optoelectronic spatial light modulators (SLMs) for the 2-D neuron unit and training term planes. Each neuron unit incorporates dual channel encoding to allow for the representation of bipolar input and output signals, and comprises two integrated detectors, two modulators and integrated electronics. The neuron unit input and output signals are represented in the optical system by intensity. The interconnections are based on a 3-D holographic material with a novel incoherent/coherent recording and reconstruction technique that permits simultaneous updates of all weights during each iteration. In addition, the interconnections utilize a unique double angular multiplexing arrangement to minimize interchannel crosstalk and throughput losses, in which each pixel of the object beam SLM is illuminated by a set of mutually incoherent beams, each at a different angle. This implementation technique is explained in the remainder of this section.

A key feature of this implementation strategy is the use of an array of individually

coherent sources that are mutually incoherent to generate an array of coherent beam pairs used for holographic recording and reconstruction in the interconnection network. Consider the problem of recording two holograms, object A recorded with reference beam x_j and object B recorded with reference beam $x_{j'}$, as shown in Figure 15.24(a). The objects A and B could each be a 2-D array of data. In order to write both holograms simultaneously, A and x_j originate from the same coherent source and are mutually coherent; similarly for B and $x_{j'}$. However, B and x_j originate from a different source than A and x_j , so that each pair is incoherent with respect to the other pair. In this way, there are no extra (crosstalk) holograms written, such as that between A and $x_{j'}$, or between x_j and $x_{j'}$. This technique can be used for more than two multiplexed holograms, in which case a separate source is assumed for each hologram written.

FIGURE 15.24 Incoherent/coherent technique for recording and reconstructing multiple holograms simultaneously, in which all solid lines represent mutually coherent beams, and all broken lines represent a separate set of mutually coherent beams: (a) recording; (b) reconstruction; and (c) holographic representation, in which each hologram represents the fanout from a given neuron unit.

During reconstruction, the holograms are illuminated by the same set of reference beams x_j and $x_{j'}$. This simultaneously reconstructs the arrays A and B (Figure 15.24(b)). If the arrays are in registry upon reconstruction, a pixel-by-pixel incoherent sum will be achieved in the output array. If we now consider each reference beam x_j to be the output of a neuron unit at the input to an interconnection layer, then each reconstructed hologram corresponds to the fan-out from one neuron unit, with a contribution to each pixel in the output array proportional to the weight of the corresponding interconnection. This is depicted in Figure 15.24(b) and (c), in which the two signals fanning in to a given neuron unit are derived from separate, mutually incoherent optical sources. Note that this

technique provides an incoherent sum for the potential of each neuron unit (Equation (26) or Equation (27), depending on the chosen representation), as desired.

Another critical as well as unique feature of the photonic architecture described herein is a "double angular multiplexing" technique in which one input node or pixel in the object beam path has multiple beams passing through it at different angles. Thus, a set of angularly multiplexed beams is introduced for each object beam node δ_i , as shown in Figure 15.25. A three-fold angularly multiplexed fan-in from x_1 , x_2 , and x_3 to yield neuron unit potential ρ_1 is depicted in this figure; solid lines represent mutually coherent beams (all dashed lines represent a mutually coherent set as well; similarly for mixed dashed lines). Note that this multiplexing technique eliminates the fan-in beam degeneracy characteristic of collinear geometries referred to above in the subsection "Photonic Interconnections". Thus, the ensuing cross-coupling terms are absent, and a much more accurate set of weights can be recorded and reconstructed at each iteration.

FIGURE 15.25 Doubly angularly multiplexed volume holographic optical interconnection, designed to circumvent the effects of beam degeneracy. The mutually incoherent input beams ($\{x_j\}$) are angularly multiplexed over j , as are the corresponding sets of output beams from the training plane ($\{\delta_i^{(j)}\}$) generated by the coherent sources S_j , to produce an angularly multiplexed fan-in at each summed output, thus yielding the neuron activation potentials $\{\rho_i\}$.

A photonic architecture for neural network implementation that utilizes these principles is shown in Figure 15.26, for the case of Hebbian learning ($\delta_i = y_i$). The components shown in the figure comprise one module; inputs and outputs refer to this particular module. Only feedforward connections are shown. The upper spatial light modulator, SLM_1 , generates the training terms δ_i that also represent neuron unit outputs in this case. The lower spatial light modulator, SLM_2 , is the array of input neuron units. An array of coherent but mutually incoherent sources is used to illuminate the system; they are provided

by a mutually incoherent laser diode array or by a coherent beam passing through an SLM that temporally modulates the phase of each pixel independently. (It can be shown that the latter method is equivalent to the former for the particular type of holographic recording and reconstruction used herein.) A volume holographic material stores the requisite weighted interconnections, and can implement either fixed or adaptive interconnections depending on the material used.

FIGURE 15.26 Photonic architecture for neural network implementation that incorporates a parallel source array, double angular multiplexing, and incoherent/coherent recording and reconstruction; the Hebbian case is depicted.

Both spatial light modulators in Figure 15.26 consist of an array of pixels, each of which comprises three elements: (1) two integrated detectors for input of positive and negative parts of the neuron unit activation potential, (2) integrated electronic circuitry to provide the neuron unit (sigmoid or hard threshold) nonlinearity, and (3) two hybrid or monolithically integrated modulators for separate optical readout of the positive and negative neuron unit outputs. The SLMs, as shown, are read out in transmission, and have detectors situated so as to receive optical inputs on the right face of the SLM.

In the learning phase, the shutter is open as shown schematically in Figure 15.27. Light from each source S_j is approximately collimated so that it illuminates the entire array on SLM_1 , at an angle dependent on the position of the j^{th} source. Thus, for an N by N array of sources, there are N^2 beams reading out the contents of SLM_1 simultaneously, each at a different angle; the entire array of terms $\{y_i\}$ is encoded onto each of these beams. Each such beam then interferes only with its corresponding reference beam x_j , derived from the same source and encoded by SLM_2 , in the holographic medium. This writes the set of desired weight update terms $\alpha x_j y_i$.

FIGURE 15.27 Photonic architecture for neural network implementation: recording configuration. This configuration implements the learning function in the photonic architecture of Figure 15.26. The sets of beams emitted from the source array (two are shown) interfere in the volume holographic medium to update the weights stored in the interconnection holograms.

During the computation phase, the shutter is closed to prevent learning as shown schematically in Figure 15.28. The array of sources is imaged onto SLM_2 as a set of readout beams, so that each individual source corresponds to one pixel (neuron unit) on the SLM. The SLM modulates each beam so that the transmitted beam has an intensity proportional to the output value of the corresponding neuron unit. Thus, the j^{th} source illuminates the j^{th} pixel of this SLM, providing the signal x_j that becomes a reference beam to read out the j^{th} hologram. This hologram reconstructs an array of spots, similar to that depicted in Figure 15.24, that contribute to the input of each neuron unit in the output plane. The optics is set up so that this array is imaged onto the detector array. In the complete neural network architecture of Figure 15.26, additional optical elements (mirror M_2 , lens L_4 and beamsplitter BS_2) are used to displace the detector array plane to the detector side of SLM_1 , providing the neuron unit activation potentials. In addition, this beam is sent through beamsplitter BS_2 to a subsequent layer in the next module or to the output layer.

FIGURE 15.28 Photonic architecture for neural network implementation: reconstruction configuration. This configuration implements a single forward pass of the computing function in the photonic architecture of Figure 15.26. The lower set of beams acts as a set of reference beams, and generates a set of weighted output arrays that are imaged onto the detector array. Each stored hologram is reconstructed by a single neuron unit x_j , and fans out with appropriate weights to illuminate the detector array. The full set of reconstructed

holograms sums within each pixel to yield the neuron activation potentials $\{\rho_i\}$.

A generalized architecture that incorporates learning algorithms of the form of Equation (24) is shown in Figure 15.29. Instead of SLM_1 , as in Figure 15.26, this architecture utilizes a training term (δ_i) generator that is implemented via one or more optoelectronic SLMs. In general, target values t_i , actual neuron unit outputs y_i , or possibly activation potentials ρ_i may be provided as inputs to the training term generator. The physical arrangement of optical beams passing through the training term generator (from left to right) is the same as that shown passing through SLM_1 in Figure 15.26. Lateral and feedback connections can be incorporated by including an optical feedback path from the output of the hologram to the input side of SLM_2 .

FIGURE 15.29 Generalized photonic architecture for neural network implementation, including provision for the generation of arbitrary training terms (δ_i).

For many applications, both SLM_1 and SLM_2 can be fabricated using the same technology. Let's consider the case of a sigmoidal response with bipolar inputs and bipolar outputs. The electronics within each neuron unit can take the difference between the two detector inputs to yield the (bipolar) neuron potential. It can then perform the sigmoidal nonlinearity, and send the result to appropriate (positive channel or negative channel) modulator(s). For example, we have fabricated a number of silicon chips that integrate the necessary control electronics with appropriate detectors. One possible circuit that has been designed to incorporate the necessary functionality is shown schematically in Figure 15.30. Outputs from the two photodetection stages (V_{in1} and V_{in2}) are differentially amplified using two pairs of CMOS transistors ($M_1 - M_3$, $M_2 - M_4$), generating two separate and complementary outputs. The differential amplifier has been designed to saturate for

large values of the input signal difference, producing the upper asymptotic limit behavior characteristic of the sigmoid function. Each output signal is then inverted and clipped by another CMOS transistor pair ($M_{21} - M_{22}, M_{11} - M_{12}$), which asymmetrizes the transfer curve and adds the lower asymptotic limit of the sigmoid function. Finally, each output signal is inverted yet again and shifted in level by a dual transistor sub-unity gain amplifier stage ($M_{23} - M_{24}, M_{13} - M_{14}$), producing complementary output signals (V_{out1} and V_{out2}) that control the dual channel modulation elements. External provision is made in each pixel (neuron unit) for the adjustment of the voltage offset (zero crossing point) of each characteristic curve. This external bias adjustment allows for post-fabrication fine tuning of the overall response of the circuit, given process-induced variations in device characteristics.

FIGURE 15.30 Schematic diagram of a dual-input, dual-output differential amplifier that effects a sigmoid-like transfer characteristic.

A sample set of characteristic curves measured from one of these chips is shown in Figure 15.31, with the voltage on the second detector input channel as the parameter. These curves show the differential function of the dual channel circuit, as well as the desired sigmoidal response characteristic. A 6×6 array of these $100 \times 100 \mu m$ neuron units has also been fabricated with excellent uniformity.

FIGURE 15.31 Experimentally obtained transfer characteristics from the circuit shown in Figure 15.30, showing the output voltage in both channels (V_{out1} and V_{out2}) as a function of one input voltage (V_{in1}), with the other input voltage (V_{in2}) as a parameter.

The modular nature of this photonic neural net architecture can be inferred from Figure 15.29; the upper right SLM is SLM_2 of the subsequent module. Feedback paths from one module back to previous modules can be added, if desired, in a relatively straightforward manner. Bipolar signals are incorporated by the dual channel nature of the SLMs, with positive and negative channels for each neuron unit. Since each neuron unit has two physical outputs and two physical inputs, each interconnection between two neuron units physically consists of four separate weighted connections (positive modulator to positive detector, positive modulator to negative detector, *etc.*). Thus, even though each physical weight is nonnegative in value, their combination permits effective implementation of bipolar functional weights. In fact, the extra degrees of freedom provided by four independent weights can require revised weight update rules to ensure convergence of the learning process [Petrisor, 1990].

With the current spatial light modulator design at $100\ \mu m \times 100\ \mu m$ per neuron unit, 10^4 neuron units per cm^2 can be implemented on each SLM. By constructing an SLM as a mosaic of such arrays, a $3\ in \times 3\ in$ SLM could implement approximately 5×10^5 neuron units. Each "tile" or small array within such a mosaic need not be carefully aligned with respect to the other tiles, as the optical system just images the array back onto itself; in the case of multiple modules or lateral/feedback connections, there is, however, a requirement that all SLMs are similarly tiled, within an appropriate tolerance figure. Note that the current design utilizes only $2\ \mu m$ feature sizes in CMOS; this could eventually be scaled down by a factor of 4 in each dimension, yielding more than an order of magnitude increase in the number of neuron units implementable per unit chip area (or an equivalent reduction in the overall size with the same number of neuron units).

It should now be clear that this architecture can be generalized to implement certain other neural models. The use of electronic circuitry for the neuron unit function and training term generation provides significant inherent flexibility. For example, we have completed preliminary designs of units for forward and backward propagating signals in a backpropagation-style multilayer neural network. Although the optical weight updates in the holographic medium are restricted to outer-product terms (Equation (24) in the architecture as shown, variants of the architecture may permit other learning scenarios.

Finally, we consider the important question of making duplicates of a network that has already been trained. Since a volume hologram may store on the order of 10^{10} independent weighted interconnections, the preferred technique is to make direct copies of the multiplexed volume hologram. Here we describe a technique for copying such a multiplexed volume hologram in one step [Jenkins, 1990c]. To our knowledge this has never previously been achieved, but the use of incoherent/coherent holographic recording and reconstruction makes this in principle quite straightforward. Figure 15.32 shows an optical setup for duplicating the hologram. The master hologram is illuminated with the same set of reference beams as those employed during exposure; all of the mutually incoherent reference beams illuminate the master volume hologram simultaneously, recalling all of the stored holograms in parallel. The source array is imaged so that it generates an identical set of reference beams on the secondary (copy) holographic medium. Similarly, the reconstructed object beams are also imaged, so that they are incident on the secondary holographic medium, with amplitude and phase identical to that during recording of the master hologram. The appropriate pairs of beams interfere in the secondary holographic medium, making a complete copy of the original hologram. (As shown in Figure 15.32, the copy will actually be a spatially inverted version of the original. A slight variant of the optical system depicted in the figure can produce a copy that is identical to the original.) Thus it is conceivable to mass produce copies of a previously trained interconnection pattern, without ever knowing exactly what the interconnection weights are.

FIGURE 15.32 Optical layout for copying the entire contents of a three-dimensional volume holographic optical element (VHOE) into a second VHOE, utilizing a two-dimensional array of individually coherent, but mutually incoherent sources.

We conclude this section with a brief summary of the current implementation status of this particular photonic approach to neural network fabrication. For the neuron unit

arrays, 6×6 arrays of dual-channel detectors integrated with neuron function electronics have been fabricated in silicon and operate correctly. Individual multiple quantum well (*InGaAs/GaAs*) modulators have been successfully fabricated and tested, and exhibit drive voltages compatible with the electronics. The novel doubly angularly multiplexed incoherent/coherent interconnection technique has been tested experimentally at the level of two inputs/two outputs, and simulated at the level of four inputs/four outputs all with very favorable results [Jenkins, 1990c; Asthana, 1990b; Asthana, 1990c]. In addition, several learning algorithms that incorporate some of the unique features of the optical hardware have been successfully designed and simulated. Large 2-D arrays of laser diodes that are not mutually coherent have been fabricated recently [Jewell, 1990; Orenstein, 1990a; Von Lehmen, 1990]. Photorefractive crystals are routinely grown commercially, and can be purchased from vendors for use at visible as well as infrared wavelengths. In addition, the basic requisite features of the doubly angularly multiplexed incoherent/coherent holographic recording techniques have been demonstrated in single crystals of bismuth silicon oxide (*Bi₁₂SiO₂₀*), though not as yet at infrared wavelengths compatible with both the laser diode source array and the multiple quantum well spatial light modulators. All of the other components in the architecture (lenses, beamsplitters, *etc.*) are essentially available off the shelf.

As with any research project in progress, several questions pertaining to the photonic approach outlined herein remain partially unanswered. Consider, for example, the incoherent/coherent source array. Given the current state of the art of laser diode arrays, the total power dissipation will limit the number, maximum optical power, and spacing of the individual sources. Cross-coherence among the sources can cause undesirable crosstalk among corresponding interconnections, although in some neural network models a small to moderate degree of interconnection crosstalk is not likely to cause intolerable degradation in performance. Fortunately, a larger spacing of sources implies that each laser can output a higher power, and also assures a higher degree of mutual incoherence. Other remaining questions include the achievable contrast ratio and uniformity of the spatial light modulators; suitable monolithic or hybrid techniques for integrating detectors, electronics and modulators; optimization of the learning algorithm relative to the chosen holographic ma-

terial's storage and erasure time constants; and linearity and limitations of the hologram copying process. The next section discusses fundamental and technological limitations of the photonic hardware and their impact on the performance of photonic neural network architectures.

FUNDAMENTAL PHYSICAL AND TECHNOLOGICAL LIMITATIONS OF NEURO-OPTICAL COMPUTATION

Even though we are relatively early on in the development of viable neuro-optical computing systems, it is not too early to begin asking questions about the ultimate boundaries that may impact our future achievements. This line of inquiry can have a two-fold impact. First, discovery of inherently *fundamental physical limitations* that affect all forms of computation can, if correctly applied to the neural computational paradigm, both provide us with an ultimate goal worthy of achievement, and perhaps warn us in advance of architectural choices that will prove unworthy of technological implementation. Second, careful analysis of the *technological limitations* (device performance boundaries within a given technological implementation) that affect system performance can provide us with necessary guidance in choosing among many possible implementation strategies. The goal, of course, is to come up with the right combination of implementation strategy and technological choices to achieve the highest computational throughput (or perhaps learning rate) based on any one of a number of metrics. In this section, then, we discuss both the fundamental physical and technological limitations that impact the future performance of neuro-optical computational systems.

The Energy Metric

Your brain is truly a remarkable instrument from a computational point of view (as well as from many other points of view!). Although estimates (as well as individuals!) vary, it

is thought that your brain consists of about 10^{11} neurons, each interconnected (in certain regions of the brain) to $10^3 - 10^4$ other neurons [Changeux, 1985; Dowling, 1987; Hubel, 1979]. The human brain exhibits both short and long term memory, performs sophisticated image analysis in fractions of a second, operates as an effective associative memory *integrated over a whole lifetime of learning, and yet operates on a power budget that is only a fraction of the power dissipated by the average light bulb in your home* [Iversen, 1979]. In order to accomplish this, the active switching elements, the neurons, operate at an average power level about seven orders of magnitude lower than that characteristic of VLSI logic circuits [Mead, 1989b]. If this were not possible, it's likely that you'd be running a temperature even *without* the flu!

This discussion points to one of many possible metrics by which computational systems can be judged: energy (or power) dissipation. In fact, many modern supercomputers are limited in performance *precisely* because of power dissipation boundaries, or the ability to extract the heat generated by the computational process from the volume used to perform the work. We can perhaps think of computation as broken down into three fundamental parts: *representation of information, implementation of computational complexity, and detection of the results*. From the energy metric point of view, *everything* costs energy: what goes in costs energy, what comes out costs energy, and what goes on in between costs energy too. The trick in building the computational engines of the future (neural or otherwise) will be to maximize the overall performance with a minimum expenditure of energy.

Some Quantum Limitations

By *representation of information*, we mean the choice of data representation on which computations are performed. Some examples might include the binary representation, *M*-ary representations, an analog representation, or the residue representation [Huang, 1979]. This choice has implications at the fundamental level for the energy cost to represent a number within a given probability of error. For example, if we detect an optical signal bit that is binary encoded with a so-called "ideal" detector that can tell the difference

between receiving exactly zero photons and one or more photons, it only takes ten photons on the average to guarantee that the signal is received with a probability of error of one part in a billion, or a “bit error rate (BER)” of 10^{-9} . The average photon at a typical optical communications wavelength of 1300 nm has an energy of only 1.5×10^{-19} joules, so the total energy cost per bit is 1.5 attojoules (1.5×10^{-18} joules). For a communications channel operating at 10 gigabytes (8×10^9 bits) per second, this implies a power dissipation due to representation cost alone (without worrying yet about the *transmission* or *detection* of the information) of only 0.12 microwatts. For currently available detectors, about a thousand photons are required to achieve the same BER, so the necessary representation cost increases to 12 microwatts. In most currently envisioned communications systems, this cost is overwhelmed by other factors.

But what if we chose to represent numbers in an *analog* representation instead? If we were to follow the same kinds of quantum statistical rules, we would find that to represent the number “1000”, say, with an effective bit error rate of 10^{-9} requires about 150 million photons [Tanguay, 1988]. This is about 15 million times larger than the representation cost of a single binary bit, and about 1.5 million times larger than the binary representation cost of the number 1000.

If we assume that the analog representation need only cover numbers between 0 and 1000, then the dependence of the probability of error on the number of photons used to represent the highest number (1000) is given in Figure 15.33. Interestingly, even if we are willing to give up on a couple of orders of magnitude of error probability, our energy cost isn’t reduced very much. In fact, it costs about 27 million photons to represent 1000 with 1% error, and about 11 million photons to represent it with as much as 10% error. These numbers can all be reduced by about two orders of magnitude if we are willing to give up a factor of ten in dynamic range, limiting the highest representable number to 100 instead of 1000, as shown in the Figure.

FIGURE 15.33 The single pixel probability of error $P(Err)$ as a function of the number of photons detected within each pixel, for the cases of 100 and

1000 analog grey levels.

These errors arise fundamentally from the quantum statistical nature of light, and from the fact that we just can't guarantee the number of photons in a packet of light (without resorting to exotic things like "squeezed states", which have their own practical limitations as well as other costs). In the brain, of course, it is currently thought that many (but not *all*) of the quantities involved in signal transmission, both electrical and chemical, are analog in nature.

The Incorporation of Computational Complexity

Given the fact that it is considerably more expensive to represent quantities in analog as opposed to binary form, why don't we always choose to compute in the binary representation? The answer is that many operations are less energy consumptive to perform in the binary representation, but others are not. The difference lies in the degree of computational complexity that can be implemented on a given representation for a particular computational operation within a chosen technological implementation. For our purposes here, we may define the computational complexity of a given operation as the minimum number of irreducible binary bit operations (over all possible computational algorithms and machine architectures) required to complete the calculation assuming that the data is represented in binary throughout.

For operations of low computational complexity such as transferring data from the CPU to memory or logic and control operations, computation in the binary representation tends to have a significant energy consumption advantage at the fundamental limits (as well as at the current technological limits for both electronic and photonic processors). On the other hand, for operations of high computational complexity such as the two-dimensional Fourier transform that require a very large number of irreducible binary operations to perform (for the optimum algorithm), computation in the analog representation tends to

exhibit lower overall energy consumption, particularly in photonic implementations.

The Hybrid Representation Concept

In the case of neural networks, a number of characteristic types of computational operations are typically employed, including for example the calculation of weight updates and storage of updated weights, the fan-out of neuron unit outputs, the multiplication of fanned-out outputs by weights, the communication of weighted signals, the fan-in and summation (or differencing) of weighted inputs, and the thresholding of summed inputs to form neuron unit outputs. These operations span a wide gamut of computational as well as physical complexity. As such, we suggest that optimum neural network performance from an energy metric viewpoint may turn out to be best achieved with a hybrid representation, in which the signal representation is essentially binary for certain functions, and essentially analog for others.

An example of the use of this hybrid representation concept is the use of a hard threshold function within each neuron unit to create a two state output (*on* and *off*), and the use of analog holographic storage for all interconnection weights, as described in a previous section. In this case, the *inputs* to each neuron unit are analog, while the *outputs* from each neuron unit are binary. Multiplications are performed in a fully hybridized representation (multiplicands are analog while multipliers are binary), but summations are fully analog (the superposition of fanned-in input intensities). Given a particular choice of implementation technology, then, the overall power budget for a given hybrid representation can be established and compared with similar power budgets for fully analog and fully binary representations.

The Inherent Costs of Interconnections

Regardless of the representation chosen, it is clear that any computational energy bud-

get must take into account the non-negligible cost of the interconnections themselves. Interconnections characteristic of neural networks are merely a form of weighted communication channels, characterized by a high degree of fan-out and fan-in. For both electronic and photonic implementations that have adaptive weights, it takes energy to calculate the weight updates, it takes energy to store the resultant updated weights, it takes energy to perform the multiplications implied by the weighting of the output signals, and it takes energy to communicate the various signals between layers.

In the electronics case, these energy costs derive from charging up the capacitances of switching devices in the various forms of memory, flipping switches in the various arithmetic operations (both addition and multiplication) for binary representations, operating linear and nonlinear devices for analog representations, and charging and discharging the capacitance associated with output line drivers as well as the capacitance associated with the physical interconnections (wires) among the various parts of the circuit. In the photonics case, the comparable energy costs derive from the generation of light by coherent optical sources, the holographic recording of weights and weight updates, the throughput losses engendered by readout of the holographically stored (and multiplexed) interconnection matrix, any throughput losses associated with the fan-out and fan-in processes, and the inter-and intralayer communication costs.

The bottom line is that in most cases, complex, highly multiplexed interconnection networks with a high degree of fan-in and fan-out are very energy consumptive, and for the neural network case may prove to be the largest energy sink.

The Inherent Cost of Detection (Switching)

No useful computational system can avoid the costs of detection, both of intermediate results that are essential to following calculations, and of the sought-after answers or output states that initiate subsequent actions or analysis. And this is the cost above all costs that we are certainly willing to pay, as answers or outputs validate the usefulness of the system and its design. As pointed out earlier, there are three primary areas in which detections

are essential: in the generation of summed inputs prior to functional transformation into individual neuron unit outputs (usually on the input side of an optically addressed spatial light modulator), in the generation of specific system outputs, and in the holographic recording of interconnection weights.

The physical process of detection *inherently* involves the dissipation of finite energy (if accomplished in finite time within prescribed uncertainties, as is appropriate for computation), since it necessarily involves the irreversible switching of the state of a physical component, as well as the guarantee that the switched state will be maintained over the time period of measurement without fluctuations due, for example, to thermally induced statistical variations. This is particularly true in highly distributed computing systems such as neural networks that depend on a certain degree of predictability and synchronization of communicated results for progressive computation. As such this is not really a fundamental physical limitation, but rather is a technological limitation imposed by the system designer, who would really like to see some intelligible output from the system in the near future.

Optimization of the Computational Architecture

The design of a computational architecture in many ways fixes the fundamental performance limitations of the system, as choices must be made about the representation of data within the architecture, the methods employed for the implementation of computational complexity, and the frequency and nature of the detections required for both intermediate and final results. For neural networks capable of sophisticated operations, optimization of the computational architecture against one or more metrics (such as total energy cost for a given computation, or total power at a given operating frequency) will necessitate an appropriate balance among the various representations employed, as well as among the physical mechanisms employed to accomplish the necessary computations. A further balance must be struck between the fraction of the computational burden that is assigned to interconnections, and the fraction that is accomplished by switching (whether logic,

arithmetic, or detection of results).

For many classes of computational problems, the neural network paradigm may prove to be nearly optimal even in the regime in which all of the individual components are assumed to be operating at their respective fundamental physical performance boundaries. Relative to a modern digital supercomputer, certainly, neural networks seemingly offer an unusual mix of hybrid representations (primarily analog), interconnections (highly multiplexed as well as weighted), and switching (infrequent relative to the rate at which interconnections are utilized). It is at the very least intriguing to imagine whether or not our biological heritage has stored within it a useful clue about highly efficient computation for truly sophisticated problems.

Technological Limitations

The choice of a technological base (or bases) within which to design a neural network with a large number of neuron units and a high degree of connectivity implies yet another set of performance constraints above and beyond the fundamental physical boundaries referred to above. These *technological limitations* may not yet have been reached within the development of a given technology, but can at least be estimated given what we know about the physics of operation of the devices in question. One such technological limitation, for example, governs the total energy dissipation density that can be tolerated on a given semiconductor substrate without either an unacceptable temperature rise that affects device performance, or resort to extraordinary cooling measures (that may prove to be unfeasible in an optical path). As a second example, the energy required to represent a single bit using current digital logic circuits integrated in silicon is about seven orders of magnitude *above* the thermal fluctuation limit [Tanguay, 1988].

In a previous section of this chapter, we discussed a particular photonic implementation strategy for neural networks that involved specific technological choices for the various types of components required by the architecture. At the present state of development of photonic computational systems, we do not have the luxury of designing everything within

a single technological base, as is the case perhaps for computational subsystems based on VLSI chips. "Optical silicon" has not yet emerged, or at the very least has not yet been identified and recognized as such, even though numerous candidates have been intensively investigated.

Perhaps the leading candidate at the current time is the compound semiconductor system based on gallium arsenide ($GaAs$) and including related ternary compounds such as indium gallium arsenide ($In_xGa_{1-x}As$) and aluminum gallium arsenide ($Al_xGa_{1-x}As$). Within this system, at least, sources, source arrays, spatial light modulators, integrated electronic circuitry, volume holographic optical elements, detectors, and detector arrays have all been fabricated and evaluated with varying degrees of success. What has *not* been established to date is the mutual compatibility of all of these elements operating within a given systems context. This demonstration of mutual compatibility in all relevant performance specifications is essential, because in a highly interconnected system the overall performance achieved is often most strongly influenced by the component with the *least* desirable characteristics. An obvious example is that of a single channel optical communications link, for which the transmission bandwidth will be delimited by the lowest bandwidth component among the source/modulator, transmission medium, and detector/amplifier.

In the remainder of this section, we briefly discuss a number of the types of technological limitations that will impact the performance of currently envisioned photonic implementations of neural networks.

With regard to sources and source arrays, the principal technological issues are the minimization of laser thresholds to allow for parallel operation of a large number of sources on a single chip, the coherence length achievable with ultrashort cavity surface emitting lasers when fabricated in an array (which impacts the holographic recording process), the uniformity of wavelength across the array (particularly for parallel readout of wavelength sensitive devices such as multiple quantum well spatial light modulators), and both the short term (process-determined) yield and the long term reliability of individual sources within a large scale array.

For the case of spatial light modulators, key technological issues include the maximum

density of neuron units that can be integrated on a monolithic or hybrid chip with appropriate detectors, control circuitry, and modulators within each pixel; the sensitivity to input intensity; the neural unit functionality (and perhaps programmability) that can be achieved at the minimum cost in real estate and energy dissipation; the contrast ratio and uniformity of the contrast ratio across the array of pixels (neuron units); the achievable dynamic range of the input/output transfer function; and the operational bandwidth that can be reached assuming a 50% duty cycle for each neuron unit (which determines the total power dissipation of the chip).

The high degree of interconnectivity envisioned for photonic implementations of neural networks hinges primarily on the achievement of appropriate functionality in the volume holographic optical elements used to record and store interconnection weights, produce fan-in and fan-out from each neuron unit, and allow for highly parallel readout of the weighted interconnection network. Key technological limitations for currently investigated photorefractive materials include the optical quality routinely achievable in large (1 cubic inch) single crystal samples, the storage capacity of the medium as determined by the *highest spatial frequency gratings that can be recorded*, the sensitivity for recording of updated weights [Johnson, 1988] at the source wavelength (which in turn determines the source power necessary to initiate weight updates during the learning phase), the potential for "fixing" of the stored interconnection weights to allow for nondestructive readout during computation, and the capability of copying the contents of the stored interconnection matrix into another holographic medium (in order to provide the capacity for mass production of fixed pattern interconnections following a *training sequence executed with a dynamic medium*). Perhaps the most important technological limitation of a given holographic recording medium will prove to be the total number of weight update cycles that can be initiated without complete erasure of the weight updates recorded during the very first training cycle. This number in effect sets an upper bound on the learning capacity of the photonic neural network.

The primary technological limitations of importance to single pixel detectors as used, for example, on the input side of optically addressed spatial light modulators, include the sensitivity of the detector/amplifier combination (which together with the spatial light

modulator gain determines the overall loop gain for a single computational iteration), the modulation bandwidth in conjunction with following circuitry, and the chip area required to achieve the desired sensitivity and bandwidth tradeoff. Depending on the technological base within which the spatial light modulator is fabricated, the potential for integration with control circuitry and in some cases the modulation elements themselves provides an additional constraint.

For detector arrays, many of the same issues apply with the additional constraints of uniformity of each performance parameter across the array, and the reliability of the full array of pixels. Another important technological issue is the frame rate for readout of the entire array at a given pixel density, which is determined both by the technological base within which the array is fabricated, and by the physical structure of the array and its readout configuration. As discussed in an earlier section, charge-coupled-device (CCD) arrays are typically read out by temporally multiplexing the contents of a full frame onto one or a few high bandwidth serial outputs. The contents of each row of stored charge packets generated during the exposure cycle are shifted out to a high speed serial readout buffer, which reads out one entire column of pixels in between each lateral shift of the rows as shown schematically in Figure 15.34.

FIGURE 15.34 Illustration of parallel-to-serial conversion in two-dimensional detector arrays such as the charge-coupled-device (CCD) array. Charge accumulated within each photosensitive region during exposure is transferred laterally by a set of row-parallel shift registers to a high speed parallel-to-serial shift register, which reads out the entire array one column at a time.

This parallel-to-serial conversion limits the frame readout rate to the maximum serial transfer rate achievable in the readout buffer, divided by the number of pixels in the array. For example, if the array is 2048×2048 pixels in size with a readout buffer operating at 200 MHz, the frame rate will be limited to about 60 frames per second. For many neural

network applications, this frame rate may be more than sufficient for access to the desired outputs (including the time required for temporal demultiplexing of the output). In cases that demand higher frame rates, the array can be segmented so that multiple readout buffers can be used, each accessing a fraction of the total number of rows in the array.

An unusual feature of identifying the technological constraints that bound *any* neural network implementation, whether it be electronic, photonic, chemical, mechanical, or all of the above, is the fact that we just don't know enough yet about the operation of highly interconnected nonlinear systems of large dimension to fully assess the impact of a *particular* constraint on the overall system operation. One example is the degree to which nonuniformities in the neuron units themselves (*e.g.* in their sensitivity, contrast ratio, or overall response function) or in the interconnection medium can be tolerated by an architecture that is to a large extent self-organizing. A second example is the necessity within a certain neural network paradigm of implementing precisely the right nonlinearity that translates summed neuron inputs to neuron outputs (or, for that matter, the very existence of nonlinearities in the recording and storage of weight updates, and in the readout of the full weighted interconnection pattern). Some of these types of questions may be amenable to simulation, but in other cases we may have to await the results from actual implementations to refine our understanding of the technological requirements.

THE FUTURE OF NEURO-OPTICAL COMPUTATION

A wide variety of photonic architectures and components are currently under intensive investigation for neural network applications. A thorough discussion of these alternative strategies and their strengths and weaknesses is unfortunately beyond the scope of this chapter. In this final section, we address the future prospects of neuro-optical computation from the point of view of those critical issues that are common to all such proposed implementation strategies.

The Critical Issues

As with any emerging technological breakthrough, successful advanced development of photonic neural networks will require that the technology prove to be *manufacturable*, in the sense that it is amenable to mass production techniques at reasonable cost; *flexible in design*, in that the technological base provides significant degrees of freedom for architectural and functional variations; and *leveraged* as much as possible by developments in related technologies that can offload a significant fraction of the development time and costs. Implicit in these three key features is the issue of the component uniformity that can be achieved over large array sizes, and the related issue of the scalability of the technology to large scale systems either by increases in the basic array sizes or by the incorporation of a modular design from the outset.

In all useful computational systems, the bottom line is to a large extent determined by the maximum amount of computational throughput capacity that can be squeezed into the smallest system volume within a tolerable energy dissipation constraint or power budget. In the case of photonic neural networks that utilize the implementation strategy outlined in a previous section of this chapter, the two most important factors that influence the computational throughput capacity are the storage capacity of the volume holographic optical element, and the operational bandwidth of the neuron units (nonlinear spatial light modulators).

The storage capacity of a holographic interconnection medium is in turn determined by the maximum number of independent weighted interconnections that can be recorded and retrieved per unit volume. Although we discussed this issue in an earlier section from a theoretical viewpoint (and in yet another section from an architectural viewpoint), we have not addressed herein the even more important question of the *actual* density of independent interconnections achievable with photorefractive (or other photosensitive real time materials) that contain a considerable number of scattering centers and exhibit diffraction efficiencies that depend strongly on the spatial frequency of the recorded grating. Demonstration of a high density of weighted interconnections with low interchannel crosstalk in a

real time holographic recording medium will provide a benchmark of achievement for photonic implementation strategies, as well as a metric by which one can more appropriately estimate eventual system performance. Demonstration of parallel weight updates at high sensitivity (requiring tolerable optical source intensities) during a complete training cycle is also necessary for a convincing proof of system feasibility.

The operational bandwidth of two-dimensional spatial light modulators that can be used as neuron unit arrays will prove to be orders of magnitude larger than the bandwidth characteristic of biological systems. Feasibility analyses, as well as preliminary device characterization studies, indicate no fundamental or technological barriers to operation of individual neuron units at bandwidths exceeding 100 MHz [Asthana, 1990b]. For a 100×100 element neuron unit to switch at 100 MHz with a 50% duty cycle, however, generates a watt of power dissipation for every 2 picojoules of switching energy required by an individual neuron unit, including the detector, amplifier, control circuitry, and modulator. Although the neuron unit arrays in most neural network architectures will not approach 50% duty cycles from full *off* to full *on* in actual operation, this still gives us a very tight energy dissipation budget, and may eventually force a lowering of the design bandwidth.

For photonic implementations of neural networks that use optical imaging and holographic interconnection systems extensively to increase the computational throughput capacity, an important question will continue to be the fraction of "unfilled" system volume dedicated to wavefront and beam propagation. Miniaturization of many sophisticated optical signal processing systems has been a focus of effort only recently, and can be expected to produce significant system volume reductions through clever (as well as careful) optomechanical engineering. For example, gradient index (GRIN) techniques have been used in conjunction with photolithographic planar processing to produce regular two-dimensional arrays of diffraction limited microlenses that can be incorporated in stacked planar optical modules with greatly reduced unfilled system volume [Iga, 1984]. There are, however, several inherent limits (both fundamental as well as technological) that provide lower bounds on the physical volume required to implement a high degree of interconnectivity among planes of neuron units.

When all is said and done, it is certainly a fair question to ask whether the physical

volume of a neuro-optical computer module would be better off densely packed with silicon chips that emulate the same functionality. All of the preliminary evidence gathered to date suggests that the answer to this question shifts rather dramatically from an emphatic "yes" in the limit of small numbers of neurons and required interconnections, to a more tenuous "no" as the number of neurons and density of required interconnections continues to increase. Perhaps the most interesting question for the future of neuro-optical computation is the clear identification of this performance boundary.

The Incorporation of Neural Paradigms

In concluding this chapter on neuro-optical computation, we assert that although the majority of preliminary demonstrations of photonic neural network architectures have seemingly focused on associative memories in general, and on variations of the Hopfield-Amari [Hopfield, 1982; Amari, 1972] network in particular, it is essential that photonic implementations have the capacity for incorporation of a wide variety of neural network architectures, computational algorithms, and learning rules. This is particularly important in view of the early stage of development that characterizes our current understanding of the operational performance of even the most fashionable neural network models, when scaled up to large numbers of densely interconnected neuron units with realistic stochastic variations in individual neuron unit performance.

Furthermore, it is likely that useful neural networks incorporated in a systems framework may require either a number of layers with different characteristics, or considerable pre-and post-processing to achieve sophisticated functionality. One example of a neural paradigm that requires such additional sophistication is the Dynamic Link Architecture of von der Malsburg discussed in Chapter 11 [Buhmann, 1991] as applied to pattern recognition problems by means of graph matching techniques. A photonic implementation of this architecture will likely require several interacting modules for complete functionality.

Currently investigated photonic architectures and components for neural network implementation do not yet enjoy the flexibility of full-fledged computer aided design and

computer aided manufacturing that is the hallmark of the silicon VLSI circuit industry. On the other hand, tremendous strides have been made in just the past few years in the simulation of even complex optical systems including the effects of both refractive and diffractive components. The next step, from extensive simulation capabilities to design automation, is under active investigation and may when taken herald the beginnings of a viable photonic-based neural network implementation technology.

SUGGESTED FURTHER READING

Abu-Mostafa, Y. S., and Psaltis, D., "Optical Neural Computers", *Scientific American*, vol. 256, no. 3, 88-95, 1987.

Applied Optics, Special Issue on Neural Networks, vol. 26, no. 23, 1 December, 1987.

Arsenault, H., Szoplik, T., and Macukow, B., Eds., *Optical Processing and Computing*, Academic Press, New York, 1989.

Collier, R. J., Burckhardt, C. B., and Lin, L. H., *Optical Holography*, Academic Press, New York, 1971.

Feitelson, D. G., *Optical Computing: A Survey for Computer Scientists*, MIT Press, Cambridge, 1988.

Goodman, J. W., *Introduction to Fourier Optics*, McGraw-Hill Book Company, New York, 1968.

Gunter, P., and Huignard, J.-P., *Photorefractive Materials and Their Applications I*, Vol-

ume 61 in Topics in Applied Physics Series. Springer-Verlag, New York, 1988; also Gunter, P., and Huignard, J.-P., *Photorefractive Materials and Their Applications II*, Volume 62 in Topics in Applied Physics Series. Springer-Verlag, New York, 1989.

Haus, H. A., *Waves and Fields in Optoelectronics*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1984.

Ishihara, S., Ed., *Optical Computing in Japan*, Nova Science Publishers, Commack, New York, 1990.

Nussbaum, A., and Phillips, R. A., *Contemporary Optics for Scientists and Engineers*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1976.

Optical Computing, Volume 9 of the 1989 OSA Technical Digest Series, Optical Society of America, Washington, D.C., 1989.

Pankove, J. I., *Optical Processes in Semiconductors*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1971.

Proceedings of the IEEE International Conference on Neural Networks, vol. III, San Diego, 1987, pp. 549-648; *Proceedings of the IEEE International Conference of Neural Networks*, vol. II, San Diego, 1988, pp. 357-442; *Proceedings of the International Joint Conference on Neural Networks*, vol. II, Washington, D.C., 1989, pp. 457-494; (The Institute of Electrical and Electronics Engineers, Inc., Piscataway, NJ).

Smith, H. M., *Holographic Recording Materials*, Volume 20 in Topics in Applied Physics Series, Springer-Verlag, New York, 1977.

Sze, S. M., *Physics of Semiconductor Devices, 2nd Ed.*, John Wiley and Sons, New York, 1981.

REFERENCES

Abu-Mostafa, Y., "Complexity in Neural Systems". Appendix D in Mead, C.A., *Analog VLSI and Neural Systems*, Addison-Wesley Publishing Company, Reading, Massachusetts, 1989, pp. 353-358.

Amari, S.-I., "Learning Patterns and Pattern Sequences by Self-Organizing Nets of Threshold Elements," *IEEE Transactions on Computers*, vol. C-21, 1197-1206, 1972.

Asthana, P., Chin, H., Nordin, G., Tanguay, A. R., Jr., Piazzolla, S., Jenkins, B. K., and Madhukar, A., "Photonic components for neural net implementations using incoherent/coherent holographic interconnections," *OC'90 Technical Digest, International Commission for Optics*, Kobe, Japan, 1990a.

Asthana, P., Chin, H., Nordin, G., Tanguay, A. R., Jr., Petrisor, G. C., Jenkins, B. K., and Madhukar, A., "Photonic components for neural net implementations using incoherent-coherent holographic interconnections", in *OSA Annual Meeting Technical Digest 1990*, Vol. 15 of the 1990 OSA Technical Digest Series, Optical Society of America, Washington, D. C., 1990b, p. 57.

Asthana, P., Nordin, G., Piazzolla, S., Tanguay, A. R., Jr., and Jenkins, B. K., "Analysis of interchannel crosstalk and throughput efficiency in highly multiplexed fan-out/fan-in holographic interconnections", in *OSA Annual Meeting Technical Digest 1990*, Vol. 15 of the 1990 OSA Technical Digest Series, Optical Society of America, Washington, D. C., 1990c, p. 242.

Blouke, M. M., Corrie, B., Heidtmann, D. L., Yang, F. H., Winzenread, M., Lust, M. L., Marsh IV, H. H., and Janesick, J. R., "Large format, high resolution image sensors", *Optical Engineering*, vol. 26, no. 9, 837-843, 1987.

Buhmann, J., Lange, J., von der Malsburg, C., Vorbrüggen, J. C., and Würtz, R. P., "Object Recognition in the Dynamic Link Architecture-Parallel Implementation on a Transputer Network," in *Neural Networks: A Dynamical Systems Approach to Machine Intelligence*, B. Kosko, Ed., Prentice-Hall, Englewood Cliffs, NJ, 1991; Chapt. 11, pp. xxx-xxx.

Carpenter, G. A. and Grossberg, S., "ART 2: self-organization of stable category recognition codes for analog input patterns," *Applied Optics*, vol. 26, no. 23, 4919-4930, 1987.

Casey, H. C., Jr., and Panish, M. B., *Heterostructure Lasers. Part A: Fundamental Principles*, in Quantum Electronics-Principles and Applications Monograph Series, P. F. Liao and P. Kelley, Eds., Academic Press, Inc., New York, 1978a.

Casey, H. C., Jr., and Panish, M. B., *Heterostructure Lasers. Part B: Materials and Operating Characteristics*, in Quantum Electronics-Principles and Applications Monograph Series, P. F. Liao and P. Kelley, Eds., Academic Press, Inc., New York, 1978b.

Chang-Hasnain, C. J., Maeda, M. W., Stoffel, N. G., Harbison, J. P., and Florez, L. T., "Surface Emitting Laser Arrays with Uniformly Separated Wavelengths". *Electronics Letters*, vol. 26, no. 13, 940-942, 1990.

Changeux, J.-P., *Neuronal Man: The Biology of Mind*, Pantheon Books, New York, 1985.

Dammann, M., and Gortler, K., "High-Efficiency In-Line Multiple Imaging by Means of

Multiple Phase Holograms", *Optics Communications*, vol. 3, no. 5, 312-315, 1971.

Drabik, T. J., and Handschy, M. A., "Silicon VLSI/ferroelectric liquid crystal technology for micropower optoelectronic computing devices", *Applied Optics*, vol. 29, no. 35, 5220-5223, 1990.

Dowling, J. E., *The Retina: An Approachable Part of the Brain*. The Belknap Press of Harvard University Press, Cambridge, Massachusetts, 1987.

Farhat, N. H., "Optoelectronic analogs of self-programming neural nets: architecture and methodologies for implementing fast stochastic learning by simulated annealing," *Applied Optics*, vol. 26, no. 23, 5093-5103, 1987.

Farhat, N. H., Psaltis, D., Prata, A., and Paek, E., "Optical implementation of the Hopfield model," *Applied Optics*, vol. 24, no. 10, 1469-1475, 1985.

Fossum, E. R., "Charge-coupled computing for focal plane image preprocessing", *Optical Engineering*, vol. 26, no. 9, 916-922, 1987.

Goodman, J. W., *Introduction to Fourier Optics*, McGraw-Hill Book Company, New York, 1968, Chapt. 4.

Gunter, P., and Huignard, J.-P., *Photorefractive Materials and Their Applications I*, Volume 61 in Topics in Applied Physics Series, Springer-Verlag, New York, 1988.

Gunter, P., and Huignard, J.-P., *Photorefractive Materials and Their Applications II*, Volume 62 in Topics in Applied Physics Series, Springer-Verlag, New York, 1989.

Haus, H. A., *Waves and Fields in Optoelectronics*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1984, pp. 63-72.

Herbulock, E. J., Garrett, M. H., and Tanguay, A. R., Jr., "Electric field profile effects on photorefractive grating formation in bismuth silicon oxide", *OSA Annual Meeting Technical Digest 1988*, Vol. 11 of the 1988 OSA Technical Digest Series, Optical Society of America, Washington, D. C., 1988, p. 143.

Hopfield, J. J., "Neural Networks and Physical Systems with Emergent Collective Computational Activity," *Proceedings of the National Academy of Sciences, USA*, vol. 79, 2554-2558, 1982.

Huang, A., Tsunoda, Y., Goodman, J. W., and Ishihara, S., "Optical computation using residue arithmetic", *Applied Optics*, vol. 18, no. 2, 149-162, 1979.

Hubel, D. H., "The Brain", *Scientific American*, vol. 241, no.3, 44-53, 1979.

Iga, K., Kokubun, Y., and Oikawa, M., *Fundamentals of Microoptics: Distributed-Index, Microlens, and Stacked Planar Optics*, Academic Press, Inc., Tokyo, 1984.

Iversen, L. L., "The Chemistry of the Brain", *Scientific American*, vol. 241, no. 3, 134-149, 1979.

Jackel, L. D., "Electronic Neural Networks," in *OSA Annual Meeting Technical Digest, 1988*, Vol. 11 of the 1988 OSA Technical Digest Series, Optical Society of America, Washington, D.C., 1988, p. 146.

Jenkins, B. K., Chavel, P., Forchheimer, R., Sawchuk, A. A., and Strand, T. C., "Architectural Implications of a Digital Optical Processor," *Applied Optics*, vol. 23 no. 19, pp. 3465-3474, 1984.

Jenkins, B. K., Petrisor, G. C., Piazzolla, S., Asthana, P., and Tanguay, A. R., Jr., "Photonic architecture for neural nets using incoherent/coherent holographic interconnections," in *OC'90 Technical Digest, International Commission for Optics*, Kobe, Japan, 1990a.

Jenkins, B. K., Tanguay, A. R., Jr., Piazzolla, S., Petrisor, G. C., and Asthana, P., "Photonic neural network architecture based on incoherent-coherent holographic interconnections", in *OSA Annual Meeting Technical Digest 1990*, Vol. 15 of the 1990 OSA Technical Digest Series, Optical Society of America, Washington, D. C., 1990b, p. 56.

Jenkins, B. K. and Tanguay, A. R., Jr., "Incoherent/coherent multiplexed holographic recording for photonic interconnections and holographic optical elements." United States Patent Application USC-2254, University of Southern California, Los Angeles, California, 1990c.

Jewell, J. L., Lee, Y. H., Scherer, A., McCall, S. L., Olsson, N. A., Harbison, J. P., and Florez, L. T., "Surface-emitting microlasers for photonic switching and interchip connections", *Optical Engineering*, vol. 29, no. 3, 210-214, 1990.

Johnson, R. V., and Tanguay, A. R., Jr., "Optical beam propagation method for birefringent phase grating diffraction," *Optical Engineering*, vol. 25, no. 2, 235-249, 1986.

Johnson, R. V., and Tanguay, A. R., Jr., "Fundamental Physical Limitations of the Photorefractive Grating Recording Sensitivity", Chapter 3 in *Optical Processing and Computing*, H. Arsenault, T. Szoplik, and B. Macukow, Eds., Academic Press, New York, 1989.

Jones, W. B., Jr., *Introduction to Optical Fiber Communication Systems*, Holt, Rinehart and Winston, Inc., New York, 1988.

Kaminow, I. P., *An Introduction to Electrooptic Devices*, Academic Press, New York, 1974.

Karim, Z., Garrett, M. H., and Tanguay, A. R., Jr., "Bandpass AR coating design for bismuth silicon oxide", in *OSA Annual Meeting Technical Digest 1988*, Vol. 11 of the 1988 OSA Technical Digest Series, Optical Society of America, Washington, D. C., 1988, p. 125.

Karim, Z., and Tanguay, A. R., Jr., "Bandpass AR coating for the photorefractive materials LiNbO_3 , BaTiO_3 , CdTe , and PLZT ", in *OSA Annual Meeting Technical Digest 1989*, Vol. 18 of the 1989 OSA Technical Digest Series, Optical Society of America, Washington, D. C., 1989a, p. 78.

Karim, Z., Kyriakakis, C., and Tanguay, A. R., Jr., "Improved two beam coupling gain and diffraction efficiency in bismuth silicon oxide crystals using a bandpass AR coating", in *OSA Annual Meeting Technical Digest 1989*, Vol. 18 of the 1989 OSA Technical Digest Series, Optical Society of America, Washington, D. C., 1989b, p. 29.

Kogelnik, H., "Coupled Wave Theory for Thick Hologram Gratings", *Bell System Technical Journal*, vol. 48, no. 9, 2909-2947, (1969).

Kressel, H., and Butier, J. K., *Semiconductor Lasers and Heterojunction LEDs*, in Quantum Electronics-Principles and Applications Monograph Series, Y.-H. Pao and P. Kelley, Eds., Academic Press, Inc., New York, 1977.

Kyriakakis, C., Karim, Z., Jung, J. J., Tanguay, A. R., Jr., and Madhukar, A., "Funda-

mental and Technological Limitations of Asymmetric Cavity MQW InGaAs/GaAs Spatial Light Modulators", in *Proceedings of the Optical Society of America Topical Conference on Spatial Light Modulators, Incline Village, Nevada*, Optical Society of America, Washington, D.C., 1990.

Lee, B. W. and Sheu, B. J., "Designs and Analysis of VLSI Neural Networks", in *Neural Networks and Fuzzy Systems: A Dynamical Systems Approach to Machine Intelligence*, Bart Kosko, Ed., Prentice-Hall, Englewood Cliffs, New Jersey, 1990; Chapt. 14, pp. xxx-xxx.

Lentine, A. L., Hinton, H. S., Miller, D. A. B., Henry, J. E., Cunningham, J. E., and Chirovsky, L. M. F., "Symmetric self-electro-optic effect device: optical set-reset latch," *Applied Physics Letters*, vol. 52, 1419-1421, 1988.

Lentine, A. L., Chirovsky, L. M. F., and D'Asaro, L. A., "Photonic Ring Counter Using Batch-Fabricated Symmetric Self-Electro-Optic-Effect Devices," *Optics Letters*, vol. 16, no. 1, 36-38, 1991.

Marrakchi, A., Hubbard, W. M., Habiby, S. F., and Patel, J. S., "Dynamic holographic interconnects with analog weights in photorefractive crystals", *Optical Engineering*, vol. 29, no. 3, 215-224, 1990.

McCormick, F. B., "Generation of large spot arrays from a single laser beam by multiple imaging with binary phase gratings", *Optical Engineering*, vol. 28, no. 4, 299-304, 1989.

McCormick, F. B., Lentine, A. L., Morrison, R. L., Walker, S. L., Chirovsky, L. M. F., and D'Asaro, L. A., "Simultaneous parallel operation of an array of symmetric self-electrooptic effect devices," in *OSA Annual Meeting Technical Digest 1989*, Vol. 18 of the 1989 OSA

Technical Digest Series, Optical Society of America, Washington, D.C., 1989b, pp. 60-61.

Mead, C. A., *Analog VLSI and Neural Systems*, Addison-Wesley Publishing Company, Inc., Reading, Massachusetts, 1989.

Mead, C. A., *op. cit.*, 1989b, p. 3.

Mead, C. A. and Mahowald, M. A., "A silicon model of early visual processing," *Neural Networks*, vol. 1, 91-97, 1988.

Miller, D. A. B., "Quantum-well self-electro-optic effect devices," *Optical and Quantum Electronics*, vol. 22, S61-S98, 1990.

Milonni, P. W. and Eberly, J. H., "Specific Lasers and Pumping Mechanisms", Chapter 13 in *Lasers*, John Wiley and Sons, New York, 1988, pp. 411-468.

Morrison, R.L., and Walker, S.L., "Binary phase gratings generating even numbered spot arrays", in *OSA Annual Meeting Technical Digest 1989*, Vol. 18 of the 1989 OSA Technical Digest Series, Optical Society of America, Washington, D. C., 1989, p. 111.

Optical Engineering, Special Issue on Charge-Coupled-Device Manufacture and Application, vol. 26, no. 9, 827-943, 1987a.

Optical Engineering, Special Issue on Charge-Coupled-Device and Charge-Injection-Device Theory and Application, vol. 26, no. 10, 963-1076, 1987b.

Orenstein, M., von Lehmen, A. C., Chang-Hasnain, C., Stoffel, N. G., Harbison, J. P.,

Florez, L. T., Wullert, J. R., and Scherer, A., "Matrix addressable surface emitting laser array", in *Proceedings of the 1990 Conference on Lasers and Electro-Optics*, Vol. 7 of the 1990 Technical Digest Series, Optical Society of America, Washington, D. C., 1990a, p. 88.

Orenstein, M., von Lehmen, A. C., Stoffel, N. G., Chang-Hasnain, C., Harbison, J. P., Florez, L. T., Clausen, E., and Jewell, J. L., "Lateral definition of high performance surface emitting lasers by planarity preserving ion implantation processes", in *Proceedings of the 1990 Conference on Lasers and Electro-Optics*, Vol. 7 of the 1990 Technical Digest Series, Optical Society of America, Washington, D. C., 1990b, p. 504.

Petrisor, G. C., Jenkins, B. K., Chin, H., and Tanguay, A. R., Jr., "Dual function adaptive neural networks for photonic implementation", in *OSA Annual Meeting Technical Digest 1990*, Vol. 15 of the 1990 OSA Technical Digest Series, Optical Society of America, Washington, D. C., 1990, p. 56.

Psaltis, D., and Farhat, N. H., "Optical information processing based on an associative-memory model of neural nets with thresholding and feedback," *Optics Letters*, vol. 10, no. 2, 98-100, 1985.

Psaltis, D., Brady, D., and Wagner, K., "Adaptive optical networks using photorefractive crystals", *Applied Optics*, vol. 27, no. 9, 1752-1759, 1988.

Psaltis, D., Brady, D., Gu, X.-G., and Hsu, K., "Optical Implementation of Neural Computers," in *Optical Processing and Computing*, H. H. Arsenault, T. Szoplik, and B. Macukow, Eds., Academic Press, New York, 1989, pp. 251-276.

Shirouzu, S., Tsuji, T., Harada, N., Sado, T., Aihara, S., Tsunoda, R., and Kanno, T., "64 x 64 InSb Focal Plane Array with Improved Two Layer Structure", *Proceedings of*

the SPIE, vol. 661, Society of Photo-Optical Instrumentation Engineers, Bellingham, WA, 1986.

Smith, H. M., *Holographic Recording Materials*, Volume 20 in Topics in Applied Physics Series, Springer-Verlag, New York, 1977.

Spatial Light Modulators and Applications, Volume 14 of the 1990 OSA Technical Digest Series, Optical Society of America, Washington, D.C., 1990.

Spatial Light Modulators for Optical Information Processing, Special Issue of *Applied Optics*, vol. 28, no. 22, 1989, pp. 4739-4913.

Streetman, B. G., *Solid State Electronic Devices*, Second Edition, Prentice-Hall, Englewood Cliffs, New Jersey, 1980.

Sze, S. M., *Physics of Semiconductor Devices*, 2nd Ed., John Wiley and Sons, New York, 1981a, pp. 312-361.

Sze, S. M., *op. cit.*, 1981b, pp. 783-784.

Sze, S. M., *op. cit.*, 1981c, pp. 407-427.

Tai, K., Fischer, R. J., Wang, K. W., Chu, S. N. G., and Cho, A. Y., "Use of Implant Isolation for Fabrication of Vertical Cavity Surface-Emitting Laser Diodes", *Electronics Letters*, vol. 25, no. 24, 1644-1645, 1989a.

Tai, K., Fischer, R. J., Seabury, C. W., Olsson, N. A., Huo, T.-C. D., Ota, Y., and Cho,

A. Y., "Room-temperature continuous-wave vertical-cavity surface-emitting GaAs injection lasers", *Applied Physics Letters*, vol. 55, no. 24, 2473-2475, 1989b.

Tanguay, A. R., Jr., "Physical and Technological Limitations of Optical Information Processing and Computing", *Materials Research Society Bulletin*, Special Issue on Photonic Materials, vol. XIII, no. 8, 36-40, 1988.

Tanguay, A. R., Jr., "Materials requirements for optical processing and computing devices," *Optical Engineering*, vol. 24, no. 1, 2-18, 1985.

von der Malsburg, C., "Goal and Architecture of Neural Computers", in *Neural Computers*, R. Eckmiller and C. von der Malsburg, Eds., Volume 41 of NATO Advanced Science Institutes Series F: Computer and Systems Sciences, Springer-Verlag, New York, 1987.

von Lehmen, A., Orenstein, M., Chang-Hasnain, C., Banwell, T., Wullert, J., Stoffel, N., Florez, L., and Harbison, J., "Rastered operation of row-column addressed vertical-cavity surface-emitting laser array", in *OSA Annual Meeting Technical Digest 1990*, Vol. 15 of the 1990 OSA Technical Digest Series, Optical Society of America, Washington, D. C., 1990, p. 15.

Wagner, K., and Psaltis, D., "Multilayer Optical Learning Networks," *Applied Optics*, vol. 26, no. 23, 5061-5076, 1987.

Warde, C., and Fisher, A. D., "Spatial Light Modulators: Applications and Functional Capabilities", Chapter 7.2 in *Optical Signal Processing*, J. L. Horner, Ed., Academic Press, Inc., New York, 1987, pp. 477-523.

Whitehead, M., and Parry, G., "High-contrast reflection modulation at normal incidence

in asymmetric multiple quantum well Fabry-Perot structure", *Electronics Letters*, vol. 25, 566-568, 1989a.

Whitehead, M., Parry, G., and Wheatley, P., "Investigation of etalon effects in GaAs-AlGaAs multiple quantum well modulators", *IEE Proceedings*, vol. 136, pt. J, no. 1, 52-58, 1989b.

Whitehead, M., Rivers, A., Parry, G., Roberts, J. S., and Button, C., "Low-voltage multiple quantum well reflection modulator with on:off ratio $> 100:1$ " *Electronics Letters*, vol. 25, no. 15, 984-985, 1989c.

Yan, R. H., Simes, R. J., and Coldren, L. A., "Wide-bandwidth, high-efficiency reflection modulators using an unbalanced Fabry-Perot structure," *Applied Physics Letters*, vol. 55, no. 19, 1946-1948, 1989.

ACKNOWLEDGEMENTS

We gratefully acknowledge the contributions to this effort provided by our faculty colleagues Christoph von der Malsburg, Joachim Buhmann, and Anupam Madhukar, and by our graduate research assistants Greg Nordin, Praveen Asthana, Howard Chin, Sabino Piazzolla, Greg Petrisor, Chris Kyriakakis, Zaheed Karim, John Rilum, and Ed Herbulock. Special thanks are also due to Gloria Bullock and Delsa Tan for their help in the preparation of this manuscript. Funding for the materials, device, and systems aspects of our research program on photonic implementations of neural networks has been provided

by the Defense Advanced Research Projects Agency, the University Research Initiative "Center for the Integration of Optical Computing" (sponsored by the Air Force Office of Scientific Research), the National Center for Integrated Photonic Technology (sponsored by the Defense Advanced Research Projects Agency), the Joint Services Electronics Program, and NTT Corporation.

PROBLEMS

1. Consider a system of beamsplitters arranged to combine a set of N input beams to form a single, collinear output beam with an intensity proportional to an equally weighted sum of all of the inputs. Choose a particular architecture for the beamsplitter arrangement, and justify it in terms of efficiency, simplicity, or minimization of component count. For the case of incoherent illumination, derive the optimal transmissivities of the beamsplitters in your arrangement, and prove that the chosen architecture generates an input-output relationship in the form of Equation (2). Repeat the analysis for the case of coherent illumination, and derive the equivalent of Equation (8) that characterizes the chosen architecture.
2. Two mutually coherent beams of intensities $|a|^2$ and $|b|^2$ are incident on a detector. The coherent superposition of the beams is given by, in one dimension,

$$A(x) = ae^{jk_1x} + be^{jk_2x}.$$

- (a) Plot the resulting intensity, $|A(x)|^2$, as a function of x .
 - (b) Show that the integral of $|A(x)|^2$ over an integral number of its periods is equal to $|a|^2 + |b|^2$, thus proving that the detector's response is equal to the *incoherent* sum of the individual beams.
3. For the case of a thin phase grating with a sinusoidal index modulation given by $n(x) = n_0 + n_1 \sin k_G x$ with $n_1 < n_0$, calculate the value of $n_1 d$ that maximizes

- the diffraction efficiency into the first diffracted order, and the maximum diffraction efficiency achievable. Assume that the grating is read out by a semiconductor laser with a wavelength of 850 nm . For a given incident intensity of the optical readout beam, calculate the ratio of the intensity diffracted into the 0^{th} , 2^{nd} , and 3^{rd} diffracted orders to that diffracted into the 1^{st} order.
4. Consider the process of diffraction from a thin *amplitude* grating with a spatially varying transmissivity and negligible phase modulation. For a sinusoidal transmittance modulation given by $t(x) = t_0 + t_1 \sin k_G x$ with $t_1 < t_0$, calculate the value of t_1 that maximizes the diffraction efficiency into the first diffracted order, and the maximum diffraction efficiency achievable. Assume that the grating is read out by a semiconductor laser with a wavelength of 850 nm . For a given incident intensity of the optical readout beam, calculate the ratio of the intensity diffracted into the 0^{th} , 2^{nd} , and 3^{rd} diffracted orders to that diffracted into the 1^{st} order. Discuss the essential differences observed between the amplitude and phase grating cases.
 5. Consider the process of holographic grating recording in a thick holographic recording medium. Assume that the grating is recorded and read out by a semiconductor laser with a wavelength of 850 nm , and that the angle included between the two recording beams is 30 degrees. What is the spatial frequency of the recorded grating? How thick must the hologram be in order to generate a grating parameter Q (as defined in the text) of 1,000? What is the approximate angular width of this recorded grating (as measured, for example, by varying the angle of incidence of the readout beam)?
 6. For the thick holographic grating described in Problem 4, calculate the amplitude of the refractive index modulation that is necessary to achieve 100% diffraction efficiency on readout. How large an absorption coefficient can be tolerated in the holographic recording medium at the readout wavelength if the absorption loss in diffraction efficiency is to be kept below 5%?
 7. Consider the holographic interconnection scheme depicted in Figure 15.9. First, derive the basic relationship for a lens that associates a given point p_1 in the input

plane with a resulting beam angle θ_1 . Given a focal length of 5 centimeters for lenses L_1 and L_2 , what is the minimum spacing required between nearest neighbor points in the input plane for a grating Q of 1,000? What Q will be required to accommodate of order 10^4 input positions?

8. Design a differentiating circuit for incorporation in an optically addressed spatial light modulator, using the principles of Chapters 14 and 15. Assume that the detector is a $p-i-n$ photodiode, and that the modulator can be treated as a purely capacitive load. If the modulator can be modeled as a parallel plate capacitor of dimensions $30 \times 50 \mu m$, with a thickness of $1 \mu m$ and a relative dielectric constant typical of gallium arsenide multiple quantum well devices ($\epsilon = 13$), estimate the bandwidth over which the differentiator is operational.
9. If storage of one synaptic weight in VLSI requires a memory element $15 \mu m \times 15 \mu m$ in size, what is the maximum number of synaptic weights that can be implemented on a $1 cm \times 1 cm$ chip? If a VLSI neural network is fully connected using one such memory element for each synapse, how many pins are required for input to, and output from, the set of neuron units? How many pins would be required for parallel input of the weights? If an optical synaptic weight can be implemented in an effective volume of $5 \mu m \times 5 \mu m \times 5 \mu m$, what is the maximum number of synaptic weights that can be implemented in a volume $1 cm \times 1 cm \times 1 cm$?
10. (a) If a neural network is simulated on a digital, sequential machine, how many multiply operations and add operations are required to simulate the computational process of a single-layer feedforward network with N neuron units and a connectivity (number of connections per neuron unit) of M ? If a multiply operation can be performed in $100 ns$ and an add operation in $25 ns$, what is the minimum time it will take for one pass through the network if $N = 10^6$ and $M = 10^4$?
- (b) For the outer product learning of Equation (24), neglecting the decay term, how many multiply operations and add operations are required for one iteration of weight updates, in terms of M and N ? For a two-layer network (1 hidden

layer), with each layer having $N = 10^6$ and $M = 10^4$, and assuming 10^4 different patterns presented 1,000 times each, how long will the network take to be trained (assuming 1 forward pass and 1 update of all weights per presentation, 100 ns per multiply operation and 25 ns per add operation)?

- (c) For the same numbers of (b), assuming each layer of a photonic system can perform a forward pass in 100 ns and a set of parallel weight updates in 1 μ s, how long will it take to be trained?
11. (a) Design an algorithm that uses only nonnegative signals outside of each neuron unit, nonnegative weights, and allows two separate inputs to and two separate outputs from, each neuron unit. It should be able to perform neural computation and weight updates for learning. Your answer should be in the form of a flow chart. You may perform subtraction and division only within each neuron unit; only addition, multiplication, interconnection and storage can be performed external to the neuron units. (No need to simulate.)
- (b) After many iterations during learning, might there be a problem with weights saturating or going out of bounds? If not, why not? If so, conjecture as to how this problem might be avoided.
12. In regard to Equation (24), find an expression for δ_i for the following algorithms:
- (a) Perceptron
 - (b) Widrow-Hoff
 - (c) Least minimum square (back propagation), for a multi-layer net for
 - (i) output layer
 - (ii) hidden layers

Assume that $\beta = 0$ for this problem.

(Note: This problem requires familiarity with neural networks not discussed in this chapter.)

13. Referring to the system of Figure 15.21, if the interconnection mask SLM were electronically addressed with a serial line, capable of transmitting analog values at 50 *MHz*, and there are 10^6 pixels (analog weights), what is the maximum frame rate? If there are 10^3 parallel lines addressing? If the SLM is optically addressed, what limits the frame rate?
14. (a) Referring to Figure 15.23, show how the following network can be drawn as a single layer network with feedback.

<< Insert Figure for Problem 14(a) >>

 (b) How can the architecture shown in Figure 15.26 be modified to include feedback connections within the module? Sketch the resulting architecture.
15. Consider a charge-coupled-device array as shown schematically in Figure 15.34, of dimension 1000×1000 pixels with a serial readout buffer that operates at a clock frequency of 100 *MHz*. Calculate the maximum frame rate achievable, and the speed required of the row shift registers. Calculate the ratio between the total storage time required of the first pixel read out to that of the last in each frame.

FIGURE CAPTIONS

Fig. 15.1 Illustration of optical addition utilizing a 50/50 beamsplitter: (a) collinear *incoherent* beam geometry; (b) collinear *coherent* beam geometry, showing input and output *amplitudes*; (c) collinear *coherent* beam geometry, showing input and output *intensities*.

Fig. 15.2 Illustration of optical addition utilizing mirrors: (a) angularly multiplexed *incoherent* beam geometry; (b) angularly multiplexed *coherent* beam geometry.

Fig. 15.3 Illustration of optical multiplication utilizing a medium with variable transparency.

Fig. 15.4 Fundamental principles of spatial light modulator function: (a) block diagram of the principal functions of an optically-addressed spatial light modulator, including the detection, functional implementation, and modulation functions; (b) schematic diagram of an $N \times N$ array of spatial light modulator pixels, in which three pixels are shown in different transmission states; (c) expanded view of the pixel array, showing an incomplete fill factor within each pixel; (d) expanded view of a single pixel within the array, illustrating one possible pixel configuration that incorporates two detector elements D_1 and D_2 , control electronics for impedance matching and functional implementation, and two modulator elements, shown here in different transmittance states.

Fig. 15.5 Examples of monolithically-integrated spatial light modulators. The chosen examples incorporate photodetectors, control circuitry, and multiple quantum well modulators within each pixel on a single gallium arsenide (*GaAs*) substrate. In (a), the control electronics and photodetector elements are fabricated following the photolithographic definition and physical isolation of the modulator elements, while in (b) a buffer (isolation) layer is used to allow fabrication and interconnection of all of the elements without chemical or ion beam etching.

Fig. 15.6 Example of a hybrid spatial light modulator, in which the photodetectors and control electronics are fabricated on a silicon substrate, and the multiple quantum well modulator elements are fabricated on a gallium arsenide (*GaAs*) substrate. The two sets of devices are bump contacted on a pixel-by-pixel basis to provide parallel electrical continuity.

Fig. 15.7 VLSI layout of a generalizable silicon-based spatial light modulator structure: (a) neuron pixel layout; (b) photograph of a single neuron unit in VLSI implementation, with probe pads substituted for the two detectors (bottom) and for contact to the two modulation elements (top); (c) photograph of a 6×6 array of neuron units on a VLSI chip that incorporates additional test circuitry.

Fig. 15.8 A simplified holographic recording configuration: case of plane wave signal and

reference beams, and a *thin* holographic recording medium: (a) recording, and (b) reconstruction with a plane wave readout beam.

Fig. 15.9 A point-to-point interconnection system, using a holographic optical element (HOE) for interconnection routing, and lenses as position-to-angle and angle-to-position encoders. In this example, the holographic optical element effectively performs an input angle to output angle transformation, such that light emitted (or transmitted) at point p_1 in the input plane (P_1) is detected at point p_2 in the output plane (P_2).

Fig. 15.10 Volume holographic recording with plane wave signal and reference beams; (a) recording, and (b) reconstruction, showing the elimination of the higher diffracted orders.

Fig. 15.11 The angular alignment sensitivity of a volume holographic optical element, as a function of the dimensionless Q -parameter defined in the text. The grating strength for all of the curves (3.14 radians) is optimized to produce 100% diffraction efficiency in the limit of large Q (Bragg diffraction regime), and is not optimized for low Q gratings. Note that the diffraction efficiency is essentially independent of angle for low Q gratings, and is very strongly peaked at the Bragg angle (7.5 degrees in this case) for high Q gratings.

Fig. 15.12 The diffraction efficiency of thin (Raman-Nath diffraction regime) and thick (Bragg diffraction regime) holographic gratings as a function of the grating strength.

Fig. 15.13 Schematic representation of a 4 input, 4 output holographic interconnection, showing 4 coherent input beams x_1 - x_4 and 4 coherent recording beams y_1 - y_4 , each of which corresponds to a desired output y'_1 - y'_4 . In (a), the sets $\{x_j\}$ and $\{y_i\}$ interfere within the volume holographic medium, recording the desired interconnection diffraction gratings. In (b), a new set of input beams $\{x_j\}$ illuminates the volume holographic medium, reading out the weighted interconnection pattern and forming appropriately weighted sums at each of the outputs $\{y'_i\}$.

Fig. 15.14 Schematic representation of the fan-out process for optical beams, for the case of one input and three outputs: (a) with beamsplitters ($BS_1 - BS_3$); (b) with a single holographic optical element containing three multiplexed (spatially superimposed) diffraction gratings.

Fig. 15.15 Schematic representation of the fan-in process for optical beams, for the case of three angularly distinct inputs and one combined collinear output beam: (a) with beamsplitters, showing the unavoidability of a throughput loss associated with the set of transmitted (and multiply reflected) beams; (b) with a single holographic optical element containing three multiplexed (spatially superimposed) diffraction gratings, showing an analogous throughput loss.

Fig. 15.16 Illustration of the generation of crosstalk in holographic optical interconnections due to beam degeneracy: recording/readout configuration. The input beams $\{x_j\}$ are assumed to have interfered within the volume holographic medium with the set of recording beams $\{y_i\}$, producing the desired set of interconnection gratings with weights w_{ij} . Illumination of the volume holographic medium with beam x_1 produces a 1 to 4 fanout into the output beams $\{y'_i\}$, as well as the zeroth order beam x'_1 . Due to the effects of beam degeneracy, power is also coupled into the zeroth order beams $x'_2-x'_4$, and crosstalk terms $\{c_i\}$ are introduced into the outputs.

Fig. 15.17 Illustration of the generation of crosstalk in holographic optical interconnections due to beam degeneracy: diffraction efficiency as a function of grating strength for the readout configuration of Figure 15.16. Shown are the depletion of the zero order beam x'_1 and the rise of the desired output beams y'_i , accompanied by a strong buildup of the cross-coupled beams $x'_2-x'_4$.

Fig. 15.18 Illustration of a surface-emitting laser diode source array [after Jewell, 1990]. In this example, the individual semiconductor laser diodes are isolated by chemical-assisted ion beam etching techniques, must be individually contacted, and emit *through* the *GaAs* substrate.

- Fig. 15.19** Schematic diagram of a photodarlington pair utilized as a high gain detector/amplifier combination.
- Fig. 15.20** Schematic diagram of a charge coupled device (CCD) photodetector array fabricated on a silicon substrate. Electrostatic potential wells are created by application of appropriate voltages to the three phase bias electrode structure, with electrical isolation provided by the gate oxide layer. Light incident through the transparent electrodes creates stored charge that can be transferred to an output signal terminal by proper sequential phasing of the bias voltages ($P_1 - P_3$).
- Fig. 15.21** Block diagram of a 1-D to 1-D photonic neural network, in which a one-dimensional neuron array is fully interconnected to a one-dimensional detector array by means of a two-dimensional interconnection mask.
- Fig. 15.22** Block diagram of a 2-D to 2-D photonic neural network, in which a two-dimensional neuron array is fully interconnected to a two-dimensional output array by means of a three-dimensional volume holographic optical interconnection mask. The input plane, output plane, and optional training plane are shown. Many variants of this geometry with similar properties are possible.
- Fig. 15.23** A single layer physical neural network with feedback, used to implement a multilayer recurrent functional network. The solid boxes indicate feedforward connections, and the broken boxes indicate lateral connections.
- Fig. 15.24** Incoherent/coherent technique for recording and reconstructing multiple holograms simultaneously, in which all solid lines represent mutually coherent beams, and all broken lines represent a separate set of mutually coherent beams: (a) recording; (b) reconstruction; and (c) holographic representation, in which each hologram represents the fanout from a given neuron unit.
- Fig. 15.25** Doubly angularly multiplexed volume holographic optical interconnection, designed to circumvent the effects of beam degeneracy. The mutually incoherent input

beams ($\{x_j\}$) are angularly multiplexed over j , as are the corresponding sets of output beams from the training plane ($\{\delta_i^{(j)}\}$) generated by the coherent sources S_j , to produce an angularly multiplexed fan-in at each summed output, thus yielding the neuron activation potentials $\{\rho_i\}$.

Fig. 15.26 Photonic architecture for neural network implementation that incorporates a parallel source array, double angular multiplexing, and incoherent/coherent recording and reconstruction: the Hebbian case is depicted.

Fig. 15.27 Photonic architecture for neural network implementation: recording configuration. This configuration implements the learning function in the photonic architecture of Figure 15.26. The sets of beams emitted from the source array (two are shown) interfere in the volume holographic medium to update the weights stored in the interconnection holograms.

Fig. 15.28 Photonic architecture for neural network implementation: reconstruction configuration. This configuration implements a single forward pass of the computing function in the photonic architecture of Figure 15.26. The lower set of beams acts as a set of reference beams, and generates a set of weighted output arrays that are imaged onto the detector array. Each stored hologram is reconstructed by a single neuron unit x_j , and fans out with appropriate weights to illuminate the detector array. The full set of reconstructed holograms sums within each pixel to yield the neuron activation potentials $\{\rho_i\}$.

Fig. 15.29 Generalized photonic architecture for neural network implementation, including provision for the generation of arbitrary training terms (δ_i).

Fig. 15.30 Schematic diagram of a dual-input, dual-output differential amplifier that effects a sigmoid-like transfer characteristic.

Fig. 15.31 Experimentally obtained transfer characteristics from the circuit shown in Figure 15.30, showing the output voltage in both channels (V_{out1} and V_{out2}) as a function of one input voltage (V_{in1}), with the other input voltage (V_{in2}) as a parameter.

Fig. 15.32 Optical layout for copying the entire contents of a three-dimensional volume holographic optical element (VHOE) into a second VHOE, utilizing a two-dimensional array of individually coherent, but mutually incoherent sources.

Fig. 15.33 The single pixel probability of error $P(Err)$ as a function of the number of photons detected within each pixel, for the cases of 100 and 1000 analog grey levels.

Fig. 15.34 Illustration of parallel-to-serial conversion in two-dimensional detector arrays such as the charge-coupled-device (CCD) array. Charge accumulated within each photosensitive region during exposure is transferred laterally by a set of row-parallel shift registers to a high speed parallel-to-serial shift register, which reads out the entire array one column at a time.

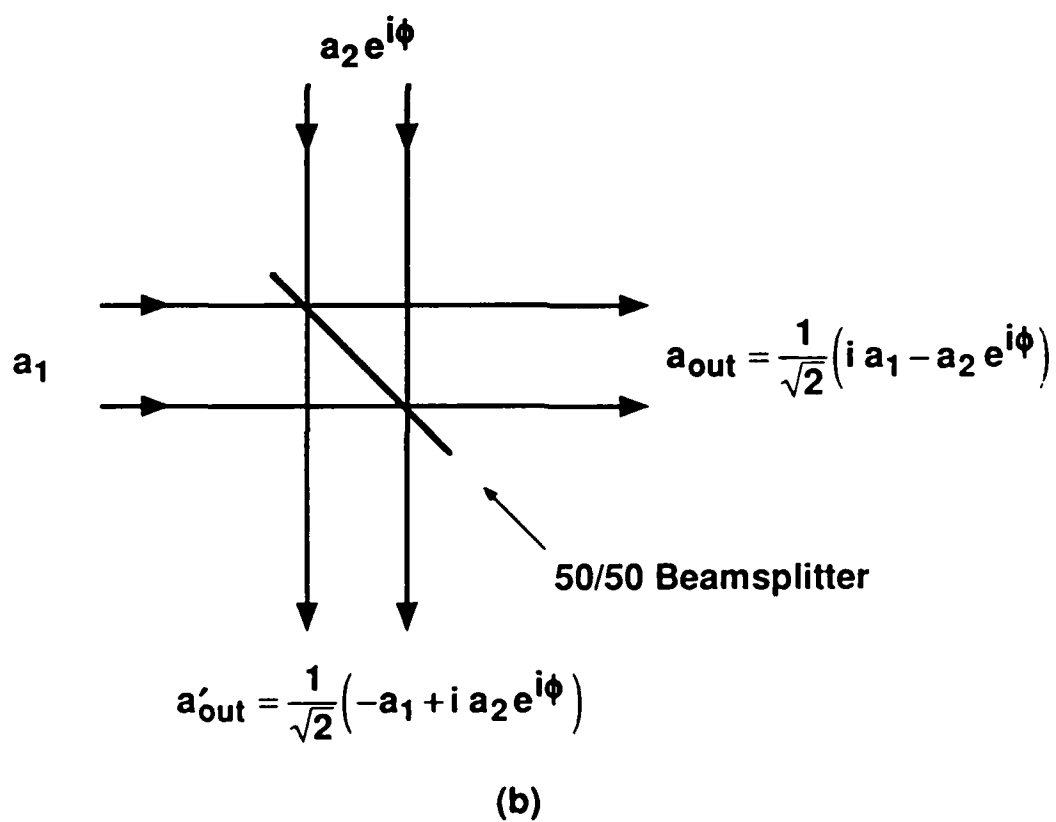
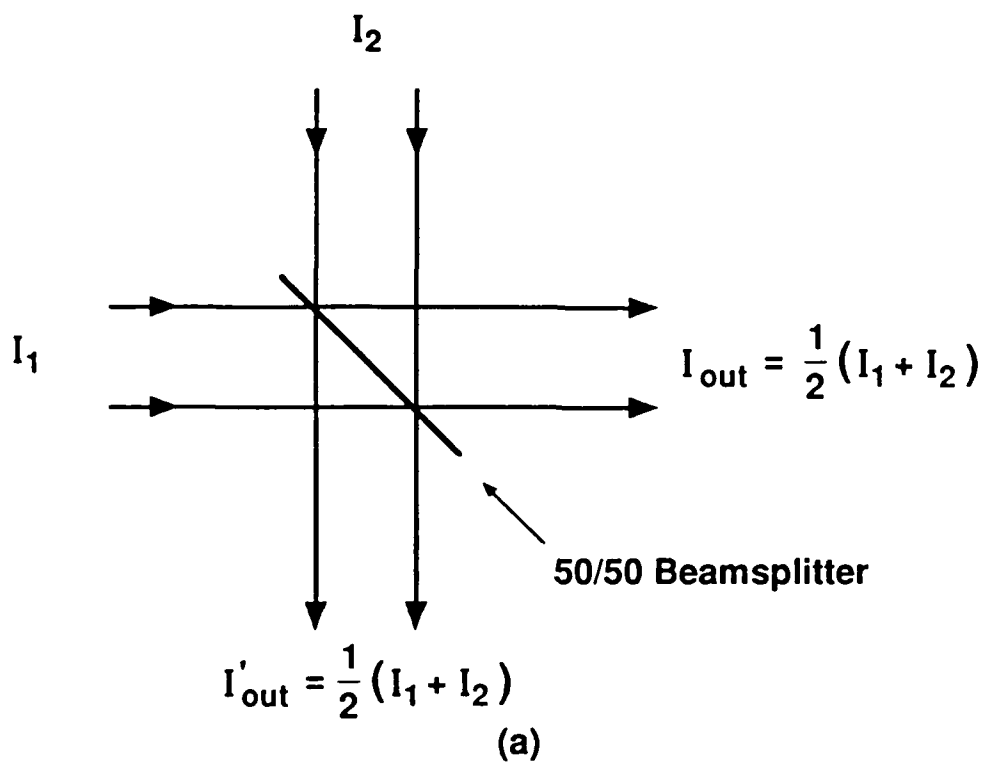


Figure 15.1 (a), (b)

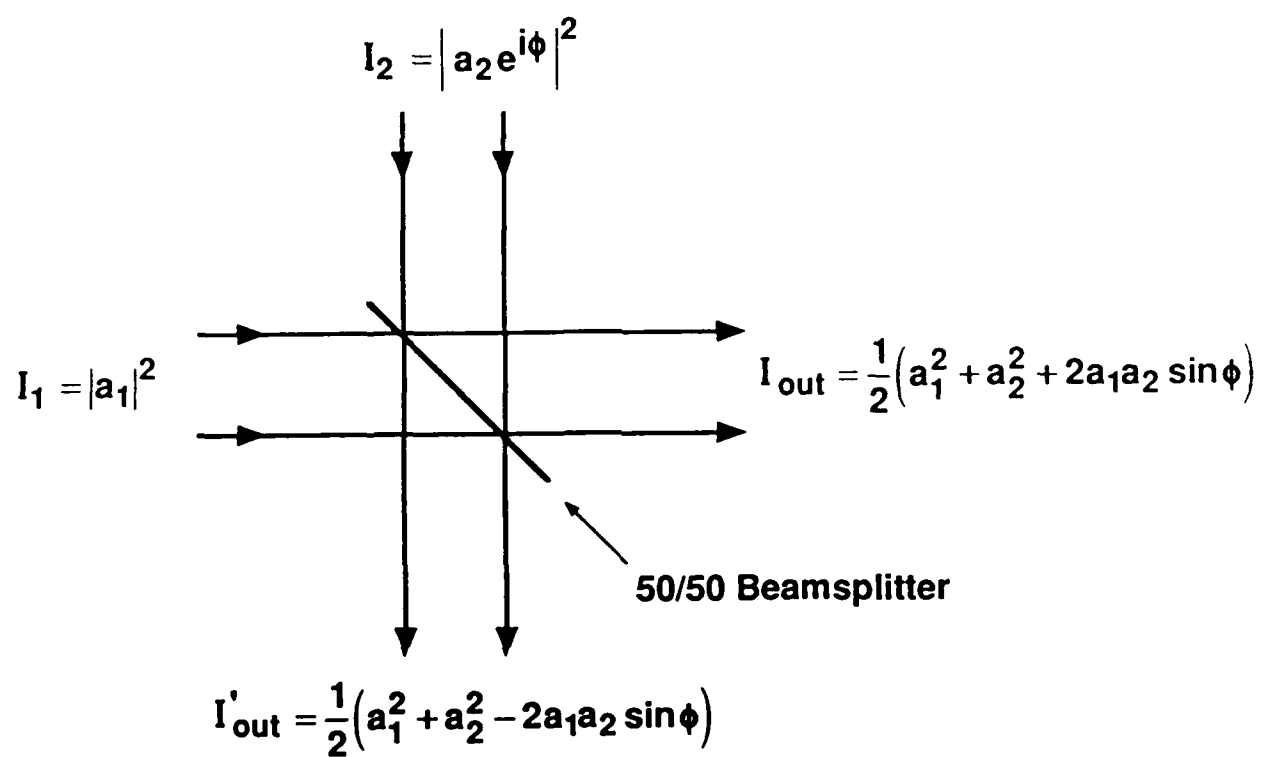
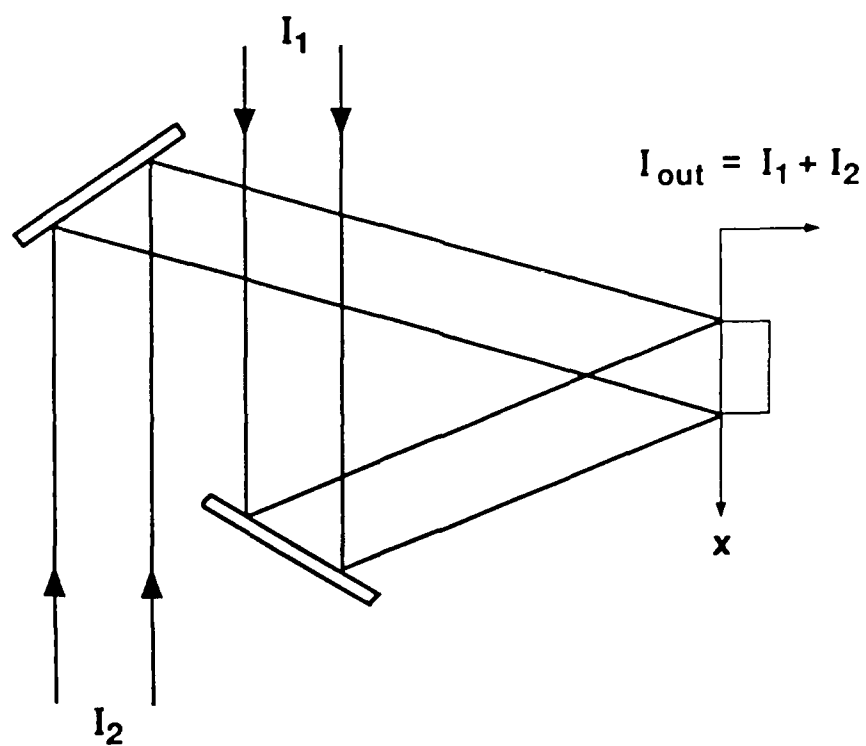


Figure 15.1 (c)

(a)



(b)

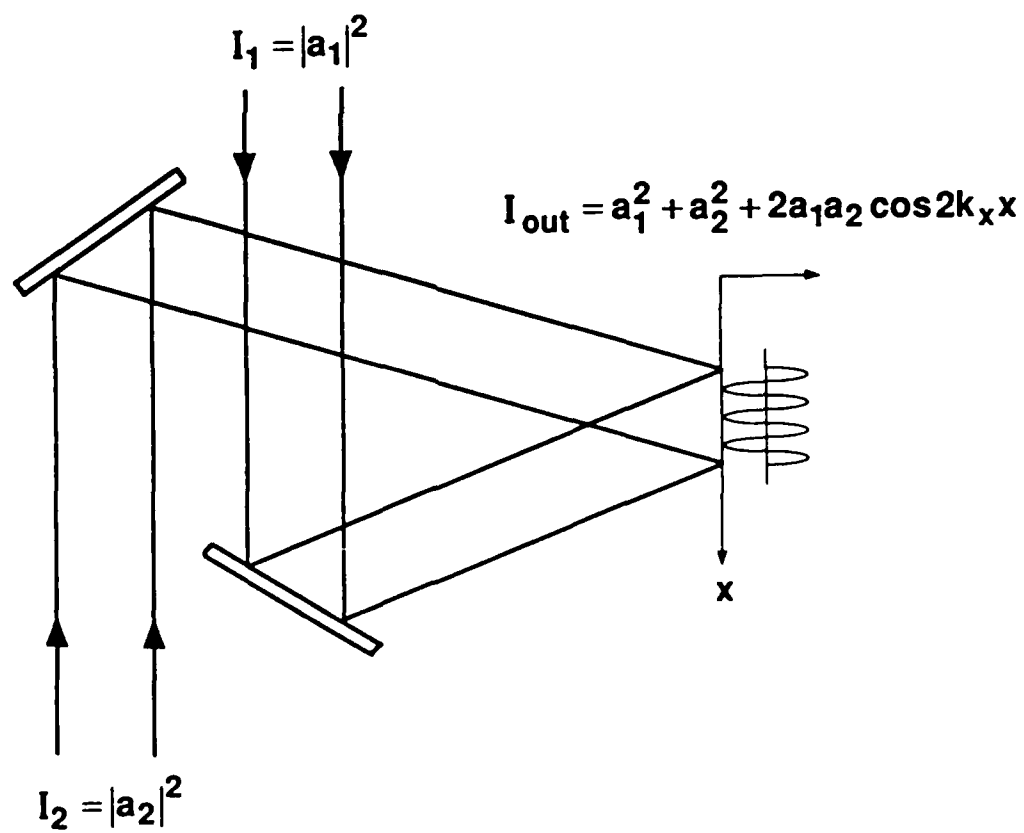


Figure 15.2

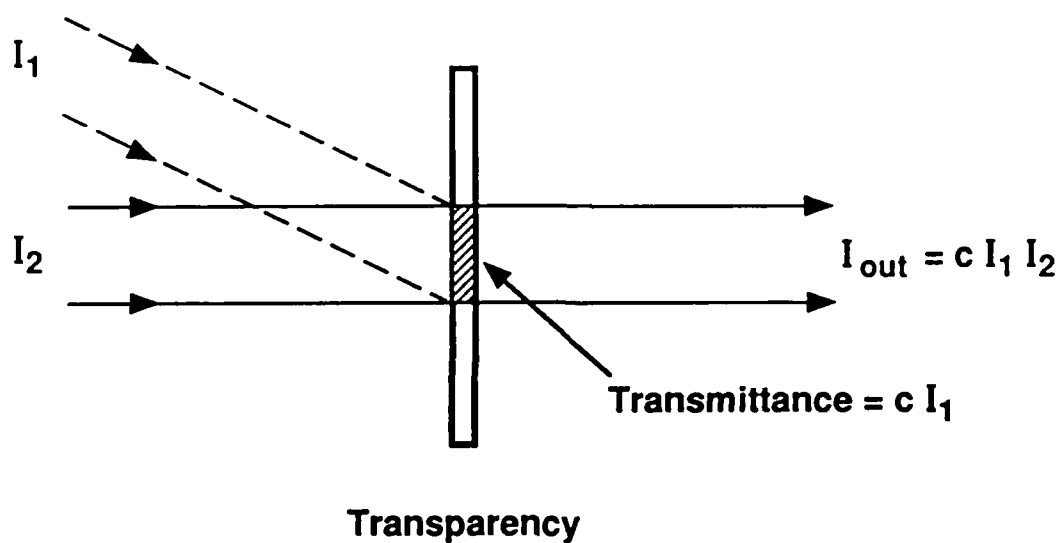
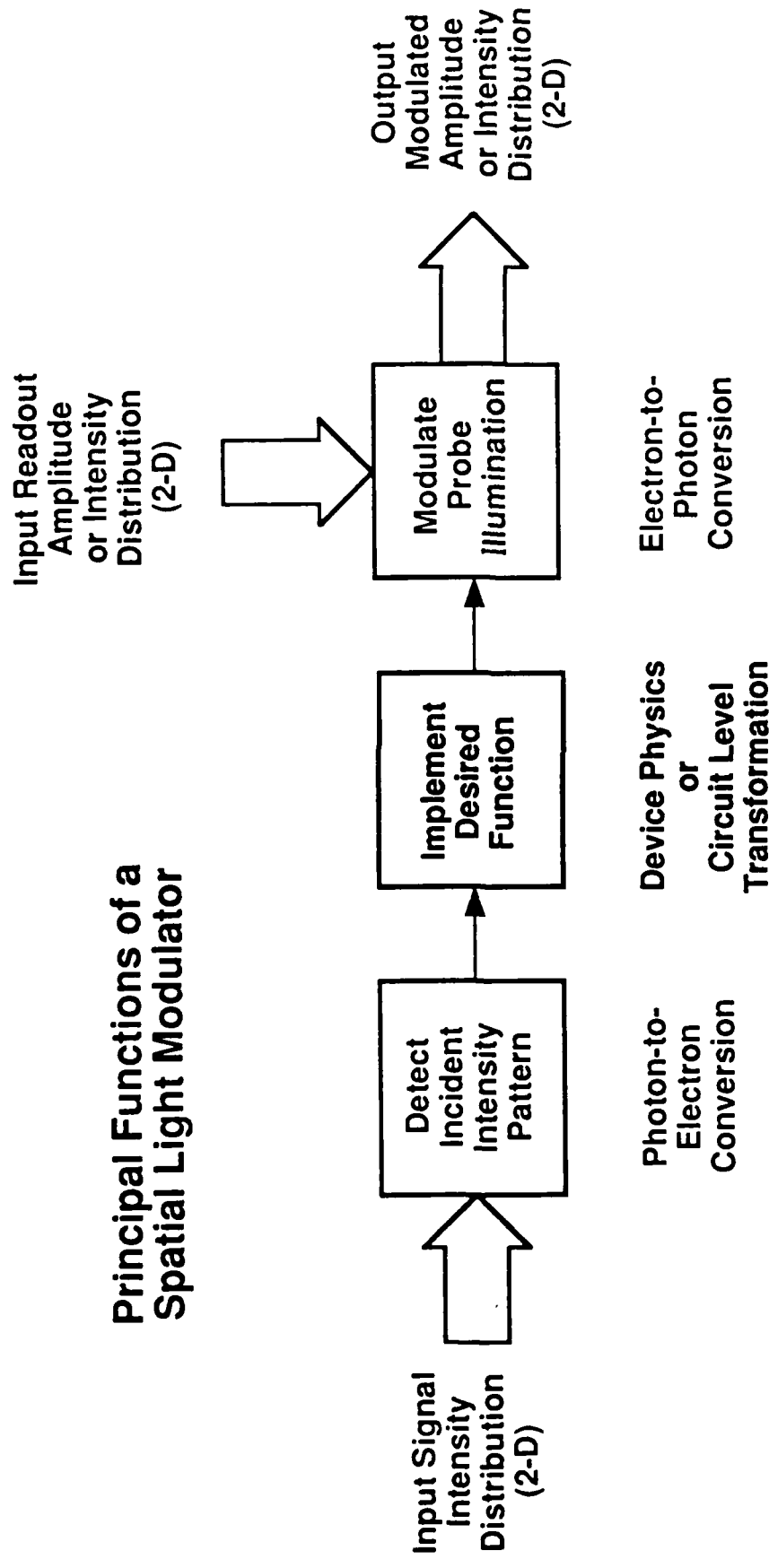


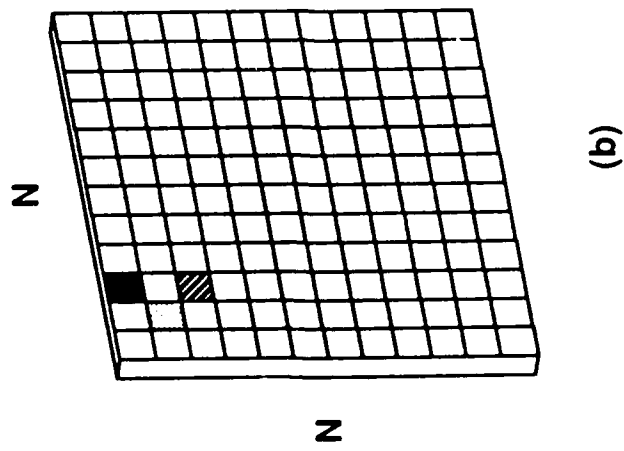
Figure 15.3

Principal Functions of a Spatial Light Modulator

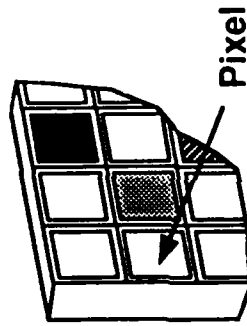


(a)

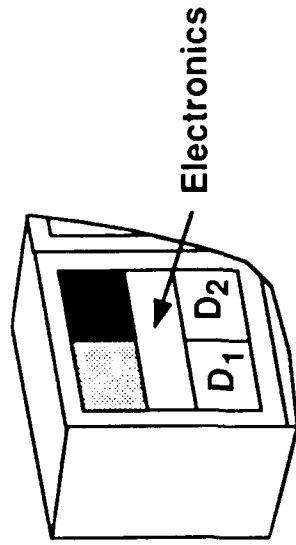
Figure 15.4(a)



(b)



(c)



(d)

Figure 15.4 (b), (c), (d)

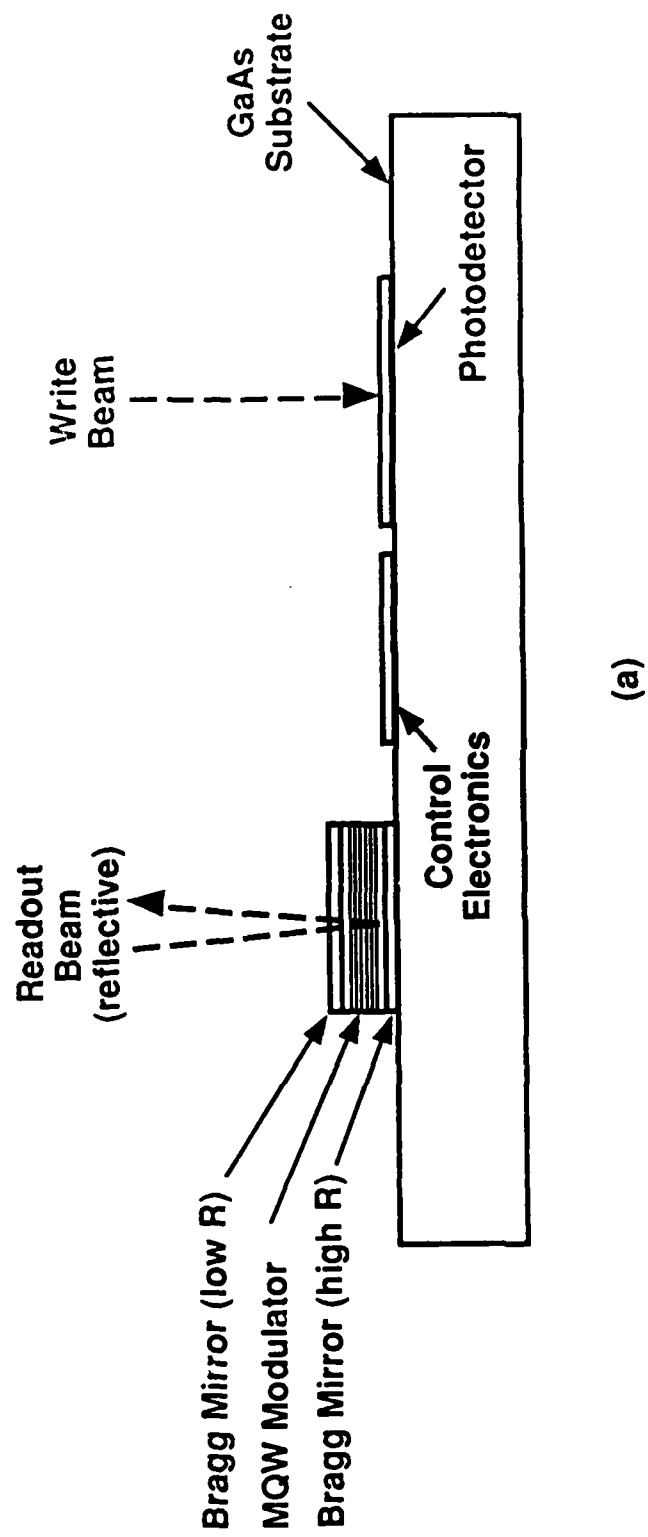


Figure 15.5(a)

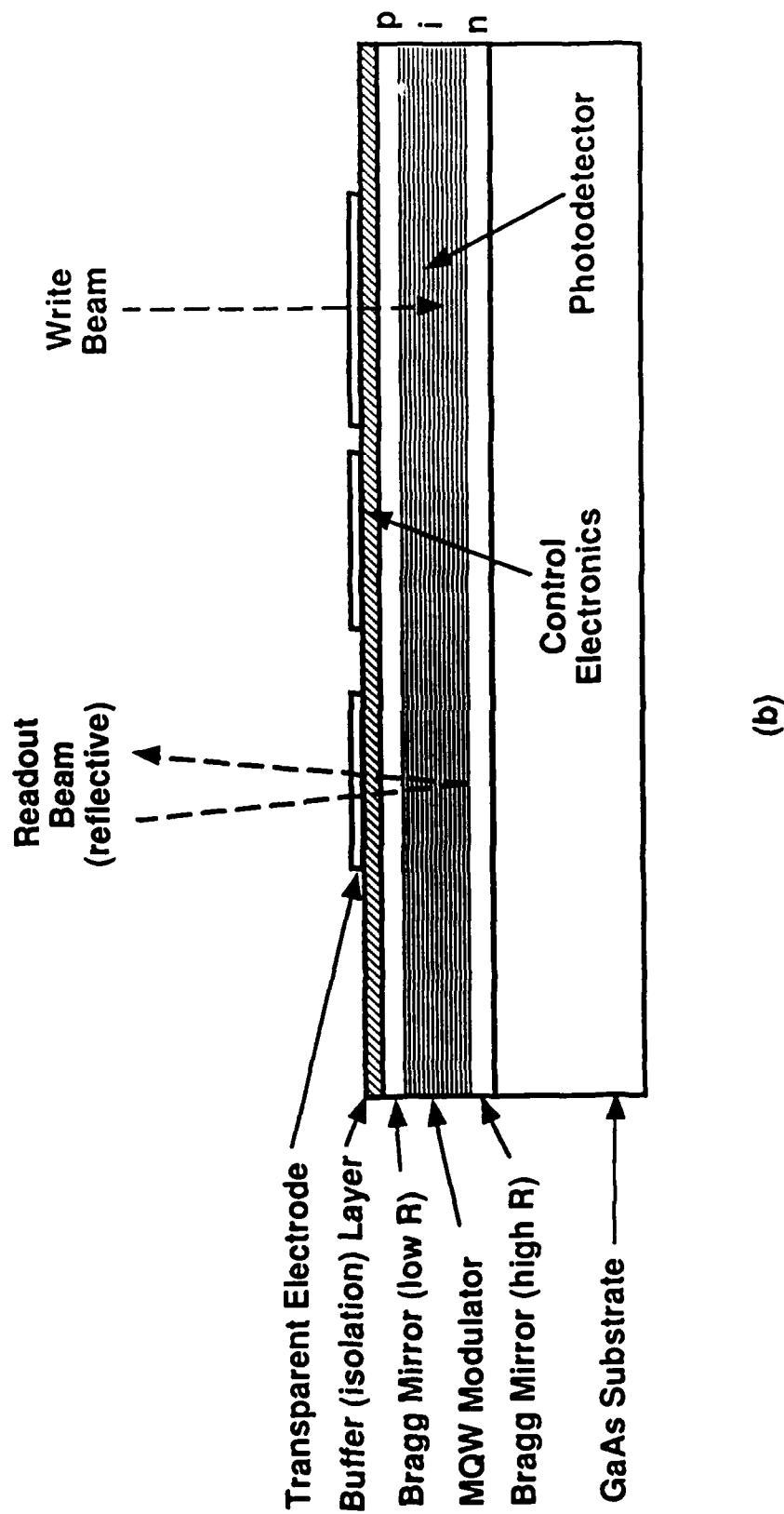


Figure 15.5(b)

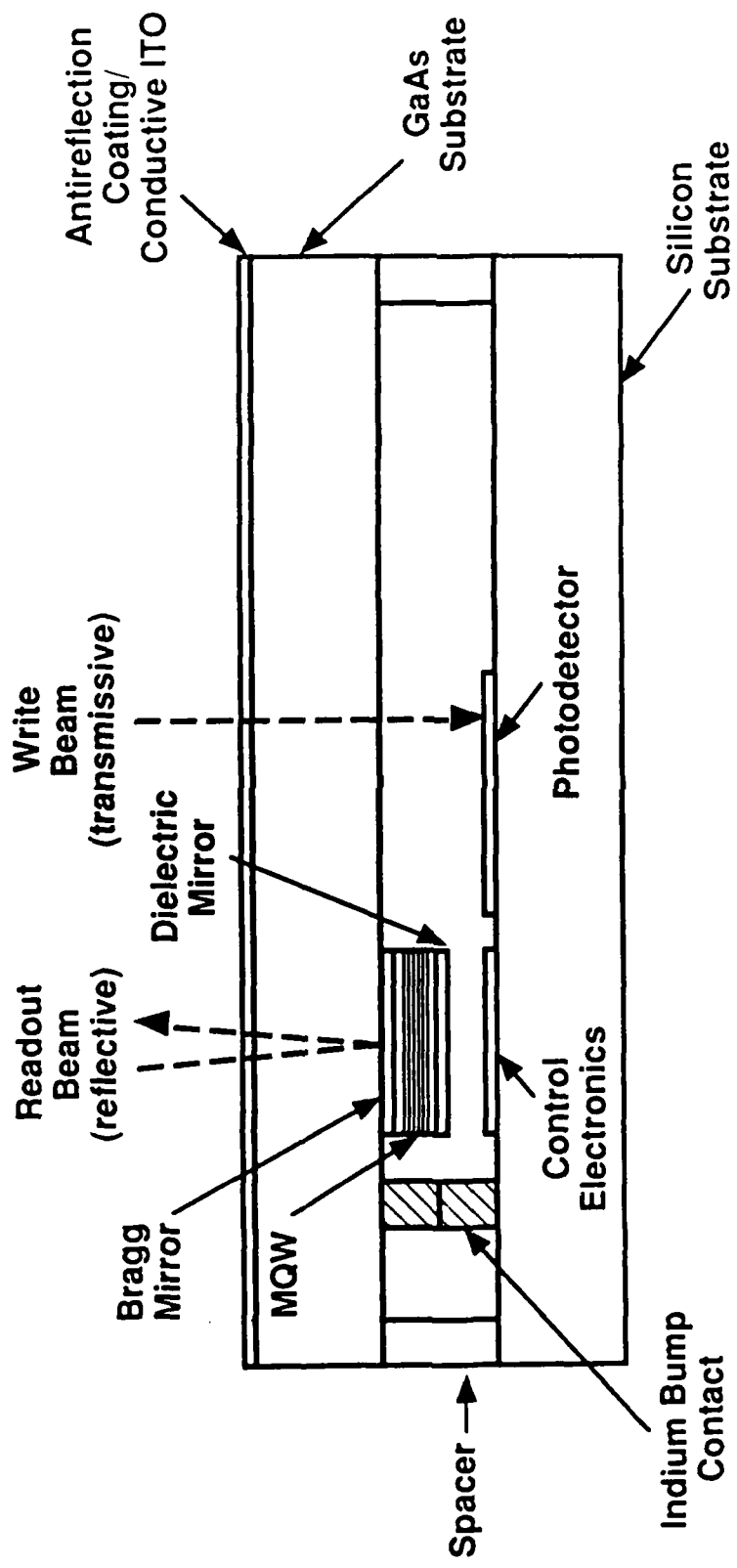


Figure 15.6

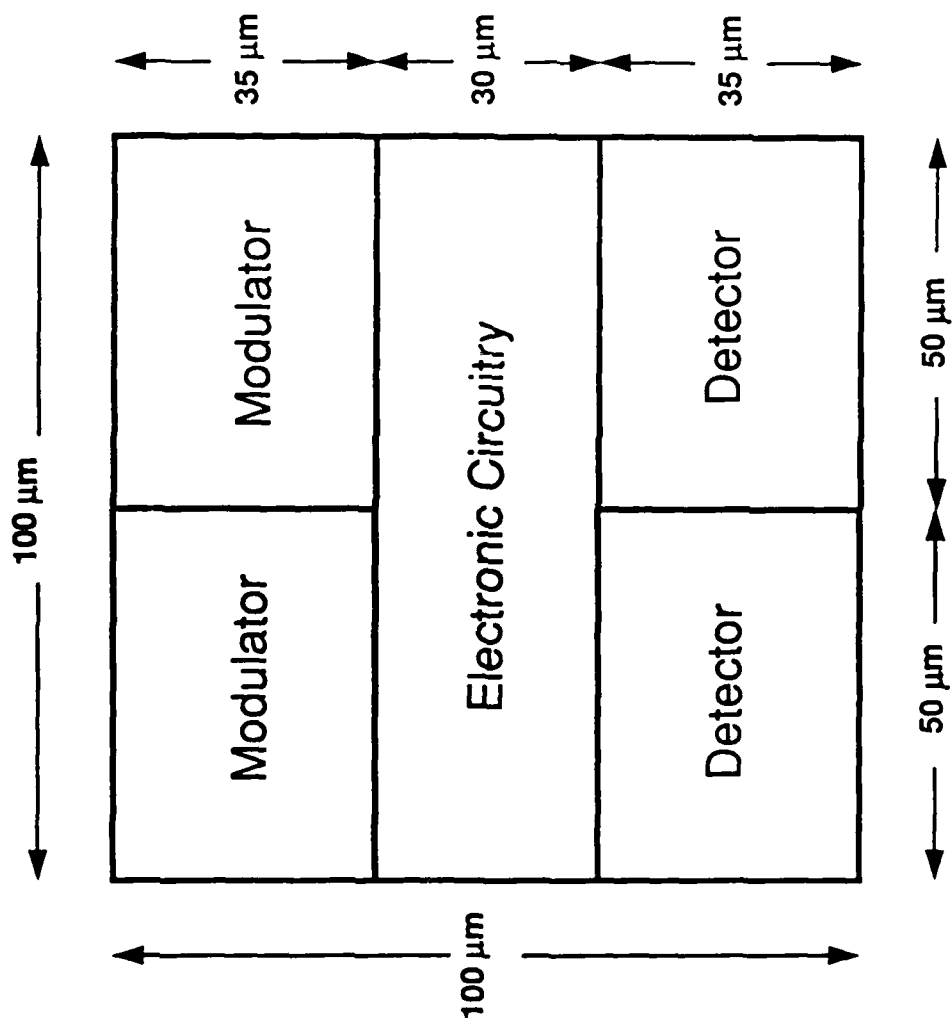


Figure 15.7 (a)

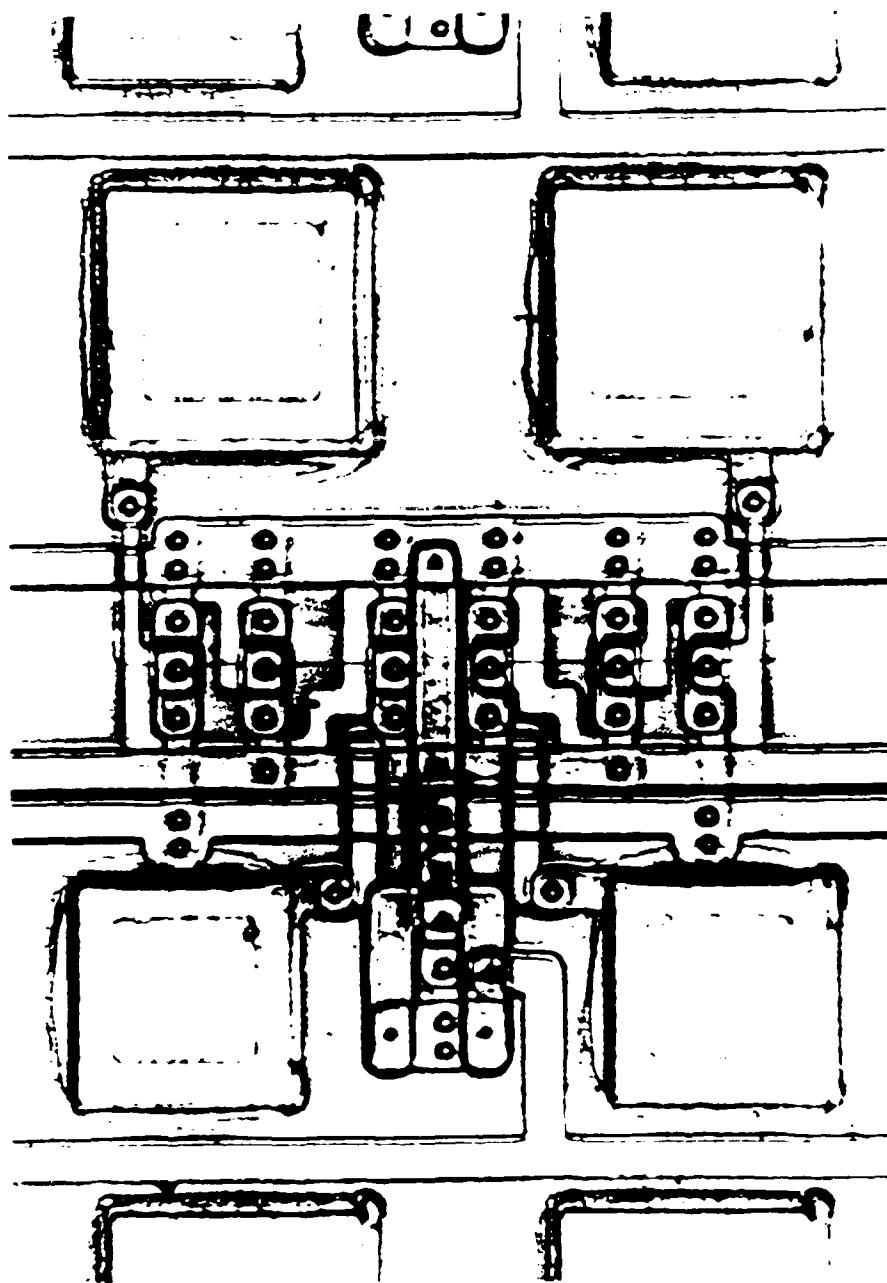


Figure 15.7 (b)

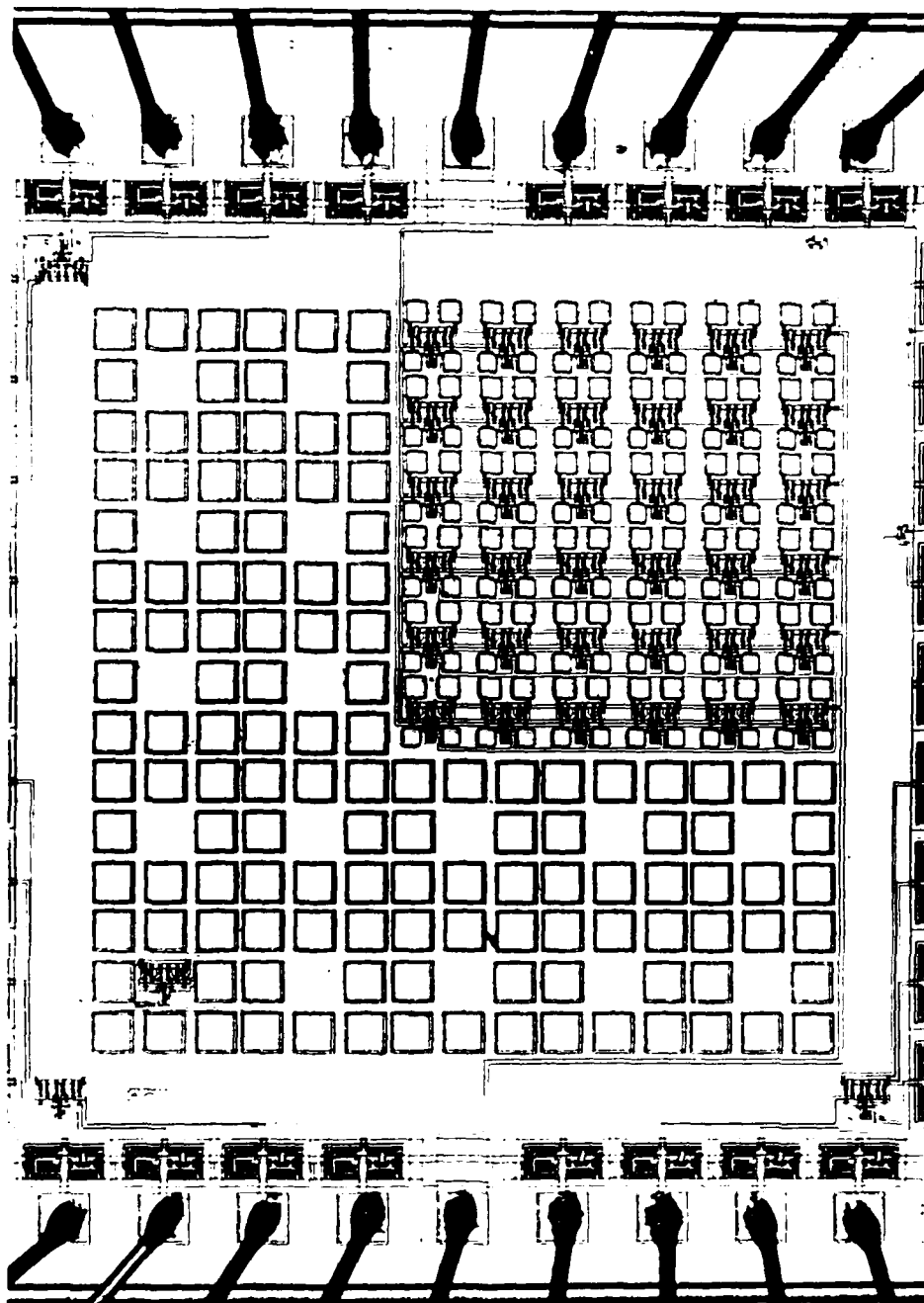
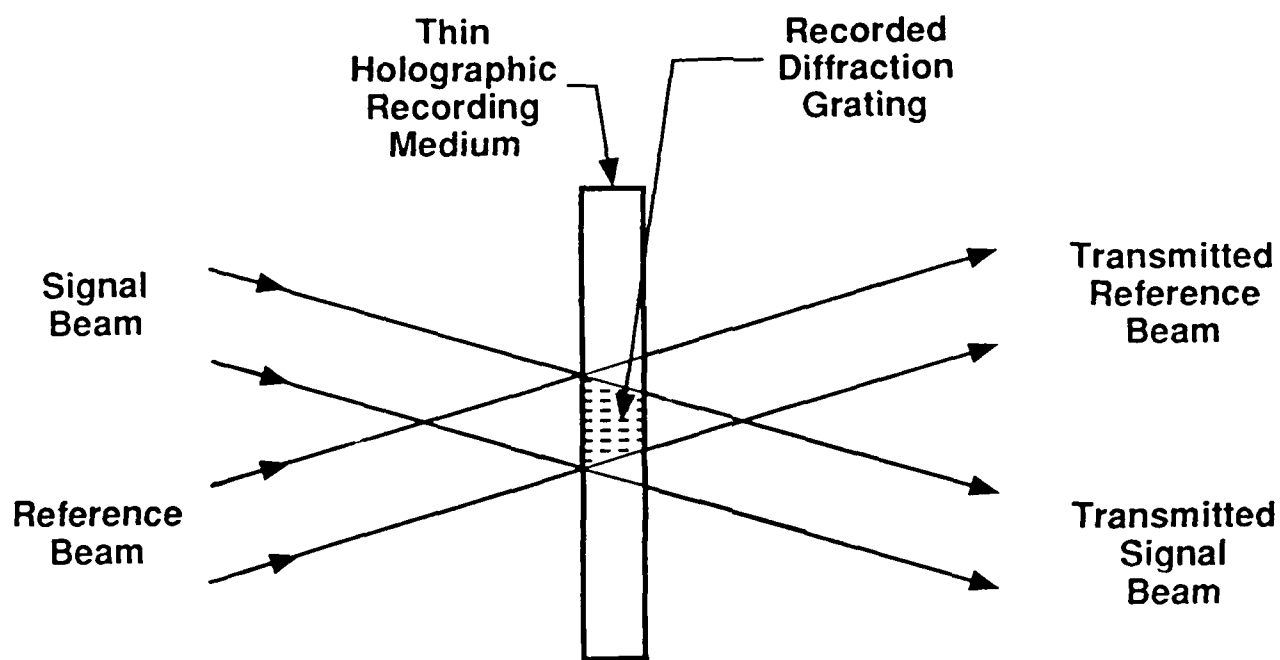
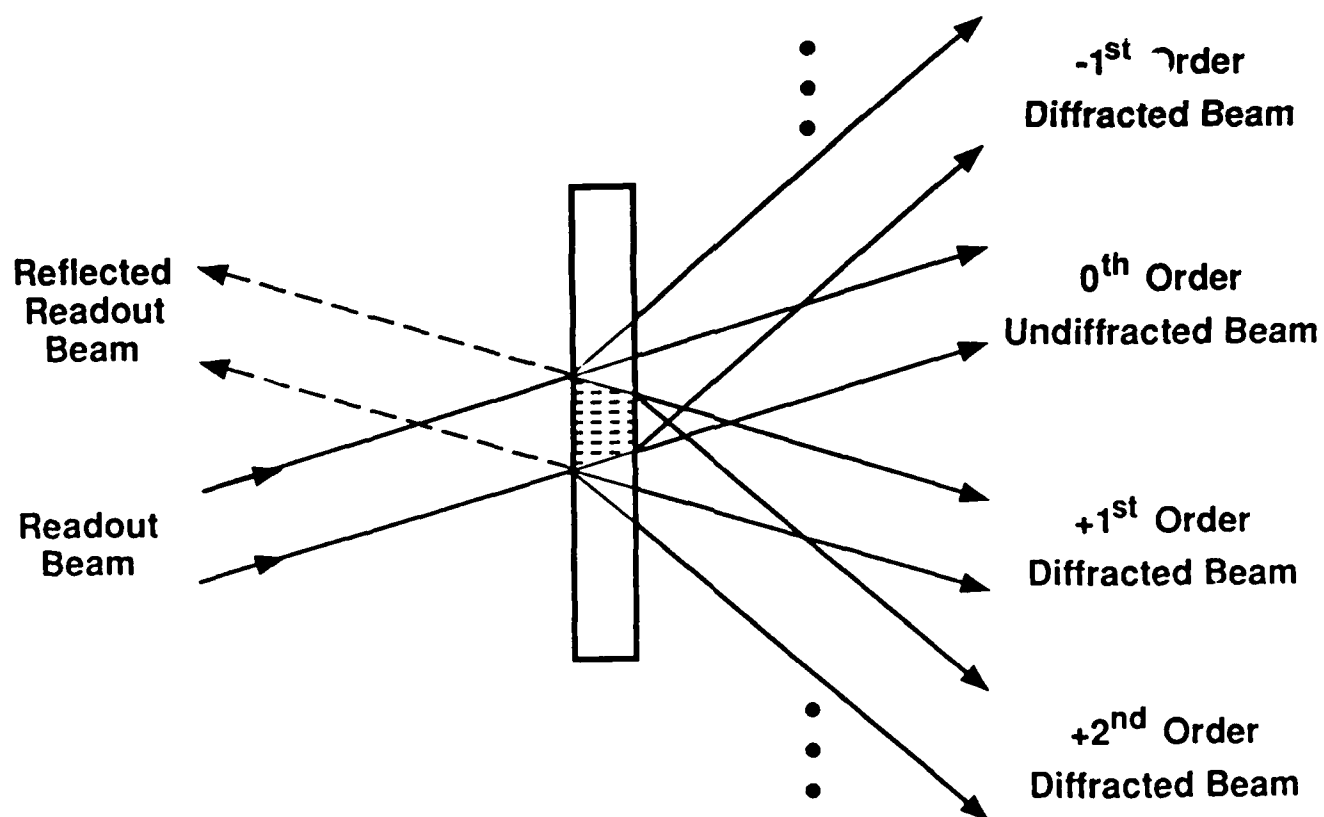


Figure 15.7 (c)



(a) Recording



(b) Reconstruction (readout)

Figure 15.8 (a), (b)

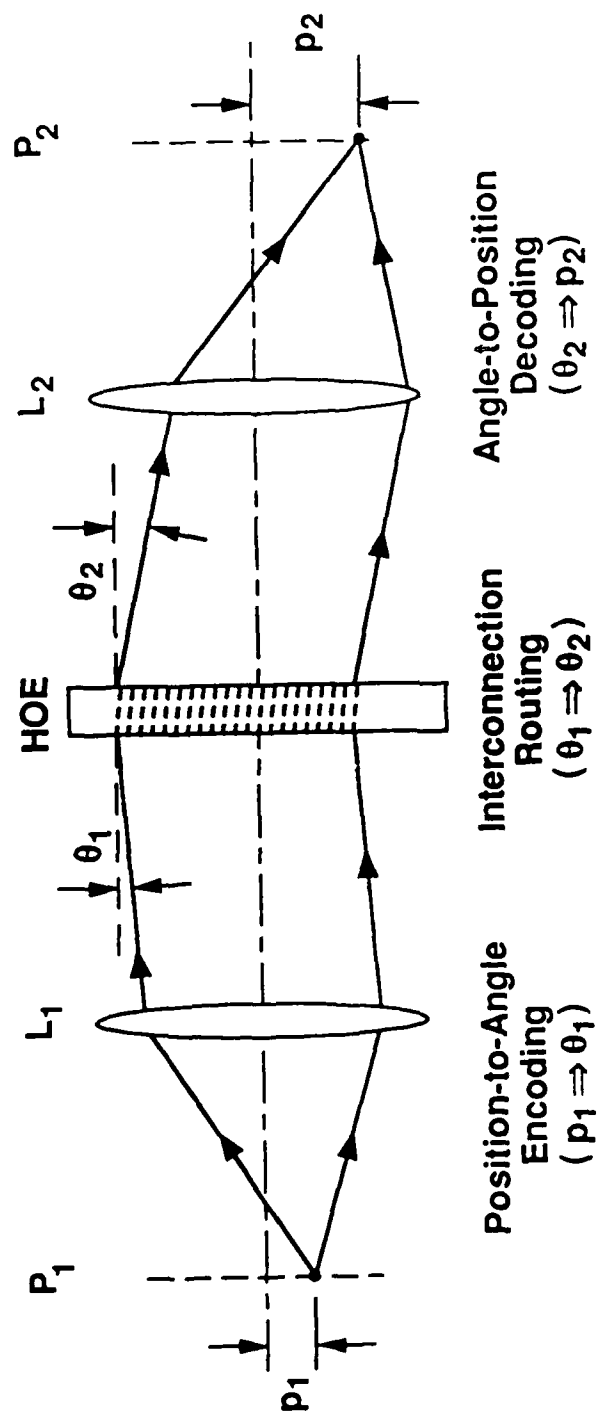
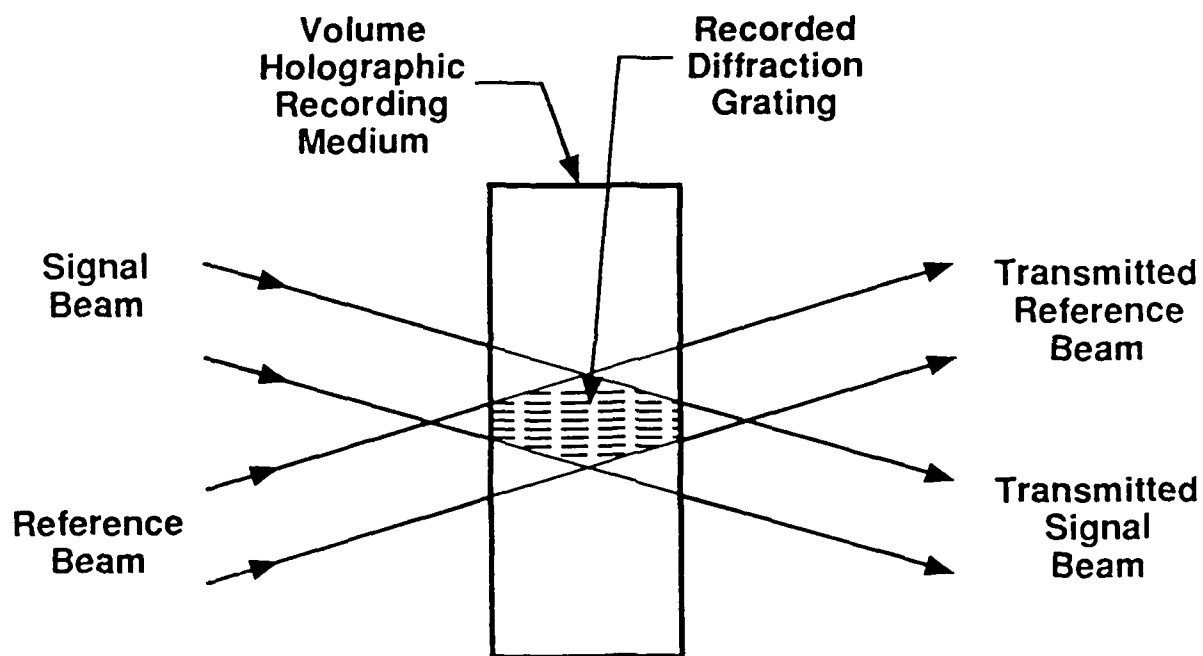
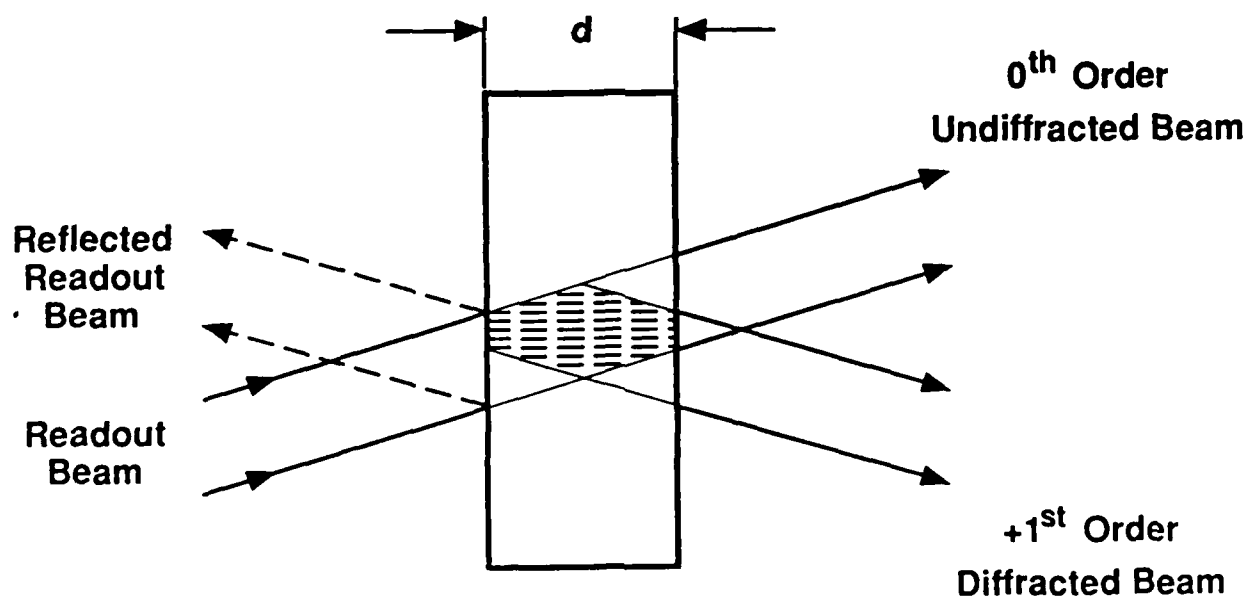


Figure 15.9



(a) Recording



(b) Reconstruction (readout)

Figure 15.10 (a), (b)

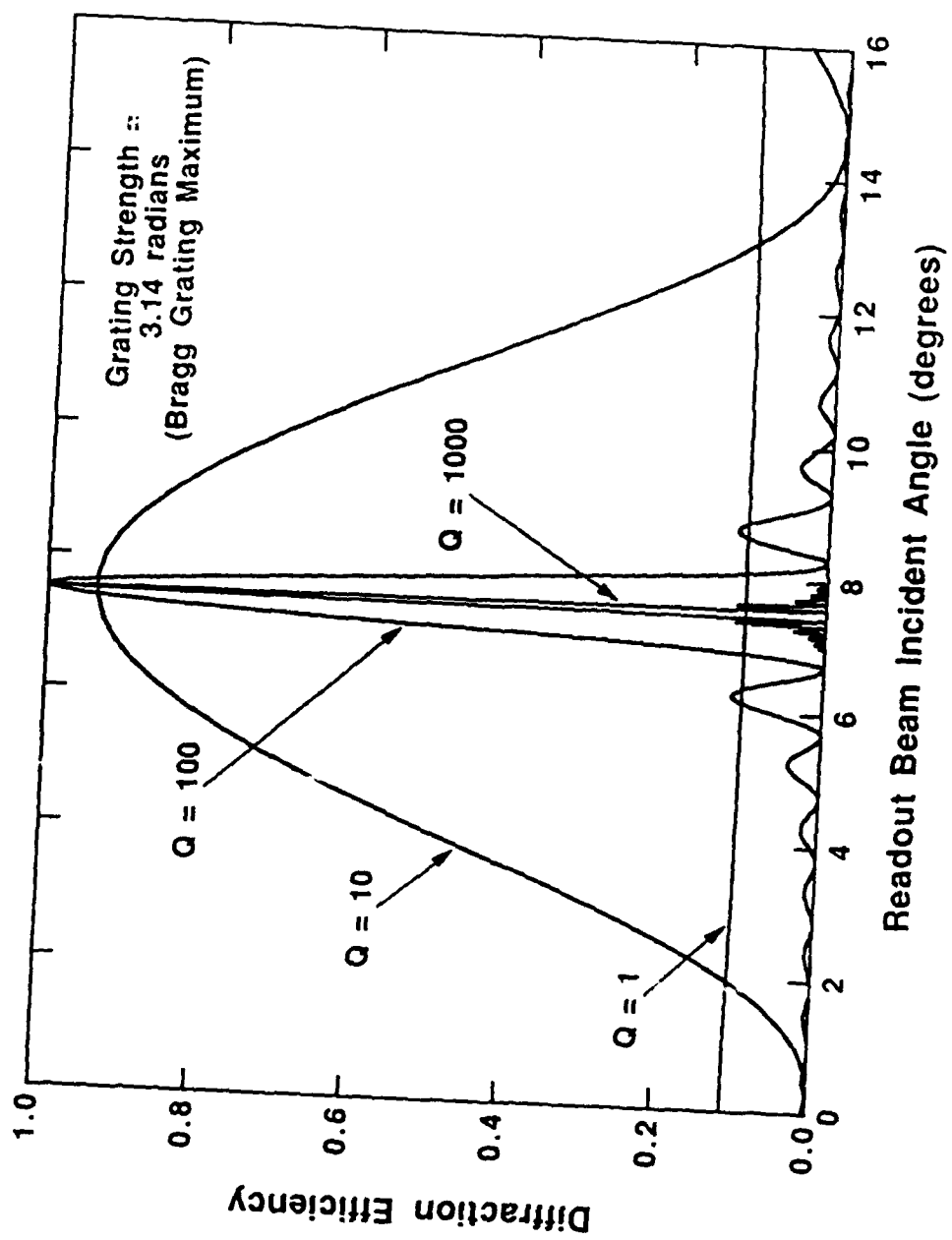


Figure 15.11

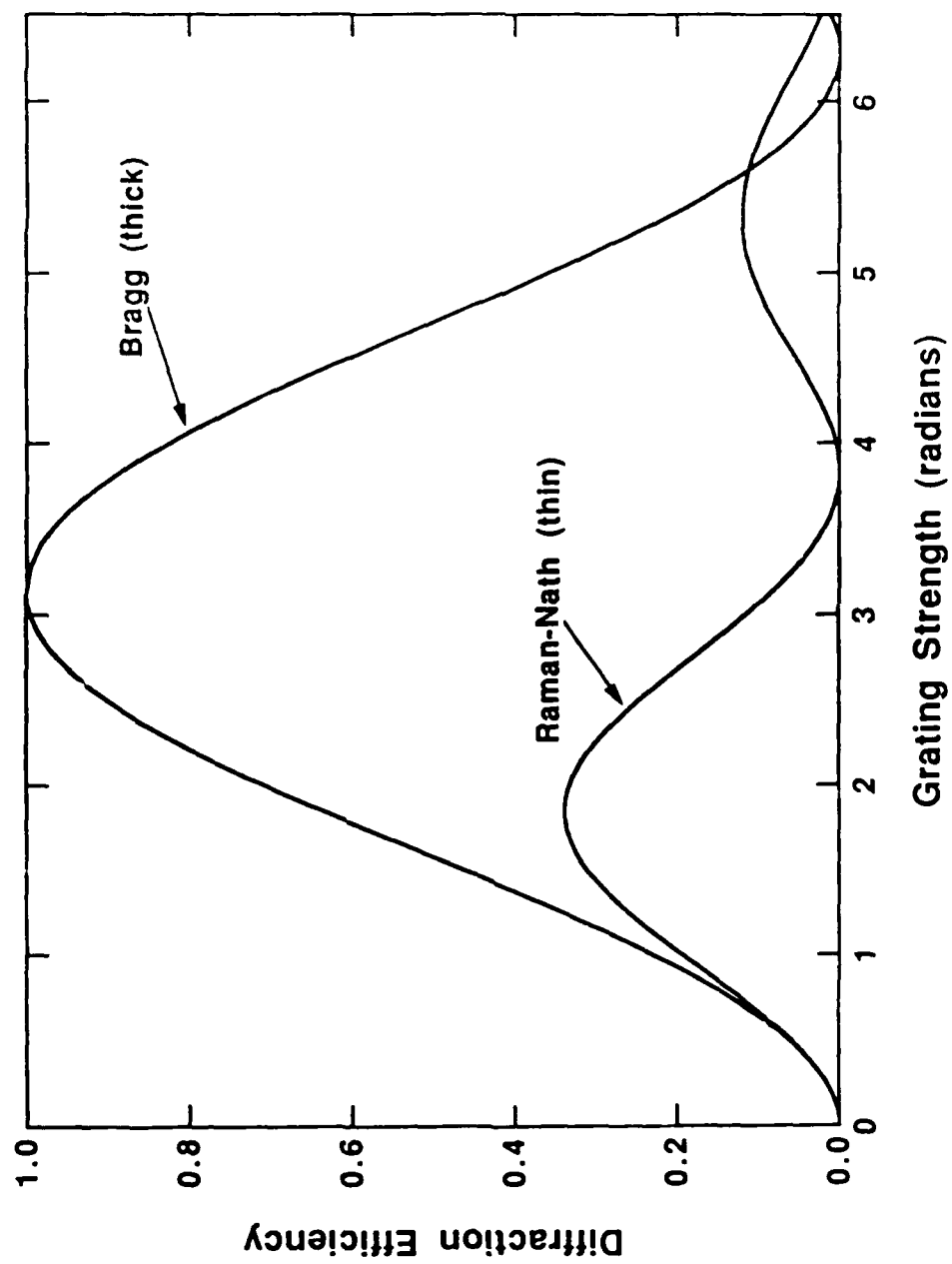


Figure 15.12

Recording Beams
for Desired Outputs

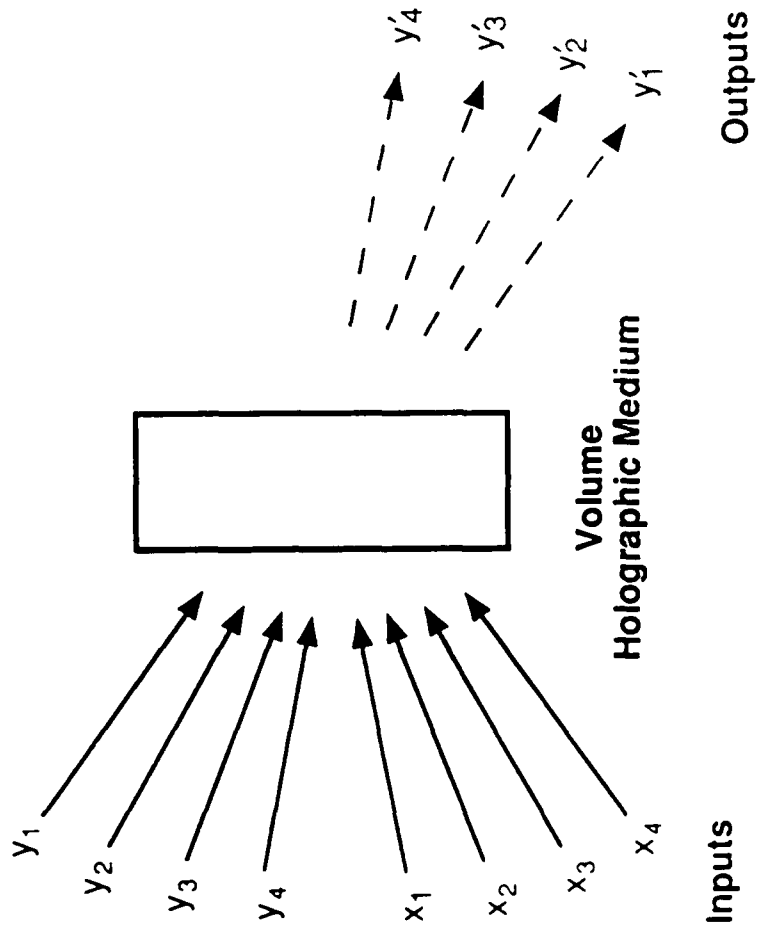


Figure 15.13(a)

Recording Beams
for Desired Outputs

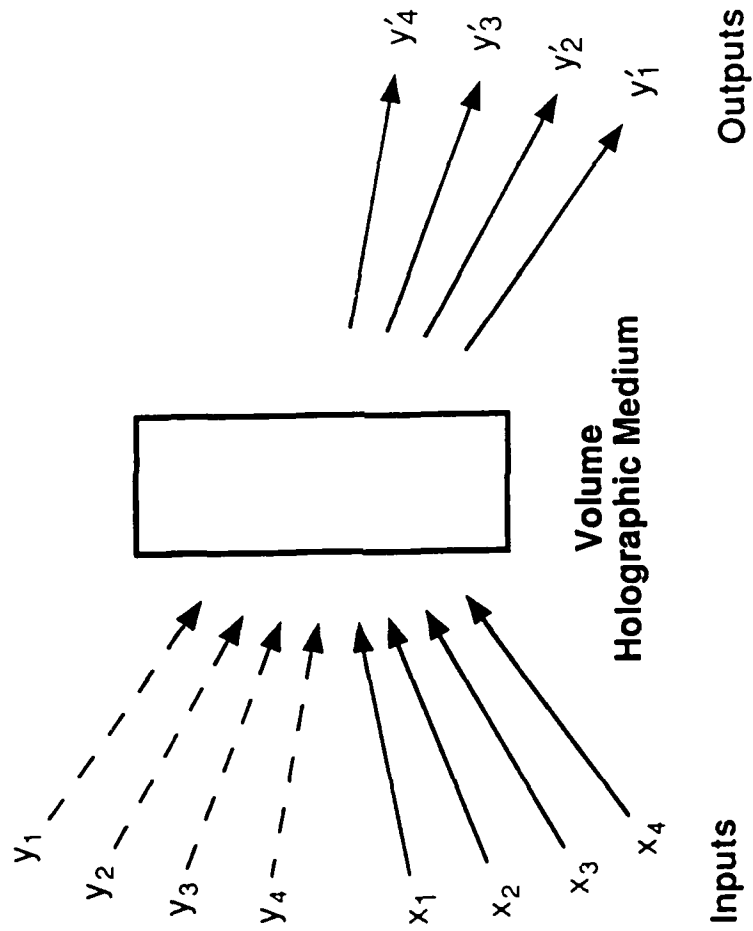


Figure 15.13(b)

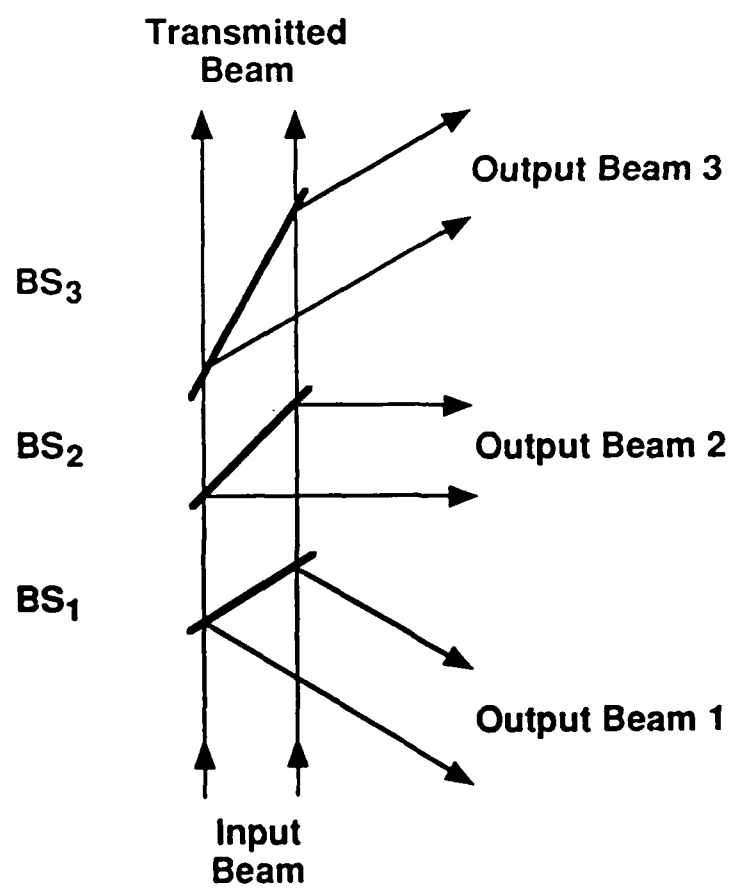


Figure 15.14 (a)

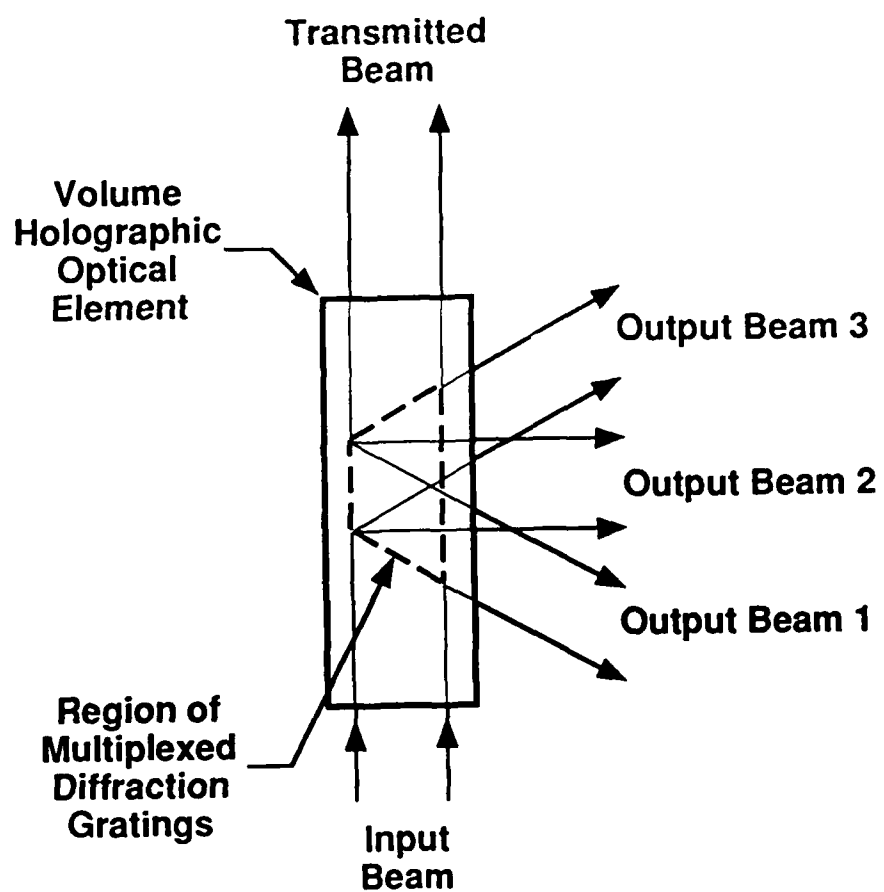


Figure 15.14 (b)

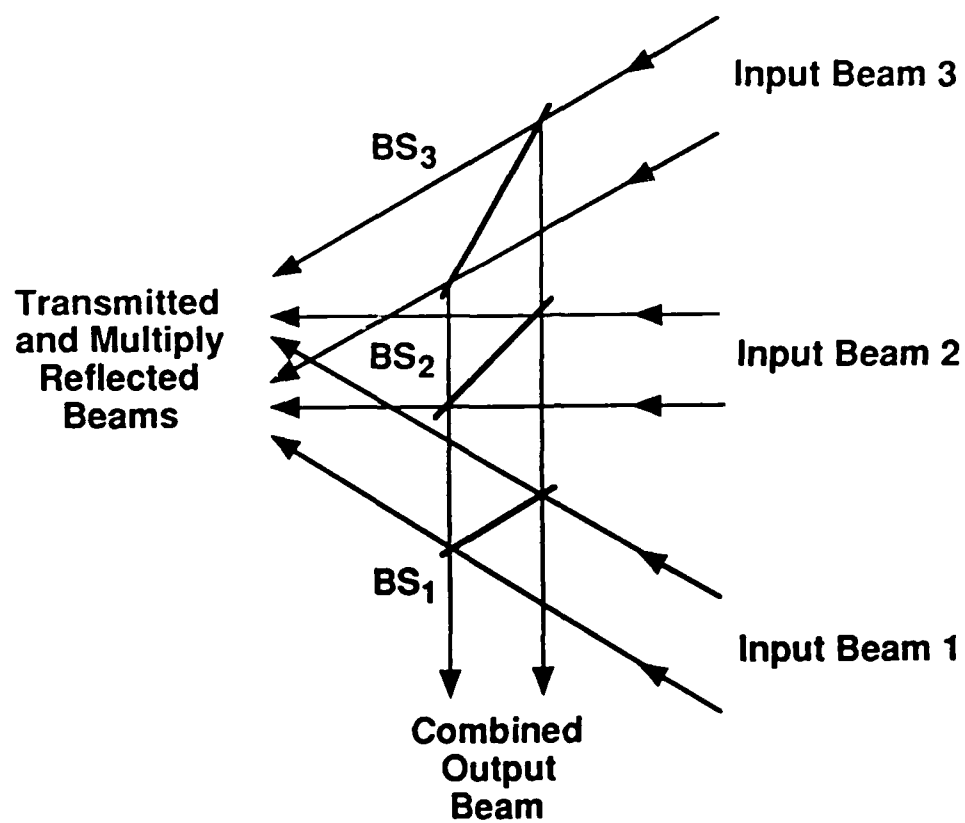


Figure 15.15 (a)

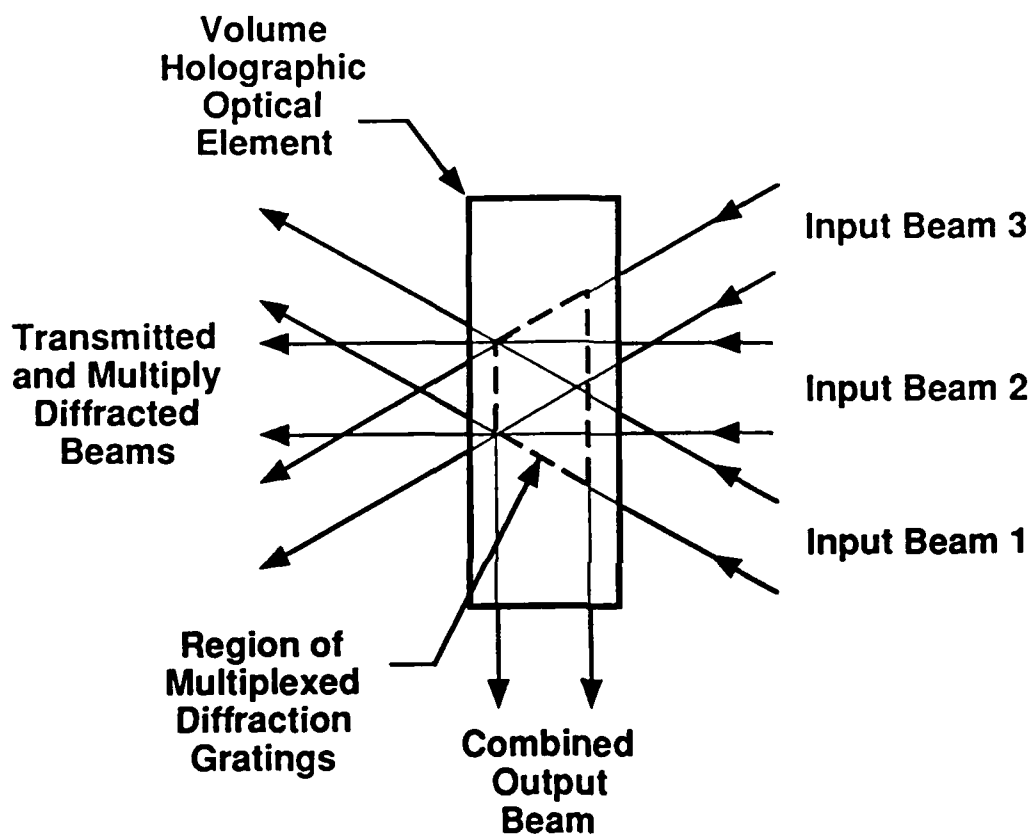


Figure 15.15 (b)

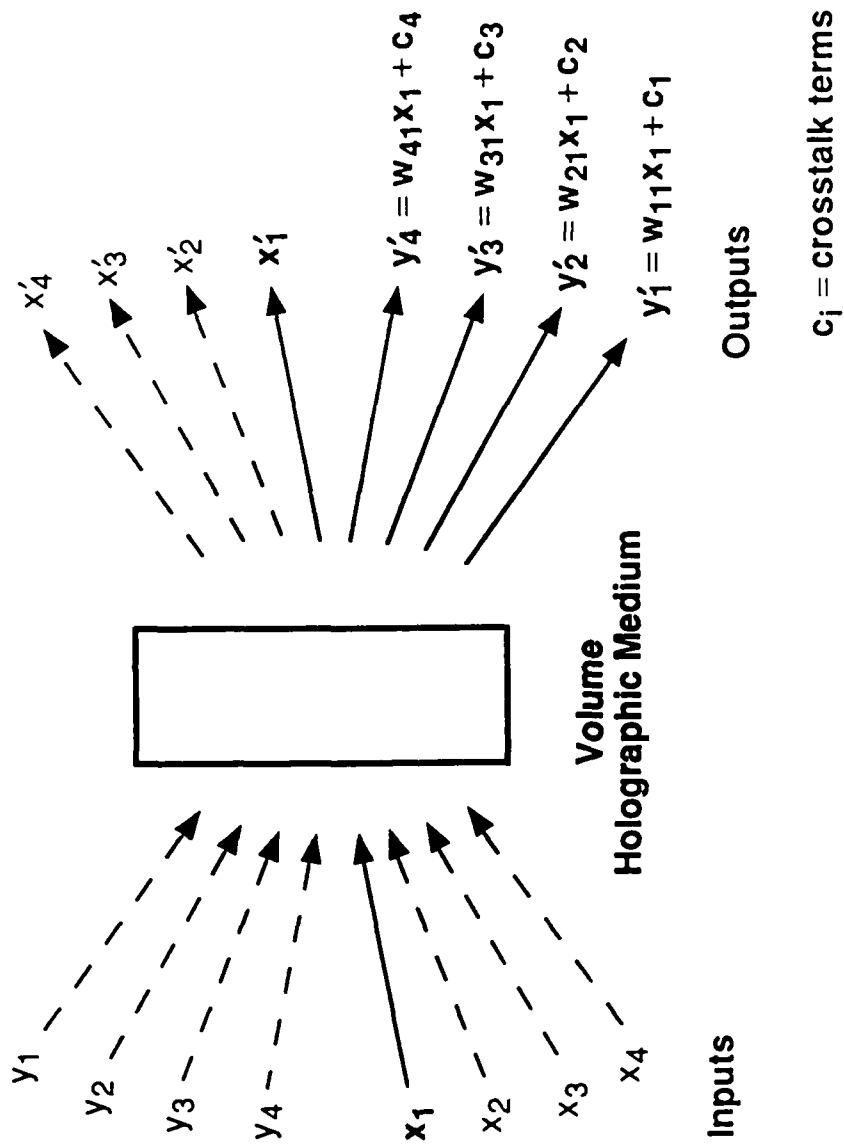


Figure 15.16

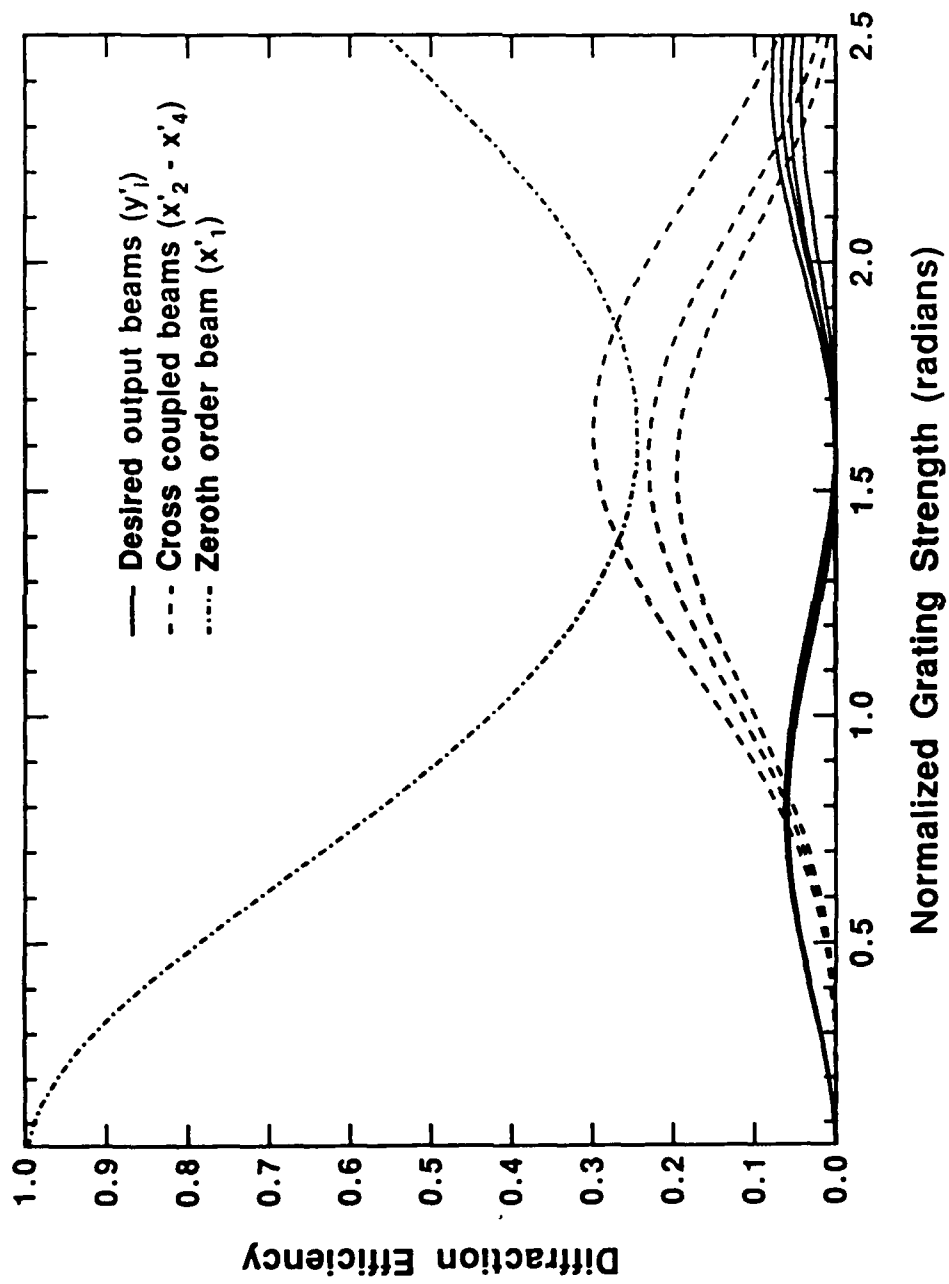


Figure 15.17

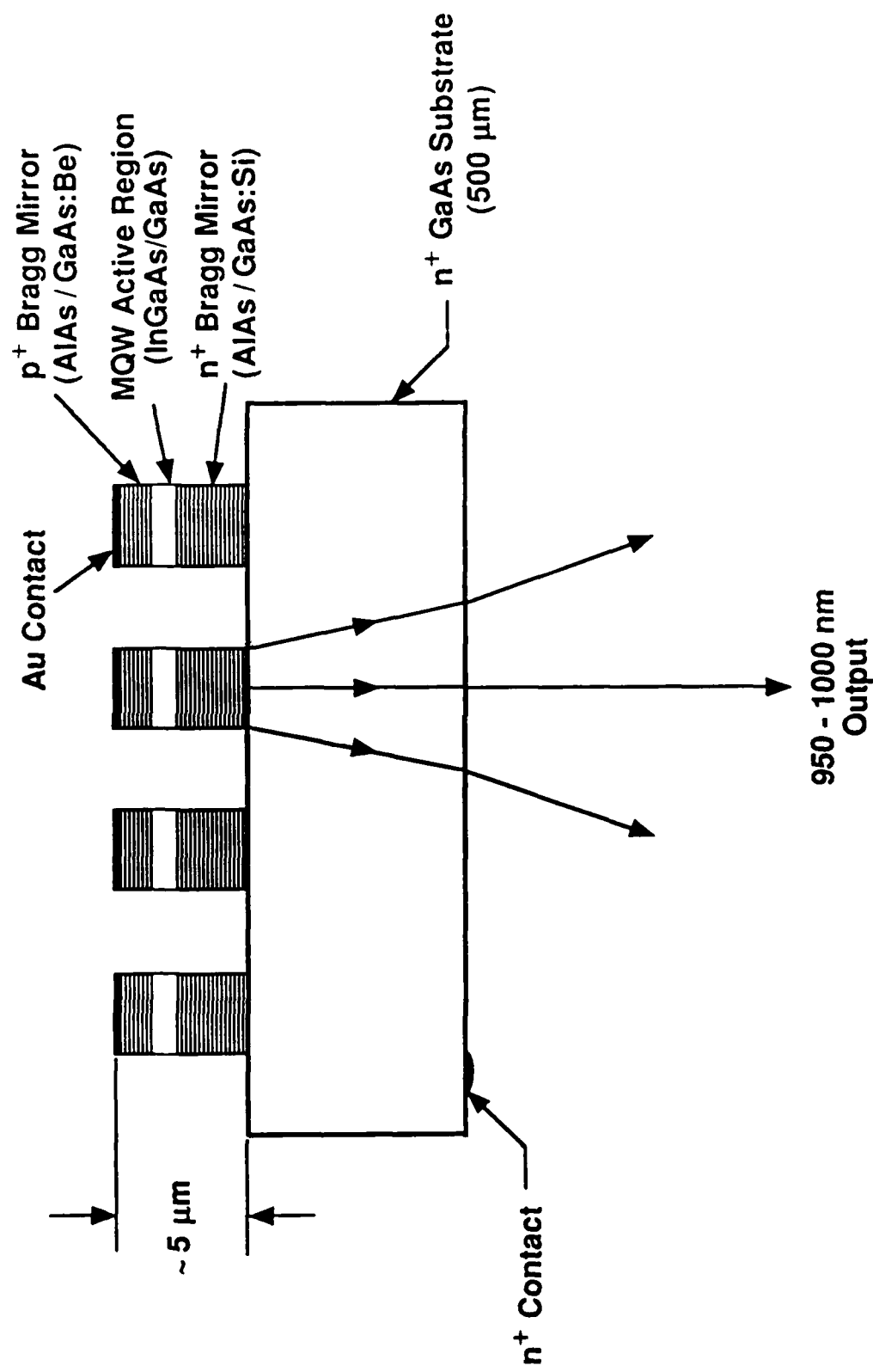


Figure 15.18

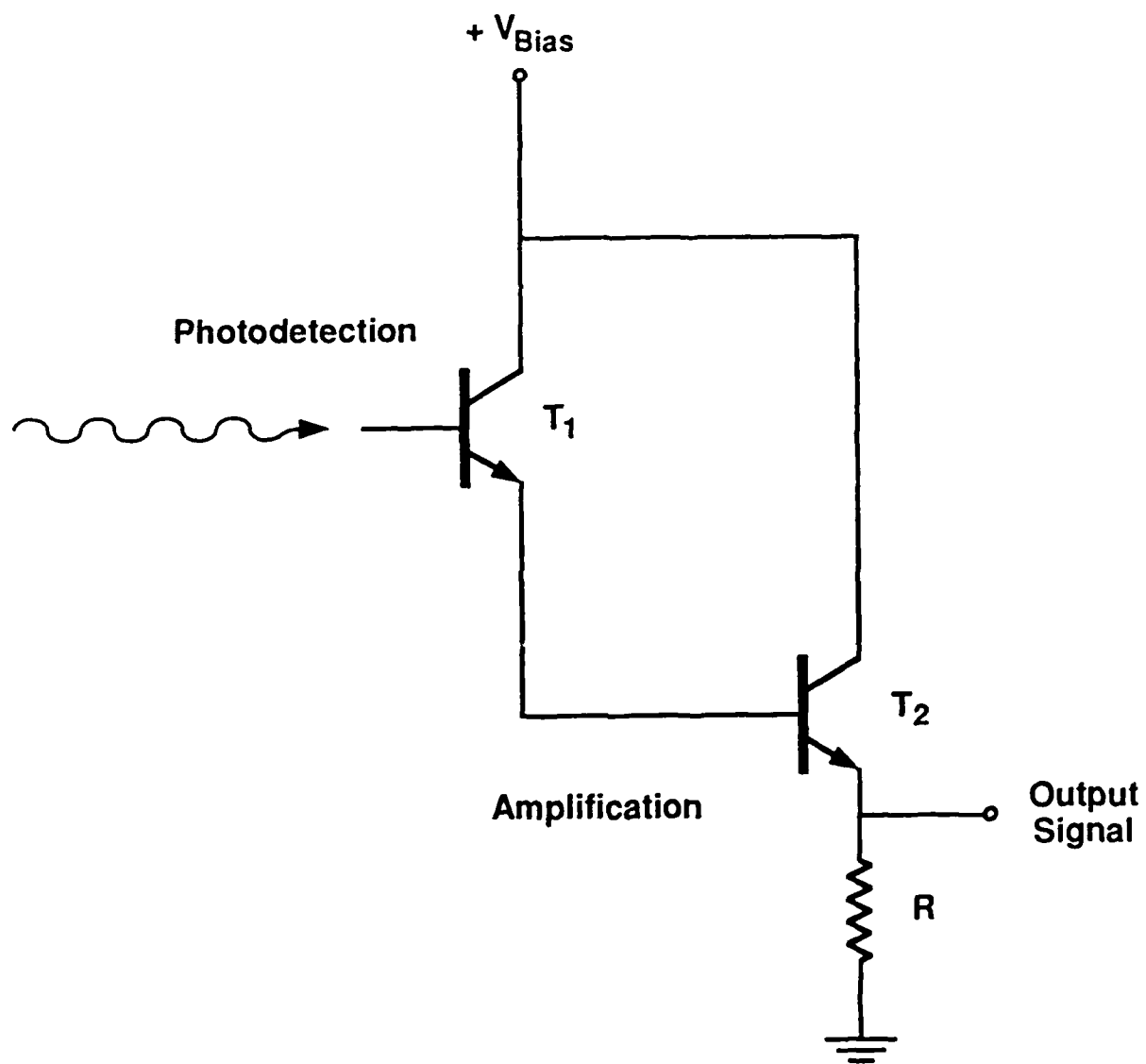


Figure 15.19

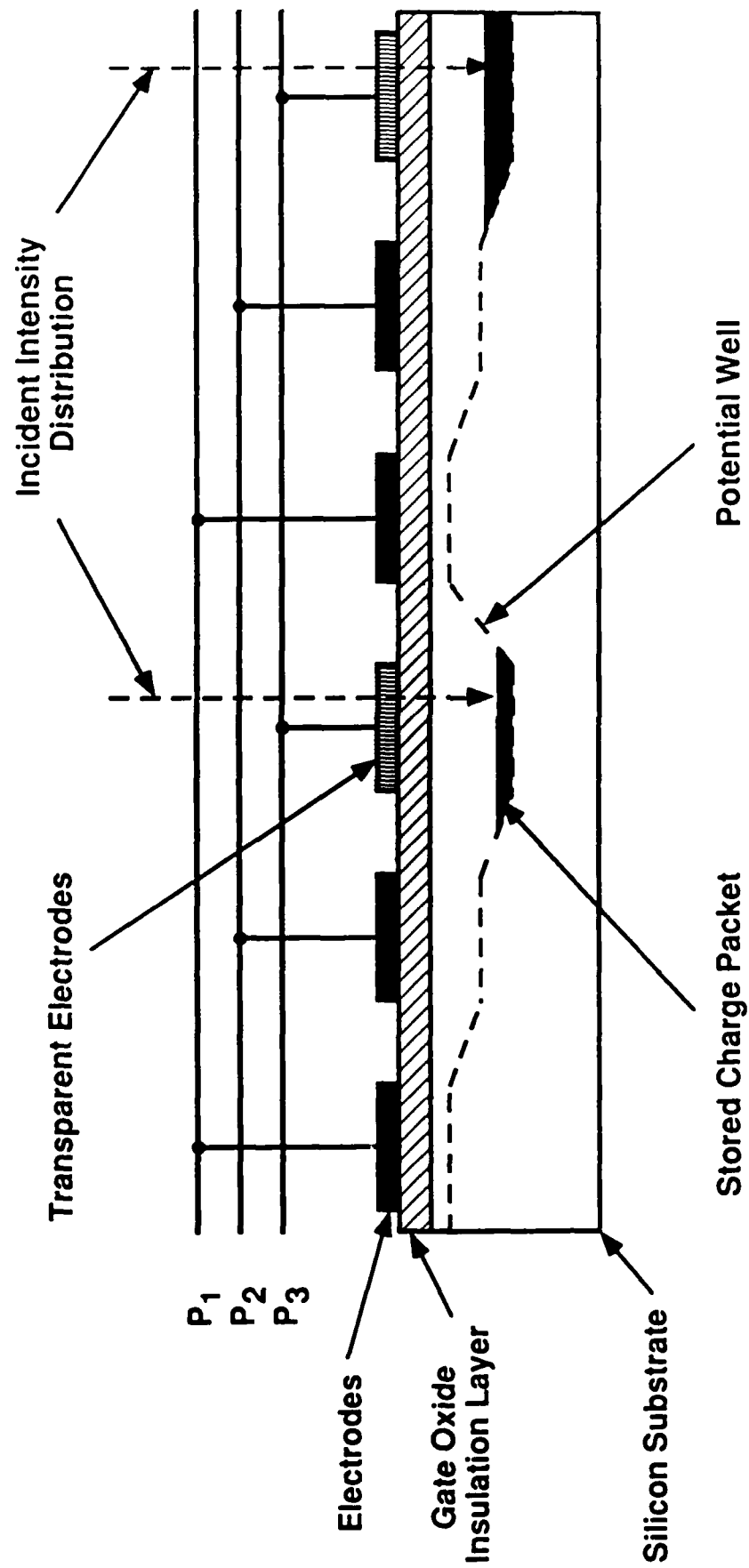


Figure 15.20

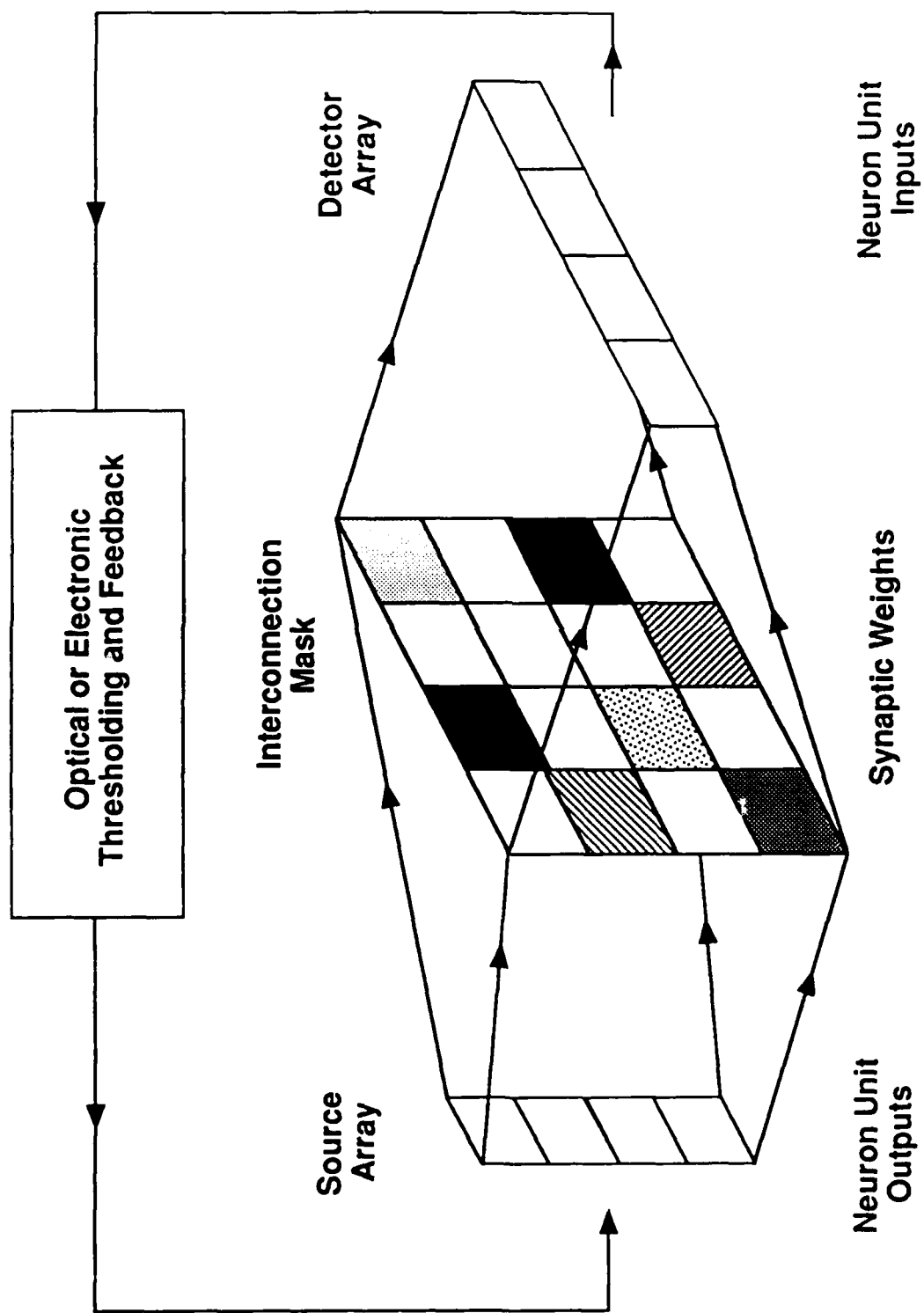


Figure 15.21

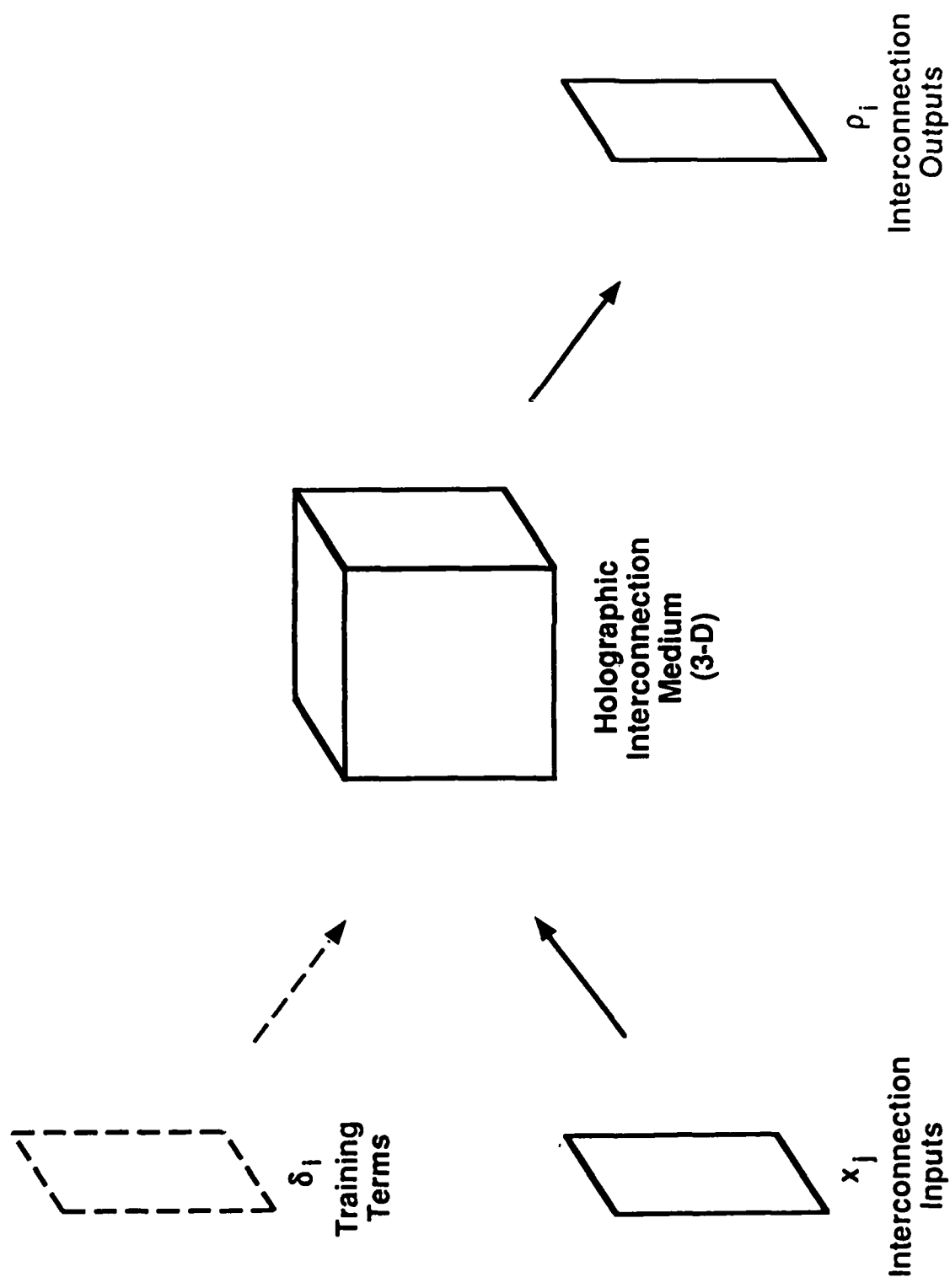


Figure 15.22

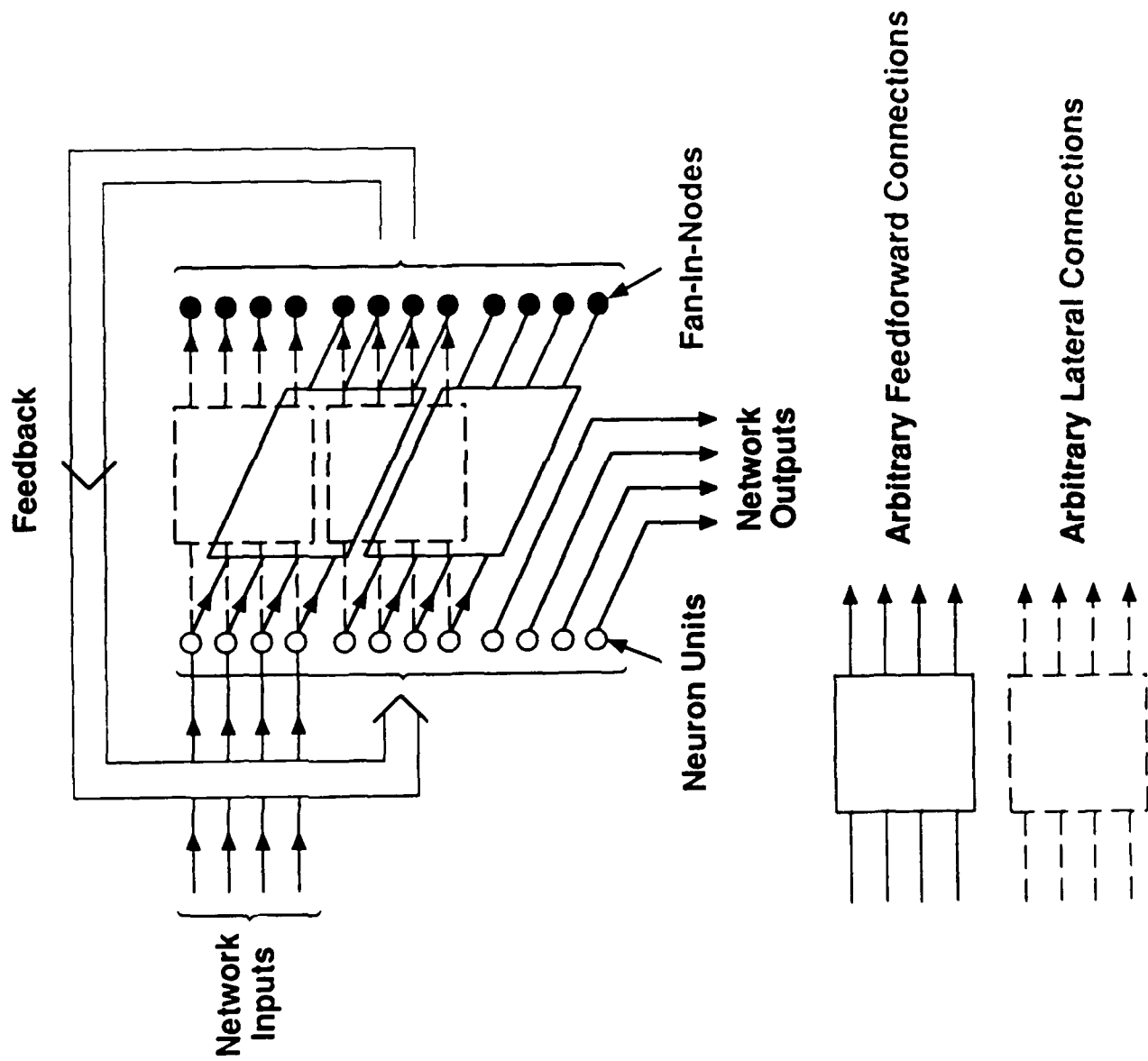
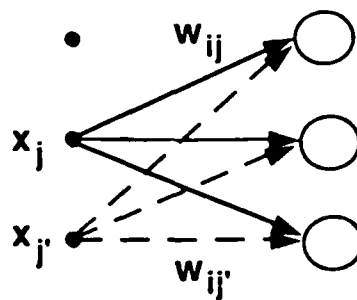
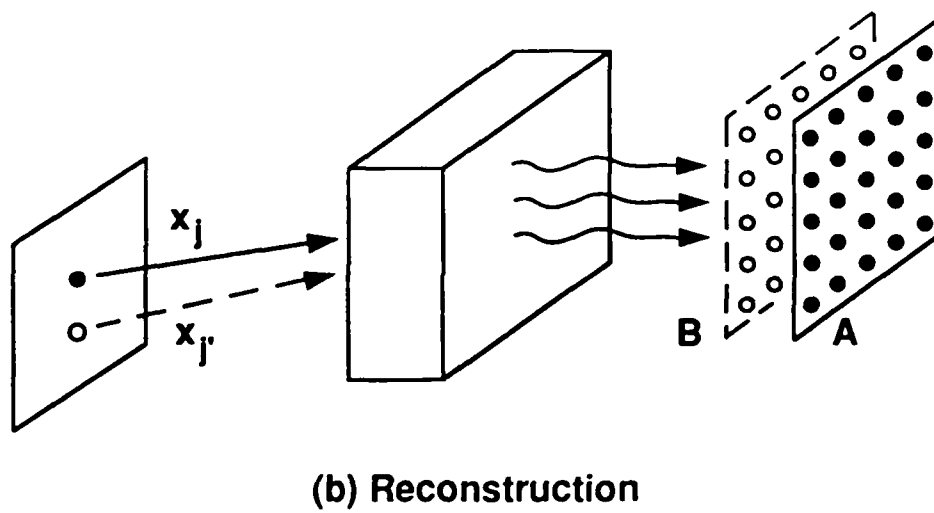
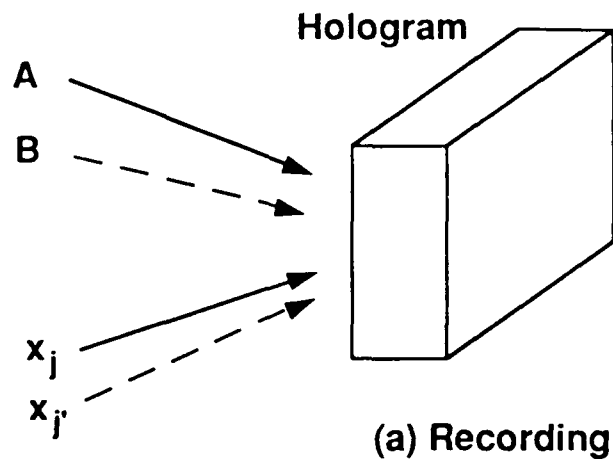


Figure 15.23



(c) Fan-out/Fan-in Interconnections

Figure 15.24

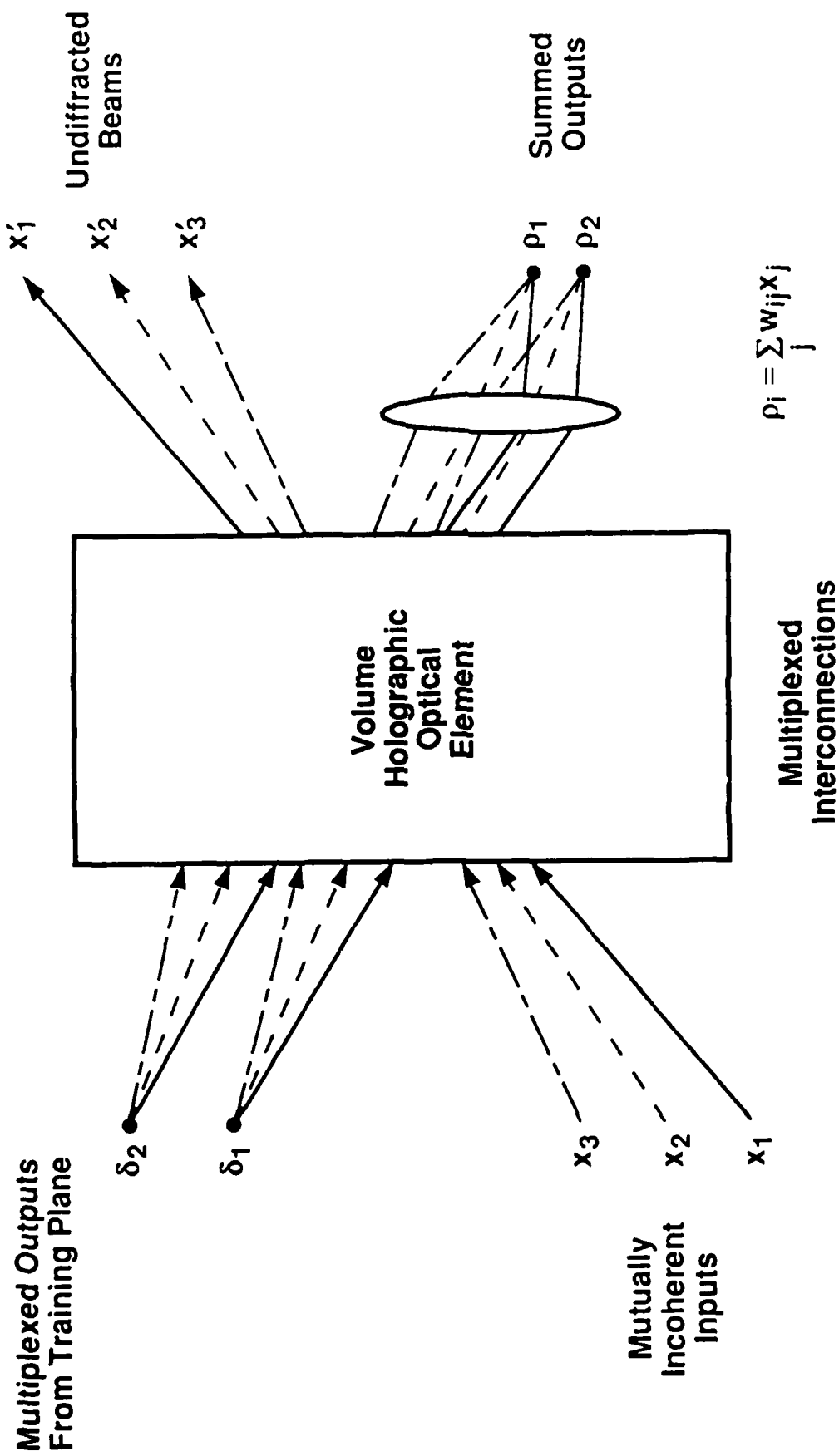


Figure 15.25

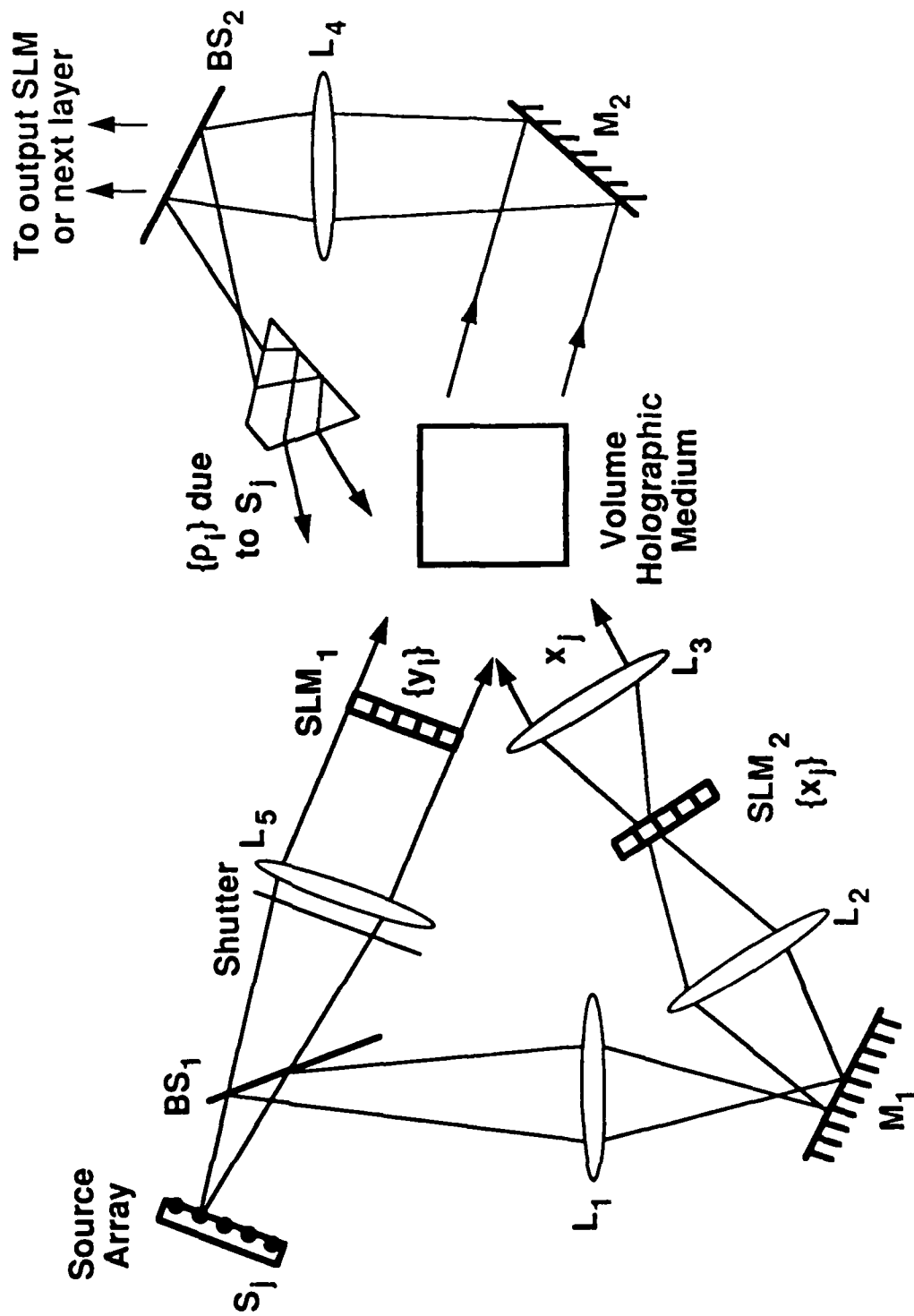


Figure 15.26

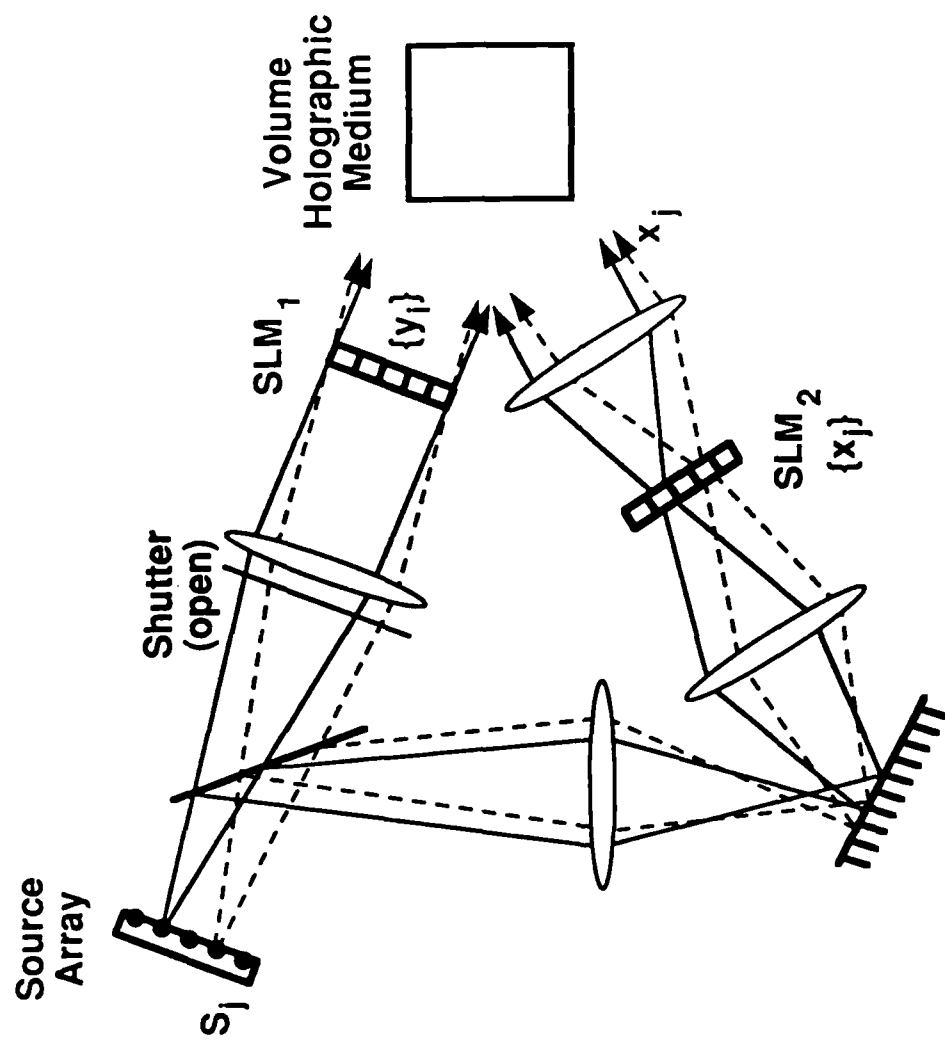


Figure 15.27

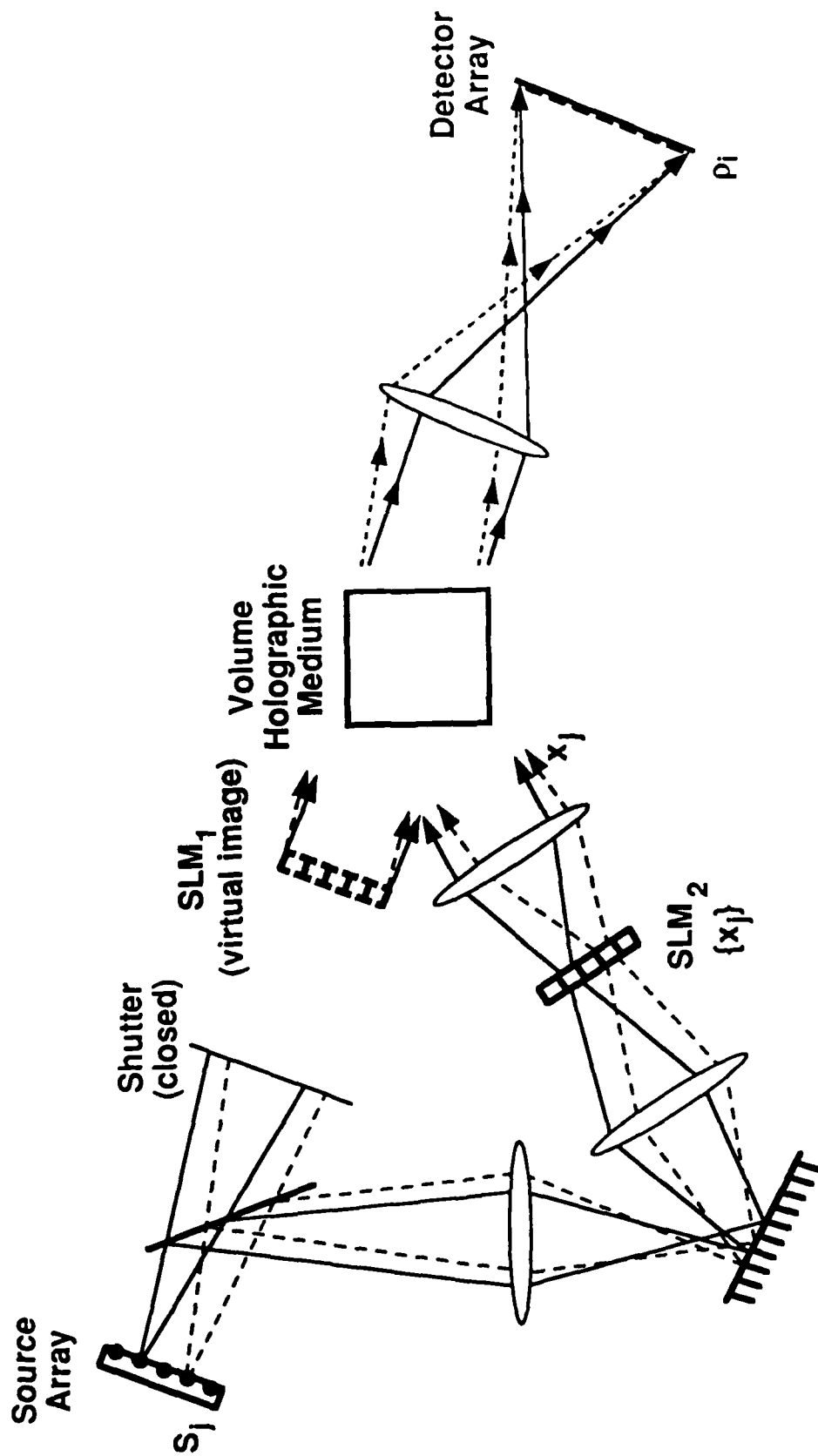


Figure 15.28

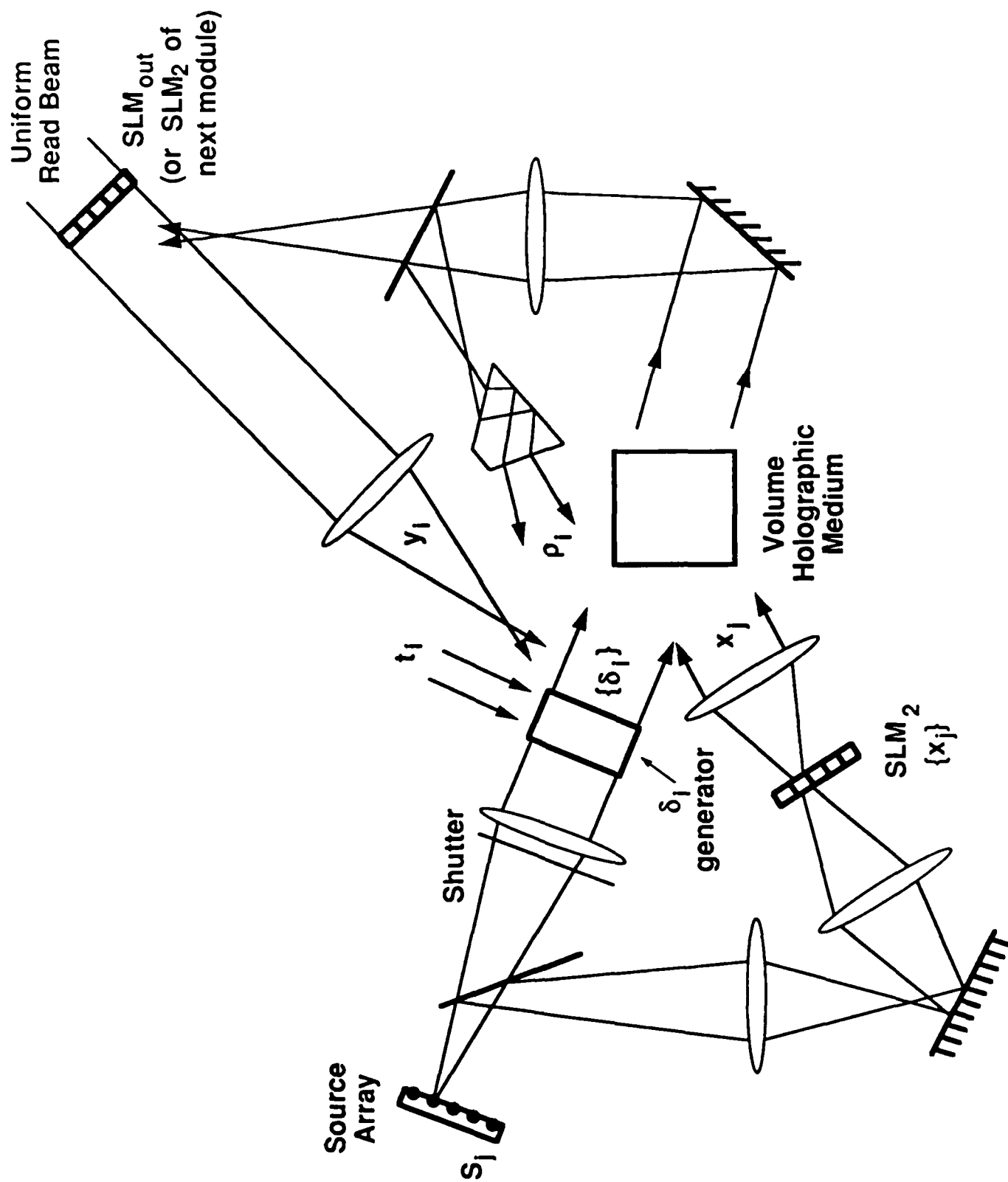


Figure 15.29

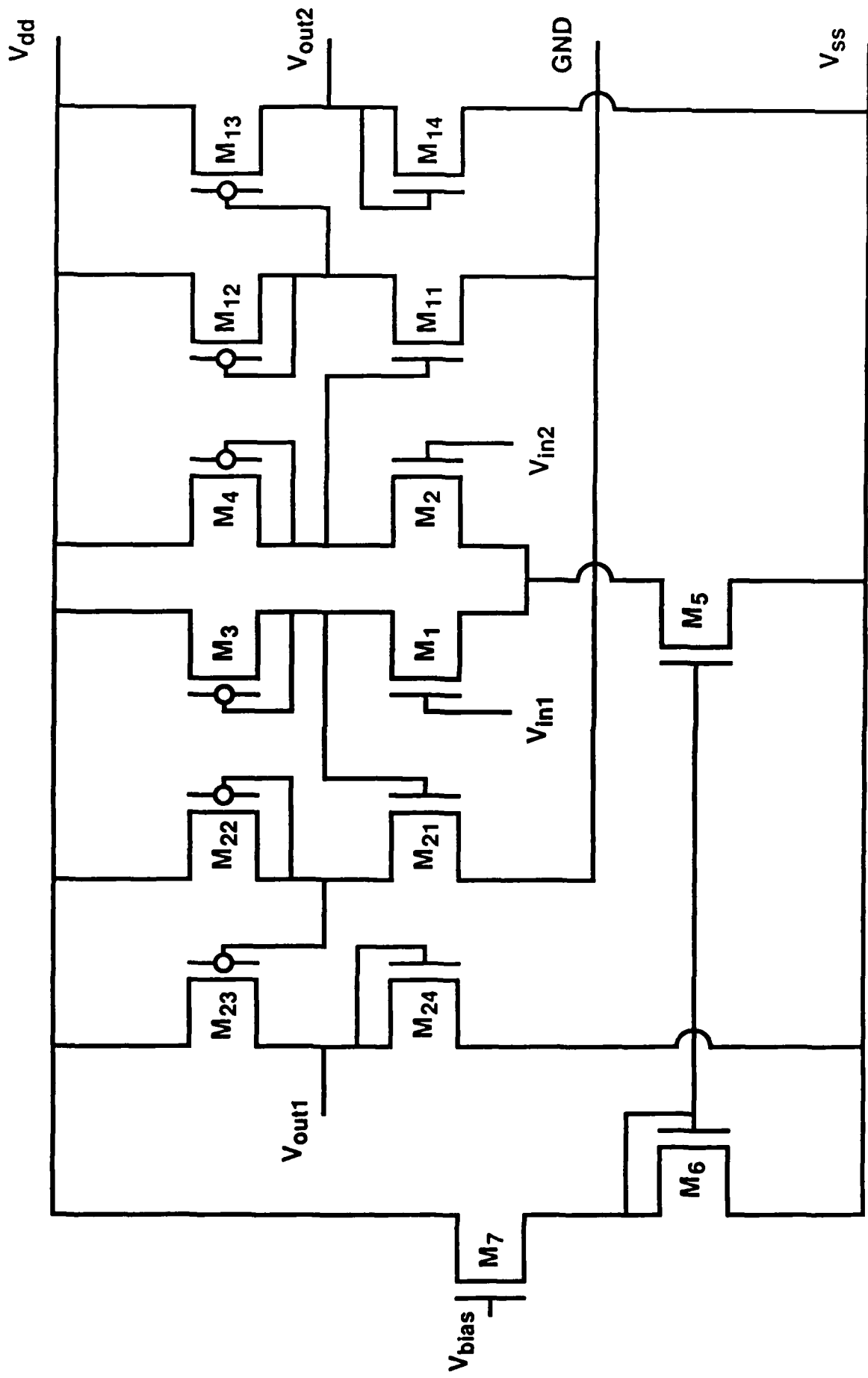


Figure 15.30

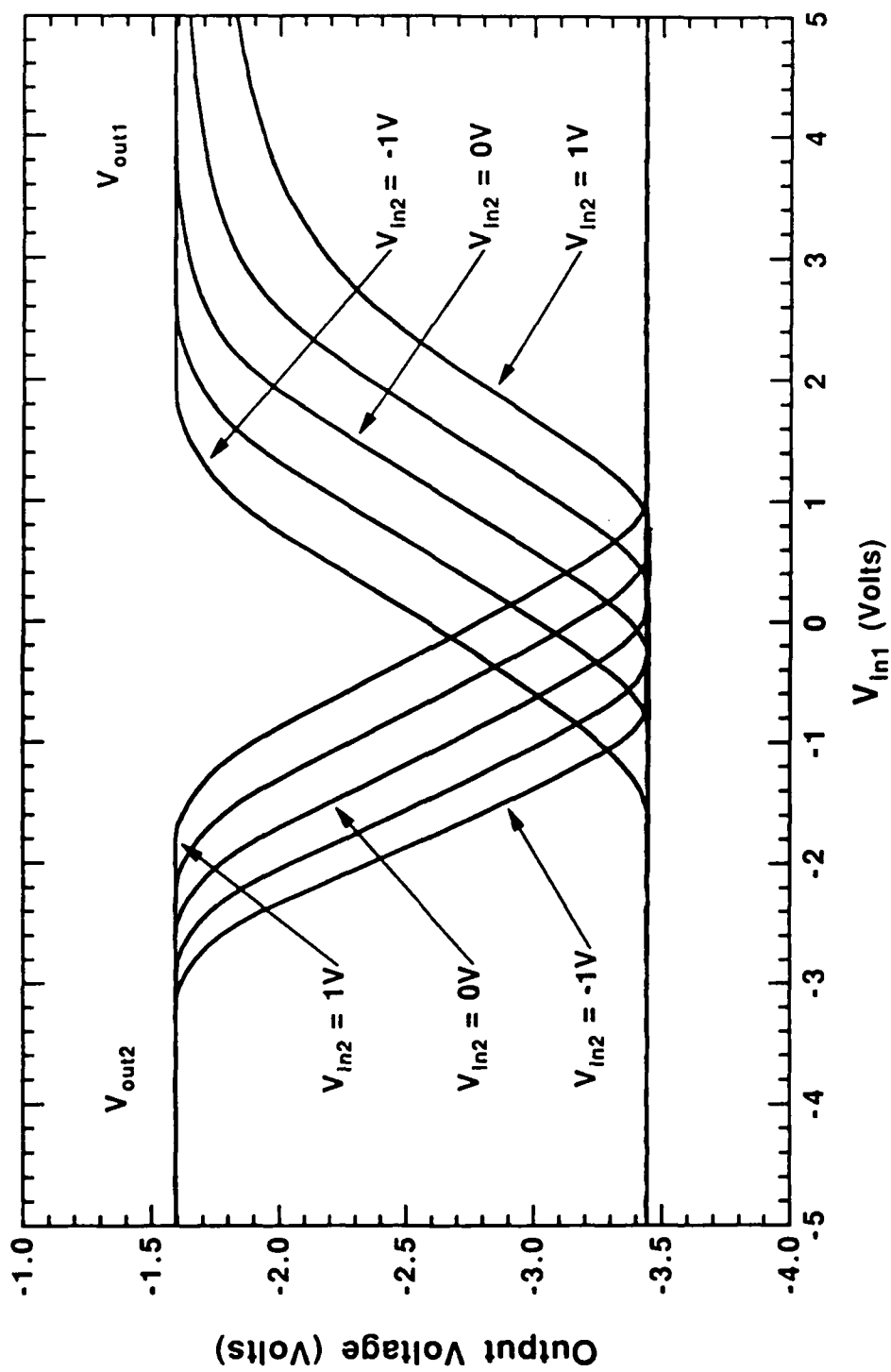


Figure 15.31

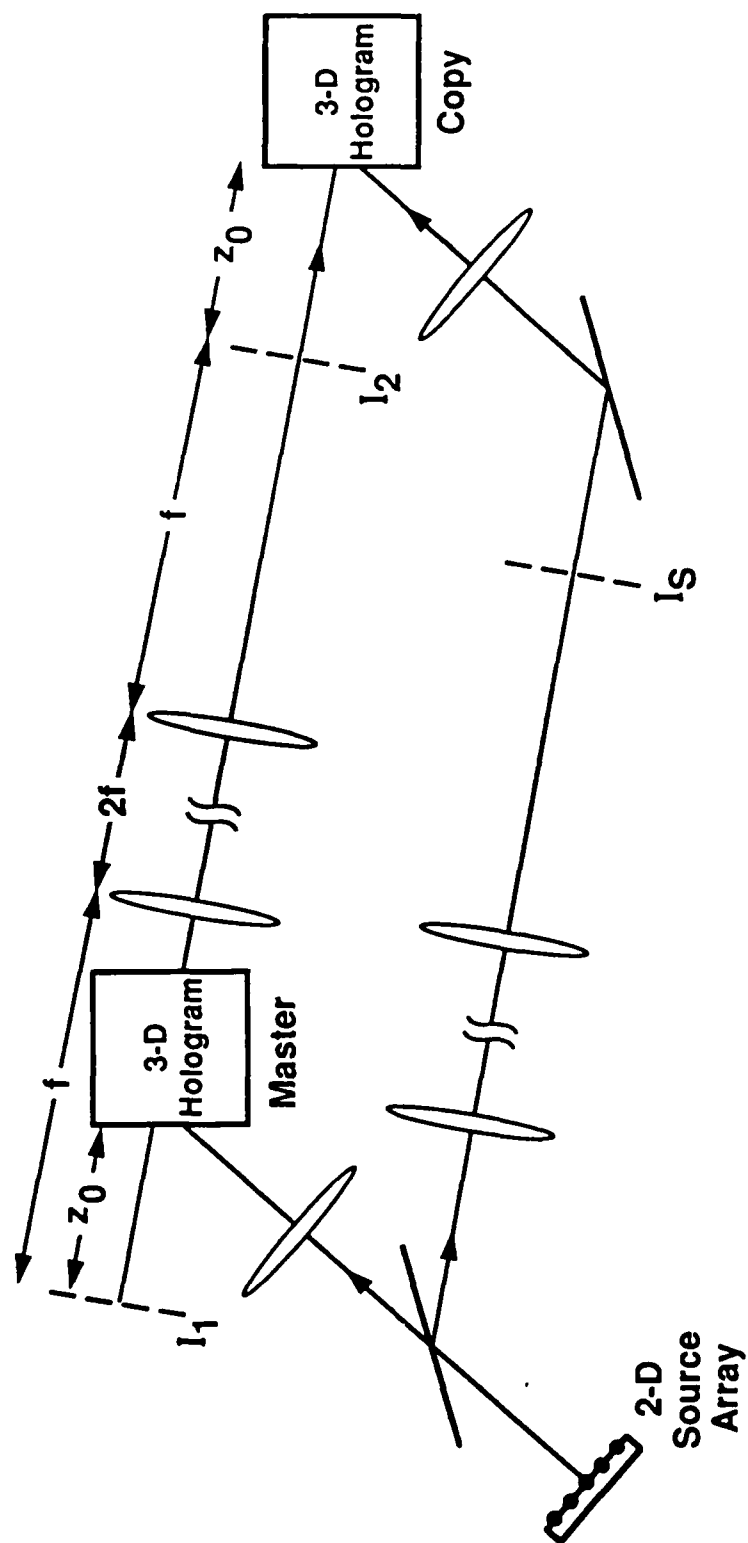


Figure 15.32

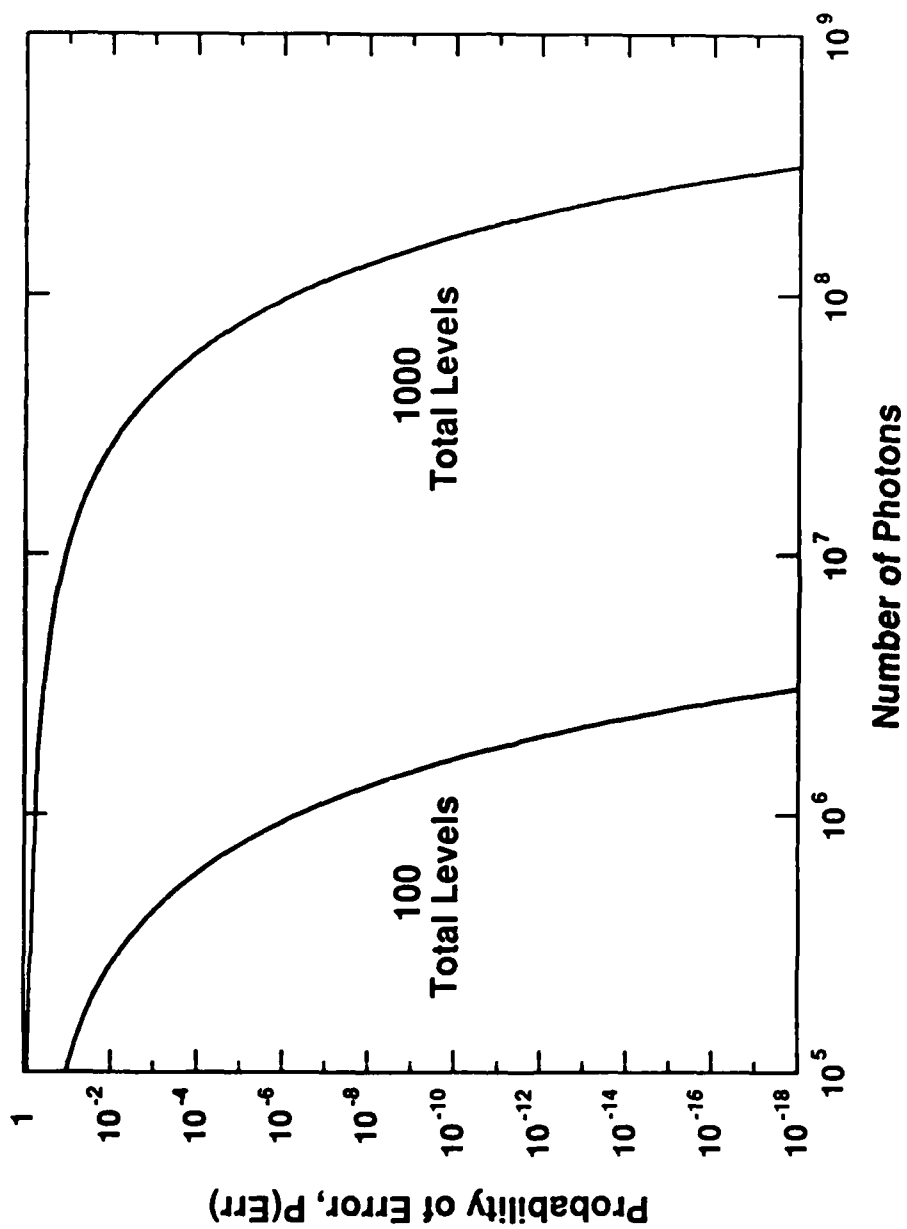


Figure 15.33

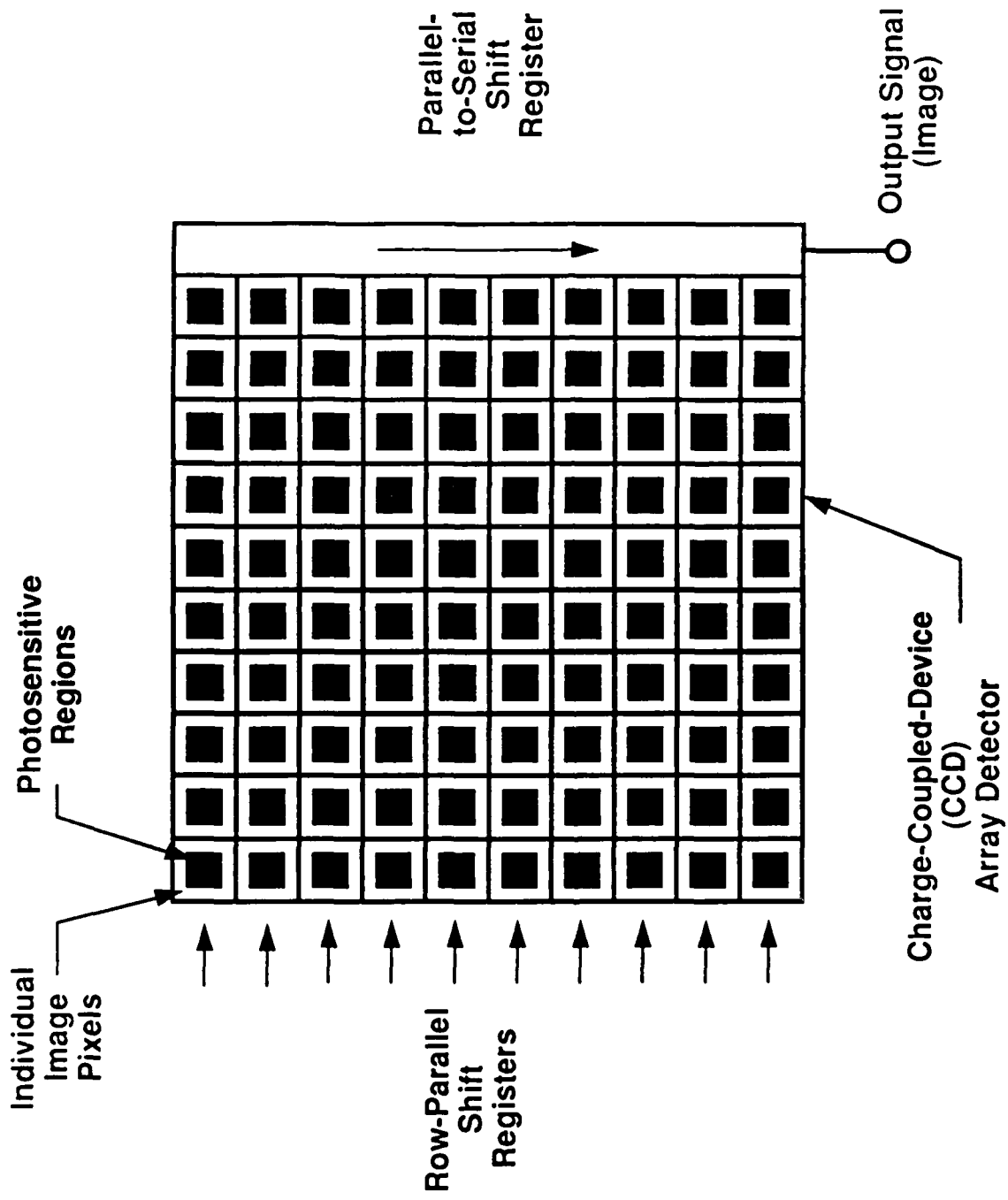


Figure 15.34



FUNDAMENTAL PHYSICAL AND TECHNOLOGICAL CONSIDERATIONS FOR SPATIAL LIGHT MODULATION

C. Kyriakakis, P. Asthana, Z. Karim, and A. R. Tanguay, Jr.

Optical Materials and Devices Laboratory,
and Center for Photonic Technology
University of Southern California
University Park, MC-0483

Los Angeles, California, USA 90089-0483

Phone: 213-743-6153

Fax: 213-746-8424

Telex: 4720490 USC LSA or 674803 UNIVSOCAL LSA

Numerous applications have been envisioned for spatial light modulators in optical information processing and computing systems [1, 2]. These applications can be summarized within the context of a generalized optical information processor or computer as shown schematically in Fig. 1 [2]. The principal functional roles of both one- and two-dimensional spatial light modulators include those of format, input, output, CPU, and memory devices. In addition, spatial light modulators can be effectively utilized to provide certain types of feedback interconnections, as for example in the case of optical crossbar switches and holographically encoded weighted interconnections.

Such a wide variety of applications has of course led to an equally wide variety of interrelated, and at times conflicting, device requirements. In this presentation, we examine these requirements from three complementary perspectives: fundamental physical limitations that affect the performance of any spatial light modulation function; the current status of spatial light modulator development with respect to such fundamental limits; and technological considerations that impact present and future device design and development. Each of these three perspectives will be discussed in detail, and is outlined briefly below.

Study of the fundamental physical limitations that affect an emerging technology is at once an exciting and somewhat sobering endeavor. The excitement arises naturally from the discovery of what we can in fact achieve; the sobering impact often occurs with the realization of what we have in fact achieved. Numerous such fundamental physical limitations pertain to the process of spatial light modulation. For example, two principal attributes

of incident wavefronts can be conveniently modulated: amplitude and/or phase. In most applications, it is desirable to modulate one or the other, but not both. Yet these two parameters are intimately related, such that modulation of one has a deterministic impact on the other through the Kramers-Kronig relations. Analysis of this interrelationship can yield fundamental limits on the phase/amplitude cross-talk anticipated for various physical device configurations, as well as appropriate figures of merit. Furthermore, such analysis reveals the potential for "dispersion engineering", in which one can imagine utilizing a technology such as compound semiconductor multiple quantum wells to tailor dispersion and/or absorption curves to match device requirements. One such example is illustrated in Fig. 2, which shows the related absorption and dispersion curves for two 5 meV linewidth Lorentzian oscillators displaced by 15 meV. As can be seen, a region of nearly linear index variation with energy can be generated at a local minimum in the absorption profile.

A second important fundamental limitation pertains to the minimum (quantum limited) energy required per unit resolution element to achieve a given level of modulation within predetermined accuracy constraints. Different limits can be derived for both analog and digital spatial light modulation [2, 3, 4, 5], as well as for the important cases of optically addressed and electrically addressed devices. For purposes of discussion, consider the case of optically addressed spatial light modulators, for which the modulation function inherently involves a detection process within each resolution element. The quantum limits for binary switching are well known, and are summarized for optical devices in Fig. 3 (after P. W. Smith, Ref. [6]). For analog modulation, quantum restrictions place much stricter boundary limitations on the minimum allowable photon flux for a given pixel resolution, error rate, and device framing rate [2, 3, 4, 5].

By comparison, analog representations (as used extensively, for example, in optical processors) require far more energy than the binary equivalent. This is due to the necessity of utilizing a much higher particle count (electrons or photons) in order to minimize the effects of quantum statistical fluctuations on the bit error rate (BER). For example, if we wish an analog representation of the number 1,000, then we require a dynamic range of at least 1,000:1. For incoherent illumination, quantum fluctuations in the emission/detection process produce a photon number distribution with a relative standard deviation of $\sigma \cong \sqrt{N}/N$. The equivalent of a 10^{-9} BER for the digital case corresponds to roughly 12 standard deviations. Therefore, the number of photons required must be greater than 1.5×10^8 from statistical considerations alone, as illustrated in Fig. 4. For a GaAs semiconductor laser characterized by a photon energy of ~ 1.5 eV, this corresponds to about $10^{10} k_B T$. To represent 1,000 optically in binary requires approximately 14 bits (10 bits for the number plus 4 bits of overhead) at 15 eV each (10 photons at 1.5 eV each, assuming direct detection and an ideal detector), or about $10^4 k_B T$.

For an optically addressed spatial light modulator operating over a dynamic range of

1000:1 with 1000 x 1000 elements in an active area of 1 cm², and at a frame rate of 1 MHz, the detected input flux must exceed 9 W/cm² in order to achieve a 10⁻⁹ BER in each pixel. Considerations such as these have significant implications for the SLM design, particularly with regard to sensitivity at a given dynamic range. These implications extend to the system design as well, in which case it is important to distinguish between single and multiple iteration algorithms (with different BER requirements at each stage). Furthermore, in many applications the algorithm is really a form of contraction mapping, in which the desired output space may be a very small subset of the transformed input space (e.g. determining whether or not a given correlation peak exceeds a predetermined threshold). In such cases, significantly relaxed demands may be appropriate for the BER of individual SLM pixels.

For envisioned systems applications involving one or more spatial light modulators, analysis of fundamental limitations such as that described above is essential for determining whether an analog or digital approach is favorable from the point of view of a fixed input power budget at a given desired computational throughput rate. In general, a given processing or computation function can be partitioned into the cost (energy or otherwise) of representation, the cost of computation, and the cost of detection and utilization of the answer. For operation at the quantum limits, analog representations are favored for architectures and algorithms that implement a high degree of computational complexity (irreducible number of equivalent binary operations) per unit detected output resolution element, whereas binary representations favor operations with a somewhat lower degree of computational complexity.

The current status of spatial light modulator development depends strongly for its assessment on the nature of the application and its resultant requirements. It is interesting to note at the outset that for certain applications (such as incoherent-to-coherent conversion with high analog accuracy), spatial light modulator technologies have been developed which approach quantum limited performance. On the other hand, a broad spectrum of applications exists for which current spatial light modulators fall far short of such ultimate performance boundaries. In making such comparisons, it is of critical importance to identify interrelated sets of performance parameters that cannot be arbitrarily separated, and to assess the conjoint figure of merit achievable within a given device technology (rather than the minimum value obtained across many different types of devices, or even within the same device under different operating conditions).

A large number of technological considerations apply to the eventual incorporation of spatial light modulators in optical information processing and computing systems. For systems of given size, spatial frequencies are inherently limited by the acceptance apertures of finite F-number lenses. Phase modulation, which is capable of much larger diffraction efficiencies at a given spatial frequency than amplitude modulation, is more difficult to

implement in imaging configurations and is more sensitive to substrate nonuniformities and polish figure. Amplitude modulation, on the other hand, can present formidable thermal dissipation problems for applications involving spatial light modulation with high optical throughput gain. Applications exist for both optically addressed and electrically addressed spatial light modulators, with distinct requirements for each. In fact, several recently conceived applications such as the utilization of spatial light modulators in neural network implementations could advantageously employ both address modes simultaneously.

Electrically addressed spatial light modulators lead quite naturally to inherently pixelated structures. Such structures can be advantageous from several points of view, including the convenient merging of optical and electrical inputs, interpixel cross-talk, strictly limited space-bandwidth product, and fixed system registration. However pixelation can yield additional difficulties such as inherently incomplete fill factors, scattering from metallic interconnections, and fixed pattern noise. Similar types of considerations apply to the utilization of reflective as opposed to transmissive device geometries.

A final technological consideration that will continue to strongly affect device design and development is that of available optical sources, both CW and pulsed. For fixed (and usually limited) external power, size, and weight considerations, the maximum achievable pulsed energy densities available set stringent requirements on the magnitudes of usable higher order material nonlinearities. Recent achievements of enhanced nonlinearities in both bulk and multiple quantum well compound semiconductor structures, as well as in nonlinear organic polymers, lend importance to the question of whether or not $\chi^{(3)}$ materials as well as $\chi^{(2)}$ materials will provide useful spatial light modulation functions in future devices.

1. C. Warde and A. D. Fisher, "Spatial Light Modulators: Applications and Functional Capabilities", in *Optical Signal Processing*, J. Horner, Ed., Academic Press, Inc., San Diego, (1987), 477-518.
2. A. R. Tanguay, Jr., "Materials Requirements for Optical Processing and Computing Devices", *Opt. Eng.*, 24(1), 2-18, (1985).
3. C. Kyriakakis, P. Asthana, R. V. Johnson, and A. R. Tanguay, Jr., "Fundamental Physical Limitations of Optical Information Processing and Computing", 1987 Optical Society of America Topical Meeting on Optical Computing, Incline Village, Nevada, (1987).
4. C. Kyriakakis, P. Asthana, R. V. Johnson, and A. R. Tanguay, Jr., "Spatial Light Modulators: Fundamental and Technological Issues", Optical Society of America Topical Meeting on Spatial Light Modulators, Lake Tahoe, Nevada, (1988).

5. A. R. Tanguay, Jr., "Physical and Technological Limitations of Optical Information Processing and Computing", Materials Research Society Bulletin, Special Issue on Photonic Materials, XIII(8), 36-40, (1988).
6. P. W. Smith, "Applications of All-Optical Switching and Logic", Phil. Trans. R. Soc. Lond., A313, 349-355, (1984).

ELEMENTS OF AN OPTICAL COMPUTER

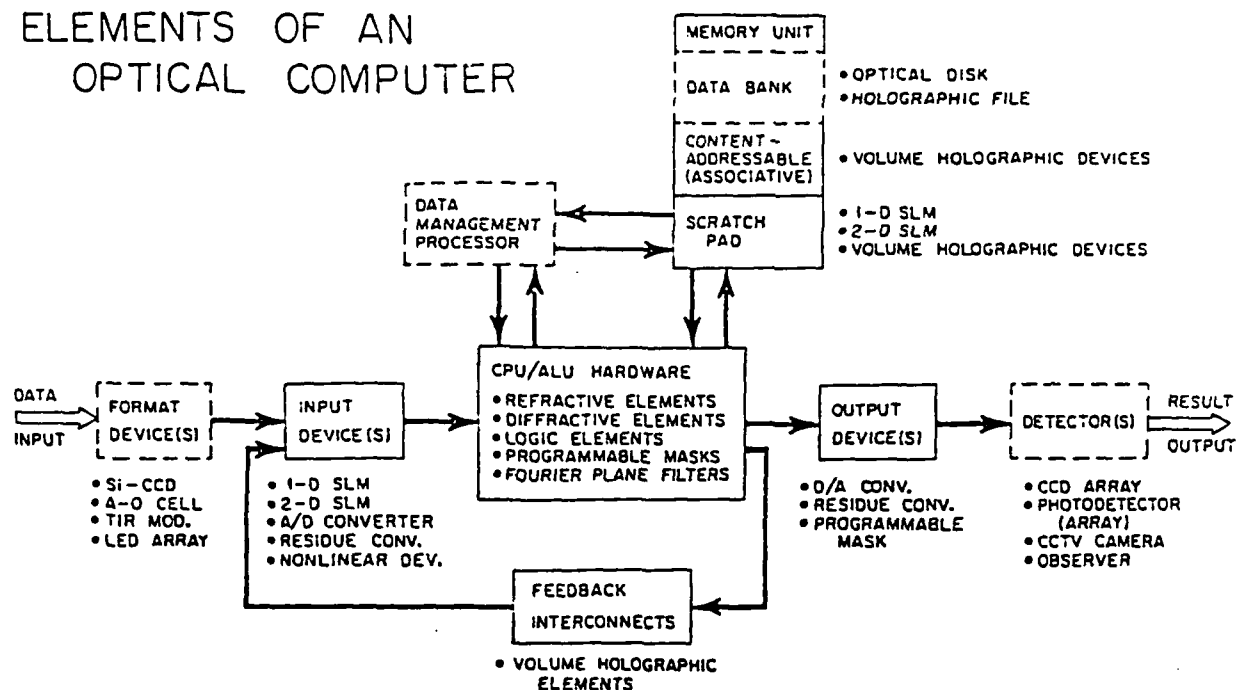


Fig. 1 Schematic diagram of the principal elements of a generalized optical processor or computer.

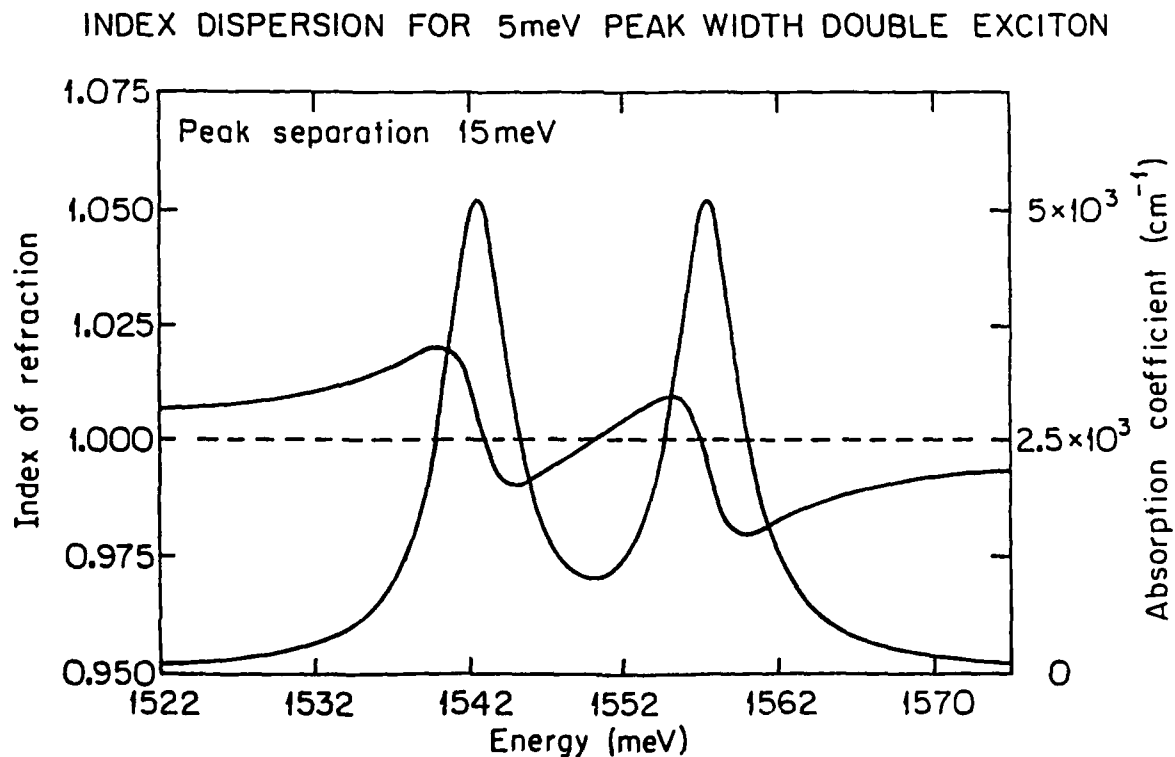


Fig. 2 Absorption and dispersion of a two oscillator system, with 5 meV linewidths displaced by 15 meV.

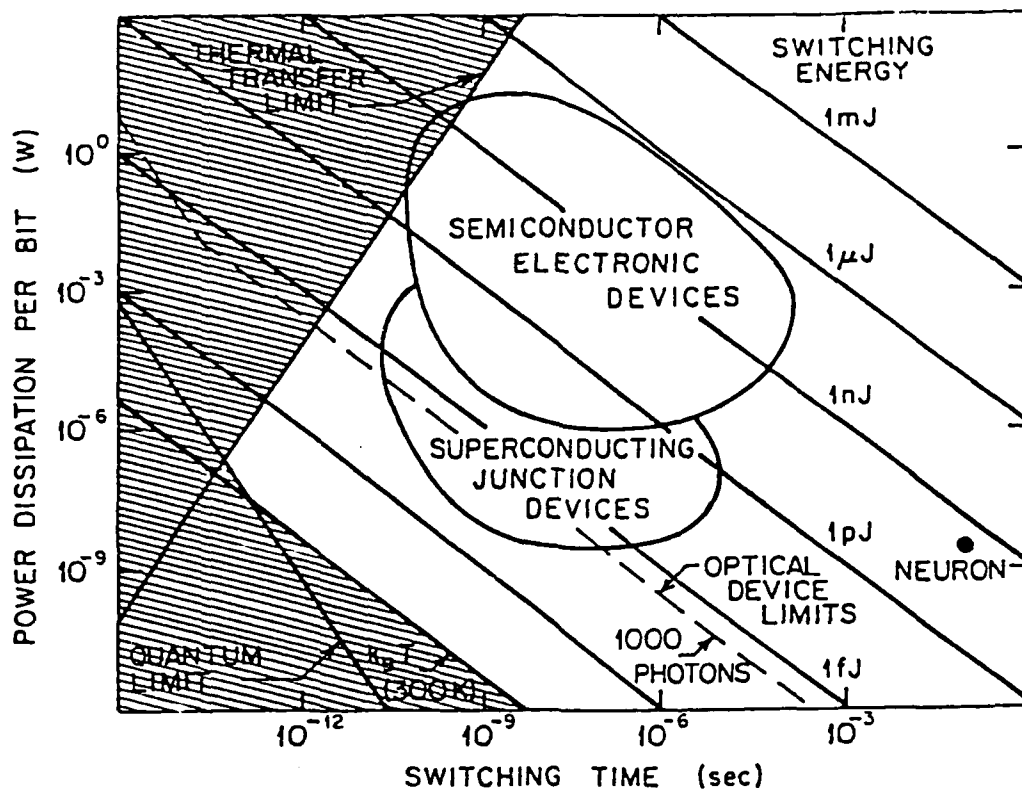


Fig. 3 Power dissipation per bit as a function of switching time, showing several fundamental and technological limitations (after Ref. 6).

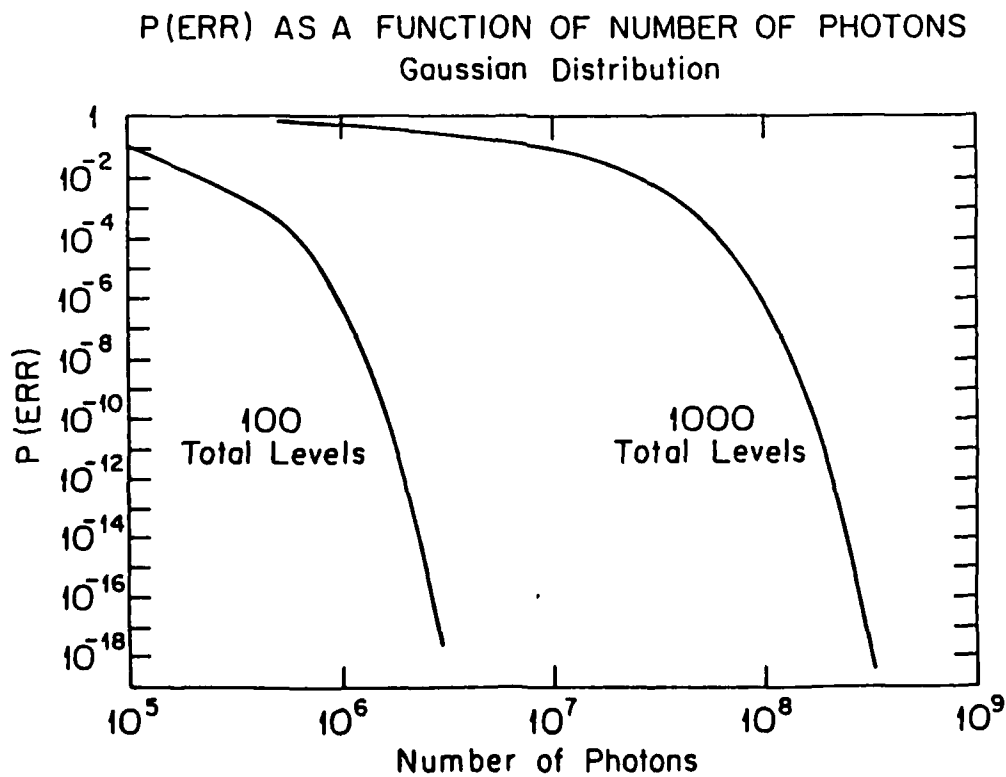


Fig. 4 The probability of error $P(\text{ERR})$ as a function of the number of photons detected in a given pixel of an optically addressed spatial light modulator, for both 100:1 and 1000:1 dynamic range.

Fundamental and Technological Limitations of Asymmetric Cavity MQW InGaAs/GaAs Spatial Light Modulators

**C. Kyriakakis, Z. Karim, J. H. Rillum, J. J. Jung,
A. R. Tanguay, Jr., and A. Madhukar**

**Optical Materials and Devices Laboratory,
and Center for Photonic Technology
University of Southern California,
Los Angeles, CA 90089-0483**

Summary

Numerous applications have been envisioned for spatial light modulators in optical information processing and computing systems. The principal functional roles of both one- and two-dimensional spatial light modulators include those of format, input, output, CPU, and memory devices. Such a wide variety of applications has of course led to an equally wide variety of interrelated, and at times conflicting, device requirements that arise from fundamental physical limitations as well as current technological drawbacks. In this investigation we have examined these requirements for a particular class of modulators, the operation of which is based on an asymmetric Fabry-Perot etalon with a multiple quantum well (MQW) structure as the cavity medium.

The use of such devices as reflection modulators based on AlGaAs/GaAs MQW structures has been investigated by a number of authors^{1,2,3}. One of the primary advantages of such structures is the considerable increase in both dynamic range of modulation and contrast ratio achievable at a given applied field strength in comparison with previously investigated transmissive configurations, in addition to the elimination of the necessity for counteretching of the (absorptive) GaAs substrate. On the other hand, the dynamic range achieved in any reflection configuration must be assessed in conjunction with the 6 dB insertion loss characteristic of the intensity (rather than polarization) beamsplitter required to implement normal incidence readout. This factor alone rescales an 80% change in reflectivity to less than 20% when considered from the perspective of *system* rather than *device* throughput. Furthermore, such a reflection configuration presents the additional systems complexity of source isolation within each stage of a (multistage) processor.

One possible means of avoiding the 6 dB insertion loss is to employ a polarizing beamsplitter in conjunction with a quarter wave plate, thereby insuring a 90° rotation of the incident polarization on reflection. This configuration adds system complexity and cost, as well as additional component requirements such as the extinction ratio of the polarizing beamsplitter, and the off-axis uniformity of the net phase retardation induced by the quarter wave plate. Such requirements must be set so as to not compromise the contrast ratio designed into the reflective spatial light modulator. Alternatively, one can employ MQW structures specifically designed to be operated at other than normal incidence, with the concomitant complications implied by the

associated anamorphic optics, lateral rescaling, depth of field, and placement of critical and conjunctive components with appropriately low f-numbers.

In order to explore asymmetric cavity MQW structures that are capable of transmissive mode operation without resort to the process difficulty and sample fragility implied by counteretching, we have examined the prospects of utilizing electric-field-addressed multiple quantum wells fabricated in the InGaAs/GaAs system on GaAs substrates. In this investigation, we have taken the perspective that identification of fundamental physical limitations to device performance is exceedingly useful as a mechanism for the delineation (and potential circumvention) of limitations that arise from purely technological choices. These considerations are of particular importance in the InGaAs/GaAs system due to the fact that the considerable lattice mismatch characteristic of useful indium compositions results in a strained layer growth environment with concomitant implications on growth morphology and the necessary incorporation of strain relief mechanisms (such as growth on patterned substrates⁴). In this paper, we describe several important design criteria that have evolved from this study, and apply them to two specific cases: (a) a transmission modulator with optimized dynamic range and contrast ratio, and (b) a reflection configuration in which the readout illumination is incident through the (antireflection coated) substrate. This latter configuration yields significantly improved performance characteristics as well as novel operational features at a considerable reduction in device fabrication complexity.

An important fundamental physical limitation pertains to the degree of spatial light modulation that can be achieved per unit change in a given modulation parameter. There are two characteristics of wavefronts that can be conveniently modulated: amplitude and phase. In many optical information processing applications it is desirable to modulate one or the other, but not both. As an example of a "pure" amplitude modulator we have utilized a Lorentzian oscillator profile to model a typical heavy hole exciton (neglecting the band edge absorption) for an InGaAs/GaAs MQW structure with a peak absorption of $10,000 \text{ cm}^{-1}$ and a linewidth (full width at half maximum) of 7.5 meV (derived from experimental data on a 27 period MQW structure with 35 monolayers of InGaAs (15% In) and 71 monolayers of GaAs grown by molecular beam epitaxy (MBE)), as shown in Fig. 1. Similarly, the response of the absorption spectrum to an applied bias due to the quantum confined Stark effect (QCSE) for a typical applied voltage of about 100 kV/cm and for a total device thickness of approximately $1.2 \text{ }\mu\text{m}$ (100 Å wells/ 200 Å barriers) is also shown (dashed) in Fig. 1. Again, experimental data were used to quantize the decrease in oscillator strength as well as the exciton broadening effects that are typical for such systems. This model illustrates the basis of operation of a normally-on amplitude modulator that utilizes electroabsorption at the chosen wavelength ($0.985 \text{ }\mu\text{m}$) to generate the off-state, resulting in a maximum achievable change in absorption of 5500 cm^{-1} . In addition to this constraint, other parameters that determine optimum performance include the device thickness and the mirror reflectivities, as well as the width and location of the Fabry-Perot resonance modes (which are also dependent on electrorefractive-induced perturbations in shape and position). Related to the device thickness is a critical cavity phasing condition, which can be set to an odd or even multiple of π , thus giving rise to two fundamentally different operational modes.

For the transmission modulator configuration, application of the oscillator model with its corresponding index dispersion data as shown in Figs. 1 and 2, in conjunction with an optimally designed asymmetric cavity, yields a combination of fundamental and technological considerations that in turn limit the dynamic range to 50% with a contrast ratio of 10:1, as shown in Fig. 3. In this case, the mirror reflectivities are both equal to 85%. It should be noted that the signal dependent phase modulation is limited in this case to approximately 0.03π . The fundamental reason for the relatively poor contrast ratio is the necessity in the transmission configuration to phase the cavity such that in the on (low absorption) state, the first and second order transmitted components add in phase. Thus, extinction in the off state relies on the development of enough voltage-induced absorption change at the operating wavelength to completely eliminate the first

order beam on a single pass through the modulator. The peak absorption coefficient characteristic of the InGaAs/GaAs system at this composition is insufficient to produce more than an 10:1 change with a reasonably wide dynamic range. It should be noted in Fig. 3 that the shorter wavelength results are somewhat artificial due to neglect of the parabolic, background and residual absorption terms.

On the other hand, by utilizing a low reflectivity (25%) incident mirror in conjunction with a high reflectivity back mirror, the best possible dynamic range achievable in reflection is 70% with a contrast ratio of greater than 25:1, as shown in Fig. 4. This configuration is characterized by an operational bandwidth of about 50 Å and an insertion loss of approximately 1.5 dB, without taking into account the system throughput loss necessitated by the incident beamsplitter. If a polarization-independent beamsplitter is employed, the system throughput loss in this case is about 8 dB with a reduced dynamic range of 17.5% at the same contrast ratio.

Should a reflection modulator configuration prove desirable, the transparency afforded by the InGaAs/GaAs MQW system on a GaAs substrate can be used to advantage by inverting the traditional device structure to incorporate the low reflectivity incident mirror as a few period Bragg mirror (typically AlGaAs/GaAs) capped during growth by the strained layer MQW structure. The device may then be completed by deposition of either a metal (affording both a contact as well as RF isolation) or dielectric stack high reflectivity mirror on the top surface. The principal advantage of this structural configuration is the elimination of the requirement for a (typically) 20 or more period Bragg reflector grown prior to the precision growth stage necessitated by the desired thickness of active MQW material. Not only does this significantly affect the device processing throughput and yield, but it also greatly enhances the quality of the final interface prior to the initiation of MQW growth.

Additional technological considerations affecting the implementation of SLM's with InGaAs/GaAs MQW structures include issues of growth thickness accuracy (such as process variability and array uniformity), which particularly affect two critical design parameters: namely, the oscillator position and linewidth, and the cavity resonance condition. Background doping concentrations are also known to be particularly important in maximizing the contrast ratio, as well as the voltage division between any incorporated Bragg mirror (with $> k_B T$ band discontinuities) and the MQW active region.

References

- [1] T. Y. Hsu, Uzi Efron, W. Y. Wu, J. N. Schulman, I. J. D'Haenens, and Yia-Chung Chang, "Multiple Quantum Well Spatial Light Modulators for Optical Processing Applications", *Opt. Eng.*, 27 (5), 372-383, (1988).
- [2] M. Whitehead, G. Parry, and P. Wheatley, "Investigation of Etalon Effects in GaAs-AlGaAs Multiple Quantum Well Modulators", *IEE Proc.*, 136 (J1), 52-58, (1989).
- [3] R.H. Yan, R.J. Simes, and L.A. Coldren, "Electroabsorptive Fabry-Perot Reflection Modulators with Asymmetric Mirrors", *IEEE Photon. Tech. Lett.*, 1 (9), 273-275, (1989).
- [4] S. Guha, A. Madhukar, K. Kaviani, and R. Kapre, "Growth of $\text{In}_x\text{Ga}_{1-x}\text{As}$ on Patterned GaAs (100) Substrates", *J. Vac. Sci. Technol. B*, 8(2), 149-153, (1990).

Device Characteristics of Optical Disc Spatial Light Modulators

John H. Rilum and Armand R. Tanguay, Jr.

Optical Materials and Devices Laboratory, and Center for Photonic Technology
University of Southern California, University Park MC-0483,
Los Angeles, CA 90089-0483

Introduction

Optical memory disc technology has recently matured and penetrated the personal computer industry market with high density and extremely high capacity serial data storage devices. The high resolution (submicron) and long life time (> 10 years) characteristic of the optical disc medium in combination with image-based recording and an appropriate parallel readout scheme also make this advanced technology a very competitive alternative for spatial light modulation.

Utilization Concept

There are numerous possible ways of utilizing the optical memory disc as a spatial light modulator (SLM). One such configuration is as a binary (half-tone) memory with a parallel readout capacity of up to 10^8 bits/cm². This use, of course, implies strict design constraints on the resolution of the readout optics in order to fully utilize this capacity. More traditional SLM functions can be incorporated by using the optical disc as a binary-encoded analog image memory¹, in which grey scale is obtained at the expense of resolution. Assume, for simplicity, a bit density of 10^8 bits/cm² (which is within reach of current optical disc technology) and an image size of 1 cm x 1 cm. The image can then be divided into 1000 x 1000 pixels, each of size 10 μ m x 10 μ m with 10 x 10 bits for grey scale generation. By random area encoding the serially written bits in each pixel, a particular grey scale level directly proportional to the total number of bits in each pixel can be obtained. The theoretical dynamic range for this case is 100:1, corresponding to approximately 6.6 grey scale bits, thus resulting in a total information density of ≈ 6.6 Mbits/cm². Note that it is now only necessary to resolve each pixel (10 μ m x 10 μ m) in order to obtain the desired continuous grey level in each pixel.

Optical Readout Approach

In order to implement parallel (two-dimensional) readout of optical memory discs, it is desirable to employ direct imaging of the optical disc to obtain the maximum achievable throughput efficiency and at the same time obtain a high contrast ratio in the output image. An alternative approach is to read out the optical disc in parallel holographically² by storing the images in the form of two-dimensional computer generated holograms. However, holographic readout implies a non-optimum trade-off between throughput efficiency and SNR for a given space-bandwidth product. An optical readout configuration which has an inherently high SNR, is independent of the groove profile (required for the tracking servo-system of most disc drivers), and can be utilized with a variety of recording layer principles (such as ablation or bump forming) can be achieved by using a differential (or shearing) interferometric approach¹. In this readout scheme, two spatially coherent images of the optical disc are created and interfered differentially by introducing a shift of one half of the minimum bit pitch along the direction of the grooves of a distance (Fig. 1a). In the output image, only local differences in phase (height) and/or reflectivity of the recording layer will be detected. Thus each bit will produce two bright marks ($\Delta/2$ apart) on a dark background in the output image (Fig. 1b). Maximum output intensity will consequently be obtained for recording with a 50% duty cycle pattern.

Optical Readout Configuration

The differential interferometric readout concept can be implemented by using a shear plate, which consists of a birefringent plate with a tilted optic axis (Fig. 2). When imaging the optical

disc through the shear plate, two images with orthogonal polarizations are produced and are shifted by an appropriate amount as determined by the recorded bit length. In order to obtain the optimum phase shift of $\pi \pm 2n\pi$ (n is an integer) between the two images, the optical path length difference between the two orthogonal polarizations can be set by tilting the shear plate, or pre-adjusted by a variable compensator. The output analyzer produces the desired interference between the two orthogonally polarized images. In practice, the shear plate tilt (and/or compensator) and the analyzer are tuned for maximum extinction of the reflected light from an unwritten area on the disc in order to obtain the highest possible contrast ratio in the output image.

Design Considerations

The optical readout configuration shown in Fig. 2 is very flexible in design. The crucial element, the shear plate, can be manufactured in several ways. For example, a uniaxial crystal plate with a tilted optic axis can be used as shown in Fig. 3. The shear, δ , of such a plate is given by

$$\delta = \frac{d(n_e^2 - n_o^2) \tan \theta_c}{n_o^2 + n_e^2 \tan^2 \theta_c},$$

in which d is the width of the plate, θ_c is the optic axis tilt angle, and n_o and n_e are the ordinary and extraordinary indices of refraction of the uniaxial crystal, respectively. Since δ is to first order directly proportional to $\Delta n = n_e - n_o$, and since the shear required (μm -range) is $\ll d$, it is desirable to choose a uniaxial crystal with a small birefringence. Quartz, for example, has an index difference $\Delta n \approx 0.009$ (at $\lambda = 632.8 \text{ nm}$), and in addition exhibits good polishing characteristics.

SLM Output Characteristics

As test patterns, three bands were written on two different types of discs: write-once ablative media discs (amorphous alloy) of the dark mark type, and erasable bump forming media (dye polymer)³. Each band was approximately 0.5 mm in width and written at a duty cycle corresponding to a specific grey scale level. The three duty cycles used were 50% (maximum output intensity), 25% and 12.5%, all written using 5 μm long bits. Using the differential interferometric readout configuration with a 5 μm shear, the three bands for both media were clearly extracted with the appropriate grey scale level factor of two between adjacent bands. The measured output intensity normalized to the highest grey scale level (the 50% duty cycle band) as a function of duty cycle for both media is shown in Fig. 4. The SLM output characteristic is seen to be very linear for both media. However, the currently obtained contrast ratio is much higher for the bump forming medium ($\approx 20:1$) than for the ablative medium ($\approx 5:1$). The main reason for the lower contrast ratio for the ablative medium is due to interference from the other (transmissive) recording layer on the two-sided optical disc. The ablative recording layer is furthermore read out through a non-AR coated polycarbonate substrate (1.2 mm thick) that may exhibit some residual birefringence. The bump forming medium, on the other hand, has a non-transmissive air incident recording layer. The contrast ratio of the bump forming medium could be further increased by using high quality (glass) substrates and higher uniformity recording layers with fewer (depolarizing) surface blemishes.

Differential Interferometric Readout Performance

The maximum throughput efficiency for an optical disc medium in an otherwise ideal differential interferometric readout configuration is given by

$$\eta_{\text{medium}} = \frac{1}{4} \left[|r_0|^2 + |r_1|^2 - 2 \cdot |r_0| \cdot |r_1| \cdot \cos \left(2\pi n \cdot \frac{2h}{\lambda} + \arg\{r_0\} - \arg\{r_1\} \right) \right],$$

in which r_0 and r_1 are the (complex valued) amplitude reflectivities of the (unwritten) recording layer and the written bit respectively, h is the height of the written bit and n is the refractive index of the optical disc substrate (for substrate incident media). The highest obtainable efficiency is achieved for an (ideal) bump forming medium (shown in Fig. 5) for which $r_0 = r_1$ and $h =$

$\pm \lambda/4n$. That is, the efficiency is directly proportional to $|r_0|^2$ with a unity constant (the efficiency factor is unity). However, for an ideal dark mark ablative medium ($r_1 = 0$ and $h = 0$), the efficiency is directly proportional to $|r_0|^2/4$; that is, the efficiency factor is 1/4. Most ablative media in fact comprise amorphous alloy recording layers with complex valued indices of refraction characterized by a non zero bit reflectivity (r_1), which may in fact boost the effective efficiency factor beyond 1/4. For our experimental readout configuration, the total measured throughput efficiency (relative to the collimated laser beam) of the SLM for the two media was 0.9% for the bump forming medium and 3.0% for the ablative medium³. These measurements however included the losses in the readout optics (polarizing beam splitter, shear plate and imaging lenses) for which the throughput efficiency was independently measured to be 0.85. The remaining losses can be divided into two separate parameters: losses due to the optical disc medium (limited recording layer reflectivity, air-substrate interface reflection loss and diffraction from grooved substrates) and loss due to the data pattern itself, which involves the shape and size of the written bits (efficiency factor, fill factor of bits, and diffraction losses due to limited numerical aperture). The optical disc medium losses are mainly determined by the reflectivity of the recording layer, which was 40% for the ablative medium and 13% for the bump forming medium. In another experiment, the bump forming medium was also coated with a 30 nm aluminum layer, which boosted the reflectivity to 86% and the total throughput efficiency to 5.8%. The total data pattern efficiencies were approximately the same for both media (0.11 for the ablative and 0.08 for the bump forming). The relatively low data pattern efficiency for the bump forming medium is due to the non-ideal shape of the bumps and a slightly lower fill factor of ≈ 0.40 versus 0.63 for the ablative medium. By increasing the fill factor and the numerical aperture of the readout optics, and by optimizing the AR-coatings, the (complex valued) recording layer reflectivity (e.g. by using separate write and parallel read wavelengths) and the bump shape (if applicable), the total throughput efficiency can be substantially increased.

References

1. J. H. Rilum and A. R. Tanguay, Jr, "Utilization of Optical Memory Discs for Optical Information Processing Applications", Tech. Dig., 1988 Annual Meeting of the Optical Society of America, Santa Clara, California, (1988).
2. D. Psaltis, M. A. Neifeld, and A. A. Yamamura, "Image Correlators Using Optical Memory Disks", Opt. Lett., 14 (9), 429-431, (1989).
3. J. H. Rilum and A. R. Tanguay, Jr, "Performance Characteristics of Optical Memory Disc Spatial Light Modulators", Tech. Dig., 1989 Annual Meeting of the Optical Society of America, Orlando, Florida, (1989).

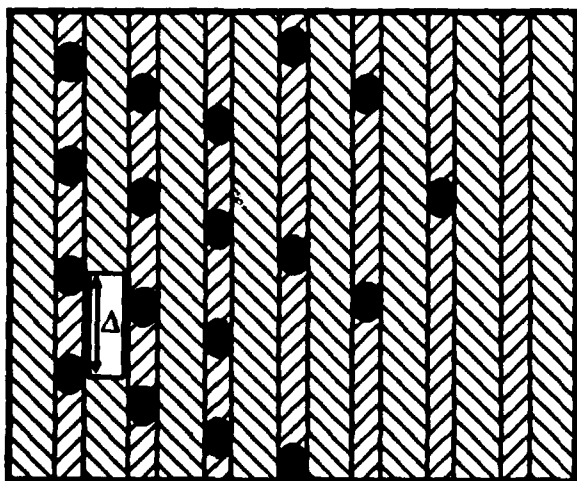


Figure 1a. Bit pattern on an optical disc.

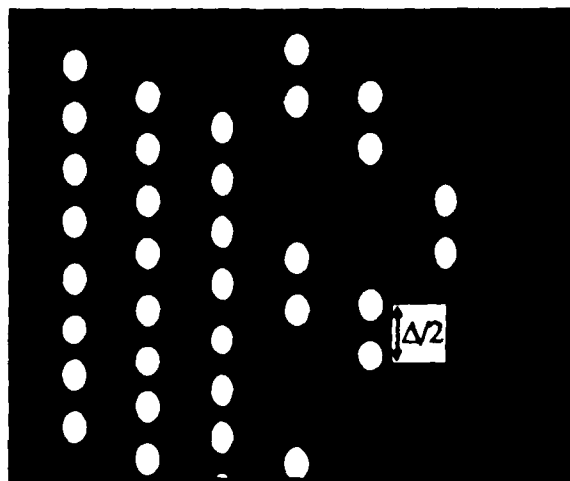


Figure 1b. Corresponding output image.

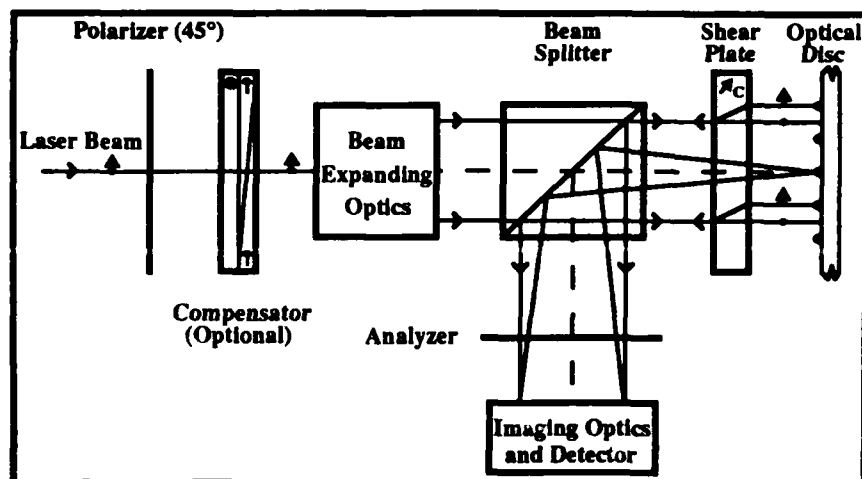


Figure 2. Optical readout configuration.

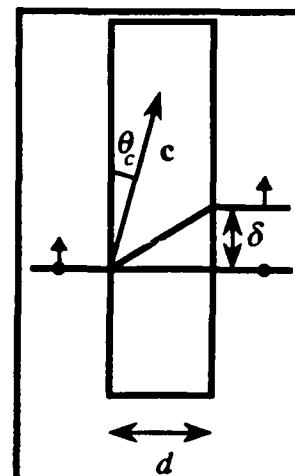


Figure 3. Shear plate.

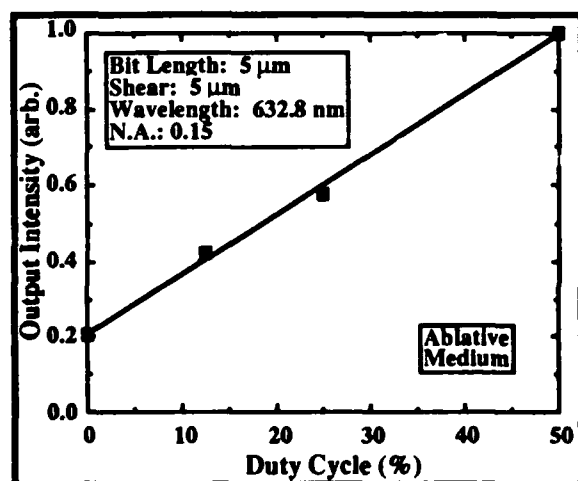


Figure 4a. SLM output characteristic.

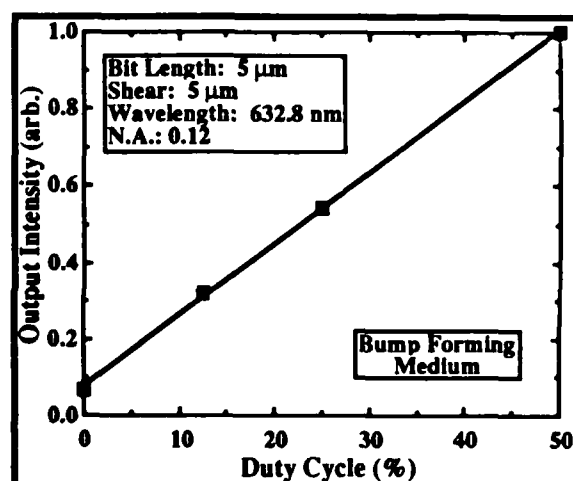


Figure 4b. SLM output characteristic.

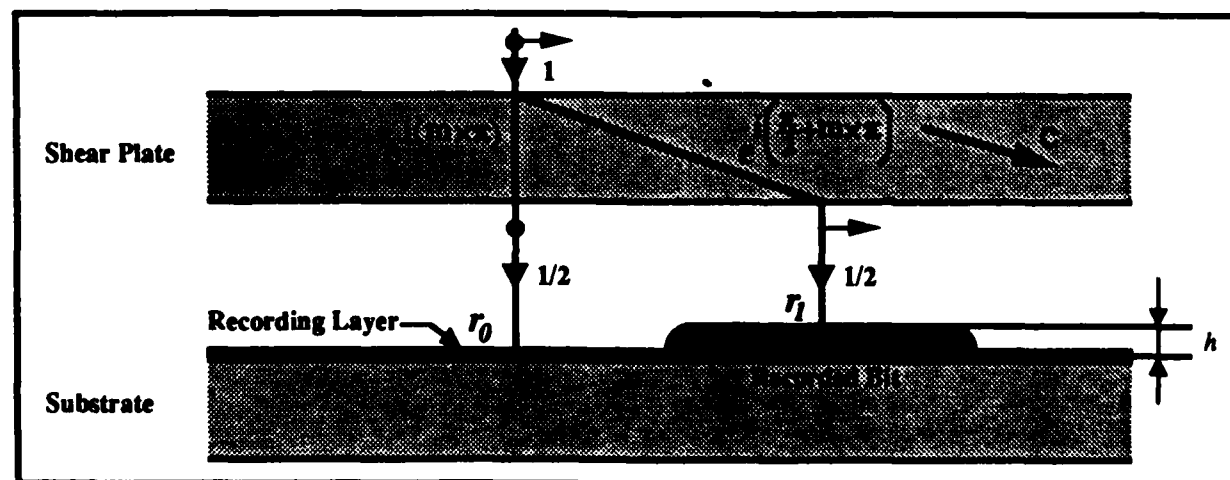


Figure 5. Readout geometry of a bump forming optical disc medium.