National Library
of Canada

Acquisitions and
Bibliographic Services Branch

395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale
du Canada

Direction des acquisitions et
des services bibliographiques

395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file - Votre référence

Our file - Notre référence

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

Canada

**OPTICAL CLOCK DISTRIBUTION USING
COMPUTER GENERATED HOLOGRAPHIC OPTICAL ELEMENTS**

by

Ryan P. MacDonald, B.Sc.

**A thesis submitted to the
Faculty of Graduate Studies and Research
in partial fulfilment
of the requirements for the degree of
Master of Engineering**

**Ottawa-Carleton Institute for Electrical Engineering
Department of Electronics
Faculty of Engineering
Carleton University**

May 1993

**© copyright
1993, Ryan P. MacDonald**

National Library
of Canada

Acquisitions and
Bibliographic Services Branch

395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale
du Canada

Direction des acquisitions et
des services bibliographiques

395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file   Votre référence

Our file   Notre référence

The author has granted an irrevocable non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of his/her thesis by any means and in any form or format, making this thesis available to interested persons.

L'auteur a accordé une licence irrévocable et non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de sa thèse de quelque manière et sous quelque forme que ce soit pour mettre des exemplaires de cette thèse à la disposition des personnes intéressées.

The author retains ownership of the copyright in his/her thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without his/her permission.

L'auteur conserve la propriété du droit d'auteur qui protège sa thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

ISBN 0-315-89824-0

Canada

Name Ryan P. MacDonald

Dissertation Abstracts International is arranged by broad, general subject categories. Please select the one subject which most nearly describes the content of your dissertation. Enter the corresponding four-digit code in the spaces provided.

Engineering, Electronics

SUBJECT TERM

0544

SUBJECT CODE

U·M·I

Subject Categories

THE HUMANITIES AND SOCIAL SCIENCES

COMMUNICATIONS AND THE ARTS

Architecture 0729
Art History 0377
Cinema 0900
Dance 0378
Fine Arts 0357
Information Science 0723
Journalism 0391
Library Science 0399
Mass Communications 0708
Music 0413
Speech Communication 0459
Theater 0465

EDUCATION

General 0515
Administration 0514
Adult and Continuing 0516
Agricultural 0517
Art 0273
Bilingual and Multicultural 0282
Business 0688
Community College 0275
Curriculum and Instruction 0727
Early Childhood 0518
Elementary 0524
Finance 0277
Guidance and Counseling 0519
Health 0680
Higher 0745
History of 0520
Home Economics 0278
Industrial 0521
Language and Literature 0279
Mathematics 0280
Music 0522
Philosophy of 0998
Physical 0523

Psychology 0525
Reading 0535
Religious 0527
Sciences 0714
Secondary 0533
Social Sciences 0534
Sociology of 0340
Special 0529
Teacher Training 0530
Technology 0710
Tests and Measurements 0288
Vocational 0747

LANGUAGE, LITERATURE AND LINGUISTICS

Language 0679
General 0289
Ancient 0290
Linguistics 0291
Modern 0401
Literature 0294
General 0295
Classical 0297
Comparative 0298
Medieval 0316
Modern 0591
African 0305
American 0352
Asian 0355
Canadian (English) 0593
Canadian (French) 0311
English 0312
Germanic 0315
Latin American 0313
Middle Eastern 0314
Romance 0401
Slavic and East European 0294

PHILOSOPHY, RELIGION AND THEOLOGY

Philosophy 0422
Religion 0318
General 0321
Biblical Studies 0319
Clergy 0320
History of 0322
Philosophy of 0469
Theology 0323

SOCIAL SCIENCES

American Studies 0323
Anthropology 0324
Archaeology 0326
Cultural 0327
Physical 0310
Business Administration 0272
General 0770
Accounting 0454
Banking 0338
Management 0385
Marketing 0501
Canadian Studies 0503
Economics 0505
General 0508
Agricultural 0509
Commerce-Business 0510
Finance 0511
History 0358
Labor 0366
Theory 0351
Folklore 0357
Geography 0351
Gerontology 0578
History 0578

Ancient 0579
Medieval 0581
Modern 0582
Black 0328
African 0331
Asia, Australia and Oceania 0332
Canadian 0334
European 0335
Latin American 0336
Middle Eastern 0333
United States 0337
History of Science 0585
Law 0398
Political Science 0615
General 0616
International Law and Relations 0617
Public Administration 0814
Recreation 0452
Social Work 0626
Sociology 0627
Criminology and Penology 0938
Demography 0631
Ethnic and Racial Studies 0628
Individual and Family Studies 0629
Industrial and Labor Relations 0630
Public and Social Welfare 0700
Social Structure and Development 0344
Theory and Methods 0709
Transportation 0999
Urban and Regional Planning 0453
Women's Studies 0453

THE SCIENCES AND ENGINEERING

BIOLOGICAL SCIENCES

Agriculture 0473
General 0285
Agronomy 0475
Animal Culture and Nutrition 0476
Armal Pathology 0359
Food Science and Technology 0478
Forestry and Wildlife 0479
Plant Culture 0480
Plant Pathology 0817
Plant Physiology 0777
Range Management 0746
Wood Technology 0306
Biology 0287
General 0308
Anatomy 0309
Biostatistics 0379
Botany 0329
Cell 0353
Ecology 0369
Entomology 0793
Genetics 0410
Limnology 0307
Microbiology 0317
Molecular 0416
Neuroscience 0433
Oceanography 0821
Physiology 0778
Radiation 0472
Veterinary Science 0786
Zoology 0760
Biophysics 0425
General 0996
Medical 0996

EARTH SCIENCES

Biogeochemistry 0425
Geochemistry 0996

Geodesy 0370
Geology 0372
Geophysics 0373
Hydrology 0388
Mineralogy 0411
Paleobotany 0345
Paleoecology 0426
Paleontology 0418
Paleozoology 0985
Polymology 0427
Physical Geography 0368
Physical Oceanography 0415

HEALTH AND ENVIRONMENTAL SCIENCES

Environmental Sciences 0768
Health Sciences 0566
General 0300
Audiology 0992
Chemotherapy 0567
Dentistry 0350
Education 0769
Hospital Management 0758
Human Development 0982
Immunology 0564
Medicine and Surgery 0347
Mental Health 0569
Nursing 0570
Nutrition 0380
Obstetrics and Gynecology 0354
Occupational Health and Therapy 0381
Ophthalmology 0571
Pathology 0419
Pharmacology 0572
Pharmacy 0382
Physical Therapy 0573
Public Health 0574
Radiology 0575
Recreation 0575

Speech Pathology 0460
Toxicology 0383
Home Economics 0386

PHYSICAL SCIENCES

Pure Sciences 0485
Chemistry 0749
General 0486
Agricultural 0487
Analytical 0488
Biochemistry 0738
Inorganic 0490
Nuclear 0491
Organic 0494
Pharmaceutical 0495
Physical 0754
Polym 0405
Radiation 0605
Mathematics 0986
Physics 0606
General 0608
Acoustics 0748
Astronomy and Astrophysics 0607
Atmospheric Science 0798
Atomic 0759
Electronics and Electricity 0609
Elementary Particles and High Energy 0610
Fluid and Plasma 0752
Molecular 0756
Nuclear 0611
Optics 0463
Radiation 0346
Solid State 0984
Statistics 0984
Applied Sciences 0346
Applied Mechanics 0984
Computer Science 0984

Engineering 0537
General 0538
Aerospace 0539
Agricultural 0540
Automotive 0541
Biomedical 0542
Chemical 0543
Civil 0544
Electronics and Electrical 0348
Heat and Thermodynamics 0545
Hydraulic 0546
Industrial 0547
Marine 0794
Materials Science 0548
Mechanical 0743
Metallurgy 0551
Mining 0552
Nuclear 0549
Packaging 0765
Petroleum 0554
Sanitary and Municipal 0790
System Science 0428
Geotechnology 0796
Operations Research 0795
Plastics Technology 0994
Textile Technology 0994

PSYCHOLOGY

General 0621
Behavioral 0384
Clinical 0622
Developmental 0620
Experimental 0623
Industrial 0624
Personality 0625
Physiological 0989
Psychobiology 0349
Psychometrics 0632
Social 0451

Carleton University
Ottawa, Canada K1S 5B6

**Thesis contains black & white
photographs &/or explanatory tables
which when microfilmed may lose their
significance. The hardcopy of the
thesis is available upon request from
Carleton University Library.**

**CARLETON UNIVERSITY
FACULTY OF GRADUATE STUDIES AND RESEARCH**

**The undersigned recommend to the Faculty of Graduate Studies
and Research acceptance of the thesis**

**"OPTICAL CLOCK DISTRIBUTION USING
COMPUTER GENERATED HOLOGRAPHIC OPTICAL ELEMENTS"**

**submitted by Ryan P. MacDonald, B.Sc.
in partial fulfilment of the requirements for
the degree of Master of Engineering.**

Thesis Supervisor

Thesis Supervisor

Chairman, Department of Electronics

**Carleton University
August 1993**

Abstract

Modern digital computers consist of hundreds of thousands of transistors interconnected with each other. As the speed and complexity of these computers continues to increase, the limitations imposed by electrical interconnections are being approached. Electrical interconnections limit the speed of communications between logic elements and occupy valuable area on both circuit boards and ICs. Free space optical interconnections have the capability of eliminating many of the problems associated with electrical interconnections.

An optical clock distribution system employing computer generated holographic optical elements is built and the procedure as well as the test results of this system is detailed in this thesis. Two holographic optical elements (HOEs) were built for use in the system: a binary-phase, reflection type and a quaternary-phase, reflection type. The frequency of the clock signal used is 1.0 GHz. The results of the system test are encouraging: without electronic amplification, the recovered clock signal has a SNR of 21 dB in the case of the binary-phase HOE, and an SNR of 28 dB when employing the quaternary-phase HOE. The increase in SNR is explained by the quaternary-phase HOE's measured efficiency of 69 %, as compared to 35 % for that of the binary-phase HOE. As well, the characteristics of the first Bessel beam produced by illumination of an HOE with light from a diode laser are reported.

Acknowledgements

I wish to thank Dr. Barry Syrett and Dr. Jacek Chrostowski for being my thesis supervisors. Much credit is due them for suggesting the research topic and guiding me through to the end of the project. I would also like to thank Dr. Simon Boothroyd for his great help in parts of this research.

I would like to thank the National Research Council of Canada for allowing me to use their vast facilities while undertaking this research. I would also like to recognize the help of all the people in Photonics section of the Autonomous Systems Laboratory at NRC, especially that from Les Chan, Alfred Dufour, and Ross Misner, and Peter Myslinski. Thanks is also due to Lyle Berndt and Carol Adams of Carleton's integrated circuit fabrication facility.

The financial assistance of the Natural Sciences and Engineering Research Council, and of DEC Canada must be acknowledged, without which this research would not have been possible.

Finally I would like to thank my parents for all of their support and encouragement over the years.

TABLE OF CONTENTS

Abstract	iii
Acknowledgements	iv
List of Symbols and Abbreviations	vii
Chapter 1 INTRODUCTION	1
1.1 Introduction	1
1.2 Thesis Objectives	4
Chapter 2 ELECTRICAL VS. OPTICAL INTERCONNECTIONS	6
2.1 Introduction	6
2.2 Electrical	7
2.2.1 Superconductors	9
2.3 A Comparison of Optics and Electronics	9
2.4 Optical Interconnections	15
2.4.1 Guided Wave Propagation	16
2.4.2 Free Space Optical Interconnections	19
Chapter 3 WAVE THEORY OF LIGHT	22
3.1 Introduction	22
3.2 Wave Theory	23
3.2.1 Interference	26
3.2.2 Diffraction	32
Chapter 4 THEORY OF HOLOGRAPHY.....	45
4.1 Introduction	45
4.2 Conventional Holography	46
4.3 Phase Holography	49
4.4 Fourier Transform Holography	54
4.5 Computer Generated Holograms	57
4.5.1 Direct Computer Generated Holograms	57
4.5.2 Computer Generated Fourier Holograms	59
4.5.3 Optimized Holograms	64
Chapter 5 ERRORS IN HOLOGRAMS	66
5.1 Introduction	66
5.2 Description of Fabrication Errors	68
5.3 Mask Alignment Error in the Detour-Phase Kinoform	69
5.3.1 The Detour Phase Kinoform.....	69
5.3.2 Alignment Error in the Detour-Phase Kinoform	77

5.4 Effects of Errors in Phase-only Holograms	84
Chapter 6 DIFFRACTION-FREE BEAMS	91
6.1 Introduction	91
6.2 Bessel Beams	92
6.2.1 Bessel-Gauss Beams	95
6.3 Original Method of Creating the J_0 Beam	98
6.4 Holographic Generation of Bessel Beams	100
6.5 Measurement and Characteristics of Bessel Beam	104
Chapter 7 DESIGN AND TEST OF OPTICAL CLOCK DISTRIBUTION SYSTEM ..	112
7.1 Optical Beam Routers	112
7.1.1 Multifacet Holograms	114
7.2 One to Four Phase Hologram	114
7.3 Electronic Circuitry for Electro-Optic Conversion	123
7.4 Characterization of HOEs	126
Chapter 8 CONCLUSIONS AND FUTURE WORK.....	133
References	135
Appendix A	140

List of Symbols and Abbreviations

A	amplitude of electric field	β	propagation constant
A_0	2-D Fourier transform of U	ΔR	reconstruction error
a	power in a diffraction order	δ	size of cell
E	electric field	η	diffraction efficiency
ϵ	radius of sphere enclosing P_0	θ_c	cellular phase function
ξ	time independent part of electric field	θ_s	step phase function
E_0	optical energy	Λ	optical path length
f	focal length	λ	wavelength
G	Green's function	ρ	radial coordinate
h	height	τ	duration of pulse
I	intensity	ϕ	phase function
J_0	zero order Bessel function of first kind	ϕ'	time independent phase
K	quantization level	ω	angular frequency
k	wave vector	φ	phase constant
L	number of cells per period	∇	gradient
l	thickness		
Mf	merit function		
n	refractive index		
O	wavefront on a surface		
P	power		
ρ	desired power		
P_E	power efficiency		
Q	slope factor		
R	distance from source		
S	surface		
T	wave period		
t	transmission function		
U	optical wave amplitude		
V	volume		
W	wavefront away from surface		
w_0	Gaussian beam waist		
Z_0	characteristic impedance		
z_{max}	maximum propagation distance		

Chapter 1

INTRODUCTION

1.1 Introduction

Since the development of the kinoform and other types of computer generated holograms in the late nineteen sixties, interest in them has been, for the most part, concentrated on their applications as spatial filters, optical memories, aberration compensators, and display elements of nonexistent objects (for example, holograms on credit cards and banknotes). Lately however, they have received renewed interest as beam routers for free space optical interconnection systems in electronic and all-optical computing. In fact, in the domain of all-optical computing, the area of free space interconnection of planes of logic elements stands out as having the greatest potential for increasing the speed of computing[1.1].

It is clear that there is a limit to the operating speed of all-electronic computers, and with continual improvements in VLSI technology that limit is quickly being approached [1.2]. The most significant bottleneck in operating speed comes from the electrical interconnection of logic elements, the logic elements themselves being quite fast[1.3]. With improvements in VLSI technology, gate delays have been reduced due to the ever shrinking minimum feature size obtainable with VLSI, while at the same time chips have grown in size, complexity and number, creating larger interconnect delays. In this way, eventually the maximum speed at which the computer can run will be limited by the interconnect delays rather than by gate delays.

In order to decrease the electronic interconnect delay, one must achieve at least one of two solutions: either increase the data rate on each existing line or, increase the number of connection lines while maintaining the same data rate. Both of these solutions are very costly and difficult to implement electronically with adequate reliability.

For several reasons optical interconnections can eliminate the bottleneck and increase the operating speed of a computer nearly a thousandfold. In typical integrated circuits and circuit boards, interconnections are made via metallic lines. Signals on metallic lines are slowed by capacitive loading, and this problem increases with line length, as the lines exhibit a capacitance per unit length, and with the size of the load capacitance that the line must drive. The problem is that before the signal can reach the intended logic element, the total capacitance along the line must first be charged up - an operation which requires precious time, so the delay is proportional to capacitance.

Another problem inherent with electrical interconnections is the excessive power consumption required to drive capacitive loads at reasonably high speeds. This effect is compounded when fanout (the delivery of the signal to more than one destination) is required. Electrical lines are also susceptible to electromagnetic crosstalk unless properly shielded (electrical lines in ICs are not). They can handle signals of only relatively small bandwidth, and they take up much valuable board area. In fact, board area and power requirements are increasingly dominated by interconnect drivers, receivers and pads. In contrast, optical interconnections can deliver wide bandwidth signals and large fanouts, while consuming little power. They are also immune to interference and crosstalk; note how two beams of light are able to pass through each other and emerge unaffected.

Optical interconnections can be implemented by optical fibre or by simple free space propagation. Connections made by free space optical propagation can be made in the space above the processors allowing parallel connections, and they have the capability of addressing any location on the board - without the need for fibre connectors.

Within a computer system, there exist five levels of interconnection:

- 1) level 1: inside an integrated circuit chip
- 2) level 2: between undiced chips on the same wafer
- 3) level 3: between different IC chips on a common board
- 4) level 4: between different boards
- 5) level 5: between different computers

At levels one and two, electrical interconnects seem to be unchallenged for performance. However, speed can be greatly increased when optical interconnects are used in levels three, four and five. This work deals with interconnections at level three, specifically using free space optical interconnections with holographic optical elements to distribute a high speed clock signal.

The reason for the poor performance of electrical interconnections appears to stem from its impedance mismatch with its logic source and termination. A small logic element will have a low capacitance, a high impedance, and will require a low switching energy ($\ll 1\text{pJ}$). Match this to the high capacitance, low impedance electrical transmission line, which requires much more switching energy (approximately 20 pJ for the transmission of a 1 volt, 1 ns pulse along a 50 Ohm line) and one can see how the problem arises. On the other hand, optical interconnections have matched impedances

[1.3]. The capacitances of the photodetector and laser diode are of the same order as that of the logic devices, and photodetectors require very little power to reproduce a signal, minimizing power requirements. The question of impedance matching will be dealt with in greater detail in the next chapter.

Apart from being useful in optical interconnection systems, holograms, or the so-called holographic optical elements (HOEs), can produce optical patterns which conventional optical elements (COEs) produce either inefficiently or impractically. Thus HOEs can be produced which are cheaper, more efficient or easier to use than some COEs performing identical functions. A novel HOE of this type will be discussed in chapter six.

1.2 Thesis Objectives

This work is aimed at assessing the applications of holographic optical elements (HOEs) as beam routers in an optical clock distribution system and as elements which transform laser diode light into non-diffracting beams. The required HOEs are produced and tested, along with the necessary electronic components of an optical interconnection system. A computer algorithm is created which produces the required pattern for producing the various holograms using standard photolithographic techniques. Such computer generated holograms are useful for creating arbitrary optical patterns which are difficult, if not impossible, to form using standard holographic techniques.

In order to communicate the advantages and limitations of optical interconnects, chapter 2 focusses on comparing and contrasting optical interconnect technologies with

those of electrical interconnections. It explains for which applications optical interconnections are best suited, and for which applications electrical interconnections should remain in use.

Chapters 3 and 4 deal with the wave theory of electromagnetic radiation and the theory of holography, respectively. As holographic theory is based on electromagnetic wave theory, these two chapters provide an introduction to holographic theory.

The production of multilevel phase holograms, and the associated fabrication errors are dealt with in chapter 5. Error models are created and transferred to computer algorithms which model the effects of the various fabrication errors on the performance of HOEs.

Chapter 6 describes non-diffracting beams and the first hologram which produces a non-diffracting beam from the light emitted by a diode laser. The design and test results of the HOE are given.

The design of an optical clock distribution system is given in chapter 7. The system HOE is produced and tested, as is the entire clock distribution system. All test results are given. Finally, chapter 8 gives the thesis conclusions as well as recommendations for further work.

Chapter 2

ELECTRICAL VS. OPTICAL INTERCONNECTIONS

2.1 Introduction

Interconnections provide communication links between logic elements; whether the communicating logic elements are the gates on a silicon chip separated by only a hundred microns, or whole computer systems half a world apart, the interconnections are a vehicle for the same basic requirement: communication. And while all levels of interconnection provide for the same need in electronic systems, not all levels will be considered in this thesis. The areas of interconnection which will be considered are those between different integrated circuits on a circuit board, although the work could be extended to the interconnections between boards via the backplane. The types of interconnections not considered in this work are those connections involving long distance communication such as terrestrial links; intrachip connections will likewise not be considered.

Long distance connections are already enjoying the benefits of optics; note the replacement of expensive copper cables, previously solely utilized by telephone companies to provide long distance service, by fibre optic lines which are cheaper, lighter, and can handle a much greater flux of information. Optical backplanes also exist which rely on fibre optics for their connections. And while optical interconnections on-chip, and between chips are not yet a reality, they are in the research stage, as this thesis can attest, and should soon be available.

The remainder of this chapter will compare and contrast conventional, and not so conventional, electrical connections with optical connections on the scale of chip to chip and smaller distances. Areas to be looked at include speed and power consumption.

2.2 Electrical

Electrical interconnections seem to be well suited for communications within individual chips; the disadvantages of electrical interconnections seem to wane as the connection distance shrinks. These interconnections are produced using standard VLSI technology, which is well established, inexpensive and reliable. Interconnects at these levels have high fanout values. Microstrips are good at this level and can handle signals with frequencies of several gigahertz. Outside of the chip, however, where electrical interconnections do not perform so well, they take the form of coaxial cable, twisted wire pairs, and printed circuit boards.

Fanout poses a problem for electrical connections, and unfortunately inside the chip the fanout is generally quite high, whereas outside the chip it tends to be low. The problem is that unless each fanout line is terminated with the correct impedance, reflections will be generated at each load. While it is possible in principle to customize each fanout with ideal terminations at each load, in practice custom designing individual interconnects is undesirable and sometimes impossible because of the physical limitations of the interconnects. Another problem associated with increasing fanout is that the propagation delay of a signal through the interconnect increases due to increasing capacitive loading [2.1]. Electrical interconnections also have trouble providing high

density interconnections. As the length of electrical interconnections, and with it the interconnection density, increase from the levels found within the chip, the disadvantages become increasingly evident; power dissipation and RC time delays increase with connection length and fanout.

Several factors limit the densities possible with electrical interconnects. First of all, the interconnects are confined to pseudo-planar structures such as printed circuit boards and backplanes - they are unable to fully utilize the third dimension. Even multi-level printed circuit boards with vias between layers have their communication links confined to the same planes as their logic elements. The limitations imposed by crosstalk restrict the spacing of adjacent lines, and the density problem is being aggravated by the increasing speeds of the logic devices. As speeds increase, signal frequencies likewise increase, creating increased susceptibility to crosstalk and higher losses from skin effect. Higher interconnect density means closer spaced lines as well as thinner lines; since the resistance of a line is proportional to its cross-sectional area, the result is higher resistance lines and a higher RC time constant associated with them.

Electromagnetic crosstalk considerations limit the proximity of adjacent connection lines and make it difficult to build lines which cross. This makes the layout of interconnection lines a tedious task, and causes the connections on a printed circuit board to take up a vast amount of area, reducing the amount available for processing.

2.2.1 Superconductors

The new breed of high temperature superconductors have potential as interconnect materials, but although seemingly obvious, it must not be assumed that the simple substitution of superconductors for copper or aluminium interconnect lines will increase communication speeds. Simply reducing the resistance of the interconnect lines will increase neither communication nor computer speed, however power consumption by communication lines will cease. Note also that the latest superconductors must be cooled with liquid nitrogen, which boils at 77 K, and that at this temperature ordinary copper is an excellent conductor [2.2].

Areas where superconductors have potential to improve the performance of computers are in interconnect density and signal dispersion. Superconductors resist the penetration of magnetic fields, and so they are resistant to electromagnetic crosstalk. As a result, superconducting interconnects can be placed closer together than their normally conductive counterparts, while remaining free from crosstalk problems. Also, the phase velocity of an electrical wave propagating along a superconducting line is independent of frequency up to extremely high frequencies. This is because the superconductor's penetration depth is independent of frequency and hence losses due to skin effect are constant to very high frequencies [2.2]. In contrast, normal conductors suffer from signal dispersion because their penetration depths are frequency dependent.

2.3 A Comparison of Optics and Electronics

It is useful, when comparing optical and electrical interconnects, to examine the

differences which exist between optics and electronics that stem from basic physics. These differences can help explain why optics is sometimes more useful than electronics, and how each may best be exploited.

First of all, there are many similarities between optics and electronics. Both may be used to carry information. Optics carries information via photons or electromagnetic waves which travel at speeds of 3×10^8 m/s in a vacuum. Electronics also uses electromagnetic waves to send information, and not the electron itself, although the waves are supported by electrons. In lossy lines the velocity of the waves might be less than 1/100 of the speed of light, depending on the line resistance and capacitance, whereas low loss transmission lines can carry the waves at close to the speed of light.

In order to perform logic functions a nonlinear interaction is required. In electronics, logic functions are performed by the electrostatic effects of electron interaction, which causes a nonlinear interaction between electrons. In the case of optical logic, the nonlinear interaction between two light beams is almost always performed by the small movements of electrons inside nonlinear optical materials [2.3]. So one can see that in both optics and electronics, communications is done through photons or electromagnetic waves, and interaction (logic) is performed by electrons.

There is, however, a major difference between optics and electronics. The frequency of optical waves is much greater than that of electronic waves. Visible optical waves have a frequency of about 5×10^{14} Hz, whereas electronic waves range from 10^6 Hz to several GHz under the right conditions. Likewise, there is also a difference in wavelengths; the wavelength of visible light is around 0.5 microns compared with 30 m

for 10 MHz and 30 cm for 1 GHz electronic waves. By using Planck's law, one can see that the energy of one photon from the visible spectrum is about 2 eV, while the energy of a 100 MHz photon is about 4×10^{-7} eV; note that room temperature thermal fluctuation energies are about 25 meV, so visible spectrum photonic energy is easily seen above the thermal noise while electronic photon energies are buried in the thermal noise.

An important consequence of the disparities in wavelengths of electronic frequency and optical frequency waves is the method of wave confinement in each case. Because optical wavelengths are so small, it is possible to build guides which are larger in cross-section than the wavelength; it is therefore possible to use dielectric guides. However, due to the large wavelength of electronic frequency waves, it is impractical to build guides where the cross-section is larger than the wavelength, hence conductors must be used to confine these waves. One problem with using conductors is that they exhibit high loss, even superconductors have losses when carrying an oscillating current. Dielectric guides, on the other hand, can have very low loss.

Another fundamental difference between optics and electronics lies in the nature of the carriers. The optical carrier, the photon, is a boson, while the electronic carrier, the electron, is a fermion. Fermions are bound by the Pauli exclusion principle, while bosons are not. This means that while no two electrons can occupy the same state, many photons may simultaneously exist in identical states. And this results in the electron's capability of interfering with other electrons, and the photon's inability to react to other photons. Note that a collision between two electrons is possible, and is indeed a common occurrence, while two photons will simply pass through each other, each unaffected by

the event. Electrons can also interact with other electrons at a distance through the mechanisms described by Maxwell's equations and known as crosstalk, whereas the only way light beams are able to crosstalk is through the scattering of light at imperfect surfaces. For this reason, light beams are immune to electromagnetic interference, and virtually free from crosstalk effects.

One distinction between optics and electronics that is often overlooked is in the impedance transformation between logic device and connection line [1.3]. Optical sources and detectors can operate as efficient impedance transformers, whereas the electrical connections between logic devices do not. The impedance transformation in optics results from the quantum nature of the process of creating one photon for each electron and vice versa. This efficient impedance transformation results in lower communication energy than that usually required using electrical connections.

The reason why electrical connections are not able to impedance match stems from the low impedance of free space (377Ω) [2.4]. In order to send logic signals from one logic device to another, a length of electrical connection equal to the length of the pulse must first be charged up; this holds even for superconductors. Since the capacitance of the lines decreases, and the impedance increases, as the separation of the two lines increases - and only logarithmically once the conductors have been separated by more than their lateral dimensions - the usual result is high capacitance per unit length (about 100 pF/m), low impedance (eg. 50Ω) electrical connections. To have lower capacitance, higher impedance lines, the conductors would have to be placed further apart, and because of limited available space this is not feasible. For example, a coaxial line having

an inner conductor of radius r_1 and an outer conductor of radius r_2 has an impedance

$$Z = \frac{Z_0}{2\pi\sqrt{\epsilon_r}} \ln\left(\frac{r_2}{r_1}\right) \quad (2-1)$$

where ϵ_r is the relative dielectric constant and Z_0 is the impedance of free space.

Lower capacitance, higher impedance lines are required to make an impedance match, saving time and energy. They are also required to eliminate reflections at each of the loads; these reflections cause ringing noise at high signal frequencies. The energy required to charge a resistive line of capacitance C to a voltage V_0 is at least [1.3]

$$E_R = CV_0^2 \quad (2-2)$$

where $(1/2)CV_0^2$ is the energy stored in the capacitor and an equal amount is assumed to be dissipated in the resistance during charging. Even in a lossless, properly terminated line, an energy of $E_L = V_0^2\tau/Z$ is required; if the time τ is the duration of the pulse, and Z_0 is the characteristic impedance of the line, the energy needed to send a TEM pulse is

$$E_L = V_0^2\tau/Z_0 \quad (2-3)$$

It may easily be shown that $E_L = CV_0^2$, where C is the capacitance of the length of the line charged at any given time by the pulse. This energy may be subdivided into $(1/2)CV_0^2$ each for the stored electric and magnetic energies, since they are equal in TEM propagation. Unfortunately, these energies are usually much greater than the switching energies of small logic devices. And given the large fractional bandwidth (bandwidth to signal frequency ratio) of electronic signals, it would be nearly impossible to properly

terminate each line with its characteristic impedance.

Optical devices, on the other hand, are capable of performing impedance transformations because they can be quantum devices. This impedance transformation matches the high impedances of small logic devices to the low impedance associated with signal propagation. For example, should a 2 μW beam fall onto a good photodiode having a 1 $\text{M}\Omega$ load across it, generating 1 μA photocurrent, then 1 V will be produced across the resistance. This is possible because the detector is a quantum detector; it produces one excited electron for each photon it absorbs. The classical power required to produce 1 V is, from equation (2-3), $P = E_L / \tau = nV_o^2 / Z_o$, which is about 2.7 mW for $n=1$, so the quantum nature of the photodetector is truly remarkable; it is a quantum detector for which the classical voltage is irrelevant. The relationship is no longer between voltage and power, but between current and power, or more specifically between the number of photons and the number of free electrons produced. And so once the load resistance is chosen, the impedance transformation is defined.

The minimum amount of optical energy E_o required to produce a voltage V_o across a photodetector of quantum efficiency β and capacitance C_D is [2.3]

$$E_o = \frac{h\omega V_o C_D}{\beta e} \quad (2-4)$$

where h is Planck's constant, ω is the angular frequency of the photon, and e is the electronic charge. This is essentially the energy required to generate enough electrons to discharge the capacitance of the photodiode. Although this is only the static result, the dynamic result is not much different as long as $RC_D \geq \tau$, where τ is the duration of the

light pulse. And, in fact, it might prove useful to use a slightly larger energy in order to allow the system to settle for a few time constants. Notice that this energy is independent of length. Recall that the communication energy required by the electrical line, $E_L = CV_o^2$, is directly proportional to line capacitance, and therefore directly proportional to length. One can therefore conclude that there exists a connection length above which optical communication requires the lesser energy, and below which electrical communication requires the lesser energy. This distance will, of course, depend on the technology being used, but is on the order of 100 μm to 10 mm [1.3].

A similar quantum impedance transformation can occur at the source end of the connection as well. An ideal quantum photon emitter will produce one photon for each excited electron. A laser diode comes close to this quantum efficiency, but they are only highly efficient while running at high energy, due to their high threshold currents. Low threshold lasers are now feasible however, and they will be the best choice for low energy communications.

2.4 Optical Interconnections

As argued in the previous section, light makes an excellent information carrier. Its ability to cross other beams of light unaffected, its high bandwidth, and its immunity from electromagnetic interference combine to make it ideal for use in communications. But optical interconnects are more complex than conventional electrical interconnects. Optoelectronic converters are required at both the source and the destination of the signal, these take the form of laser diodes and photodetectors, respectively. In order for

information to be transported from source to destination, photons must be directed from laser to detector; this criterion may be met by way of either guided wave propagation or free space wave propagation.

2.4.1 Guided Wave Propagation

There exist two basic modes of guiding a light beam to a detector: through optical fibre and through planar wave guides. These interconnection schemes are sometimes referred to as *index guided optical interconnections* because it is the guide's refractive index profile which confines the beam. In this method, light from a single source is guided to many destinations via a waveguide. The waveguide could be either an optical fibre or a planar waveguide integrated onto or into the surface of an appropriate substrate.

If fibres are used to guide the signal, they must first be aligned to the detectors. One way of aligning several fibres and detectors at once is illustrated in Figure 2-1. A V-groove for each fibre is etched into a silicon alignment fixture. The fibres then fit inside the grooves and the block then aligns each of the fibres to its own detector [2.1,2.5]. The end of the fibre is polished at an angle of about 45° so that all of the light is reflected down onto the detector. The fibre is cylindrical in cross section, the beam will be focused into an elliptical spot on the detector, and a very small detector can thus be used. The small detector will be more responsive to higher frequency signals because of its lower capacitance.

Another method of aligning individual fibres to detectors is illustrated in Figure 2-2 [2.6,2.7]. In a process called laser assisted etching, a narrow steep-walled hole is

Figure 2-1. Fiber to detector alignment using silicon alignment block with v-grooves.

Figure 2-2. Bare fiber core fixed in place inside photodiode well.

etched into silicon. Standard VLSI techniques are then used to form a photodetector at the bottom of the well. Then the cladding is stripped from the end of an optical fibre and this end is inserted into the well and bonded in place with an epoxy adhesive. When completed, the fibre is precisely and accurately aligned to the detector. The complexity of this approach, however, may prove to be too great to be economically feasible.

Planar waveguides are another possibility for guiding waves to their destination. They are attractive because they can be fabricated by the same VLSI technology which produces the rest of the chip. Rib waveguides have the ability to cross each other without causing overly great signal deterioration [2.8], something which is not possible with electrical connections since the two lines would simply short each other out. This crossover ability will vastly simplify the process of connection routing compared to that of electrical connections. However, in order to save area, the planar waveguides must exhibit bends similar to those of the copper lines on a circuit board, and only recently have waveguides been fabricated which exhibit 90° bends and suffer only minor attenuation (0.4 dB) [2.8].

Since planar waveguide production is similar to IC fabrication, the two may be easily integrated and alignment problems will then be limited to the usual alignment problems associated with VLSI technology. The completed structure would then be relatively resilient to mechanical stress. Unfortunately, planar waveguides based on present technology are quite lossy (0.3 dB/cm) [2.8], and could therefore be used only for short connections. Their two-dimensional nature limits the density of the interconnections and the space available for logic. Also, waveguides are not well suited for high fanout

connections because the regular structures required to distribute optical power equally into all of the branches of the fanout are very difficult to fabricate in waveguides.

2.4.2 Free Space Optical Interconnections

The other major group of optical interconnects is the *free space* type and, as the name implies, the light for such interconnects is not guided by refractive index variations, but is instead allowed to propagate freely. The two major categories of free space interconnections are the unfocused and focused varieties.

An unfocused optical interconnection is made by simply illuminating the entire surface on which the detectors lay with the optical signal. This can be done by placing the optical source at the focal point of a collimating lens, and then placing the detectors in the path of the collimated beam, at normal incidence, as illustrated in Figure 2-3. Unfortunately, the system is very inefficient since only a small fraction of the incident light actually strikes the detectors.

A focused interconnection, the other type of free space connection, requires a focusing element; that element can be either a lens or a focusing hologram. Holograms can be much easier to fabricate than lenses, and can be highly efficient. Focused systems, however, require a high degree of alignment accuracy to ensure that the focused beams strike the proper spots. The arrangement shown in Figure 2-4 illustrates how the focusing element focuses the light from a source and distributes it to the detectors.

There are two major classes of holograms which can focus a light beam down to a point. Conceptually, perhaps the simplest is the type recorded on film plates. With this

Figure2-3. Unfocused signal distribution system

Figure 2-4. Focused optical signal distribution using a holographic optical element.

type, two or more laser beams interfere on the plane of the film which records this interference pattern. With this type of arrangement there are a few ways to make several focusing elements on one hologram. One is to replace each of the detectors in the system by a laser source, and then illuminate the film plate with the original source and all of the replacement sources. Unfortunately, this method is difficult to implement. Another method exposes the film plate to each of the replacement sources consecutively, while maintaining exposure to the original source. However, with this method each exposure tends to partially overwrite the previous recording. Yet another method involves exposing a different part of the film to each of the replacement sources, creating a multifaceted hologram. This method works well, but the multifaceted hologram generally produces a larger spot than the holograms of the previous two methods. Unfortunately, the process of making these optically recorded holograms on a large scale is too costly.

The other class of hologram is the computer generated hologram. These may be either the reflection or transmission type. They may also be either amplitude or phase holograms, or a combination of both. However, amplitude holograms are generally not as efficient as phase holograms since amplitude holograms must absorb some of the incident light. A multifaceted computer generated hologram requires no more fabrication steps than does a single faced one. Since free space interconnections must use the third dimension, these interconnections can be made above the plane of the chip, allowing more room for both interconnections and logic gates. This type of hologram can be produced at low cost and can provide a large degree of fanout.

Chapter 3

WAVE THEORY OF LIGHT

3.1 Introduction

The holographic process involves the total reconstruction of an object wavefront, not just the intensity portion as in a photograph, but the phase information as well. Indeed the word hologram is derived from the Greek *holo*, meaning whole, and *gram*, meaning message. A hologram captures all of the visual information contained within a wavefront, or the whole message. The resulting light field from a hologram is ideally indistinguishable from the original; it is as though the original object or wave field were being observed. It is this remarkable property which enables holograms to faithfully reproduce three-dimensional images.

The term wavefront refers to the surface over which the phase of the wave field is constant at any given instant of time, that is to say it is an instantaneous sampling of the waves on some surface of constant phase. The wavefront contains all of the three dimensional information about the waves. For this reason, all that is required to reproduce a given wave as a whole is to reconstruct its wavefront on some surface. After a wave has been forced to conform to the original on a surface, it will subsequently duplicate the propagation of the original [3.1.3.2].

In order to reconstruct a wavefront, two steps are required: first a method of recording the information contained in the light field is needed, and secondly there must exist a method of extracting this information to form an image. While there are many

ways of recording the necessary information onto a hologram, they always involve two elements of basic wave theory: interference and diffraction.

3.2 Wave Theory

Although light behaves sometimes as a wave and other times as a particle, as was postulated by Planck and Einstein, only the wave nature of light is observed in the holographic process, and so only the wave nature of light will be considered in this discussion of holography. As we now know, the wave theory of light, and all electromagnetic radiation, as first developed by Maxwell is an approximation to the quantum mechanical nature of photons. However, on a large scale, the energy distribution of a vast number of photons is equivalent to that of a classical electromagnetic wave. And so the wave theory of light remains a useful representation of electromagnetic phenomena.

The holographic principle, that of reconstructing wavefronts, applies to all waves, not just light or electromagnetic waves. Any wave which may be represented by the equation

$$\nabla^2 E(r,t) - \frac{1}{v^2} \frac{\partial^2 E(r,t)}{\partial t^2} = 0 \quad (3-1)$$

where v is the speed of the wave, can be used to produce a hologram. This is the scalar version of the wave equation and cannot represent polarization effects, but is otherwise adequate for our purposes. If the solutions to the wave equation are restricted to those having the form

$$E(r,t) = \mathcal{E}(r)e^{-i\omega t} \quad (3-2)$$

in other words single frequency, or monochromatic, waves, then the wave equation simplifies to the Helmholtz equation

$$\nabla^2 \mathcal{E}(r) + k^2 \mathcal{E}(r) = 0 \quad (3-3)$$

where

$$k^2 = \frac{\omega^2}{c^2} \quad (3-4)$$

One solution to the Helmholtz equation is the plane wave

$$\mathcal{E}(r) = \mathcal{E}_0 e^{ik \cdot r} \quad (3-5)$$

where r is the spherical coordinate, k is the vector with magnitude k , and $\mathcal{E}_0$ is a constant. It has identical values for each point in any plane which is normal to k . Another solution of the Helmholtz equation is

$$\mathcal{E}(r) = \frac{A}{r} e^{ikr} \quad (3-6)$$

which has a constant value on a sphere of radius r , centred at the origin, and is therefore termed a spherical wave. A point source produces a spherical wave. As one can see, the magnitude of the electric field is inversely proportional to the distance, r , from the source, and so the intensity, being proportional to the square of the electric field, decreases with distance according to the inverse square law.

The reader will notice that the wave and Helmholtz equations are linear, and so

if $\mathcal{E}_1$ and $\mathcal{E}_2$ are solutions to one of the above two equations, then $\mathcal{E}_3 = \mathcal{E}_1 + \mathcal{E}_2$ must also satisfy the same equation. Thus the principle of superposition applies to waves. This principle states that a wave E can be thought of as a sum of individual waves E_i such that $E = \sum E_i$. This is very important to holography. To determine the wavefront at some arbitrary surface all that is needed is a knowledge of the wavefront on a surface S . If the surface S is divided into points S_i with wavefronts denoted O_i such that the wavefront on S is $O = \sum O_i$, the wavefront at some point x_0 not on S is $W = \sum W_i$, where W_i is the wavefront at x_0 due to O_i . Since each O_i is a point source, the value of the wavefront W_i due to O_i can be expressed as

$$W_i = A_i \frac{e^{ikr}}{r} \quad (3-7)$$

where A_i is the magnitude of O_i on the point S_i . The wavefront at point x_0 can now be written as

$$W(x_0) = \int_S A(x) \frac{e^{ikr(x,x_0)}}{|x-x_0|} dx \quad (3-8)$$

And so, knowledge of the wavefront at one surface is enough to determine the wavefront on some arbitrary surface [3.1,3.2]. Holography takes advantage of this by forcing a wave to conform to a predetermined wavefront on some surface, thereby duplicating the original wavefront at any arbitrary point x_0 .

Another property of light, and of all waves, is that it is periodic; monochromatic light waves oscillate at some frequency f , while travelling in free space at the speed of

light c , and have a wavelength of $\lambda=c/f$. Therefore, at distances of whole multiples of its wavelength in the direction of propagation the phase of the wave is identical. However, in order to get a better understanding of the wave nature of light, one must study the phenomena of interference and diffraction.

3.2.1 Interference

The theory of interference flows from the theory of wave superposition. Interference generally refers to the production of large and small amplitude fringes representing the near-doubling and near-cancellation of amplitudes. Where the waves are in phase, the amplitudes are added; where they are out of phase, the amplitudes subtract. In optical interference, the electric field vectors superimpose algebraically and hence the resultant irradiance differs from the sum of the component irradiances (irradiance is proportional the square of the amplitude of the electric field vector).

The phase of the light from a beam can be expressed as

$$\phi = 2\pi\left(\frac{t}{T} - \frac{R}{\lambda}\right) + \phi = (\omega t - kR) + \phi \quad (3-9)$$

where ϕ is a constant of the source. When two beams of light are superimposed after travelling equal optical distances (optical path length is the material's refractive index multiplied by the actual path length) from their respective sources, the phase difference can be written as

$$\phi_2 - \phi_1 = (\omega_2 - \omega_1)t + (\phi_2 - \phi_1) \quad (3-10)$$

If interference is to be observed, then the above quantity must not change during the time,

T, of the observation. This means that the difference between the two frequencies ω_2 and ω_1 must be zero; in other words, the frequencies must be equal. It is clear from the above expression that if the difference between ϕ_2 and ϕ_1 remains constant, it is possible that the phase difference will not change. This is equivalent to saying that interference is possible when the two sources are not in phase with each other; the fringes of the interference pattern will be slightly shifted, but they will still exist. Such sources whose phase differences are always constant are said to be mutually coherent. The length of time over which the phase difference between two beams remains fairly constant (ie. the difference changes by less than 1 rad) is called the mutual coherence time. If the time of the observation is less than the mutual coherence time the two sources are said to be mutually coherent, and if the observation time is longer, the two sources are said to be mutually incoherent [3.3]. The length of time that an individual source oscillates in a predictable fashion is called the source's coherence time. In order to observe interference between two sources then, their frequencies must be equal and there must be coherence between the two sources.

If we consider just the electric field component E of the electromagnetic wave which, as before, is expressed as $E = Ae^{i\phi}$, the concept of interference can be illustrated by means of phasor addition. However, it should be understood that the phasor notation is simply a convenient way of representing the electromagnetic field, which is not complex. This form is used because of the way in which phases may be easily manipulated, but in the end, it is the real part of the phasor expression that is the physically significant quantity.

Consider two beams having fields

$$E_1 = A_1 e^{i\phi_1} \quad \text{and} \quad E_2 = A_2 e^{i\phi_2} \quad (3-11)$$

These two fields can be added either graphically or mathematically. Mathematically the addition is performed using Euler's theorem which states that

$$e^{i\theta} = \cos\theta + i \sin\theta \quad (3-12)$$

And so the two fields may be added as

$$\begin{aligned} E &= E_1 + E_2 \\ &= (A_1 \cos\phi_1 + iA_1 \sin\phi_1) + (A_2 \cos\phi_2 + iA_2 \sin\phi_2) \end{aligned} \quad (3-13)$$

If the frequencies of the phase terms are equal, then the resultant sum E will have the same frequency as well, this can be seen more easily using vector addition.

The two fields, E_1 and E_2 , can also be represented as phasors in the complex plane as illustrated in Figure 3-1(a). Since their frequencies are equal, the two phasors will rotate about the origin with angular frequency ω , but the angular difference, $\delta = \phi_2 - \phi_1$, will remain constant. From this, the reader will see that the frequency of the resultant vector will be the same as that of its two component vectors, however its amplitude and phase will be different. Using vector addition, the two components can be combined to give the resultant as illustrated in Figure 3-1b, this operation can be written as

$$\begin{aligned} E &= E_1 + E_2 = A e^{i\Phi} \\ &= A(\cos\Phi + i\sin\Phi) \end{aligned} \quad (3-14)$$

where A is the resultant amplitude and Φ is the resultant phase. At this point it should

(a)

(b)

Figure 3-1. (a) Two electric field vectors E_1 and E_2 in the complex plane. (b) The resultant E found through vector addition.

be noted that while both methods of addition yield the same result, vector addition is used only to illustrate the process of phasor addition. A mathematical treatment is required in a rigorous case.

The field $E=Ae^{i\Phi}$, resulting from the addition of its two component fields, can be separated into real and imaginary components. The following treatment of phasor addition roughly follows that of Klein and Furtak [3.4]. The amplitude and phase of the resultant vector can be expressed as

$$A=[(Re\ E)^2+(Im\ E)^2]^{1/2} \quad (3-15a)$$

$$\begin{aligned} \Phi &= \tan^{-1}\left(\frac{Im\ E}{Re\ E}\right) \\ &= \cos^{-1}\left(\frac{Re\ E}{A}\right) \end{aligned} \quad (3-15b)$$

where

$$Re\ E = A_1 \cos \Phi_1 + A_2 \cos \Phi_2 \quad (3-16a)$$

$$Im\ E = A_1 \sin \Phi_1 + A_2 \sin \Phi_2 \quad (3-16b)$$

according to Euler's theorem. So now, the amplitude A may be expressed, using equations (2-15a) and (2-16), as

$$\begin{aligned} A &= [A_1^2(\cos^2 \Phi_1 + \sin^2 \Phi_1) + A_2^2(\cos^2 \Phi_2 + \sin^2 \Phi_2) \\ &\quad + 2A_1 A_2(\cos \Phi_1 \cos \Phi_2 + \sin \Phi_1 \sin \Phi_2)]^{1/2} \end{aligned} \quad (3-17)$$

which can be simplified by using two common trigonometric identities to give

$$\begin{aligned} A &= [A_1^2 + A_2^2 + 2A_1 A_2 \cos(\Phi_2 - \Phi_1)]^{1/2} \\ &= [A_1^2 + A_2^2 + 2A_1 A_2 \cos \delta]^{1/2} \end{aligned} \quad (3-18)$$

The resulting phase can also be determined from equations (2-15b) and (2-16a)

$$\begin{aligned}\cos\Phi &= \left[\frac{\text{Re } E}{A} \right] \\ &= \left[\frac{(A_1 \cos\omega t + A_2 \cos(\omega t + \delta))}{A} \right]\end{aligned}\quad (3-19)$$

where the phases have been set as $\phi_1 = \omega t$ and $\phi_2 = \omega t + \delta$. This can be expanded to give

$$\begin{aligned}\cos\Phi &= \frac{A_1}{A} \cos\omega t + \frac{A_2}{A} \cos\omega t \cos\delta - \frac{A_2}{A} \sin\omega t \sin\delta \\ &= \left[\frac{A_1}{A} + \frac{A_2}{A} \cos\delta \right] \cos\omega t - \left[\frac{A_2}{A} \sin\delta \right] \sin\omega t\end{aligned}\quad (3-20)$$

The time independent parts of the above equation can now be defined as

$$\cos\Phi' = \frac{A_1}{A} + \frac{A_2}{A} \cos\delta \quad (3-21a)$$

and

$$\sin\Phi' = \frac{A_2}{A} \sin\delta \quad (3-21b)$$

since it can be shown that $\cos^2\Phi' + \sin^2\Phi' = 1$ by direct substitution into equation (2-21) and use of the expression for A above, the definition is justified. We can now simplify the expression for phase as

$$\begin{aligned}\cos\Phi &= \cos\omega t \cos\Phi' - \sin\omega t \sin\Phi' \\ &= \cos(\omega t + \Phi')\end{aligned}\quad (3-22)$$

Thus the phase of the resultant field is

$$\Phi = \omega x + \Phi' \quad (3-23)$$

and one can see that the frequency ω is the same as that of the two constituent fields.

3.2.2 Diffraction

The distinction between interference and diffraction is subtle, in fact there is no significant physical difference between the two. However, it is customary to refer to interference when considering the superposition of just a few waves, and diffraction when a large number of waves are involved. Diffraction may be defined as those interference phenomena which are due to the spatial restriction (due to an obstruction which alters either the amplitude or phase of a wave) of light beams. An approximation of the phenomenon of diffraction is described by Huygens' principle. According to Huygens, each point on a wavefront can be considered a secondary source of spherical waves whose wavelength and frequency are equal to those of the primary wave at each point in space. This idea is illustrated in Figure 3-2. As these secondary wavelets propagate outward from their respective points on the wavefront, a new wavefront is formed which is defined by the envelope of the secondary wavelets. This technique does not faithfully represent the physical reality however, and it has its shortcomings: it does not depend on wavelength, and as such would predict the same diffraction pattern for waves of all wavelengths. When using Huygens' principle, only the contributions from each point

Figure 3-2. Huygens' construction for the propagation of light.

source which lies on the surface of the envelope are used to construct the new wavefront, otherwise those parts of the spherical wavelets which propagate back toward the source would form a back-wave, and this is simply not observed.

Fresnel used the concept of interference to modify Huygens' principle. The resulting Huygens-Fresnel principle states that every unobstructed point on a wavefront acts as a secondary source of spherical waves having frequencies identical to that of the primary source; the superposition of all these wavelets gives the intensity distribution of the field at any point beyond the wavefront. Figure 3-3 illustrates a plane wave incident upon a small aperture. According to the Huygens-Fresnel principle, each point on the aperture acts as a coherent secondary source of light. The maximum optical path length difference $\Lambda_{\max}$ from any two points in the aperture will be

$$\Lambda_{\max} = |\overline{AP} - \overline{BP}| \quad (3-24)$$

where P is the observation point. This maximum path difference corresponds to a source point at each edge of the aperture, A and B respectively, and this distance is always less than or equal to the width of the aperture. In the case where the two distances are equal, the observation point P must be on the aperture. In order to have some degree of constructive interference at the observation point, the wavelength of light used must be much greater than $\Lambda_{\max}$ [3.5], assuming of course that all of the point sources in the aperture are coherent, and therefore mutually in phase. When the wavelength of light is much greater than the size of the aperture, as in Figure 3-4 the wavelength is therefore also much greater than $\Lambda_{\max}$, therefore giving some degree of constructive interference

Figure 3-3. Diffraction at a small aperture.

Figure 3-4. Diffraction at a large aperture

everywhere beyond the aperture. When the wavelength is much less than the size of the aperture, the only region where $\lambda \gg \Lambda_{\text{max}}$ is in a small band extending directly out in front of the aperture. Out past this band some of the wavelets will interfere destructively, signalling the beginning of the shadow zone. A truly sharp and abrupt shadow zone will therefore only occur for wavelengths of light where $\lambda \rightarrow 0$, or where λ is much smaller than the size of the aperture.

The Huygens-Fresnel principle was given more credence by Kirchhoff; he was able to show that it was a direct consequence of the wave equation. It should be noted that Kirchhoff treated diffraction as a scalar phenomenon in his mathematical model. Fortunately, the scalar theory is an approximation which works quite well given that: 1) the diffracting aperture is large compared to the wavelength, and 2) the diffracted waves are observed far away from the aperture [3.5]. The above two conditions are usually met in practical cases. The subsequent derivation of the Fresnel-Kirchhoff diffraction formula roughly follows that of Goodman [3.6].

The solution to the Helmholtz equation can be simplified by using Green's theorem, which states

$$\iiint_V (G \nabla^2 U - U \nabla^2 G) dV = \iint_S \left(G \frac{\partial U}{\partial n} - U \frac{\partial G}{\partial n} \right) dS \quad (3-25)$$

where $U(P)$ and $G(P)$ are complex functions of position, S is a closed surface containing a volume V , and n represents the outward normal direction on each point of the surface S . If the function G also satisfies the Helmholtz equation, then

$$\iiint_V (G \nabla^2 U - U \nabla^2 G) dV = \iiint_V (G k^2 U - U k^2 G) dV = 0 \quad (3-26)$$

and therefore, by substituting Equation (3-26) into Equation (3-25), Equation (3-25) may be written as

$$\iiint_V (G \nabla^2 U - U \nabla^2 G) dV = \iint_S (G \frac{\partial U}{\partial n} - U \frac{\partial G}{\partial n}) dS = 0 \quad (3-27)$$

Kirchhoff chose as a Green's function a unit amplitude spherical wave expanding about an observation point P_0 .

$$G(P_1) = \frac{\exp(ikr_{01})}{r_{01}} \quad (3-28)$$

where P_1 is an arbitrary point in space, and r_{01} is the distance from P_0 to P_1 . A Green's function and its first and second derivatives must, however, be continuous within the volume V , whereas there is a discontinuity at P_0 for the above spherical wave. To preclude this problem, the integration does not include a spherical volume of radius ϵ around P_0 , and the surface of integration becomes a composite surface $S' = S + S_\epsilon$, as illustrated in Figure 3-5. So now the integral of the right side of Equation (3-27) can be written as a sum of two integrals

$$\iint_S (G \frac{\partial U}{\partial n} - U \frac{\partial G}{\partial n}) dS + \iint_{S_\epsilon} (G \frac{\partial U}{\partial n} - U \frac{\partial G}{\partial n}) dS = 0 \quad (3-29)$$

and since G is known, we may write

Figure 3-5. Kirchhoff surface of integration.

$$\frac{\partial G(P_1)}{\partial n} = \cos(n, r_{01}) \left(ik - \frac{1}{r_{01}} \right) \frac{\exp(ikr_{01})}{r_{01}} \quad (3-30)$$

where $\cos(n, r_{01})$ represents the cosine of the angle between the surface normal n and the vector r_{01} joining P_0 and P_1 . For the case where P_1 is on S_ϵ

$$G(P_1) = \frac{\exp(ik\epsilon)}{\epsilon} \quad \text{and} \quad \frac{\partial G(P_1)}{\partial n} = \frac{\exp(ik\epsilon)}{\epsilon} \left(\frac{1}{\epsilon} - ik \right) \quad (3-31)$$

and if U and its derivatives are continuous, making ϵ arbitrarily small gives

$$\begin{aligned} \lim_{\epsilon \rightarrow 0} \iint_{S_\epsilon} \left(G \frac{\partial U}{\partial n} - U \frac{\partial G}{\partial n} \right) ds &= \lim_{\epsilon \rightarrow 0} 4\pi\epsilon^2 \left[\frac{\partial U(P_0)}{\partial n} \frac{e^{ik\epsilon}}{\epsilon} - U(P_0) \frac{e^{ik\epsilon}}{\epsilon} \left(\frac{1}{\epsilon} - ik \right) \right] \\ &= -4\pi U(P_0) \end{aligned} \quad (3-32)$$

substitution of the above result into Equation (3-29) gives

$$\begin{aligned} U(P_0) &= \frac{1}{4\pi} \iint_S \left(\frac{\partial U}{\partial n} \left[\frac{e^{ikr_n}}{r_{01}} \right] - U \frac{\partial}{\partial n} \left[\frac{e^{ikr_n}}{r_{01}} \right] \right) ds \\ &= \frac{1}{4\pi} \iint_S \left(\frac{\partial U}{\partial n} G - U \frac{\partial G}{\partial n} \right) ds \end{aligned} \quad (3-33)$$

which is known as the *integral theorem of Helmholtz and Kirchhoff*, and can be used to determine the field at any point P_0 in terms of the boundary values of the wave on any closed surface around that point.

In order to solve for the diffraction of a point source by an aperture A , equation (3-33) is used, with the surface of integration being $S=S_1+S_2+S_3$ as illustrated in Figure 3-6. To make the integration simpler, the integration over surface S_2 is disposed of on

Figure 3-6. Kirchhoff's formulation of diffraction by an aperture.

the following grounds. Assume that the wave does not exist for all time. Then, by making the radius of S_2 arbitrarily large, the disturbance U and its derivatives will vanish on this surface [3.6]. Now the integration can be performed over the plane S_1+S_3 . To further simplify the integration, Kirchhoff imposed the following two boundary conditions: (1) the field distribution U and its derivative $\partial U/\partial n$ on the aperture S_3 are as if the screen were absent, and (2) the field U and $\partial U/\partial n$ in the geometrical shadow vanish. These conditions are called Kirchhoff's boundary conditions, and allow the integration to be performed over just the imaginary surface of the aperture S_3 . Usually $r_{01} \gg \lambda$, this along with Kirchhoff's boundary conditions allows us to approximate Equation (3-31) as

$$\frac{\partial G(P_1)}{\partial n} \approx jk \cos(n, r_{01}) \frac{\exp(jkr_{01})}{r_{01}} \quad (3-34)$$

This allows equation (3-33) to be written as

$$U(P_2) = \frac{1}{4\pi} \iint_{S_3} \frac{\exp(jkr_{01})}{r_{01}} \left[\frac{\partial U}{\partial n} - jk U \cos(n, r_{01}) \right] ds \quad (3-35)$$

If the aperture were illuminated by a point source

$$U(P_1) = A \frac{\exp(jkr_{21})}{r_{21}} \quad (3-36)$$

where P_2 is the source at a distance r_{21} from a point P_1 on the aperture and where $r_{21} \gg \lambda$ then the point source can similarly be approximated as a term with a cosine function.

These substitutions give

$$U(P_2) = \frac{A}{j\lambda} \iint_{S_1} \frac{\exp[jk(r_{21} + r_{01})]}{r_{21}r_{01}} \left[\frac{\cos(n, r_{01}) - \cos(n, r_{21})}{2} \right] ds \quad (3-37)$$

which is called the *Fresnel-Kirchhoff diffraction formula*, and only applies to the diffraction of a point source.

Another way in which to formulate scalar diffraction theory which makes calculating the efficiencies of certain types of holograms easy, is to make use of the angular spectrum of plane waves. Should the Fourier transform of the complex field distribution across any plane be taken, it would be evident that the various Fourier components resemble plane waves travelling in different directions. These component plane waves are called the angular spectrum of plane waves. To determine the field amplitude at any other point, the contributions of all the plane waves are added, taking account of their phase shifts. A more detailed account of this approach to diffraction theory is found in Ratcliffe [3.7].

A wave travelling in the positive z direction and incident on the x - y plane of a Cartesian coordinate system produces a complex field $U(x,y,0)$ in that plane. Across the x - y plane, the two-dimensional Fourier transform of the complex field U is

$$A_o = \iint_{-\infty}^{\infty} U(x,y,0) \exp[-i2\pi(f_x x + f_y y)] dx dy \quad (3-38)$$

Such a transformation can be considered a breaking up of a more elaborate function into an array of simpler functions. In fact, the original function U can be written as a sum of functions by taking the inverse transform of its spectrum A_o .

$$U(x,y,0) = \iint_{-\infty}^{\infty} A_o(f_x, f_y) \exp[i2\pi(f_x x + f_y y)] df_x df_y \quad (3-39)$$

Recalling that a plane wave with direction cosines (α, β, γ) can be written

$$E(x,y,z) = \exp[i\frac{2\pi}{\lambda}(\alpha x + \beta y + \gamma z)] \quad (3-40)$$

where $\alpha + \beta + \gamma = 1$, the function $\exp[i2\pi(f_x x + f_y y)]$ can be regarded as a plane wave having direction cosines

$$\begin{aligned} \alpha &= \lambda f_x \\ \beta &= \lambda f_y \\ \gamma &= [1 - (\lambda f_x)^2 - (\lambda f_y)^2]^{1/2} \end{aligned} \quad (3-41)$$

where the amplitude of that plane wave is $A_o(f_x, f_y) df_x df_y$, evaluated at $(f_x = \alpha/\lambda, f_y = \beta/\lambda)$.

Equation (3-38) can now be written as

$$A_o(\frac{\alpha}{\lambda}, \frac{\beta}{\lambda}) = \iint_{-\infty}^{\infty} U(x,y,0) \exp[-i2\pi(\frac{\alpha}{\lambda}x + \frac{\beta}{\lambda}y)] dx dy \quad (3-42)$$

This is called the *angular spectrum* of the wave $U(x,y,0)$ [3.6].

However, if the angular spectrum is to be calculated via a computer algorithm, it must be formulated using the discrete Fourier transform. If the spatial period of the wave on the x-y plane are T and S respectively, the discrete angular spectrum is

To determine the fraction of light which falls into each diffraction order, now all that has to be done is to determine the power spectrum by squaring the absolute value of the

$$A\left(\frac{\alpha}{\lambda}, \frac{\beta}{\lambda}, z\right) = S^{-1} T^{-1} \int_0^S \int_0^T U(x, y, z) \exp\left[-i2\pi\left(\frac{\mu}{S}\sigma + \frac{\nu}{T}\tau\right)\right] d\sigma d\tau \quad (3-43)$$

angular spectrum [3.6].

If U instead represents the complex amplitude transmission function of some optical element placed in the x - y plane, A will represent the angular spectrum of the field immediately behind the element, assuming the element is illuminated by a plane wave propagating in the positive z direction. So the fraction of light falling into the ν^{th} diffracted order from a one-dimensional optical element having a complex amplitude transmission function $t(\tau) = \exp[i\theta(\tau)]$ is

$$a(\nu) = \left| T^{-1} \exp[-i\theta(\tau)] \exp\left[2\pi i \nu \frac{\tau}{T}\right] d\tau \right|^2 \quad (3-44)$$

By substituting $\tau = T x$ the expression simplifies to

$$a(\nu) = \left| \int_0^1 \exp[2\pi i \nu x - i\theta(T \cdot x)] dx \right|^2 \quad (3-45)$$

which can be evaluated for different $\theta(x)$.

Chapter 4

THEORY OF HOLOGRAPHY

4.1 Introduction

A hologram is a type of optical element, which means it performs some predetermined operation on the light which is incident upon it. The effects of optical elements may be described by either ray optics, as are lenses and mirrors, or by interference and diffraction, the phenomena described in the preceding chapter. Holograms are one type of optical element whose effects on light must be described by interference and diffraction.

The word *holography* means whole image. It is the art of wavefront reconstruction; a hologram records both amplitude and phase information of a wave on each point of a two dimensional surface. The essential principle - that of the reference beam - was discovered by Gabor [4.1] before the invention of the laser. Since the development of the laser however, holography has become a much less daunting task than it was at the time of Gabor's original experiments. Laser sources are readily available and their coherent light is ideal for making holograms.

While physical photodetectors, such as film and the eye, do not record phase information, only light intensity, a reference beam which has the same frequency as the object wave will produce an interference pattern with the object wave when both are incident on the recording medium. That is how phase information is recorded. The reference beam is also required to reproduce the object wave.

4.2 Conventional Holography

Holography is based on mixing the wave whose information you want to store (called the *object wave*) U_o with a recording/decoding wave (called the *reference wave*) U_r and recording the interference pattern produced. When recording the interference pattern on a film plate, a transparency of complex amplitude transmittance t is produced, as in Figure 4-1(a), where t is proportional to the intensity of the sum of the mixed waves. The transmittance is given by

$$\begin{aligned} t &\propto |U_o + U_r|^2 \\ &= |U_r|^2 + |U_o|^2 + U_r^* U_o + U_r U_o^* \\ &= I_r + I_o + U_r^* U_o + U_r U_o^* \end{aligned} \quad (4-1)$$

where I_r and I_o are the intensities of the reference wave and object wave, respectively. By looking at this hologram transmittance function, one can see that it contains information about the object wave U_o .

To decode the information stored in the hologram and reconstruct the object wave, the reference wave U_r is again used to illuminate the hologram, as in Figure 4-1(b). The wave which results has a complex amplitude on the hologram plane of [4.2]

$$U = tU_r \propto U_r I_r + U_r I_o + I_r U_o + U_r^2 U_o^* \quad (4-2)$$

The first two terms on the right-hand side of the above equation represent the reference wave, although modulated by the sum of the intensities of the two waves. The third term is the object wave multiplied by the intensity of the reference wave. The fourth term is

Figure 4-1. The holographic principle. (a) Recording a hologram on film and (b) reconstructing the object wave.

a conjugated version of the object wave multiplied by U_r^2 , and is called the *conjugate wave*. So three separate waves are produced upon reconstruction of the object wave: the reference wave, the object wave, and a conjugate wave. To determine what behavior the conjugate wave exhibits, we must examine more closely the hologram transmission function and the reconstructed object wave. Suppose the time independent reference wave can be described by

$$U_r = U_{or} e^{i\phi} \quad (4-3)$$

on the hologram plane, and the object wave can be described by

$$U_o = U_{oo} e^{i\phi_o} \quad (4-4)$$

where this is again the time independent field. The reconstructed object wave can now be expressed, after some algebraic manipulation, as [3.4]

$$U = U_{or}(U_{or}^2 + U_{oo}^2)e^{i\phi} + U_{or}^2 U_{oo} e^{i(2\phi - \phi_o)} + U_{or}^2 U_{oo} e^{i\phi_o} \quad (4-5)$$

where, in this case, the first term is the reference wave, the second term is the conjugate wave and the third term is the object wave. As shown by the above expression, the conjugate wave has an amplitude which is directly proportional to that of the object wave, but its phase term contains the negative of the phase of the object wave. The conjugate wave contains all of the necessary information about the object, but presents it in a sort of reverse order [3.4]. Moreover, the reconstructed wave's phase term also contains a 2ϕ term, which provides angular separation of the reconstructed object wave and the conjugate wave.

Perhaps conceptually the simplest hologram is that produced when a coherent, monochromatic plane wave is incident upon a point object, causing the wave at that point in space to diffract and produce a spherical wave. These two waves can then produce an interference pattern on a film plate, for example, consisting of a series of concentric rings resembling a Fresnel zone plate, as illustrated in Figure 4-2. Like a Fresnel zone plate, when the reference beam is incident upon the finished hologram, the light will be focused down to a point at the same distance from the hologram as the original point object was from the film plate. This type of hologram, known as the in-line or Gabor hologram, was the first type of hologram ever produced.

The conjugate wave produces a real image (where the rays actually cross at a point opposite the hologram from the reference wave source) of the object where points in the foreground appear to be in the background and vice versa. For this reason, the real image is of limited value. The reconstructed object wave on the other hand produces a virtual image (where the rays only appear to cross) where the object was when the hologram was recorded as shown in Figure 4-3.

4.3 Phase Holography

A phase-only expression of the interference pattern may also be written, as opposed to the amplitude-only recordings previously considered. Such an expression contains all of the information in the interference expression in a phase-only format. With these holograms, called *kinoforms* [4.3], the interference pattern is not recorded as amplitude transmittance on film, but as relief, or height, variations. The kinoform alters

Figure 4-2. (a) Arrangement for recording an in-line point object hologram. (b) Interference pattern recorded on film from (a).

(a)

(b)

Figure 4-3. The recording (a) and reconstruction (b) of a holographic image of a point object showing the object and conjugate waves.

the phase of the reference beam, by way of its phase variation, and in so doing reproduces the object field. While such patterns are difficult, if not impossible, to record on film they may be produced by etching a pattern into a substrate. When two fields U_o and U_r are superposed and the resultant field is again squared as in the previous case, without taking the absolute value this time, the resultant field may be broken down into real and imaginary components as

$$U_i(x,y) = A_o \cos \phi_o(x,y) + A_r \cos \phi_r(x,y) + i(A_o \sin \phi_o(x,y) + A_r \sin \phi_r(x,y)) \quad (4-6)$$

where equal amplitudes are assumed for the object and reference waves. The phase of this expression can then be expressed as

$$\exp[\phi_i(x,y)] = \exp\left[\tan^{-1}\left(\frac{\sin \phi_o(x,y) + \sin \phi_r(x,y)}{\cos \phi_o(x,y) + \cos \phi_r(x,y)}\right)\right] \quad (4-7)$$

and by applying some trigonometric identities, this may be reduced to

$$\exp[\phi_i(x,y)] = \exp\left[\frac{\phi_o(x,y) + \phi_r(x,y)}{2}\right] \quad (4-8)$$

The above phase term must be squared if it is to be proportional to the phase equivalent of intensity. Thus the complex amplitude transmission function which represents the interference pattern is

$$\exp[\phi_i(x,y)] = \exp[\phi_o(x,y) + \phi_r(x,y)] \quad (4-9)$$

The assumption of equal amplitude object and reference beams is justifiable because 100% efficiency is always strived for, or in other words the amplitude of the

reproduced object beam is desired to be as close to that of the reference beam as possible.

The reconstructed object wave is again obtained by illuminating the hologram with the reference wave, giving on the hologram plane a wave with complex amplitude

$$U(x,y) = U_{or}(x,y)e^{i\phi_o(x,y)}e^{-i(\phi_o(x,y)+\phi_o(x,y))} = U_{or}(x,y)e^{-i\phi_o(x,y)} \quad (4-10)$$

Note that the reconstructed wave consists only of the object wave - there is no conjugate or reference wave terms present. Because they direct light only into one order, kinoforms are able to reconstruct an object wave with efficiencies which approach 100%.

In order to produce a kinoform, the phase information of the hologram must be stored somehow on some medium. The phase information of the hologram may be created by introducing an object to the reference beam which will cause the correct phase delay at each point on the hologram plane. The easiest way to create a phase delay is to insert a piece of optically dense material of refractive index n into the path of the reference beam, so that the optical path length is increased. The phase delay $\Delta\phi$ effected on the wave which travels through the material, of thickness l , is

$$\Delta\phi = 2\pi l \frac{(n-n_o)}{\lambda} \quad (4-11)$$

where λ is the wavelength of the reference beam, and n_o is the refractive index of the surrounding medium.

Another way to introduce a spatially variable phase delay to the reference beam is through reflection. By subtracting a height h from the mirror, a phase delay

$$\Delta\phi = \frac{4\pi h}{\lambda} \quad (4-12)$$

can be impressed upon the reference beam. Two equal terms add to produce the phase delay in this case because the wave must travel the added distance twice: once on incidence and once after reflection. A reflection hologram can create the same image as the previous transmission hologram, but the image will be on the opposite side of the hologram surface.

4.4 Fourier Transform Holography

Fourier transform holography utilizes object and reference waves which have similar radii of curvature in order to store in the hologram the Fourier transform of the object wave. This method is designed so that the reference wave cancels out the quadratic dependence of the phase of the object wave on position within the hologram plane which appears in conventional holograms such as zone-lens types [3.4]. The simplest method of obtaining object and reference waves of similar curvature is to use a converging lens to collimate the object wave so that it consists of a sum of plane waves, while using a plane wave for the reference beam. In order to view this type of hologram another lens is required to focus the image. The lenses in this scheme act as optical Fourier transformers. The first lens takes the Fourier transform of the object wave, and the second takes the inverse Fourier transform.

To illustrate the Fourier transform properties of a lens, consider the case illustrated in Figure 4-4. The optical element $t(x,y)$ is in front of the lens at a distance d_0 and is illuminated by a normally incident plane wave of amplitude A . The field at plane y_r for a sufficiently large lens, is [3.6]:

$$U(x_r, y_r) = \frac{A}{i\lambda f} e^{\frac{i\pi}{\lambda f} (1 - \frac{d_0}{f}) (x_r^2 + y_r^2)} \iint_{-\infty}^{\infty} t(x_t, y_t) e^{-i\frac{2\pi}{\lambda f} (x_t x_r + y_t y_r)} dx_t dy_t \quad (4-13)$$

which is proportional to the two-dimensional, spatial Fourier transform $T(x_r, y_r)$ of $t(x, y)$, the constant of proportionality being the phase factor preceding the transform integral. The above relation holds assuming that the Fresnel approximation is valid over the distance d_0 between the optical element and lens. This approximation states [3.6]

$$(L_t + L_r)^4 \ll \frac{4\lambda d_0^3}{\pi} \quad (4-14)$$

where L_t and L_r are the maximum radial extent of t and T . Note that for the special case of $d_0 = f$ the initial phase factor vanishes, leaving an exact Fourier transform of the optical element on the focal plane.

To record a Fourier transform hologram, a film plate is placed in the Fourier transform plane of the above arrangement, recording the Fourier transform of the image. Thus a lens is also required to perform the inverse Fourier transform and generate the original wavefront.

Figure 4-4. One of several Fourier transforming configurations of a lens system.

4.5 Computer Generated Holograms

A computer generated hologram is a quantized amplitude or phase transmission hologram that is produced by calculating the interference pattern created by a fictitious object beam and a reference beam on some chosen surface and recording this function on some medium. The record may be in the form of a quantized amplitude mask or a multi-level phase relief structure, or as a combination of both. These holograms can be designed to operate as either reflection or transmission type holograms.

4.5.1 Direct Computer Generated Holograms

This type of hologram is the computer generated version of the conventional hologram - no lens is required. The interference pattern at the plane of the hologram is calculated and then quantized. When two spherical waves, having electric fields denoted by

$$E_o(x,y) = Ae^{i\phi_o}, \quad \text{where} \quad (4-15a)$$

$$\phi_o = k[(x-x_o)^2 + (y-y_o)^2 + (z-z_o)^2]^{1/2}$$

and

$$E_r(x,y) = Ae^{i\phi_r}, \quad \text{where} \quad (4-15b)$$

$$\phi_r = k[(x-x_r)^2 + (y-y_r)^2 + (z-z_r)^2]^{1/2}$$

interfere on the hologram plane, the field produced can be determined by substituting equations (4-15) into equation (4-6) and applying a few trigonometric identities. The resultant field is

$$E_t(x,y) = 2A \cos\left[\frac{1}{2}(\phi_o - \phi_r)\right] e^{i\frac{1}{2}(\phi_o + \phi_r)} \quad (4-16)$$

Now, the E field above must be squared to get a result which is proportional to intensity, then the amplitude and phase of the expression must be quantized. When making an amplitude hologram the phase of the expression is quantized to unity, and when a phase hologram is desired, the amplitude is quantized to unity. In either case the total E field E_t must be squared:

$$E_t^2(x,y) = 2A^2 (1 + \cos(k(\phi_o - \phi_r))) e^{ik(\phi_o + \phi_r)} \quad (4-17)$$

To quantize the above expression, the argument of either the amplitude or the phase term is set equal to $2\pi m/K$. In the case of phase quantization, the result is $k(\phi_o + \phi_r) = 2\pi m/K$ where m is any positive integer and K is the number of quantization levels. What this means is phase values between $2\pi(k-1)/K$ and $2\pi k/K$ are quantized to $2\pi k/K$, where k is the index of phase quantization and is an integer which ranges from 1 to K . In this way, a phase profile resembling that of a staircase is produced.

Such a quantization can produce an analytic solution which is used to produce the actual hologram. Other more complex arrangements involving wavefronts which are not simply spherical however, may not yield an analytic solution upon quantization. When this is the case, the most appropriate way to quantize the information is to divide the area of the hologram into a number of pixels, and then iteratively calculate and quantize the field at each pixel location. However, this method can generate an enormous amount of data when there are many pixels in the hologram [4.4].

The diffraction efficiency of quantized holograms is given by [4.5,4.6]

$$\eta = \left(\frac{\sin(\pi/K)}{\pi/K} \right)^2 \quad (4-18)$$

For a binary phase hologram, where $K=2$, the maximum efficiency will be 40.5%. In the case of a binary amplitude hologram, which consists of areas of near total absorption and areas of near zero absorption only which together make up areas called zones, $K=2$ and the electric field component from each zone of the hologram will be half that of the binary phase hologram. Since the electric field is half the value of that in the case of the binary phase hologram the transmitted power is $(1/2)^2$, so the maximum efficiency is only $(1/2)^2 40.5\% = 10.1\%$, which is one quarter of the maximum efficiency of a binary phase hologram. The efficiencies of other multi-level holograms are given in Figure 4-5, and it shows that efficiencies near 100% can be reached for high K . In other words, holograms with large numbers of quantization levels can have near perfect efficiencies.

4.5.2 Computer Generated Fourier Holograms

A computer generated Fourier hologram is similar to an ordinary Fourier hologram, except that the interference pattern between the object and reference waves is calculated and quantized rather than physically recorded on film. Calculation of the Fourier transform of the object's image usually involves using the discrete Fourier transform of the sampled image.

A computer generated Fourier transform hologram must record the value of the

Figure 4-5. Efficiencies of multilevel phase-only holograms

Fourier transform of the image at each point on the Fourier transform plane. One of the first methods of doing this was detour-phase holography [4.4], which represents the complex value of each Fourier transform sample point by its own cell. The amplitude information is represented by the size of a transparent rectangular aperture in the cell, whereas the phase information is recorded in the position of the aperture within the cell. As illustrated in Figure 4-6, the size of each cell is $\delta_x \times \delta_y$, and the aperture is centred vertically within the cell. The width of each aperture remains constant, while the height of the aperture varies and is denoted h_{nm} , n and m being, respectively, the horizontal and vertical indices of each cell. The horizontal distance from the center of the aperture to the center of the cell is denoted c_{nm} .

Building such a pattern of cells containing apertures creates an amplitude transmittance which can be expressed as

$$t(x,y) = \sum_n \sum_m p[(x - n\delta_x - c_{nm})/w] p[(y - m\delta_y)/h_{nm}] \quad (4-19)$$

where the function $p(x)$ is defined by

$$p(x) = \begin{cases} 1 & \text{for } |x| \leq \frac{1}{2} \\ 0 & \text{otherwise} \end{cases} \quad (4-20)$$

When this hologram is placed in the optical system of Figure 4-4, the Fourier transform of the hologram will be produced by the lens in the system and the desired image will be visible on the back focal plane of the lens. This may be demonstrated analytically by examining the Fourier transform of the transmission function

(a)

(b)

Figure 4-6. The detour phase hologram. (a) an individual cell and (b) an array of such cells.

$$\begin{aligned}
 T(u,v) &= \iint_{-\infty}^{\infty} t(x,y) e^{-i2\pi(ux+vy)} dx dy \\
 &= \sum_n \sum_m G_{nm}(u,v) e^{-i2\pi(nd_x u + md_y v)}
 \end{aligned}
 \tag{4-21}$$

where the function G_{nm} is given by

$$G_{nm}(u,v) = \frac{\sin \pi w u}{\pi u} \frac{\sin \pi h_{nm} v}{\pi v} e^{-i2\pi c_{nm} u}
 \tag{4-22}$$

As one can see, the amplitude of $G_{nm}(u,v)$ is dependent upon the height of the aperture h_{nm} , and so the amplitude can be adjusted by changing h_{nm} . Likewise, the phase of $G_{nm}(u,v)$ is determined by the horizontal location of the aperture c_{nm} , and so it can be controlled by setting c_{nm} .

Unfortunately, detour-phase holograms, while being simple to construct, are very inefficient at routing light. Since only a small portion of each cell transmits light, only a small amount of light can get through to form the image, and of this light a lot goes through the hologram unaffected, forming a bright spot, called the DC order, where one would occur in the absence of the hologram. Fourier holograms also produce a conjugate image on the opposite side of the DC component, further decreasing the amount of light routed to the image. In order to separate the desired and conjugate images, a carrier frequency is introduced, causing the images to be off-axis and separated in space.

4.5.3 Optimized Holograms

Instead of recording the actual field distribution or calculating a desired field distribution, it is possible to create an optimum hologram field distribution by an iterative trial and error method, according to a computer algorithm. In this way, the hologram evolves from an essentially random state into a highly structured one.

One very popular method of creating optimized holograms, which is capable of solving very complex optimization problems, is that of simulated annealing. A quick description is given here; the reader will find a more rigorous treatment in [4.7]. The method deals with the problem of minimizing the so-called cost function $C(x_i)$ of the many variables x_i . The system to be optimized is treated as a physical system, and its solution space is searched for those x_i which give the minimal value to the cost function; this state is called the *ground state*.

To reach the ground state, a gradual cooling of the physical system from a very high temperature T is simulated. However, the cooling process must be sufficiently slow so the system does not become caught in local minima of the cost function. The cooling process begins with the system in a random configuration x_i . The configuration is then altered slightly by a random amount Δx_i and the new cost function is calculated. If the new cost function has a lower value than the old one, the new configuration is accepted, otherwise the new configuration is accepted with probability $\exp[\Delta C/T]$, where $\Delta C = C(x_i) - C(x_i + \Delta x_i)$. If the perturbation is accepted, the new configuration is changed by some random amount, and the new cost function is tested, otherwise the old configuration is again altered by a random amount and its new cost function is tested. This process is

repeated enough times at the same temperature to guarantee that thermodynamic equilibrium is reached, then the temperature is slightly lowered and the whole process is repeated until the temperature reaches absolute zero. If the rate of temperature decrease has been slow enough, the final state reached will be the ground state of the physical system in question. Often, however, it is not necessary to continue lowering the temperature to absolute zero. Rather, in most systems the minimum cost function configuration has a high degree of degeneracy, that is to say many configurations exist which exhibit cost functions very close to the absolute minimum cost function, all of which are effectively equivalent. The probability that one of these pseudo-minimal cost function states will be reached before the absolute minimal state can be extremely high, and the iteration process may be stopped as soon as one of these states has been achieved.

Applying this method to creating holograms, the cost function can be some function involving hologram efficiency and reconstruction error so that the former is maximized while at the same time the latter is minimized. Holograms of this type are capable of giving very high efficiencies along with low reconstruction errors, however calculating them is very computationally intensive.

Chapter 5

ERRORS IN HOLOGRAMS

5.1 Introduction

The production of the holograms studied in this thesis involves standard VLSI techniques, and as such many of the same errors common to VLSI processing techniques also exist in the fabrication process of holograms. The lithographic production of a staircase profile on a substrate involves mask to pattern registration, pattern transfer, and etching steps. Each of these steps has associated with it a particular type of error. The main classes of fabrication errors resulting from these lithographic processes are mask alignment error, line error, and etch depth error. These are the types of errors which will be considered in the following sections, but first it is necessary to briefly describe the fabrication process of a multilevel kinoform.

The process of making a K-level hologram, where $K=2^N$, consists on N sequential lithographic masking steps. Each masking stage consists of producing a duplicate pattern in photoresist on the substrate with a binary amplitude mask, and then selectively etching the uncoated areas; this process is illustrated in Figure 5-1. The etch depth of the n th stage ($n=1,...,N$) is such that a phase delay

$$\theta_n = \frac{2\pi}{2^n} \quad (5-1)$$

is created in the uncoated areas.

Figure 5-1. Lithographic steps of kinoform fabrication. The substrate is successively masked and etched.

5.2 Description of Fabrication Errors

Fabrication errors associated with kinoforms may be grouped into three major categories: etch depth errors, line errors, and alignment errors. Etch errors are simply the result of allowing the etching process to go either too far or not far enough; consequently the etch depths are either too deep or too shallow. Etch errors may be further subdivided into two categories: systematic and random etch errors. Systematic etch errors refer to a systematic variation in the proper etch depth across the surface of the kinoform. These errors might be caused by variations in the widths of the etch zones, whereby narrower etch zones (often near the periphery) are etched at a slower rate than are wider etch zones. Systematic etch errors might also be caused by spatial variations in the etcher itself, for example uneven ion beam milling. Random etch errors, on the other hand, are just that - random; these errors refer to random fluctuations in the etch depths across the entire surface of the kinoform. Generally, random etch errors are of a much lower magnitude than those of systematic etch errors.

Line errors can take the form of either lateral fluctuations in the positions of the etch zones, due to errors in the mask used to produce the kinoform, or they can be errors in the widths of the etch zones caused by either overexposure or underexposure of the photoresist during the lithographic production of the hologram. The former class of line errors are known as mask errors, while the latter are called linewidth errors and are the more common.

The last class of fabrication error considered here is alignment error. Alignment error refers to those errors caused by mask misalignment, and occurs whenever one mask

is not aligned exactly with the pattern etched into the substrate. Errors caused by non-vertical walls are not considered here because the wall heights at the $.830\text{ }\mu\text{m}$ wavelength used in this work are so small compared to the width of the narrowest step available with the present technology that the effect would be unnoticeable. However, the presence of non-vertical walls may actually help to decrease the effects of lateral errors; this will be discussed later.

5.3 Mask Alignment Error in the Detour-Phase Kinoform

The detour-phase kinoform is a type of cellular, periodic, multiple-phase Fourier transform hologram, introduced by Turunen et al [1.1], which is characterized by a set of generalized detour-phase values. When the number of quantization levels is taken to extremes, the hologram reduces to two well-known hologram classes. The detour-phase kinoform becomes a binary phase-only hologram when the quantization drops to two phase levels, and as the number of quantization levels approaches infinity, the hologram becomes an off-axis kinoform. In this way the detour-phase kinoform fills the void which exists between two common classes of computer generated holograms. Since the detour-phase kinoform has a number of distinct phase levels, the production of such a hologram requires several masking steps, allowing for the possible occurrence of alignment errors. The design of a detour-phase kinoform begins with an optimized on-axis kinoform.

5.3.1 The Detour-Phase Kinoform

To design a detour-phase kinoform, one first requires an on-axis kinoform as a

foundation; non-linear optimization techniques such as simulated annealing are generally used to create the required kinoforms, but any on-axis kinoform will do. For the sake of brevity, only one-dimensional holograms are considered here, although the equations can be easily generalized to characterize two-dimensional holograms. The angular spectrum of an on-axis kinoform, with period equal to unity, has components defined as

$$A(m) = \int_0^1 \exp[i\theta(x) - i2\pi mx] dx \quad (5-2)$$

immediately after the hologram, assuming an incident plane wave, where m is the index of the diffraction order and $\exp[i\theta]$ is the complex amplitude transmission function of the kinoform. The power delivered to each of the diffraction orders is then given by

$$P(m) = |A(m)|^2 \quad (5-3)$$

In order to make the kinoform an on-axis one, the phase of the complex amplitude transmission function is expressed as

$$\theta_s(x) = \theta_l, \quad x \in [(l-1)/L, l/L] \quad (5-4)$$

where l is an integer having values from 1 to L and θ_l is everywhere between 0 and 2π . To obtain the angular spectrum for this kinoform, equation (5-4) is substituted into equation (5-2), with the result

$$A_s(m) = \frac{1}{L} \exp(i\pi \frac{m}{L}) \text{sinc}(\frac{m}{L}) \sum_{l=1}^L \exp[i(\theta_l - 2\pi \frac{ml}{L})] \quad (5-5)$$

where $\text{sinc}(x) = \sin(\pi x)/(\pi x)$.

To optimize the performance of the hologram, suitable non-linear optimization techniques can be used to obtain a desired irradiance distribution within a given set of diffraction orders. Simulated annealing has been used successfully with a merit function of the form

$$Mf(\theta_i) = \sum_{m=M_1}^{M_2} [P(m) - \Phi(m)]^2 \quad (5-6)$$

where Φ is the desired power spectrum in the set of diffraction orders $m=M_1, \dots, M_2$. To ensure that the performance of the hologram is fully optimized, the number of cells, L , per period should satisfy $L \geq M_2 - M_1 + 1 = M$, where M is the number of diffracted orders in the image area. In fact, the efficiency of the kinoform increases with the ratio L/M [1.1].

Since the desired irradiance distribution in optical computing applications is almost always binary (i.e. either on or off) we can call the diffracted orders with $\Phi(m)=1$ the on beams and those with $\Phi(m)=0$ the off beams. The diffraction efficiency of the on beams may now be written as

$$P_E = \sum_{m=M_1}^{M_2} \Phi(m) P(m) \quad (5-7)$$

and one way of expressing the reconstruction error into the orders within the image area is

$$\Delta R = \max_{m=M_1, \dots, M_2} \left| \Phi(m) \left[1 - \frac{NP(m)}{P_E} \right] + \gamma [1 - \Phi(m)] \frac{NP(m)}{P_E} \right| \quad (5-8)$$

where N is the total number of on beams. The constant γ allows different weighting of the off beams in relation to the on beams in reconstruction error calculations. Values of

0.1-0.2 have been successfully used for γ in the past; in this way, a greater weight is given to the on beams since in optical computing, nonzero off beam intensities are usually less destructive than nonuniform on beam intensities. As the reader has probably gathered, this is only one definition of reconstruction error and others do exist.

With the use of a suitable nonlinear optimization algorithm to generate the phase relief of the cells, the reconstruction error can be reduced to a very low value (<1%) and the efficiency (one-dimensional) of the kinoforms can range from 75% to as high as 97%, at least in theory. Unfortunately, it has been shown that the on-axis kinoform is very sensitive to errors in the phase relief values and as such, with existing technology it is extremely difficult, if not impossible, to fabricate these kinoforms with sufficient accuracy in the phase levels to realize the theoretical performance of the kinoform.

One way of avoiding the high accuracy requirement inherent in the phase relief values of the on-axis kinoform is to transform this vertical accuracy requirement into a horizontal accuracy requirement, which is what Turunen et al [1.1] did when they proposed their detour-phase kinoform. To turn an on-axis kinoform into a detour-phase kinoform, a carrier frequency or linear phase term must first be added to the phase function θ_0 , creating the sawtooth phase function θ_c as seen in Figure 5-2. This new phase function can be expressed as

$$\theta_c(x) = \theta_0 + 2\pi Q L x - 2\pi Q(l-1), \quad x \in \left[\frac{(l-1)}{L}, \frac{l}{L}\right] \quad (5-9)$$

where l is an integer which ranges from 1 to L , and for simplicity Q , the slope factor, is

Figure 5-2. (a) linear phase term added to on-axis kinoform (heavy line) gives off-axis kinoform (dashed line). (b) heavy line is off-axis kinoform after modulo 2π operation, this is the detour phase kinoform together with the generalized detour phase values D_i .

also an integer. Since Q is an integer, the slope of the ramp is quantized; it is this slope that causes a lateral shift of the spectrum by QL diffraction orders, and so now the kinoform is of the off-axis variety. By inserting the expression for $\theta_c(x)$ into Equation (5-2), the expression for the angular spectrum can be found:

$$A_c(m) = \frac{1}{L} \exp[i\pi(Q + \frac{m}{L})] \text{sinc}(\frac{m}{L} - Q) \sum_{l=1}^L \exp[i(\theta_l - 2\pi \frac{ml}{L})] \quad (5-10)$$

As electromagnetic radiation is inherently periodic, there is no need for the phase relief function $\theta_c(x)$ to have values greater than 2π (the period of the incident radiation), and so the phase relief $\theta_c(x)$ may be changed to $\theta_c(x)$ modulo 2π without losing generality. The distance between the beginning of the cell and the point where the phase relief equals 2π is called the generalized detour phase and is expressed as

$$D_l = \frac{(1 - \frac{\theta_l}{2\pi})}{QL} \quad (5-11)$$

Notice that the phase information in θ_l is contained within the detour phase value. Unfortunately, fabricating a blazed hologram of this type is even more difficult than it is to make the optimized on-axis kinoform [1.1].

One way to make the kinoform easier to produce is to quantize the phase function into steps defined as

$$\theta_q(x) = \frac{2\pi k}{K}, \quad \theta_c(x) \in [2\pi \frac{(k-1)}{K}, 2\pi \frac{k}{K}] \quad (5-12)$$

where k is an integer which ranges from 1 to K . This staircase approximation is illustrated in Figure 5-3.

Figure 5-3. Staircase approximation of a continuous function; the heavy line is a step function approximation to the continuous function of the dashed line.

Many methods exist for expressing the angular spectrum of the phase relief function $\theta_q(x)$, one way is to expand the complex amplitude transmission function into a Fourier series as

$$\exp[i\theta_q(x)] = \sum_{p=-\infty}^{\infty} c_p \exp[ip\theta_c(x)] \quad (5-13)$$

where the coefficients c_p are the same as those used in the expansions of other K-level relief functions (ref. 17-18)

$$c_p = \exp(i\pi \frac{p}{K}) \text{sinc}(\frac{p}{K}) \quad (5-14)$$

where p is an integer. When equation (5-13) above is inserted into equation (5-2), the angular spectrum is, after some manipulation, expressed as

$$\begin{aligned} A_q(m) = & \frac{1}{L} \exp[i\pi(Q + \frac{m}{L})] \sum_{p=-\infty}^{\infty} c_p \exp[i\pi(p-1)Q] \\ & * \text{sinc}[\frac{m}{L} - pQ] \sum_{l=1}^L \exp[i(p\theta_l - 2\pi \frac{ml}{L})] \end{aligned} \quad (5-15)$$

One can see that as the number of quantization levels K increases toward infinity, $U_q(x)$ becomes equivalent to $U_c(x)$, the angular spectrum components for the sawtooth phase function. From this equation, one can see that the detour-phase kinoform consists of an identical set of steps within different cells, the only difference between cells is the amount of lateral shift, relative to some point within each cell, that is applied to its set of steps. Each set of steps represents a distorted kinoform grating of period $1/QL$ which is superimposed on top of the true kinoform grating (period is unity).

With respect to the detour-phase kinoform, the phase relief information of the on-axis kinoform, which is contained within the phase-transfer function θ_l , is found in the

generalized detour phase. Therefore the detour-phase kinoform does not require high accuracy in the phase relief to preserve the phase information, whereas the on-axis kinoform does. The detour-phase kinoform requires accurate horizontal spacing for the detour phase values in order to preserve phase information. When using standard photolithographic techniques to produce the kinoform, this means having an accurate mask and mask aligner - two things which are readily available from the microelectronics industry. So by converting the relatively simple phase-transfer function $\theta_s(x)$ into the much more complicated function $\theta_q(x)$, one can immensely simplify the process of recording the desired phase information onto a medium. Producing the phase relief for $\theta_s(x)$ with the required accuracy is extremely difficult if not impossible, while fabricating the detour phase values can be done precisely using standard lithographic techniques.

5.3.2 Alignment Error in the Detour-Phase Kinoform

There exist three main types of errors associated with the photolithographic fabrication of a multilevel kinoform [5.1, 5.2]. The first class encompasses errors in the masks and, if applicable, the undesired slant, due to isotropic etching, of walls which were designed to be vertical, both of which result in the horizontal displacement of phase-level boundaries. The second class are those errors in the etch depths themselves; these are equivalent to errors in the phase delay function of the kinoform. Finally, the last class consists of errors in the alignment of the different masks used in the lithographic steps to fabricate the kinoform. The first two classes of errors, as they apply to the detour-phase kinoform, are covered extensively by Turunen et al. [1.1]. In this discussion, only

alignment errors in the detour-phase kinoform are considered.

As the reader will recall, the physical effect of mask alignment error on the phase relief of a kinoform was discussed in a previous section of this dissertation. However to analyze the effects of this type of error, a mathematical expression for the phase relief is needed. As Figure 5-4 illustrates, the phase relief profile of a kinoform suffering from mask alignment error can be thought of as a superposition of two or more profiles, where the superposition function is performed on the phase dimension. One phase function is that of the desired kinoform, the others correspond to the phase errors introduced by each mask which has been subsequently misaligned to the first one. The phase of the desired kinoform is a staircase function and may be expressed as

$$\theta_c(k) = 2\pi \frac{k}{K} \quad (5-16)$$

where K is the number of levels in the kinoform and k is an integer which ranges from 1 to K . For now, the phase relief of the error terms will simply be grouped together as $f(k)$.

As was the case with the error-free hologram, the function $\theta_q(x) + f(x)$ is a periodic function of $\theta_c(x)$ and so the complex-amplitude transmission function can be expressed in the form of a generalized Fourier series

$$\exp[i(\theta_q(x) + f(x))] = \sum_{p=-\infty}^{\infty} c_p \exp[ip\theta_c(x)] \quad (5-17)$$

where the Fourier coefficients can be determined by evaluating the following expression

Figure 5-4. The kinoform profile with alignment error in (a) can be thought of as a superposition of the two functions in (b).

$$c_p = \int_0^1 \exp[i(\theta_q(x) + f(x))] \exp[-i2\pi px] dx \quad (5-18)$$

where p is again an integer.

In the case of a four-level detour-phase kinoform, the expansion coefficients may be found by breaking up the integral in Equation (5-18) above to form

$$c_p = \sum_{k=1}^K \left[\int_{\frac{k-1}{K}}^{\frac{k-1}{K} + \Delta} \exp[i(\theta_q(x) + f(x))] \exp[-i2\pi px] dx \right. \\ \left. + \int_{\frac{k-1}{K} + \Delta}^{\frac{k}{K}} \exp[i\theta_q(x)] \exp[i2\pi px] dx \right] \quad (5-19)$$

Here, the expression for the Fourier coefficients has been broken up into a sum of several integrals. Within each step (where the value of k is constant) the integral has been split into the sum of two integrals; one integration being performed over those values of x where $f(x)$ is nonzero, and the other being performed over those values of x where $f(x)$ is zero. The expansion coefficients are then expressed as a summation over all the steps (all values of k from 1 to K). The width of each area where $f(x)$ is nonzero is Δ , and the value of $f(x)$ within these regions is a constant

$$f(x) = \frac{2\pi}{K} (-1)^{k+1} \quad \frac{k-1}{K} \leq x \leq \frac{k-1}{K} + \Delta \quad (5-20)$$

and as before, the desired phase transmission function, $\theta_q(x)$ is

$$\theta_q(x) = \frac{2\pi k}{K} \quad k=1,2,3,\dots,K \quad (5-21)$$

With a little algebraic manipulation, this expression for the expansion coefficients can be

simplified into

$$\begin{aligned}
 c_p = & \Delta \text{sinc}(p\Delta) \exp[i\pi p\Delta] \sum_{k=1}^K \exp[i2\pi(k \\
 & + (-1)^{k+1} - K\Delta p + p - pk)/K] + (\frac{1}{K} - \Delta) \text{sinc}(p(1/K - \Delta)) \\
 & * \exp[i\pi p(\frac{1}{K} - \Delta)] \sum_{k=1}^K \exp[i2\pi k(1-p)/K]
 \end{aligned} \quad (5-22)$$

By substituting the expression for the Fourier expansion coefficients into Equation (5-2) the components of the angular spectrum of this hologram can be written as

$$\begin{aligned}
 A(m) = & \sum_{l=1}^L \sum_{p=-\infty}^{\infty} \frac{c_p}{L} \exp[ip\theta_l - i2\pi \frac{ml}{L}] \\
 & * \exp[i\pi pQ + i\pi \frac{m}{L}] \text{sinc}(\frac{m}{L} - pQ)
 \end{aligned} \quad (5-23)$$

The effects of mask misalignment on the efficiency and reconstruction error of the detour-phase kinoform can now be evaluated by substituting Equation (5-23) into Equations (5-7) and (5-8) respectively.

Typical results of computer simulation of the effects of mask alignment error on a four level detour-phase kinoform are depicted in Figure 5-5. Note that the efficiency of the kinoform decreases gradually with the growth of the lateral offset of the second mask (the mask alignment error) until the size of the alignment error equals the size of the spacing between adjacent steps and then begins to grow again. As Figure 5-5 illustrates, as the size of the mask alignment error approaches the size of the spacing between adjacent steps, the staircase gradually assumes an overall slope which is still up but in the opposite lateral direction, and so forms a mirror image of the original staircase. As the alignment error continues to grow past this point, the staircase begins to resume

fraction of kinoform period
(a)

fraction of kinoform period
(b)

fraction of kinoform period
(c)

Figure 5-5. Theoretical plots of the effects of alignment error on diffraction efficiency and reconstruction error for three different quaternary-phase detour-phase kinoforms. Solid lines represent efficiency, and dashed lines show reconstruction error. (a) $L=25$, $Q=1$; (b) $L=25$, $Q=2$; (c) $L=30$, $Q=1$.

its original shape and slope, and as such the efficiency of the hologram likewise increases.

The reader will recall that the detour-phase kinoform is a reconfiguration of the phase information contained within an optimized on-axis kinoform, the key word being optimized. A non-linear optimization algorithm is used to search the solution space for the phase relief structure that gives the lowest value for the appropriate merit function. However, there exists more than one phase relief structure which will give a near optimal value for the merit function (the optimum configuration is highly degenerate) and as the number of different solutions is infinite, the algorithm may only locate a few of the near optimal solutions in the time allotted. Of these solutions it will choose the best. Therefore, many phase relief structures exist which are capable of performing the same function with nearly identical performance, and the effect of alignment error will have different effects on each of the different holograms. Now recall the definition chosen for the reconstruction error of a detour-phase kinoform; it requires the maximum value of a function which is applied to all spectral orders within the image area, and so there is no reason for this value to be the same for different phase relief structures. Therefore, while it is possible to predict how the reconstruction error of each kinoform will change with the magnitude of the alignment error, each different kinoform, while performing identical functions, will react differently in terms of reconstruction error. The only generalization that can be made regarding reconstruction error is that the reconstruction error will be low at low values of alignment error and will approach an infinite value as the magnitude of the alignment error approaches the value of the distance separating adjacent phase relief steps on the kinoform. With this information, it should be possible to create an iterative

optimization algorithm which will design a detour-phase hologram which has optimum tolerance to alignment error.

Although each of the different kinoforms will have a slightly different reaction to errors in mask alignment, one can generalize that when the minimum feature size of the kinoform (the distance separating steps k and $k+1$) is held constant, increasing the value of Q or L will greatly enhance the kinoform's tolerance to alignment error. For example, a detour phase kinoform which has $Q=1$, $L=25$, and $K=4$ can only tolerate a mask misalignment of 12% of its minimum feature size and still have less than 2% reconstruction error, while the equivalent kinoform with $Q=4$, $L=25$, and $K=4$ can withstand a misalignment of the second mask of nearly half the kinoform's minimum feature size and still have a reconstruction error of less than 2%. Of course, when the alignment error is this severe the efficiency of the kinoform will degrade substantially; the efficiency must necessarily be zero when the mask alignment is off by the full amount of the step size. Figures 5-5 illustrate the relationship between alignment error and efficiency, and between alignment error and reconstruction error. Note that the curves for efficiency all exhibit the same trend.

5.4 Effects of Errors in Phase-only Holograms

By using the angular power spectrum in a way similar to that used in calculating the efficiency of the detour-phase kinoform, the efficiency of any periodic, quantized phase hologram can be calculated. To calculate the effects of fabrication errors on kinoform efficiency, simple error models based on the aforementioned error descriptions

were built. Next, based on these models, the different hologram profiles were simulated, and their corresponding efficiencies were calculated using the discrete Fourier power spectrum. This is a scalar calculation and as such gives no polarization effects. It is also independent of grating period and wavelength, but requires computational time which is orders of magnitude less than that required for the full vector approach.

To simulate the fabrication errors, an error term was added to the desired phase relief profile. The Fourier power spectrum along with error term is given by

$$|A(v)|^2 = |T^{-1} \int_0^T e^{i\phi(v) + \text{error}} e^{-i2\pi \frac{v}{T} \tau} d\tau|^2 \quad (5-24)$$

This is the one-dimensional power spectrum and is used justifiably because all of the holograms dealt with here are essentially superpositions of one-dimensional holograms. For the linewidth and alignment errors, all errors are assumed to propagate with equal magnitude in a direction perpendicular to the zone boundaries. While this is, in reality, the case for exposure errors, it is not necessarily the case for all alignment errors. The assumption represents the worse case for alignment error and was made to simplify the efficiency calculations.

All errors were modeled by adding an error term to the desired phase relief profile. The etch errors were modeled by simply adding or subtracting an amount of phase relief to that portion of the kinoform which was etched, the unetched portion remaining unchanged. To model the linewidth errors, the width of the etched portion was varied, simulating exposure error - the most common type of line error. And finally, alignment error was modeled by adding a periodic error function to the phase relief profile. All

error models are illustrated in Figure 5-6. Notice the narrow peaks created by the lateral errors in the quaternary phase kinoform. The non-vertical walls created by the isotropic etching process will actually help to reduce the size of these unwanted peaks, reducing the effects of lateral errors.

Three simulations each were completed for the etch and exposure errors: one each for binary level kinoforms, and two each, one for errors in each of the two masking steps, for quaternary level kinoforms. Only one simulation was performed for mask alignment error as it originates only from errors in masking steps of second level and higher, and this discussion deals only with kinoforms having a maximum of four phase levels.

The results of the simulations are illustrated in Figure 5-7. Note that the graphs depicting the relationship between etch depth error and efficiency all show that, at peak efficiency, the kinoforms are rather tolerant of errors in etch depth. Etching can easily be done to within 10% of the desired depth and these graphs show very little degradation in efficiency over that range, so errors in etch depth should be of little concern.

Unfortunately, the kinoforms seem to be not as tolerant of lateral errors, with one exception: exposure errors in binary level kinoforms. The simulations predict that they will be quite tolerant of exposure errors. Quaternary level kinoforms however, seem to be quite intolerant of exposure error as Figure 5-7 illustrates. Likewise, quaternary level kinoforms seem to be just as intolerant of alignment error. As such, care must be taken during the fabrication of these kinoforms to ensure that these types of errors are kept to a minimum.

2 of/de 2

(a)

(b)

(c)

Figure 5-6. Fabrication errors; shaded region represents ideal profile, heavy line is profile with errors. (a) systematic etch error (b) linewidth error (c) alignment error.

Figure 5-7(a). Effects of fabrication errors on diffraction efficiency for the binary-phase hologram.

Figure 5-7(b). Effects of etch errors on diffraction efficiency for the quaternary-phase hologram

Figure 5-7(c). Effects of exposure error on diffraction efficiency for the quaternary phase hologram.

Chapter 6

DIFFRACTION-FREE BEAMS

6.1 Introduction

Recently, a new type of coherent light beam, discovered by Durnin and called diffraction-free beams or Bessel beams, has been studied and shown to be quite interesting [6.1-6.9]. These beams represent a set of monochromatic solutions to the free-space Helmholtz equation and are not subject to the effects of diffraction. Mathematically, they are described by the zeroth-order Bessel function of the first kind, J_0 . This fundamental Bessel beam is characterized by a transverse intensity distribution which is constant along the direction of propagation and is proportional to $J_0^2(\alpha, \rho)$, where $k^2 = (\omega/c)^2 = \alpha^2 + \beta^2$ is the wave number and $\rho^2 = x^2 + y^2$ is the radial coordinate. More specifically, the intensity distribution, $I(x, y, z)$ may be stated as

$$I(x, y, z > 0) = I(x, y, 0) \quad (6-1)$$

where z is the direction of propagation, in other words, the beams are not subject to diffractive spreading. Such a beam would be very useful in performing precision alignment operations. The beam could even be used to transport energy to a target albeit not very efficiently, for only a small fraction of the beam's energy can strike a single point in space, the rest of the energy being focused on other points along the line of the beam.

While a true Bessel beam has an infinite extent and contains an infinite energy, and is therefore not experimentally realizable, it has been shown both numerically and

experimentally that finite-aperture approximations of these Bessel beams can be produced which exhibit negligible diffraction over distances which are orders of magnitude greater than the Rayleigh range of Gaussian beams of the same initial FWHM [6.1-6.9]. The Rayleigh range is that distance, $\pi r^2/\lambda$, where r is the radius of the aperture, beyond which diffractive spreading becomes increasingly noticeable. For example, a Gaussian beam emanating from an aperture of radius r will diverge at an angle proportional to λ/r at distances much greater than $\pi r^2/\lambda$ from the aperture.

6.2 Bessel Beams

Diffraction is a well understood physical phenomenon. It affects all classical waves and, as Heisenberg explained, it plays a role in quantum mechanical uncertainty relations as seen in the diffraction of particles such as electrons by a narrow opening. Diffraction is predicted by the Helmholtz equation

$$\nabla^2 E(r) + k^2 E(r) = 0 \quad (6-2)$$

$$k^2 = \frac{\omega^2}{c^2} \quad (6-3)$$

where

and $E(r)$ is the amplitude of the electric field. A solution of the Helmholtz equation for $E(r)$ will give a monochromatic solution of the paraxial wave equation

$$\nabla^2 E(r,t) - \frac{1}{c^2} \frac{\partial^2}{\partial t^2} E(r,t) = 0 \quad (6-4)$$

One class of solutions of the Helmholtz equation, discovered by Durnin [6.1], is a class of diffraction-free mode solutions. While this in itself is not novel (plane waves are diffraction free mode solutions of the Helmholtz equation), this property married with the new solution's characteristic well-defined beam and small beam radii is remarkable. These Bessel beams can have a central spot radius whose size is on the order of one wavelength of its constituent light [6.1].

The derivation of Durnin's Bessel beam solution begins with the Helmholtz equation; a solution to this equation must be found which satisfies equation (6-1). To simplify things, the form of the solution will be factored into terms which are dependent on z (the direction of propagation), time t , and the spatial variables which are perpendicular to z , respectively; thus a solution of the following form is desired:

$$E(r,t)=U(x,y)Z(z)\exp(-i\omega t) \quad (6-5)$$

where, because of equation (6-1), $|Z|$ must be independent of z . When equation (6-5) is substituted into the Helmholtz equation, the result can be simplified into the following two differential equations:

$$\frac{d^2 Z}{dz^2} + \beta^2 Z = 0, \quad (6-6)$$

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \alpha^2 \right) U(x,y) = 0, \quad (6-7)$$

where $\alpha^2 + \beta^2 = (\omega/c)^2$. In addition, the field $E(r,t)$ must agree with the Sommerfeld radiation condition (that the field corresponds to an outgoing wave as $z \rightarrow \infty$, or that the field vanishes at least as fast as a diverging spherical wave). This gives

$$Z(z) = C \exp(i\beta z), \quad (6-8)$$

where C is a constant and β is real and positive.

The Whittaker-type [6.10] solution to equation (6-7) is

$$U(x,y) = \int_0^{2\pi} F(\phi) \exp[i\alpha(x\cos\phi + y\sin\phi)] d\phi \quad (6-9)$$

where $F(\phi)$ is an arbitrary complex function. Now, in order for $U(x,y)$ to remain finite, α must be real and β must satisfy $0 < \beta \leq \omega/c$. Combining equations (6-8), (6-9), and (6-5) and setting $CF(\phi) = A(\phi)$ gives

$$E(x,y,z,t) = \exp(i(\beta z - \omega t)) \int_0^{2\pi} A(\phi) \exp(i\alpha(x\cos\phi + y\sin\phi)) d\phi \quad (6-10)$$

When β is real, the solution above is a non-diffracting field.

In order for the beam to exhibit axial symmetry, $A(\phi)$ must be constant for all ϕ ; when this is the case, the amplitude of the field is proportional to

$$E(x,y,z,t) = \exp(i(\beta z - \omega t)) \int_0^{2\pi} \exp(i\alpha(x\cos\phi + y\sin\phi)) \frac{d\phi}{2\pi} \quad (6-11)$$

which may be written as

$$E(r,t) = \exp(i(\beta z - \omega t)) J_0(\alpha \rho), \quad (6-12)$$

where $\rho^2 = x^2 + y^2$ and J_0 is the zeroth-order Bessel function of the first kind. When $\alpha = 0$ the result is just the plane wave, however when $0 < \alpha < \omega/c$ the solution represents a diffraction-free beam, since it has the same intensity distribution $J_0^2(\alpha, \rho)$ in every plane

normal to the z -axis and an intensity profile which is proportional to $1/\alpha\rho$. When α becomes greater than ω/c , the field becomes evanescent [6.1]. The width of the half maximum of the central spot is approximately $1/\alpha$ and therefore is at its smallest when α is at its largest, that is when α is ω/c or $2\pi/\lambda$; at this value of α the central spot exhibits a diameter of approximately $3\lambda/4$ [6.1].

Since the intensity of the beam envelope varies inversely with ρ , the central beam is sharply peaked, but its intensity profile in any transverse plane never dies off, and so the true J_0 beam contains an infinite energy. In fact, the energy contained in each ring between two consecutive zeros of the zeroth order Bessel function approximately equals the energy contained within the central lobe [6.1] of a true Bessel Beam. Therefore, it is impossible to create a true Bessel beam over an entire plane, as this would require infinite energy. It is, however, quite possible to create such a beam which transverses a finite area.

6.2.1 Bessel-Gauss Beams

A slight variation on the Bessel beam is the Bessel-Gauss beam [6.11], in which the optical field has the form of a Bessel function multiplied by a Gaussian profile at the plane at which the beam is created. Unlike the Bessel beam, the Bessel-Gauss beam delivers a finite power, and a better approximation than that of the Bessel beam can be realized experimentally thanks to the fast lateral decay of the gaussian profile.

The field of the Bessel-Gauss beam at its plane of origin (say the plane $z=0$) can be stated mathematically as

$$E(r,0) = AJ_0(\beta r)\exp[-(r/w_0)^2] \quad (6-13)$$

where A is an amplitude constant and w_0 is the waist of the Gaussian term at the plane $z=0$. The following derivation of the field of the Bessel-Gauss beam follows that of Gori, Guattari, and Radovani [6.11]. The field at any plane $z>0$ can be evaluated by using the paraxial approximation and the following Fresnel diffraction integral

$$E(r,z) = (-ik/z)\exp[i(kz + kr^2/(2z))] \\ \times \int_0^\infty E(\rho,0)\exp[ik\rho^2/(2z)]J_0(k\rho r/z)\rho d\rho \quad (6-14)$$

where $k=2\pi/\lambda$ and λ is the wavelength of the light being used. The integral in the above expression can be solved with the help of the following integral formula [6.12]

$$\int_0^\infty \exp(-\alpha x)J_0(2\gamma\sqrt{x})J_0(2\delta\sqrt{x})dx \\ = (1/\alpha)I_0(2\gamma\delta/\alpha)\exp[-(\gamma^2+\delta^2)/\alpha] \quad (6-15)$$

where α , γ , and δ are parameters which are able to assume the role of real or complex values. I_0 is the modified Bessel function of the first kind and zero order, and is related to J_0 in the following manner [6.12]

$$I_0(x) = J_0(ix) \quad (6-16)$$

By substituting equation (6-13) into equation (6-14), and then using equations (6-15) and (6-16), the result is

$$E(r,z) = -(ikA/2zQ)\exp[ik(z+r^2/2z)] \quad (6-17)$$

$$\times J_0(i\beta kr/2zQ)\exp[-(1/4Q)(\beta^2+(kr/z)^2)]$$

where

$$Q = \frac{1}{w_0^2} - \frac{ik}{2z} \quad (6-18)$$

From here, simple algebra and the fact that J_0 is an even function can be used to rewrite equation (6-17) in the form of

$$E(r,z) = (Aw_0/w(z))\exp[i[(k-\beta^2/2k)z-\Phi(z)]] \quad (6-19)$$

$$\times J_0[\beta r/(1+iz/L)]\exp[(-1/w^2(z) + ik/2R(z))(r^2+(\beta z/k)^2)]$$

where $L=kw_0^2/2$ and

$$(6-20)$$

$$w(z) = w_0\sqrt{1+(z/L)^2} \quad (6-21)$$

$$\Phi(z) = \arctan(z/L) \quad (6-22)$$

$$R(z) = z+L^2/z$$

These last equations are identical to those used to describe ordinary Gaussian beams, where L is the Rayleigh range, w gives the spot size, Φ is the phase-shift, and R is the radius of curvature of the Gaussian wavefront.

6.3 Original Method of Creating the J_0 Beam

One can see from equation (6-10) and Figure 6-1 that the Bessel beam consists of a sum of plane waves, each having the same amplitude and each making the same angle with the z-axis while the azimuthal angles range from 0 to 2π . The angle made with the z-axis by each plane wave is $\theta = \arcsin(\alpha\lambda/2\pi)$ [6.1]. Therefore, in order to produce a non-diffracting beam in practice, a superposition of plane waves having the aforementioned traits is required. Furthermore, they must be limited in space by a finite aperture. To create such a superposition of plane waves, and hence a finite-aperture approximation of a zeroth-order Bessel beam, Durnin et al. [6.2] conceived the optical arrangement depicted in Figure 6-2. The arrangement consists of a coherent plane wave, normally incident upon a thin annular slit which is located in the focal plane of a positive lens. The light emanating from each point around the slit ideally acts as a point source which the lens transforms into a plane wave. In this way, the set of plane waves generated has wave vectors all lying on the surface of an imaginary cone; this may be regarded as the defining characteristic of the Bessel beam. Geometrical optics tells us that, in this arrangement, $\alpha = k \sin\theta = \pi d/\lambda F$, where $k = 2\pi/\lambda$, d is the diameter of the circular slit, λ is the wavelength of the incident plane wave, and F is the focal length of the lens. One can also see that a shadow zone will preclude the existence of the Bessel beam past a distance $z_{\max} = R \tan\theta$ from the lens aperture, where R is the radius of the lens. Since the radius of the central beam is approximately $r = 1/\alpha$, one can write $1/r = k \sin\theta$, or $\sin\theta = 1/rk$, and since $\sin\theta = \tan\theta$ for small θ , one can approximate the maximum propagation distance as $z_{\max} = Rrk = 2\pi Rr/\lambda$. Comparing this distance to the Rayleigh range

Figure 6-1. A Bessel beam is the result of the superposition of plane waves whose wave vectors lie on the surface of a cone.

Figure 6-2. Durnin's original optical arrangement for generating a Bessel beam.

of a typical Gaussian beam, $z_r = \pi r^2 / \lambda$, one can see that the propagation distance of the Bessel beam will be much greater than the Rayleigh range of a conventional beam of radius $1/\alpha$ for $d \gg 1/\alpha$.

The main problem with this method of Bessel beam generation lies in the severe loss of light in the system, and consequent low efficiency. Of the total energy incident upon the apparatus in the form of a plane electromagnetic wave, the amount which is actually used in the creation of the Bessel beam is that amount which passes through the very thin annular slit which shapes the intensity profile of the incident wave. Needless to say, this amount is orders of magnitude smaller than that present in the incident plane wave. Another problem associated with Durnin's original optical arrangement for creating Bessel beams stems from limiting the transverse extent of the beam by an abrupt aperture. Because of diffraction effects at the aperture, the peak intensity of the beam does not remain constant. As the Bessel beam propagates away from its source, the intensity of the central peak oscillates about its initial value with ever increasing amplitude and decreasing frequency until it reaches a point, not quite at its maximum range, where it sharply declines [6.1].

6.4 Holographic Generation of Bessel Beams

In an effort to alleviate the problem of low Bessel beam construction efficiency, Turunen et al. [6.3] proposed to generate such a beam by illuminating a holographic optical element with a plane wave source; the arrangement they used is illustrated in Figure 6-3. The hologram has the same effect on phase as an axicon (a solid glass cone)

Figure 6-3. The original optical arrangement used by Turunen to holographically create a Bessel beam, showing the beam's propagation distance $Z_{\max}$ along the propagation axis z .

and its complex-amplitude transmission function is described as

$$t(\rho) = \exp\left(i\Phi \frac{(\rho)}{\rho_0}\right) \quad (6-23)$$

where ρ_0 is the spacing between adjacent rings of the hologram. In fact the phase relief of the hologram is precisely the complex-amplitude transmission function of a thin linear axicon. And so, just like an axicon, the hologram converts an incident plane wave into the required superposition of plane waves with a common angle to the propagation axis.

One problem with this type of holographic Bessel beam generation is that the beam intensity is proportional to the propagation distance z [6.7]. Physically this can be easily understood by realizing that in any plane perpendicular to the propagation axis the received light comes from a thin annulus in the plane of the hologram, the distance from the hologram plane being directly proportional to the radius of the annulus. The area of this annulus is proportional to its own radius, and the amount of light received is in turn proportional its area. Thus as the radius of the annulus grows, the amount of light directed to the Bessel beam also grows, hence the linear dependence of intensity on propagation distance.

To alleviate the problem of the linear growth of axial light intensity, the disparities in light energy contributed by the different annuli must be eliminated. This was first accomplished experimentally by Davidson et al. [6.7] Their arrangement consists of two holograms; the first transforms a uniform intensity distribution plane wave into a $1/r$ distribution on the plane of the second hologram. It is the second hologram which corrects for the phase distortion of the first hologram and adds a conical phase of the

form $\exp(-i2\pi r/r_0)$ to produce the non-diffracting beam. However, as this system requires a pair of holograms, it is necessarily larger and more difficult to align than a single hologram system. Moreover, both of the previous holograms require a collimated beam, which adds another degree of complexity to the system.

The arrangement presented in this dissertation involves the use of a single phase-only hologram to produce a Bessel-Gauss beam from the elliptical Gaussian beam of an ordinary laser diode [6.13]; the Sharp LT015MD laser was used. It is the first hologram which uses the light from an uncollimated laser diode to produce such a beam. And the laser diode source produces a beam which can approximate the $1/r$ intensity profile required to keep the Bessel beam's intensity approximately constant over a wide range of values of distance along the propagation axis.

The laser diode source, when used correctly, should also be capable of reducing the magnitude of the periodic fluctuations in peak intensity of the Bessel beam by reducing diffraction effects at the aperture. The Gaussian profile of the laser diode beam, which gradually decays to zero intensity for sufficiently large r , means that edge effects will be greatly reduced for sufficiently large apertures [6.14].

The complex-amplitude transmission function of the hologram was obtained by mathematically superimposing the phase information from the fields of the point source of the laser diode together with the necessary set of plane waves having wave vectors lying on the surface of a cone required to produce a Bessel beam. The phase transmission function of the hologram in a typical on-axis arrangement is

$$t(r) = \exp(-i2\frac{\pi}{\lambda}(\beta r + \sqrt{f_o^2 + r^2})) \quad (6-24)$$

where f_o is the distance separating the point source and the hologram plane, r is the radial coordinate and β is a constant that determines the angle at which the wave vectors from the set of plane waves intersects the z -axis (propagation direction). An off-axis hologram can be made by adding a linear phase term to the phase in equation (6-24). The resulting phase transmission function would be:

$$t(r) = \exp(-i2\frac{\pi}{\lambda}(\beta r + \sqrt{f_o^2 + r^2} + vy)) \quad (6-25)$$

6.5 Measurement and Characteristics of Bessel Beam

The hologram used to produce the Bessel beam is a reflection type HOE, and it was incorporated into the optical experimental arrangement shown in Figure 6-4. The design distance from the laser diode to the hologram was 50 mm giving an actual distance of 59.4 mm to compensate for the presence of a 25.4 mm beamsplitting cube. The beam parameter β was chosen to be 24050 m^{-1} giving a central spot diameter [6.3] of $4.81/\beta = 200 \text{ } \mu\text{m}$. The radius of the hologram is $R = 2.5 \text{ mm}$. The predicted maximum range of the beam is $z_{\text{max}} = Rk/\beta = 788 \text{ mm}$ where $k = 2\pi/\lambda$, and $\lambda = 0.83 \text{ } \mu\text{m}$. At the design distance of 50 mm from the hologram, the laser gave a $1/e$ spot size of 20.4 mm in one direction, and 7.06 mm in the orthogonal direction. Since these dimensions are much greater than the aperture radius of 2.5 mm, this beam is similar to the finite aperture Bessel beam approximation. Figure 6-4 has a scale drawing of the central portion of the

Figure 6-4. Schematic of experimental arrangement showing a scale view of the central portion of the hologram used in the experiments as well as a predicted intensity profile for the Bessel beam. The plane mirror together with the rest of the apparatus created a Michelson interferometer and was used only during interference experiments described in the text.

hologram, where the shaded regions represent areas having a thickness greater by $\lambda/4$ than that of the unshaded regions. The rings at the outer edge of the hologram are the thinnest and represent the minimum feature size of $\sim 7 \mu\text{m}$ for this hologram.

To measure its characteristics, the Bessel beam HOE was illuminated by a laser diode operating at $\lambda = 830 \text{ nm}$, with a 25.4 mm beamsplitting cube placed between the laser and the hologram. The beam profile was then measured at 2 cm intervals with a CCD camera, beginning at 12 cm from the hologram face. The measured efficiency of the hologram was found to be 33% which comes close to the theoretical maximum of 40.5% for a binary coded phase hologram. The central spot was measured by blocking out the rest of the beam with a pinhole. The central spot was found to contain 9.9% of the total beam energy. A plane mirror was later added to the apparatus to form a variation on the Michelson interferometer in order to examine the phase characteristics of the Bessel beam.

Figure 6-5 illustrates the beam $1/e^2$ intensity radius of the central spot at different points along the optical axis, as well as the $1/e^2$ intensity radius of a Gaussian beam of similar beam waist. The $1/e^2$ intensity radius of the Bessel beam is approximately $100 \mu\text{m}$ until the beam nears its ray optics maximum propagation distance of 0.78 m. This is the distance where the rays from the outermost regions of the hologram finally intersect, and beyond which none of its rays intersect the optic axis, creating a shadow zone. For comparison, a Gaussian beam with the same spot size has a confocal beam parameter of only 7.57 cm. Figure 6-6 presents the measured as well as predicted axial intensity from a distance of 12 cm from the hologram face up to a distance of 80 cm. The predicted axial intensity was calculated using the Fresnel-Kirchhoff diffraction

Figure 6-5. Measured peak to peak distance from central spot to first ring of the Bessel beam at various distances along the beam's propagation length. The solid line shows the calculated spot size of a Gaussian beam whose beam waist is equal to that of the central peak in the Bessel beam. The measured peak to peak data is also shown at a scale of $\times 10$. The measured data is approximately 1.6 times larger than the $1/e$ amplitude radius for the central peak in the Bessel beam.

Figure 6-6. The data points represent measured intensity of the beam's central spot averaged over five trials. The solid line represents the theoretical axial intensity.

formula in cylindrical coordinates

$$I(\rho) = \left(\frac{\pi}{\lambda z} \right)^2 \left| \int_0^R \exp[ik\sqrt{\rho^2 + \rho'^2 + z_0^2}] J_0(\rho \cdot \rho' k/z) \rho' d\rho' \right|^2 \quad (6-26)$$

where ρ is the radial coordinate in the observation plane, ρ' is the radial coordinate in the plane of the hologram, and z_0 is the axial distance from the hologram. The expected linear growth with distance is seen in Figure 6-6 together with an oscillation in the intensity whose amplitude and period increase with distance [6.3, 6.7]. Repeated measurements gave an error bar for the intensity values which for three standard deviations was less than 5%. Figure 6-7 shows photographs of the Bessel beam at a distance of 50 cm from the holographic element. Figure 6-7(a) shows a photograph of the Bessel beam profile with its intense central spot and ~9 outer rings. As mentioned, the Bessel beam has a transverse electric field amplitude which is proportional to the zeroth-order Bessel function of the first kind. The changes in sign of the Bessel function correspond then to borders separating areas of positive electric field amplitude from areas of negative electric field amplitude, and hence these borders separate areas having a phase shift of π radians [6.9]. The borders exist between each of the bright rings, so that each bright ring will be π out of phase with the next. To observe this characteristic, the Michelson interferometer in Figure 6-2 was created. Light from the laser diode is reflected from the plane mirror and forms a plane wave (large radius of curvature) at the observation plane which interferes with the Bessel beam, creating an interference pattern. Moving the plane mirror through a displacement of $\lambda/4$ ($\lambda/2$ optical path length change) shifted the interference pattern, showing that every ring of the Bessel beam is π radians

out of phase with its neighbouring ring. The two photographs of Figures 6-7(c) and 6-7(d) were produced showing, respectively, interference patterns where the odd-numbered rings and the even-numbered rings are suppressed.

Figure 6-7. Photographs of the beam taken at a distance of 50 cm from the hologram face. (a) Photo of the Bessel beam, and (b) an oscilloscope trace of a horizontal section through the center of the ring structure in (a). (c) and (d) are Michelson interferograms of the Bessel beam. With the central spot as the zeroth ring, in (c) the odd numbered rings are suppressed, while (d) shows the even numbered rings suppressed.

Chapter 7

DESIGN AND TEST OF OPTICAL CLOCK DISTRIBUTION SYSTEM

7.1 Optical Beam Routers

If the image stored in a hologram is that of a point source of light, then upon reconstruction with the proper reference beam, the real image of this point source will have a high amount of optical power directed into a relatively small area. The effect is the same as that produced when focussing an optical beam with a convex lens or a concave mirror; for this reason these types of holograms are often called holographic lenses.

Any of the different types of holograms discussed in the previous chapter can be used as beam routers. However some holograms make better beam routers than others. Because their efficiencies are so high, phase holograms are better suited to the job than amplitude holograms. If lenses can be tolerated in the system, Fourier transform holograms can be ideal beam routers. Unfortunately lenses add to the cost and complexity of an optical beam routing system, and take up valuable space as well, so rarely can their use be justified. Although it can be relatively inexpensive to produce one, photographically produced holograms have high unit production costs when high production volumes are desired. Conventional type, computer generated quantized phase holograms possess the most favourable characteristics for beam routers. They are highly efficient, require little space and nothing other than a reference beam to operate, and are easily reproducible on large scales.

When using conventional type phase holograms for beam routers, the simplest HOE (Holographic Optical Element), as mentioned in section 4.2, has zone boundaries identical to those of the Fresnel zone plate. The Fresnel zone plate records the interference pattern produced by a normally incident plane wave and a spherical wave. Using equation (4-11) and setting

$$\begin{aligned} U_r &= U_{or} e^{j\phi_r} \\ U_o &= U_{oo} e^{j\phi_o} \end{aligned} \quad (7-1)$$

we get

$$I = |U_{or}^2 e^{j2\phi_r} + U_{oo}^2 e^{j2\phi_o}| \quad (7-2)$$

By setting $U_{or}=U_{oo}$ and applying a few trigonometric identities, the above expression may be rewritten as

$$I = 2U_{or}^2 (1 + \cos(\phi_r - \phi_o)) e^{j(\phi_r + \phi_o)} \quad (7-3)$$

Note that the two phase terms of the interfering waves are subtracted in the amplitude portion of the interference pattern, whereas they are added in the phase portion. By setting $\phi_r=0$ to satisfy the condition of a normally incident plane reference wave, and $\phi_o=k(x^2+y^2+z_o^2)^{1/2}=kr$ for a spherical object wave, we get

$$I = 2U_{or}^2 (1 + \cos(-kr)) e^{jkr} \quad (7-4)$$

so the corresponding Fresnel zone plate will have an amplitude variation proportional to $1+\cos(-kr)$ and the corresponding phase-only hologram will have a phase variation $\exp(jkr)$. By quantizing the above expression so that there are only two allowable phase

levels per 2π radians, the result is $kr=n\pi$, where n is the nearest integer. If x^2+y^2 is defined as R^2 on the plane $z=0$, $kr=n\pi$ can be rewritten as

$$R^2 = \left(\frac{n\lambda}{2}\right)^2 - z_o^2 \quad (7-5)$$

which describes the radii of the zones.

7.1.1 Multifacet Holograms

So far only HOEs which produce single spots have been discussed. A multifacet HOE consists of several distinct areas or facets, each of which contain a spot producing hologram. The multifacet HOE is used to direct the light of a source beam into many spot images. Each facet of the hologram focuses light from the reference beam onto a different location, and in doing so provides optical fanout. Figure 7-1 shows a multifacet HOE. The size of each facet is determined by the irradiance distribution function of the reconstructing reference beam and the position of the facet. If equal power output beams are desired, then the powers incident on each facet must be equal. By knowing the irradiance distribution function of the reconstructing beam, and the desired position of the hologram facet, a facet size and shape can be chosen which receives the desired power.

7.2 One to Four Phase Hologram

Two one-to-four phase holograms are produced: one is a binary phase HOE, while the other is a quaternary phase HOE. Each HOE focusses the $0.83 \mu\text{m}$ spherical

Figure 7-1. A multifacet binary amplitude hologram. This hologram produces 4 spots from one input beam.

light wave from a laser diode to the same in line four spot pattern. The theoretical efficiency of the binary phase hologram is 40.5% and that of the quaternary phase HOE is 81%. The phases of each of the input and four output beams can be described as

$$\Phi = \exp(ikr) \quad , \text{where} \quad (7-6)$$

$$r = \sqrt{(x-x_o)^2 + (y-y_o)^2 + (z-z_o)^2}$$

and x_o , y_o , and z_o are different for each beam. According to equation (7-3) the phase of the resulting transmission function is the sum of the object and reference wave phase terms, and so for the binary phase hologram, the locus of points describing the boundaries between half zones is given by

$$\frac{n\lambda}{2} = \sqrt{(x-x_{o1})^2 + (y-y_{o1})^2 + (z-z_{o1})^2} + \sqrt{(x-x_{o2})^2 + (y-y_{o2})^2 + (z-z_{o2})^2} \quad (7-7)$$

where n is an integer. The above equation describes a family of ellipses when n is larger than a critical value. This critical value is the number of half wavelengths in the distance between the input beam source to the centre of the hologram to the focus of the output beam. A similar equation gives the locus of points describing the boundaries separating quarter zones for the quaternary phase hologram.

A computer program written in FORTRAN called TORCH.F was created to generate the masks necessary to produce a multilevel phase hologram of the single facet or multifacet variety. This program is listed in Appendix A.1. It is a user- friendly program which can generate the information necessary to create a focussing hologram given either a collimated or spherical input beam. The program works by solving

equation (7-7) where the phases are computed from information input by the program user. The program computes the data for one quadrant of one facet at a time. The area assigned to each facet is input by the user. The minimum zone number of each hologram is calculated by the computer; the program prompts the user for the size of the minor axis of the elliptical hologram and then computes the maximum zone number and minimum feature size of the mask based on this information. The data the program generates is written to a file in AutoCAD .DXF format so that the mask file may be viewed in AutoCAD before it is fabricated.

The data generated by TORCH.F is used to create an electron beam mask. The e-beam mask generating facility at Northern Telecom has high resolution and is capable of producing step sizes as small as $0.1\mu\text{m}$. The Northern Telecom facility accepts data in GDS2 CAD format, so the .DXF format files generated by TORCH.F must be converted to GDS2 format. One program which performs the necessary conversion is available from Octal Inc. This is the program currently used to convert the files generated by TORCH.F. Any shapes drawn in GDS2 must be closed polygons having no more than 200 vertices. To comply with this limitation, TORCH.F approximates the ellipses it draws with a set of polygons.

The holograms were designed to be placed at a distance of 150 mm away from the diode laser source. The distance of 150 mm was chosen to allow plenty of room for a CCD camera and an optical power meter to have access to the four spots. However, the distance between hologram and laser source could easily be shortened to less than 10 mm in a practical system. The holograms were designed to produce an array of four

spots all in line with the laser diode source. The spot nearest to the laser is 10 mm away, measured in the plane of the laser and detectors, and each successive spot is 8 mm further away. Each hologram facet is a portion of an ellipse as shown in Figure 7-2, and each facet has a surface area of such size that each collects approximately equal optical power. The arrangement illustrated in Figure 7-3 is used to evaluate the HOEs.

The surface area of each facet was determined by integrating the intensity function of the Gaussian beam from the laser diode over a chosen area to determine the optical power falling within that area. This process was repeated until an area for each of the four facets was determined such that each facet received roughly equal optical power. The shape of each facet was determined by a combination of the straight-line boundaries between the facets and the outermost ring of each hologram facet which represents the smallest feature size producible by Carleton's Department of Electronics fabrication facility.

Pulse broadening is an unavoidable characteristic of holographic beam routers. An optical pulse which is focussed by a hologram is broadened in time because of the differential path length available to photons travelling from the laser to the hologram and from there to the detector. A photon which reflects off the innermost ring of the hologram travels a shorter distance from laser to detector than one which is reflected off of the outermost ring. The maximum path difference for the optical arrangement used in this work is 183 μm . This distance corresponds to a delay of 0.61 ps in free

Figure 7-2. Outline of the facets used in both the binary and quaternary phase 1:4 multifacet holograms.

Figure 7-3. Arrangement for testing the 1:4 HOEs. The data collected by the CCD camera is sent to a PC for processing.

space, which can be considered insignificant compared to the pulse period of 1.0 ns.

The clock skew of the system is the time difference between the events of the first and last pulses striking their respective detectors. This time can be calculated using the maximum path difference for the light forming the first and last spots, and the velocity of light in free space. In the case of the system presented, the maximum path difference is 1.59 mm which corresponds to a delay of 5.3 ps. This delay corresponds to 0.5% of the 1.0 ns period of the clock signal. This type of system can have a maximum path difference of up to 150 mm which corresponds to a delay of less than 500 ps between signals.

The holograms were fabricated at Carleton University's Department of electronics fabrication lab by first growing a layer of SiO_2 onto a silicon wafer, a quarter wavelength thick layer for the binary phase hologram and a $3/8$ wavelength thick layer in the case of the quaternary phase hologram. Figure 7-4 illustrates the procedure. Next a layer of photoresist was spun onto the wafer, and the wafer was masked and exposed to UV radiation. The exposed wafer was developed and then bathed in a chemical etchant, eroding away the exposed photoresist. Next, the exposed regions of the SiO_2 layer were etched down to a depth of one quarter wavelength, completing the pattern transfer from the first mask to the wafer. In the case of the quaternary phase hologram the above process, beginning with spinning the wafer with photoresist, must be repeated while using the second mask instead of the first. The final step is to give the hologram a high reflectance by evaporating a layer of aluminum on the wafer. The aluminum layer is 80 nm thick, corresponding to 4.6

Figure 7-4. Lithographic steps in making a quaternary phase hologram.

skin depths in aluminum for radiation of wavelength $0.83\ \mu\text{m}$. Freshly evaporated aluminum has a maximum normal-incidence reflectance of 86.7% for radiation of $0.83\ \mu\text{m}$ wavelength [7.1] giving a binary phase hologram made with an aluminum coating a maximum efficiency of $0.405 \times 0.867 = 35.1\%$, and a similar quaternary phase hologram a maximum efficiency of $0.81 \times 0.867 = 70.2\%$.

7.3 Electronic Circuitry for Electro-Optic Conversion

In this optical clock distribution system an electrical, sinusoidal clock signal operating at 1 GHz is converted into an optical signal for distribution. At the receiving end of the signal, the optical signal must be converted back to an electrical one in order to communicate with the existing electronics. Optoelectronic components, such as laser diodes and photodiodes, are needed to perform the necessary conversions. Furthermore, additional circuitry is necessary to control the operation of the various optoelectronic components.

The laser diode used in this system is a Sharp model number LT015MD. This laser was chosen because of its high maximum optical output power of 40 mW, and its low attenuation at 1 GHz modulation frequency (<1dB). The Sharp LT015MD operates at a frequency of 830 nm at a case temperature of 25 °C and at 836 nm at a case temperature of 50 °C. Such a wavelength change will cause the photons hitting the outer ring of one of the holograms to be displaced a distance of $75\ \mu\text{m}$ from the focal point at the detector plane. This displacement of photons at the detector plane linearly decreases to zero displacement as the point where the photon strikes the hologram moves inward

toward the center. The effect of a change in wavelength, thus, is a broadening of the focused spot on each of the detectors.

A simple drive circuit was built as shown in Figure 7-5(a) which allowed the laser to be biased in the middle of its output power range at 77.7 mA, while a high frequency sine current wave of amplitude 31.6 mA modulated the laser. The DC biasing circuit has a resistance of 59 Ω including the 3 Ω resistance of the laser diode, and the AC biasing circuit has a resistance of 50 Ω including the resistance of the laser. The AC biasing circuit produces the 31.6 mA current signal from a 1.58 V , 1.0 GHz voltage sine wave and drives the laser from near minimum to maximum forward current when biased at 77.7 mA DC or, equivalently, 4.6 V across the biasing circuit. When the forward current is at its maximum value of 109 mA, the laser's optical power output is 40 mW, and when the forward current is at its minimum value of 46.1 mA the laser output is nearly zero. At maximum power output, roughly 1.46 mW of optical power was measured to strike each facet of the hologram. This small received power is due to the large angular spread of the incident beam together with the long distance from the laser to the hologram. The laser drive circuit is a simple one and contains no automatic power controller, and so the average optical output power will vary with the case temperature of the laser.

At the receiving end of the optical clock signal, an Antel Optronics model FS1002 silicon p-i-n photodiode converts the optical signal into a current signal. It has a responsivity of approximately 0.25 A/W to 830 nm radiation modulated at a frequency of 1 GHz, and its active area has a diameter of 200 μm . Figure 7-5(b)

(a)

(b)

Figure 7-5. (a) Drive circuit for the laser diode and (b) biasing circuit for each of the photodiodes.

illustrates the simple biasing circuit used for the photodiode, which produces a voltage waveform of frequency 1.0 GHz from the current signal produced by the photodiode. The biasing circuit is terminated into the 50 Ω input of either an HP 54120B Digitizing Oscilloscope or an Advantest R3361A spectrum analyzer which are both used to observe the recovered signal.

7.4 Characterization of HOEs

To predict the spot size of each of the hologram faces, a program using the Fresnel-Kirchhoff diffraction integral on a hologram having the outline of a full ellipse of equal major axis to those in the produced hologram was written and executed. The results, as well as the measured spot profiles are displayed in Figures 7-6 and 7-7. The focused spots from the binary hologram are approximately 100 μm in diameter, and are twice as large as the predicted spot size of $\sim 40 \mu\text{m}$. These enlarged spots were the result of charge spillover in a saturated CCD array. Unfortunately, time constraints precluded the option of recollecting spot profile data for the binary hologram, and thus the spot profiles for the binary phase hologram are of limited use, however they are presented nonetheless. The focused spots from the quaternary phase hologram, however, were enlarged five times by a positive lens, so that the image formed on the CCD array was large enough to avoid charge spillover effects. The spot profiles from the quaternary phase hologram were very close in width to that of the predicted profile, as shown in Figure 7-7(b). A comparison of the intensities of the predicted spot profile with those of the measured profiles was not performed

(a)

(b)

Figure 7-6. (a) Relative intensity profiles of the four spots produced by the binary phase HOE. (b) Normalized intensity profiles of first spot, measured and predicted.

(a)

(b)

Figure 7-7. (a) Normalized measured intensity profiles of each of the four spots produced by the quaternary phase hologram. (b) Normalized intensity profiles of the fourth spot, the solid line represents a theoretical profile and the data points represent the measured profile.

because the necessary equipment was not available. The CCD array used to take the spot profile data could give only relative intensity, and was not calibrated to give absolute intensity.

As mentioned previously, the measured maximum power incident upon each facet of the holograms was approximately 1.46 mW. The maximum power measured in each spot produced by the binary phase hologram was roughly 0.51 mW, giving a measured efficiency of 35%, this is close to the maximum efficiency of 35.1% for an aluminum coated binary phase hologram. The measured maximum power in each spot from the quaternary phase hologram was 1.0 mW, corresponding to a measured efficiency of 69% and approaches the maximum efficiency of 70.2% for a quaternary phase hologram with an aluminum coating operating at an 830 nm wavelength.

The maximum power of 0.51 mW produced by each facet of the binary phase hologram is incident on the photodetector in the receiver circuit of Figure 7-5(b), and the detectors responsivity of 0.25 A/W produce a current signal of 0.1275 mA maximum current. The signal is measured across the 50 Ω impedance of the HP digital oscilloscope and so becomes a voltage signal with a theoretical maximum of 6.375 V. Since the laser is biased so that the minimum output power during its 1 ns period is nearly zero, a voltage sine wave of frequency 1 GHz and amplitude 3.19 mV is expected. Figure 7-8a shows a photograph of the actual oscilloscope trace showing a voltage signal of frequency 1 GHz and amplitude 2.4 mV. Likewise, the 1.0 mW maximum optical signal from the quaternary phase hologram will produce, via the receiver circuit, a voltage signal with an ideal maximum of 12.5 mV. Figure 7-8(b)

(a)

(b)

Figure 7-8. Photographs of the oscilloscope traces showing the recovered clock signal using (a) the binary-phase HOE, and (b) the quaternary-phase HOE.

shows the actual oscilloscope trace produced by the optical signal from the quaternary phase hologram; it has an amplitude of approximately 5 mV and a voltage difference between peak and trough of about 10 mV. It is believed that the low experimental voltage readings can be attributed to signal attenuation in the drive circuit for the laser diode and in the biasing circuit for the photodiode. The signal strength compared to the total noise in the system is shown in the spectrum analyzer traces of Figure 7-9. Note that the signal from the binary phase hologram is 21 dB above the noise, while that from the quaternary phase hologram is 28 dB above the noise level.

(a)

(b)

Figure 7-9. Photographs of spectrum analyzer traces showing the strength of the recovered clock signal and the background noise using (a) the binary-phase HOE, and (b) the quaternary-phase HOE.

Chapter 8

CONCLUSIONS AND FUTURE WORK

As a result of the research presented in this thesis, it seems apparent that an optical clock distribution system employing HOEs can effectively route a 1.0 GHz optical clock signal from an optical source to the appropriate detectors. The optical source used was an ordinary diode laser operating at an 0.83 μm wavelength, and the detectors used were simple silicon p-i-n photodiodes. The HOEs used were binary and quaternary, one to four, phase only, reflection holograms. The measured efficiency of the binary and quaternary holograms were 35% and 69% respectively.

The recovered clock signal was converted, optoelectronically by the photodiodes, into an electrical signal without amplifying this signal. It was found that the recovered signal had a SNR of 21 dB in the case of the binary phase HOE, and a SNR of 28 dB for the quaternary phase HOE, indicating that the quaternary phase hologram is roughly twice as efficient as the binary phase type..

The HOEs used to split and route the optical clock signal work by simply focussing the light from the laser diode. Another HOE was also presented which produces a so-called Bessel beam from the laser diode light. Such a beam has a profile which can be described by the zeroth-order Bessel function of the first kind. This Bessel beam has a central spot having a diameter of 200 μm , which is constant $\pm 5\%$ for the beam's propagation distance of 78 cm.

Future work in this area will involve the construction of a clock distribution

system incorporating transmission HOEs etched into a diffractive/reflective substrate [8.1]. Such a method can be more spatially efficient, and could route a signal to a farther lateral destination than can the system presented in this dissertation. An interconnection system using Bessel beams will also be built. One method of using the Bessel beam in an interconnection system involves connecting elements on several daughter cards via an optical backplane. Another involves the use of a Dove prism lying over the circuit board to guide the Bessel beam from the HOE to the detector. Both of these schemes will be considered further.

Recent developments in quantized holography have lead to a new type of binary phase hologram which achieves, theoretically, nearly 100% diffraction efficiency at Bragg incidence [8.2]. This type of hologram has features whose transverse dimension is smaller than the wavelength of the light used on them, and are able to diffract this light at high efficiencies when it satisfies the Bragg condition. Such a design is an alternative approach which avoids the problems of alignment which are inherent in multilevel holograms.

REFERENCES

- [1.1] J. Turunen, J. Fagerholm, A. Vasara, and M.R. Taghizadeh, "Detour-phase kinoform interconnects: the concept and fabrication considerations," *J. Opt. Soc. Am. A* **7**, 1202-1208 (1990).
- [1.2] J.W. Goodman, F.J. Leonberger, S. Kung, and R.A. Athale, "Optical interconnections for VLSI systems," *Proc. of the IEEE* **72**, 850-866, (1984).
- [1.3] D.A.B. Miller, "Optics for low-energy communication inside digital processors: quantum detectors, sources, and modulators as efficient impedance converters," *Opt. Lett.* **14**, 146-148 (1989).
- [2.1] P.R. Haugen, S. Rychnovsky, A. Husain, "Optical interconnects for high speed computing," *Opt. Eng.* **25**, 1076-1084 (1986).
- [2.2] H. Kroger, C. Hilbert, U. Ghoshal, D. Gibson, and L. Smith, "Applications of superconductivity to packaging," *IEEE Circuits and Devices Magazine* **5**, 16-21 (1989).
- [2.3] J.W. Goodman, "Optical interconnections in computing," in *Computation and Cognition. Proceedings of the First NEC Research Symposium*, 22-31 (1991).
- [2.4] N. Streibl, K. Brenner, A. Huang, J. Jahns, J. Jewell, A.W. Lohmann, D.A.B. Miller, M. Murdocca, M.E. Prise, and T. Sizer, "Digital optics," *Proc. of the IEEE* **77**, 1954-1969 (1989).
- [2.5] L.D. Hutcheson, P. Haugen, and A. Husain, "Optical interconnects replace hardware," *IEEE Spectrum*, 30-35, March 1987.
- [2.6] P.R. Prucnal, "VLSI optical interconnection networks," *SPIE* **703**, Integration and

packaging of optoelectronic devices, 106-115 (1986).

[2.7] J.A. Neff, "Major initiatives for optical computing," *Opt. Eng.* **26**, 2-9 (1987).

[2.8] C.T. Sullivan, "Optical waveguide circuits for printed wire-board interconnections," *SPIE 994 Optoelectronic Materials, Devices, Packaging, and Interconnects II*, 92-100 (1988).

[3.1] H.J. Caulfield, Sun Lu, The Applications of Holography, Wiley-Interscience, Toronto 1970.

[3.2] H.M. Smith, Principles of Holography, Wiley-Interscience, Toronto, 1969.

[3.3] P.W. Milonni, J.H. Eberly, Lasers, Wiley-Interscience, Toronto, 1988.

[3.4] M.V. Klein, T.E. Furtak, Optics, John-Wiley and Sons, Toronto, 1986.

[3.5] E. Hecht, Optics, Second Edition, Addison-Wesley Publishing Co., Reading, Massachusetts, 1987.

[3.6] J.W. Goodman, Introduction to Fourier Optics, McGraw-Hill Book Co., San Francisco, 1968.

[3.7] J.A. Ratcliffe, "Some Aspects of Diffraction Theory and Their Application to the Ionosphere," in A.C. Strickland (ed.), Reports on Progress in Physics, XIX, The Physical Society, London, 1956.

[4.1] D. Gabor, "A new microscopic principle," *Nature* **161**, 777 (1948).

[4.2] B.E.A. Saleh, M.C. Teich, Fundamentals of Photonics, Wiley-Interscience, Toronto, 1991.

[4.3] L.B. Lesem, P.M. Hirsch, J.A. Jordan, Jr., "The Kinoform: A new wavefront reconstruction device," *IBM J. Res. Develop.* **13**, 150-155 (1969).

- [4.4] W.H. Lee, "Computer generated Holograms: techniques and applications," *Progress in Optics*, E. Wolf editor, XVI, 119-232 (1978).
- [4.5] J.W. Goodman, A.M. Silvestri, "Some effects of Fourier-domain phase quantization," *IBM J. Res. Develop.* 13, 478-484 (1970).
- [4.6] T.J. McHugh, D.A. Zweig, "Recent Advances in Binary Optics," in Holographic Optics: Optically and Computer Generated, Proc. SPIE 1052, 85-90 (1990).
- [4.7] S. Kirkpatrick, C.D. Gelatt, and M.P. Vecchi, "Optimization by simulated annealing," *Science* 220, 671-680 (1983).
- [5.1] J.A. Cox, T. Werner, J. Lee, S. Nelson, B. Fritz, and J. Bergstrom, "Diffraction efficiency of binary optical elements," SPIE 1211 Computer and Optically Formed Holographic Optics, 116-124 (1990).
- [5.2] M.W. Farn, J.W. Goodman, "Effect of vlsi fabrication errors on kinoform efficiency," SPIE 1211 Computer and Optically Formed Holographic Optics, 125-136 (1990).
- [6.1] J. Durnin, "Exact solutions for nondiffracting beams. I. The scalar theory," *J. Opt. Soc. Am. A* 4, 651-654 (1987).
- [6.2] J. Durnin, J. J. Miceli Jr., and J. H. Eberly, "Diffraction-free beams," *Phys. Rev. Lett.* 58, 1499-1501 (1987).
- [6.3] J. Turunen, A. Vasara, and A. T. Friberg, "Holographic generation of diffraction-free beams," *Appl. Opt.* 27, 3959-3962 (1988).
- [6.4] R. M. Herman, and T. A. Wiggins, "Production and uses of diffractionless beams," *J. Opt. Soc. Am, A* 8, 932-942 (1991).

- [6.5] L. C. Laycock, and S. C. Webster, "Bessel beams: their generation and application," *GEC J. Research* **10**, 36-51 (1992).
- [6.6] Takehiko Hidaka, "Generation of a diffraction-free laser beam using a specific Fresnel zone plate," *Japanese Journal of Applied Physics* **30**, 1738-1739 (1991).
- [6.7] N. Davidson, A. A. Friesem, and E. Hasman, "Efficient formation of nondiffracting beams with uniform intensity along the propagation direction," *Opt. Commun.* **88**, 326-330 (1992).
- [6.8] A. Vasara, J. Turunen, and A. T. Friberg, "Realization of general nondiffracting beams with computer-generated holograms," *J. Opt. Soc. Am. A* **6**, 1748-1754 (1989).
- [6.9] Y. Lin, W. Seka, J. H. Eberly, H. Huang, and D. L. Brown, "Experimental investigation of Bessel beam characteristics," *Appl. Opt.* **31**, 2708-2713 (1992).
- [6.10] E.T. Whittaker, "On the partial differential equations of mathematical physics," *Math. Ann.* **57**, 333-355 (1902).
- [6.11] F. Gori, G. Guattari, and C. Padovani, "Bessel-Gauss Beams," *Opt. Commun.* **64**, 491-495 (1987).
- [6.12] I.S. Gradshteyn, and Tyzhik, Table of Integrals, Series, and Products, Academic Press, New York, ch. 6 (1980).
- [6.13] R.P. MacDonald, J. Chrostowski, S.A. Boothroyd, and B.A. Syrett, "Holographic formation of a diode laser 'non-diffracting' beam, " *Appl. Opt.*, In Press (1993).
- [6.14] P. L. Overfelt, and C. S. Kenney, "Comparison of the propagation characteristics of Bessel, Bessel-Gauss, and Gaussian beams diffracted by a circular aperture," *J. Opt. Soc. Am. A* **8**, 732-745 (1991).

- [7.1] G. Hass, in R. Kingslake, ed., Applied Optics and Optical Engineering, vol. III, 309-330, Academic Press, New York, 1965.
- [8.1] K.H. Brenner, and F. Sauer, "Diffractive-reflective optical interconnects," *Appl. Opt.* **27**, 4251-4254 (1988).
- [8.2] J. Turunen, P. Blair, J. M. Miller, M.R. Taghizadeh, and E. Noponen, "Bragg holograms with binary synthetic surface-relief profile," *Opt. Lett.* **18**, 1022-1024 (1993).

Appendix A

Torch.f

c Torch.f
c This program computes the equations for a holographic optical
c element and produces the first level mask. A dark field must be
c specified.

c Ryan MacDonald January, 1992

```
real cx,pi,min,max,angle,phi,xlo,xhi,ylo,yhi,xo,yo,zo,xi,yi,zi,  
+mask,l,m  
open(unit=10,file='holomask.dxf',status='new')
```

```
pi= 3.1415927
```

c ecc is the eccentricity of the ellipse, that is the major axis
c divided by the minor axis.

```
print*, 'Welcome to Torch. Please enter the position, in'  
print*, 'x-y-z coordinates, of the source and detector. The'  
print*, 'hologram will lie in the plane z=0.'  
print*, ' '  
print*, 'Enter x coordinate of source'  
read*, xs  
print*, 'Enter y coordinate of source'  
read*, ys  
print*, 'Enter z coordinate of source'  
read*, zo  
print*, 'Would you like to produce a collimated beam (enter 1)  
+ or a focused beam (enter 2)?'  
read*, beam  
if (beam.gt. 1.5) then  
print*, 'Enter x coordinate of detector'  
read*, xd  
print*, 'Enter y coordinate of detector'  
read*, yd
```

```

print*, 'Enter z coordinate of detector'
read*, zi
yi= sqrt((xd-xs)**2+(yd-ys)**2)
center=(yi)*zo/(zo+zi)
else
  print*, 'For the collimated beam, enter the refractive index
+ of the medium that the collimated beam will be propagating
+ through.'
  read*, m
  print*, 'Enter the angle, in radians, the beam makes with
+ the x-axis'
  read*, xx
  xd=cos(xx)
  print*, 'Enter the angle, in radians, the beam makes with
+ the y-axis'
  read*, yy
  yd=cos(yy)
  print*, 'Enter the angle, in radians, the beam makes with
+ the z-axis'
  read*, zz
  beta= sqrt(1-(cos(zz))**2)
  print*, 'beta=', beta
  center= zo*tan(asin(m*sin(zz)))
  print*, 'center=', center
endif
print*, 'Please indicate which mask level this mask is for.
+An integer number (1,2,3,...) is required.'
read*, mask
print*, 'Enter the wavelength of light emitted by the source.'
print*, ' '
read*, l

write (10,5) 0
write (10,11) 'SECTION'
write (10,5) 2
write (10,12) 'ENTITIES'
write (10,5) 0

yo= 0.0
print*, 'wavelength=', l

```

- c Calculate minimum number of wavelengths from source to hologram to
- c detector.

```

if (beam.gt.1.5) then
min=(sqrt(center**2+zo**2)+sqrt((yi-center)**2+zi**2))/1+.01
c   max= min+3
else
min= (sqrt(center**2+zo**2)-m*beta*(center+yo))/1+.01
endif
print*, 'min=', min
print*, 'distance to hologram center is', center
print*, 'Enter the radius you wish the hologram to have.'
read*, r

c Calculate maximum number of wavelengths from source to hologram to
c detector.
if (beam.gt.1.5) then
max= (sqrt(center**2+r**2+zo**2)+sqrt((yi-center)**2+zi**2+
+ r**2))/1
size= 1/(2*mask*(sin(atan(r/zo))+sin(atan(r/zi))))
if ((xd-xs).GT.-0.00001.AND.(xd-xs).LT.0.00001) then
angle= 0.0
else
if ((xd-xs).GT.0.0) then
angle= atan((yd-ys)/(xd-xs))-pi/2.0
else
angle= atan((yd-ys)/(xd-xs))+pi/2.0
endif
endif
print*, 'angle=', angle
c   Angle determines the angle of the y-axis or major axis of
c   the ellipse with the y-axis of the coordinate space.
else
max= (sqrt(center**2+r**2+zo**2)-m*beta*(center+yo))/1
c   min=max-5
size= 1/(2*mask*(sin(atan(r/zo))))
if ((xd).GT.-0.00001.AND.(xd).LT.0.00001) then
angle= 0.0
else
if ((xd).GT.0.0) then
angle= atan((yd)/(xd))-pi/2.0
else
angle= atan((yd)/(xd))+pi/2.0
endif
endif
print*, 'angle is=', angle
endif

```

```

print*, 'The minimum linewidth of this hologram is', size
print*, 'If this linewidth is too small, use a smaller radius.'

```

```

c   if ((xd-xs).GT.-0.00001.AND.(xd-xs).LT.0.00001) then
c       angle= 0.0
c   else
c       if ((xd-xs).GT.0.0) then
c           angle= atan((yd-ys)/(xd-xs))-pi/2.0
c       else
c           angle= atan((yd-ys)/(xd-xs))+pi/2.0
c       endif
c   endif
c   print*, 'angle=', angle
c       Angle determines the angle of the y-axis or major axis of
c       the ellipse with the y-axis of the coordinate space.
phi= 0.0000001
c   Phi is the starting angle for the calculations.
cx= xs
print*, 'enter the spatial limits of the hologram face'
print*, 'enter the minimum limit in the x-direction'
read*, xlo
print*, 'enter the maximum limit in the x-direction'
read*, xhi
print*, 'enter the minimum limit in the y-direction'
read*, ylo
print*, 'enter the maximum limit in the y-direction'
read*, yhi

print*, 'Calculating hologram mask.'

ticket=1

call polyline(min,max,xs,ys,angle,phi,xlo,xhi,ylo,yhi,mask,
+yi,yo,zi,zo,l,beta,beam,m)

phi=phi + 90
call polyline(min,max,xs,ys,angle,phi,xlo,xhi,ylo,yhi,mask,
+yi,yo,zi,zo,l,beta,beam,m)

phi=phi + 90
call polyline(min,max,xs,ys,angle,phi,xlo,xhi,ylo,yhi,mask,
+yi,yo,zi,zo,l,beta,beam,m)

phi=phi + 90

```

```
call polyline(min,max,xs,ys,angle,phi,xlo,xhi,ylo,yhi,mask,
+yi,yo,zi,zo,l,beta,beam,m)
```

```
print*, 'end'
```

c Now close dxf file

```
write (10,8) 'ENDSEC'
write (10,5) 0
write (10,17) 'EOF'
```

```
5 format (I3)
8 format (A6)
11 format (A7)
12 format (A8)
16 format (A1)
17 format (A3)
end
```

```
*****
```

```
subroutine polyline(min,max,xs,ys,angle,phi,xlo,xhi,ylo,yhi,
+mask,yi,yo,zi,zo,l,beta,beam,m)
```

```
*****
```

c This subroutine computes the verticies of polylines necessary
c to make a hologram, and produces a dxf file containing this
c information.

```
real min,max,xs,ys,cy,angle,theta,rminor,ecc,xi,yi,x,y,n,t,l,
+aa,bb,cc,mask,lest,m
integer tag,ticket,flag
```

```
pi= 3.1415927
```

c Calculate the ring of hologram for each n, where n is the number
c of wavelengths from source to hologram to detector.

```
do 10 n=min,max,1/mask
flag= 0
mark= 1
```

```
do 20 t=0,1
if (beam.gt.1.5) then
```

```

      lest= -(-2*zi**2*zo**2+zi**4+zo**4+yo**4-4*yi**3*yo+
+2*zi**2*yo**2+yi**4+2*yi**2*zo**2+6*yi**2*yo**2-4*yi*yo**3+
+2*yi**2*zi**2-4*yi*zi**2*yo-4*yo*zo**2*yi+2*yo**2*zo**2+
+(n+t/(2*mask))**4*l**4+(-2*yo**2-2*zi**2-2*zo**2-2*yi**2+
+4*yi*yo)*(n+t/(2*mask))**2*l*l)/(-4*(n+t/(2*mask))**2*l*l+
+4*yi**2-8*yi*yo+4*yo**2)
      rminor=sqrt(lest)
      ecc= ((n+t/(2*mask))*l)/sqrt((n+t/(2*mask))**2*l**2+2*yi*yo
+-yi**2-yo**2)
      cy= (-yo*zo**2-yi*zi**2+yi*zo**2+yi*yo**2+zi**2*yo
++yi**2*yo-yi**3-yo**3+(yi+yo)*(n+t/(2*mask))**2*l*l)/
+(2*((n+t/(2*mask))**2*l*l+2*yi*yo-yi**2-yo**2))
      else
      lest= (2*yo*(n+t/(2*mask))*l*m*beta-zo**2+(n+t/(2*mask))
+**2*l**2+m**2*beta**2*yo**2+m**2*beta**2*zo**2)
+/(1-m**2*beta**2)
      rminor=sqrt(lest)
      ecc= 1/(sqrt(1-m**2*beta**2))
      cy= ((n+t/(2*mask))*l*m*beta+yo)/(1-m**2*beta**2)
      endif

tag= 1

      betr= phi + 90
      aa= phi*phi**(-t)*betr**t
      bb= betr*betr**(-t)*phi**t
      if (bb.lt.0.0001) then
      bb= 0.0
      endif
      cc= (-1)**t
      do 30 theta= aa, bb, cc
      xp= rminor*cos(theta*pi/180)
      yp= rminor*ecc*sin(theta*pi/180) +cy
      x= xp*cos(angle) - yp*sin(angle) +xs
      y= xp*sin(angle) + yp*cos(angle) +ys
      if ((x.gt.xlo).and.(x.lt.xhi).and.(y.gt.ylo).and.(y.lt.yhi))
+ then
      if (flag.eq.0) then
      write (10,9) 'POLYLINE'
      write (10,6) 8
      write (10,13) '0'
      write (10,6) 66
      write (10,7) 1
      write (10,6) 10

```

```

write (10,14) '0.0'
write (10,6) 20
write (10,14) '0.0'
write (10,6) 30
write (10,14) '0.0'
write (10,6) 70
write (10,7) 1
write (10,6) 0
mark=0
flag= 1
endif

write (10,15) 'VERTEX'
write (10,6) 8
write (10,13) '0'
write (10,6) 10
write (10,4) x
write (10,6) 20
write (10,4) y
write (10,6) 30
write (10,14) '0.0'
write (10,6) 0
lastag= tag
tag= 0

```

- c Now, add the points of the ellipse which lie on the borders of the
c quadrant of the hologram.

```

elseif ((x.lt.xlo).and.(tag.eq.0)) then
  y= ((y-ylast)/(x-xlast))*(xlo-xlast)+ylast
  x=xlo
  write (10,15) 'VERTEX'
  write (10,6) 8
  write (10,13) '0'
  write (10,6) 10
  write (10,4) x
  write (10,6) 20
  write (10,4) y
  write (10,6) 30
  write (10,14) '0.0'
  write (10,6) 0
  tag= 1

```

```

elseif ((x.gt.xhi).and.(tag.eq.0)) then

```

```

    y= ((y-ylast)/(x-xlast))*(xhi-xlast)+ylast
    x= xhi
    write (10,15) 'VERTEX'
    write (10,6) 8
    write (10,13) '0'
    write (10,6) 10
    write (10,4) x
    write (10,6) 20
    write (10,4) y
    write (10,6) 30
    write (10,14) '0.0'
    write (10,6) 0
    tag= 1

elseif ((y.lt.ylo).and.(tag.eq.0)) then
    x= ((x-xlast)/(y-ylast))*(ylo-ylast)+xlast
    y= ylo
    write (10,15) 'VERTEX'
    write (10,6) 8
    write (10,13) '0'
    write (10,6) 10
    write (10,4) x
    write (10,6) 20
    write (10,4) y
    write (10,6) 30
    write (10,14) '0.0'
    write (10,6) 0
    tag= 1

elseif ((y.gt.yhi).and.(tag.eq.0)) then
    x= ((x-xlast)/(y-ylast))*(yhi-ylast)+xlast
    y= yhi
    write (10,15) 'VERTEX'
    write (10,6) 8
    write (10,13) '0'
    write (10,6) 10
    write (10,4) x
    write (10,6) 20
    write (10,4) y
    write (10,6) 30
    write (10,14) '0.0'
    write (10,6) 0
    tag= 1

```



```
endif  
xlast= x  
ylast= y
```

```
30 continue  
if ((lastag.eq.1).and.(tag.eq.0)) then  
  write (10,15) 'VERTEX'  
  write (10,6) 8  
  write (10,13) '0'  
  write (10,6) 10  
  write (10,4) x  
  write (10,6) 20  
  write (10,4) y  
  write (10,6) 30  
  write (10,14) '0.0'  
  write (10,6) 0  
endif
```

```
20 continue  
if (mark.eq.0) then  
  write (10,15) 'SEQEND'  
  write (10,6) 8  
  write (10,13) '0'  
  write (10,6) 0  
endif
```

```
10 continue  
return  
6 format (I3)  
7 format (I6)  
9 format (A8)  
13 format (A1)  
14 format (A3)  
15 format (A6)  
4 format (F11.7)  
end
```

END

2007-94

FIN