Micro-Optic Elements for a Compact Opto-electronic Integrated Neural Coprocessor

by

William Frederick Herrington Jr.

B.S., University of Michigan Ann Arbor (2002)
S.M., Massachusetts Insitute of Technology (2005)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2015

Author . Signature redacted ....
Department of Electrical Engineering and Computer Science
September 15, 2014

Signature redacted

Certified by...
Cardinal Warde
Professor of Electrical Engineering
Thesis Supervisor

Accepted by.............. Signature redacted ....
Leslie A. Kolodziejski
Chair, Department Committee on Graduate Theses

Micro-Optic Elements for a Compact Opto-electronic Integrated Neural Coprocessor

by

William Frederick Herrington Jr.

Submitted to the Department of Electrical Engineering and Computer Science
on September 16, 2014, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Electrical Engineering and Computer Science

Abstract

The research done for this thesis was aimed at developing the optical elements needed for the Compact Opto-electronic Integrated Neural coprocessor (COIN coprocessor) project. The COIN coprocessor is an implementation of a feed forward neural network using free-space optical interconnects to communicate between neurons. Prior work on this project had assumed these interconnects would be formed using Holographic Optical Elements (HOEs), so early work for this thesis was directed along these lines. Important limits to the use of HOEs in the COIN system were identified and evaluated. In particular, the problem of changing wavelength between the hologram recording and readout steps was examined and it was shown that there is no general solution to this problem when the hologram to be recorded is constructed with more than two plane waves interfering with each other. Two experimental techniques, the holographic bead lens and holographic liftoff, were developed as partial workarounds to the identified limitations. As an alternative to HOEs, an optical element based on the concept of the Fresnel Zone Plate was developed and experimentally tested. The zone plate based elements offer an easily scalable method for fabricating the COIN optical interconnects using standard lithographic processes and appear to be the best choice for the COIN coprocessor project at this time. In addition to the development of the optical elements for the COIN coprocessor, this thesis also looks at the impact of optical element efficiency on the power consumption of the COIN coprocessor. Finally, a model of the COIN network based on the current COIN design was used to compare the performance and cost of the COIN system with competing implementations of neural networks, with the conclusion that at this time the proposed COIN coprocessor system is still a competitive option for neural network implementations.

Thesis Supervisor: Cardinal Warde
Title: Professor of Electrical Engineering

Acknowledgments

The list of people to whom I owe thanks for the help and support I've received in completing this thesis would fill several pages, so I'll have to acknowledge them by group. I would like to thank the professors I've worked with, in teaching and research positions, for their support and guidance. I would also like to thank the members of MIT Lincoln Laboratory Group 83 and the staff at MTL for their help and patience as I worked through the lithographic processes used to complete this research. To the current and former members of both Professor Warde's Photonic Systems Group and Professor Ram's Physical Optics and Electronics Group, thank you for teaching me new ways to look at problems and strange ways to build things. Finally, I would like to thank my family for the continued encouragement and enthusiasm that made this possible.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The Photonic Systems Group at the Massachusetts Institute of Technology has been developing the Compact Opto-electronic Integrated Neural Coprocessor (COIN coprocessor). The COIN coprocessor is a hybrid optical-electronic neural network based co-processor and the project aims to combine the advantages of optical interconnects, including fast summing and high interconnect density, with electronics to implement aspects of neural networks difficult to achieve optically, such as flexible, non-linear transfer functions. The advantage of the COIN coprocessor over other optical-electronic hybrid neural network schemes will be the compact interconnects and the high density of neurons obtained by integrating the detector, driver, and optical source elements on a single chip. The work done for this thesis was directed at the development of several types of optical interconnect elements for use in the COIN system. Although the interconnects were designed with the neural coprocessor in mind, the techniques and interconnect elements developed for this thesis will have application in other systems utilizing optical interconnects.

This chapter will review the background and motivation for the development of the optical interconnects by presenting the previous work done on hybrid optical-electronic neural networks within and outside of the Photonic Systems Group. As part of the background material a brief overview of Neural Network theory will be presented. The chapter will end with an brief outline of the remainder of the thesis.

1.1 Neural Network Overview

Many advances have been made in the design, implementation, and programming of microprocessors. These advances have allowed the application of computationally intensive algorithms to a wide variety of problems which were difficult or impossible to solve a generation ago. However, certain classes of problems in computer science still remain difficult to address using traditional programming techniques. In particular, problems in feature extraction, classification, and discrimination run into difficulties due to the nature of defining the features to be extracted, the classes to be identified, and the exact distinction between data sets for discrimination. A family of techniques, collectively known as Machine Learning, hopes to address these and other issues by "training" systems to perform these tasks without explicitly defining an algorithm for performing the tasks.

A machine learning technique consists of a model with a number of parameters that can be adjusted to change the models input-output behavior. The parameters are adjusted during the training stage to set the behavior of the system. In "supervised learning", the type of learning assumed for most work on the COIN coprocessor, during the training stage the system is presented with known input / desired output pairs. Ideally the trained model will then give the correct outputs when presented with the inputs used during training and will generalize to give the correct outputs when presented with inputs that are different from but similar to the inputs used during the training phase. The exact details of how a system is trained and implemented will depend on the system and task for which it is being trained [6]. One machine learning technique suitable for direct implementation in hardware is the Artificial Neural Network.

The inspiration for artificial neural networks, at least for the perceptron model of Rosenblatt [7], and the source of much of the terminology used, is an analogy with how the brain processes information. The fundamental unit of processing power in a modern computer could be argued to be the transistor, or at a slightly higher level the logic gate. The basic functional unit of the brain as currently understood is the neuron [8]. Neurons in the brain are not arranged into logic gates computing boolean algebra like the transistors in a microprocessor, but instead are heavily interconnected with each other and the behavior of one neuron influences the behavior of other neurons in a complex way. It is clear that networks of neurons are capable of performing complex computational tasks, and some promising results have been achieved with networks of artificial neurons.

An artificial neural network is composed of a number of artificial neurons connected to each other through synapses. A number of network topologies are possible, and in general communication between neurons can be bi-directional along the synapses. The operation of a biological neuron is very complicated and depends on several variables that may include past neuron activity, the availability of chemicals at the synapse between neurons, and the local activity of other neurons. Fortunately the exact modeling of biological neurons is not necessary to preform useful work. For the COIN coprocessor the neurons will be arranged in planes, with synapses connecting neurons between planes, and with signals propagating in one direction only along the synapses making the COIN coprocessor a feed-forward neural network. The simplified neuron model used in the COIN coprocessor, and commonly used for artificial neural network research, is diagrammed below in Figure 1-1a.

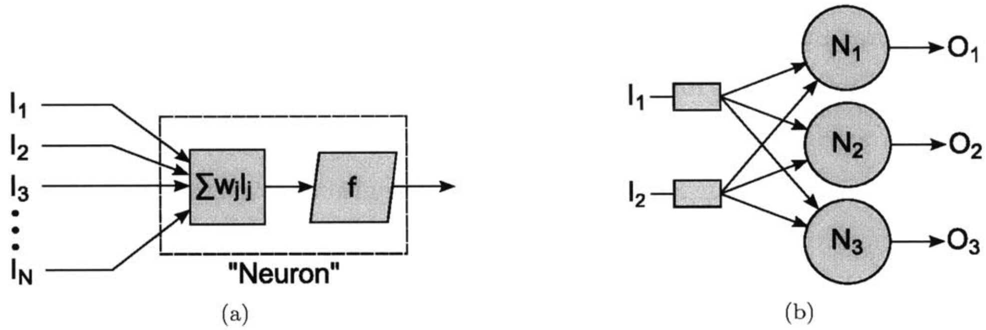


Figure 1-1: (a) A single neuron (b) A single layer Perceptron. In each case a set of inputs, I_i, are sent to the neuron(s). The inputs are summed, and the sum fed into a nonlinear "activation" function. For the single neuron there is one output, but for the single layer Perceptron there are several outputs.

The operation of an artificial neuron can be divided into two steps. The artificial neuron first computes a weighted sum of its inputs and then passes the result to a non-linear activation function which determines its output value. The simplest activation function is a step function at some threshold value, with the output of the neuron turning on whenever the sum of the input values is larger than the threshold value. McCulloch and Pitts [9, p. 129] showed that networks of artificial neurons of this type could be used to implement a Turing machine (they can compute the same set of numbers computable by a Turing machine). A single neuron can also be used to preform linear classification, and a great deal of work describing this behavior was done by Rosenblatt who referred to this basic operation as a 'perceptron' [7]. A simple network of several of these neurons sharing the same inputs can be used for multi-class classification problems. Early work with the perceptron used a step function activation as described above but other activation functions can be used. In particular, continuous and differentiable sigmoid type activation functions are used to simplify the training process for multi-layer systems.

All machine learning techniques require some set of variables which can be adjusted during training to allow the machine to "learn". In the artificial neural network model the function of the neuron is dependent on the choice of weights and activation thresholds, and it is these values which are modified during training. For a single neuron, or a single layer network as diagrammed above, one approach to training the network is to use supervised learning. A sample set of inputs and desired outputs is selected. The weights and threshold values are initialized to some value, commonly zero or random. A sample from the training set is input into the network and the output computed. The difference between the actual output and desired output is used to adjust the weights, and the amount of change depends on the current weight and the difference between the current output and the desired output. This process is repeated with a new sample from the training set and continued until the error for samples in the training set is small enough. Detailed treatment of this process can be found in references [7], [10].

There are some limits to the types of functions single layer perceptrons can be trained to implement. For example a single artificial neuron as described above cannot be trained to compute the exclusive-or of its inputs because the output sets are not linearly separable. Minsky and Papert explored some other limits of single layer perceptrons in the book 'Perceptrons' [10], first published in 1969. An interesting limit identified in the book is the inability of a single layer of diameter limited neurons, neurons connected to only some subset of inputs, to compute a function which requires knowledge of the state of every input [10, p.12-13]. Research published after 'Perceptrons' showed that at least some of the limits identified in 'Perceptrons' can be overcome by moving from a single layer system to a multi-layer system. A simple multi-layer system is shown in Figure 1-2.

The multi-layer perceptron is characterized by the presence of one or more "hidden" layers, so called because their outputs are hidden from the user. For the network diagrammed in Figure 1-2, the input layer distributes the network inputs across the first layer of neurons which make up the hidden layer. The hidden layer neurons compute their outputs and send

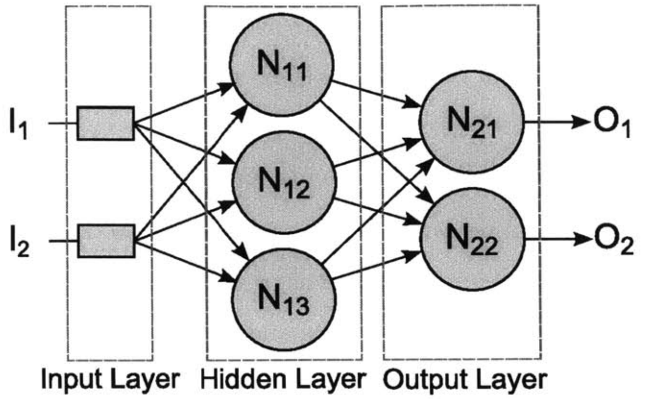


Figure 1-2: Basic Multi-layer Neural Network. Depending on the nomenclature used, this will be either a three layer or two layer neural network. The multi-layer system is fundamentally different from the single layer perceptrons in that it has one or more "hidden" layers with outputs that are hidden from the user in normal applications.

them to the next layer of neurons which make up the output layer. The output neurons sum their inputs and generate the two output signals O_1 and O_2. The multi-layer perceptron outlined above is only one type of artificial neural network, but closely matches the layout of the COIN coprocessor.

Training in a multilayer network is more complicated than training in a single layer network but follows the same basic procedure. The set of training inputs is presented to the network and the network output is computed. The difference between the desired output and the actual output is used to adjust the weights of the connections. The complication comes from the fact that the 'error' is only known at the output layer, which is not easily traced to the weights and threshold values in previous layers of the network. The 'error' is propagated backwards through the network taking into account current connection weights and threshold functions to allow updating, giving rise to the term 'back-propagation' to describe training in multilayer networks. When the activation function of neurons is non-linear the training process is a problem in non-linear optimization. Iterative gradient descent methods are useful for this type of optimization, but require differentiable activation functions.

Much of the research in artificial neural networks uses networks simulated in a traditional computer, but the simple nature of individual neurons makes it possible to implement neural networks in hardware which replicates the network structure. It is possible to build a physical implementation of a neural network using discrete or integrated electronics.

Depending on the network topology and electronics chosen to build the network, purely electronic networks may run into issues with wiring the connections between neurons, cross-talk between connections, and device speed. Transitioning to optical interconnects, and in some cases purely optical networks, is one way to sidestep some of these problems.

Optical and hybrid optical network processing have been active areas of research since at least the 1980's [11]. Optical implementations of neural networks have included Hopfield model networks [11], an optical perceptron [12], and second-order networks using quadratic interconnects [13]. These networks were often built using standard optical elements, relied heavily on optical processing of signals, and as a consequence were quite large, offered limited processing capability, and were difficult to scale [3]. More recent attempts have shifted from optical processing to electronic processing for threshold, summing, and weight implementation, while keeping optical interconnect elements [3]. By exploiting the advantages of both electrical processing and optical interconnects it is hoped that networks of much smaller size and much higher computational power will be possible.

1.2 COIN System Overview and Previous Work

The current architecture for the COIN coprocessor is shown in Figure 1-3. The proposed COIN system is a multi-layer, feed-forward, artificial neural network with diameter limited neurons using optical interconnects for connections between planes. The current connection scheme uses nearest neighbor connections, but connections beyond nearest neighbors are being considered. The system is designed to have multiple physical layers. Five layers were considered in the work of Travis Simpkins [14]. The controller allows the output of the last layer to be cycled to the input of the first layer for problems which require more layers than physically present in the system. The controller also sets the layer weights and thresholds, allowing the system to be reconfigured to implement different neural networks for different tasks.

The system is designed to take advantage of the strengths of both optical and electronic signal processing. In particular, optical signals are easily summed with a photodetector and allow for high interconnect densities. By providing the output with a source for each interconnect the connection weighting can be easily adjusted by changing the driving level of the source. Finally, using electronic circuits to control weighting and thresholding allows

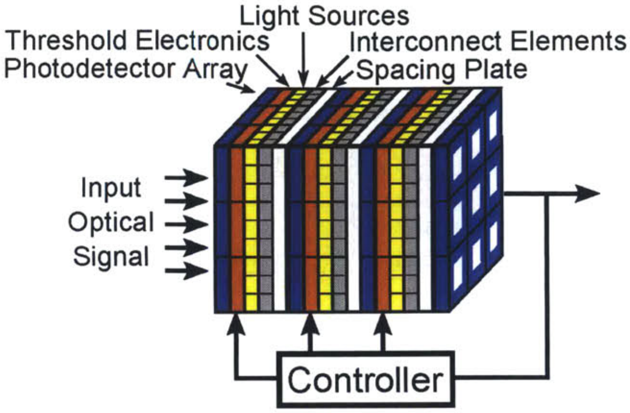


Figure 1-3: Diagram of the COIN system. Three full layers are shown, but the structure can be repeated to make a more powerful network.

for a much more compact system than all optical approaches.

A block diagram of the basic COIN system neuron is shown in Figure 1-4. Each neuron has a photo-detector for input, a silicon circuit to implement the neuron transfer function, and an array of sources with drivers to connect the neuron to other neurons in the next plane. Each source within a specific neuron connects to a different neuron in the next plane through some optical element. Connection weights can be implemented by controlling the drive level of the optical sources. Since the connections are weighted at the neuron outputs, the weighted sum needed at the input of the neuron is just the total light intensity at the neuron input. This allows the sum at the neuron input to be computed with a simple photo-detector. The neuron block diagram drawn assumes that weights are all positive, but inhibitory connections with negative weights are desirable. Implementing negative connection weights will be discussed in the next chapter.

Most of the theoretical Neural Network research in the Photonic Systems Group was done by Ben Ruedlinger [15] and Travis Simpkins [14]. Ruedlinger's initial work involved VCSEL arrays obtained from the US-Japan Joint Optoelectronics Project. He characterized the VCSELs performance and beam shape and used this information to refine the holographic interconnect design. Ruedlinger also developed the first photodetectors and amplifiers to be fabricated for the project. As a final contribution to the project, he developed a MATLAB model of the proposed network and demonstrated that it was capable of performing simple letter recognition.

Travis Simpkins's work on the COIN system was focused on the development of the

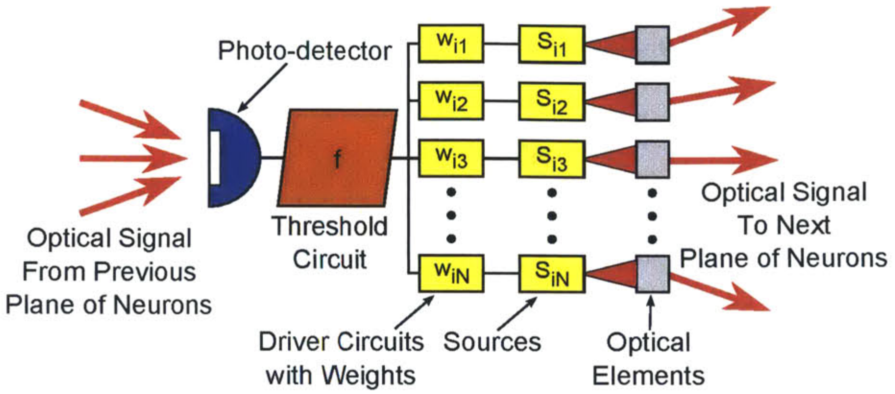


Figure 1-4: Block diagram of the COIN neuron. The components are color coded to match the system diagram in Figure 1-3. The signal at the input is light from the previous plane of neurons, and the signal at the output is light going to the next plane of neurons.

neuron and synapse circuits necessary to implement the summing and threshold functions in the network [14]. The circuits Simpkins designed were fabricated and tested. As part of his thesis, he developed models for the neuron behavior based on the measured behavior of the fabricated circuits. Simpkins integrated these models with MATLAB neural network models to investigate theoretical system performance for different network architectures and the impact of neuron failure in the proposed network. He used his model to show the network was capable of performing simple human face recognition.

Between Ruedlinger and Simpkins, the first-generation electronics for the COIN coprocessor have been designed and tested. They also implemented simulations of the basic network architecture to show the proposed network would be capable of some information processing tasks. The last element needed to demonstrate a prototype co-processor is an appropriate system of interconnects.

1.3 Optical Interconnects for the COIN coprocessor

The research in this dissertation is aimed towards creating the interconnects needed for the final version of the COIN coprocessor, but the techniques developed for this thesis will be applicable to any system needing similar optical interconnect elements. Simpkins's work on the COIN system has treated these optical interconnects as "wires" with perfect connections between neuron outputs and neuron inputs. Past work in the Photonic Systems Group on optical interconnection elements was focused on thick holographic optical elements, and this work will be discussed in Chapter 3. There are a number of possible interconnect types and this thesis will expand on the previous work and examine two types of optical interconnects: those using holographic optical elements and those using zone plate based optical elements.

The optical elements designed for this work are significantly smaller than those previous fabricated in the Photonic Systems Group resulting in new complications and design challenges. Some of these design challenges include edge effects due to the small apertures in use, alignment and packing density for holographic interconnects, and general fabrication and alignment issues due to the sub millimeter size of the new optical interconnects. The design challenges related to specific elements are discussed in chapter associated with the element type.

The material in this thesis can be divided into three areas: COIN development, Holographic Optical Element design and experiment, and Zone Plate element design. Chapter 2 will present a summary of the current COIN design, as developed for this thesis, with an analysis of how the COIN design choices impact which optical elements could be considered suitable for the network. Chapters 3 and 4 present the research work aimed at the use of Holographic Optical Elements in the COIN system. Chapter 3 covers the theoretical results, including an analysis of the problems caused when attempting to record a hologram at one wavelength use it at another wavelength. Chapter 4 is focused on some experimental work done with Holographic Optical Elements including the development of a holographic bead lens and a technique for holographic lift-off which could allow exceptionally thin holographic optical elements. Holographic optical elements have some problems, and may not be suitable at the large array sizes envisioned for the COIN coprocessor, so an effort was made to identify some alternative optical elements that might be better suited for the COIN system. Chapter 5 presents one such element, a phase-modulating zone plate, that can be fabricated using common lithographic techniques. Chapter 6 presents an analysis of a potential COIN system based on the design choices outlined in Chapter 2 and the optical elements developed for this thesis. The analysis in Chapter 6 is aimed at developing a comparison between the anticipated COIN system and alternative implementations of the neural network architecture, and shows that the COIN system is still a competitive option. Chapter 7 wraps up this thesis with a summary of the results and suggestions for future work on the COIN coprocessor project in general and on the optical elements in particular.

Chapter 2

COIN System Design

The goal of the COIN coprocessor project is to create a useful neural network based co-processor, competitive with other co-processor options, for applications which can take advantage of neural network algorithms. An ideal COIN system would be fast and computationally powerful, capable of implementing enough neurons for a wide variety of network tasks. Additionally, the system should be physically compact and draw as little electrical power as possible. Design choices including network size, light source type and wavelength of operation, and interconnect scheme and geometry are influenced by these design goals. In turn, the design choices for these aspects of the COIN system will impact the design of the COIN optical elements and even determine which elements might be suitable for use in the system. This chapter will review the COIN coprocessor system design and how it impacts the optical element selection process. The chapter will start by examining the photosensor choice and the resulting optical power budget.

2.1 Photosensor Choice and Modeling

A key aspect of the COIN coprocessor system is the use of photosensitive elements, integrated with the logic circuits, to optically sum the inputs to a given neuron. To simplify the COIN construction process and create a more rugged system it has been assumed that the photosensor elements would be fabricated on the same silicon substrate as the neuron thresholding and output driving electronics. In silicon it is possible to fabricate PN, PIN, and avalanche photodiodes along with phototransistors. All of these sensor elements rely on junctions between differently doped regions in the silicon substrate and detect light through electron-hole pair generation caused by photon absorption. A practical implication of this is that the silicon detector elements will have a long wavelength response limit related to the bandgap of silicon and a short wavelength limit due to high absorption of photons by silicon at energies much higher than the band gap, allowing an approximate wavelength range of operation from 400nm to 1000nm [1, p. 437].

Avalanche photodiodes offer the best sensitivity of the photosensor elements, but require reverse bias voltages in the 30-200V range to operate [2, p.331]. Supplying the high reverse biases needed would complicate the system design, so either photodiodes or phototransistors will be used as the optical sensor elements for the COIN neurons. The photodiode options are easier to model, so this chapter will assume the COIN co-processor will use photodiodes as the sensor elements. However, when the integrated photo sensor and thresholding electronics are designed it may be possible to simplify the sensor-to-threshold interface by using phototransistors rather than photodiodes and this design choice should be revisited.

A detailed discussion of the physics behind semiconductor photodiode operation is outside the topic of this thesis, but a basic outline is appropriate. The electrical properties of a semiconductor can be changed by doping, a process that replaces some of the semiconductor atoms with dopant atoms that have a different number of valence electrons than the semiconductor. If the dopant atom has fewer valence electrons than the semiconductor atom it replaces then doping creates a P-type region which is electrically neutral but behaves as if it contained positive charge carriers available to conduct current. If the dopant atom has more valence electrons than the semiconductor atom it replaces then doping creates an N-type region which is also electrically neutral but behaves as if it contained some excess electrons available to conduct current. A PN junction is formed by creating a P-doped region in contact with an N-doped region. Carriers in each doped region will diffuse to the adjacent region where they combine with the opposite charge carriers creating a depletion region with no free charge carriers. The diffusion of charge carriers causes a local charge build up resulting in an electrical potential across the depletion region. This process continues until the forces which drive the diffusion of the charge carriers exactly balance the force due to the electrical potential across the region. The sketch of the band diagram for a PN junction is shown in Figure 2-1.

When a photon is absorbed by an atom in the junction it will create an electron-hole pair. If the photon is absorbed in the depletion region the electron and hole created will

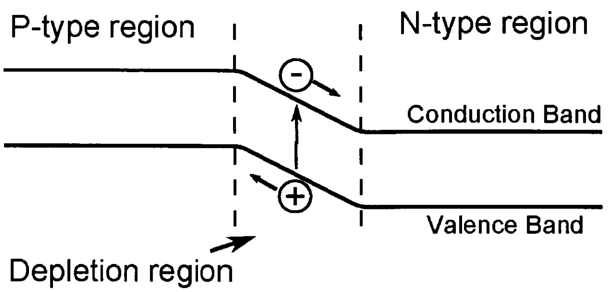


Figure 2-1: A band diagram for a PN photodiode. Absorbtion of a photon in the depletion region generates and electron hole pair which then drift across the junction under the influence of the field across the depletion region. Although each carrier has a charge q, the net current is equivalent to one electron crossing the entire junction since the individual carriers do not cross the full junction width [1, p. 436].

drift across the depletion region under the influence of the junction potential (and any applied potential) and result in a charge flow of q through an external load [1, p.436]. If the photon is absorbed outside the depletion region, the created minority charge carrier must first diffuse to the depletion region before drifting across the depletion region. Outside the depletion region there are a number of charge carriers available for recombination, so photons absorbed outside the depletion region do not always result in charge flow through an external load. Ideally, all of the photons incident on the junction will be absorbed in the depletion region so they all result in charge flow. A PIN junction is created by separating the P and N regions of the junction with an intrinsic region of undoped semiconductor. In the PIN structure most of the potential drop across the device, due to either the junction potential or an applied field, will occur across the intrinsic layer. Making the intrinsic layer thick enough to ensure it absorbs most of the photons incident on the device will ensure most photon absorption events result in a measurable charge flow.

Photodiodes can be operated as either photoconductive or photovoltaic elements. In photovoltaic mode the photodiode produces a voltage across its terminals that is proportional to the power of the light shining on the junction. In photoconductive mode a reverse bias is applied across the diode and there is a current flow through the diode proportional to the power of the light shining on the junction. Photovoltaic operation has the advantage of relatively low noise but is slower in response than photoconductive operation. Generally photovoltaic operation is limited to signals with bandwidths less than 10kHz [2, p. 334]. Photoconductive mode has the advantage of higher bandwidths, but because the diodes have a non-zero current under reverse bias photoconductive operation will have higher levels of noise [2, p. 334]. Since one of the design goals of the COIN system is to create a fast network the photodiodes will be operated in photoconductive mode, and the signal levels will be designed to accommodate the noise levels associated with photoconductive operation.

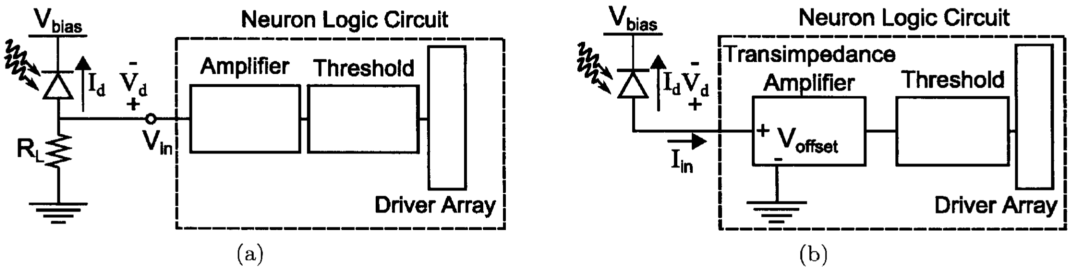


Figure 2-2: Two ways to interface a photodiode to the neurons logic circuit. In (a) the input to the neuron is a voltage caused by the diode photocurrent flowing through a load resistor, $V_{in} = -I_d R_L$. In (b) the input to the neuron is the photocurrent, $I_{in} = -I_d$.

Two methods of interfacing the photodiode to a neuron logic circuit are shown in Figure 2-2. In 2-2a, the first stage of the neuron logic circuit is a voltage amplifier and the input to the neuron is the voltage generated by photocurrent flowing through the load resistor, R_L. In this case the diode voltage is given by $V_d = V_{in} - V_{bias}$ and because of its orientation the diode is reverse biased when $V_{in} < V_{bias}$. With this circuit the voltage across the diode changes as the current through the diode changes. The second option, illustrated in 2-2b, uses a transimpedance amplifier as the first stage of the neuron logic circuit and takes the photocurrent directly as the neurons input. In this case the diode voltage is given by $V_d = V_{offset} - V_{bias}$ and the diode will be reverse biased when $V_{offset} < V_{bias}$. An advantage to using a transimpedance amplifier at the neuron input is that the offset voltage can be made constant, or nearly so, for a wide range of photocurrents keeping the diode at a constant reverse bias. However, this thesis will assume the neuron logic circuit operates as in 2-2a to develop a noise analysis and power budget without a detailed treatment of the neuron logic circuits. Additionally, this thesis will assume the current flow into the neuron logic circuit is negligible so that the voltage, V_{in}, can be computed by analyzing just the diode and load resistor portion of the circuit.

The electrical behavior of a photodiode is similar to that of a traditional junction diode with the addition of a photocurrent due to the current generated when the diode is illumi-

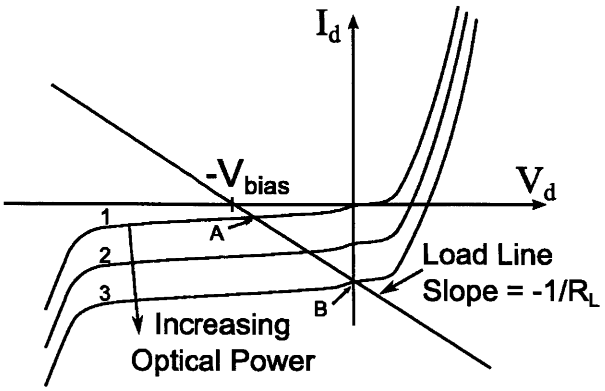


Figure 2-3: A load line analysis for the circuit in Figure 2-2a. At zero applied optical power the IV curve of the photodiode (line 1) is the same as a regular diode. As the optical power incident on the junction increases the IV curve shifts down, and the bias across the diode decreases.

nated. From Equation 7.60 in [16, p.169], the current-voltage relationship for a photodiode can be modeled as

$$I_d = I_o\{e^{qV_d/kT} - 1\} - I_L$$

The first term in the equation is the exponential model for a normal junction diode, and the second term is a shift in the I-V curve due to the photocurrent generated by illuminating the junction. Following the standard diode current and voltage notation the photocurrent has a negative value and shifts the I-V curve down as illustrated in Figure 2-3. The photocurrent, I_L, is computed as product of the power incident on the junction, P_{opt}, and the responsivity of the diode, $\mathbb{R}$. The responsivity of a commercial diode will generally be specified by the manufacturer for some range of input wavelengths and diode bias values. This parameter will have to be measured for the diodes fabricated for the COIN. From the circuit diagram the diode voltage will be $V_d = V_{in} - V_{bias} = -(I_dR_L + V_{bias})$ assuming no current flows into the amplifier. Thus the current, voltage, and optical power relation for the diode resistor pair can be written as

$$I_d = I_o\{e^{-q(I_dR_l+V_{bias})/kT} - 1\} - P_{opt}\mathbb{R} \tag{2.1}$$

Equation 2.1 can be solved numerically, but the graphical solution in Figure 2-3 is sufficient to illustrate the behavior of the circuit. For a given optical power incident on the diode, the operating point is indicated by the diode current and voltage at the intersection of the load line and the diode I-V curve associated with that optical power. When there is no optical power incident on the diode there is a small reverse current through the diode and the diode reverse bias is slightly smaller than $-V_{bias}$ as indicated by point A in the diagram. As the illumination power increases the photo-current also increases and the reverse bias decreases. For some optical power the reverse bias across the diode voltage will reach 0V, indicated by point B in the diagram. To keep the diode reverse biased at some minimum bias $V_{d,min}$, the optical power incident on the diode will have to be limited. Recognizing that $V_{in} = V_d + V_{bias}$, then the current associated with this maximum power will be

$$I = \frac{V_{d,min} + V_{bias}}{R_L} = -I_{d,max}$$

The maximum operating power can then be calculated from Equation 2.1 as

$$P_{max} = \frac{I_\circ}{\mathbb{R}}\{e^{qV_{d,min}/kT} - 1\} - \frac{-I_{d,max}}{\mathbb{R}} = \frac{I_\circ}{\mathbb{R}}\{e^{qV_{d,min}/kT} - 1\} + \frac{V_{d,min} + V_{bias}}{R_L\mathbb{R}} \qquad (2.2)$$

For P_{max} above to be real the minimum reverse bias should be set in the range $-V_{bias} \leq V_{d,min} \leq 0$. To determine the lowest detectable signel levels for the system the noise behavior of the photodiode circuit must be analyzed.

2.1.1 Noise and the Minimum Signal Level

At this time reverse biased PIN photodiodes are the best choice for photosensitive elements in the COIN co-processor system. The next step in the system design is to form an estimate for the optical source operating point based on the minimum detectable power levels and an optical power budget for the inter-neuron connections. The minimum detectable power levels are set by the signal to noise ratio at the photodiodes. Noise at the photodiodes can be the result of physical processes or the result of crosstalk between connections in the system. This section will only address the noise due to device physics. The exact properties of the COIN photodiodes will only be known after they have been designed and fabricated so to get an estimate for the optical power budget an example photodiode will be used. For this section the COIN photodiodes will be assumed to have the same properties as the Hamamatsu S9055-01 silicon PIN photodiode. This diode was chosen because its active area, a circle 100µm in diameter, is close to the expected active area of the COIN photodiodes. The relevant properties of this diode are listed in Table 2.1. The noise analysis in this section follows the approach presented in section 11.7 of [1].

Capacitance	0.5 pF
3dB Bandwidth (with a 25 Ω load)	2 GHz
Active Area	0.1mm diameter
Dark Current	1 pA
Wavelength for Peak Sensitivity	700 nm
Photosensitivity between 600nm and 800nm	> 0.32 A/W

Table 2.1: Hamamatsu S9055-01 PIN photodiode specifications from the manufacturer's specification sheet [5]. Capacitance, 3dB Bandwidth, and dark current are specified for 2 V reverse bias operation. The peak in detector response is 0.4 A/W and occurs at a wavelength of 700nm.

A schematic model of the photodiode circuit is shown in Figure 2-4. The circuit elements associated with the photodiode properties are enclosed in the dashed box. R_d and C_d are the resistance and capacitance across the junction of the photodiode, and R_s is the series resistance of the diode. Well designed diodes for visible light operation will have a junction resistance high enough to be neglected [17, p.1471], and the bandwidth of the circuit will be determined by the diode capacitance, diode series resistance, and load resistance R_L. The bandwidth of 2GHz was specified for a 2V reverse bias and 25Ω load. Equation 11.7-1 from [1] can be used to retrieve the series resistance for the diode at the specified 2V reverse bias.

$$R_s = \frac{1}{\omega_m C_d} - R_L = \frac{1}{2\pi \times (2 \times 10^9)(0.5 \times 10^{-12})} - 25 = 134\Omega$$

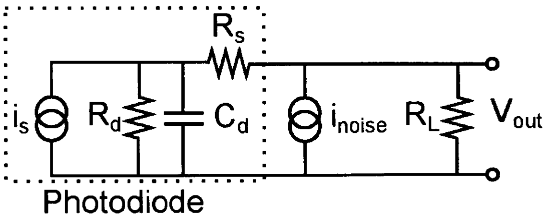


Figure 2-4: Equivalent noise circuit for the photodiode.

As stated earlier, the Hamamatsu photodiode has an active area with a size similar to the expected active area of the COIN photodiodes. Since the junction capacitance of a photodiode is proportional to its area, the COIN photodiodes should have a small junction capacitance just like the Hamamatsu photodiode. The small junction capacitance allows the photodiodes to operate at GHz frequencies, so the COIN coprocessor could be made to operate with a GHz layer clock if the bottlenecks in getting data into and out of the network can be overcome. From the analysis of competing network implementations in Chapter 6 it is clear that the COIN coprocessor does not need to achieve a gigahertz layer clock to remain competitive, a clock in the 50 MHz range should be sufficient. Given the small junction capacitance, this allows an increase in the load resistance while still maintaining an adequate system bandwidth. The increased load resistance will increase the signal level at the input to the neuron logic circuit and decrease Johnson noise which will be discussed shortly. To achieve a 500MHz 3dB bandwidth, the load resistance could be set to

$$R_L = \frac{1}{\omega_m C_d} - R_e = \frac{1}{2\pi \times (500 \times 10^6)(0.5 \times 10^{-12})} - 134 = 500\Omega \qquad (2.3)$$

Before the noise can be evaluated an expected input signal should be specified. For this analysis, following the method outlined in Yariv [1], the signal will be assumed to be sinusoidal with a modulation frequency of 50MHz so that the modulation frequency is much smaller than the system bandwidth and so that current through the junction capacitance can be neglected from the analysis. This would be consistent with a desired bit rate of 50×10^6 bits/s [1, p.451], and can be taken as the frequency of operation of the COIN system. The expected current from the photodiode is then given by equation 11.7-11 from [1, p.443], with the diode responsivity $\mathbb{R}$ substituted for $e\eta/h\nu$ and modulation index set to $m = 1$, as

$$i_s(t) = 1.5P\mathbb{R} + 2P\mathbb{R}e^{i\omega_m t}$$

This current will have a mean-square value of $\overline{i_S^2} = 2(P\mathbb{R})^2$. There are two types of noise associated with reverse biased operation of PIN photodiodes and each noise source can be represented by an equivalent noise current. The first source of noise will be shot noise due to the random generation of mobile charge carriers inside the diode. Shot noise will be proportional to the average current flowing through the load resistance. For the oscillating signal presented above there are three sources contributing to the average current: (1) the dark current, i_{dark}, present when the diode is reverse biased regardless of illumination, (2) the signal current due to the optical power associated with the signal, and (3) a DC current caused by the background optical power. For operation with a frequency bandwidth of $\Delta\nu$ the mean squared current amplitude of the shot noise is given by equation 11.7-13 from [1],

$$\bar{i}^2_{N1} = 3e(P_{opt} + P_B)\mathbb{R}\Delta\nu + 2ei_{dark}\Delta\nu \tag{2.4}$$

The second source of noise due to device physics is Johnson noise, thermal noise which causes random fluctuations in the voltage across the photodiode load. From [1] equation 11.7-14 the Johnson noise can be expressed as

$$\bar{i}^2_{N2} = \frac{4k_BT_e\Delta\nu}{R_L} \tag{2.5}$$

The term k_B is the Boltzmann constant and T_e is the equivalent temperature of operation. For this analysis the temperature of operation will be assumed to be 325K, just over 50°C, the physical temperature of the system. The temperature of operation could be set higher than the physical operating temperature to take into account the noise behavior of the amplifier following the load, but that will skipped for this section as the noise behavior of the amplifier at the neuron input is currently unknown. The noise currents is 2.4 and 2.5 can be combined into an equivalent mean squared noise current to set the noise level in the system.

$$\bar{i}^2_{eq} = \bar{i}^2_{N1} + \bar{i}^2_{N2}$$

The signal to noise ratio (SNR) for the system can be expressed with the mean squared expected signal current and equivalent mean squared noise current as

$$SNR = \frac{\bar{i}^2_s}{\bar{i}^2_{eq}} = \frac{2(P_{opt}\mathbb{R})^2}{3e(P_{opt} + P_B)\mathbb{R}\Delta\nu + 2ei_{dark}\Delta\nu + 4kT_e\Delta\nu/R_L} \tag{2.6}$$

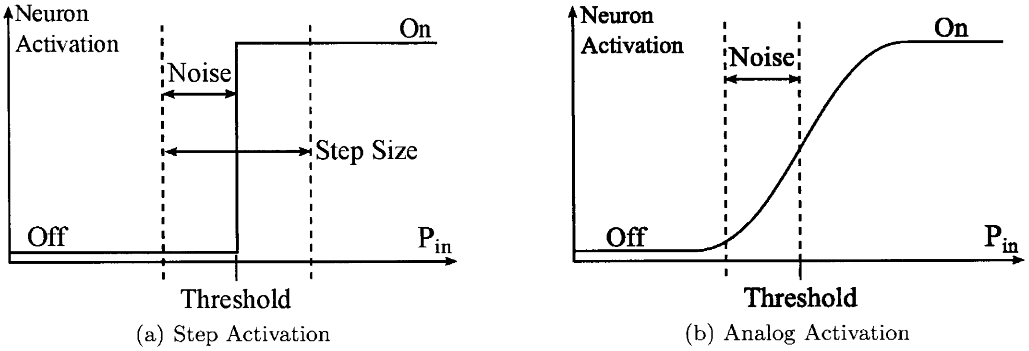
The minimum detectable power is the power at a given operating point for which the SNR is 1. From Equation 2.6 it is clear that this power level will depend on the load resistance, operating temperature, and background optical power. The load resistance was found to be 500Ω using Equation 2.3, and the operating temperature is assumed to be 325K. Background optical power could be due to light leaks in the system, or it could be due to additional inputs to a given neuron. Using the values from Table 2.1 and assuming an operating frequency of 50MHz the optical power corresponding to an SNR of 1 is 0.10μW when the background optical power is set to 0W. However, if the neuron is operated right at the optical power limit of 3.125mW established by Equation 2.2 the optical power corresponding to an SNR of 1 is 0.30µW. These optical powers are computed at the photodiode surface. If the optical interconnect elements are η% efficient in directing light from a source to these photodiodes, then the minimum power level at the source is

$$P_{min,source} = P_{min} \times 100/\eta$$

At this point a range of operating power levels can be set for the COIN sources. As a worst case estimate, assume the optical elements are 50% efficient in directing light from a source to a detector. This is between the currently expected binary zone plate efficiency (30%) and the expected multi-step zone plate efficiency (70% to 90%, depending on step size) discussed in Chapter 5. The optical power at the photodiodes should not exceed 3.125mW to keep the diode properly reverse biased. If the system is designed so that this power level is reached when all nine nearest neighbor inputs to one photodiode are fully on, then each optical source will have to supply a maximum optical power of 0.70mW. If the minimum signal level for an input is set so that the SNR would be one at the photodiode with no other inputs activated, and assuming the same 50% optical element efficiency, then the minimum signal level is 0.20µW. These limits establish a maximum range of operation, and the next section will discuss how to set the signal levels within this range.

2.1.2 Designing the Signal Levels and Activation Function

The previous sections established both the maximum power level at a neuron input which keeps the photodiode in the desired reverse bias range and the minimum power level at the photodiode which can be distinguished from noise in the system. These limits establish the range of possible optical signals which could be used in the COIN coprocessor, but do not specify how that range should be used. To create a flexible network the neuron hardware should be designed to allow the connection drive levels to be modified by uploading new drive weight values to the neuron. Previous work on the COIN project has assumed a binary weight register and the use of a digital to analog converter in the drive electronics to allow fine control of the drive level across the entire range of possible optical signals. The minimum increment in output level is set by the least significant bit of the weight, and choosing a drive level to associate with this weight requires consideration of the neuron transfer function. The two example neuron activation functions in Figure 2-5 illustrate two approaches to neuron activation. Figure 2-5a shows a binary neuron activation, where the neuron output is either completely on or completely off. Figure 2-5b shows an analog neuron activation where the output smoothly changes from fully off to fully on as the input signal increases.



(a) Step Activation

(b) Analog Activation

Figure 2-5: Neuron Activation functions and noise. The step activation in (a) has only two states: off and on. For the analog activation in (b), the neuron output slowly changes from "Off" to "On" and has a region of partial activation between the two states.

The step response is the easier of the two to analyze. A single connection, with high enough weight, should be able to flip the state from a zero output to a one output. The operating levels of the source should be chosen so that if the power level is one increment below the threshold then the expected system noise should not cause the state to flip from zero to one, and when the power level is one increment above the threshold the expected system noise should not cause the state to flip from one to zero. In each case the distance from the signal level to the threshold should be equal to the expected noise, so one step in operating level should change the signal power by twice the expected noise.

An alternative to the step response activation would be a smoothly changing analog activation function as shown in Figure 2-5b. Rather than switch from fully off to fully on an input change of one minimum step, the neuron gradually turns on as the input increases over several steps. In this case the impact of the noise on the activation function will depend on the slope of the activation function around the operating point. When the slope of the activation function is small the impact of noise on the output state will be small, but the transition from fully off to fully on will require a larger change in the optical power at the input of the neuron.

Of the two activation functions the step activation has the more stringent requirement on step size so it will be used to estimate the signal levels. As seen in the previous section, the expected noise will depend on the state of all of the optical inputs to the neuron. To ensure the step size condition is meet regardless of the state the other inputs the noise level should be computed with all nine input connections fully on. Assuming an N bit binary weight scheme, nine inputs from the nearest neighbors as discussed in Section 2.4, and with a step size of twice the maximum expected noise then $P_{max} = 9 \times 2^N \times P_{step}$ with P_{step} equal to the optical power corresponding to an SNR of 2 when the total input power is P_{max}. Solving for the step size is an iterative process. Using the parameters from the previous section, and an 8 bit weight scheme, P_{step} at the photodiode is 0.21μW for operation at 50MHz. Assuming a worst case of 50% efficient optical elements this translates to a power range of 0.42μW to 108μW at the source.

Choosing the step sized based on a signal to noise ratio of two at full power may be a little optimistic since the noise behavior of the electronics in the neuron logic circuit has been neglected. If the step is instead set to give a SNR of four at the photodiode, then for 50% efficient optical elements the power range at the source is 0.7μW to 180μW for 50MHz operation. The maximum power with this scheme is still smaller than the maximum single neuron power computed earlier, so the step size could be increased from this value at the expense of increasing the systems total power consumption.

2.2 Light Source Choices

The COIN network will feature a large number of neurons, and a large number of connections. As a result the network will need very large arrays of light sources and driver circuits. An ideal light source for the COIN co-processor would have the following physical and electrical properties:

- the light source and driver circuit should be physically compact so a large number of neurons can be packed into a small package

- the light source and driver circuit should be energy efficient to limit waste heat generation and allow low power operation of the network

- the driver circuits should be flexible to allow changing connection weights electronically
- the driver circuits should be capable of maintaining stable output power levels

Of more importance to the work in this thesis are the optical properties of the potential light sources. An ideal source, from the perspective of the optical elements, would have a narrow spectrum, moderate coherence length, and a beam shape which is smooth and moderately divergent. The spectral width is important for optical elements which rely on interference, such as diffraction gratings or holograms, whose performance and behavior is highly wavelength dependent. Similarly the coherence length of the optical sources is important for holographic optical elements which rely on interference between multiple reflections within the hologram. The shape and divergence of the source beam set the distance between the source layer and optical elements, and highly divergent sources will be more difficult to focus and steer than sources which are more slowly diverging (see Chapter 5 for more information). Finally, based on the results of the previous sections sources should be chosen that can provide power levels of 0.7 to 180 µW per connection somewhere in the wavelength range of 400 to 1000 nm. Sources considered for the COIN system have included laser diodes, semiconductor light emitting diodes, and organic light emitting diodes. The advantages and disadvantages of the various light source options are discussed below.

2.2.1 Laser diode light sources

If the optical properties of the light sources were the only consideration then an ideal source would be a laser operating with a single longitudinal and single transverse mode. The narrow spectrum and well defined beam behavior makes it easy to design optical elements for these sources. The individual drivers and lasers can be made quite compact, and Vertical Cavity Surface Emitting Lasers (VCSELs) have been fabricated in large two dimensional arrays which initially appear ideal for use in the COIN system. While the optical properties and availability of large two dimensional arrays are quite attractive, there are two potential problems with the use of lasers as sources in the COIN co-processor system: integration of the lasers with the silicon detector-threshold-driver wafers and the low efficiency of lasers when operated below to just above threshold as illustrated in Figure 2-6.

For all practical purposes at this time there are no laser structures suitable for use in the COIN system that can be directly fabricated on the silicon wafers used for the detector, threshold, and driver electronics. At this time there are two potential approaches to integrating lasers with the silicon wafers. The first is flip-chip bonding of the silicon drive electronic wafers with GaAs VCSEL array wafers. In this technique the VCSEL laser wafer and silicon wafer are processed separately and then bonded together. The downside of this technique is that the wafers used in fabricating the VCSEL lasers are much more expensive than the silicon wafers used for the rest of the electronics. The second approach is the OptoPill laser-silicon integration technique developed in Professor Fonstad's research group at MIT [18]. In this technique the VCSEL lasers are fabricated and then separated from their substrates. The silicon wafers are fabricated with holes into which the VCSEL structures can be deposited. In theory this technique should allow more efficient use of the expensive VCSEL wafers, but currently the yield is low as some VCSEL assemblies do not work after being deposited and soldered into the silicon wafers.

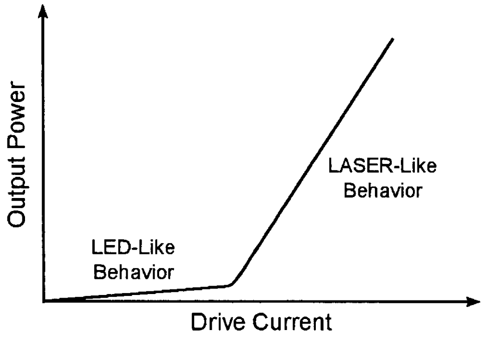


Figure 2-6: Laser output power versus drive current. Note the two regions of operation, LED like behavior at low currents and LASER behavior at higher currents. [2, p.133]

If the fabrication and integration problems could be solved there is still an issue of the low efficiency of lasers at low power levels. A typical light-current curve might look like something like the curve shown in Figure 2-6. Diode lasers operate by injecting a large enough current into a PN junction so that there are a large number of electrons and holes in the active region to support stimulated emission of photons. At lower currents there are not enough carriers available and spontaneous emission is more likely, making the output light behave more like an LED than a laser. The slope efficiency of the laser is the ratio of the light output to current input. In general semiconductor lasers can have fairly high slope efficiencies above threshold, but have poor efficiency below threshold. A detailed discussion of semiconductor lasers can be found in Chapter 5 of the book "Lasers and Optoelectronics" [2].

2.2.2 Semiconductor LEDs

Semiconductor Light Emitting Diodes (LEDs) are a second candidate for sources in the COIN co-processor system. Similar to a diode laser, light is generated in LEDs through the recombination of electrons and holes in a diode junction. Traditional LED structures have a few disadvantages when compared to diode lasers for use in the COIN-coprocessor. In particular, they will have a wider spectral bandwidth and produce a highly diverging beam. Resonant Cavity LEDs (RCLEDs) are created by placing the active region of the LED inside an optical cavity. The optical cavity is designed to resonate around the wavelength of the expected LED emission, and the cavity enhances emission at these wavelengths. RCLEDs will have narrower spectra than traditional LED structures and will have a more directional emission spectra [19, p.255]. However, the spectra of RCLEDs will be wider than that of a diode laser. The advantage of the RCLED is that at the low output powers where VCSEL lasers would also be operating as LEDs the RCLEDs are more efficient [19, p.259] [20]. Diode lasers can have high slope efficiencies beyond their laser threshold, so the choice of LED or Diode laser will depend on the operating point. A comparison is shown in Figure 2-7. Assuming both devices have the same forward voltage drop, if the signal levels place the operating point to the left of the intersection then the LED would be the most energy efficient choice. If the signal levels place the operating point to the right of the intersection then the diode laser would be the most efficient choice.

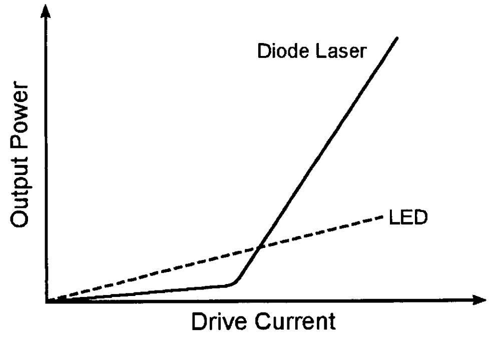


Figure 2-7: Comparison of the output power versus drive current of a diode laser (solid line) and an LED (dashed line)

While semiconductor LEDs may offer a significant improvement on electrical efficiency at low optical power levels they still have the diode laser problem of integration with the silicon wafers used for the COIN electronics for the same reason, LEDs cannot be fabricated directly in silicon. The same solutions, flip-chip bonding and OptoPill type assembly, are applicable but the downsides of higher cost and yield issues are still present. Organic LEDs are an option which could provide the benefits of LED efficiency with an improvement in integration.

2.2.3 Organic LEDs

Organic light emitting diodes (OLEDs) may be an interesting alternative to the semiconductor lasers and LEDs discussed above. Instead of using an inorganic semiconductor junction, OLEDs use and electroluminescent organic compound sandwiched between conductive layers. Like LEDs, it is possible to fabricate resonant cavity OLEDs with the same benefits in improved spectral purity and narrower emission. However, OLEDs are generally less efficient (especially at shorter wavelengths) than current semiconductor LEDs. The unique advantage of organic LEDs is that the OLED structure is based on organic compounds which can be deposited directly onto standard silicon wafers, making OLED-silicon integration much simpler than combining either semiconductor LEDs or laser diodes with the COIN silicon electronics.

2.2.4 Source Summary

Each of the source options have advantages and disadvantages. Lasers are ideal from an optical perspective, but have serious disadvantages in power dissipation and fabrication. Resonant Cavity LEDs can be extremely efficient in energy usage at low powers but suffer from the same fabrication issues as semiconductor lasers and will have wider spectra and a larger beam divergence. OLEDs offer a potential solution to the fabrication issues of the semiconductor lasers and LEDs, but are less efficient than the semiconductor LEDs. Table 2.2 summarizes the features of the various sources.

For the purposes of the optical element design in this thesis, it will be assumed that the COIN source produces a single mode Gaussian beam. This assumption is best for a small class of VCSEL lasers, the other optical sources will produce more complicated beams. When the exact properties of the final sources are known then the optical elements can and

Source Type	Power efficiency	Integration with Silicon	Spectral Width	Beam Behavior
Laser	Poor	Poor	Excellent	Excellent
LED	Excellent	Poor	Poor	Poor
RC-LED	Excellent	Poor	Acceptable	Acceptable
OLED	Moderate	Excellent	Poor	Poor
RC-OLED	Moderate	Excellent	Acceptable	Acceptable

Table 2.2: Potential COIN Light Sources

should be redesigned using the approach outlined in this thesis.

2.3 Wavelength of Operation

Detector considerations have established the wavelength of the sources should be in the 400-1000 nm range, and all three of source types can be fabricated to produce light across this range. For the optical elements considered in this thesis, the characteristic feature size of the optical elements is related to the wavelength of operation. As the wavelength decreases the feature size also decreases, making fabrication difficult. On the other side, as wavelength and feature size are increased the total size of an optical element will also increase and may make the optical elements impractically large for the COIN system. A balance between the two extremes would place the wavelength of operation in the 600-800nm range (red to near IR). The optimal wavelength of operation will depend heavily on the properties of the COIN detectors and sources, so it would be unproductive to narrow it much further. Fabrication issues related to operation across this range will be presented for the optical elements designed in this thesis.

2.4 Connection geometry and weight implementation

As discussed in section 2.1, the COIN coprocessor uses a simple photo-detector to sum the inputs to a neuron which requires that the connection weights must be applied at the driving end of the connection from the previous layer. To maintain network flexibility and to allow re-training for different tasks the weights for a given connection cannot be fixed. There are two possible solutions as illustrated in Figure 2-8: either drive each connection with its own source or use one output light source and include a spatial light modulator to change the intensity illuminating each connection. Due to potential differences in source driver circuits it is not immediately obvious that one solution is more energy efficient than the other. However, replicating source driver circuits will be simpler than producing a spatial light modulator layer to modulate the connection weights, so the COIN system will use individual sources for each connection with variable source driver circuits to implement the connection weights.

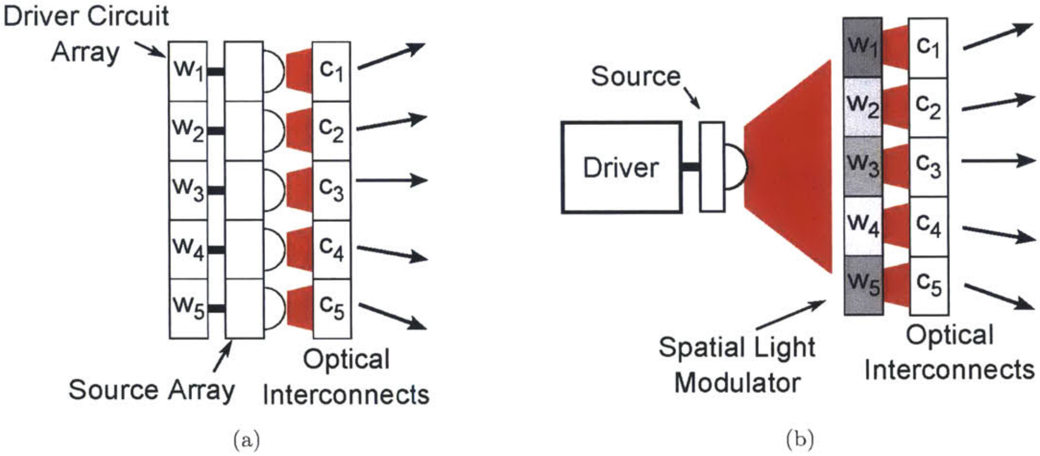


Figure 2-8: (a) Interconnects driven by individual sources and modulated by controllable driver circuits, (b) Interconnects driven by one light source per neuron and modulated by a Spatial Light Modulator

The total number of connections per neuron will depend on the interconnection scheme. While a fully connected network, in which each neuron is connected to all neurons in the next plane, should offer the highest performance it would require very large number of optical connections per neuron. Past work in the Photonic Systems Group [14] indicates that a nearest neighbor connection scheme will allow sufficient network flexibility to make the COIN coprocessor capable of some interesting neural network applications. In a nearest neighbor connection scheme each neuron in a plane will have connections with the nine neurons closest to it in the next plane. With independently driven connections this will require a 3x3 array of sources and a 3x3 array of interconnect elements at the output of each neuron.

The connection geometry also determines the off axis steering behavior of the optical elements. If maximum packing density is assumed, and the detectors are centered on the input side of the neuron, a quick sketch shows that the edge elements will steer the light from their sources off axis a distance of two times the width of a single optical element and

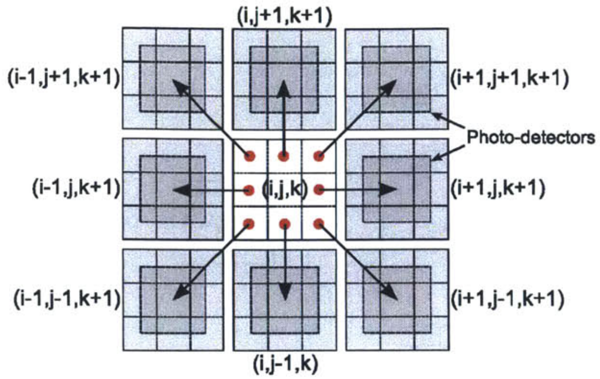


Figure 2-9: Nearest Neighbor Connection Geometry. There are only three unique connections in the nearest neighbor connection geometry. Straight ahead (not shown), Edge, and Diagonal. The four "Edge" connections are identical, but rotated in steps of 90 degrees. Similarly the four "diagonal" connections are the same but rotated in steps of 90 degrees.

that diagonal elements will steer the light from their sources off axis a distance of $2\sqrt{2}$ times the width of a single optical element. The steering angle with respect to the system axis will be set by this behavior and the separation of the optical elements of one plane and the detectors of the next plane along the system axis.

The nearest neighbor connection geometry chosen for the COIN coprocessor simplifies the design of the optical interconnects by limiting each neuron to nine connections, as discussed previously, and by requiring only three optical element types to form all of the connections. The three connection types are center, edge, and diagonal as illustrated in Figure 2-9. The interconnections for the four "Edge" elements can be described by one interconnect element rotated by 90 degree steps, and the same can be said of the "Diagonal" elements.

2.5 Inhibitory Connection Weights

Previous work in the Photonic Systems Group involving neural network training for the COIN coprocessor has assumed the availability of negative connection weights. While it is sometimes possible to train networks without negative weights, the ability to implement an inhibitory weight should result in a more flexible neural network. To create a negative weight optically would require mutually coherent optical sources and a method of modulating the phase and/or polarization of the signal light. Creating such a system would greatly increase the cost and complexity of the COIN coprocessor, so an alternative is needed. Three approaches to encoding negative weights discussed in this section.

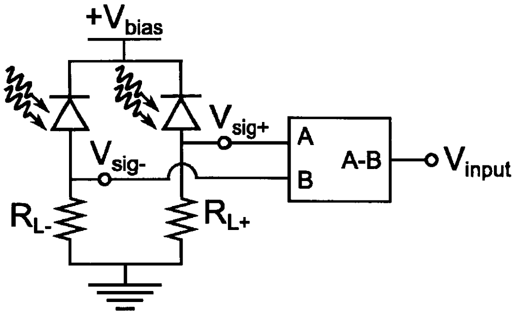


Figure 2-10: Negative weights implemented using multiple optical channels per neuron.

The most obvious modification to allow negative weights in the COIN coprocessor is to add a second photodiode to the input of each neuron. This approach is diagrammed in Figure 2-10. Optical signals representing positive inputs to a neuron are directed to one photodiode while optical signals representing negative weighted connections are directed to the second photodiode. A subtracting circuit is used to compute the difference between the two sets of inputs. To allow each connection to take positive or negative weights requires that either the optical elements connecting the neurons be steerable, or that each connection consists of two sources and two optical elements: one directed at the positive input and one directed at the negative input of the receiving neuron. Steerable optical elements are significantly more complicated than fixed elements, and doubling the number of source elements and optical elements in the COIN would increase both the network size and cost. A more efficient approach would be to use one optical channel and use some encoding scheme to indicate if a connection is positive or negative.

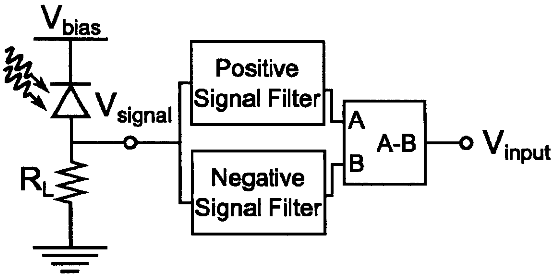


Figure 2-11: Separating positive and negative weights by drive frequency.

It is possible to use a single optical input and use some encoding scheme to separate the signals with a positive connection weight from those with a negative connection weight. This approach is diagrammed in Figure 2-11. One method of encoding the sign of the signal would be to modulate optical signals representing positive connection weights at one frequency and modulate optical signals representing negative connection weights at a second frequency. If reference signals at the two frequencies are distributed to the neurons from some set of master clocks then generating the modulated optical signals is straightforward. At the neuron input these signals could be separated using bandpass filters at the appropriate frequencies. Fabricating these filters in the area allotted to the COIN neurons may be difficult, and this approach may require an increase in the neuron area with the associated increase in network size and cost. The modulation frequencies for the weight signals in this scheme must be higher than the system operating frequency, so this approach requires an increase in the bandwidth of the optical input system compared to the approach of using two optical channels.

Another encoding approach would be to separate the positive and negative connection weight signals in time. A system clock cycle could be divided into a period during which only positive connections are active and a period during which only negative connections are active. The circuitry required to compare the inputs during the two time period may be simpler than the bandpass filters needed for frequency encoding. A simple approach might be to charge one capacitor during the positive signal period, a second capacitor during the negative signal period, and then subtract the two voltages to get the neuron input level. Time encoding, like frequency encoding, requires an increase in the optical input bandwidth over the two optical channel scheme.

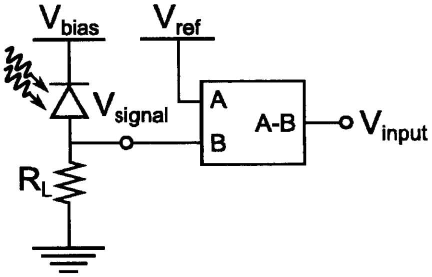


Figure 2-12: Using a bias to achieve effective negative weights.

The last approach to implementing negative weights is to bias the system so that the default source state is halfway on rather than off. In this scheme, the source is set so that when the neuron is off the source is turned on to some bias level. For connections with negative weights increasing neuron activation results in a decrease in output, and for positive connection weights the source output increases with neuron activation. At the neuron input, shown in Figure 2-12, the signal level on the optical input is compared with a reference voltage representing the expected "zero input" level. The signal voltage will be lower than the reference voltage when the sum of the negative connection inputs is higher than the sum of the positive connection inputs. This scheme requires a more complicated source drive system than the other approaches and will result in a higher power consumption, but it does not require an increase in the optical input bandwidth or duplicating the optical channels.

At this time the two best candidates for allowing negative weights in the COIN co-processor are the time encoding technique and the biasing technique. Both approaches have the advantage of not requiring duplicate optical channels, but each method requires more complicated electronics than previously designed for the COIN system. The choice of which approach to use should be made during the next round of neuron logic development, and will depend on which method is easiest to integrate into the rest of the COIN neuron logic.

Chapter 3

Design Considerations for Holographic Interconnects

Prior to this thesis the COIN coprocessor design has assumed that Holographic Optical Elements (HOEs) would be used to form the optical interconnects that that allow neurons in one plane of the network to communicate with neurons in the next plane. Previous students have developed techniques to design HOEs for the COIN coprocessor and successful fabrication of HOEs with widths on the order of 2 to 5 mm has been achieved [21] [22] [4]. The final COIN system will require optical elements with widths in the 100µm to 200µm range. Much of the early work done for this thesis was directed towards refining the hologram design and fabrication process to create holographic optical elements of the sizes needed for a COIN coprocessor. This chapter will start with a brief introduction to holographic optical elements and the previous work done in the Photonic Systems Group on using HOEs in the COIN coprocessor. The chapter will then move in to the new results developed for this thesis, starting with a review of the change of wavelength problem and then examining some of the limits encountered when scaling holograms to small sizes. The results in this chapter show that holographic interconnects may not be the best choice for the COIN coprocessor, motivating the development of some alternatives to traditional HOEs which are presented in the Chapters 4 and 5.

3.1 Introduction to Holograms and Prior Work

Holography is a technique for capturing three dimensional information about an object by recording the interference pattern between light scattered from the object and a reference beam. After recording and processing the hologram can be "read out" by illuminating the hologram with a replica of the reference beam. The hologram will convert some portion of this read out beam into a replica of the light scattered from the object. Most discussions of holography start by examining a thin hologram formed by recording the interference pattern between the two beams in one plane. Thin holograms are useful for illustrating the basic properties of holograms and provide a motivation for Holographic Optical Elements but only convert a small portion of the input read-out beam to the desired output beam. Thick holograms, truly three dimensional recordings of the interference pattern, allow for more of the read-out light to be converted into the replica object beam. Before starting a mathematical treatment of holography the basic notation used to express the fields forming the interference patters should be established.

3.1.1 Notation for Waves and Phasors

The results presented in this chapter were derived under the assumptions of classical electromagnetism. The optical power levels expected in the interconnect system are high enough that when the light sources are on there will be sufficient photons in the system so that the classical description of light as a wave is appropriate. This dissertation will use the following conventions: scalar quantities will be represented by plain letters, vector quantities by bold font, and complex quantities by a tilde over the variable. For example, the scalar electric field magnitude is E, the real electric field vector is $\mathbf{E}$ and the complex electric field phasor is $\tilde{\mathbf{E}}$. Monochromatic waves will be represented using the standard phasor notation. The real wave in space and time will be expressed using a phasor that encompasses the spatial variation of the field and a separate term describing the time dependence of the field.

$$\mathbf{E}(\mathbf{r}, t) = \mathbf{Re}\{\ \tilde{\mathbf{E}}(\mathbf{k}, \mathbf{r}) e^{-j\omega t}\} \tag{3.1}$$

The time dependent term for a monochromatic plane wave will be assumed to be $e^{-j\omega t}$, so the phasor for a plane wave traveling in the positive $\hat{k}$ direction will be $\tilde{\mathbf{E}} = \hat{e} E_o e^{j\phi} e^{j\mathbf{k}\cdot\mathbf{r}}$. The field for a plane wave traveling in the positive $\hat{k}$ direction is then given by

$$\mathbf{E}(\mathbf{r},t) = \mathbf{Re}\{\ \tilde{\mathbf{E}}(\mathbf{k},\mathbf{r})e^{-j\omega t}\} = \mathbf{Re}\{\hat{e}E_o e^{j\phi} e^{j\mathbf{k}\cdot\mathbf{r}} e^{-j\omega t}\} \tag{3.2}$$

Maxwell's equations describe the relationship between electric and magnetic fields in classical electromagnetism. For the field $\mathbf{E}(\mathbf{r},t)$ in Equation 3.2 to represent a physically possible field it must be a solution to Maxwell's equations. This will be the case when the magnitude of the vector $\mathbf{k}$ satisfies $|\mathbf{k}| = k = \omega/v_p = 2\pi n/\lambda_\circ$. The plane wave is one propagating wave solution to Maxwell's equations. Two other propagating wave solutions will be used in this thesis: the spherical wave and the Gaussian beam. The phasor representation for these two waves is given below.

Spherical Wave :

$$\tilde{\mathbf{E}}(\mathbf{r}) = \frac{A}{|\mathbf{r}-\mathbf{r}_\circ|} e^{\pm jk|\mathbf{r}-\mathbf{r}_\circ|} e^{j\phi_A} \tag{3.3}$$

Gaussian Beam :

$$\begin{aligned}\tilde{\mathbf{E}}(x,y,z) = E_o\sqrt{\frac{\omega_{ox}}{\omega_x(z)}}\sqrt{\frac{\omega_{oy}}{\omega_y(z)}}exp\Bigg(j[kz-\eta(z)] - x^2\left(\frac{1}{\omega_x^2(z)} - \frac{jk}{2R_x(z)}\right) \\ - y^2\left(\frac{1}{\omega_y^2(z)} - \frac{jk}{2R_y(z)}\right)\Bigg)\end{aligned} \tag{3.4}$$

For the work in this thesis, the spatial variation of the wave or the sum of several waves is the interesting quantity. Phasor representation provides an easy way to add the fields from multiple waves while ignoring the time evolution of the waves. When an expression for the actual electric field is needed, it can be recovered by finding the real part of the product of the phasor and $e^{-j\omega t}$ as is shown in Equation 3.1.

3.1.2 Thin Holograms

A thin hologram is a recording of the interference pattern in one plane created by two or more propagating waves. When the interference pattern is created by two distinct beams one beam is generally called the reference beam and the other is called the object beam. A basic outline of how these holograms are formed and read out is given below, a more detailed treatment can be found in reference [23]. Consider the interference pattern formed across a small region in the $z = 0$ plane by the intersection of a plane wave traveling along the $\hat{\mathbf{z}}$ axis, $\tilde{\mathbf{E}}_1 = \hat{e}_1 E_1 e^{jk_{1z}z}$, and a spherical wave converging to the point $(x,z) = (+x_\circ, +z_\circ)$,

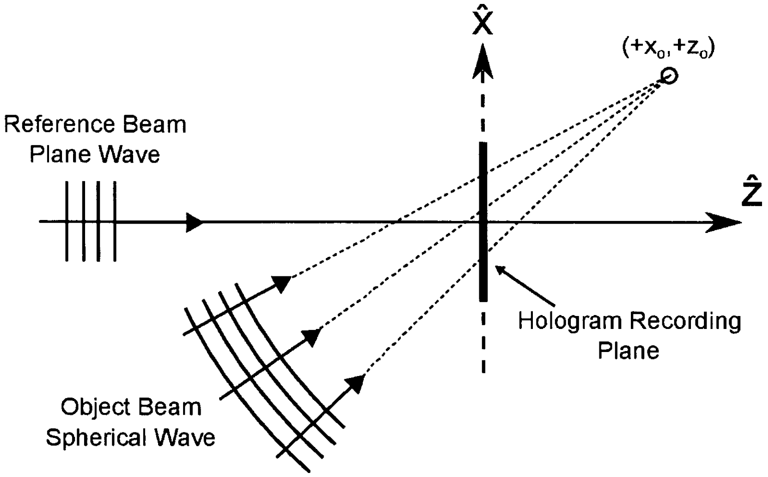


Figure 3-1: Thin Hologram Recording. The hologram in this example is formed with a plane wave reference beam and a spherical wave object beam.

$\tilde{\mathbf{E}}_2(\mathbf{r}) = \hat{e}_2 \frac{A}{|\mathbf{r}-\mathbf{r}_o|} e^{-jk|\mathbf{r}-\mathbf{r}_o|}$ as illustrated in Figure 3-1. For this discussion the plane wave will be called the reference beam and the spherical wave will be called the object beam. To simplify the problem assume that over the extent of the hologram both waves are polarized in the $\hat{\mathbf{y}}$ direction, and that the two beams are mutually coherent so their fields add rather than their intensities. The intensity of the sum of the two beams in the $z = 0$ plane is

$$\begin{aligned} < I > \propto \left| \tilde{\mathbf{E}}_1 + \tilde{\mathbf{E}}_2 \right|^2 &= (\tilde{\mathbf{E}}_1 + \tilde{\mathbf{E}}_2) \cdot (\tilde{\mathbf{E}}_1 + \tilde{\mathbf{E}}_2)^* \\ &= |E_1|^2 + \frac{|A|^2}{|\mathbf{r}-\mathbf{r}_o|^2} + \frac{E_1 A}{|\mathbf{r}-\mathbf{r}_o|} e^{+jk|\mathbf{r}-\mathbf{r}_o|} + \frac{E_1 A}{|\mathbf{r}-\mathbf{r}_o|} e^{-jk|\mathbf{r}-\mathbf{r}_o|} \end{aligned}$$

Although the expression contains complex terms the sum is a real quantity and could be recorded on a photographic plate, provided the resolution of the plate was high enough to capture the features of the pattern. If the photographic recording process is linear with respect to the intensity on the plate then the transmission of the plate can be written as

$$t(x,y,z) = \frac{E_{out}}{E_{in}} = \alpha \left(|E_1|^2 + \frac{|A|^2}{|\mathbf{r}-\mathbf{r}_o|^2} + \frac{E_1 A}{|\mathbf{r}-\mathbf{r}_o|} e^{+jk|\mathbf{r}-\mathbf{r}_o|} + \frac{E_1 A}{|\mathbf{r}-\mathbf{r}_o|} e^{-jk|\mathbf{r}-\mathbf{r}_o|} \right)$$

The transmission function represents the ratio of the field at the output side of the photographic plate to the field at the input side of the photographic plate. A developed photographic plate is a passive material, it cannot amplify the light at its input face, so the transmission function can only take values in the range $0 \leq t(x, y, z) \leq 1$. The recorded interference pattern could be "read out" by illuminating the plate with a replica of one of the initial two beams. For readout with a scaled replica of the plane wave, $\tilde{\mathbf{E}}_1 = \hat{\mathbf{y}} E_1 e^{jk_{1z}z}$, the field at the output of the plate is given by

$$\begin{aligned}
\tilde{\mathbf{E}}_{out} = C\ \tilde{\mathbf{E}}_1 t(x, y, z)|_{z=0^+} =& \hat{\mathbf{y}}\alpha C \left(|E_1|^2 + \frac{|A|^2}{|\mathbf{r} - \mathbf{r}_o|^2} \right) E_1 \\
&+ \hat{\mathbf{y}}\alpha C E_1^2 \frac{A}{|\mathbf{r} - \mathbf{r}_o|} e^{jk\sqrt{(x-x_o)^2+z_o^2}} \\
&+ \hat{\mathbf{y}}\alpha C E_1^2 \frac{A}{|\mathbf{r} - \mathbf{r}_o|} e^{-jk\sqrt{(x-x_o)^2+z_o^2}}
\end{aligned} \tag{3.5}$$

The expression for the field at the photographic plate output consists of three terms, each representing a different wave leaving the hologram. Under the approximation that the region of the hologram is small compared to the distance between the hologram and the original focus point the term $\frac{|A|^2}{|\mathbf{r}-\mathbf{r}_o|^2}$ is approximately constant across the hologram. So the first term in the field at the output is

$$\hat{\mathbf{y}}\alpha C \left(|E_1|^2 + \frac{|A|^2}{|\mathbf{r} - \mathbf{r}_o|^2} \right) E_1 \approx \hat{\mathbf{y}} C' E_1. \tag{3.6}$$

This is simply a scaled version of the illuminating field and is generally called the undiffracted beam because it represents the portion of the illuminating beam which continues propagating as if there were no hologram. The second term in Equation 3.5 is

$$\hat{\mathbf{y}}\alpha C E_1^2 \frac{A}{|\mathbf{r} - \mathbf{r}_o|} e^{jk\sqrt{(x-x_o)^2+z_o^2}}. \tag{3.7}$$

In the $z = 0$ plane this field could represent a spherical wave diverging from either $(x, z) = (x_o, z_o)$ or $(x, z) = (x_o, -z_o)$. Knowing that the resulting field should continue propagating in the $+\hat{\mathbf{z}}$ direction, it becomes clear that this is a spherical wave diverging from $(x, z) = (x_o, -z_o)$. This spherical wave is related to the conjugate of the object beam. The last term in Equation 3.5 is

$$\hat{\mathbf{y}}\alpha C E_1^2 \frac{A}{|\mathbf{r} - \mathbf{r}_o|} e^{-jk\sqrt{(x-x_o)^2+z_o^2}}. \tag{3.8}$$

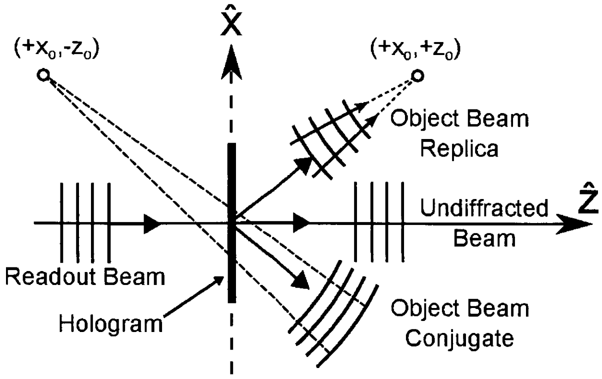


Figure 3-2: Thin hologram Readout. The hologram is read out with a plane wave, a replica of one of the beams used to make the hologram. At the output side of the hologram there are three beams: the undiffracted beam, one beam that is a replica of the object beam, and one beam related to the conjugate of the object beam.

Applying the same reasoning as used on the second term, this field represents a spherical wave converging to $(x, z) = (x_o, z_o)$. Where the second term was related to the conjugate of the object beam, this term is a scaled replica of the object beam. The readout geometry and three output beams are illustrated in Figure 3-2.

Recording the interference pattern of a plane wave and a spherical wave creates a transmission object which will take a plane wave at its input and produce a beam at its output which is steered off axis and focused to a point. The resulting transmission object is called a holographic optical element. Proper choice of the recording beams can create holographic optical elements for collimation, focusing, and steering behaviors. One downside to the use of thin holograms as holographic optical elements is apparent from the result above. Ideally an optical element would transform 100% of the incident power into the desired beam, but the two additional beams at the output mean that it is not possible to put 100% of the power into the desired beam. In the case of simple gratings the maximum efficiency of a thin hologram which modulates amplitude, as described in this section, is 6.25% and the efficiency of one which modulates phase is 33.9% [23, p.61]. The diffraction efficiency of thin holograms is low, but for the COIN coprocessor a larger problem is the presence of the two extra beams.

Consider the case of a hologram which is supposed to connect neuron (i, j) in plane one to neuron $(i+1, j+1)$ in plane two. Assume that the optical source driving the interconnect produces a collimated beam and the hologram is designed to focus the collimated beam to the photodetector at the input to neuron $(i+1, j+1)$. The desired hologram could be recorded using a plane wave as the reference beam, and a spherical wave focusing to the correct spot as the object beam. When read out with the plane wave from the optical source it will produce three beams. The replica of the object beam forms the connection as desired. Due to the undiffracted beam some of the signal directed at neuron $(i+1, j+1)$ will be received by the photodetector for neuron (i, j) in plane two. The third beam at the output of the hologram is a conjugate of the focusing object beam. The conjugate of the focusing beam is a diverging wave and it may be received by several photodetectors in plane two depending on how fast it diverges and the input sensitivity of the photodetectors. Both of the undesired beams form connections between neuron (i, j) in plane one with neurons in plane two. These extra connections create crosstalk between neurons in the network. It may be possible to train the network even in the presence of crosstalk terms but the best solution would be to eliminate them to the extent possible. By making the hologram "thick" it is possible to significantly reduce the power in these unwanted beams.

3.1.3 Thick Holograms

"Thick" or "Volume" holograms are formed by recording the interference of the two recording beams in a material that is thick compared to the period of the interference pattern. The physically thick recording material captures a section of the interference pattern in three dimensions. Early treatments of this type of hologram include Kogelnik's coupled wave approach [24] with modeled the behavior of plane waves passing through volume plane gratings. From Kogelnik's work the diffraction efficiency of a thick amplitude hologram may be as large as 3.7%, and the diffraction efficiency of a thick phase hologram may be as large as 100% [23, p.61]. Further work has shown that when the object beam is not a plane wave the diffraction efficiency will decrease [25]. The practical limit of diffraction efficiency for non-plane-wave phase holograms recorded using photographic emulsions is 60%, and for phase holograms recorded in photopolymer the practical limit is 90% [23, p.96].

3.1.4 Previous COIN holography

Early work on the optical interconnects focused on developing approaches to the design of the HOEs and the recording equipment needed to create the holograms. Miloš Komarčević began the development of a hologram writer that could be used to create the arrays of holographic optical elements needed for the COIN coprocessor [21]. His initial work was done using silver halide holographic plates with a recording layer approximately 7µm thick. These plates produce holograms that behave more like "thin" holograms than "thick" holograms. He was able to achieve 10% diffraction efficiency using BB-PAN plates and Kodak D-19 processing in simple holograms that were written and read out at the same wavelength and had no beam focusing capability, effectively plane gratings. At the time of Komarčević's work it was assumed that the optical sources in the COIN coprocessor would be vertical cavity surface emitting lasers (VCSELs), but the holographic recording material he used was not sensitive at the wavelengths available from these lasers. This required creating hologram recording geometries at one wavelength that result in the proper hologram for use at another wavelength. This change of wavelength problem will be discussed in detail in section 3.2. Using basic Bragg diffraction theory Komarčević was able to solve this problem for the case of plane waves. Miloš graduated in 2000 and his research is summarized in his thesis [21].

The work on the interconnect elements was continued by Marta Ruiz-Llata [22]. Ruiz-Llata advanced the interconnect work by developing recording geometries to produce holographic optical elements capable of both steering and focusing, as required by the COIN coprocessor. To solve the change of wavelength problem with the diverging and converging beams required by the new geometries, Ruiz-Llata applied the plane wave solution at the edge of the HOEs and used the result to predict the focusing positions of the two beams used to record the hologram, illustrated in Figure 3-3a [3]. To achieve greater diffraction efficiency a thick holographic photopolymer material was used, Polaroid ULSH 500-7A-49 (Aprilis HMC-050-6-15-A-200). This material had a published thickness of 200µm as opposed to the 7µm thickness of the BB-PAN material used previously. With the new photopolymer recording material holograms were recorded that could direct 60% of the transmitted light into the desired output beam. The last student to work on the optical interconnect design was Gilles Hennenfent who refined Ruiz-Llata's writing geometry to allow for the fact that the photopolymer is sandwiched between two plates of glass [4], as illustrated in Figure 3-3b. He also refined the hologram writing system and recorded holograms using Slavich PFG-03M that could direct up to 47% of the transmitted light into the desired output beam.

All of the holograms fabricated during this stage of COIN development had dimensions

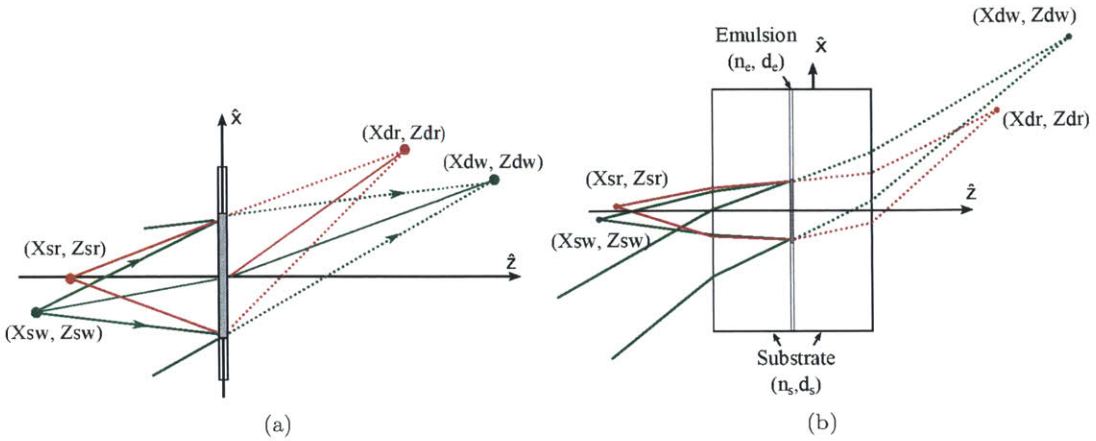


Figure 3-3: (a) Ruiz-Llata, changing geometry for focusing beams [3], (b) Hennenfent, changing geometry to account for substrate thickness and index [4]. The solid red lines indicate the desired readout beam, the solid green lines indicate the computed recording beams. All beams are incident on the hologram from the left side of the diagram during recording and readout. The dotted lines on the right side indicated the focusing positions of the desired readout spot (red) and focusing spot defining the focusing recording beam (green). These geometries were found by examining the behavior the beams at the edges of the hologram as pictured.

on the order of 2-5 mm, at least one order of magnitude larger than desired for the final COIN coprocessor. Also, all of the prior Photonic Systems Group students had approached the problem of designing hologram geometries under the assumption that at any point in the hologram the beams involved could be modeled as plane waves. Part of the early work for this thesis involved reexamining the change of wavelength problem to see what impact it would have on holograms written with finite beams.

3.2 Change of Wavelength Problem

When the wavelength used to record the hologram and the wavelength used to read out the hologram are the same, the design of the hologram is simple: make the reference beam be a replica of the light produced by the source and make the object beam be a replica of the desired output beam focusing to the correct point. Unfortunately it is not always possible to record the holographic optical element at the readout wavelength. The majority of the available holographic recording material is designed for use in the visible to ultra violet, but many interesting light sources produce light outside of this range. For this reason it is interesting to ask if the desired interference pattern due to the readout geometry at the readout wavelength can be recorded at a different wavelength. When the interference pattern is formed by two plane waves this is possible, and the result is investigated below. However, it will be shown that when the pattern becomes the result of three plane waves interfering there is no general solution to the change of wavelength problem.

3.2.1 Two Plane Wave Solution

The hologram formed by two plane waves interfering is simply a planar grating. This result is well known and gratings of this type have been investigated extensively. Rather than start with the known solution, the interference pattern will be derived here to develop insight into the change of wavelength problem. Let the waves have the following phasor representations:

$$\tilde{\mathbf{E}}_1 = \hat{e}_1 E_1 e^{j\mathbf{k}_1\cdot\mathbf{r}} e^{j\phi_1}, \qquad \tilde{\mathbf{E}}_2 = \hat{e}_2 E_2 e^{j\mathbf{k}_2\cdot\mathbf{r}} e^{j\phi_2}.$$

When the two waves are propagating in different directions, as illustrated in Figure 3-4, the vectors $\mathbf{k}_1$ and $\mathbf{k}_2$ will define the plane of propagation. For notational convenience let the plane of propagation be the (x, z) plane. Defining Θ_i to be the angle between the z axis and the $\mathbf{k}$ vector describing plane wave i allows $\mathbf{k}_i \cdot \mathbf{r}$ to be rewritten as

$$\begin{aligned}\mathbf{k}_i \cdot \mathbf{r} &= \left[\frac{2\pi}{\lambda}(\hat{\mathbf{x}}\sin\Theta_i + \hat{\mathbf{z}}\cos\Theta_i)\right] \cdot [x\hat{\mathbf{x}} + y\hat{\mathbf{y}} + z\hat{\mathbf{z}}] \\ &= \frac{2\pi}{\lambda}(x\sin\Theta_i + z\cos\Theta_i)\end{aligned}$$

The total electric field is the sum of the field due to each wave.

$$\begin{aligned}\tilde{\mathbf{E}}_{total} = \tilde{\mathbf{E}}_1 + \tilde{\mathbf{E}}_2 &= \hat{e}_1 E_1 e^{j\frac{2\pi}{\lambda}(x\sin\Theta_1 + z\cos\Theta_1)} e^{j\phi_1} + \hat{e}_2 E_2 e^{j\frac{2\pi}{\lambda}(x\sin\Theta_2 + z\cos\Theta_2)} e^{j\phi_2} \\ &= \hat{e}_1 E_1 e^{j\alpha_1} + \hat{e}_2 E_2 e^{j\alpha_2} \qquad (3.9)\end{aligned}$$

with $\alpha_i = \frac{2\pi}{\lambda}(x\sin\Theta_i + z\cos\Theta_i) + \phi_i$.

Holographic recording materials respond to the time average intensity of the field rather than the field itself. The time average intensity is proportional to the magnitude squared

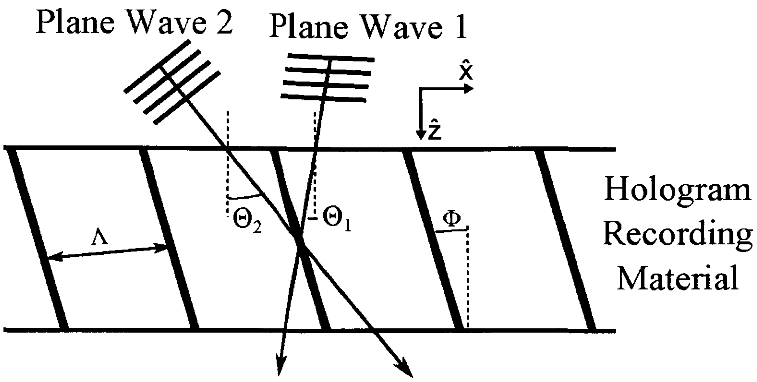


Figure 3-4: Interference of two plane waves

of the total field. If the waves are linearly polarized then the polarization vectors, $\hat{e}_i$, are real and the time average intensity will be

$$\begin{aligned} < I > \propto | \tilde{\mathbf{E}}_{total}|^2 &= (\hat{e}_1 E_1 e^{j\alpha_1} + \hat{e}_2 E_2 e^{j\alpha_2}) \cdot (\hat{e}_1 E_1 e^{-j\alpha_1} + \hat{e}_2 E_2 e^{-j\alpha_2}) \\ &= E_1^2 + E_2^2 + E_1 E_2 (\hat{e}_1 \cdot \hat{e}_2) e^{j(\alpha_1 - \alpha_2)} + E_1 E_2 (\hat{e}_1 \cdot \hat{e}_2) e^{-j(\alpha_1 - \alpha_2)} \\ &= E_1^2 + E_2^2 + 2 E_1 E_2 (\hat{e}_1 \cdot \hat{e}_2) \cos(\alpha_1 - \alpha_2) \end{aligned} \tag{3.10}$$

Figure 3-4 shows the relevant geometry for this interference pattern. The expression in (3.10) results in a sinusoidally varying intensity pattern. The sinusoidal intensity pattern has some period, Λ, and the fringes are slanted at some angle ϕ to the surface of the recording material. The pattern is uniform in $\hat{\mathbf{y}}$, so the fringes formed are planes. The fringe period and slant angle are

$$\Lambda = \frac{\lambda}{2 \sin \frac{\Theta_1 - \Theta_2}{2}}, \qquad \phi = \frac{\Theta_1 + \Theta_2}{2}. \tag{3.11}$$

In all of these expressions refraction at the surface of the hologram has been neglected, the angles and wavelengths relevant to the hologram geometry are those defined inside the recording material. Assume that the interference pattern in (3.10) represents the desired hologram when evaluated at the readout wavelength λ_R, but for some reason the the hologram cannot be recorded at this wavelength and must instead be recorded at a different wavelength, λ_W. The only spatially varying term in (3.10) is the argument of the cosine, $\alpha_1 - \alpha_2$. Matching the pattern at the new wavelength requires matching the argument of the cosine at the new wavelength.

$$\begin{aligned}
\alpha_1 - \alpha_2\big|_{\lambda_R} &= \frac{2\pi}{\lambda_R}\{x\sin\Theta_{R1} + z\cos\Theta_{R1} - x\sin\Theta_{R2} - z\cos\Theta_{R2}\} + \phi_{R1} - \phi_{R2}\\
\alpha_1 - \alpha_2\big|_{\lambda_W} &= \frac{2\pi}{\lambda_W}\{x\sin\Theta_{W1} + z\cos\Theta_{W1} - x\sin\Theta_{W2} - z\cos\Theta_{W2}\} + \phi_{W1} - \phi_{W2}
\end{aligned}$$

The difference $\alpha_1 - \alpha_2$ can be grouped into three terms: one static term due to the phase difference of the two waves at the origin, one term which varies with x, and one term which varies with z. The grouped terms will be matched at both wavelengths when the following three conditions below are satisfied.

$$\phi_{W1} - \phi_{W2} = \phi_{R1} - \phi_{R2} \tag{C1}$$

$$\frac{2\pi}{\lambda_R}(\sin\Theta_{R1} - \sin\Theta_{R2}) = \frac{2\pi}{\lambda_W}(\sin\Theta_{W1} - \sin\Theta_{W2}) \tag{C2}$$

$$\frac{2\pi}{\lambda_R}(\cos\Theta_{R1} - \cos\Theta_{R2}) = \frac{2\pi}{\lambda_W}(\cos\Theta_{W1} - \cos\Theta_{W2}) \tag{C3}$$

The following trigonometric identities can be used to simplify the matching conditions.

$$\sin a - \sin b = 2\sin\left(\frac{a-b}{2}\right)\cos\left(\frac{a+b}{2}\right) \tag{A}$$

$$\cos a - \cos b = -2\sin\left(\frac{a-b}{2}\right)\sin\left(\frac{a+b}{2}\right) \tag{B}$$

Apply Trig identity (A) to condition (C2):

$$\sin(\frac{\Theta_{R1} - \Theta_{R2}}{2})\cos(\frac{\Theta_{R1} + \Theta_{R2}}{2}) = \frac{\lambda_R}{\lambda_W}\sin(\frac{\Theta_{W1} - \Theta_{W2}}{2})\cos(\frac{\Theta_{W1} + \Theta_{W2}}{2}) \tag{C2'}$$

Apply Trig identity (B) to condition (C3):

$$\sin(\frac{\Theta_{R1} + \Theta_{R2}}{2})\sin(\frac{\Theta_{R1} - \Theta_{R2}}{2}) = \frac{\lambda_R}{\lambda_W}\sin(\frac{\Theta_{W1} + \Theta_{W2}}{2})\sin(\frac{\Theta_{W1} - \Theta_{W2}}{2}) \tag{C3'}$$

Conditions (C2') and (C3') can be combined by taking the ratio of the two.

$$\tan\left(\frac{\Theta_{R1}+\Theta_{R2}}{2}\right) = \tan\left(\frac{\Theta_{W1}+\Theta_{W2}}{2}\right)$$

This will be satisfied when

$$\Theta_{W1}+\Theta_{W2} = \Theta_{R1}+\Theta_{R2}+2\pi N, \tag{3.12}$$

where N is some integer. This result means that the two plane waves at the recording wavelength, λ_W, must have the same average angle with respect to the z axis as the two plane waves at the readout wavelength, λ_R. The $2\pi N$ term can be ignored since adding an even multiple of 2π is equivalent to doing nothing, and adding an odd multiple of 2π is equivalent to reversing the direction of propagation of both waves which does not change the interference pattern. Equation (3.12) can be used to simplify condition (C2') by recognizing that the cosine terms are identical when $\Theta_{W1}+\Theta_{W2}=\Theta_{R1}+\Theta_{R2}$.

$$\sin\left(\frac{\Theta_{W1}-\Theta_{W2}}{2}\right) = \frac{\lambda_W}{\lambda_R}\sin\left(\frac{\Theta_{R1}-\Theta_{R2}}{2}\right) \tag{C2''}$$

The new condition can be rewritten as

$$\Theta_{W1}-\Theta_{W2} = 2\arcsin\left[\frac{\lambda_W}{\lambda_R}\sin\left(\frac{\Theta_{R1}-\Theta_{R2}}{2}\right)\right] \tag{3.13}$$

The right hand side of (3.13) will be real whenever $|\frac{\lambda_W}{\lambda_R}\sin(\frac{\Theta_{R1}-\Theta_{R2}}{2})| \leq 1$. For $\lambda_W \leq \lambda_R$ this condition will always be true, and equations (3.12) and (3.13) will define the angles of two write beams such such that the interference pattern produced at wavelength λ_W is the same as the interference pattern produced at wavelength λ_R.

This result can also be derived by examining the fringe period and grating slant given in (3.11), and requiring that the recording geometry result in a grating slant and period that is the same as the expected grating slant and period that would be generated with the readout geometry. Equation (3.12) sets the average angle at wavelength λ_W to be equal to the average angle at wavelength λ_R, so the grating slant ϕ is the same at either wavelength. The requirement that the grating period be the same at both wavelengths is equivalent to stating

$$\Lambda = \frac{\lambda_R}{2\sin\left(\frac{\Theta_{R1}-\Theta_{R2}}{2}\right)} = \frac{\lambda_W}{2\sin\left(\frac{\Theta_{W1}-\Theta_{W2}}{2}\right)},$$

which can be rewritten as

$$\lambda_R \sin\left(\frac{\Theta_{W1}-\Theta_{W2}}{2}\right) = \lambda_W \sin\left(\frac{\Theta_{R1}-\Theta_{R2}}{2}\right). \tag{3.14}$$

The condition described by (3.14) is exactly the condition described by (3.13). The general approach to the change of wavelength problem for two plane waves can be summarized as follows. First find the grating slant angle, the two recording beams will be symmetric about this plane. Next, find the desired grating spacing. When the readout wavelength is larger than the recording wavelength, $\lambda_R \geq \lambda_W$, the write beams will be oriented closer to the fringe plane than the readout beams and there will always be a pair of $(\Theta_{W1}, \Theta_{W2})$ that solve (3.12) and (3.13). When the readout wavelength is shorter than the recording wavelength, $\lambda_R \leq \lambda_W$, there will be a pair of $(\Theta_{W1}, \Theta_{W2})$ that solve (3.12) and (3.13) when $|\frac{\lambda_W}{\lambda_R}\sin(\frac{\Theta_{R1}-\Theta_{R2}}{2})| \leq 1$. When there is a solution for $\lambda_R \leq \lambda_W$ the write beams will be oriented further from the fringe plane than the readout beams.

3.2.2 Beyond two plane waves

The preceding section established the solution to the change of wavelength problem when the hologram to be recorded is a simple plane grating. The holographic optical elements envisioned for the COIN coprocessor must be able to steer and focus light from a divergent source and will be more complicated than a plane grating. The approach to more complicated holograms within the Photonic Systems Group has been to assume that at each point in the hologram plane the two beams are locally plane waves with steering angles determined by the source point and desired focusing point. The hologram geometry is then computed by applying the two plane wave solution to the edges of the holographic optical element. This is illustrated in the Figure 3-5.

This approach was applied to the case of a small (100µm square) hologram and the expected interference pattern at the readout wavelength and the interference pattern at the recording wavelength were compared. The two interference patterns were quite different. This motivated an attempt to find a more general solution to the change of wavelength prob-

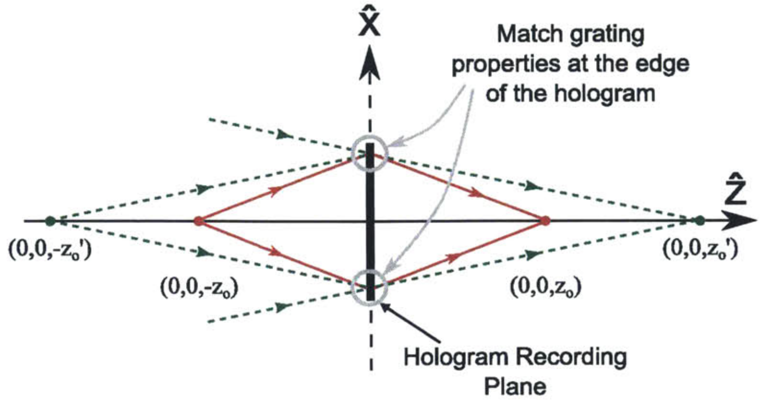


Figure 3-5: One possible result from the two point solution. For this case the source and focus points are symmetric about the hologram. Solving for the grating period and angle results at the edge of the hologram predicts new source and focus points to be used in the recording system. If the hologram is supposed to be read out with red light, but must be recorded with green light, the source and focus points for the recording beams (dashed green lines in the diagram) are further from the hologram than the source and focus points at the readout wavelength (solid red lines in the diagram).

lem for holograms more complicated than plane gratings. A convenient way to represent a propagating wave is through its angular spectrum, which is related to the Fourier Transform of the field due to the wave at some reference plane [26, Section3.10]. The resulting angular spectrum represents the propagating wave as a weighted sum of plane waves. As a move towards an "sum of plane waves" representation, the change of wavelength problem was investigated for an N-plane-wave interference pattern and it was found that there is no general solution when $N > 2$. This section will look at the case for three plane waves, and how it can be expanded to a general N plane wave case.

Consider three plane waves at wavelength λ_R, propagating nominally along the +z axis with angles Θ_1, Θ_2, Θ_3 as defined in the previous section. The phasors for the three waves are

$$\tilde{\mathbf{E}}_1(x,y,z) = \hat{\mathbf{e}}_1 E_1 e^{j\mathbf{k}_1\cdot\mathbf{r}} e^{j\phi_1} = \hat{\mathbf{e}}_1 E_1 e^{j\alpha_1},$$
$$\tilde{\mathbf{E}}_2(x,y,z) = \hat{\mathbf{e}}_2 E_2 e^{j\mathbf{k}_2\cdot\mathbf{r}} e^{j\phi_2} = \hat{\mathbf{e}}_2 E_2 e^{j\alpha_2},$$
$$\tilde{\mathbf{E}}_3(x,y,z) = \hat{\mathbf{e}}_3 E_3 e^{j\mathbf{k}_3\cdot\mathbf{r}} e^{j\phi_3} = \hat{\mathbf{e}}_3 E_3 e^{j\alpha_3}.$$

The time average intensity is proportional to the magnitude squared of the total field, which is just the sum of the individual fields.

$$\begin{aligned} < I > \propto \mid \tilde{\mathbf{E}}_{total}|^2 = E_1^2 + E_2^2 + E_3^2 &+ 2 < \hat{\mathbf{e}}_1 \cdot \hat{\mathbf{e}}_2 > E_1 E_2 \cos(\alpha_1 - \alpha_2) \\ &+ 2 < \hat{\mathbf{e}}_1 \cdot \hat{\mathbf{e}}_3 > E_1 E_3 \cos(\alpha_1 - \alpha_3) \\ &+ 2 < \hat{\mathbf{e}}_2 \cdot \hat{\mathbf{e}}_3 > E_2 E_3 \cos(\alpha_2 - \alpha_3) \end{aligned}$$

As in the two plane wave case the resulting $< I >$ is real and the spatially varying terms are the arguments of the cosines. Unlike the two plane wave case, there are now two expressions containing α_1, two for α_2, and two for α_3. If the $\mathbf{k}$ vectors of the three waves are restricted to the (x, y) plane, then the logic from the two plane wave case can be applied and the condition that $\alpha_1 - \alpha_2$ be the same at both wavelengths gives the equations

$$\begin{aligned} \frac{\Theta_{W1} + \Theta_{W2}}{2} &= \frac{\Theta_{R1} + \Theta_{R2}}{2} + N\pi \\ \frac{\Theta_{W1} - \Theta_{W2}}{2} &= \arcsin(\frac{\lambda_W}{\lambda_R} \sin(\frac{\Theta_{R1} - \Theta_{R2}}{2})) \end{aligned}$$

Adding these two equations together gives an expression for Θ_{W1}

$$\Theta_{W1} = \frac{\Theta_{R1} + \Theta_{R2}}{2} + \arcsin(\frac{\lambda_W}{\lambda_R} \sin(\frac{\Theta_{R1} - \Theta_{R2}}{2})) \tag{3.15}$$

Repeating this for the condition that $\alpha_1 - \alpha_3$ be the same at both wavelengths gives a second equation for Θ_{W1}

$$\Theta_{W1} = \frac{\Theta_{R1} + \Theta_{R3}}{2} + \arcsin(\frac{\lambda_W}{\lambda_R} \sin(\frac{\Theta_{R1} - \Theta_{R3}}{2})) \tag{3.16}$$

Consider the case that $\lambda_W < \lambda_R$, and place wave number 1 in the middle of waves 2 and 3 so that $\Theta_{R2} < \Theta_{R1} < \Theta_{R3}$. Equation (3.15) implies that wave one at the recording wavlength, λ_W, should move towards wave number two, while Equation (3.16) implies that wave one at the recording wavelength should move towards wave number three. It is only possible to satisfy both (3.15) and (3.16) when the two wavelengths are equal. This analysis has kept the current wave numbering scheme so that $\alpha_{W1} - \alpha_{W2}$ is paired with $\alpha_{R1} - \alpha_{R2}$, but this is not necessary. Any matching of $(\alpha_{W1} - \alpha_{W2},\ \alpha_{W1} - \alpha_{W3},\ \alpha_{W2} - \alpha_{W3})$ to $(\alpha_{R1} - \alpha_{R2},\ \alpha_{R1} - \alpha_{R3}\ ,\alpha_{R2} - \alpha_{R3})$ will still generate two conflicting conditions for the orientation of each plane wave. Reversing the direction of the wavelength change, so that $\lambda_W > \lambda_R$, results in a similar set of conflicting conditions on each beam angle.

The approach can be applied to the case of N plane waves, giving with a total field in (3.17) and the time average intensity in (3.18). From (3.18) it is clear that when there are N plane waves interfering there will be $N(N-1)/2$ cosine terms in the intensity function. In general there will be more equations than free variables, the system will be overdetermined, and there is no guarantee of a solution that will recreate the interference pattern in three dimensions. There are a few important points to note about this result. First, just because the interference pattern is not exactly duplicated does not mean that the hologram cannot be read out with light at λ_R. If the interference pattern is close enough to the desired pattern then the hologram should produce roughly the correct output, albeit with some aberrations and lower diffraction efficiency. Secondly, this is only a restriction on matching the interference pattern in all three dimensions. The next section will show that it is possible to exactly match the interference pattern in two dimensions. In some situations a two dimensional match may be good enough. Finally, this is a restriction on a single step reconstruction of the interference pattern!

$$\tilde{\mathbf{E}}_{total} = \sum_{n=1}^{N} \hat{\mathbf{e}}_n E_n e^{j\mathbf{k}_n \cdot \mathbf{r}} e^{j\phi_n} = \sum_{n=1}^{N} \hat{\mathbf{e}}_n E_n e^{j\alpha_n} \tag{3.17}$$

$$< I > \propto \mid \tilde{\mathbf{E}}_{total}|^2 = \sum_{n=1}^{N} E_n^2 + \sum_{L=1}^{N} \sum_{M=L+1}^{N} 2 < \hat{\mathbf{e}}_L \cdot \hat{\mathbf{e}}_M > E_L E_M \cos(\alpha_L - \alpha_M) \tag{3.18}$$

Examining the interference pattern shows that it is the sum of a set of cosines and a dc term. From the two plane wave case, each of these terms can be matched individually by one pair of plane waves. Holograms can be recorded over multiple exposures. Under the right conditions, each pair of plane waves could be used to expose the hologram eventually building up a the proper hologram. To do this would require a recording material that, from the point of view of the illuminating beams, does not change during exposure but only changes during the development process. Additionally, there are $N(N-1)/2$ cosine terms so this process would require a large number of exposures for all but the smallest number

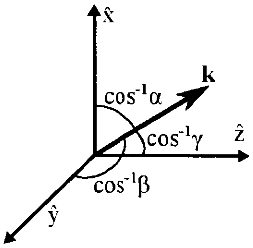


Figure 3-6: The wavevector **k** and the direction cosines α, β, and γ.

of interfering beams. When the exact hologram in three dimensions is not needed, the two dimensional match in the next section will be a better approach.

3.2.3 Solution in Two Dimensions

The previous section showed that there will not, in general, be a solution to the problem of changing the wavelength between hologram recording and readout because it will not be possible to recreate in three dimensions at λ_W the interference pattern caused in three dimensions by at λ_R when the interference pattern is formed by the intersection of more than two plane waves. However, that does not mean the interference pattern cannot be recreated in two dimensions. For a physically thin hologram, or for interference patterns which are slowly changing in one dimension, matching the pattern in two dimensions may be enough to create an acceptable hologram. To develop the two dimensional solution, start with the angular spectrum representation of the interference pattern field at λ_R. The notation for the angular spectrum and transforms in this section is taken from Goodman [26, Section 3.10].

$$\tilde{\mathbf{E}}(x, y, z) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \tilde{\mathbf{A}}_{\lambda_R}(f_x, f_y, z) \exp\left[j2\pi(f_x x + f_y y)\right] \partial f_x \partial f_y \tag{3.19}$$

Equation (3.19) expresses the electric field at some plane z as a sum of plane waves with spatial frequencies f_x and f_y, with each plane wave weighted by the term $\tilde{\mathbf{A}}(f_x, f_y, z)$. For a given plane wave the spatial frequencies f_x and f_y are related to the direction of the waves propagation k through the direction cosines illustrated in Figure 3-6, [26].

$$f_x = \frac{\alpha}{\lambda_R}, \qquad f_y = \frac{\beta}{\lambda_R}, \qquad \gamma = \sqrt{1 - (\lambda_R f_x)^2 - (\lambda_R f_y)^2} \tag{3.20}$$

Equation (3.19) defines the electric field as the inverse transform of the angular spectrum, so the angular spectrum can be computed by a Fourier Transform of the field at the same plane [26]. For convenience define the hologram plane to be the $z = 0$ plane. Then the angular spectrum in (3.19) can be computed from the field in the plane $z = 0$ by

$$\tilde{\mathbf{A}}\left(\frac{\alpha}{\lambda_R}, \frac{\beta}{\lambda_R}, 0\right) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \tilde{\mathbf{E}}(x, y, 0) \exp\left[-j2\pi\left(\frac{\alpha}{\lambda_R}x + \frac{\beta}{\lambda_R}y\right)\right] \partial x \partial y \qquad (3.21)$$

The transforms above are only computed over two dimensions, rather than three as might be expected. This is because the plane waves in the representation are constrained by the requirement that $|\mathbf{k}|^2 = k_x^2 + k_y^2 + k_z^2$. This allows the three dimensional integral to drop to a two dimensional integral. Now consider a field at λ_W in the $z = 0$ plane with an angular spectrum given by

$$\tilde{\mathbf{A}}_{\lambda_W}(f_x, f_y, 0) = \tilde{\mathbf{A}}_{\lambda_R}(f_x, f_y, 0) \qquad (3.22)$$

Using Equation (3.19) the field for this angular spectrum is

$$\tilde{\mathbf{E}}(x, y, z) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \tilde{\mathbf{A}}_{\lambda_W}(f_x, f_y, z) \exp\left[j2\pi\left(f_x x + f_y y\right)\right] \partial f_x \partial f_y \qquad (3.23)$$

In the $z = 0$ plane this field will have the same amplitude as the desired field. It might not be clear that this is a different field at first glance because the difference is hidden in the notation. The angular spectra, as a function of spatial frequency, were defined to be the same. But the direction of propagation for a wave with spatial frequency (f_x, f_y) will be different for a wave at wavelength λ_R than for a wave with the same (f_x, f_y) at wavelength λ_W. Expressing the angular spectra as a function of the direction cosines gives

$$\tilde{\mathbf{A}}_{\lambda_W}(\alpha, \beta, 0) = \tilde{\mathbf{A}}_{\lambda_R}\left(\alpha\frac{\lambda_W}{\lambda_R}, \beta\frac{\lambda_W}{\lambda_R}, 0\right) \qquad (3.24)$$

As the waves propagate, the amplitude spectrum changes. The amplitude spectrum at some plane z can be found from the amplitude spectrum at $z = 0$ using (3.25), which depends on the wavelength of the plane waves making up the field [26]. So, although the two fields are identical when $z = 0$ they will be different at any other value of z.

$$\tilde{\mathbf{A}}_{\lambda}\left(\frac{\alpha}{\lambda}, \frac{\beta}{\lambda}, z\right) = \tilde{\mathbf{A}}_{\lambda}\left(\frac{\alpha}{\lambda}, \frac{\beta}{\lambda}, 0\right) \exp\left\{j\frac{2\pi}{\lambda}\sqrt{1 - \alpha^2 - \beta^2}\right\} \qquad (3.25)$$

The technique outlined above could be summarized as follows: First, using the desired source and object beams compute the field at the plane of the hologram. Find the angular spectrum of this field, and using the angular spectrum find the amplitudes of the plane waves that will be used to replicate the field at the new wavelength. Since the field at λ_R in the hologram plane that is exactly the same as the electric field at λ_W in the same plane they will have the same time average intensities and create the same two dimensional patterns. However, as the fields propagate away from the hologram plane they will become different. There is one catch to the method outlined above related to the size of the features in the interference pattern. If the interference pattern is recorded at some λ_R, then it may contain fringes with a spacing on the order of half of the wavelength, $\frac{\lambda_R}{2}$. This occurs when the interference pattern contains counter propagating waves. As long as the recording wavelength λ_W is smaller than the readout wavelength then this should not be a problem. In the case where the readout wavelength is shorter, the technique above will only match the pattern if it does not contain features smaller than $\lambda_W/2$.

The exact two dimensional match discussed here could also be used to build up an approximate three dimensional match to the desired interference pattern. This could be accomplished by building up the three dimensional pattern as a stack of two dimensional matches. The stack could be built up using discrete thin holograms, or by adding a new layer of recording material between exposures. The downside to the first approach is that each discrete thin holograms would have to be positioned very accurately with respect to all of the other holograms in the stack. The second approach requires a recording material that can be applied in layers, with each new layer not damaging the previous layer, and with the exposed layers not being optically different from the unexposed layers.

3.2.4 Change of wavelength summary

The results of this thesis indicate that there is no guaranteed solution to the change of wavelength problem when the interference pattern to be replicated is formed by more than two plane waves. However, in most cases there will be a pattern that exactly matches the desired interference pattern in two dimensions. In the case of thin holograms, or small wavelength changes, this may not be an impediment to using holographic optical elements in the COIN coprocessor but recording the holograms at a different wavelength than they will be used at always complicates the design of the holograms. To create the best holographic optical elements the readout and recording wavelengths should be the same. The remainder of this chapter will examine some other limits on hologram behavior as the size of the holograms is reduced to the expected size range for the COIN coprocessor.

3.3 Issues of Small Holograms

How small should the optical elements in the COIN coprocessor be? To estimate the cost and physical size of a COIN network it was assumed that the optical elements would be roughly 200µm square. This size was chosen because it represents the size range of the optical elements fabricated for the material in Chapter 5 of this thesis. From the perspective of network cost and physical size there is an advantage to using optical elements that are as small as possible. However, there are physical limits imposed on the hologram dimensions due to the nature of the recording process and holograms themselves. This section will examine some of the issues of reducing the size of the holograms and increasing the size of the hologram arrays. For most of these limits the exact nature of the small finite beams will be ignored, and it will be assumed that they propagate as perfectly collimated finite waves with flat phase fronts. This approach approximates the problem as that of a simple plane grating which can be recorded as the interference pattern of two plane waves. The readout geometry for such a grating is shown in Figure 3-7.

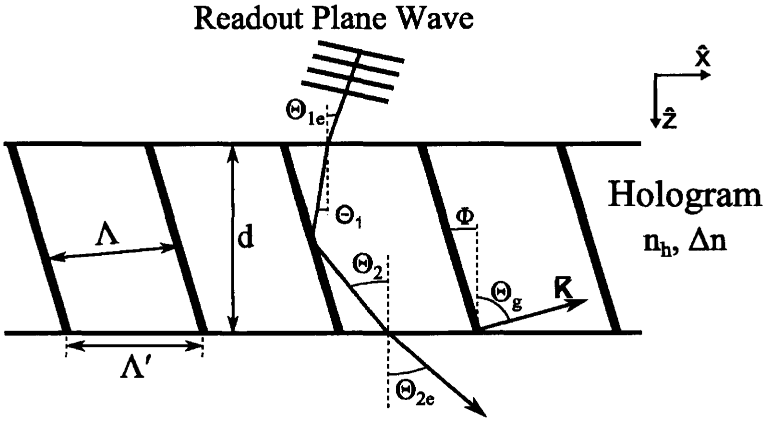


Figure 3-7: Hologram Readout Geometry

Designing this holographic grating requires two pieces of information: the angle of incidence of the expected readout beam, Θ_{1e}, and desired steering angle, Θ_{2e}. Both of these angles are defined outside the hologram itself. Analysis of the resulting holographic structure requires the internal angles which can be found from the external angles and the average index of refraction of the recorded hologram. To simplify the expression for diffraction efficiency in subsection 3.3.2, the grating vector $\mathbf{K}$ and associated angle Θ_g, are also defined in addition to the grating slant angle ϕ used in previous sections of this chapter.

3.3.1 Minimum Element Width

Holograms are by nature diffractive elements. Good performance requires some minimum width, and a commonly accepted lower limit is about 10 grating periods. For a plane wave interference pattern the important quantity is Λ', the projection of the grating period along the hologram surface. The projection can be computed from the physical grating period, Λ, and the angle of the grating relative to the surface of the hologram, ϕ. Both of these quantities will depend on the desired readout angle, steering angle, and hologram refractive index. Using the convention that angles are positive measured clockwise from the $\hat{\mathbf{z}}$ direction the quantities ϕ, Λ, and Λ' are given by

$$\phi = \frac{\Theta_1 + \Theta_2}{2}, \qquad \Lambda = \frac{\lambda/n_h}{2\sin\{(\Theta_1 - \Theta_2)/2\}},$$

and

$$\Lambda' = \frac{\Lambda}{\cos(\phi)} = \frac{\lambda/n_h}{2\sin\{(\Theta_1 - \Theta_2)/2\}\cos\{(\Theta_1 + \Theta_2)/2\}} \tag{3.26}$$

Table 3.1 gives the minimum element width computed for several steering angles under the assumptions that the desired readout beam is normal to the surface, $\Theta_1 = 0$ and that the average index of refraction of the recorded hologram is $n_h = 1.5$ for both $\lambda = 632.8$nm and $\lambda = 980$nm. As expected, the required minimum width decreases with increasing steering angle since the fringe spacing decreases as the steering angle increases. At both of the wavelengths examined the minimum element width is fairly small even for shallow steering angles. This requirement is a lower limit to the size of the holographic optical element and there is an advantage in increasing the actual size beyond this whenever possible.

Steering Angle	$\lambda = 632.8$nm Λ', µm	$\lambda = 632.8$nm Element Width, µm	$\lambda = 980$nm Λ', µm	$\lambda = 980$nm Element Width, µm
10°	2.429	24.294	3.762	37.624
20°	1.233	12.335	1.910	19.102
30°	0.844	8.437	1.307	13.067
40°	0.656	6.563	1.016	10.164

Table 3.1: Minimum Element Width to achieve 10 fringes for various steering angles assuming normal readout and $n_h = 1.5$ at two wavelengths.

3.3.2 Hologram Thickness Requirement

The introduction to holography provided at the start of this chapter stated that "thin" holograms will have lower diffraction efficiencies than "thick" holograms, but did not answer the question of how thick a hologram has to be to achieve good diffraction efficiency. A first estimate of the proper hologram thickness can be obtained by applying the result of Kogelnik's paper "Coupled Wave Theory for Thick Hologram Gratings" [24]. For a lossless slanted dielectric grating the diffraction efficiency, η, as a function of grating thickness is

$$\eta = \sin^2\left[\frac{\pi \Delta n d}{\lambda}\left(\cos^2(\Theta_1) - \frac{\lambda}{\Lambda n_h}\cos(\Theta_1)\cos(\Theta_g)\right)^{-1/2}\right] \tag{3.27}$$

where λ is the free space wavelength of the readout beam, Λ is the fringe period, Θ is the angle of the readout beam in the material, Δn is the index modulation in the material, and n_h is the bulk index of the material. Kogelnik's paper works under the assumption of plane waves and infinitely wide gratings, so the maximum diffraction efficiency will be 100%. Rewriting equation (3.27) for $\eta = 100\%$ and solving for the hologram thickness, d, gives

$$d = \frac{1}{2}\frac{\lambda}{\Delta n}\left(\cos^2(\Theta_1) - \frac{\lambda}{\Lambda n_h}\cos(\Theta_1)\cos(\Theta_g)\right)^{1/2} \tag{3.28}$$

The thickness needed to obtain 100% diffraction efficiency depends on both the steering angle and the index modulation that can be achieved in the recording material. The implied hologram thickness for a range of Δn is plotted below in Figure 3-8. The analysis in Kogelnik's paper neglects Fresnel reflection losses at the input and output faces of the hologram, so the total diffraction efficiency will be less than 100% at the thickness indicated by equation (3.28).

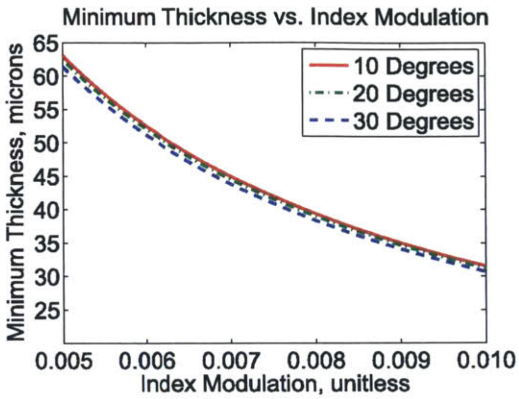


Figure 3-8: Grating thickness for 100 % diffraction efficiency as a function of index modulation under the assumptions of $\lambda = 632.8\text{nm}$ and $n_h = 1.5$.

3.3.3 Packing Density

Since the write beams are finite in extent, the interference pattern of the two beams will also be finite in extent. The write beams must be wider than the actual hologram to ensure the interference pattern extends through the entire depth of the hologram. Additionally, to allow each holographic optical element to be recorded independently the widths of the beams used during the recording process must be limited to an area as wide as the hologram spacing. This maximum width will be denoted the "subpixel width". Figure 3-9 shows the geometry involved in computing the usable width of the holographic elements, the width of the interference pattern which occurs through the entire depth of the recording material, for a hologram written with two collimated beams. Based on the internal angles of the write beams the usable width can be computed using equation 3.29.

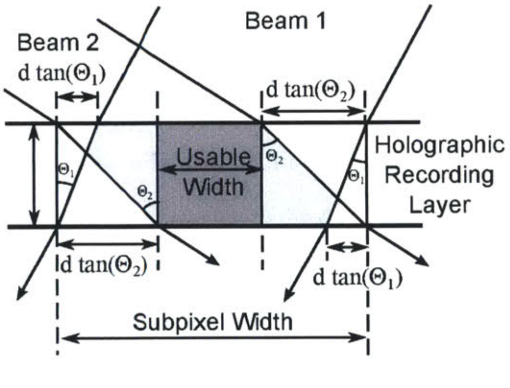


Figure 3-9: Geometric construction used to determine the usable holographic element width.

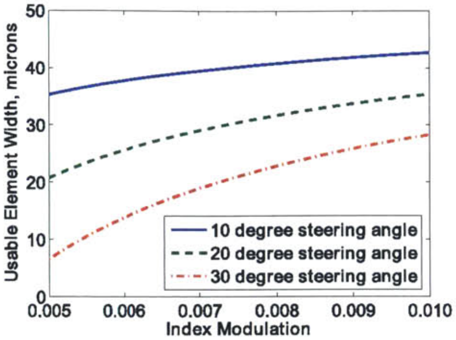


Figure 3-10: Usable element width as a function of achievable index modulation for 50µm wide subpixels.

$$W_{usable} = W_{subpixel} - 2d \times max(\tan \Theta_1, \tan \Theta_2) \tag{3.29}$$

The example in Figure 3-9 has $\Theta_1 < \Theta_2$. Both write beams are as wide as they can be without extending into adjacent holograms. The region in which the hologram extends the full depth of the holographic layer (the usable width) is shown shaded in a dark gray. There is overlap between the two beams, and a hologram is recorded, in the lightly shaded region. However, this hologram does not extend through the depth of the recording layer.

Equation 3.29 is plotted in Figure 3-10 for the same range of index modulation as in the previous section. Smaller steering angles result in a larger usable width for a given maximum element width. Also, higher index modulation allows a higher usable width because an increase in the index modulation decreases the required hologram thickness. Ideally the holographic optical elements would use as much of the space possible at any given optical element pitch, implying that the smallest possible steering angles should be used. However, the lower steering angles have a larger fringe spacing and will require larger minimum element widths, as discussed in Section 3.3.1. For some index modulation values the element width required by the grating spacing may be larger than the usable element width dictated by the steering angle and the optical element pitch.

3.3.4 Recording time

One issue that has not been addressed during previous work is the time required to record the hologram array. To record a hologram, the recording material must be loaded into the recording system and the material must be exposed at a certain energy. The exposure energy is generally determined experimentally. The exposure time is set by the ratio of the exposure energy to the power of the beams at the recording plane. The hologram is formed from the interference pattern of the two recording beams, so a small change in phase of one of the beams during the exposure time will shift the pattern during the recording process. To minimize the amount of movement during exposure, after loading the recording material the system is generally left to settle for 2-5 minutes before exposure.

The COIN system will require the fabrication of large arrays of holograms. The 428x428 neuron COIN, the largest that would fit on one wafer as discussed in Chapter 6, would require just over 1.6 million holograms assuming a nearest neighbor connection scheme. If the setting time and exposure could be reduced to one second it would take just over 450 hours to record the holographic optical elements needed for one plane of the COIN. Serial recording of the massive number of holograms need for a COIN system is inefficient, so a massively parallel hologram writer would be needed to create a fast COIN hologram production system. The search for a fabrication technique to create a massively parallel hologram writer was part of the motivation for the lithographic mask based optical element design discussed in Chapter 5.

3.4 Holographic Theory Summary

Work done for this thesis has advanced the holographic optical element design used in the Photonic Systems Group. In particular the limits associated with writing small holograms due to finite beam dimensions were established. Also, the change of wavelength problem was examined for cases more complicated than the two plane wave case used to generate write geometries in previous work and it was shown that there will not be an exact solution when the number of plane waves needed to describe the pattern is greater than two. For the COIN coprocessor this means that holographic optical elements will only be a viable interconnect choice when the wavelength of operation is in the range of a given holographic materials sensitivity. In addition to the theoretical results presented in this chapter, some experimental holographic work was performed for this thesis. The experimental work will be described in the next chapter.

Chapter 4

Holographic Optical Elements: Experimental Results

Work done for this thesis has advanced both the theoretical limits for Holographic Optical Elements (HOEs) in the COIN coprocessor system, which was discussed in the last chapter, and the fabrication of HOEs for the COIN coprocessor, which will be discussed in this chapter. Two fabrication techniques were developed as part of the work for this thesis: the Holographic Bead Lens and Holographic Lift-off. Both techniques were motivated by the availability of a liquid photopolymer for hologram recording that becomes solid after exposure and development. The manufacturer suggested that this material could be exposed in a 'glass - recording material - glass' sandwich, and one of the glass substrates removed off after development.

The holographic bead lens is a non-planer holographic optical element that consists of a plane grating recorded inside a hologram structure that has a lens like surface profile. This technique replaces one layer of glass in the recording sandwich with a mold, which can be removed after development, to form the liquid recording material into a lens like shape. The combination of a lens surface profile with an internal phase grating creates an optical element which can steer and focus input light, but only requires recording a plane grating. This may offer a solution to the change of wavelength problem discussed in the previous chapter by sidestepping the need to write gratings more complicated than plane gratings, and may reduce the recording time required for a COIN element array by allowing multiple holographic optical elements to be written at one time with a simple exposure system.

The second technique developed, Holographic Lift-off, allows for the creation of very thin holographic optical elements by separating the recording layer of the hologram from its substrate. This technique was developed as an extension of the bead grating lens when it was recognized that sometimes the bead lenses would release from the glass slide they were created on, and could be glued to a new glass substrate. This chapter will start with a discussion of the Holographic Lift-off Technique and then cover the Holographic Bead Lens. This is a reverse of the order the in which the techniques were developed, but results in a more logical presentation. Fabrication of the holographic bead lens required fabrication of small lens molds. The techniques developed to fabricate the lens molds will be presented at the end of this chapter.

4.1 Holographic Lift-off

Holographic recording material has traditionally been available as a layered structure consisting of a recording layer and one or more support layers. A typical silver halide hologram plate might have a emulsion layer that is 7-10μm thick which records the hologram and a glass substrate approximately 2.5mm thick to provide physical support to the emulsion layer and to make handling the recording plates easier. The thickness of the support layer limits how close other optics in a system can be placed to the hologram. When it is desirable to have close access to both sides of the hologram this may be a problem.

A simple solution is to fabricate the hologram recording plate with a thinner structure, and some of the emulsions from Slavich are available on acetate film as thin as 180μm. Another approach to this problem might be to fabricate a hologram recording layer within the system the HOE is being designed for, and recording the hologram in-place. This is an attractive option when a physically thin holographic optical element is desired and the system geometry is conducive to recording in place. Back reflections, opaque substrates, access to the desired hologram position, and compatibility issues between system components and the development process may make in place recording impractical. The third option is the holographic lift-off technique developed for this thesis. In this technique a holographic recording structure is created with one or more support layers temporarily attached to the recording layer. After exposure and development the recording structure is disassembled resulting in a hologram that is only as thick as the recording layer.

4.1.1 Lift-off Design Considerations

Before presenting the lift-off experiment performed for this thesis a more general discussion of the lift-off process design will be given. The lift-off process is different from the traditional holographic process because it involves separating the components of the hologram recording structure. This complicates both the material selection and hologram recording geometry. The material selection considerations will be discussed first.

Commercial holographic plates are not designed with the lift-off process in mind, so it makes sense to create a recording structure specifically for the lift-off process. This requires selection of the support material, recording material, and some method of temporarily attaching the two together. The recording emulsion should be physically strong, at least after developing the hologram, so that the recording can survive being separated from the support layers. Photopolymer recording materials are a good choice because they are fairly robust after curing, and stronger than the gelatin used in silver halide recording processes.

The substrate material should be optically clear and easy to handle. Glass is the obvious choice, and was the substrate material chosen for the experiment described in the next section. To allow the recording layer of photopolymer to separate from the substrate with minimal damage to the hologram a release agent should be used. Three basic approaches to the release agent have been identified: a release agent that bonds to the substrate and not the recording material, with a release agent that bonds to the recording material but can be separated from the substrate, or with a release agent that can be dissolved using a solvent that does not damage the recording layer. These options are illustrated in Fig. 4-1.

After choosing a recording material, release agent, and substrate material the recording structure can be designed. If the recording material is solid, or can be made solid before exposure, the recording structure could be as simple as a layer of recording material, a layer of release agent, and a substrate. However, if the recording material is a liquid that does not harden until after the hologram is recorded, or if it must be isolated from the atmosphere during the exposure and development process, then a second substrate will be needed to provide additional support to the recording layer and/or to isolate the recording layer during exposure. This is called sandwich mode exposure. In addition to supporting and isolating the recording material, sandwich mode exposure allows the recording layer thickness to be controlled by the use of spacers between the substrates. Applying different

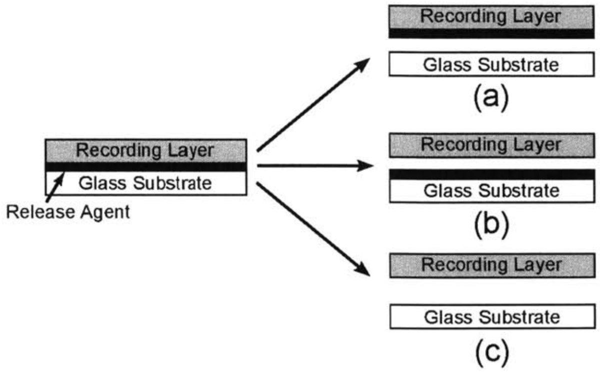


Figure 4-1: Release Agent Options. (a) a release agent which sticks to the recording layer but not the glass substrate, (b) a release agent which sticks to the substrate but not the recording layer, and (c) a release agent which can be dissolved in a solvent that does not damage the recording layer.

release agents to the two substrates in a sandwich mode exposure can control which side releases first during the separation process.

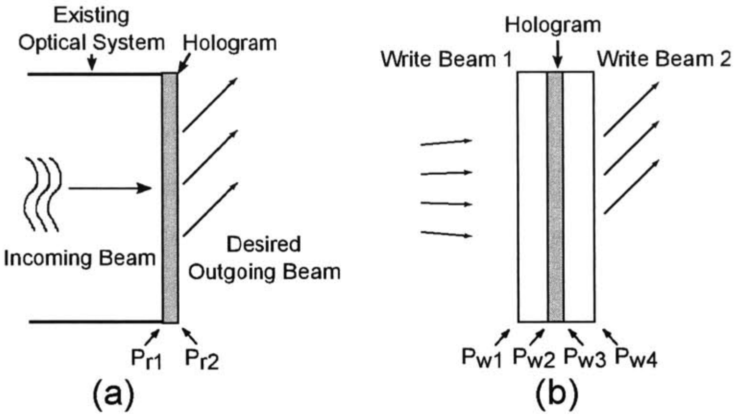


Figure 4-2: Change in substrate between hologram write and hologram readout

The materials used in a lift-off holographic process must be carefully chosen to survive the separation of the recording layer from the other layers in the recording structure. Disassembling the hologram recording structure between writing the hologram and using the hologram also introduces some changes to the recording system design. Consider the case in Figure 4-2. For a conventional holographic optical element the entire block in the recording system would be installed in the readout system. To generate the correct hologram the write system would need to match the field due to the incoming beam in plane P_{r1} with write beam one in plane P_{w1} and the field due to the outgoing beam in plane P_{r2} with write beam two in plane P_{w4}. With holographic liftoff, only the recording layer is transfered between the write and readout systems, so now the write system must match the field at plane P_{r1} to the field at plane P_{w2} and the field at plane P_{r2} with the field at plane P_{w3}. To avoid this complication in the experiment described in the next section, the decision was made to record a plane grating as the test hologram.

4.1.2 Experiment

The recording material used for the lift off experiment was Polygrama SM-532TR photopolymer. This material is designed for recording at 532 nm, can resolve more than 5000 lines/mm, and is capable of a diffraction efficiency of 88% as per the manufacturer's data sheet [27]. The material is supplied as a liquid and is developed by UV light exposure which also cures the polymer into a solid. SM-532TR must be isolated from oxygen during the exposure and development process, and it remains a liquid until after the UV curing step, so a sandwich mode exposure was required. Glass microscope slides were used as the substrate material as they are easy to handle and readily available. The manufacturer's data sheet suggested using Rain-X, a commercial hydrophobic surface coating [28], as a release agent on one of the glass plates during recording to allow that plate to be removed after hologram development. In this experiment Rain-X was used as a release agent on both glass slides to allow both glass slides to be removed after hologram development. The Rain-X coating bonds with the glass and prevents the recording material from bonding with the glass. Dow Sylgard ®, a castable Polydimethylsiloxane (PDMS), was also used as a release agent. This material was very effective in preventing the recording layer from sticking to the glass, but did not result in good holograms. A summary of the lift-off process used for this experiment is given in Figure 4-3.

Sandwich mode exposure allows the thickness of the recording layer to be controlled by using spacers. A diagram of the recording sandwich is given in Figure 4-4. A variety of spacer materials were used during this experiment including plastic shim material, polyimide tape, cellophane tape, and clear packing tape. The tape spacer options resulted in an easier to assemble recording sandwich because the adhesive allowed the spacer to stick to the slides.

This experiment was aimed at determining the feasibly of the lift-off process rather than the production of any particular holographic optical elements. To avoid the design

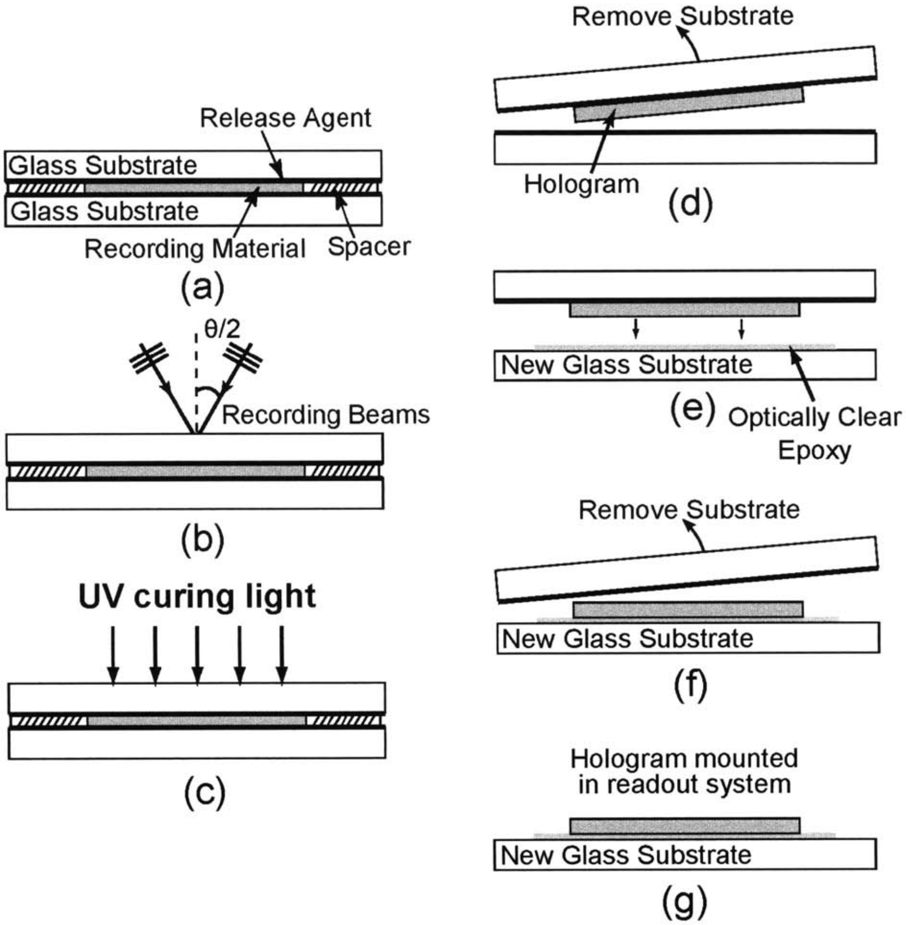


Figure 4-3: Sandwich Mode Lift-off process. (a) a recording sandwich is created using glass substrates treated with Rain-X. (b) The hologram is recorded. (c) The hologram is developed. For the photopolymer used in this experiment development was achieved by prolonged exposure to UV light. (d) After developing one layer of the recording sandwich is removed. (e) The hologram can now be attached to its new system with clear epoxy. (f) After fixing the hologram to its new system, remove the second substrate. (g) The final hologram in its new system, ready for use.

Spacer Material	Thickness
Plastic Shim Material	13 to 760 µm
Cling Film	15 µm
Aluminum Foil	12 to 20 µm
Packing Tape	75 to 80 µm
Cellophane Tape	45 to 60 µm
Polyimide Tape	55 µm

Table 4.1: Spacer Material Thickness. Thickness was measured with an outside micrometer. The range in the aluminum foil, packing tape and cellophane tape occurs across brands, within a roll the thickness is consistent. Only one polyimide tape and one cling film sample were measured.

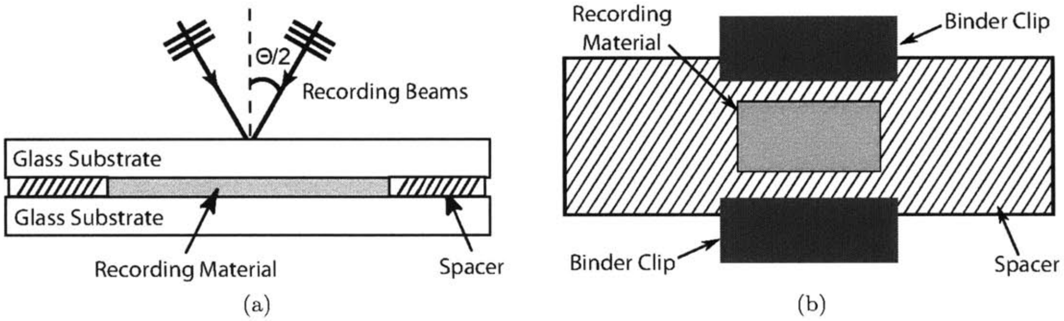


Figure 4-4: (a) Side view of the recording structure showing the glass substrates and spacer material. (b) Front view of the recording structure. The spacer material is arranged for form a well of recording material in the center of the glass slides. Binder clips are used to hold the recording structure together during exposure and development.

complications discussed in the previous section the holograms recorded for this experiment were simple plane gratings formed as the interference pattern between two collimated laser beams. The geometry used to record these holograms is shown in Fig. 4-5. One recording beam was set to be normal to the surface of the recording material, the other incident at an angle of 30 degrees. The laser used for this experiment was a frequency double ND:YAG laser producing light at 532nm, the recommend exposure wavelength for SM-532TR.

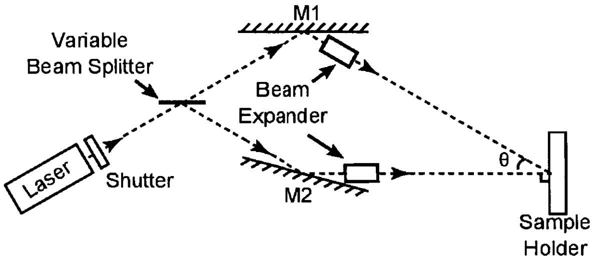


Figure 4-5: Hologram writing geometry

The photopolymer recording material creates a phase hologram by changing the refractive index across the material. The variation in index across the hologram can be expressed as the sum of a background index, n_b, and a spatially varying deviation from the background index, $\Delta n(x, y)$, giving an index of $n(x, y) = n_b + \Delta n(x, y)$. Assuming a background index of $n_b = 1.5$, the angle between the two recording beams inside the material is $\Theta = \arcsin(\sin 30/1.5) = 19.47°$. The period of the recorded grating is then given by

$$\Lambda = \frac{\lambda_{rec}}{2\sin(\Theta/2)} = \frac{532 \times 10^{-9}}{2\sin(9.735)}\text{m} = 1.57\mu\text{m}, \tag{4.1}$$

giving a fringe frequency of

$$f_g = \frac{1}{\Lambda} = 636\frac{lines}{mm}, \tag{4.2}$$

which is well within the published resolving power of $5000\frac{lines}{mm}$ [27]. At the time of the experiment the recording material was nearing the end of its usable life. Based on a series of test exposures, an exposure of 320mJ/cm^2 was used for this experiment, which is much larger than the manufacturer's suggested exposure of $15 - 30\text{mJ/cm}^2$.

Development of the Polygrama material is achieved by UV exposure. For the series of experiments described in this paper a high power UV LED (Thorlabs M365D1), with a wavelength centered at 365nm, was used. Following the manufacturer's instructions the

Test Number	Before Separation	After Separation	After Mounting
1	25%	19.3%	12%
2	25.8%	28.2%	18.5%
3	26.1%	26.4%	Damaged
4a	26%	15.7%	Continued next line
4b	See Line 4	25.8%	11.8%
5 SG1	34.7 %	36.8 %	31.8 %
6	38.0%	34.9 %	27.98%, Fogged
7	6.2%	Damaged	Damaged
8 SG2	35.3 %	38.2%	31.5 %

Table 4.2: Plane grating test results, diffraction efficiency at various stages of the liftoff process

material was exposed to the UV light until clear. This was achieved with 2.5 to 3 hours of exposure to the UV led operating at approximately one third its rated power, and a distance of 5cm.

To determine if the hologram is damaged during the liftoff process the diffraction efficiency was measured at three points in the liftoff process after development: before separation of the layers, after the removal of one glass substrate, and after mounting the hologram to the new substrate and removing the second recording substrate. The diffraction efficiency was calculated as the ratio of the power of the light in the first diffracted beam to the total power of the light leaving the hologram. This is consistent with the definition used in [24].

$$\eta = \frac{\text{Power in First Order Beam}}{\text{Total Power leaving hologram}} \tag{4.3}$$

Several adhesives were tested to glue the recording layer to the new glass slide after separation. These include Norland UV curing optical cements, a spray adhesive, and superglue. A summary of the measurements is given in Table 4.2. A picture of the first hologram to survive the liftoff process is given in Figure 4-6.

4.1.3 Results

The diffraction efficiency measurements for the eight liftoff attempts are given in Table 4.2. The results in the table were achieved with an exposure of 320mJ and used Rain-X on both glass slides as a release agent. All of the attempts used two layers of packing tape as spacers for a recording layer thickness of approximately 150µm, except for experiment 4 which used three layers of tape for a thickness of 225µm. Attempts one, two, and three attempted to mount the hologram to its new slide using Norland 68 Optical Adhesive. This worked ok, but the mounted hologram would separate from the slide within a day. Attempt four initially used spray adhesive to attach the hologram, but this did not hold well enough to separate the remaining glass slide. The hologram from attempt four was cleaned and remounted using Norland 61, which was also used to mount sample six. Norland 61 seemed to hold initially, but again failed within a few days. Attempts five and eight used super-glue to mount the recording layer, and this seemed to produce the best result with the mounted hologram remaining on the slides for at least a week after mounting.

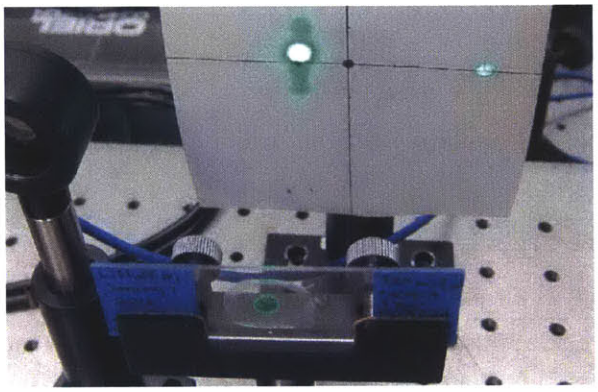


Figure 4-6: An example of a "lifted" hologram showing diffraction.

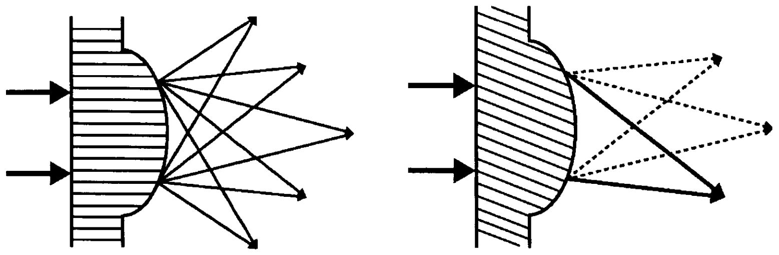
4.1.4 Liftoff Summary

Through proper choice of recording material, support material, and release agent the lift-off process allows the creation of holograms that are exactly as thick as their recording layer. This offers an interesting approach to holographic optical element design for applications in which close access to both sides of the HOE is needed. The results of this experiment show that the process is possible, though there is room for refinement. In addition to the lift-off process the use of liquid holographic recording materials also allows the development of hologram structures that are not simple planes, which will be discussed in the next section.

4.2 Holographic Bead Lens

Chapter 3 of this thesis showed that there will not, in general, be a single exposure solution to the change of wavelength problem when the desired hologram is more complicated than a plane grating. The best solution to this problem is to record holographic optical elements at the wavelength at which they are to be used. From a brief argument involving the hologram recording process chapter three also argued that for large arrays of holographic optical elements recording the elements individually and in series will take an excessively long time. The solution to this is to create a hologram recording system that can record several holograms at one time. A nearest neighbor COIN network would require three distinct optical elements and each of these optical elements needs to steer and focus its input light, so each element type requires a non-trivial recording system. Making an array recording system capable of writing several of each optical element type at once is a difficult task. The holographic bead lens discussed in this section offers a possible solution to both the change of wavelength problem and the problem of creating a massively parallel hologram recording system.

Rather than a traditional planar hologram the holographic bead lens is a hologram with the surface profile of a lens. The basic idea is to use the lens shape to collimate/focus the input light, allowing the light to be steered with a simple plane grating. Since solutions to the change of wavelength problem exist for plane gratings this type of holographic optical element could be recorded at wavelengths different from the operating wavelength. Plane gratings are recorded by the interference pattern of two plane waves. In theory one pair of wide plane waves at the proper orientation could record an entire interconnect plane worth of one hologram in one recording step. To record the optical elements for a plane of N COIN style nearest neighbor neurons would require roughly $9N$ hologram recording steps if the holograms are recorded one at a time, but with this approach recording the elements could be accomplished in eight recording steps: four steps to record the edge connection holograms, and four steps to record the diagonal connection holograms. The center connection element would not need a hologram recording step in this scheme. This section will examine two types of holographic bead lens, depicted in Figure 4-7. The normal grating lens in Figure 4-7a has a grating oriented perpendicular to the input plane, and produces a number of output beams. The slanted grating lens in Figure 4-7b has the internal grating slanted to direct most of the light into one output beam. The power in the output beams will be a function of the total thickness of the grating and the achievable index modulation within the material. To increase the power in the diffracted beams, both lenses have some constant thickness section attached to boost the average grating thickness.



(a) Normal Grating (b) Slanted Grating

Figure 4-7: Two holographic optical elements with a spherical front surface, the lines inside the structure indicate planes of equal phase. In case (a) the holographic grating is written normal to the flat surface of the object, and in case (b) the grating is slanted with respect to the flat surface of the element. By properly slanting the grating in (b) we can get Bragg readout and most of the output power should be placed in the +1-diffracted order. Only the three primary output beams are shown.

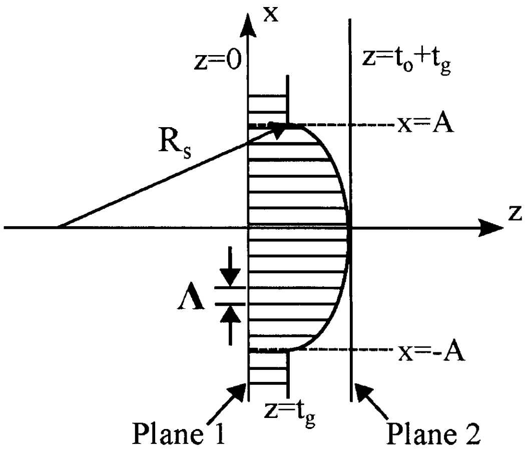


Figure 4-8: Bead Geometry for the case in figure 4-7a

4.2.1 Analytic Model For a Thin Normal Grating Lens

Modeling the normal grating lens requires finding a transmission function that relates the electric field at plane one to the electric field in plane two in Figure 4-8. If the lens portion of the structure is thin relative to the radius defining the surface ($t_o << R_s$) then the effect of refraction at the curved surface of the lens can be ignored and the assumption can be made that the light propagates parallel to the z-axis from the lens surface to plane two. The approximation works because the change in angle at the output of the lens as a function of position is small and because the distance between the second surface of the lens and plane two is small. For the propagation geometry shown, with the index change in the grating perpendicular to the direction of light propagation, it can also be assumed that light propagates straight through the grating and picks up a phase which varies across the grating due to the changing index. Under these approximations the transmission function for the normal grating lens can be written as

$$T(x,y) = e^{-j\phi(x,y)}, \tag{4.4}$$

with

$$\phi(x,y) = \frac{2\pi}{\lambda}\left[(t_l(x,y)+t_g)n(x,y) + n_{air}(t_o + t_g - t_l(x,y))\right]. \tag{4.5}$$

In these equations $t_l(x,y)$ and $n(x,y)$ represent the thickness of the lens portion of the structure and the index of the material index at point (x,y). Without loss of generality, assume the plane grating is oriented so that the refractive index change is along the $\hat{\mathbf{x}}$ direction. For the case of phase gratings recorded as the interference pattern of two plane waves, the index modulation will have the form

$$n(x) = n_m + \Delta n \cos\left(\frac{2\pi}{\Lambda}x\right), \tag{4.6}$$

where n_m is the average index of refraction for the photopolymer, Δn is the modulation of the index of refraction, and Λ is the period of the grating. For a spherical surface the thickness of the lens as a function of position, $t_l(x,y)$, will depend on the radius defining the spherical surface, R_s, and the maximum thickness of the lens, t_o.

$$t_l(x,y) = \sqrt{R_s^2 - (x^2+y^2)} - (R_s - t_o) \tag{4.7}$$

Substituting (4.6) into (4.5) and re-grouping terms gives

$$\begin{aligned}\phi(x,y) &= \frac{2\pi}{\lambda}[n_a(t_o+t_g) + (n_m - n_a)(t_g + t_l(x,y))] + \frac{2\pi}{\lambda}[\Delta n(t_l(x,y)+t_g)\cos(\alpha x)]\\ \phi(x,y) &= \phi_1(x,y) + \phi_2(x,y)\end{aligned} \tag{4.8}$$

So the transmission function for the normal grating lens is

$$T(x, y) = e^{-j\phi_1(x,y)}e^{-j\phi_2(x,y)}. \tag{4.9}$$

When illuminated with a plane wave propagating along $\hat{\mathbf{z}}$, the field at plane two is simply the product of the field amplitude and the transmission function.

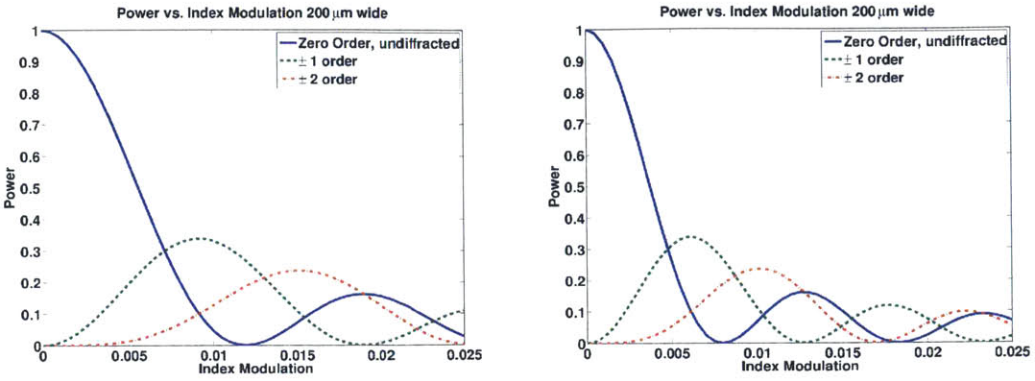
$$\tilde{\mathbf{E}}_2 = E_o T(x, y) = E_o e^{-j\phi_1(x,y)}e^{-j\phi_2(x,y)} \tag{4.10}$$

The first term, $\phi_1(x, y)$, corresponds to the phase due to a thin lens with focal length $F = R_s/(n_m - n_{air})$. At each point in plane two the second term of (4.5) can be expanded using the Jacobi-Anger identity.

$$e^{-j\phi_2(x,y)} = \sum_{m=-\infty}^{\infty} j^m J_m\left(\frac{2\pi}{\lambda}\Delta n[t_l(x, y) + t_g]\right) e^{jm\cos(2\pi x/\Lambda)} \tag{4.11}$$

In (4.11) J_m is a Bessel function of the first kind of order m. In the case of an infinite width grating of fixed thickness this expression would represent a sum of plane waves, propagating with spatial frequencies $f_{xm} = \frac{m}{\Lambda}$ and amplitude $J_m(kt_o\Delta n)$. When the thickness of the grating is changing with position (4.11) still represents a sum of waves with different spatial frequencies, but now the waves have an amplitude that varies in the output plane. Knowing that each term in the Bessel function expansion is associated with a different plane wave provides a way to estimate the power in the various orders by integrating the expansion terms in the output plane. A plot of this power estimate as a function index modulation is shown for two different bead grating lenses in Figure 4-9. In both cases the lenses are designed to be 200µm wide with 4mm focal lengths. In 4-9a the grating preceding the lens is 20µm thick, and in 4-9b the grating is 30µm thick.

The power in the diffracted orders depends on both the index modulation and the thickness of the lens and grating. If the focal length remains constant, decreasing the width of the lens decreases the maximum thickness of the lens portion of the structure. At the same time, when the focal length of the lens is small there will be a larger change in lens thickness across the optical element. So there will be more variation in the power associated with each order across the lens. In the normal grating lens geometry the power directed into the positive and negative diffracted orders is the same. This symmetry may be desirable in some situations, such as in an optical broadcast neural network [3].



(a) (b)

Figure 4-9: Relative power of the 0, ±1, and ±2 diffraction orders as a function of index modulation for two bead grating lenses. Both lenses are designed to have a 4mm focal length and 200μm lens diameter. In (a) the grating preceding the lens is 20μm thick, and in (b) the grating before the lens is 30μm thick. For the same range of index modulation, increasing the grating thickness allows for more power to be transfered into the diffracting orders.

In the focal plane of the lens the expected pattern can be computed by numerically evaluating the diffraction integral with the input field given above. The diffraction integral is similar to a Fourier transform. This implies that the output at the focal plane of the holographic bead lens will be related to the convolution of the two terms in (4.10). The lens term in the transmission function implies that the output beams will focus, approximately, at the plane $z = F = R_s/(n_m - n_{air})$. The second term implies that there will be several spots in the output plane. Since the amplitude of the waves in the lens output plane is not uniform the spots in the output plane will have a shape different from that due to an aperture of the same width. With the transmission function as defined above, it is possible to numerically evaluate the diffraction integral in the output plane to predict spot size. Figure 4-10 shows the result of the diffraction pattern, as predicted by a numerical evaluation of the diffraction integral, across the focal plane of the 200μm wide, 4mm focal length lens, with a 20μm grating preceding the lens. In 4-10a, the index modulation has been set to 0.005 and in 4-10b the index modulation has been set to 0.01. Details of the numerical evaluation of the diffraction integral are given in the appendix.

For some network geometries the Normal Grating Lens may be a good choice of interconnect element. Due to its low diffraction efficiency and the fact that it produces multiple output beams it is not well suited to the COIN coprocessor network geometry. One way

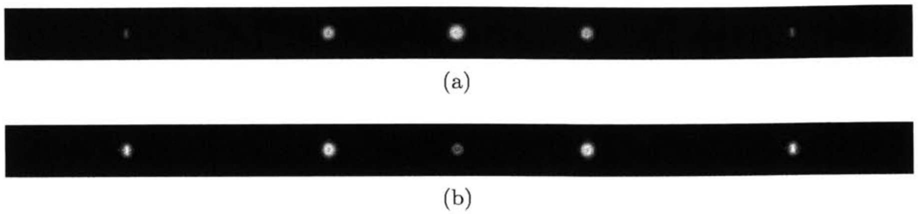
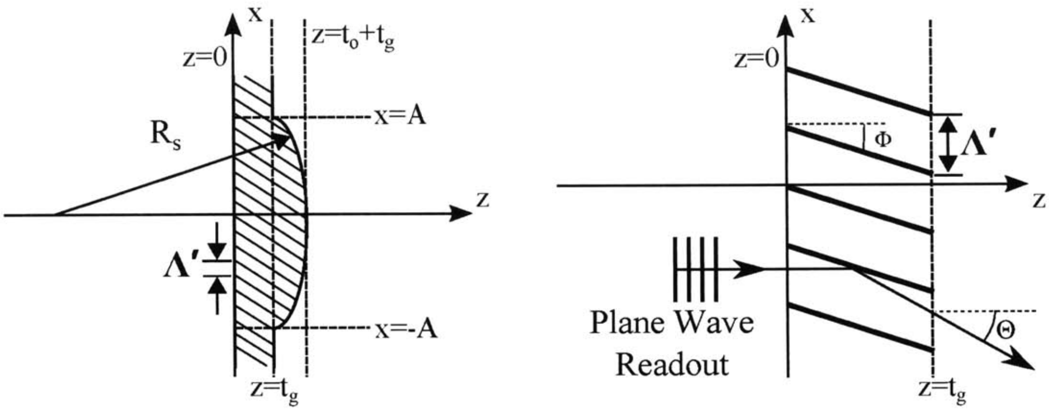


Figure 4-10: Simulation of a 200µm wide, 4mm focal length normal grating lens with a steering angle of 20°, and an additional grating thickness of $t_g = 20\mu m$. In (a) the index modulation is set to 0.005, and in (b) the index modulation is set to 0.01. Brightness in this image is proportional to log_{10}(intensity), to help bring out some of the low intensity diffraction features.

to increase the diffraction efficiency and eliminate the extra output beams is to create a structure more like a thick hologram by slanting the grating for Bragg readout.

4.2.2 Modeling A Slanted Grating Lens



(a) Bead Geometry for a slanted grating lens.

(b) Grating Geometry

Figure 4-11: (a) The slanted grating lens. (b) Plane grating portion of the lens.

The goal of the optical elements in the COIN coprocessor is to maximize the power directed to the desired neuron input. For holographic elements this means maximizing the power into the first diffracted order. Switching from a normal grating, discussed in the previous subsection, to a thick slanted grating will increase the power directed into the first order output and decrease the power directed into all other outputs. The geometry of the full structure and the grating are given in Figure 4-11. A COIN optical element might have a designed focal length of 0.75mm, and a total width of 200µm. Assuming an effective index of 1.5, this lens would only be 3.4µm thick at the center of the structure. Assuming an index modulation of $\Delta n = 0.005$, from Equation 3.27 the grating would have to be 63µm thick to achieve 100% diffraction efficiency. The grating thickness required to achieve a high diffraction efficiency is almost twenty times the thickness of the lens structure. This section will examine two approaches to modeling the slanted grating lens. In the first the two components will be separated and examined independently. The second approach will approximate the result by modeling each point in the structure as a plane grating that can be described by Kogelnik's coupled wave model.

Independent Lens and Grating

In cases where the grating is much thicker than the lens structure it makes sense to separate the problem into two parts: a slanted plane grating and a lens. At the output of the grating section there will be two beams, one propagating in the direction of the illuminating beam and one in the direction of the the first order diffracted beam. The relative amplitudes can be determined from [24]. The two output beams can then be propagated through the thin lens structure to get the field at the output of the slanted grating lens, and the diffraction integral applied to determine the field in the focal plane of the lens.

This approach will work best when the grating thickness is much larger than the lens thickness. In this case the slanted plane grating is what converts the input light to the diffracted beam. In modeling the slanted grating lens as a plane grating and a lens without grating, any scattering due refractive index changes inside the lens portion caused by the recording process will be missed. Also, because the base assumption is that the output from the grating is limited to the undiffracted and plus one diffracted orders any higher order diffraction will be excluded from the predicted output. Finally, there will also be some error due to the fact that the read out beam is actually a Gaussian and not a plane wave.

Approximate as stepped grating

An alternate approach to modeling the slanted grating lens is to approximate the surface as a series of plane gratings of varying thickness, and applying Kogelnik's approach to each grating. In this approach the amplitude of the undiffracted and first order diffracted beam will vary across the surface of the structure. Since Kogelnik's result depends on infinitely wide gratings this model will be most accurate when the lens structure is relatively flat. Regardless of how flat the lens structure is there will still be some error in this approximation due to the limited extent of the holographic gratings and the fact that the illuminating beam is not a plane wave.

4.2.3 Experiment

To test the viability of holographic bead lenses two sets of experiments were performed. The first set was aimed at creating normal grating lenses, and used well slides and pressed acrylic for lens molds. The second set of experiments was aimed at recording slanted grating lenses, and mainly used machined and vapor polished acrylic lens molds. The general exposure-develop process used in this set of experiments was the same as in the Holographic Lift-off experiments, with the exception that one of the glass substrates was replaced with a lens mold in the Holographic Bead Lens experiments. The bead lens process is illustrated in Figure 4-12. For the bead grating lenses, the UV development light was provided by a quartz halogen lamp during some experiments and a black-light-blue fluorescent bulb during others. The quartz halogen lamp worked, but also flooded the samples with light at the recording wavelength in addition to the UV curing wavelength. The fluorescent black light was able to cure the samples but required a long develop period. Problems with these sources motivated the move to the UV LED for the lift-off experiment.

Normal Grating Lens

A series of holographic bead lenses were formed using SM-514TR green sensitive photopolymer made by Polygrama Photopolymers [27]. This material is similar to the SM-532TR used for the holographic lift-off experiment, except its peak sensitivity for recording is at 514nm. For this experiment holographic recording was performed at 488nm, the blue line of an argon ion laser. The experiment used a two beam recording system, similar to that used in the lift-off experiment, but with the plate holder oriented so that the recording beams were symmetric about the plate holder normal. The recording geometry is shown in Figure 4-13.

Two types of lens molds were used during this experiment. The first were microscope well slides, roughly 15mm wide and 100µm deep at the well center. The well slides successfully molded the photopolymer into a lens shape, and coating the wells with Rain-X was adequate to ensure the lenses would release from the mold. Unfortunately the wells were too deep, creating a large variation in recording material thickness across the hologram. This made

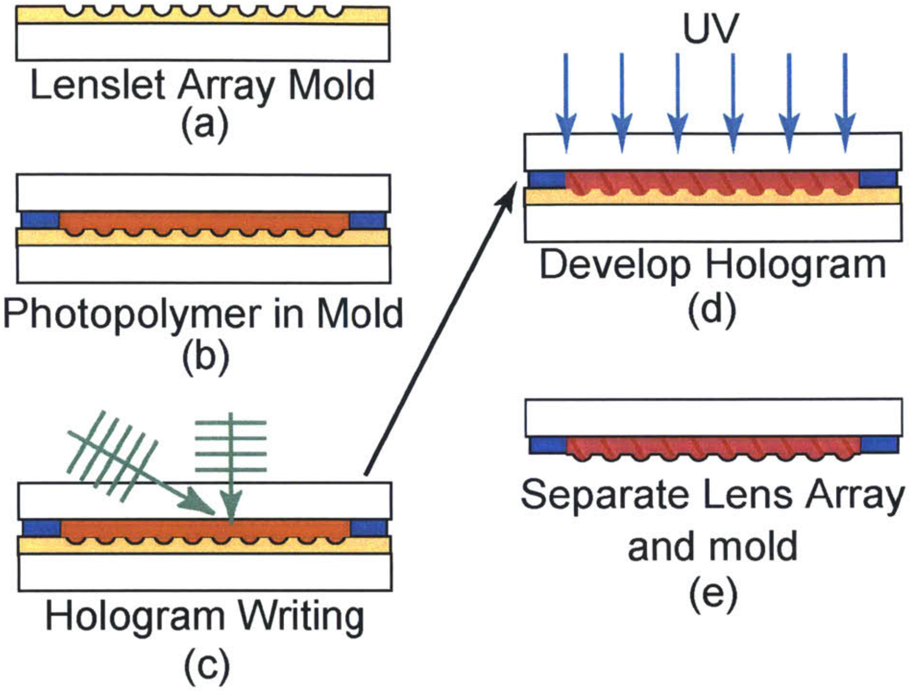


Figure 4-12: Holographic Grating Lens process diagram. This is similar to the lift-off process, except it starts with a treated lens mold (an array is illustrated here). In (b) the recording sandwich is assembled using the treated lens mold, a plain slide, and the recording photopolymer. (c) The hologram is recorded. (d) The photopolymer is developed by prolonged exposure to UV light. (e) The lens mold is removed leaving a holographic grating lens, or grating lens array.

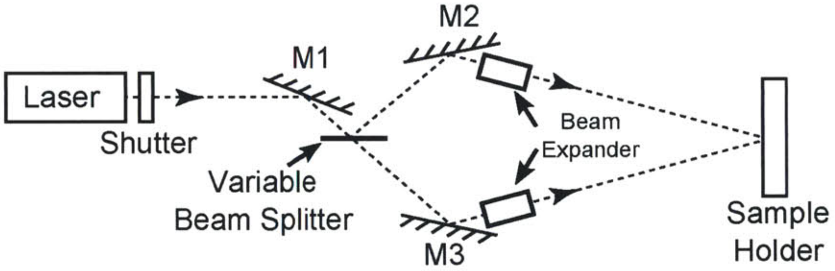
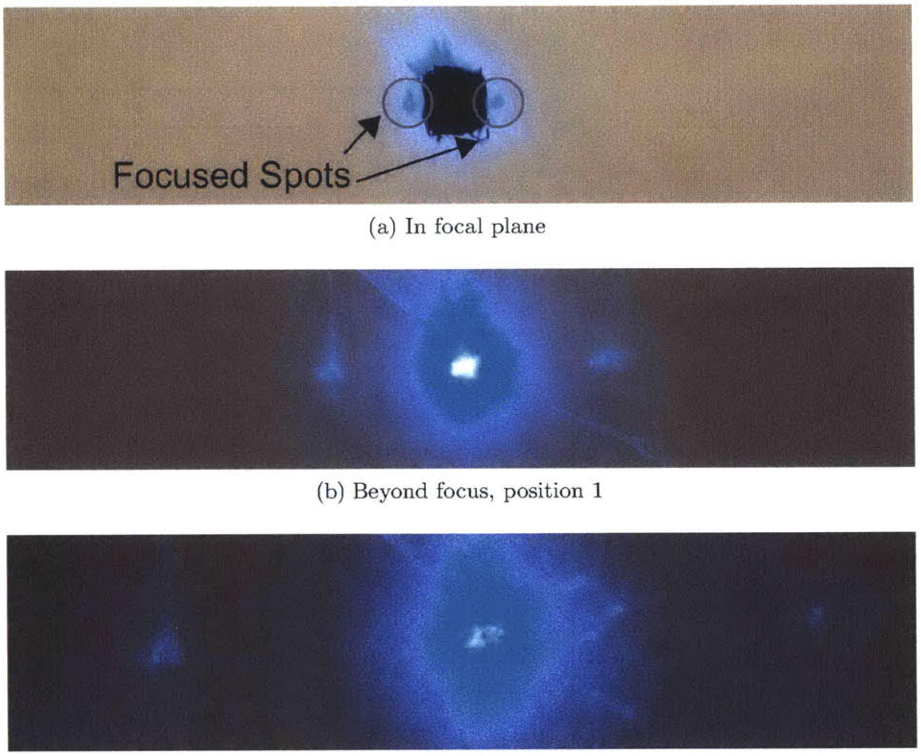


Figure 4-13: Recording system used for the Normal Grating Lens experiment. The beams are set to be 20° apart when they hit the plate holder. The plate holder was adjustable, and set so that the recording beams were symmetric about the plate normal.

correct exposure and development with the two beam system impossible as there was no way to vary the exposure across the well to match the thickness variation.

The second type of lens mold used for this experiment was formed by pressing a ball bearing into a heated piece of acrylic. This resulted in a shallower well that was easier to expose/develop, but the surface of the well was not smooth enough to form a smooth hologram. An image of the diffraction pattern at three planes for one of the holograms is shown in Figure 4-14.



(a) In focal plane

(b) Beyond focus, position 1

(c) Beyond focus, position 2

Figure 4-14: (a) Image of the focal plane. The center spot has been blocked because it saturates the camera used to make the images. (b) and (c) show the beams diverging as the distance from the hologram is increased. The spots visible from left to right are the -1 order diffracted beam, the zero order undiffracted beam, and the +1 order diffracted beam.

The holographic bead lens produced during this experiment did exhibit focusing and steering of the incoming laser light, but were not very effective in transferring light to the diffracted beams. The diffraction efficiency was about 0.7% for the first order spots. The next set of experiments was aimed at the slanted grating lens, with the goal of improving both the diffraction efficiency of the grating lenses and the focusing ability of the lenses.

Slanted Grating Lens

The normal grating lens experiment produced holographic bead gratings that did demonstrate focusing and steering, but the diffraction efficiency was quite low and the resulting bead lenses produced a lot of scattered light. Two problems were identified: (1) the wide range of hologram recording material thickness across the hologram, and (2) rough surfaces in the lens molds which create rough surfaces in the final hologram bead lenses. To eliminate some of these problems, and as an attempt to achieve higher diffraction efficiency, a series of Slanted Grating lenses was made using a new lens mold made from acrylic that had been milled and vapor polished. An image of one of these lens molds is given in Figure 4-15.

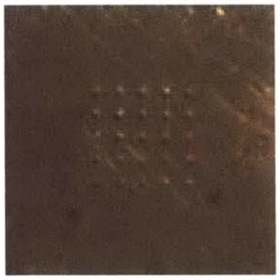


Figure 4-15: Lens mold array used for the slated grating lens experiment. This mold was made by milling lens wells using a ball end mill, and then vapor polishing to smooth the well surface.

This series of grating lenses was recorded using the SM-532TR photopolymer used in the holographic lift-off experiment described earlier in this chapter, and a similar hologram recording geometry. These lens wells were designed to have a much shorter focal length than the previous experiment, on the order of a few millimeters, and were much smaller in width than in the previous experiment. To increase the diffraction efficiency, a spacer was used to create a structure with a $20 - 50\mu m$ grating followed by lens profile. To examine the image plane of this lens the imaging system in 4-16 was built. This system was used to project an image of the focal plane to a screen, shown in Figure 4-17.

The resulting slanted grating lenses exhibited much better focusing than in the normal grating case, but were not thick enough to limit the diffraction to just the first order and undiffracted beam. In 4-17b several diffraction orders are visible. Increasing the thickness of the grating portion of the structure should increase the diffraction efficiency of the structure and eliminate all but the first order and undiffracted light from the output.

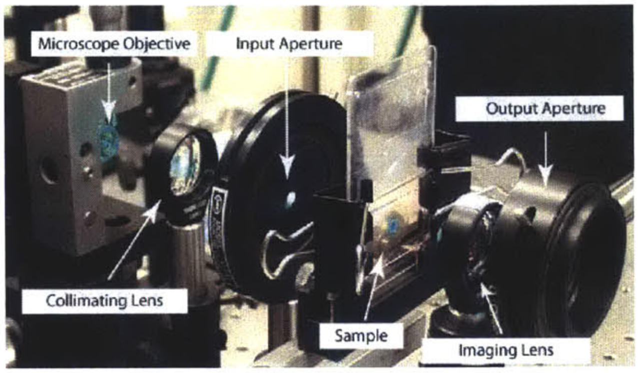


Figure 4-16: Bead Lens imaging system.

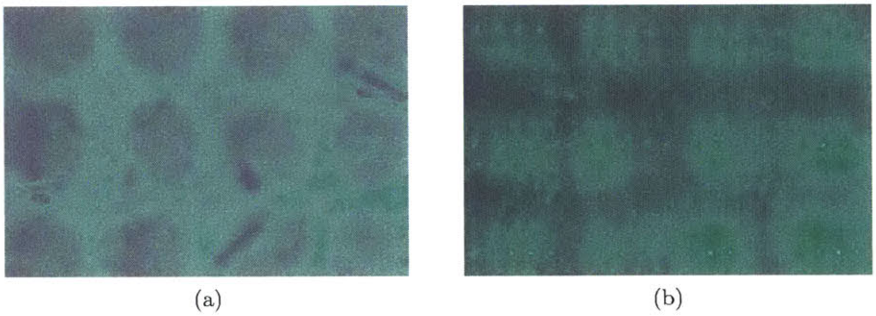


Figure 4-17: (a) Imaging the plane containing the lenses, seen here as dark circles. (b) Imaging the focal plane showing several diffracting orders.

4.2.4 Summary, Holographic Bead Lenses

Bead holographic lenses offer an interesting alternative to traditional planar holograms. By separating the focusing and steering properties of the element they offer a work around to the change of wavelength problem discussed in the previous chapter, since they only require a holographic plane grating rather than a focusing holographic optical element. Example holographic bead lenses have been fabricated for this thesis which exhibit the focusing and steering behaviors expected, but have much lower diffraction efficiency than desired. Improvements in the molding process, and redesign of the elements to use larger plane grating sections, should result in an improved behavior.

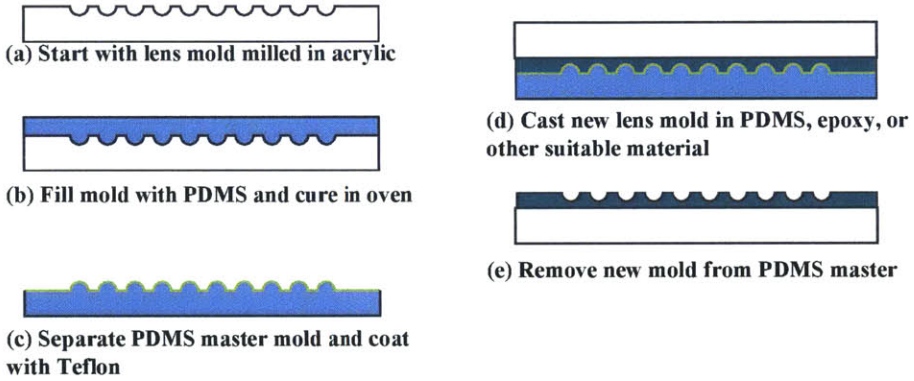
4.3 Lens Well Fabrication

Fabrication of the bead lenses described in the previous section required lens wells which could serve as molds for the bead lenses. Since there were no readily available lens wells some techniques were developed to fabricate the lens wells needed for the experiments. The techniques used to create the lens wells include the use of partially filled SU-8 photoresist rings, grayscale lithographic techniques, PDMS replication of existing lenses, and an acrylic machining and vapor polishing technique. The acrylic machining and vapor polishing technique was developed outside the group, but produced the best lens molds currently in use. This section will discuss the techniques used and some possibilities for future lens well fabrication.

4.3.1 Machined Lens Molds

Initial work on the holographic bead lens project used large glass well slides purchased from a laboratory supplier. The depth of these slides proved to be too large to create a good hologram. Smaller lens wells were created with help from Professor Rajeev Ram's group at MIT. These wells were created by machining wells in optically clear acrylic using a ball end mill and then vapor polishing the resulting surface. The surface profile of the resulting well is determined by the radius of the ball end mills. End mill radii of 0.005", 0.02", and 0.04" were used during these experiments. Control over the depth of the lens was limited by the accuracy in determining the height of the acrylic surface relative to the end of the mill, but the goal was to create lens wells roughly $10 - 20\mu m$ deep.

The plastic lens molds were easily scratched due to both the softness of the acrylic, and because the recording material seems to slightly dissolve the plastic. After a few uses it became impossible to separate the lenslets from the molds without damage. To create a more robust lens mold the 0.005" lens mold was replicated using PDMS, using a process diagrammed in Figure 4-18. The lens molds associated with the 0.02" and 0.04" radius ball end mills were replicated using a similar process, but omitting the Teflon coating indicated in stage (c) of the diagram. These were duplicated into clear epoxy, which does not stick to the PDMS even without the Teflon coating.



(a) Start with lens mold milled in acrylic

(b) Fill mold with PDMS and cure in oven

(c) Separate PDMS master mold and coat with Teflon

(d) Cast new lens mold in PDMS, epoxy, or other suitable material

(e) Remove new mold from PDMS master

Figure 4-18: PDMS replication of lens mold. This technique was developed by Professor Ram's Physical Optics Group at MIT.

Of the lens molds investigated, the machined plastic molds were by far the easiest to use. Although they quickly wore out, they were physically strong and optically clear when new. The PDMS replicas of these molds were easy to separate during the de-molding process, but were physically soft and required additional support during hologram recording and development. Although the machine lens molds proved the easiest to use, other techniques were used to create lens molds with more flexible surface profiles.

4.3.2 Partially filled SU-8 Rings

One approach to creating a well shape is to partially fill a small cylinder with liquid and use the resulting meniscus. This approach was motivated by the technique developed by E-H Park et al [29]. In their paper, wells were formed and then partially filled or overfilled using a micro-injector and tapered fiber. The approach described here was to form the wells using standard photolithography and SU-8 photoresist, and then spin on additional photoresist to partially fill the wells. The experimental work described in this section was performed in the Group 83 clean-room facility at MIT Lincoln Laboratory. The technique is illustrated in Figure 4-19. The goal was to use the spin speed to control the amount of material deposited in the wells.

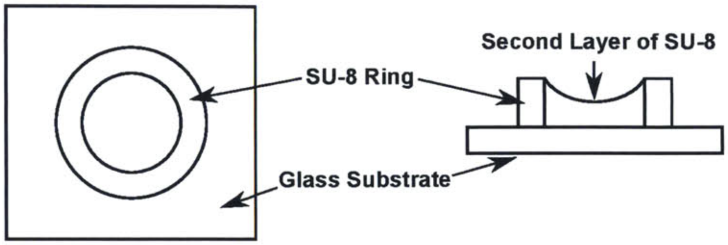
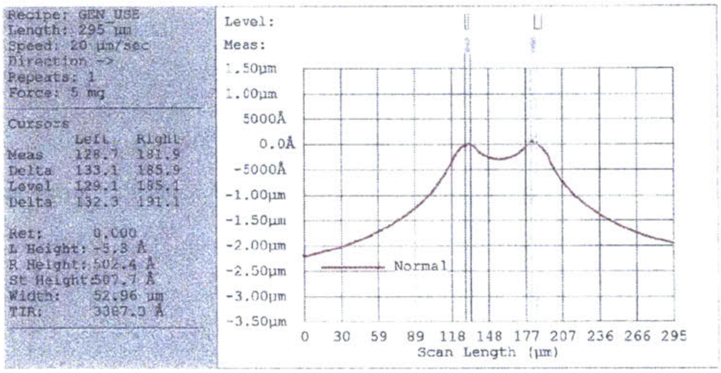


Figure 4-19: Top and Side View of a ring used to form an SU-8 lens well

The SU-8 rings used were 50µm to 200µm in diameter, and roughly 8µm tall. After developing the rings an additional layer of SU-8 was spun on the substrate at a speed chosen to give a layer roughly 4µm thick. The 50µm rings showed the most repeatable pattern, but the lenses formed were generally less than $0.5\mu m$ deep. Assuming a spherical surface profile this represents a surface radius of 625µm, which translates to a 1.25mm focal length for a lens made using this mold with an index of 1.5. A characteristic profile of the mold is shown in Figure 4-20.



```
Recipe: GEN_USE
Length: 295 µm
Speed: 20 µm/sec
Direction ->
Repeats: 1
Force: 5 mg

Cursors
        Left   Right
Meas    128.7  181.9
Delta   133.1  185.9
Level   129.1  185.1
Delta   132.3  191.1

Ref:         0.000
L Height: -5.3 Å
R Height: 502.4 Å
St Height:507.7 Å
Width:       52.96 µm
TIR:         3387.3 Å
```

Figure 4-20: Profile data for a characteristic lens well formed using the "Partially Filled Ring" technique.

Changing the spin speed did result in slightly different profiles. The lens molds formed using this technique exhibited some serious drawbacks including limited control over the focal length and large changes in resist level outside the rings (the slope from the ring edge to surface level). Ideally the lens mold would be flat except where the lens wells were located, so this approach was abandoned.

4.3.3 Grayscale Lithography using Gaussian Beams

Grayscale lithographic take advantage of the fact that resist thickness in any area after processing is dependent on the exposure energy applied to the area. In traditional lithographic techniques a layer of photo-resist material is spun onto a substrate giving a layer of near constant thickness. A photo-mask or similar technique is used to control exposure of the resist with the goal of completely exposing some areas and not exposing others. When using a positive resist the completely exposed areas will have no remaining resist after processing, and areas that were not exposed will be left with a full thickness layer of resist. The energy required for a complete exposure of the resist depends on the individual resist and the resist thickness. If the exposure energy is not high enough there will be resist remaining after processing.

The grayscale exposure technique takes advantage of the fact that a partial exposure results in a partial removal of resist. By controlling the exposure energy across the area of the mask it is possible to generate a two dimensional variation of resist thickness. A single exposure grayscale mask can be fabricated by controlling chrome thickness across the mask or by changing the density of small apertures in the chrome, much smaller than the desired features, to control the average exposure in a given area. Both of these techniques will be limited to creating features much larger than the mask writing resolution.

It is possible to create the gray-scale pattern using the diffraction pattern of UV laser light leaving a single mode optical fiber. The beam exiting a single mode optical fiber is roughly Gaussian so it has a natural intensity variation without requiring a mask. This eliminates the feature size limits imposed by mask writing techniques. Since the beam exiting a single mode fiber is diverging, control of both the fiber-tip to resist surface distance and the exposure time will allow control of the width of the exposed area and depth at the center of the exposed area. The experimental development of this technique was performed with the assistance of Group 83 at MIT Lincoln Laboratory.

Experimental system

The light source used for this experiment was as 405nm laser coupled to to a single mode fiber. The output beam from the single-mode fiber is roughly Gaussian, which is a rough approximation of a parabola near the center of the beam. The mode field diameter for a

405nm single mode fiber is roughly 3.5µm, resulting in a highly diverging beam. Varying the distance from the end of the fiber to the surface of the resist changes the width of the pattern on the resist allowing for control in the lens diameter. Varying the exposure time can control the depth of the of the resulting profile. The two parameters are not completely independent so some iterative testing of exposure and fiber-surface distance is necessary if an exact profile is needed.

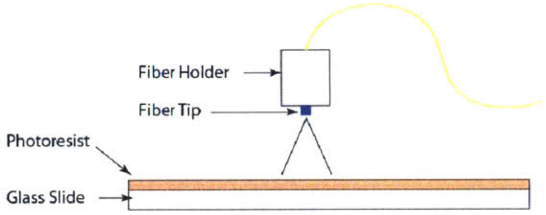


Figure 4-21: Setup for fiber diffraction pattern lens writing

To test a number of exposure energies a lens array writer was built using a pair of translation stages in an x-y configuration to position the glass substrate while the fiber remained stationary. The lens writer is illustrated in Figure 4-21. A glass slide coated with unexposed AZ-1529 was placed on the movable platform and the assembly positioned below a fiber holder mounted on a three axis manual translation stage. The translation stages were controlled by a Labview program which allowed rapid and repeatable changes in position. This combination was used to investigate exposure times from 50ms to 16s at a laser power of 1 μW (measured at the output of the fiber). With a fiber-resist distance under 1mm, an exposure of $300 - 400$ms at a laser power of 1µW resulted in the lens mold profile shown in 4-22. Slightly shorter exposures resulted in shallower wells, but wells shallower than approximately 1µm showed significant surface roughness around and within the well. Slightly longer exposures resulted in deeper wells, up to the depth of the photoresist. Very long exposures resulted in complete removal of a cylinder of photoresist.

The mold was not cross-linked to avoid material reflow. Instead the mold recorded in AZ-1529 was replicated in PDMS using a process similar to the one described for the machined lens molds. To check that the surface profile was maintained during the mold replication process we filled the new PDMS mold with Norland 61 optical adhesive, capped the mold with a glass slide, and then UV cured the adhesive. After curing, the mold-adhesive-glass sandwich was separated leaving the Norland adhesive attached to the glass

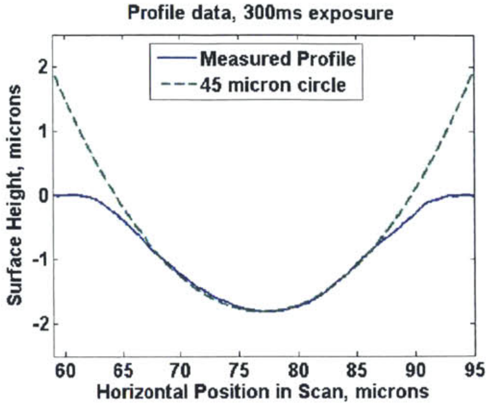


Figure 4-22: Profile data from characteristic lens mold formed using "fiber grayscale" technique

slide. The resulting adhesive-based lenses were imaged with a 10x microscope objective in the system used to image the slanted gratings. The system showed focusing indicating that the process could be used to make lens molds for the holographic bead lens project. Images from the optical cement lens test are given in Figure 4-23.

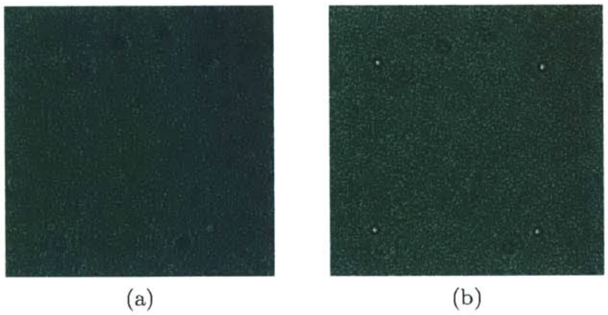


Figure 4-23: Imaging the optical cement lenses created from the AZ-1529 gray-scale process. (a) Imaging the plane containing the lenses, seen here as dark circles. (b) Imaging the focal plane showing the focused spots.

4.4 Holographic Experimental Work Conclusions

As part of the work for this thesis some experimental holographic techniques were developed: holographic lift-off, and the holographic bead lens. The results in this chapter show that these techniques are possible to implement but they will require some refinement before they could be used in a system like the COIN coprocessor. As part of the holographic bead lens project lens molds were fabricated and a new technique for lens mold fabrication using a gray-scale exposure was developed. The results of the fiber based gray-scale exposure show that it is a viable technique to produce lens wells for future use in holographic bead lens experiments.

Chapter 5

Zone Plates for COIN Interconnects

Chapters three and four examined the use of holographic optical elements to create the interconnects in the COIN coprocessor system and identified some of the issues which make fabricating large arrays of holographic optical elements difficult. Holographic optical elements in the COIN network perform the same basic function as traditional optical elements in any system, they transform some input field to some desired output field. Traditional optical elements were not considered during the early work on the COIN system due to the difficulty of fabricating the small structures necessary to make these optical elements at the sizes needed in the COIN coprocessor. Advances in lithographic techniques and a general reduction in the minimum feature size of lithographic masks, have made fabricating very small surface relief structures possible. Moving to a lithographic fabrication process for the optical interconnects will greatly improve the scalability of the COIN system by allowing entire arrays of optical elements to be created with a single exposure step.

As part of the work done for this thesis, the potential use of traditional optical elements was revisited. The requirements of the COIN system are more flexible than those of imaging systems, so optical elements not normally considered for imaging applications were investigated. The most interesting element, and the main topic of this chapter, is a structure similar to a Fresnel Zone Plate. Based on simulation results, the zone plate optical element may be adequate for the needs of the COIN interconnects. This chapter will start with the design of one set of potential COIN interconnect geometries. From the interconnect geometries a set of potential zone plate like optical elements will be developed, and their potential for use in the COIN coprocessor will be evaluated. As part of the work done on this topic a lithographic mask featuring several binary zone plate elements was fabricated. This lithographic mask was used to create phase-modulating zone plates made of photo-resist. The chapter will end with a discussion of that experiment.

5.1 Designing a COIN Interconnect Geometry

In order to design the optical elements for the COIN coprocessor an interconnect geometry describing relative positions of the optical sources, optical elements, and detector elements must be developed. For a nearest neighbor connection scheme there are only three unique optical elements. Since the layers of the COIN are assumed to be planar and the neurons are assumed to be equally spaced, four main design parameters define the connection geometry. These design parameters are the distance between the optical source plane and the optical element plane, the distance between the optical element plane and the detector plane, the optical element width, and the horizontal spacing between neurons. This section will design a COIN interconnect geometry, using 100µm wide optical elements, starting from some assumptions about the optical source being used, and working through the remaining design. The interconnect geometry developed in this section formed the basis for the experimental zone plate described at the end of this chapter, so some of the design choices were made with the goal of creating a structure that could be fabricated and tested.

5.1.1 Modeling the Optical Sources

Knowledge of the optical behavior of the interconnect source, in particular its divergence and power distribution, is needed to determine the relative position of the optical element plane and source plane. At this time the exact behavior of the COIN source light is unknown, because the exact sources have not been determined. In order to make a first attempt at designing optical elements for the COIN some assumption must be made about the sources. For the purposes of this section the source will produce a zero order, circularly symmetric, Gaussian beam at a wavelength of 632.8nm and a beam waist of $\omega_o = 2.3$µm. This set of values was chosen to match the output of a single mode optical fiber operated at the red helium-neon laser line, which was one of the sources used in the experiment described at the end of this chapter. From reference [1], when changed for an assumed time dependence of $e^{-j\omega t}$, the electric field phasor for a circularly symmetric Gaussian beam propagating in the $+\hat{\mathbf{z}}$ direction is

$$\tilde{\mathbf{E}}(x,y,z) = E_o\frac{\omega_o}{\omega(z)}exp\left\{j[kz-\eta(z)] - r^2\left(\frac{1}{\omega^2(z)} - \frac{jk}{2R(z)}\right)\right\}, \tag{5.1}$$

with

$$k = \frac{2\pi}{\lambda}, \tag{5.2}$$

$$R(z) = z\left(1+\frac{z_o^2}{z^2}\right), \tag{5.3}$$

$$\eta(z) = \arctan\left(\frac{z}{z_o}\right), \tag{5.4}$$

$$\omega^2(z) = \omega_o^2\left(1+\frac{z^2}{z_o^2}\right), \text{ and} \tag{5.5}$$

$$z_o = \frac{\pi\omega_o^2 n}{\lambda}. \tag{5.6}$$

The radius of the Gaussian beam, $\omega(z)$, changes with distance from the beam waist. The beam waist is ω_o in the equations above, and is set to occur at $z = 0$. The dimension r is the distance from the axis of propagation, $\hat{\mathbf{z}}$. From (5.1) the field decreases as the observation point moves away from the axis, so the intensity of the beam will also fall as the observation point is moved further from the axis. This expression is for a beam propagating along z, but with a coordinate transform it can be used with any propagation direction.

Power through an aperture

A good optical interconnect element should collect as much of the source light as is practical. Two design parameters, the width of the individual optical elements and the distance between the beam waist and optical element plane, determine how much of the source light is intercepted by the interconnect element. Assuming illumination with a single mode Gaussian Beam, the power passing through an aperture of radius R_a is given by

$$P(R,z) = P_\circ \left[1 - e^{-2R_a^2/\omega^2(z)}\right] \tag{5.7}$$

This is a well known result and is plotted in Figure 5-1. For the square aperture of the COIN optical elements the fraction of the incident power can be found by integrating the intensity over the aperture area. For a radially symmetric zero order Gaussian the intensity at some position (r, z) is given by

$$I(r,z) \propto |E(r,z)|^2 = |E_\circ|^2 \left(\frac{\omega_\circ}{\omega(z)}\right)^2 e^{-2r^2/\omega^2(z)}.$$

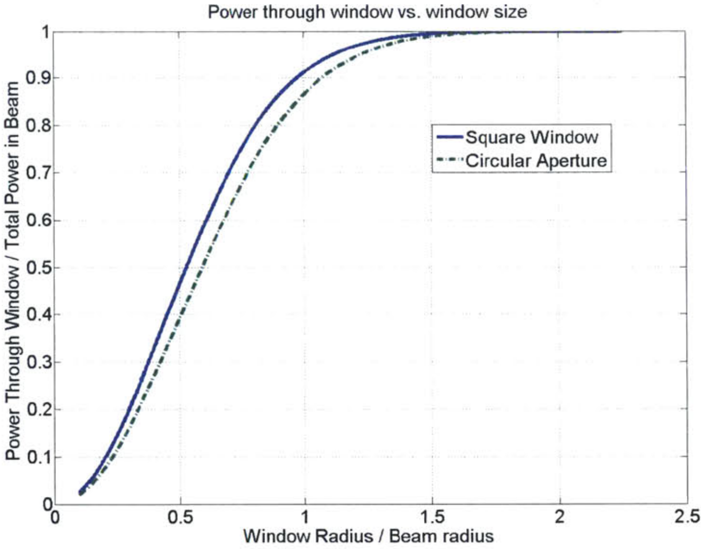


Figure 5-1: Power through a square window or circular aperture. For the square window the "Radius" is half the window width.

From (5.7) it is not possible to capture, with a finite width element, exactly 100% of the beam power because for a theoretical Gaussian beam the field is decaying but still exists out to infinity. Rather than attempt to capture all of the light in this theoretical beam, it is sufficient to design the optical element to capture some large fraction of the light in the theoretical beam. From Figure 5-1, any window size greater than the beam radius will intercept more than 90% of the total power, and any window size greater than about $1.1\omega(z)$ will capture more than 95% of the incident beam. The packing density of the COIN neurons has been assumed to be square, and the optical elements have also been assumed to be square. From 5-1 the square window will capture more power than the largest circular aperture that can be inscribed within the square, as expected. When comparing square apertures and circular apertures that capture the same fraction of the incident power it is important to note than the square window will capture a fraction of this power from regions of lower intensity than the circular aperture. As a compromise, the design for this thesis will place the square window so that the window width is $2.5\omega(z)$ at the optical element plane. With this positioning the maximum intensity at the edge of the optical element will be less then 1% of the peak intensity, and the element will intercept more than 97% of the beam power.

Source to Optical Element Distance

From (5.5) the beam radius varies with the distance from the beam waist. Setting the distance between the source plane and optical elements will set the beam radius at the optical element for a given source. Equation (5.5) can be rearranged to give the source-element distance needed to achieve a given beam radius at the optical elements.

$$z = \frac{\pi\omega_o^2}{\lambda}\sqrt{\left(\frac{\omega(z)}{\omega_o}\right)^2 - 1} \tag{5.8}$$

If the physical width of the optical element is W, and this physical width is related to the beam radius by $W = F\omega(z)$, then (5.8) becomes

$$z = \frac{\pi\omega_o^2}{\lambda}\sqrt{\left(\frac{W}{F\omega_o}\right)^2 - 1} \tag{5.9}$$

Assuming an element width of 100µm, and setting the width equal to $2.5\omega(z)$ as discussed above, the source to element distance for the zero order Gaussian under discussion is then

$$z = \frac{\pi(2.3\times10^{-6})^2}{632.8\times10^{-9}}\sqrt{\left(\frac{100\times10^{-6}}{2.5\times2.3\times10^{-6}}\right)^2 - 1} = 455.8\mu\text{m} \tag{5.10}$$

5.1.2 Off axis steering distances

A nearest neighbor interconnection scheme requires three optical elements: one for forward connections, one for the edge connections, and one for the diagonal connections as illustrated in Figure 5-2. The off axis steering behavior will depend on the relative location of the detectors and connection elements. As indicated in the diagram, it may be desirable to add some empty space between the elements in adjacent neurons or between the optical elements within a given neuron. Allowing some spacing between elements helps to ensure that the power from an optical source only illuminates the optical element directly in front of that source. It may also be the case that some space is required by the requirements of the electronics being used in the system, or the optical sources themselves. Limits to the electronic and source packing may require neurons to be larger than implied by the optical element size.

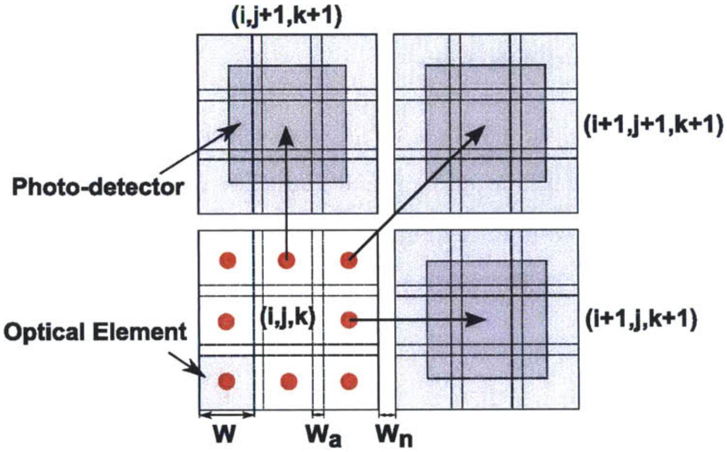


Figure 5-2: Three unique connections for the nearest neighbor connection geometry. Straight ahead (not shown), Edge, and Diagonal. The four "Edge" connections are identical, but rotated in steps of 90 degrees. Similarly the four "diagonal" connections are the same but rotated in steps of 90 degrees.

Let the space between the adjacent optical elements for two neurons be W_n, and let the space between two adjacent optical elements within a neuron be W_a. To hit the center of the detector in the next plane, an edge interconnect must steer its input light two times the individual element width, plus one W_n between neurons and one W_a between the edge and center focusing element of the next neuron. For an optical element width of W, this distance is $2W + W_n + W_a$. Following the same logic, the diagonal interconnection elements must steer their input light off axis a distance of $\sqrt{2}\,(2W + W_e + W_n)$. If W_e and W_n are both set to 10μm, and the element width is set so that at the element plane $W = 2.5\omega(z)$, then the maximum intensity from an adjacent source drops to less then 0.004% of the peak source intensity for the interconnect element being discussed. This results in an off axis steering distance of 220μm for the edge connection elements, and an off axis steering distance of

$220\sqrt{2}$µm for the diagonal connection elements.

5.1.3 Optical Element to Detector Plane Spacing

The remaining design parameter for the interconnect geometry is the distance between the optical element plane and the detector plane, denoted d_f. The off axis steering distance is set by the optical element width and empty space, as discussed above, so the choice of d_f determines the steering angle for the interconnect elements and will indirectly determine what types of output beams can be created. As d_f is reduced the steering angles needed will increase, and the phase difference the optical elements must create will also increase. The choice of this distance is somewhat arbitrary, but will be impacted by fabrication issues discussed in the next section. For this thesis, a value of $d_f = 1.5$mm will be used which allows the zone plate base element discussed in the next section to be fabricated using a standard lithographic mask.

The interconnect geometry is summarized in Table 5.1. Each of the four edge connections can be formed by rotating the one edge connection geometry by multiples of 90°, and similarly each of the four diagonal connections can be formed by rotating the one diagonal connection geometry by multiples of 90°. In the next section the interconnect geometry described in the table will be used to design the connection elements.

Element Type	d_f [mm]	Off Axis Distance X [µm]	Off Axis Distance Y [µm]
Center Connection	1.5	0	0
Edge Connection	1.5	220	0
Diagonal Connection	1.5	220	220

Table 5.1: Possible COIN connection elements, as designed in this section. Each element will have a a width of 100µm, and will take as its input the beam from a single mode fiber operated at 632.8nm, positioned to fill the optical element.

5.2 Designing COIN optical elements

The COIN optical elements will be designed as thin phase only objects. Consider the operation of a basic optical element, for instance the lens in Figure 5-3. The lens can be assumed to be a "thin" optical element if a ray traveling parallel to the optical axis passing through some position (x_1, y_1) in the input plane exits the lens at approximately the same position in the output plane, $(x_2, y_2) \approx (x_1, y_1)$. This is equivalent to neglecting any ray

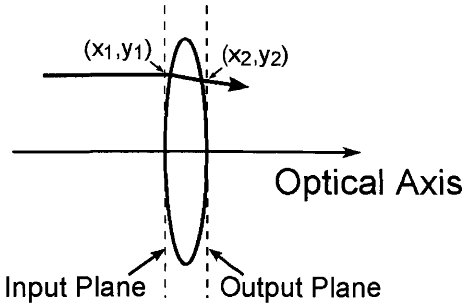


Figure 5-3: A Thin Lens

displacement due to refraction at the lens-air interfaces. Under this assumption, the electric field phasor at the output face of the optical element can be related to the field phasor at the input face by the elements transmission function $\tilde{t}(x, y)$.

$$\tilde{\mathbf{E}}_{out} = \tilde{t}(x, y)\ \tilde{\mathbf{E}}_{in} \tag{5.11}$$

When the input and output fields are known, the desired transmission function can be computed by rearranging (5.11).

$$\tilde{t}(x, y) = \frac{\tilde{\mathbf{E}}_{out}}{\tilde{\mathbf{E}}_{in}} = \frac{|\mathbf{E}_{out}(x, y)|}{|\mathbf{E}_{in}(x, y)|} e^{j(\phi_{out}(x,y) - \phi_{in}(x,y))} \tag{5.12}$$

The COIN optical elements will be designed as phase objects. This means that they will only modulate the phase of the light passing through the element, neglecting the change in amplitude from (5.12). So, the desired transmission function for the COIN optical elements is given by

$$\tilde{t}(x, y) = e^{-j\phi(x,y)} = e^{j(\phi_{out}(x,y) - \phi_{in}(x,y))}, \tag{5.13}$$

where $\phi_{in}(x, y)$ is the phase distribution describing the light from the optical source at the element input plane, and $\phi_{out}(x, y)$ is the phase distribution of describing the phase distribution of the desired output beam evaluated at the plane of the optical element.

5.2.1 Designing the output beam

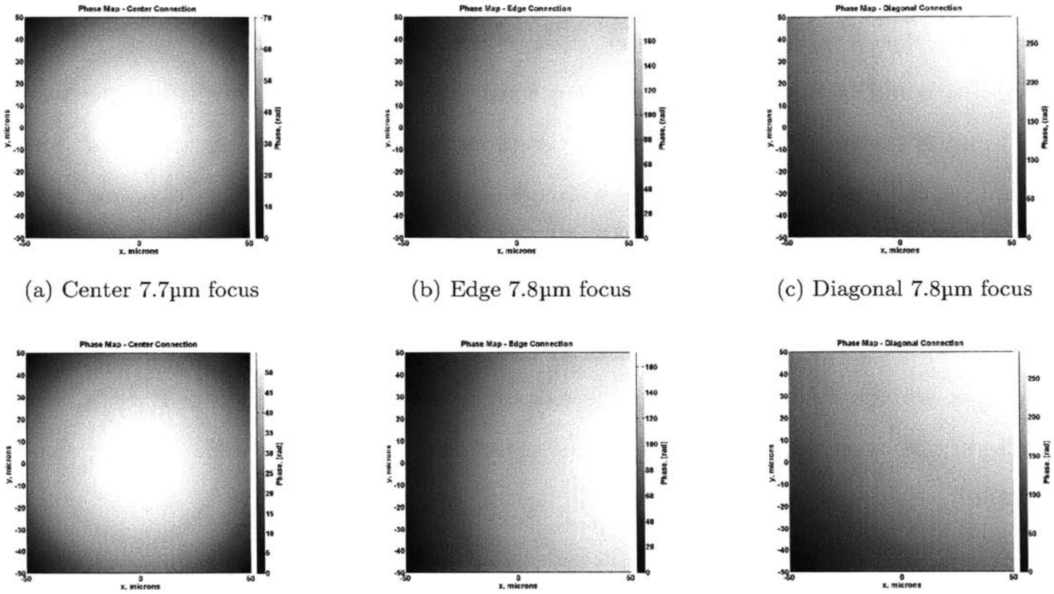
The parameters of the optical element input beam are set by the optical source and the position of the optical element relative to the source. The interconnect geometry discussed in the previous section sets the direction of the desired output beam, but no restriction has been placed on either the beam waist or the position of the beam waist. The optical elements could be designed for any beam waist size and position, but it makes sense to try and match the output field to the input field as much as possible. At the input plane the beam radius is set to $W/2.5$ by the source-element distance chosen, so this can be used as the radius of the output beam at the element plane. To minimize the spread of the beam in the detector plane it makes sense to attempt to focus the output beam at the detector plane. The off axis steering distance and the optical element to detector plane distance together set the distance between the element plane and the detector. Placing the beam waist at the detector, (5.5) can be re-arranged to give the value of the beam waist at the detector.

$$\omega_o^2 = \frac{\omega^2(z)}{2} \pm \sqrt{\frac{\omega^4(z)}{4} - \left(\frac{z\lambda}{\pi n}\right)^2} \tag{5.14}$$

For the chosen interconnect geometry, there are two possible beam waists for each connection type. For the center focusing element, $z = 1.5$mm and the beam could be focused to a spot with $\omega_o = 39.2$µm or $\omega_o = 7.7$µm. For the edge connection the distance between the optical element and detector is $z = 1.516$mm, resulting in spot sizes of $\omega_o = 39.2$µm or $\omega_o = 7.8$µm. For the diagonal connection the distance is $z = 1.532$, resulting in spot sizes of $\omega_o = 39.2$µm or $\omega_o = 7.88$µm. Since the edge and diagonal beams are not propagating perpendicular to the optical element surface, they will not have a circularly symmetric pattern in the optical element plane. To exactly match the fields, the resulting transmission object would have to modulate the intensity of the pattern in addition to the phase.

5.2.2 Phase Profiles and Fabrication Issues

With the beam geometries set it is possible to evaluate (5.13) to find the desired phase transmission function for the COIN optical elements. The ideal phase functions for the $7.7 - 7.8$µm and 39.2µm focal spot elements is shown in Figure 5-4. From the Figure, the range of phase shift that the COIN elements need to achieve is on the order of 100π radians. Rather than the total phase at any point, the important quantity is how the phase varies across the surface of the optical element. These elements are designed to operate at one wavelength, so the phase change can be reduced modulo 2π. The phase wrapped versions



(a) Center 7.7µm focus

(b) Edge 7.8µm focus

(c) Diagonal 7.8µm focus

(d) Center 39.2µm focus

(e) Edge 39.2µm focus

(f) Diagonal 39.2µm focus

Figure 5-4: Ideal phase function for the 100µm wide COIN optical elements. In general there is a smaller range of phase needed for the elements designed for a 39.2µm focal spot.

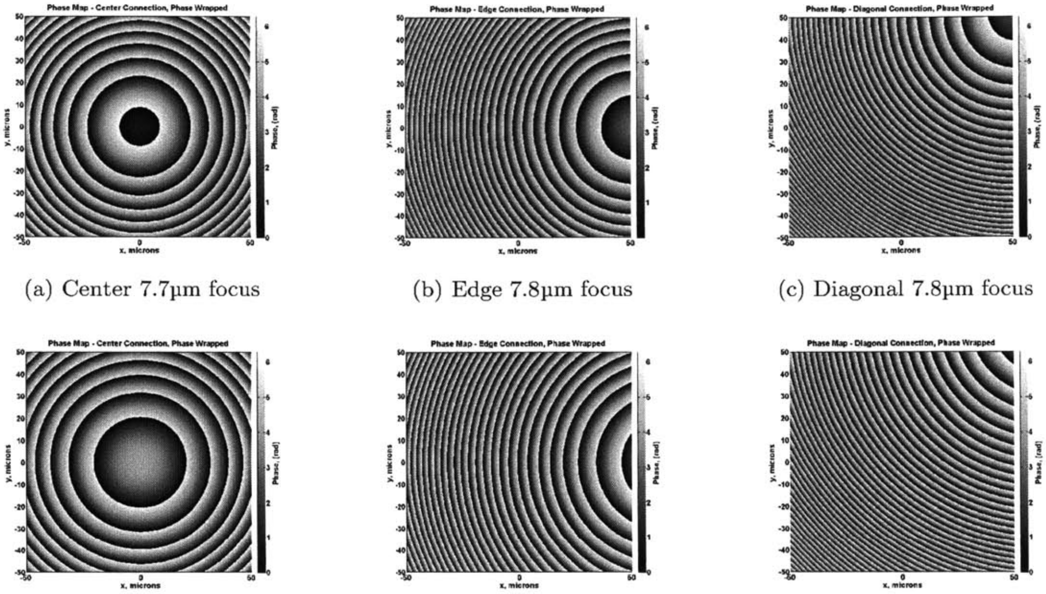
of the phase profiles are shown in Figure 5-5.

An ideal COIN interconnection would be one that exactly recreates either the continuous or wrapped phase profiles, either through a change in surface height as in a traditional lens or through a change in optical index as in a graded index lens. Assuming a material with an index of n_m in air, the surface height variation Δd needed to create a given phase shift $\Delta\phi$ can be computed as

$$\Delta d = \Delta\phi \frac{\lambda}{2\pi(n_m - 1)} \tag{5.15}$$

To achieve a 100π phase shift with a material of index $n_m = 1.5$ operated in air would require a surface height variation of 100λ. For the COIN elements this would require a change in thickness of approximately 63µm across the surface of the 100µm wide optical element. The phase wrapped solution would only require a material thickness range of 2λ, approximately 1.3µm for the COIN elements, which should be easier to achieve.

A technique like gray-scale lithography could be used to fabricate these features. Gray-scale lithography uses a photoresist coated substrate exposed in such a way that the exposure



(a) Center 7.7µm focus

(b) Edge 7.8µm focus

(c) Diagonal 7.8µm focus

(d) Center 39.2µm focus

(e) Edge 39.2µm focus

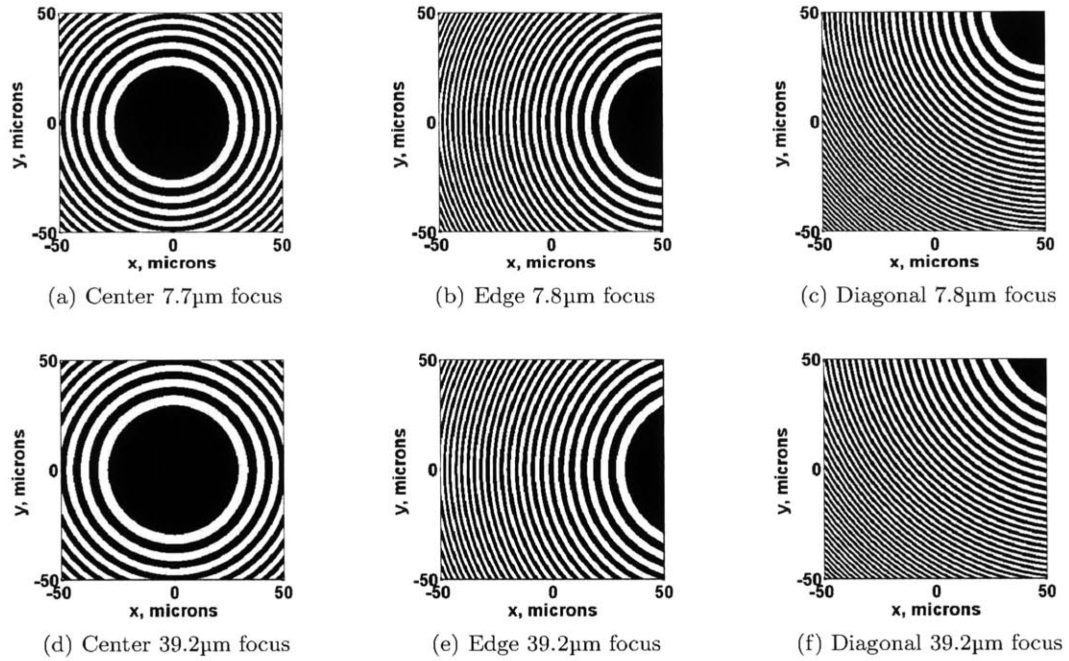
(f) Diagonal 39.2µm focus

Figure 5-5: Wrapped phase function for the 100µm wide COIN optical elements. Since the 39.2µm spot elements have a smaller range of phase, the phase wrapped rings are larger than those of the $7.7 - 7.8$µm spot elements

varies across the surface of the resist. Developing the resist transforms the variation in exposure into a variation in surface height which can either be used as is, or etched to transfer the profile to the substrate. The downside to this technique is that it requires control of the exposure on a scale small enough to recreate the desired exposure profile. This would be difficult to achieve for the sizes of the COIN coprocessor elements. Rather than attempt to match the exact phase profiles, approximate profiles can be used.

5.2.3 Zone Plates

The COIN optical elements do not need to perfectly focus the light from the sources onto the proper detectors, though that would be nice, but rather they need to direct as much light as practical to the detectors while still being easy to fabricate. The actual shape of the focal spot does not matter as long as a large portion of the light from the source eventually reaches the proper detector, ideally while putting as little light as possible onto neighboring detectors. To this end some gross approximations of the phase function were examined which could be created with one or two lithographic steps, using standard binary masks rather than gray-scale masks. Fresnel Zone Plates motivated this work. Fresnel Zone Plates are diffractive optical elements which focus light either by blocking portions of the input wave which would interfere destructively at the focal spot or by shifting the phase of those regions by π so that the interfere constructively at the focal spot. The first elements examined were binary phase plates created from the desired phase transfer function.



(a) Center 7.7µm focus

(b) Edge 7.8µm focus

(c) Diagonal 7.8µm focus

(d) Center 39.2µm focus

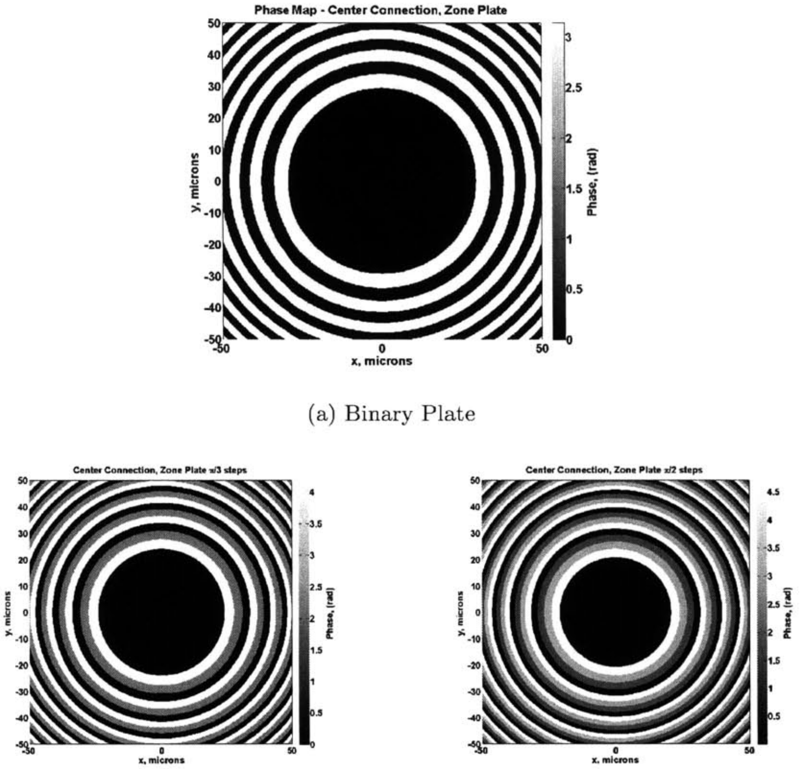
(e) Edge 39.2µm focus

(f) Diagonal 39.2µm focus

Figure 5-6: The phase functions for the 100µm wide COIN optical elements, reduced to zones within which the desired phase function varies by less than π. The two phase levels are zero and π, and the plate will work with either choice of polarity.

The binary phase plate was designed by segmenting the phase function into zones, within which the desired phase change did not vary by more than π. These zones, computed using the phase profiles in Figure 5-4, are shown in Figure 5-6. The width of the zones in the binary phase plate is related to how fast the desired phase function is changing in the area. The phase function for the larger focal spot elements varies more slowly than the phase function for the more tightly focused spots. Additionally, the further away the focal spot is located the slower the desired phase function changes across the element. The distance between the optical element plane and detector plane of 1.5mm was chosen so that for a 100µm wide binary plate, focused to a 10µm spot, would result in zones that were not much smaller than 1µm wide. This choice was made to match the resolution of the features in the zone plate to the resolution of our mask supplier, who's standard process can resolve lines separated by about 1μm.

The binary phase plate can be seen as an extreme example of turning the continuous desired phase variation into a series of discrete steps from zero to 2π. The binary phase plate allows two levels, 0 and π. The approximation is very course, but the binary phase plate could be fabricated with one exposure step in a standard lithographic process. Allowing an additional exposure step allows the approximate phase function to take up to two more values. A version of the zone plate based on three levels (0, $\pi/3$ and $2\pi/3$) and a version based on a four level approximation (0, $\pi/2$, π, $3\pi/2$) for the center connection element is shown in Figure 5-7, for comparison with the binary zone plate.



(a) Binary Plate

(b) $\pi/3$ phase steps

(c) $\pi/2$ phase steps

Figure 5-7: Zone plates based on phase steps of π, $\pi/3$, and $\pi/2$.

Both the $\pi/3$ and $\pi/2$ step zone plates could be created using two lithographic exposures. The difference is that the minimum sized element in the $\pi/3$ step plate will be roughly 2/3 the size of the minimum sized element in the binary plate, while the $\pi/2$ plate would require a minimum element size of roughly 1/2 the minimum element size of the elements in the binary plate. The difference in minimum feature size will make it easier to fabricate the $\pi/3$ step plate than the $\pi/2$ step plate of the same optical element plane to detector plane distance. However, with a minimum mask feature size established the optical element to detector spacing could be adjusted so that all three zone plate options could be fabricated.

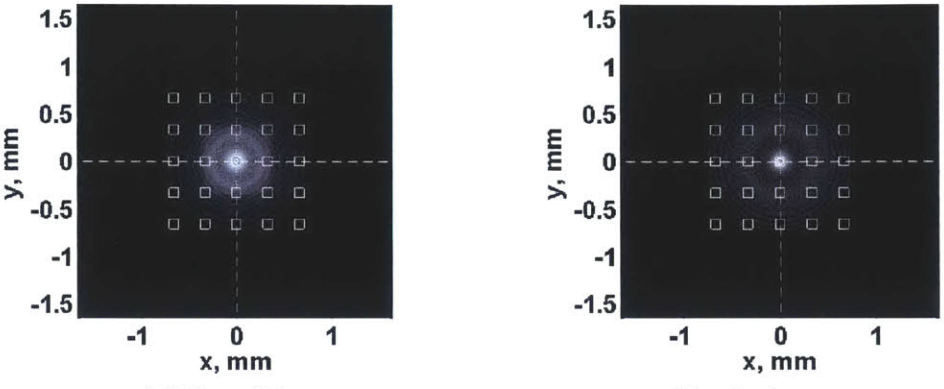
5.3 Evaluation of the Zone Plate Elements

To evaluate the zone plate elements designed in the previous section the expected intensity distribution in the detector plane was computed by numerically evaluating the diffraction integral for the expected phase and amplitude distribution defined by the illuminating Gaussian beam and the phase objects. From the intensity distribution in the detector plane the expected signal levels at each of the 25 next-to-nearest neighbor connections was computed as a function of the total power in the plane.

Based on the results of this simulation, it appears that the zone plate based elements could be used for the COIN coprocessor. The basic zone plate element, with zones set to alternate a phase shift of 0 and π should form connection elements that could be used in the COIN coprocessor. The efficiency of the binary zone plate is lower than either the $\pi/3$ or $\pi/2$ step plate, but can still be made to direct over 40% of the output light to the desired detector for each of the three connection types. The largest disadvantage of these simple elements is that they result in large cross talk terms relative to their signal levels. For example, the binary plate edge connection for the 39.2µm focal spot has a crosstalk term that puts 2.21% of its output light into the wrong neuron. This is especially important as this connection only puts about 42% of its input light to the correct connection. The high level of crosstalk will have to be accounted for when training the neural network.

As expected, moving to either the $\pi/3$ or $\pi/2$ phase step plates results in both higher signal levels at the desired output neuron and lower crosstalk terms. The $\pi/3$ step plate can achieve at least 60% of the output light directed to the proper detector, while pushing the cross talk terms down below 1% of the total output power. The $\pi/2$ step plate can push this minimum efficiency up to 79%, and the cross talk terms even lower.

The next three pages summarize the results of the evaluation for the three connection types. For each connection type the detector plane output is shown for the 39.2µm binary and $\pi/2$ step phase plates. The example diffraction patterns are shown with a log scale intensity to show some of the lower intensity features. Each summary also includes the crosstalk and connection matrix for the best ($\pi/2$ step, 7µm focal spot) and worst (Binary plate, 39.2µm focal spot) elements investigated to give a feel for the range in cross talk and connection efficiency. Finally each summary ends with table comparing the desired connection level and maximum crosstalk level for all six plates examined for each connection.



(a) Binary Plate

(b) $\pi/2$ phase steps

Figure 5-8: Detector Plane for Binary and $\pi/2$ step plates designed for 32.9µm focal spots. The neuron driving the connection is indicated by the cross hairs. The white boxes indicate the location of detectors in the plane.

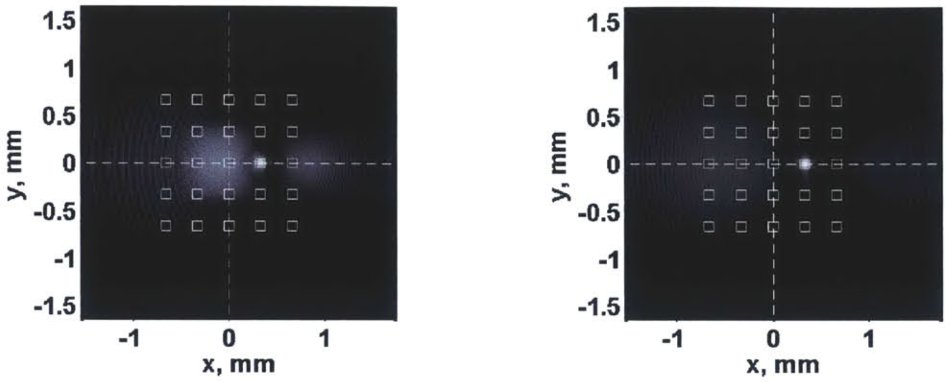
	i-2	i-1	i	i+1	i+2
j+2	0.01	0.01	0.31	0.01	0.01
j+1	0.01	0.07	0.31	0.07	0.01
j	0.31	0.31	31.45	0.31	0.31
j-1	0.01	0.07	0.31	0.07	0.01
j-2	0.01	0.01	0.31	0.01	0.01

	i-2	i-1	i	i+1	i+2
j+2	0.01	0.02	0.03	0.02	0.01
j+1	0.02	0.06	0.07	0.06	0.02
j	0.03	0.07	81.78	0.07	0.03
j-1	0.02	0.06	0.07	0.06	0.02
j-2	0.01	0.02	0.03	0.02	0.01

Table 5.2: Connection and crosstalk for center connection elements, 100µm wide detectors and optical elements. Left: 39.2µm designed focal spot, binary phase plate. Right: 7.7µm designed focal spot, $\pi/2$ phase steps

Plate Type	Focal Spot Size [µm]	Fraction to Desired Connection	Maximum Cross Talk Term
Binary	7.7	42.66	0.45
Binary	39.2	31.45	0.31
$\pi/3$	7.7	70.03	0.24
$\pi/3$	39.2	63.33	0.24
$\pi/2$	7.7	81.78	0.07
$\pi/2$	39.2	79.44	0.11

Table 5.3: Comparison of Center Connection elements. The fraction to desired connection and maximum cross talk terms are given out of 100.



(a) Binary Plate

(b) $\pi/2$ phase steps

Figure 5-9: Detector Plane for Binary and $\pi/2$ step plates designed for 32.9µm focal spots. The neuron driving the connection is indicated by the cross hairs. The white boxes indicate the location of detectors in the plane.

	i-2	i-1	i	i+1	i+2
j+2	0.01	0.01	0.00	0.00	0.00
j+1	0.05	0.17	0.24	0.00	0.00
j	0.07	1.11	2.21	42.66	0.24
j-1	0.05	0.17	0.24	0.00	0.00
j-2	0.01	0.01	0.00	0.00	0.00

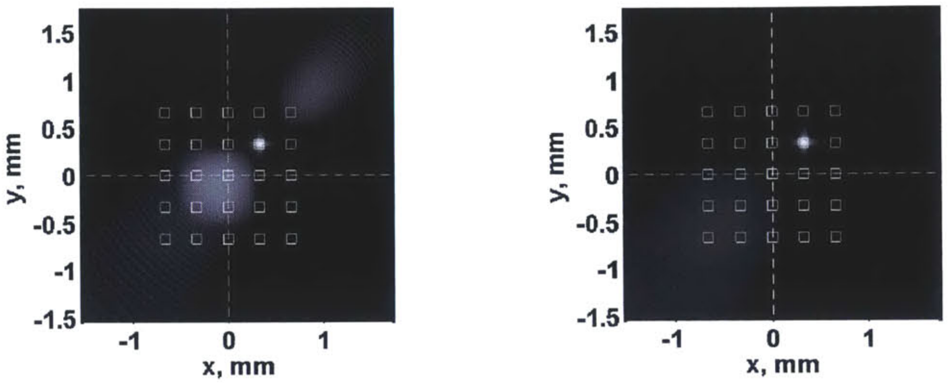
	i-2	i-1	i	i+1	i+2
j+2	0.02	0.02	0.01	0.00	0.00
j+1	0.07	0.06	0.02	0.00	0.01
j	0.10	0.10	0.03	86.58	0.01
j-1	0.07	0.06	0.02	0.00	0.01
j-2	0.02	0.02	0.01	0.00	0.00

Table 5.4: Connection and crosstalk for Edge focusing element, 100µm wide detectors. Left: 39.2µm designed focal spot, binary phase plate. Right: 7.7µm designed focal spot, $\pi/2$ phase steps

Plate Type	Focal Spot Size [µm]	Fraction to Desired Connection	Maximum Cross Talk Term
Binary	7.8	42.55	1.87
Binary	39.2	42.66	2.21
$\pi/3$	7.8	73.07	0.40
$\pi/3$	39.2	72.06	0.56
$\pi/2$	7.8	86.58	0.10
$\pi/2$	39.2	85.12	0.14

Table 5.5: Comparison of Edge Connection elements. The fraction to desired connection and maximum cross talk terms are given out of 100.

Diagonal Connection Elements



(a) Binary Plate

(b) $\pi/2$ phase steps

Figure 5-10: Detector Plane for Binary and $\pi/2$ step plates designed for 32.9μm focal spots. The neuron driving the connection is indicated by the cross hairs. The white boxes indicate the location of detectors in the plane.

	i-2	i-1	i	i+1	i+2
j+2	0.00	0.00	0.00	0.00	0.07
j+1	0.00	0.01	0.01	42.93	0.00
j	0.01	0.83	1.76	0.01	0.00
j-1	0.04	0.52	0.83	0.01	0.00
j-2	0.06	0.04	0.01	0.00	0.00

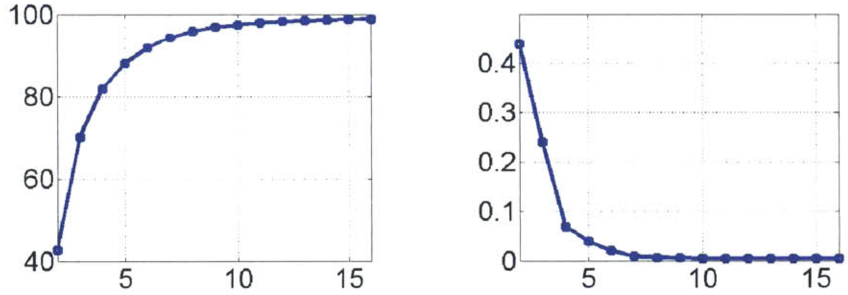
	i-2	i-1	i	i+1	i+2
j+2	0.00	0.00	0.00	0.00	0.00
j+1	0.00	0.00	0.00	89.34	0.00
j	0.03	0.03	0.01	0.00	0.00
j-1	0.07	0.06	0.03	0.00	0.00
j-2	0.08	0.07	0.03	0.00	0.00

Table 5.6: Connection and crosstalk for diagonal connection element, 100μm wide detectors. Left: 39.2μm designed focal spot, binary phase plate. Right: 7.7μm designed focal spot, $\pi/2$ phase steps

Plate Type	Focal Spot Size [μm]	Fraction to Desired Connection	Maximum Cross Talk Term
Binary	7.8	43.81	1.56
Binary	39.2	42.93	1.76
$\pi/3$	7.8	75.03	0.37
$\pi/3$	39.2	73.82	0.49
$\pi/2$	7.8	89.34	0.08
$\pi/2$	39.2	87.27	0.11

Table 5.7: Comparison of Diagonal Connection elements. The fraction to desired connection and maximum cross talk terms are given out of 100.

The analysis above covers the one and two step zone plates currently under consideration. As mask technology changes it may be feasible to include more phase levels than available in a one or two step process. Figure 5-11 shows the predicted connection efficiency and the maximum cross talk term for a center focusing plate fabricated with up to 16 phase levels. The predicted efficiency approaches 100% as the number of phase levels is increased, and the crosstalk terms approach zero. The two step process, allowing up to four phase levels, should produce optical elements that would work in the COIN system. If it becomes feasible to fabricate additional phase levels, the additional control over the phase will allow for even more efficient interconnection elements.



(a) Efficiency as a function of the number of phase levels

(b) Maximum crosstalk term as a function of the number of phase levels

Figure 5-11: Center connection efficiency and crosstalk as a function of the number of phase levels allowed. The binary plates fabricated for this thesis offer two phase levels, the $\pi/3$ step design offers three phase levels, and the $\pi/2$ step design offers four phase levels.

5.4 Experimental Zone Plate result

To test the results of the simulation, a set of zone plate based interconnect elements were designed and a lithographic mask was fabricated. In addition to the COIN interconnect elements discussed in the previous sections, larger interconnect elements with longer focal lengths and a set of interconnects using plane waves as their inputs were also fabricated. The mask was fabricated as a chrome on quartz mask by Benchmark Technologies. A summary of the elements fabricated for test is given in Table 5.8.

Width [μm]	Along Axis Distance [mm]	Off Axis Distance [μm]	Input Beam Type
100	1.5	0 220 $220\sqrt{2}$	Gaussian
200	6	0 440 $440\sqrt{2}$	Gaussian
300	6	0 660 $660\sqrt{2}$	Plane Wave

Table 5.8: Fabricated Zone Plate Elements. To test the zone plane concept a lithographic mask was fabricated with the zone plates designed for the nine geometries in the table. For each element width there is a center, edge, and diagonal connection element. The 100 μm and 200 μm wide elements were designed to use a Gaussian beam from a single mode fiber as their inputs. All elements were designed for a 10μm focal spot and for use at a wavelength of 632.8nm.

5.4.1 Resist Selection and Processing

The goal of the experiment was to validate the simulation and design approach discussed in the previous sections, so to eliminate the difficulties in etching a glass substrate the experiment was designed to use patterned photoresist as the optical element. An ideal photoresist for this application would be optically clear, physically strong, and have a known refractive index. Initial experiments using the standard photoresist SU-8 were not successful, but good results were obtained using the positive photoresist SPR-700-1. The manufacturer specified index of refraction for this resist around 632.8nm is approximately 1.66, and to achieve a phase change of π between areas with resist and areas without resist would require a resist thickness of

$$d = \Delta\phi \frac{\lambda}{2\pi(n-1)} = \pi \frac{632.8\text{nm}}{2\pi(1.66-1)} = 480\text{nm}. \tag{5.16}$$

Stock SPR-700-1 is too thick to spin coat a layer this thin, so the SPR-700-1 was thinned with cyclopentanone, a solvent commonly used to thin photoresist. A mixture of 30mL SPR-700-1 and 10mL cyclopentanone was prepared and several glass substrates were prepared by spinning on the mixture at various speeds. The resulting resist thickness vs. spin speed curve is shown in the diagram below. Keep in mind that this curve is based on a small number of samples (roughly one sample at each indicated speed) and was only used to get into the right range of resist thickness.

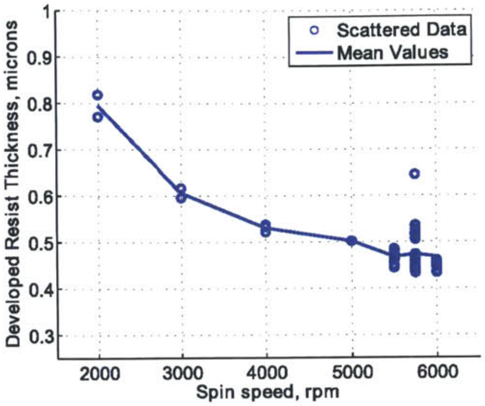


Figure 5-12: Measured resist thickness vs. Spin speed for the thinned SPR-700-1. The solid line follows the mean values, but the scatter plot of all measurements is shown. Notice that at most spin speeds there were only one or two data points taken.

After determining the proper spin speed for the thinned photoresist, a series of test exposures were made with the lithorgraphic mask and the thinned resist using the Microsystems Technology Laboratories contact aligner in the Exploratory Materials Laboratory. The proper exposure time was determined to be between 2.5 and 3 seconds for the patterns on the lithographic mask. If the exposure was reduced to two seconds the pattern did not expose well, and if the exposure was time was increased to 4 seconds scattering at a certain feature size begins to overexpose and degrade parts of the pattern as illustrated in the pictures below.

The resulting process can be summarized as follows

1. On spin coater, mount glass substrate and wash with Acetone and then Isopropanol to clean substrate

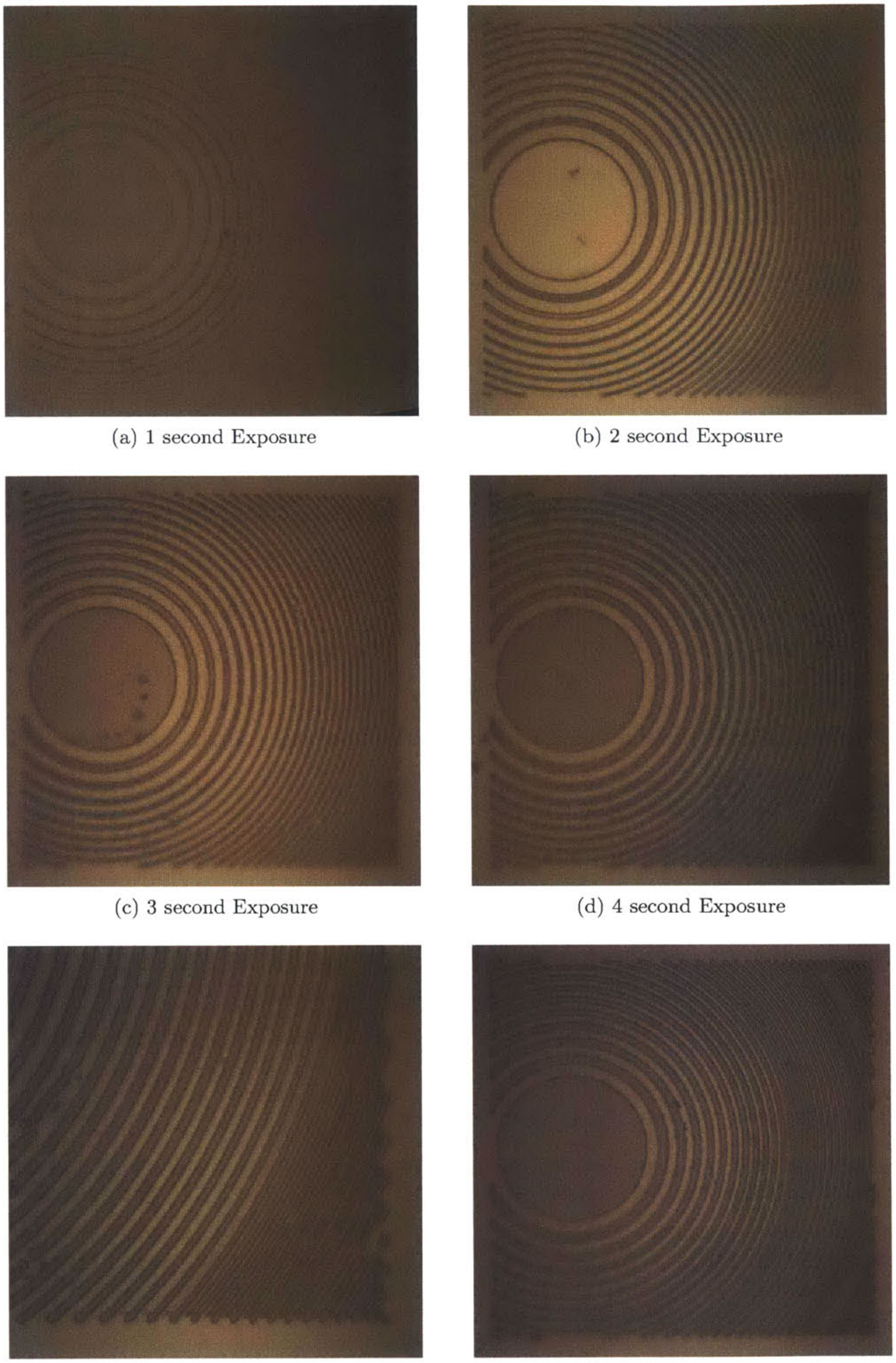
2. Spin on Hexamethyldisilazane (HMDS) as an adhesion promoter

3. Spin on thinned photoresist, 30s at 5750 to 6000 rpm

4. Softbake the substrate and resist for 4 minutes on a hotplate at 115°C

5. Expose for 2.5 to 3 seconds under hard contact

6. Develop in MF-CD-26 developer for 30-60 seconds (until resist stops flowing from surface)



(a) 1 second Exposure

(b) 2 second Exposure

(c) 3 second Exposure

(d) 4 second Exposure

(e) 4 second Exposure View 2

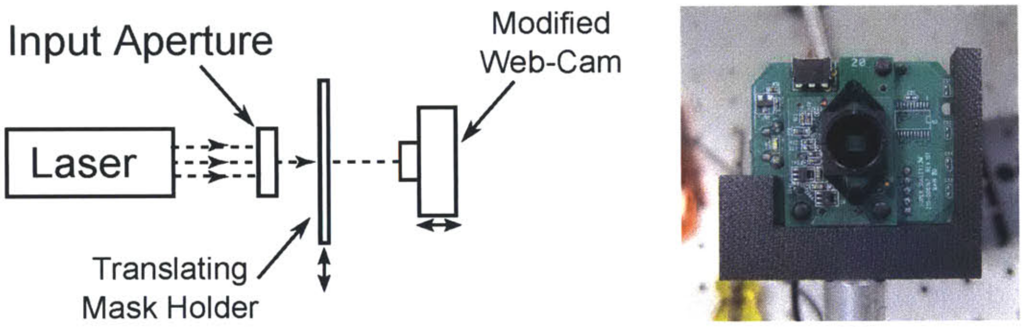
(f) 6 second Exposure

Figure 5-13: Exposure test for the 500nm thick edge elements. Exposures longer than three seconds begin to show some evidence of overexposure. The overexposure is apparent in the corners of the 4-second exposure sample.

Additionally, the best results were obtained when using a round substrate (18mm round glass coverslips) and removing the edge bead during the resist spin step.

5.4.2 Testing the results

The Fresnel Zone concept offers two ways to focus light: as an amplitude plate created by blocking alternating zones, and as a phase plate by changing the phase of alternating zones. The lithographic mask itself is usable as an amplitude zone plate, so the zone plates designed for this thesis can be tested in two ways: (1) using the lithographic mask as an amplitude zone plate, and (2) using the patterned photoresist as a phase zone plate. To test the zone plates an imaging system was built, diagrammed below, to place the sensor array of a web-cam at the approximate focal plane of the optical elements. The web-cam housing was modified to remove all obstructions in front of the camera sensor plane, allowing the sensor plane to be positioned very close to the test system which was necessary to capture the focused spots from the zone plates. This subsection will present the images captured by this system and compare them with the expected patterns from simulation. The first set of zone plates to investigate is the set designed around plane wave illumination.

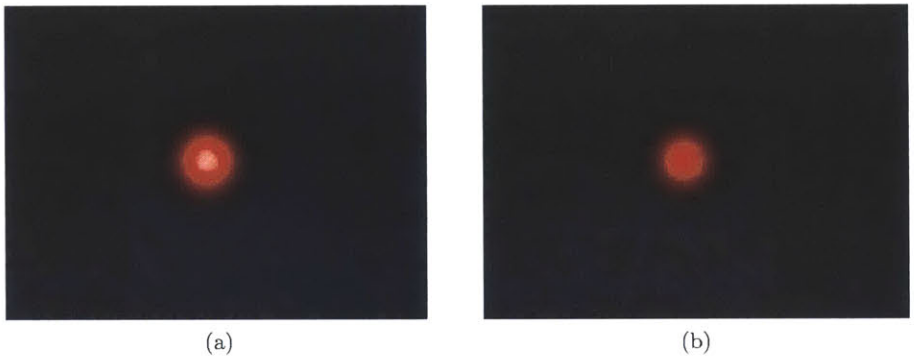


(a)

(b)

Figure 5-14: A diagram of the plane wave test system, and a picture of the modified web-cam used. For the Gaussian beam test system the light from the laser was coupled into a single mode fiber optical fiber, and the optical fiber mounted in a translation stage near the mask holder.

The software used with the web-cam in the test system does not allow access to the raw information from the sensor plane but an attempt was made to calibrate the web-cam reported intensity to actual intensity value. Figure 5-15 shows two images of the expanding beam from the optical fiber used in the Gaussian Beam zone plate tests. The only change made between the two images was positioning a neutral density filter at the output of the laser. The chosen neutral density filter has a transmission of 12.6%. Regions were identified within which the initial image was neither saturated nor too dim, and the red values in these regions were compared between the two images. The web-cam output is not linear with input intensity and it does not appear to be a standard gamma encoding of the form $V_{out} = AV_{in}^{\gamma}$ since the retrieved γ value depends on the limits used to define the regions. As a result, there is no clear translation between the image brightness and the actual optical power other than higher image intensities are associated with higher optical powers. So the images in the comparisons below should be used to develop a feel for the diffraction pattern caused by the zone plates and the positioning and size of the focal spots.

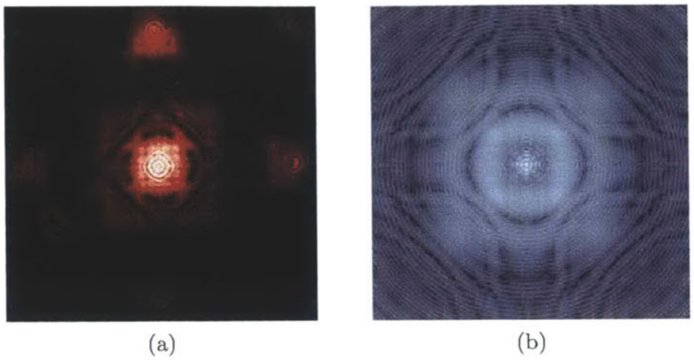

(a)

(b)

Figure 5-15: Web-cam intensity calibration images. The only change between the two images was the addition of a 12.6% transmission filter at the laser for the (b) image.

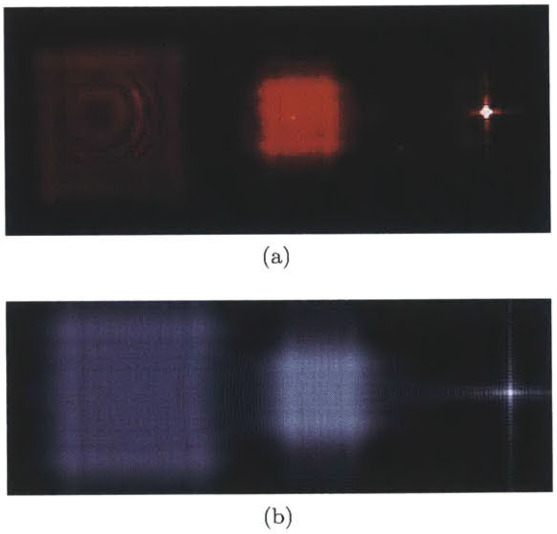
Plane Wave Zone Plates

The Plane Wave Zone Plate elements were tested individually as amplitude zone plates, and in arrays as phase-modulating zone plates. The phase-modulating plates are clear, so testing a single optical element would require masking the area around the zone plate to block out background light. The polarity of the lithographic mask was chosen such that the region around the zone plates is chrome, which blocks out all of the illuminating laser beam except for the portion passing through the apertures defining the zones. This allows the lithographic mask to be used as an amplitude modulating zone plate for testing purposes. As the following images show, the experimental images seem to match the behavior of simulations for the amplitude modulating plane wave zone plates.



(a) (b)

Figure 5-16: The center focusing plane wave input zone plate. (a) Image of focal plane, (b) Logorithmic pcolor plot of the simulation result.



(a)

(b)

Figure 5-17: The edge connection plane wave input zone plate. (a) Image of focal plane, (b) Logorithmic pcolor plot of the simulation result.

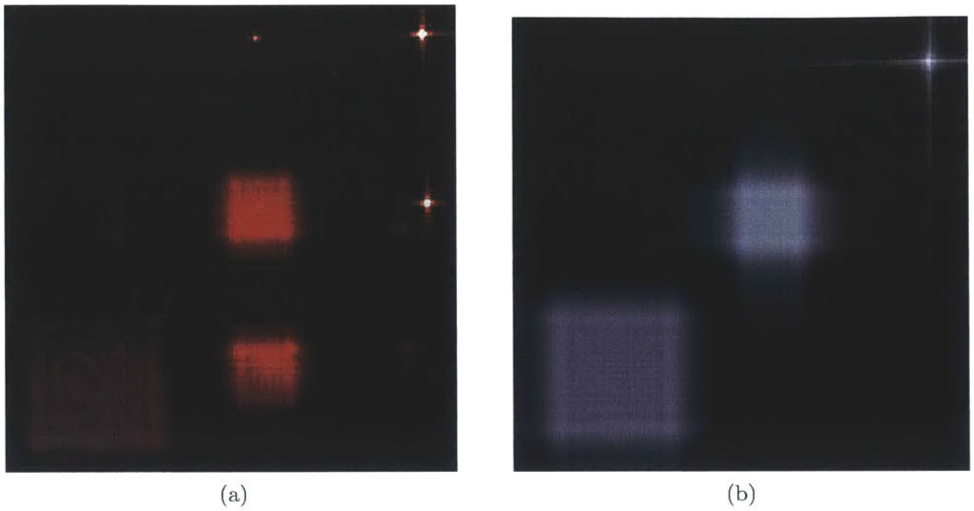
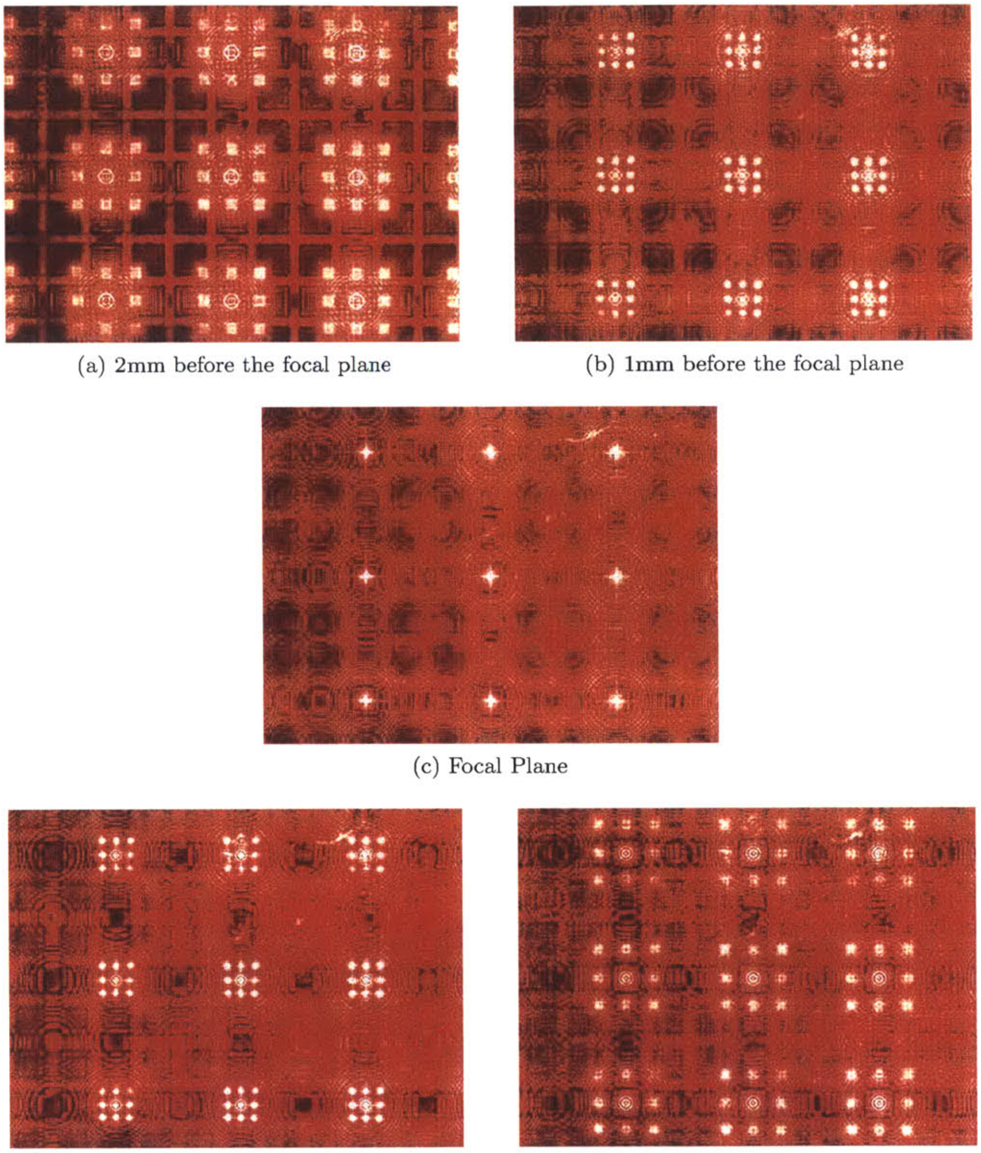


Figure 5-18: The diagonal connection plane wave input zone plate. (a) Image of focal plane, (b) Logorithmic pcolor plot of the simulation result. In the image, the diffracting spot and undiffracted beam from a few adjacent elements are visible because more than one element was illuminated during the test.

The nature of the experimental system is such that exact measurements of the focal plane position are not practical. For each type of input, the lithographic mask also included combined arrays of the three element types as they would exist in a nearest neighbor interconnect scheme. The neuron-array region for the plane wave interconnects was printed in photoresist to create a neuron array phase plate. The following set of images show a scan of the output of this phase plate at positions before, at, and beyond the focal plane of the individual interconnect elements. The focal plane for all of the optical elements occurs at the same distance from the zone plate, and the focal spots for sets of interconnect elements overlap each other showing that they behave as expected from the simulation and design.



(a) 2mm before the focal plane

(b) 1mm before the focal plane

(c) Focal Plane

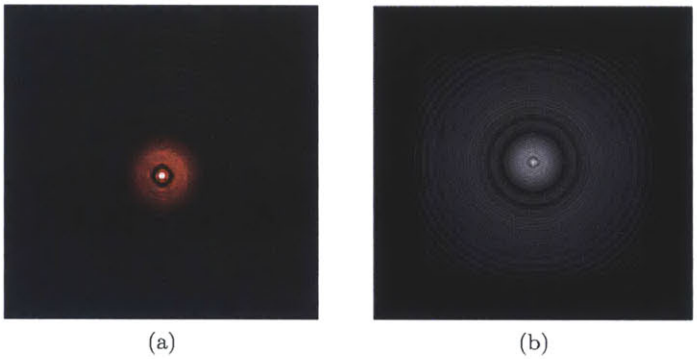
(d) 1mm after the focal plane

(e) 2mm after the focal plane

Figure 5-19: Images from the plane wave neuron array test plate. The images are spaced approximately 1mm apart moving away from the sample. The set of images shows that the light from the different types of zone plates converge to the correct positions in the focal plane, Figure 5-19c. Note that even with the stripped web-cam it was not possible to put the web-cam sensor right at the zone plate plane, so that even by the first image the light directed towards the focusing spots is already outside of its original zone plate region.

Large Gaussian Beam Zone Plates

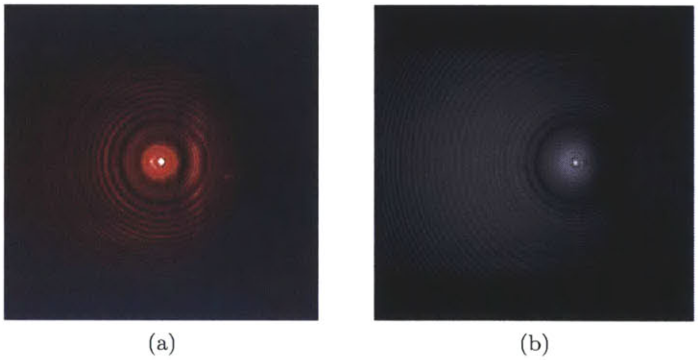
The remaining zone plates were designed around a Gaussian beam illumination profile. This naturally limits the illuminating light to the region of the zone plate making it possible to test individual elements as phase-modulating plates. Due to the short readout condition and the thickness of the lithographic mask it was not possible to image the focal plane of these elements when using the lithographic mask as an amplitude-modulating plate. As with the plane wave test plates, these zone plates show a decent match between the fabricated phase plate and the pattern expected from the simulation.



(a)

(b)

Figure 5-20: The center focusing 200µm wide Gaussian Input Zone Plate. (a) Image of focal plane, (b) Logorithmic pcolor plot of the simulation result. The simulation result assumes the plate was fabricated with a resist of index $n = 1.66$ and thickness $d = 528.3$nm, which approximate the experimental plate.



(a)

(b)

Figure 5-21: The edge connection 200µm wide Gaussian Input Zone Plate.. (a) Image of focal plane, (b) Logorithmic pcolor plot of the simulation result. The simulation result assumes the plate was fabricated with a resist of index $n = 1.66$ and thickness $d = 451.2$nm, which approximate the experimental plate.

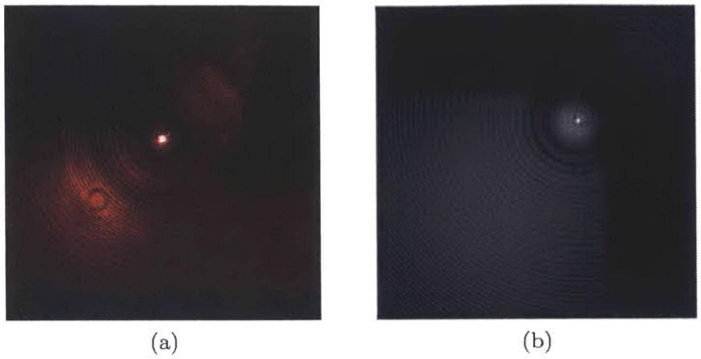

(a) (b)

Figure 5-22: The diagonal connection 200µm wide Gaussian Input Zone Plate. (a) Image of focal plane, (b) Logorithmic pcolor plot of the simulation result. The simulation result assumes the plate was fabricated with a resist of index $n = 1.66$ and thickness $d = 505.8$nm, which approximate the experimental plate.

Small Gaussian Beam Zone Plates

Similar to the large Gaussian Beam test plates, the 100µm wide test plates were tested using photoresist phase plates and illumination from a single mode fiber at 632.8nm. The designed illumination condition requires that the illuminating fiber tip be placed approximately 0.5mm from the zone plate, and the designed focal length for these plates was 1.5mm, with a designed focal spot width of 10µm. To capture these images the zone plate to web-cam distance was set to the designed focal length and then the fiber tip to zone plate was tuned to give a good focal spot. This set of zone plates does not match the simulated versions as well as the large Gaussian zone plates in the previous section. Since the change in feature size between the two is not particularly large, most of this error is probably due to the difficulty of achieving the correct positions of the plate, illuminating fiber, and web-cam.

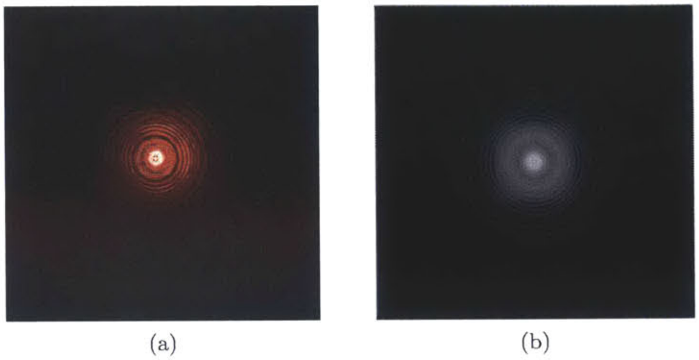


Figure 5-23: The center focusing 100µm wide Gaussian Input Zone Plate. (a) Image of focal plane, (b) Logorithmic pcolor plot of the simulation result. The simulation result assumes the plate was fabricated with a resist of index $n = 1.66$ and thickness $d = 467.7$nm, which approximate the experimental plate.

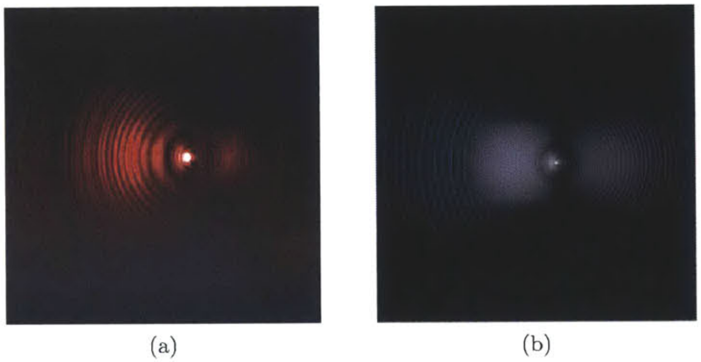


Figure 5-24: The edge connection 100µm wide Gaussian Input Zone Plate. (a) Image of focal plane, (b) Logorithmic pcolor plot of the simulation result. The simulation result assumes the plate was fabricated with a resist of index $n = 1.66$ and thickness $d = 469.0$nm, which approximate the experimental plate.

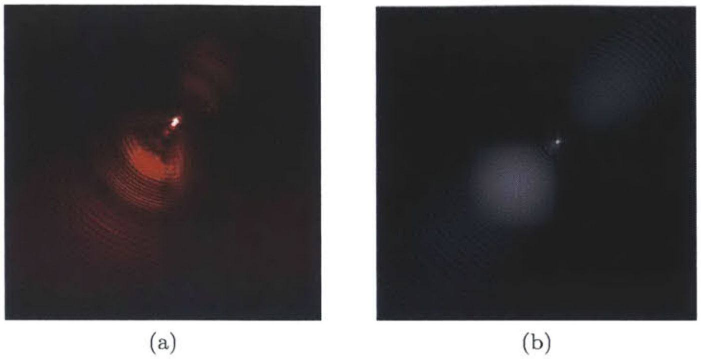
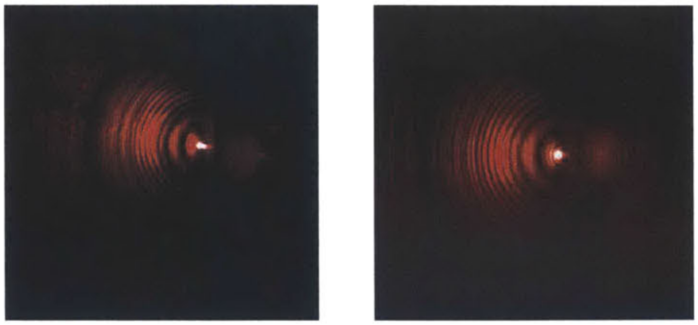


Figure 5-25: The diagonal connection 100µm wide Gaussian Input Zone Plate. (a) Image of focal plane, (b) Logorithmic pcolor plot of the simulation result. The simulation result assumes the plate was fabricated with a resist of index $n = 1.66$ and thickness $d = 463.2\text{nm}$, which approximate the experimental plate.

5.5 LED Illumination

The design of the zone plates in this chapter assumed monochromatic light. This may be an acceptable approximation for laser light but it is not a great approximation for light from LEDs which tend to emit a wider spectrum of light than lasers. Since the COIN system may use LED based light sources rather than laser sources a brief test was made to see if the zone plate elements fabricated for this thesis would work with an LED light source. To test the elements a standard red LED from Radio Shack, with a peak emission wavelength of 660nm was coupled into the same single mode fiber used for the testing shown in the previous subsections and used to illuminate the optical elements. An image of the 100µm wide edge focusing element under LED illumination is given in Figure 5-26a. The same element illuminated by laser light is shown in Figure 5-26b (the same image is used Figure 5-24a). Although the LED wavelength is very different than the design wavelength, and the emission bandwidth is wider than the laser bandwidth, the optical element still focuses the input light. Under LED illumination the focal spot appears to be slightly elongated and it appears that the non-focal-spot light is slightly brighter.



(a) LED Illumination

(b) HeNe Illumination

Figure 5-26: The edge connection 100µm wide Gaussian Input Zone Plate. (a) Image of the focal plane for 660nm LED illumination. (b) Image of focal plane for Helium Neon laser illumination.

5.6 Summary of Zone Plate work

In this chapter a zone plate based approach to optical element design for the COIN coprocessor was developed. The zone plates can be fabricated using standard lithographic techniques widely used in industry, greatly simplifying the process of scaling the number of neurons in the COIN coprocessor. A lithographic mask was fabricated to allow the creation of the binary version of the zone plates designed in this chapter. The lithographic mask was used directly as an amplitude modulating zone plate to test the zone plate designs based around plane wave input light. To test the Gaussian Beam based zone plates the lithographic mask was used to print phase-modulating plates in photoresist. The focal plane images show a good match between the simulations used to design the zone plates and the experimentally fabricated plates.

Binary phase modulating zone plates do not seem to be a great choice for the COIN coprocessor. While they may work, they are not very efficient at directing light from the sources to the detector elements. A better approach would be to modify the designs developed in this thesis to allow phase steps of either $\pi/3$ or $\pi/2$. Both the $\pi/3$ and $\pi/2$ phase step plates could be fabricated using two exposure steps, with each exposure step using a different lithographic mask. This process would require masks with features finer than required for the binary zone plates and accurate alignment of the masks between exposures, but it would greatly increase the fraction of light directed to the correct detector elements and would almost eliminate the crosstalk terms due to stray light.

Chapter 6

Anticipated COIN system Performance: Comparison with other systems

There are several ways to create an artificial neural network in addition to the COIN system being developed within the Photonic Systems Group. Application Specific Integrated Circuit (ASIC) chips have been built which implement neural network functions, off the shelf Field Programmable Gate Arrays (FPGAs) can be configured to operate as a neural network, and networks can always be simulated on traditional computers or on graphics processing units (GPUs) within a traditional desktop personal computer. Building a COIN coprocessor is not a simple task, so it is worth asking "In what ways can a COIN coprocessor compete with an alternate neural network implementation?" Answering that question requires builiding a model of a COIN coprocessor based on the results of prior work and the optics designed for this thesis. At the end of the work done for this thesis, a COIN network could

- use PIN photodiodes operated under a modest reverse bias as the photosensitive elements,
- use resonant cavity OLED source elements operating at the red end of the visible wavelength spectrum,
- use phase modulating Zone Plates to steer and focus the light from the sources,

- and possibly use a negative weight scheme requiring only one optical channel.

With these design choices in mind it is possible to predict some of the operating parameters of a COIN coprocessor including network cost, power consumption, and operating frequency. Potential areas the COIN coprocessor may offer an improvement over the alternative network implementations include the number of neurons available in the network, the speed of a the COIN network in processing information, the cost of the network, and the power consumption of the network. The competing network options operate with a range of network sizes, topologies, and clock cycles. In order to fairly compare the COIN coprocessor with the other network implementations this section will develop a model to calculate the cost, power, and size of various COIN networks based on the number and arrangement of neurons in a network. This model will then be used to generate the parameters of roughly equivalent COIN networks for comparison with the ASIC, FPGA, and GPU based network options.

6.1 COIN Coprocessor

The COIN coprocessor will consist of two separate systems: a hybrid optical-electronic neural network and a network control and interface device to allow data, weights, and network inputs to be loaded to the network. The exact nature of the network controller will depend on the network size and how the network will be used. A neural network with a small number of neurons, in an application with a low frequency of inputs, could use a simple micro-controller as the network control. As the number of neurons in the network grows or the rate at which the network must process inputs or update weights increases, a more complicated network controller will become necessary. This thesis will assume the network controller will take the form of a single board computer, similar to the BeagleBone Black, which runs a Texas Instruments ARM Cortex-A8 processor. The current Manufacturer's Suggested Retail Price (MSRP) for the BeagleBone Black is $45, it may draw up to 5W of electrical power under heavy activity, and has a physical size of roughly 87mm by 54mm with a height less then 10mm when stripped of header connections and jacks. This should be overkill for most operations so the physical size, cost, and power consumption computed below are conservative estimates of the COIN network parameters.

6.1.1 COIN Size

An estimate of the size and cost of the physical neural network portion of the coprocessor can be made by examining the size of the optical elements and the cost of the neural circuit fabrication. Based on the results in Chapter 5 of this thesis, it is possible to create optical interconnect elements roughly 100µm square using standard lithographic techniques. Allowing for 10µm dead space between adjacent optical elements within a neuron, and 10µm of dead space between neurons, a COIN neuron will occupy an area 330µm wide by 330µm tall (320µm for the optical elements plus 10µm dead space between neurons). Travis Simpkins's work on the circuits for the COIN coprocessor (which was called the CONCOP at that time) produced neural logic circuits with an area of approximately 0.26mm × 0.26mm per neuron [14, p.120], slightly smaller than the size indicated by the optical elements developed for this thesis. Figure 6-1 shows the in-plane layout of the COIN neuron optical elements. The photodetector and neuron circuitry are omitted for simplicity.

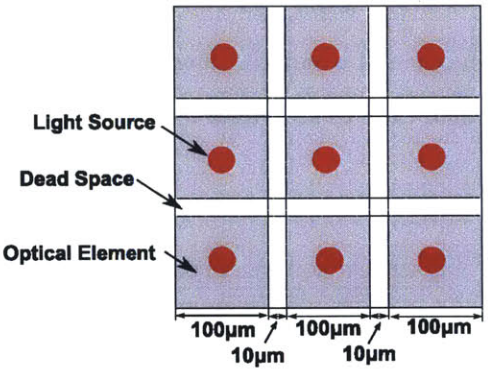


Figure 6-1: In plane layout of the COIN neuron showing the optical element width and dead space between adjacent optical elements. The effective width of a neuron will also include any deadspace between neurons.

Based on the results in Chapter 5 a light-source to optical-element distance of 0.5mm and an optical-element to light sensor distance of 1.5mm are reasonable. Assuming the COIN detector/ logic layer and optical source layer are both fabricated on 500µm thick substrates, then a single COIN layer is roughly 3mm thick. A side view of the COIN neuron showing the relative dimensions of the layers is given in Figure 6-2. From this analysis, a network of X layers and M×N neurons per layer, the physical size of the neural network portion of a COIN coprocessor will be 0.33M × 0.33N × 3X [mm]. The next parameter to examine is the cost of a COIN network.

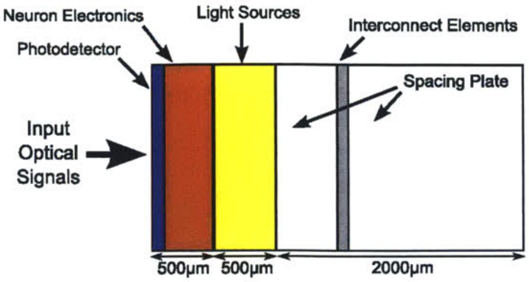


Figure 6-2: A side view of a COIN neuron showing the relative sizes of the different components. The photodetectors and COIN neuron electronics are fabricated in one 500μm thick silicon wafer even though the diagram seems to indicate two materials.

6.1.2 COIN Cost

Estimating the cost of the COIN network requires some estimate for the number of neurons that can be fabricated on a given wafer and the cost of the wafers. Assuming a square network, the number of neurons per wafer is set by the integer number of neurons that can fit along one edge of the square inscribed on the round wafer.

$$\#\text{ Neurons / wafer} = \left\lfloor \frac{\text{Inscribed square edge length}}{\text{neuron edge length}} \right\rfloor^2 = \left\lfloor \frac{\text{wafer diameter}/\sqrt{2}}{0.33} \right\rfloor^2$$

Assuming a wafer diameter of 200mm, then the largest square array that could be fabricated on one wafer would be 428 × 428 neurons. If the wafers used for the COIN neural logic circuits can be bought, patterned, and processed at a rate of \$50K per 100 wafers then the electronics for a COIN coprocessor would cost 0.27¢ per neuron with 428^2 neurons per wafer. Assuming a low cost light source and optical elements compatible with standard semiconductor processing techniques, the total cost of the light sources and optical interconnects combined should be approximately the same as the electronics. This would make the COIN neuron price 0.55¢ per neuron before adding the controller needed to interface the neural network with the host computer. Assuming the single board computer discussed earlier could control 100K neurons, then the total cost of a COIN network with N_{total} neurons can be computed by adding the cost of the neurons to the cost of the controller.

$$\begin{aligned}\text{Total Cost} &= \text{Neuron Cost} + \text{Controller Cost}\\ &= (N_{total} \times \$0.0055) + \left(\left\lceil \frac{N_{total}}{100 \times 10^3} \right\rceil \times \$45\right)\end{aligned}$$

This equation gives an estimate of the total cost to manufacture a COIN network. To give a fair comparison with the commercially available products discussed below, this cost will be multiplied by a factor of 5 to allow for some manufacturer profit and retail markup.

$$\text{Total Cost} = \left[(N_{total} \times \$0.0055) + \left(\left\lceil \frac{N_{total}}{100 \times 10^3} \right\rceil \times \$45\right)\right] \times 5$$

6.1.3 COIN Power Consumption

The next parameter for comparison is the power consumption of the COIN coprocessor. In Section 2.1.1 the maximum optical power needed for a connection, as measured at the source, is 180µW. Since each COIN neuron has 9 connections to its nearest neighbors, and assuming each connection is set to half the maximum value on average, then each COIN neuron must supply approximately 810µW of optical power. If the electrical to light conversion efficiency of the sources is 30%, and the neural logic and source driver circuits together use as much electrical power as the light source, then the power consumption of the COIN neurons can be computed as

$$\text{Power Per Neuron} = 2 \times \frac{\text{Optical Power}}{\eta_{opt}} = 2 \times \frac{810\mu\text{W/Neuron}}{0.3} = 5.4\text{mW/Neuron}$$

Using a power budget of 5W for each single board computer the total power used in a COIN system is

$$\begin{aligned}\text{Total Power} &= \text{Neuron Power} + \text{Controller Power}\\ &= (\#\ \text{Neurons} \times 5.4\text{mW}) + \left(\left\lceil \frac{\#\text{Neurons}}{100 \times 10^3} \right\rceil \times 5\text{W}\right)\end{aligned}$$

6.1.4 COIN Speed

The last point of comparison is the network speed which can be expressed as the frequency at which the network processes a layer of neurons or the number of neuron layers the network can process per second. The small size of the photosensor elements, discussed in Section 2.1, will allow them to respond to input changes up to 1GHz so speed of the network will limited to the speed of the neuron logic circuits and light source driver circuits. A modest goal for the COIN neural logic circuits would be 50MHz operation, meaning the logic circuits would need to be fast enough to respond to an optical input modulated at 50MHz and the COIN coprocessor would be able to process one layer of neurons every 20×10^{-9} seconds.

Using the equations developed in this section it is possible to estimate the parameters of any given COIN network. For each of the alternative network implementations discussed in the next three sections an approximately equivalent COIN network with the same number of neurons, or capable of performing the same network tasks, will be designed and the cost, size, power consumption, and speed of the equivalent COIN will be compared with the alternative network implementation. This approach was chosen to offer a fair comparison of the COIN and the alternative networks, and in general the COIN network should be designed around a specific task.

6.2 Dedicated Neural Network Chips

Application specific integrated circuit (ASIC) chips have been developed which implement neural network behavior. One such ASIC neural network is the CM-1K, a 1024 neuron chip produced by CogniMem technologies (formerly Recognetics) [30]. The chip is available as a standalone device, or on a board with four CM-1Ks for 4096 neurons, and the boards can be stacked in series or parallel to expand the network. The CM-1K operates at 27MHz and it is designed to implement Radial Basis Function or K-Nearest Neighbor classifiers. At $94 per chip, the CM-1K costs about 9¢ per neuron. With a power consumption of 300mW @27 MHz operation, the CM-1K consumes 290µW per neuron. The CM-1K is designed for recognizing patterns up to 256 bytes long and it takes 38 clock cycles, or 1.4×10^{-6} seconds after the input is presented to the chip, for the identification result to be ready [31].

The base hardware of the COIN network is not suited to direct implementation of the Radial Basis Function or K-Nearest Neighbor classifiers, so another approach must be made.

If the 256 byte pattern is encoded as 256 one byte inputs to a COIN network of suitable size, it should be possible to train the network to recognize the input. To load the input into the network the input plane would have to be at least 16 × 16 neurons in size. Since the COIN neurons are not fully connected, it would take 16 layers for for data from the input in one corner to propagate to the opposite corner. If the network were made 32 layers deep then data should be able to propagate from one corner to the opposite corner and back, and the network should be capable of being trained to recognize the 256 byte inputs of the CM-1K. Depending on the application the network may not need all of the layers from this calculation, so it could be better to build a physical system with a smaller number of layers and provide some mechanism to feed back the result of the last layer to the input of the first layer. Assuming a two clock cycle penalty for cycling the outputs from the last layer to the inputs of the first layer, a 50MHz clocked COIN with a 16 × 16 × 4 neural network could identify a 256 byte input in 1.28×10^{-6} seconds.

The CM-1K has 1024 neurons, so an alternate approach would be to make a COIN network with planes of 32×32 neurons and replicate the 16×16 input vector four times onto the input layer of this COIN network. In this case it would take eight layers for data from any one of the 16×16 inputs to reach any other neuron in the network, and a conservative network depth would then be 16 layers. Assuming the full 16 layers of this network were built, the COIN network would be able to identify a 256 byte input in 0.320×10^{-6} seconds, roughly four times faster than the CM-1K. Additionally, the inputs to this network could be pipelined to improve performance. At each clock cycle a new input could be loaded to the first layer of the network and after an initial 16 clock cycles to generate the first classification result an additional classification result would be ready each clock cycle.

A summary of the network parameters of the CM-1K and two competing COIN coprocessor options is given in Table 6.1. From the table it is clear that the CM-1K is smaller, draws less power, and costs less than the two COIN options. However, both competing COIN networks are faster. The COIN network options are also more flexible in that they offer a general neural network system instead of being limited to the two classifier algorithms of the CM-1K. In situations where the total power consumption of the system is critical, and with applications that map well to the CM-1K's classifier algorithms, the CM-1K is the most attractive option. When the application does not map well to the two CM-1K classifier algorithms, or where network speed is more important than power consumption or network cost, one of the COIN coprocessor options should be chosen.

Network Implementation	Total Cost	Input Identification Time [seconds]	Network Size (physical)	Total Power Consumption
Dedicated Network ASIC (CM-1K)	$94	1.4×10^{-6}	32 x 44 x 3 [mm]	0.3 [W]
COIN (16x16x4 network)	$253	0.960×10^{-6}	5.3 × 5.3 × 12 [mm] +87 × 54 × 10 [mm]	10.5 [W]
COIN (32x32x16 network)	$675	0.320×10^{-6}	10.6 × 10.6 × 48 [mm] +87 × 54 × 10 [mm]	93.5 [W]

Table 6.1: Neural Network ASIC (CM-1K) and equivalent COIN comparison table. The Input Identification Time for the 16×16×4 COIN assumes a penalty of two clock cycles to move the output of the second layer back to the input of the first layer.

6.2.1 FPGA Based Neural Processor

Another option for implementing a neural co-processor would be to simulate the network on a high performance Field Programmable Gate Array. At the time of writing the largest Xilinx FPGA available is the Virtex-7 offering 2 million logic cells each operating a 6-input look up table and running with a system clock of 200MHz [32]. Like the COIN network, the Xilinx Virtex-7 needs some supporting electronics to interface with the outside world. For the purposes of this comparison the assumption will be that the power consumption and physical size of a Virtex-7 in a neural network application will be similar to that of the Xilinx Virtex-7 VC707 evaluation board. The Virtex-7 evaluation board has a physical size of approximately 14 × 27 × 1.5 [cm], and has a power supply rated to deliver about 60W. From the Xilinx benchmark applications, a reasonable power draw is closer to 40W when running at full speed and using a large number of the FPGA's logic cells.

A simple implementation of a neuron would require one 8-bit full adder unit, two 8-bit registers (one for the threshold level, one for the intermediate result), and one 8-bit output register for each connection. Using a basic design the 8-bit full adder would require 4 logic cells, the 8-bit registers would require approximately 1 cell per register, and nearest neighbor connections then the basic neuron would require 15 logic cells. Under these assumptions the Virtex-7 would be capable of simulating 133,333 neurons. These neurons could be arranged into a single layer of 365x365 neurons, but that would require re-loading the FPGA with threshold and weight information every clock cycle. A more interesting arrangement of these neurons would be a 10 layer, 115x115 neuron per layer network. Note that this is a very simplistic assumption about how many gates are needed and how the network should be built in the FPGA, so the actual power of the network may be higher or lower than what was computed here. The current price for the Virtex-7 with approximately 2 million logic cells is $17000, about 13¢ per neuron. Unlike the CM-1K chip, this network would not be fully connected but would have the same connectivity as the COIN system with nearest neighbor connections. Assuming the adder and look up tables operate in one clock cycle, it would take 8 clock cycles to add the 9 inputs, one clock cycle to check the look up table, and one clock cycle to activate the outputs for a total of 10 clock cycles for a neuron to perform its function. This gives a neuron clock rate of 20MHz, or a ten layer processing time of 0.5×10^{-6} seconds.

In this case the equivalent COIN system could be a system with the same number and arrangement of neurons running at 20MHz or it could be a system with fewer neurons running at a higher speed. A five layer, 115x115 neuron per layer, system with a network clock of 50MHz would be able to simulate a 10 layer network with a ten layer processing time of 0.28×10^{-9} seconds. A comparison between the FPGA neural network and the 115x115 neuron by 5 layer COIN is given in Table 6.2. From the comparison table the COIN Network has an advantage in terms of the network cost and speed, but consumes significantly more electrical power than the FPGA based network. The power consumption could be decreased by reducing the number of physical layers at the expense of slowing down the network.

Network Implementation	Total Cost	Ten Layer Processing Time	Network Size (physical)	Total Power Consumption
FPGA Neural Network (Xilinx Virtex-7)	$17000	0.5×10^{-6}	14 x 27 x 1.5 [cm]	40 [W]
COIN (115x115x5 network)	$2043	0.28×10^{-9}	38 × 38 × 15 [mm] +87 × 54 × 10 [mm]	362 [W]

Table 6.2: FPGA (Xilinx Virtex-7) and equivalent COIN Network Comparison Table

6.3 Computer Neural Network Simulation

Traditionally, artificial neural networks have been simulated on computers. Available RAM and processor speed are two important factors in determining how fast a given computer can simulate a neural network. Estimates of both the number of instructions required to replicate the function of one neuron and the memory required to hold the simulation state of the neuron are required to calculate the size of a network that can be implemented on a given computer. One of the goals of the COIN coprocessor project is the development of a compact neural network as a co-processor, so the comparison in this section will be limited to simulation hosted on a graphics processing unit. There would be little point in comparing the COIN coprocessor to the 55 peta-flop Tianhe-2 at the National Supercomputing Center in Guangzhou, China, or the 27 peta-flop Titan at Oak Ridge National Laboratory in the United States. In the last decade performance of computer video cards has been enhanced by the use of large arrays of simple processors in the cards graphics processing unit (GPU). Scientists have been reprogramming these video cards to take advantage of the parallel processor arrays in the GPUs to speed up some types of computation, and manufactures have responded by releasing versions of their video cards built specifically for scientific computing applications. On such card is the Tesla K10 from nVidia. The K20 has a physical size of 11cm x 27cm x 2.5cm and consumes 225 Watts of electrical power.

Casting the operation of a neuron in terms of floating point operations, the operation of a single neuron would require 9 multiply operations to compute its inputs ("What is the weight of this input?" times "Is the neuron associated with this input turned on?"), 8 addition operations to sum the inputs, and one look up table operation to determine its output for a total of 18 operations per neuron. This is just a first approximation, and neglects the instructions needed to move data in memory. The same neuron would require the storage of 9 weights, one threshold, and one output state. So eleven single precision floating point numbers, or 44 bytes of data.

The Tesla K20 GPU can execute approximately 3.5×10^{12} single precision floating point operations per second and has an on-board memory of 5GB [33] and a cost of $3000 (not including host computer). The floating point performance corresponds roughly to computing the output of 1.9×10^{11} neurons every second. Assuming the goal is to create a network with a 50MHz clock, the network would consist of 3800 COIN style neurons. With a memory of 5GB the Tesla K20 could hold the state of roughly 113×10^{6} neurons at once, so the factor limiting performance is the time needed to compute a layers output rather than the memory needed to hold the layers state.

Table 6.3 summarizes the comparison between the GPU based neural network simulation and an equivalent COIN network. When comparing the GPU to a COIN of the same total number of neurons, the COIN outperforms the GPU based network simulation in both the total cost and power consumption. The two networks have the same clock speed by design. One potential advantage to the GPU based simulation is the flexibility to re-organize the network structure. With the GPU simulated network the number of neurons per layer and number of layers can be changed so long as the total number of neurons is lower than 3800, keeping the same network speed, or neural networks with fewer than 3800 neurons could be simulated at a higher speed than the COIN network. One way to shift the advantage back to the COIN is to build a larger COIN network. A 62 × 62 × 10 COIN network has 10 times as many neurons as the GPU network when run at 50MHz, with a similar power consumption and a slightly higher cost.

Network Implementation	Total Cost	Network Speed	Network Size (physical)	Total Power Consumption
Graphics Processing Unit (Tesla K20)	\$3000	50 MHz	11 x 27 x 2.5 [cm]	225 [W]
COIN (3800 neurons)	\$330	50 MHz	20.46 × 20.46 × 3 [mm] +87 × 54 × 10 [mm]	25.5 [W]
COIN (62 × 62 × 10)	\$1282	50 MHz	20.46 × 20.46 × 45 [mm] +87 × 54 × 10 [mm]	212.6 [W]

Table 6.3: Graphics Processing Unit based neural network and COIN equivalent comparison

6.4 Is the COIN competitive?

The comparisons developed in this chapter show that the COIN network can be competitive with the alternate methods of implementing a neural network co-processor. In general a properly designed COIN style neural co-processor should offer a faster neural network than the competing options. In most cases the COIN coprocessor will also have a competitive network cost. However, the improved performance of the COIN comes at the expense of a higher power consumption. No single COIN layout would be competitive with every alternate network. When choosing between neural co-processor options it will be important to design the network, and identify any critical network parameters such as power consumption, before comparing the neural co-processor options.

Chapter 7

Summary and future work

The work done for this thesis has significantly advanced the state of the optical element design for the Compact Opto-electronic Neural co-processor project, and the state of the COIN coprocessor project in general. The work done for this thesis can be divided into three general areas. The first is the COIN system design work presented in both Chapter 2 and Chapter 6. The second area is the use of holographic optical elements in the COIN coprocessor. The third area is the development of new optical interconnect elements based on zone plates. A brief summary of the work in each of these areas is presented below. After summarizing the work done for this thesis this section will lay out the path going from the current state of the project to a fully functional COIN coprocessor.

7.1 COIN system design

As part of the work done for this thesis the COIN system design was refined. Chapter 2 of this thesis presented the current COIN system design. Based on the choice of photodetector elements and system clock, Chapter 2 established an optical power budget and estimated noise levels. Based on the required power levels an argument can be made that the COIN system should move towards a resonant cavity LED type source, rather than the laser sources that have been assumed throughout previous work. To achieve the greatest system flexibility some provision should be made to allow negative weights in the system. Three options were developed for future work.

Using the optical power budget from Chapter 2 and an estimate of the final optical element size, Chapter 6 developed an estimate of both the cost and power consumption of the COIN system. Using these estimates a comparison of a potential COIN system with alternative neural network implementations was made. The general comparison was over network cost, power consumption, and speed. In general the COIN system could be made competitive in any two of these three areas, at the expense of performance in the third.

7.2 Holographic Work

Work done for this thesis has advanced both the theory of holographic element design for the COIN coprocessor and created some experimental techniques which may be useful for future holographic optical element work. The results of the theoretical work on holographic optical elements were presented in Chapter 3. In particular the limits associated with writing small holograms due to finite beam dimensions were established. Also, the change of wavelength problem was examined for cases more complicated than the two plane wave case used to generate write geometries in previous work and it was shown that there will not be an exact solution when the number of plane waves needed to describe the pattern is greater than two. For the COIN coprocessor this means that holographic optical elements will only be a viable interconnect choice when the wavelength of operation is in the range of a given holographic materials sensitivity.

The results of the experimental holographic work were presented in Chapter 4. Two new techniques were developed: holographic lift-off, and the holographic bead lens. The holographic liftoff technique allows for the creation of holographic optical elements that are only as thick as the recording layer holding the hologram. The holographic bead lens uses a mold and liquid recording material to create holographic optical elements that have a lens like surface profile. This allows the creation of a hybrid holographic-refractive optical element that can steer and focus light, but only requires recording a plane grating. To create the lens mold needed for the holographic bead lens a few techniques were developed, including a gray-scale lithographic technique which allows the creation of a range of lens well shapes using a fairly simple system and standard lithographic resist processing.

7.3 Zone Plate optical elements

From the perspective of the COIN coprocessor project, the most important contribution made by this thesis is the development of zone plate based optical elements for use as COIN optical interconnects. These zone plate optical elements could be fabricated using standard lithographic techniques and binary masks. With $\pi/2$ phase steps, the zone plate masks should result in optical interconnects with over 80% connection efficiency and very low cross talk between connections. Additionally, moving to a standard lithographic fabrication process results in a COIN system that can be scaled to very large numbers of neurons.

7.4 Going Forward, Optical Interconnects

At this time it looks like the most promising optical interconnect elements for the COIN system are the multi step zone plate interconnects designed in Chapter 5. Fabrication of these structures will not be a trivial task, and development of a reliable fabrication process should be a priority. However there have also been recent advancements in grayscale lithographic techniques, which may shortly make possible the fabrication of the phase wrapped versions of the transmission functions developed in Chapter 5. The next generation of optical elements should be designed after the optical source properties are finalized and they should be designed specifically for the elements. If at that time the gray-scale lithographic techniques allow for fabrication at the scales needed for the COIN interconnects, the use of phase zone plates should be re-evaluated.

7.5 Going Forward, COIN system

The COIN coprocessor system relies on four main components: (1) optical sources and drivers for the sources, (2) optical interconnect elements, (3) optical detectors, and (4) thresholding electronics. Previous work in the Photonic Systems Group has developed suitable thresholding electronics. By following the results of this thesis it will be possible to tailor optical interconnect elements to the system sources. The last pieces missing to implementing a COIN coprocessor are the optical sources and the detector electronics. Advancing these two areas is needed before a final COIN design can be established. At this time, the best possible path would be to start with the detectors and detector electronics. The noise properties of these circuits will require a re-design of the optical power budget, which may change the selection for the best type of source. Once the detector and electronics are established the optical sources should be revisited, and once the sources are finalized a new set of optical elements should be designed with the final system parameters in mind.

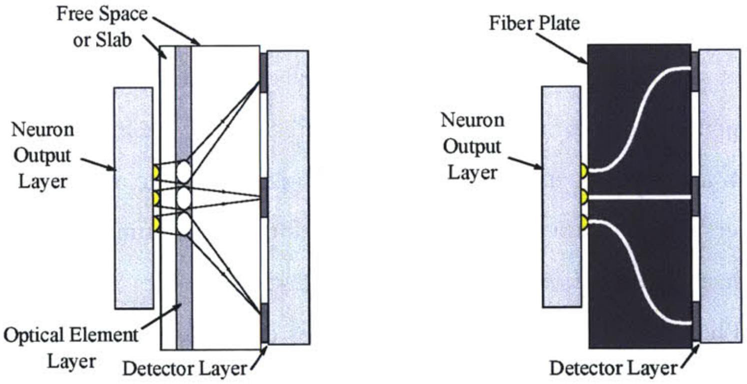
7.6 Conclusions

This thesis has advanced the state of the optical element design for the COIN coprocessor by re-examining the basic assumptions about the co-processor system, establishing some theoretical limits to the use of holographic optical elements in the system, developing new techniques to fabricate these elements, and finally by designing a new type of COIN interconnect using zone plates. The work done for this thesis was aimed at developing interconnect elements for the COIN coprocessor, but the general results will be applicable to any similar interconnect application. The experimental holographic techniques in Chapter 4, while ultimately not the best choice for the COIN system, should be considered when designing holographic optical elements.

Appendix A

Fiber Plate and COIN test system

In addition to the free space optical interconnects discussed in the main body of this thesis interconnects using fiber optics to route the light from source to detector were also investigated. The basic approach is diagrammed in Figure A-1. At this time fiber optic interconnects do not seem like a viable option for the COIN co-processor due to the difficulties in running the large number of fiber connections needed for even a moderately sized COIN, but for small networks it it possible to assemble a fiber based connection system in a reasonable amount of time. To get an estimate of the amount of work involved in creating a fiber interconnect based COIN, a fiber connection plate was built for this thesis. The plate was then incorporated in a COIN testbed system was built to simulate a single COIN layer. A picture of the testbed system is shown in Figure A-2.



(a) Free Space Interconnects

(b) Fiber Plate Inteconnects

Figure A-1: A free space optic connection scheme and a fiber connection scheme. In the fiber based connection scheme a fiber plate provides support to and controls the position of the optical fibers making the neuron connections.

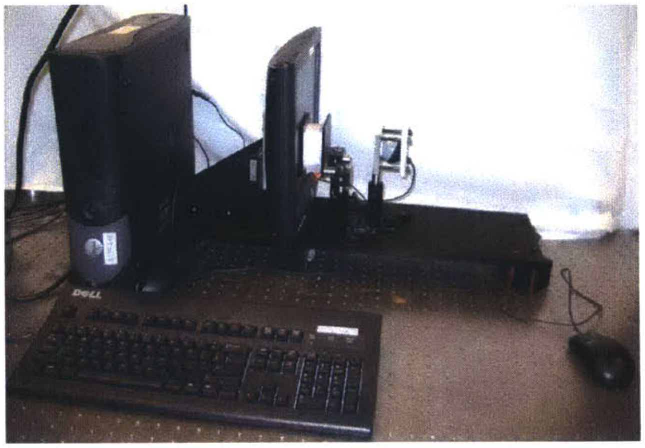
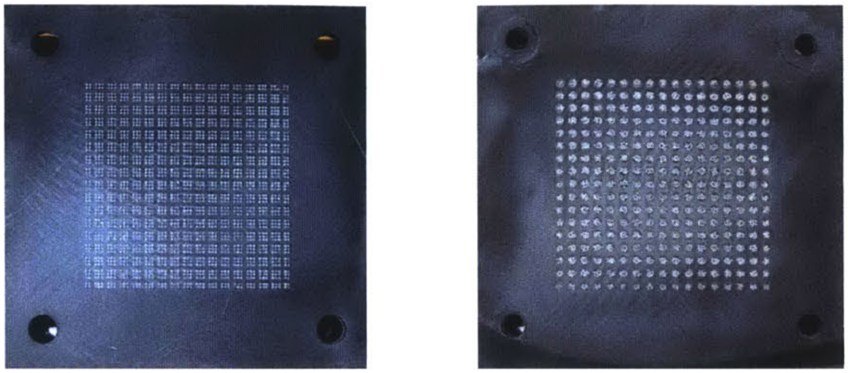


Figure A-2: Side view of first generation testbed system

A.1 The Fiber Interconnection Plate

An interconnect system can be made using optical fibers. Optical fibers are low loss, allowing most of the source light to reach the detectors, and with proper design will have very low crosstalk between fibers. The disadvantages of optical fibers are the need to couple light into the fibers, which can be difficult for some light sources, and the need of some support structure to accurately position the optical fibers relative to the detectors and optical sources. For this thesis a fiber connection plate was built using plastic optical fiber 0.5mm in diameter. The plate was designed to implement the connections for a 17 × 17 neural network using nearest neighbor connections. The input and output faces of the plate are shown in Figures A-3, and A-4.

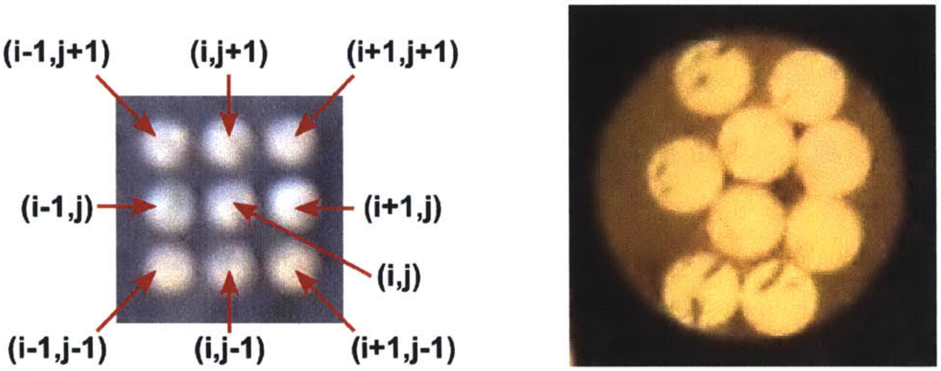
On the input side of the plate each 0.5mm fiber is glued into a single hole. The individual fibers are segmented into groups of up to nine fibers, and each fiber is used to form a connection with a different neuron in the output plane. The edge and corner neuron groups have fewer fibers because they have fewer neighbors. The number of fibers needed for an $N \times M$ neuron connection plate is $(3N - 2) \times (3M - 2)$. The plate pictured had 2401 inputs. The output side of the plate features one hole per neuron, and all of fibers representing connections to a given neuron share one hole. The fiber connections were routed by hand and then glued in place with epoxy. To create smooth input and output faces, after securing the fibers with epoxy the faces were fly-cut, sanded, and polished with plastic polishing compound.



(a) Input

(b) Output

Figure A-3: The input and output faces of the fiber optic test plate. The input face is has 49x49 holes each with an individual fiber, grouped by neuron. The output face has 17x17 holes with three to nine fibers, one fiber for each connection to the input of that neuron.



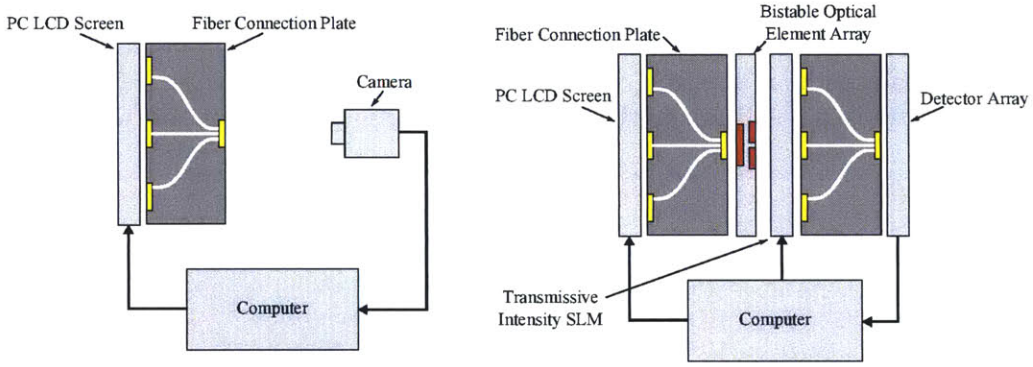
(a) Input

(b) Output

Figure A-4: One neuron on the input and output sides of the fiber optic plate. Each fiber on the input side represents a connection to the a neuron on the output side as indicated in the diagram. Each hole on the output side contains up to nine fibers, one for each connection to a neuron on the input side of the plate. A detector of sufficient area placed over a hole in the output face will recieve light from all the inputs to a given neuron, allowing it to perform a sum of the inputs.

A.2 The COIN testbed system

The fiber plate described in the previous section was built for use in a COIN testbed system. This first generation testbed was designed to implement a single layer of a COIN style nearest neighbor network and consisted of four main components: (1) an LCD display to generate the neural layer outputs, (2) a fiber optic plate to connect outputs from layer n to layer n+1, (3) a web-cam to image the output plane of the fiber optic plate, and (4) a computer to process data from the webcam and generate images to be displayed on the LCD screen. A diagram of this system is shown in Figure A-5a. In this system almost all of the neural processing was performed by the computer.



(a) First Generation Testbed

(b) Second Generation Testbed

Figure A-5: System Diagrams for the first and second generation COIN testbed systems.

A second test system has been designed, implementing two layers of a COIN network. This test system will use an array of bistable optical elements, developed by Wegene Tadele, which use analog circuitry to recreate the neuron activation function. In the second generation test system the neural processing for at least one layer of the network will be performed by neuron-like processing units.

Appendix A

Evaluating the diffraction Integral

To simulate the optical elements developed for this thesis a method of numerically evaluating the diffraction integral was needed. The general form of the diffraction integral with no far-field or paraxial approximations made is

$$U(\mathbf{r}_2) = \int U(\mathbf{r}_1) \left[\frac{\cos(\mathbf{n}, \mathbf{r}_2 - \mathbf{r}_1) - \cos(\mathbf{n}, \mathbf{r}_1)}{2} \right] \frac{e^{jk|\mathbf{r}_2 - \mathbf{r}_1|}}{|\mathbf{r}_2 - \mathbf{r}_1|} \partial \mathbf{r}_1$$

The terms $\cos(\mathbf{n}, \mathbf{r}_i)$ represent the cosine between the vectors $\mathbf{n}$, the normal vector for the diffracting aperture, and $\mathbf{r}_i$, the vector from the system origin to a given point. Points in the input plane are associated with $\mathbf{r}_1$ vectors and points in the observation plane are associated with $\mathbf{r}_2$ vectors. The integral is evaluated in the plane of the input aperture making $\cos(\mathbf{n}, \mathbf{r}_1) = -1$. An auxiliary function $\tilde{\mathbf{h}}(\mathbf{r})$ can be introduced to simplify the notation.

$$\tilde{\mathbf{h}}(\mathbf{r}) = \left[\frac{\cos(\mathbf{n}, \mathbf{r}) + 1}{2} \right] \frac{e^{jk|\mathbf{r}|}}{|\mathbf{r}|}$$

Under these substitutions the diffraction integral becomes

$$U(\mathbf{r}_2) = \int U(\mathbf{r}_1) h(\mathbf{r}_2 - \mathbf{r}_1) \partial \mathbf{r}_1$$

In this form it is clear that the diffraction integral is the convolution of the input field and the auxiliary function $\tilde{\mathbf{h}}(\mathbf{r})$. From the convolution theorem the same result can be had by computing

$$U(x_2, y_2) = \mathcal{F}^{-1}\{\mathcal{U}_1 \cdot \mathcal{H}\}$$

The Fourier transform of the output field is the product of the Fourier transform of the input field and the Fourier transform of the function $\tilde{\mathbf{h}}(\mathbf{r})$. Computing the output field using this approach will be faster than computing the field by directly evaluating the diffraction integral.

A.1 Linear Convolution using the FFT

The convolution theorem holds for the Discrete Fourier Transform (DFT) as well as the continuous Fourier Transform. This is important since for the structures investigated in this thesis the fields to be evaluated are computed numerically so only the Discrete Fourier Transform of the fields is available. Use of the DFT introduces one complication in computing the convolution. The Discrete Fourier Transform computation assumes that the functions being transformed are periodic, repeating exactly outside of the window representing the sample of the function. A direct product of the Discrete Fourier Transform of two sets of data will produce the circular convolution of the functions represented by the two data sets. Since the linear convolution is desired the data sets must first be padded with zeroes, and then the result truncated to give the region where the circular convolution of the padded sets is equal to the linear convolution of the original data sets. If the first data set consists of N points, and the second data set consists of M points then the linear convolution should have a length of N+M-1 points. If the individual data sets are padded with zeros to the length N+M-1 before transforming, then the first N+M-1 data points in the resulting convolution correspond to the linear convolution of the two original data sets. Padding beyond this length is not necessary, but may be useful since some implementations of the Discrete Fourier Transform are faster when the length of the input data is a power of 2. Regardless of how far the data sets are zero padded, the first N+M+1 points still represent the linear convolution of the two data sets, there is no change in accuracy.

The two data sets do not need to be the same length, or cover the same input range, for the convolution to work correctly. If the convolution between two functions $f(x)$ and $g(x)$ is to be computed, and the transform of f is computed using some range of samples $-A_1 \leq x \leq B1$ and the transform of g is computed using some range of samples $-A_2 \leq x \leq B_2$

then the discrete convolution will be non-zero for the range $-(A_1+A+2) \leq t \leq (B_1+B_2)$. To get the proper result, the two functions should be sampled at the same frequency over their input range.

For any computer there will be a limit to the size of the data set that can fit in the computers memory and trying to convolve data sets larger than this will drastically slow the computation. To evaluate the diffraction integral over a larger output plane than can be held in memory, it is possible to compute the result for small sections of the output plane and stitch the sections together. The results shown in this thesis were computed using the overlap-add method. The output plane was divided into non-overlapping regions and the discrete convolution for each region was computed. The results, with proper offsets, were added together to form the full output plane.

The auxiliary function $\tilde{\mathbf{h}}(\mathbf{r})$ becomes large for small values of $|\mathbf{r}|$. However, there cannot be more light in the output plane than there was in the input plane. To enforce this, the result of the diffraction integral is normalized by the total intensity in the output plane. The total power is then 1, and the sum of the intensity over any region represents the fraction of the output plane power in that region. The underlying assumption of this normalization is that the majority of the light described by the input field actually propagates to the output plane. If the output plane is too small, or if the input field is scattered at high angles, then this approach will overestimate the power in the output plane.

A.2 Testing the Propagation Results

To test the numerical implementation of the diffraction integral outlined above, a series of example problems were run and their results compared to the expected values. For the optical elements examined in this thesis, most of the work is in the near field as opposed to the far field. Standard diffracting apertures do not have simple solutions in the near field, so instead Gaussian Beams were used as input fields. Gaussian beams have the advantage of having relatively narrow widths and well defined propagation behavior. The code was tested with both expanding and focusing Gaussian beams. From the test cases, a sampling of 1001 points per 100µm is high enough to give good results.

Bibliography

[1] A. Yariv, *Optical Electronics in Modern Communications.* New York: Oxford University Press, 5th ed., 1997.

[2] A. K. Maini, *Lasers and Optoelectronics: Fundamentals, Devices and Applications.* Chichester, West Sussex, United Kingdom: John Wiley & Sons Ltd., 2013.

[3] M. Ruiz-Llata, H. Lamela, and C. Warde, "Design of a compact neural network based on an optical broadcast architecture," *SPIE*, vol. 44.5, 2005.

[4] G. Hennenfent, *Production of Holographic Optical Interconnection Elements for a Compact Optoelectronic Neural Network co-Processor.* Thesis, Ecole Nationale Superieure de Physique de Strasbourg, Strasbourg, 2003.

[5] H. Photonics, "s9055_series_kpin1065e04.pdf." `http://www.hamamatsu.com/us/en/product/category/3100/4001/4103/index.html`. Accessed: 12/16/2013.

[6] C. M. Bishop, *Pattern recognition and machine learning / Christopher M. Bishop.* Information science and statistics, New York : Springer, c2006., 2006.

[7] F. Rosenblatt, *Principles of neurodynamics; perceptrons and the theory of brain mechanisms.* Spartan Books, 1962.

[8] M. Bear, B. Connors, and M. Paradiso, *Neuroscience: Exploring the Brain.* Philadelphia, PA: Lippincott Williams & Wilkins, 3rd ed. ed., 2007.

[9] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bulletin of Mathematical Biophysics*, vol. 5, pp. 115–133, December 1943.

[10] M. Minsky and S. Papert, *Perceptrons : an introduction to computational geometry.* MIT Press, 1988, c1969.

[11] N. H. Farhat, D. Psaltis, A. Prata, and E. Paek, "Optical implementation of the hopfield model," *Applied Optics*, vol. 24, no. 10, pp. 1469–1475, 1985.

[12] J. Hong, S. Campbell, and P. Yeh, "Optical pattern classifier with perceptron learning," *Applied Optics*, vol. 29, no. 20, pp. 3019–3025, 1990.

[13] L. Zhang, M. Robinson, and K. Johnson, "Optical implementation of a second-order neural network," *Optics Letters*, vol. 16, no. 1, pp. 45–47, 1991.

[14] T. L. Simpkins, *Design, Modeling, and Simulation of a Compact Optoelectronic Neural Coprocessor.* PhD dissertation, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, Feb. 2006.

[15] B. F. Ruedlinger, *Fundamental Building Blocks for a Compact Optoelectronic Neural Network Processor.* Phd thesis, Massachusetts Institute of Technology, May 2003.

[16] C. G. Fonstad, *Microelectronic Devices and Circuits.* McGraw-Hill series in electrical and computer engineering. Electronics and VLSI circuits, New York: McGraw-Hill, Inc., 1994.

[17] M. B. F. Hans Melchior and F. R. Arams, "Photodetectors for optical communication systems," *Proceedings of the IEEE*, vol. 58, pp. 1466–1486, October 1970.

[18] J. M. Perkins, T. L. Simpkins, C. Warde, and C. G. Fonstad, "Full recess integration of small diameter low threshold vcsels within si-cmos ics," *Optics Express*, vol. 16, no. 18, pp. 13955–13960, 2008.

[19] E. F. Schubert, *Light-Emitting Diodes.* Cambridge University Press, second ed., 2006.

[20] E. F. Schubert, N. Hunt, R. J. Malik, M. Micovic, and D. L. Miller, "Temperature and modulation characteristics of resonant-cavity light-emitting diodes," *Journal of Lightwave Technology*, vol. 14, no. 7, pp. 1721–1729, 1996.

[21] M. Komarčević, "Production of holographic optical interconnection elements," thesis (m.eng), Massachusetts Institute of Technology, 2000.

[22] M. Ruiz-Llata, *Diseño e Implementación de Redes Neuronales Optoelectrónicas. Aplicación en Sistemas de Visión.* Phd thesis, Universidad Carlos III de Madrid, Spain, May 2005.

[23] P. Hariharan, *Optical Holography: Principles, techniques, and applications.* Cambridge Studies in Modern Optics, Cambridge ; New York: Cambridge University Press, second ed., 1996.

[24] H. Kogelnik, "Coupled wave theory for thick hologram gratings," *The Bell System Technical Journal*, vol. 48, p. 2909, 1969.

[25] J. Upatnieks and C. Leonard, "Efficiency and image contrast of dielectric holograms," *Journal of the Optical Society of America*, vol. 60, pp. 297–305, March 1970.

[26] J. W. Goodman, *Introduction to Fourier Optics.* Greenwood Village, Colorado: Roberts and Company Publishers, 3rd ed., 2005.

[27] Polygrama, "Polygrama." `http://www.polygrama.com/Index.htm`. Accessed: 14/11/2012.

[28] RainX, "Rainx." `http://www.rainx.com`. Accessed: 1/9/2014.

[29] E.-H. Park, K. Moon-Jung, and K. Young-Se, "New fabrication technology of convex and concave microlens using uv curing method.," in *IEEE Lasers and Electro-Optics Society 1999 12th Annual Meeting*, vol. 2, pp. 639–640, 1999.

[30] Cognimem, "Cognimem technologies." `http://www.cognimem.com/`. Accessed: 07/21/2013.

[31] Cognimem, "Tm_cm1k_harware_manual.pdf." `http://www.cognimem.com/`. Accessed: 07/28/2013.

[32] Xilinx, "Virtex-7 fpga family." http://www.xilinx.com/products/silicon-devices/fpga/virtex-7.html, 2013. Accessed: 7/29/2013.

[33] nVidia, "Personal supercomputer workstations with tesla gpus." http://www.nvidia.com/object/tesla-servers.html/, 2013. Accessed: 7/28/2013.