Fundamental Building Blocks for a Compact Optoelectronic Neural Network Processor

by

Benjamin Franklin Ruedlinger

B.S., Physics
Case Western Reserve University (1999)
M.S., Computer Engineering
Case Western Reserve University (1999)

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

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2003

Author .....
Department of Electrical Engineering and Computer Science
May 19, 2003

Certified by.....
.....
Cardinal Warde
Professor
Thesis Supervisor

Accepted by.....
.....
Arthur C. Smith
Chairman, Department Committee on Graduate Students

Fundamental Building Blocks for a Compact Optoelectronic Neural Network Processor

by

Benjamin Franklin Ruedlinger

Submitted to the Department of Electrical Engineering and Computer Science
on May 19, 2003, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

The focus of this thesis is interconnects within the Compact Optoelectronic Neural Network Processor. The goal of the Compact Optoelectronic Neural Network Processor Project (CONNPP) is to build a small, rugged neural network co-processing unit. This processor will be optimized for solving various signal processing problems such as image segmentation or facial recognition. These represent a class of problems for which the traditional logic-based architectures are not optimized. The CONNPP utilizes the processing power of traditional electronic integrated circuits along with the communications benefits of optoelectronic interconnects to provide a three-dimensionally scalable architecture. The topic of interconnects includes both optoelectronic interconnects between processing planes as well as wire based interconnects within the processing planes. The optoelectronic inter-plane interconnects allow the CONNPP to achieve 3-dimensional scalablility. These interconnects employ an array of holograms designed and fabricated using an analytic model that has been developed. This analytic model takes into account reading and writing the hologram at different wavelengths as well as non-idealities of the optoelectronic devices being used. The analytic model was tested and showed agreement with experiment to within 10% of the calculated values. The photodetectors used for the testing of this system have been designed within standard process technologies using a lateral PiN structure and a novel lateral BJT structure. In addition, highly-linear transimpedance amplifiers were designed to condition the signal from the photodetector for further processing. Finally, a new chip-level interconnect architecture has been proposed for the CONNPP that utilizes a system bus to provide global connectivity between the neurons within the plane of the chip. Software models were built to simulate various chip-level connectivity schemes. These simulations show the significant potential benefits of the global bus-based chip-level architecture that has been proposed.

Thesis Supervisor: Cardinal Warde
Title: Professor

Acknowledgments

First, I would like to thank Professor Cardinal Warde for his guidance and metoring over the past 4 years. Without his vision for the Compact Optoelectronic Neural Network Processor Project, the work presented here would not be possible.

In addition, I would like to thank Professor Clifton Fonstad. I have worked with Professor Fonstad over the past four years and he has aided in the design and testing of the circuits and devices produced during my tenure at MIT.

To Dr. Thomas Knight, thank you for serving on my doctoral committee and providing very useful feedback. Your treks over to Building 13 have been much appreciated over the past year and a half.

Thank you to my wife, Emily Nelson, who I both met and married at MIT. She deserves the largest thanks as she has put up with me for the past four years of graduate school. Without her love and support, I don't know if I could have completed this journey.

Travis Simpkins has been very instrumental as both a colleague and a friend over the past several years. As suitemates at CWRU, I doubt we could have conceived of the fact that years later we would be sharing an office at MIT as Ph.D. candidates. Thank you for all your time spent over the last couple years on both a professional and a personal level.

Thank you also to my parents, Richard and Jeanne Ruedlinger. Without them, of course, absolutley none of this would be possible. Over the past twenty-seven years, they have done a great job of supporting me in all my endeavours. For this reason, it is only appropriate that they also share in the accomplishment.

Contents

List of Figures

Chapter 1

Introduction

As an introduction to the main topic of this thesis, basic background on neural networks will be given. In addition to background, a discussion of the benefits of neural networks over traditional logic based computing for certain classes of problems will also be discussed. Next, the issue of computing interconnects will be examined. Traditionally, electrical wires of some sort have been used to connect computing nodes. However, as computing speeds increase, electrical interconnects are potentially less effective. Are optical interconnects a viable alternative to electrical interconnects in some cases? Optical interconnects have already taken the lead in transcontinental communication (fiber optics). They may well prove to be the next step in chip level interconnects also.

After brief discussions of neural networks and optical interconnects, an overview of the Compact Optoelectronic Neural Network Processor Project (CONNPP) [1] will be given. The goals for this project as well as implementation details will be examined in order to provide the context for the work discussed in this thesis. Finally, a brief overview body of work completed and the following chapters of this document will be given.

1.1 Why Neural Networks?

The human brain is in some ways one of the most complex and accomplished processors in existence. Yet, in other ways, it is inefficient, unwieldy, and slow. Humans are extremely adept at pattern recognition, visual navigation, and making inferences with incomplete data sets. However, a simple computational task such as multiplying three digit numbers in our heads can confound even the most brilliant person. The fact of the matter is that our brains are not wired for mathematical computation. As an evolving species, mathematical computation was not a feature that helped humans survive. It was much more important to be able to recognize friend from foe than it was to multiply three digit numbers.

In order to compensate for the lack of computational prowess, people have sought to design and build machines to help them overcome this weakness. These machines which started as simple calculators with a single accumulator have grown into extremely complex general purpose computers with multiple processing units and vast amounts of memory. Although these processors have increased their capabilities considerably[1], they all still use the same basic theory of operation. By converting all data within the processor to binary format, Boolean algebra (AND, OR, NOT, etc.) is then used to implement all the complex operations within these traditional logic-based processors. This has been extremely successful and has produced applications which could not even be conceived of when the pioneers were designing the first calculator-like processors.

Although these logic-based processors can be used for a variety of applications, there are several classes of problems with which traditional processors have trouble. These problems have not been able to be reduced to a set of simple logic-based rules that can be used to solve them. A few of these problems include speaker-independent voice recognition, natural language processing, and orientation and scale independent visual object recognition. While computer scientists have been trying to reduce these

[1] Over the past fifty years, traditional logic-based processors have increased their computational power by almost seven orders of magnitude. ENIAC could carry out 357 floating point operations per second (FLOPS) while today's Intel Pentium 4 processor can do over 2×10^9 FLOPS

problems to logic-based rules for over thirty years, the results in many cases are still unacceptable.

Ironically, these classes of problems which logic-based processors have difficulty with are some of the same problems for which the human brain is seemingly optimized. A perfect example of this is speaker-independent voice recognition. This task consists of taking strings of phonetic sounds and parsing them into the words to which they correspond. This task is further complicated by the variation in dialect and individual speech patterns between speakers. A particular word might sound very different when spoken by a person from the United States than when spoken by someone from the United Kingdom. As English speakers we are able to understand these speech pattern differences and still recognize the words that are being spoken. Logic-based computers, however, have a very difficult time with this task. Although millions (if not billions) of dollars of funding and research has gone into solving this problem with traditional logic-based processors, the results have not been encouraging. After decades of research on this specific topic, there has yet to be produced a logic-based processor system which can even come close to matching the speaker independent voice recognition capabilities of a child.

In light of this fact, researchers began looking at the human brain for inspiration in creating processors to handle these types of problems. The human brain is an extremely complex biological system which is far from being completely understood. Even so, it is still possible to use those workings of the brain which are understood to design and build models replicating a very limited subset of its functionality. Artificial neural networks (ANNs) seek to model the low-level functionality of the brain either programmatically or through hardware.

The basic processing unit of the brain is a single celled structure called a neuron. Each neuron has many inputs from other neurons and also many outputs to other neurons. These inputs and outputs allow the neuron to communicate with other neurons through chemical interactions. Communication occurs by chemical reactions within the neuron that slightly alter the electrical potential of the inputs and outputs connecting to different neurons.

The basic model of a neuron consists several parts. In the first step, weighted inputs are presented to the neuron in the form of potentials. The neuron sums these potentials presented to the inputs. It then uses this summed value as the input to a function ($f(x)$). This function can take many forms such as a sigmoid function or linear function. The neuron then sets the potential of its outputs to reflect the result of this function. The output lines (as well as the input lines) have weights which determine how strong the connection is between the neurons.

An artificial neuron seeks to replicate this functionality either in software or hardware. A diagram of such an artificial neuron can be seen in Figure 1-1.

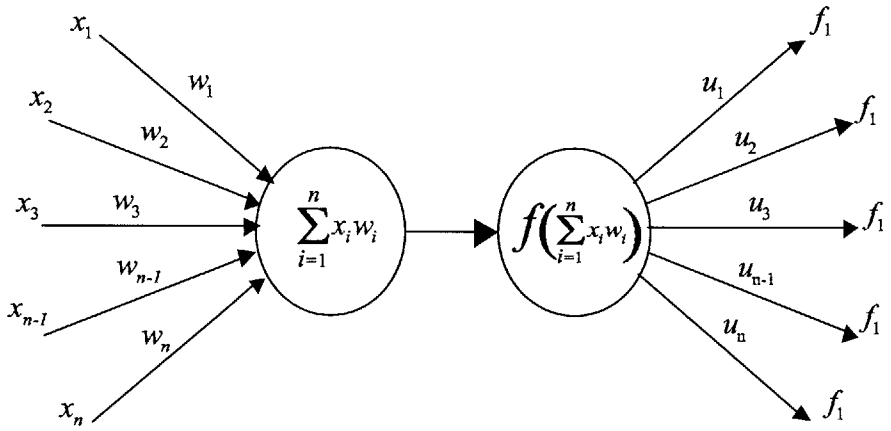

Figure 1-1: Illustration of one model of an artificial neuron

In this diagram, x_i represents the i^{th} input to the neuron and w_i represents the weight of the connection between these neurons. Each input (x_i) is multiplied by its connection weight (w_i) and these values are summed. This sum is used as the argument for the transfer function $f(x)$. The output of this transfer function (f_1) is put on the outputs where it will be multiplied by the weights of those connections (u_i).

These neurons, which act as independent processing units, are then connected together in a dense network. With each neuron being able to communicate with a large number of other neurons, information is able to propagate through the network efficiently. Furthermore, the connection weights allow each neuron to know how

important each piece of information is that it is receiving. With relatively simple processing units, this dense interconnection network provides much of the power of both biological and artificial neural networks.

In artificial neural networks, these artificial neurons are many times connected together in layered structures. Each layer might have a different transfer function or different interconnection scheme. When there is a layer structure present, there are three distinct types of layers; an input layer, hidden layers, and an output layer. The input signals to the neural network are connected directly to the input layer neurons. The input layer is then connected to the first hidden layer. The first hidden layer is in turn connected to the second hidden layer. This proceeds until the last hidden layer is connected to the output layer. The output layer then presents the output of its neurons as the result of the neural network's computation. An example of the layer structure can be seen in Figure 1-2.

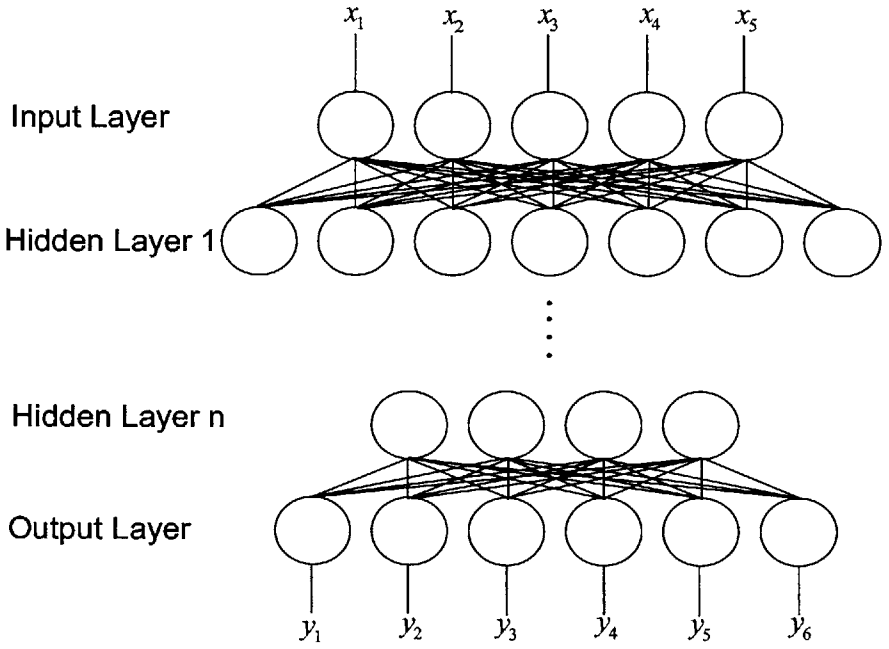

Figure 1-2: Layer structure of an artificial neural network

One of the most important characteristics of neural networks is their ability to learn. While traditional logic-based processors are programmed with explicit instruc-

tions on how to accomplish a particular task, neural networks are trained. To do this, a selection is made of inputs for which the outputs are known. These inputs are then presented to the input layer of the neural network. When the input propagates through the network and a result is available at the output, it is then compared to the desired output. If the actual output exactly matches the desired output, nothing is changed within the neural network. In the more likely event that the actual output does not exactly match the desired output, the error between the actual output and the desired output is calculated. This information is then used with a particular training algorithm to slightly change the connection weights. The connection weights are modified in a manner to make it more likely the desired output is produced the next time that input is presented to the neural network. The network is generally trained repeatedly with different input/output data sets. The goal of the training is to minimize the error between the actual output and the desired output. When an acceptable level error has been achieved, the network can then be presented with inputs that have unknown outputs. If the training data is sufficiently representative of the total data set and if the network has been effectively trained, the neural network will produce a correct output when presented with an unknown or incomplete input.

This ability of neural networks to operate effectively with noisy, slightly different, or incomplete data is extremely important. This is one of the main features that distinguish neural networks from traditional logic-based processors. Logic-based processors must use strict sets of incredibly complicated mathematical and statistical rules in order to work with data input that is not presented as expected. Through their numerous simple processing units, a dense network of connections, and the ability to learn, neural networks are much more adept at solving problems which require some amount of inference. The problems discussed previously such as speaker independent voice recognition and visual navigation are problematic for logic-based processors precisely because they do require some level of inference. As a result, solving these classes of problems is exactly the goal of artificial neural network research.

1.2 Electrical vs. Optical Interconnects

Put most simply, interconnects transfer information between computing nodes within a system. These computing nodes may be logical blocks within a microprocessor, the CPU and system memory, or even remote systems. As more and more nodes, running at faster and faster speeds, are fit into smaller spaces, the electrical interconnects currently in use have significant trouble scaling in both density and speed. One example of this is system buses connecting the CPU and system memory in server and desktop computing systems. The current state of the art has the system bus running at less than 20% of the core CPU speed (Figure 1-3) [2]. This represents a major bottleneck in transferring data within a PC.

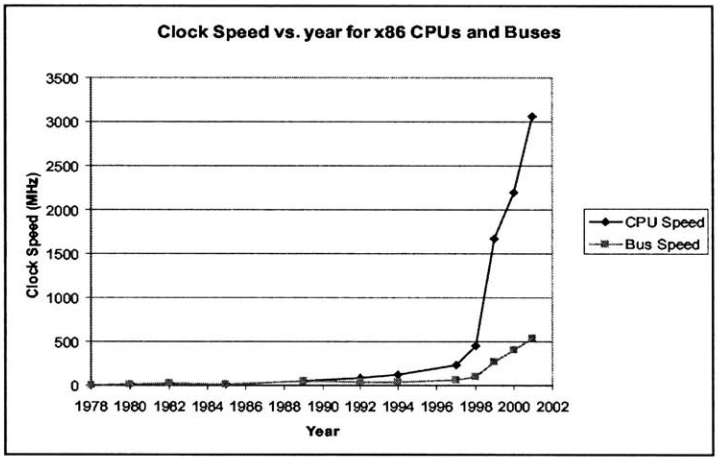

Figure 1-3: CPU and system bus speed as a function of time

The fact of the matter is that there are significant limitations to electrical interconnects that are starting to be approached. One of the most critical of these limitations is aspect ratio considerations of electrical interconnects. It is well known that all non-superconducting wires have both resistance and capacitance determined by the cross-sectional area and length of the wire. This RC time constant therefore determines the fastest speed at which the interconnect can be run. For a typical interconnect made of aluminum or copper, at room temperature, the bandwidth limitation (B) in bits/s is as follows

$$B \cong 10^{16}\frac{A}{l^2}\text{bits/s.} \tag{1.1}$$

Where A is the cross-sectional area of the wire and l is the length of the wire [3]. This may not seem so bad until it is considered that the cross-sectional areas of on-chip wires are on the order of micrometers squared while the length of the interconnects can be as long as centimeters. With CPU clock speeds extending into the gigahertz range and process technologies shrinking into the nanometer range, these aspect ratio considerations represents a formidable problem.

Timing and synchronization within microprocessors represents another area where electrical interconnects are running into issues. System clock skew is defined as the difference in arrival time of the system clock signal between two registers. This difference in arrival times of the system clock can result in race conditions and malfunctioning logic [4]. Even though much care is given to designing microprocessors such that nodes are equidistant from clock drivers (using hierarchal H-trees), clock skew still remains a problem. Much of this clock skew is caused by thermal effects. The resistivity of the electrical interconnects varies significantly with temperature. As different logical blocks are switched on and off, different interconnects are heated up, causing potential clock skew [5]. Industry standards traditionally dictate that the clock skew remain less than 10% of the total clock cycle [6]. As the speed of microprocessors increases, the clock period becomes smaller. As a result, the clock skew becomes more of a problem as it represents an increasing portion of the clock cycle.

In addition to the problem of clock skew, power consumption of the is also a major consideration. Today's typical x86 based processors can consume between 60 and 100 Watts of power with as much as 40% of that power going to the clocking and synchronization circuitry. This power causes significant heating of the microprocessor which can lead to the thermal-based resistivity changes and transistor threshold variation. Once again, as process technology scales down, designers are able to pack more and more logical blocks on a chip, all of which need to be synchronized with a

system clock. As a result, power consumption of clock generation and distribution is a continuing problem.

Optical interconnects represent a complete paradigm shift that can reduce or eliminate many of the previously discussed issues faced by electrical interconnects. Optical interconnects, in the most general sense, use light instead of electrical wires to transfer information between two computing nodes [7]. The information is usually encoded as an intensity modulation.

The simplest optical interconnect consists of a light emitter, the medium through which the light propagates, and a receiver to detect the optical signal. The emitting device is most usually a laser or LED, though technically it can be any light emitting device. The light can be modulated either directly by the emitter (i.e. modulation the drive current of a diode laser) or by an external modulation device. The propagation medium can be either free-space or a dielectric. Many times, a device will be used to direct the light to a particular place through the propagation medium. This can be done with devices such as waveguides, MEMS mirrors, or holograms. Finally, the receiver of the optical interconnect most generally a photodetector which converts the intensity modulated optical signal into a electrical current modulation.

Because optical interconnects use fundamentally different mechanisms of operation, they are able to overcome some of the inherent shortfalls of electrical interconnects. For instance, the bandwidth limiting aspect ratio problem in electrical interconnects discussed previously does not exist in optical interconnect systems. In fact, for well-engineered optical materials, there is negligible loss or distortion of the optical signal over the distances in question. Consequently, as the process technologies yield smaller features, the potential bandwidth of optical interconnects remain unaffected.

The timing and synchronization issues faced by electrical interconnects can also be improved upon with the use of optical interconnects. As stated above, even though microprocessors are generally designed such that clocked registers are equidistant from the clock drivers, thermal effects can still cause impedance mismatch that results in large clock skews. While the index of refraction of the propagation material

(somewhat analogous to the resistivity) may change with temperature, it does not cause the nearly the same magnitude of clock skew as would be seen in an electrical interconnect system [5].

Furthermore, optical interconnects may be able to utilize more exotic geometries that would almost completely eliminate the non-uniform heating of the propagation medium. Currently, it is difficult to get high speed signals in and out of a microprocessor due to the large capacitances of the bond pads, bond wires, and package leads. As a result, the clock must be generated on-chip for modern microprocessors. This directly contributes to the large power consumptions discussed above. Optical interconnects allow the possibility of generating a high-speed clock signal off chip and bringing it on-chip using light. In this scenario, the clock speed is only limited by the response of the emitting and detecting devices. Additionally, moving clock generation and distribution off-chip frees up significant amounts valuable chip real estate.

In electrical clock distribution, significant redesigns are necessary when new logical blocks are added or a process shrink is done. These redesigns are needed in order to accommodate impedance matching and signal reflections so that synchronization is performed correctly [5]. Due to the fact that these effects are absent in an optical implementation, the optical system may be able to operate at either 5 MHz or 5 GHz without any redesign.

While optical interconnects do have many benefits over electrical interconnects, there are still issues to be resolved before wide spread use of the technology is seen. The largest hurdle in the adoption of optical interconnects is incompatibility with silicon based processes. Silicon is an inherently poor optoelectronic material due to its indirect bandgap structure. While silicon photodetectors perform moderately well, silicon light emitters are in their infancy. As a result, other optoelectronic devices must be grown on or bonded onto chips. Widespread acceptance of optical interconnects would require retooling of fabrication facilities and major process flow changes for the major foundries such as TSMC, IBM, or Intel. Other issues which confront optical interconnects include quality of optoelectronic devices and lack of optomechanical interface hardware.

1.3 The Compact Optoelectronic Neural Network Processor Project

The goal of the Compact Optoelectronic Neural Network Processor Project (CONNPP)?? is to produce a small, rugged co-processing unit that can be used in conjunction with a standard PC. The PC will be able to offload appropriate tasks such as pattern recognition, visual navigation, or sensor fusion to this co-processing unit. Furthermore, the co-processing unit will be optimized, using neural network architectures, to solve problems of this nature. First, some of the operational characteristics of this co-processing unit will be discussed. After that, the specifications and implementation of this processor will be examined in more detail.

The idea of optoelectronic neural network processors is by no means a new one. There have been many proposed and implemented optical neural network architectures over the past twenty years [8–10]. One characteristic that these implementations share is that they take up the better part of an optical table in terms of physical size. While this size is perfectly acceptable for research purposes, it is less than practical for commercial systems. For this reason, one of the goals of the CONNPP is to make its optoelectronic neural network processor the size of a CD-ROM drive. This will be accomplished through the use of monolithically fabricated optoelectronic integrated circuits (OEICs).

Advances in the integration of optoelectronic devices with standard CMOS electronics have only recently made this level of integration possible. MIT Professor Clifton Fonstad, along with his group, have pioneered a process known as Aligned Pilar Bonding (APB) that allows optoelectronic devices to be grown on separate substrates and subsequently bonded onto standard CMOS chips [11–13]. This, as will be seen, opens up a realm of possibilities in terms of new and innovative systems that can be built.

One of the key features of the OEICs designed and built for the CONNPP is that they will accept optical input from one side of the chip and provide optical output from the other side of the chip. This allows the OEICs to be cascaded in the third

dimension. For the past thirty years, microprocessors have been fabricated on a single 2-dimensional substrate and connected via system buses to memory and peripherals. In the last several years, researchers have begun to think about cascading processors in the third dimension. Groups in both academia and industry are actively investigating the possibility of interconnecting these 3-dimensionally cascaded processors either electrically or optically [14]. The Compact Optoelectronic Neural Network Processor Project seeks to leverage the benefits of optical interconnects in order to create a very dense system of inter-chip interconnects (almost 15000 channels/cm^2).

Neural networks are a prime candidate for 3-dimensional cascading because of their interconnection density requirements. In order for neural networks to work effectively, the neural nodes must be able to communicate with a significant number of other neurons [15]. The most effective neural network is a globally connected network where every neuron is connected to every other neuron. While global interconnects are very effective, they are also quite impractical. For global interconnects, the number of connections needed goes as n^2 where n is the number of neurons in the network. As a result, a less dense interconnect scheme for neurons is necessary to make the systems more practical. In the human brain for example, each neuron connects to approximately 10,000 other neurons. While this may seem like a large number, it is actually only .0001% of the 10 billion neurons in the brain. Interconnection schemes between neurons both within the plane of the OEIC and between OEIC planes are a major consideration for the CONNPP. This is one of the central topics addressed by this thesis.

1.3.1 High-level overview of the CONNPP

As discussed previously, optics can offer significant benefits in some cases over electronics when it comes to interconnect technology. Processing and computation, however, are a different matter. Over the past thirty years, millions of man-hours and billions of dollars have gone into integrated circuit design and research. The results of this time and money have increased the processing power per square centimeter of integrated circuits by many orders of magnitude. While many optical neural network

architectures and implementations have used optics for both processing and communication [9, 16, 17], it was decided to use the power of electrical processing units coupled with the benefits of optical interconnects thereby utilizing the best of both worlds.

Artificial neurons like those described in Section 1.1 will be created on a 2-dimensional OEIC chip using circuits. Neurons within this 2-dimensional array are electrically connected to one antoher. Because the chips have been fabricated to accept optical input on one side and produce optical output on the other side, they can be cascaded in the third dimension. This allows the neurons in one OEIC plane to communicate optically with neurons in the next OEIC plane. Holographic elements are used to steer the light between inter-plane neurons. Figure 1-4 shows a simplified version of the cascaded processor.

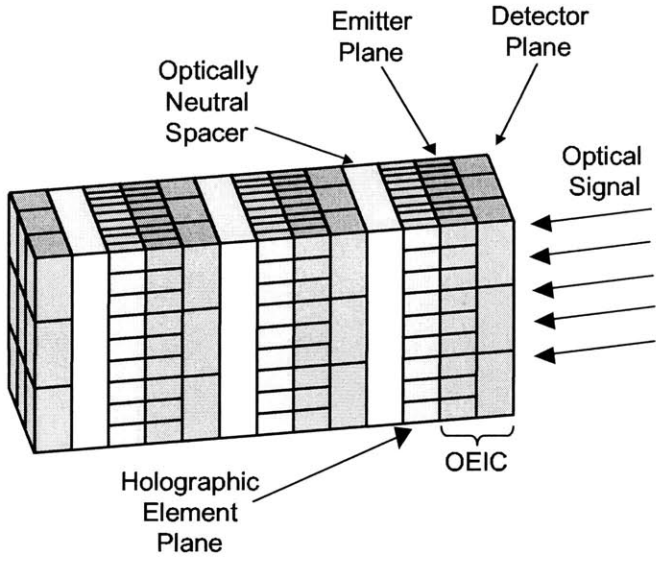

Figure 1-4: A simple example of the layer structure of the CONNP

In this figure, the layered structure of the Compact Optoelectronic Neural Network Processor (CONNP) is seen. A version of the processor with only nine neurons per layer (3×3) is shown for visual simplicity. Each "layer" of the processor consists of four planes: a detector plane, an emitter plane, a holographic element plane, and an optically neutral spacer. The detector and emitter planes are fabricated on a

single substrate and can be grouped together as the optoelectronic integrated circuit (OEIC). After the emitter plane, there is a holographic element plane that directs and structures the light beams from the emitters. Finally, there is an optically neutral spacer that allows the beams passing through the holograms to propagate at the correct angles to reach their destination on the following detector plane. This four plane "layer" can then be cascaded. Figure 1-4 shows three such layers.

As was mentioned previously, in addition to being compact, ruggedness is also a requirement for the CONNPP. It would be very undesirable for the processor to come out of alignment when moved or shipped. In order to make the processor rugged, the layers will be precisely aligned and then cemented into place using an optically neutral glue or epoxy. Once all the desired layers are cemented into place, the processor will no longer be subject to possible misalignment when it is bumped or jarred.

In addition to the optical interconnects between layers, there are also electrical interconnects to the the OEICs. These electrical interconnects allow the CONNP to interact with the PC controlling it. In addition to presenting input to the CONNP, the PC can potentially be responsible for training the neural network. During training cycle, the input will be presented to the CONNP and it will produce an output. This output can then be evaluated by the PC and the corrections to the connection weights can be calculated. These new updated weights can then be scanned into the CONNP for the next training cycle. Additionally, this allows the PC to store connection weights for the CONNP after it has been fully trained. By doing this, the collection of weights produced for a particular training session can be loaded just as a program is loaded by a traditional processor.

1.3.2 Low-level overview of the CONNPP

Once the basic structure of the CONNP is understood, it can be examined at a lower level to uncover some of the challenges faced in its design. It is useful to examine a single neuron in order to better understand what is happening. In Figure 1-5, a single neuron view of each plane is shown.

The detector plane for a single neuron consists of a single photodetector. The

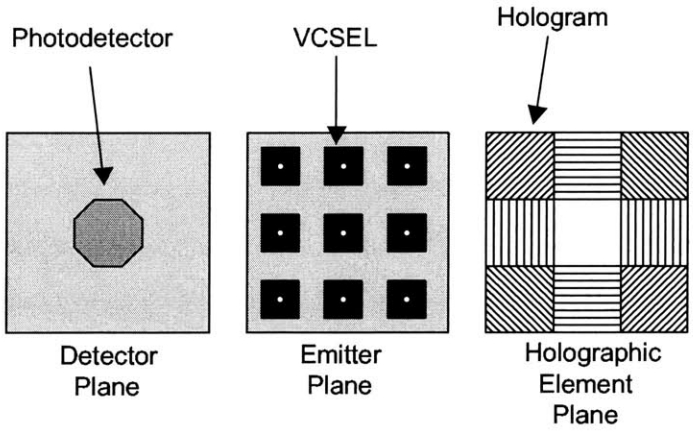

Figure 1-5: Single neuron view of the CONNP planes

emitter plane, on the other hand, has nine vertical cavity surface emitting lasers (VCSELs). The beams if these nine VCSELs then pass through nine separate holograms in the holographic element layer. The number nine comes about due to the optical interconnect scheme used.

In order to fix the number of inter-plane optical interconnects while still taking advantage of optical density benefits, an interconnect scheme known as "nearest-neighbors" has been chosen for the CONNPP. This means that a neuron in one layer can communicate with its nearest-neighbor neurons in the next layer. As the neurons are laid out in a grid structure, this allows a neuron in layer 'A' to communicate with nine neurons in layer 'B'. Illustrating this concept, Figure 1-6 shows a one-dimensional cross section of a single neuron communicating with multiple neurons in the next layer. Extending this one-dimensional symmetry to two-dimensions, the 1:9 fan-out ratio can be seen.

Another important characteristic of the individual neuron to discuss is its size. The specification for the CONNPP calls for the artificial neurons on the OEIC to be $250\mu\text{m}\times250\mu\text{m}$. This leads to a neuron density of about 1600 neurons/cm^2. As mentioned, each neuron contains nine VCSELs, a photodetector, and electronic circuitry. While the size of the VCSELs has not yet been determined, the size of the photodetector will be between $40\mu\text{m}\times40\mu\text{m}$ and $75\mu\text{m}\times75\mu\text{m}$. Furthermore, each

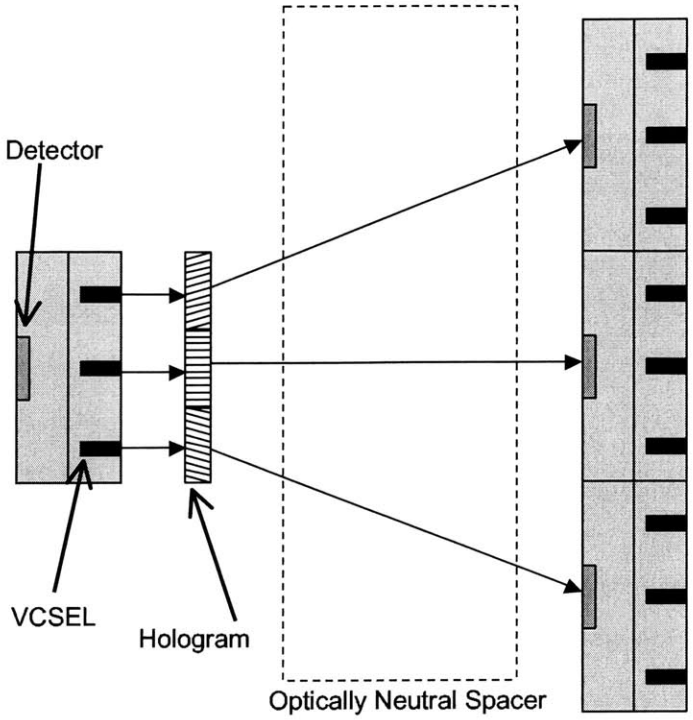

Figure 1-6: One-dimensional cross-section of a single neuron communicating with multiple neurons in the next layer

neuron also has nine individually written holograms associated with it. This means that the size of the individual holograms is approximately $83\mu m \times 83\mu m$.

1.4 Thesis Summary

The remainder of this thesis consists of three chapters encompassing the body of work completed. Chapter 2 explores holographic interconnects for the Compact Optoelectronic Neural Network Processor. Background on holograms is given, analytical holographic writing models are developed, and several systems are tested. Chapter 3 focuses on circuits and optoelectronic devices for the CONNPP. In this chapter, the two chips designed and their constituent components will be discussed. Finally, Chapter 4 examines a newly proposed hardware connection model for the artificial neurons in the CONNP. The suggested hardware architectural changes will be discussed along with neural network simulations showing the impact of these architectural changes. In

all sections, suggestions for future work on each of these topics will also be included.

Chapter 2

Holographic Interconnects

As discussed in Chapter 1, micro-fabricated wires have traditionally been used as interconnects within microprocessors; however, as the clock speed and complexity of microprocessors increases exponentially, these wire interconnects have a difficult time keeping up. The parasitic capacitance and resistance of the wires limits the clock speed at which these wire interconnects can be driven. Likewise, the large distances that the interconnects must travel relative to the clock period and propagation velocity of the signal can cause very large skews in the arrival times of signals. Holographic interconnects allow for the possible elimination of many of the challenges that plague traditional wire interconnects.

Section 2.1 will give the reader the necessary background on holograms and how they are fabricated. The research presented here fits within the larger context of the Compact Optoelectronic Neural Network Processor Project which relies heavily on holographic interconnects. Section 2.2 will build on the CONNPP overview given in Chapter 1 and show how holographic interconnects are implemented

Using some of the implementation-based assumptions made in Section 2.2, a model will be developed in Section 2.3 that will determine the conditions needed to fabricate the holographic interconnections for the Compact Optoelectronic Neural Network Processor Project.

Three different setups were tested to fabricate holographic interconnections. Section 2.4 will present these different setups as well as show results from the holograms

created.

Finally, Section 2.5 will synthesize the learnings from the previous sections into suggestions for future work.

2.1 Holographic Interconnect Overview

Holograms are the fundamental element of holographic interconnects. As such, a brief introduction to holograms and their fabrication will be given. Various systems in which holographic interconnects are useful will be shown.

2.1.1 What are holograms?

When an optical wavefront interacts with an object, the wavefront's amplitude and phase is changed based on the physical structure of the object. The complex wavefront resulting from the wavefront-object interaction now contains a vast amount of information about the structure of the object (Figure 2-1). It is this structural information encoded on the wavefront that the human eye is able to interpret as the visual characteristics of an object (i.e. shape, texture, relative size, etc.).

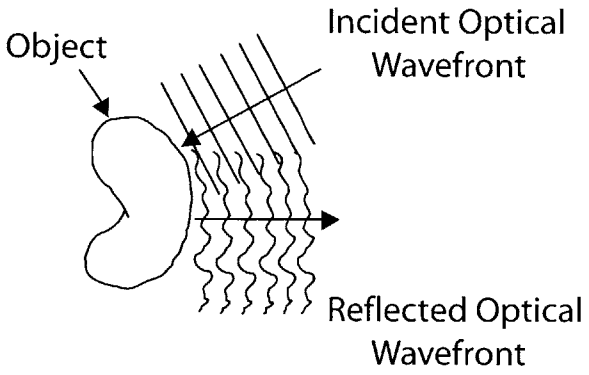

Figure 2-1: A plane wave reflects off an object and the reflected wave then contains information about the structure of the object.

It is important to note that the human eye is only able to detect the intensity of the wavefront, not the phase. Also, because of its small size, the eye is only able

to capture a very small part of the total wavefront intensity at once; hence, it only receives a very small amount of the total information contained in the wavefront about the object at any given moment. When the eye moves to a different position however, it now intercepts a different portion of the wavefront and receives other information about the object. Having two eyes allows humans to obtain two separate portions of the wavefront simultaneously. The brain is then able to fuse the information from each eye into a single scene with more information than either of the constituent images. Three-dimensional perception is a direct result of the ability to fuse the two sets of wavefront information captured by the eyes.

A camera is analogous to a single eye. Like the eye, it has a very small aperture through which part of the wavefront can pass. Therefore, the camera captures a small portion of the wavefront at a single moment in time and records the intensity of that portion of the wavefront on film (Film, like the eye, can detect intensity, not phase). As a result, only that particular view, from that particular moment can be recovered from the photograph no matter where the eye moves with respect to the photograph.

A hologram, on the other hand, is fundamentally different from a photograph. Whereas the photograph only contains intensity information, a hologram contains both intensity and phase information about an object. For a photograph, it was said that because only a small portion of the wavefront passed through the aperture of the camera, only a very small portion of the information was able to be recorded. Because holograms are recorded in a fundamentally different way, the hologram is able to record a large part of the information contained in the wavefront.

Another way to think about holography is as wavefront reconstruction. If an optical wavefront was produced that was identical to the reflected optical wavefront in Figure 2-1, it would appear to the eye (or any recording medium) as if the object was in the same position as Figure 2-1. This is precisely what a hologram does. It takes an optical wavefront as its input and it outputs a reconstructed wavefront with properties identical to the object's reflected optical wavefront (Figure 2-2). Therefore, just like the object, the hologram is able to produce an image that has 3-dimensional qualities such as depth of field and parallax.

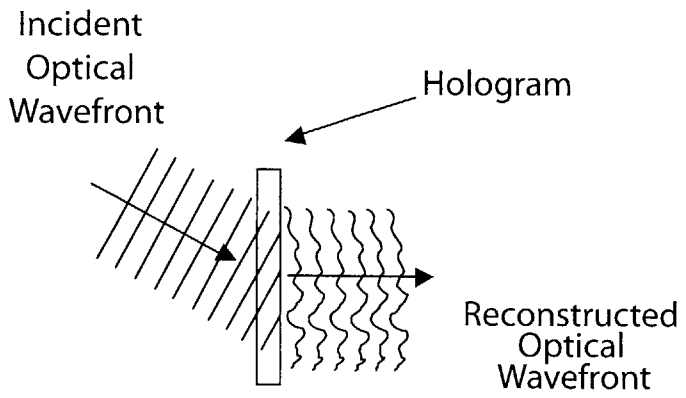

Figure 2-2: A plane wave is incident on the hologram and the reconstructed object wavefront results

2.1.2 How are holograms made?

As stated previously, holograms and photographs are made in fundamentally different ways. Photography relies on ray optics and simple imaging properties of lenses. Holography, on the other hand, relies on the use of wave optics along with interference and diffraction effects. This is why holograms, unlike photographs, are able to record both intensity and phase information.

For now, a very high level explanation of how holograms are made will be given. A more rigorous treatment dealing with the theory and mathematics behind making holograms will be given in Section 2.3.

To make a hologram, two mutually coherent optical waves, a reference wave and an object wave, are interfered in the plane of an intensity sensitive emulsion (Figure 2-3). The reference wave is usually a plane wave, while the object wave is a wave containing phase and amplitude information about an object (i.e. the reflected optical wavefront from Figure 2-1). As discussed before, photographic and holographic emulsions are not able to directly record phase information. However, by interfering the reference wave and the object wave, the phase information contained in the object wave is transformed into amplitude information and is able to be recorded in the emulsion.

Due to the fact that the emulsions are extremely light sensitive, all light except the object and reference waves must be eliminated from the system. As a result,

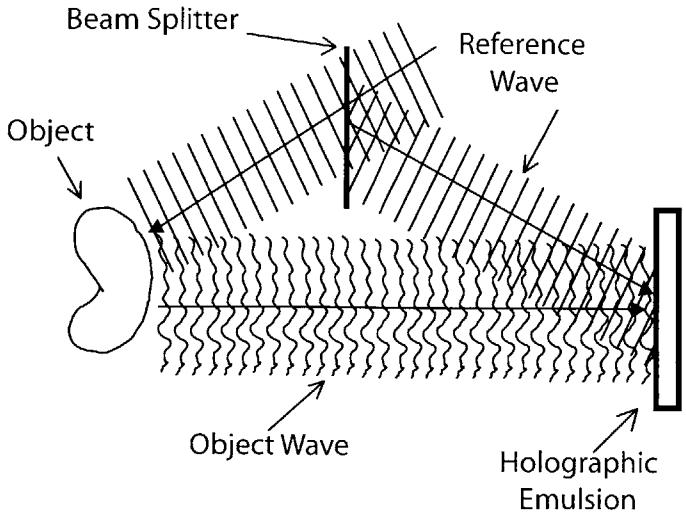

Figure 2-3: The object and reference waves interfere at the holographic emulsion to record a hologram.

holograms are usually recorded in a dark room. When the room is completely dark, the light source providing the coherent waves (usually a laser) is turned on and the emulsion is exposed to the interference pattern. When the photographic emulsion has been exposed for the proper amount of time and chemically processed in the correct manner, the interference pattern will be recorded in the emulsion.

To playback the hologram, the reference wave is input into the emulsion. If the hologram were an ideal Bragg transmission hologram made with a thick emulsion (Discussed further in Section 2.3), the object wave would be present on the opposite side of the emulsion. This is the situation depicted in Figure 2-2.

2.1.3 Holographic interconnects explained

Holographic interconnects are a specific implementation of optical interconnects. The same basic emitter, propagation medium, detector structure that defines optical interconnects in general is also used for holographic interconnects. The characteristic that distinguishes them as their own class of optical interconnect is a holographic element in the propagation medium portion of the system. This holographic element can be used to direct the light wave, change the structure of the light wave, or encode

additional information in the light wave. These operations are quite important to optical interconnects as a whole.

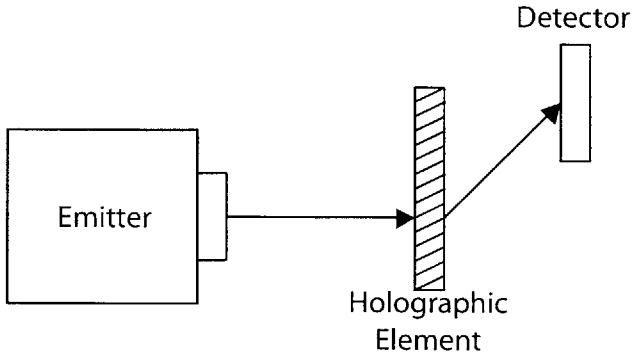

Figure 2-4: A simple holographic interconnect with a free-space propagation medium

Changing the direction of a light wave is one of the most crucial functions to optical interconnects in general. While there are many different ways to do this, holograms provide some of the most elegant solutions. It is well known that light (a plane wave for example) travels in a straight line unless acted upon by an outside optical element. For instance, a mirror use simple laws of reflection (angle of incidence equals angle of reflection) to send the beam in the desired direction. While a mirror is the most basic element used to change the direction of a beam of light, it is not always the most practical element for the job.

Imagine a situation where there is a significant density of beams to be sent in different directions as in Figure 2-5(a). This is a situation where the emitters are on one chip and the detectors are on another chip. The goal is to direct the light from emitter "A" to detector "A", from emitter "B" to detector "B", etc. While it might be possible to place mirrors in the beam paths to correctly map the emitters to the detectors (Figure 2-5(b)), it would most likely be impractical. If separation of the beams is small (tens or hundreds of microns), placing and aligning the four mirrors in the individual beams would be an extremely difficult and time-consuming task. Another alternative is the use of holograms. A hologram can be made using previously discussed techniques to direct the different beams in different directions. It can then be placed in the paths of the beams and aligned as a single element

(Figure 2-5(c)).

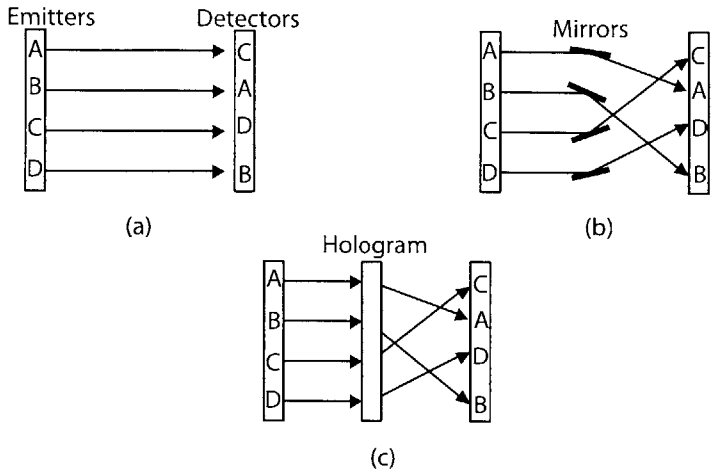

Figure 2-5: (a)Emitters and detectors without any light steering element (b)Mirrors as a light steering element. Notice that detectors and emitters now mapped correctly. (c) A single hologram as the light steering element

Because holograms affect the phase of an incoming light wave, it is possible to cause structural changes in the wave. These structural changes are known as beam conditioning or wavefront shaping [18]. For instance, if an optical signal were being broadcast to a large number of nodes, it might be useful to adjust the divergence of the beam in order to reach all the nodes with a single beam. A beam conditioning hologram would be able to accomplish this.

Because holograms are able to encode both phase and amplitude information, it is possible to use them to modulate the amplitude of the resulting wave. The limitation is that the modulation imparted to the optical wave is static, or not changing with time. Once a hologram is written, it will only be able to amplitude modulate an incoming beam in one particular way. While this might not seem useful at first glance, one can imagine situations where the interconnection hologram could represent a cryptographic hardware key. Without the correct combination of amplitude modulations for the various interconnects, the device in question would not be able to function.

Additionally, holograms are able to perform the functions described above simul-

taneously. This increases the class of problems that holographic interconnects are suited to solve. In the next section, a system for which holographic interconnects are ideally suited will be explored.

2.2 Holographic Interconnections for the Compact Optoelectronic Neural Network Processor Project

In Chapter 1 an overview of the Compact Optoelectronic Neural Network Processor Project (CONNPP) was given. This section will focus on giving the reader a more in depth view of the holographic interconnect elements described in the overview. Specific challenges facing the holographic interconnect elements for the Compact Optoelectronic Neural Network Processor Project will also be discussed (i.e. device non-idealities). Later sections will show ways in which to overcome some of these challenges.

Having seen the basics of the compact optoelectronic neural network processor project architecture, some of the components that make up the holographic interconnect (VCSEL, hologram, and detector) can be examined in more detail.

2.2.1 Vertical cavity surface emitting lasers (VCSELs)

While extremely small semiconductor lasers have been available for many years, they have suffered from some distinct shortcomings. One of the most significant shortcomings that these devices display is the fact they they are edge-emitting. This edge-emitting characteristic impacts the traditional semiconductor laser in several ways. First, the fact that the coherent light is emitted from the edge of the structure (parallel to the substrate) means that the devices must be cleaved from the wafer [19] (Figure 2-6). This limits the ability to produce two-dimensional arrays of the devices. Secondly, because of the geometry of the laser cavity, these lasers generally produce an elliptical beam that must be conditioned with non-spherical optics (i.e cylindrical lenses).

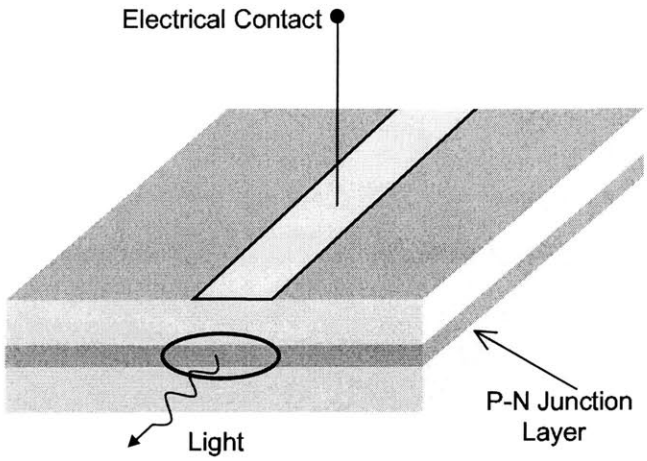

Figure 2-6: Edge emitting semiconductor laser

Vertical cavity surface emitting lasers are designed in such a way that these problems are eliminated. First, as the name implies, the light is emitted in the vertical direction (normal to the substrate) rather than from the edge of the substrate. This can be seen in Figure 2-7. The fact that the light is emitted vertically allows two-dimensional arrays of VCSELs to be easily fabricated. Also, because the contact/aperture of the VCSEL is patterned on the top of the structure via photolithographic means, the output beam can be circular.

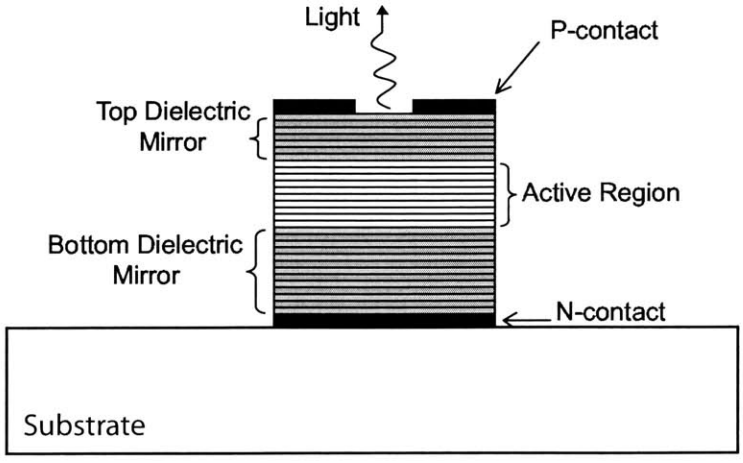

Figure 2-7: Structure of a vertical cavity surface emitting laser (VCSEL)

While some of the problems associated with edge-emitting lasers have been solved

by VCSELs, some still remain. One such problem is beam divergence. Because these devices have extremely small apertures (on the order of 1-10μm) [19] diffraction effects come into play. As is known from diffraction theory, as the aperture size decreases, the divergence increases [18].

For the initial development phase of the CONNPP, two VCSEL arrays were obtained from the U.S.-Japan Joint Optoelectronics Project (JOP) run by the Optoelectronics Industry Development Association (OIDA). These VCSEL arrays contained 64 individually addressable VCSEL structures in an 8×8 configuration on a 250μm pitch operating at 850nm. Also received from JOP were several laser driver chips designed especially for VCSELs.

One of the most critical aspects of VCSEL performance for holographic interconnects is the beam profile. The beam profile takes into account both the divergence of the beam and the mode structure of the beam. Figure 2-8 shows the beam profile from one of the VCSELs in the OIDA/JOP 8×8 array. Each of the five pictures shown is the same VCSEL with a different input current. The power intensity gradually increases from Figure 2-8(a) to Figure 2-8(e).

In looking at Figure 2-8, the most noticeable characteristic is that the mode structure of the VCSEL changes as the power is increased. In Figure 2-8(a), there is only a central bright spot which looks similar to a Gaussian distribution. As the power is increased, the mode structure evolves to a two lobed pattern in Figure 2-8(b), and eventually in Figure 2-8(e) to a ring-like pattern with a local minimum in the center. While the change in modes can be recognized in Figure 2-8, the effect was even more pronounced when viewed in person. Also of note is the fact that the transition between the different modes shown in Figure 2-8 is not completely smooth. As the current into the VCSEL is gradually increased, the pattern will change subtly. At some points, however, the VCSEL appears to "snap" to a completely different mode pattern. The most likely explanation for the mode changing in general and the mode "snapping" in particular is that as the current into the VCSEL is increased, the VCSEL heats up. The VCSEL heating up causes the laser cavity to change shape, supporting different modes. As the device heats up and expands, when a particular

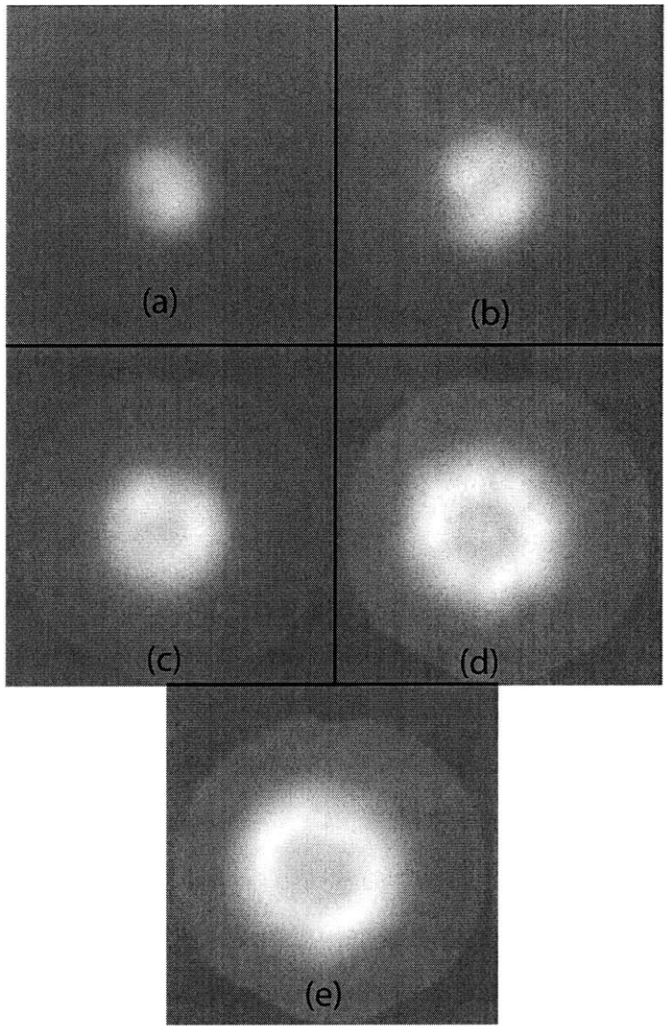

Figure 2-8: Beam profiles of a single VCSEL at different operating powers

mode cutoff is passed, the VCSEL has the potential to "snap" into a different mode pattern.

This changing mode structure of these VCSELs could have a very different impact on different types of systems. If these VCSELs were to be used in an imaging system that required them to be dynamically modulated, the changing mode structure could seriously distort the images produced. The application examined here, holographic interconnects, might be somewhat more forgiving. As long as the light emitted from the VCSEL hits the corresponding detector, the holographic interconnect accomplishes its goal. If the holograms to direct and shape the light are fabricated in such a way that

they direct as much light as possible to the detectors, the changing mode structure should not have much effect on the performance of the holographic interconnects.

The other part of the beam profile to examine is the divergence angle. Looking at Figure 2-8, it appears as if the divergence increases with power. In this situation, the divergence and the mode structure seem to be somewhat coupled. In Figure 2-8(a), the beam appears to be a simple Guassian. As a result, the intensity of the beam trails off significantly as the angle from the center of the beam increases. On the other hand, examining Figure 2-8(e), it can be seen that there is local minimum near the center of the beam and most of the energy is concentrated towards the outside edge of the beam. This modal change in intensity distribution will result in a change in observed divergence angle. The question becomes how does this observed change in divergence angle affect how the holograms are designed.

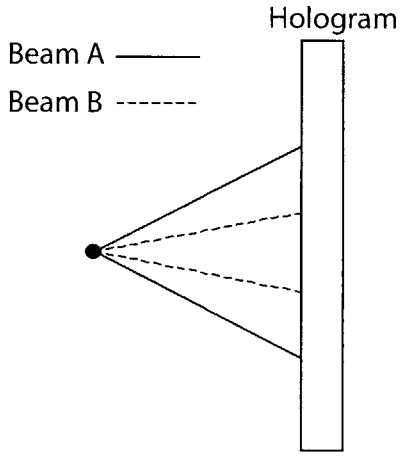

Figure 2-9: Two beams from the same origin point with different divergence angles hitting a hologram

If the position of the VCSEL is fixed in the system and the hologram is fabricated to accommodate the largest observed divergence, a change in divergence angle should not affect the system. This situation is illustrated in Figure 2-9. Two diverging beams ("A" and "B") with the same point of origin but different divergence angles are shown hitting the hologram. As can be seen in the figure, the spatial frequency components that make up beam "B" are a subset of the spatial frequency components that make

up beam "A". As a result, if the hologram works effectively for beam "A", it must also work effectively for beam "B".

The beam with the greatest divergence (Figure 2-8(e)) is used as the divergence of the VCSEL. A half-angle divergence of $\sim 7.3°$ was measured. This value was calculated by determining the width of the beam at the $1/e^2$ point. Therefore, according to the previous analysis, if the hologram is created to accommodate a read beam with a half angle divergence of $\sim 7.3°$, the hologram should work well for the various modes and divergence angles encountered.

2.2.2 Spacing between emitter planes and holographic elements

Because of the non-zero divergence angle of the VCSELs used in the CONNPP, there will be definite limitations on the spacing between the emitter plane and the holographic element plane of the processor. The CONNPP specifies that the pitch of the neurons for the processor will be 250μm. From the previous discussion of the architecture, it is known that each of these $250\mu\text{m}\times250\mu\text{m}$ neurons will contain nine VCSEL emitters. These nine VCSELs will then be directed and conditioned by nine individual holographic elements. These are the same nine holographic elements per pixel that are shown in Figure 1-5. Since the full pixel dimension is $250\mu\text{m}\times250\mu\text{m}$, this means that the individual hologram dimensions must be $83\mu\text{m}\times83\mu\text{m}$. As the VCSEL is moved further away from the holographic element plane, the footprint of the VCSEL beam on the hologram will increase. This is shown in Figure 2-10.

Notice that when the VCSEL is positioned at z_1, the beam footprint fits fully within the middle of the three holograms. When moved to z_2, the VCSEL beam now hits not only the middle hologram, but also the neighboring holograms. This will result in crosstalk between the channels. Given that the VCSELs discussed in the previous section had a half angle divergence of $\sim 7.3°$, this means that the maximum allowed distance distance between the emitter plane and the holographic element plane is about 324μm or about 1/3 mm. Please note that the $\sim 7.3°$ measurement

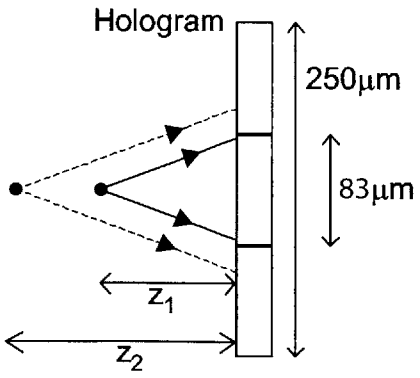

Figure 2-10: How distance of VCSEL from hologram affects crosstalk

was the half-angle at the $1/e^2$ point. If more of the beam needs to be confined to the 83μm hologram, the distance between planes without crosstalk will be less than 324μm.

2.2.3 Uniformity and linearity of VCSELs

As the VCSELs will be used as analog outputs from one layer and analog inputs to another layer, it is important to examine the linearity and uniformity of the devices. To do this, an experiment was setup to measure the optical power as a function of current through the VCSEL. This experiment was done for three different VCSELs on a single 8×8 array chip. For each device the current was increased to its maximum value. The beam from the VCSEL was then focused through a lens, onto a power meter. By aligning the detector at its maximum intensity (i.e. maximum beam size) it is insured that the maximum amount of light will hit the detector at all intensities. The current was then reduced to 0mA and swept to its maximum of 16mA. The results can be seen in Figure 2-11.

While the three different VCSEL curves look reasonably close together, there are potential issues due to uniformity and linearity. In terms of uniformity, ideally all three curves would be identical and have a region that was perfectly linear. As can be seen in the graph, this is not the case. Raw errors of .32mA can be observed between different devices. By averaging the three power values for each current, a curve is

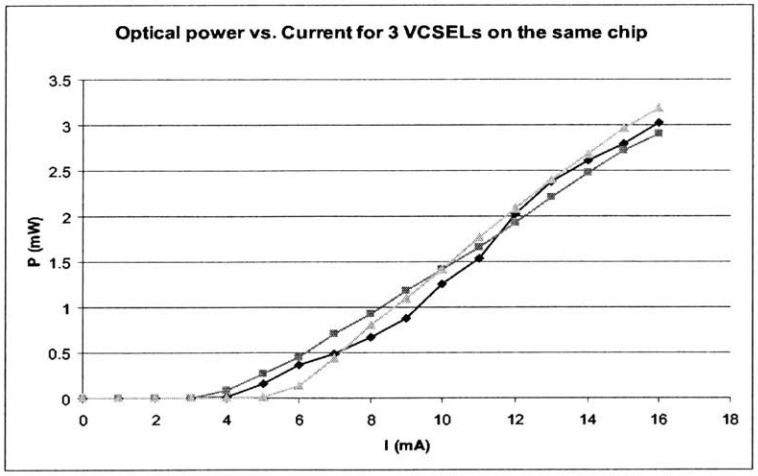

Figure 2-11: Optical power emitted from three different VCSELs on a single chip as a function of input current

obtained that best fits the data. Even in this case, errors as large as .18mW can be seen. This means that if these three devices were being used in a system, a maximum resolution of .18mW could be achieved. This means that due to uniformity issues alone, the VCSELs are limited to 5-bit resolution in the linear region between 6mA and 16mA.

In order to look at the linearity of the devices, the data from 6mA to 16mA is again examined. This time, a linear fit is done for each curve. The standard error between the linear fit and the actual data is then calculated. For the three curves shown, the standard errors from their linear fits are 12.4%, 4.7%, and 2.7%. Please keep in mind that this is a best case scenario because a linear fit was done for each curve. If instead the power values are averaged at each point and the linear fit is then applied to this averaged curve, the standard errors for the three original curves will increase. Because of the 12.4% standard error in the linearity, the resolution is limited to about 3-bits.

Because of these uniformity and linearity based resolution issues, it is important to keep very tight control over the performance of the VCSEL devices. The VCSELs tested here are older (1997) devices provided through a government program. As a result of this, they are no longer state-of-the-art. The VCSEL devices for the CON-

NPP will be fabricated in house. It will be important for the parameters discussed here to be tracked with the VCSELs fabricated in-house in order to optimize the performance of the holographic interconnects.

As was discussed in Chapter 1, neural networks offer the benefit of being marginally fault tolerant. This fact may allow the CONNP to overcome some of the VCSEL accuracy limitations discussed above. A much more thorough neural network architecture study will be needed to determine which parameters the system is particularly sensitive to and which can be ignored.

2.2.4 Photodetectors

As discussed previously, the detector plane of the holographic interconnect contains a single photodetector for each neuron in the plane. In the final CONNPP implementation, the photodetector is specified to be a gallium arsenide (GaAs) pin diode. Just as the VCSELs, the photodetectors will be fabricated on a substrate separate from the neural network circuits. Once the devices are fabricated, a process known as Aligned Pilar Bonding will be used to transfer the devices from the GaAs substrate to the silicon on sapphire substrate that contains the neural network circuits.

The size of the photodetectors is not explicitly stated in the specifications of the CONNPP. There is a definitive tradeoff that can be made with the size of the photodetectors and alignability of the beams on the photodetectors. An important fact to remember is that according to the architecture, the beams from the nine nearest neighbors in the previous plane must be aligned through the holographic element and then onto the single photodetector. If the photodetectors are made large (i.e. $100\mu m \times 100\mu m$) it will be much easier to align the nine beams onto the single detector. A bigger target is easier to hit. On the down side, making the detectors large also takes up valuable chip real estate. With $100\mu m \times 100\mu m$ photodetectors and $250\mu m \times 250\mu m$ neurons, the photodetectors consume 16% of the total neuron area. This does not leave much room for VCSELs and neural circuitry. If the detectors are made very small ($25\mu m \times 25\mu m$) they consume only 1% of the total neuron area, but they will be much more difficult to align. Once a test system is built after the next

iteration of chips, a study will be done to assess the alignment accuracy achievable. This will in large part determine the necessary photodetector size.

2.2.5 Holographic element specifications

The holographic interconnect system consists of three elements: the emitter, the holographic element, and the detector. Now that the characteristics of the emitters (VCSELs) and detectors (GaAs pin devices) are known, the holographic element specifications can be derived. The goal of each holographic element is to direct as much light from its corresponding VCSEL to its intended photodetector. The emitter plane and holographic element plane are aligned such that they are parallel. As the VCSELs emit light vertically, the beams from the VCSELs will be normal the holographic interconnect plane. As can be seen in Figure 1-6, most of the beams exiting the hologram are not normal to the plane. This means that the hologram must change the direction of the VCSEL beam.

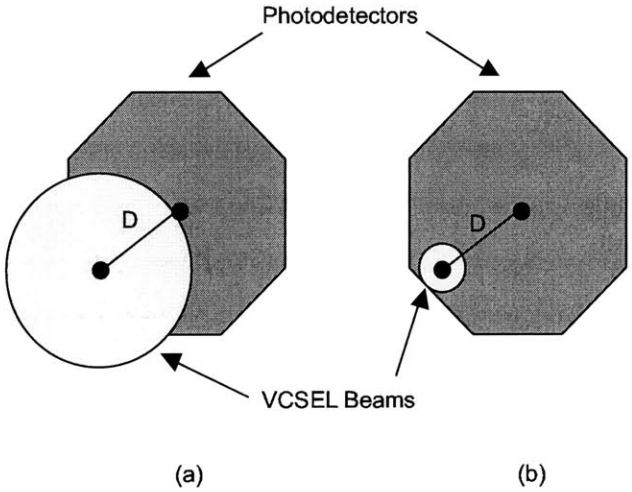

Figure 2-12: Two beams with the same energy, but different sizes hitting a photodetector

It was shown previously that light emitted from the VCSELs can have up to a $\sim$ 15° full angle divergence. It was also discussed that the photodetector sizes will be somewhere between 25μm$\times$25μm and 100μm$\times$100μm. However, the beam

must propagate a minimum distance in order to travel laterally to hit the nearest neighbor photodetector. As a result, if nothing is done to condition the beam, the divergence will cause the beam to be significantly larger than the photodetector. In some cases, this can result not only in a degraded signal at the photodetector, but also crosstalk among the channels. For this reason, it is important to also use the hologram to correct the divergence of the beam. While it might work to only correct the divergence (i.e. plane wave output), it might prove useful to make the beam convergent. By focusing the beam onto the detector plane, maximum alignment flexibility is achieved. This is illustrated in Figure 2-12.

In Figure 2-12(a) larger beam is seen incident on the photodetector. The larger beam is misaligned by a distance D. As a result, a significant portion of the beam does not hit the detector. In Figure 2-12(b) the same amount of energy is incident on the detector plane, but it is focused to its minimum size. Once again beam is misaligned by a distance D. Because the beam is so much smaller, all of the incident light still hits the photodetector. For this reason, it is best to have the beam focused in the plane of the photodetector for the production system.

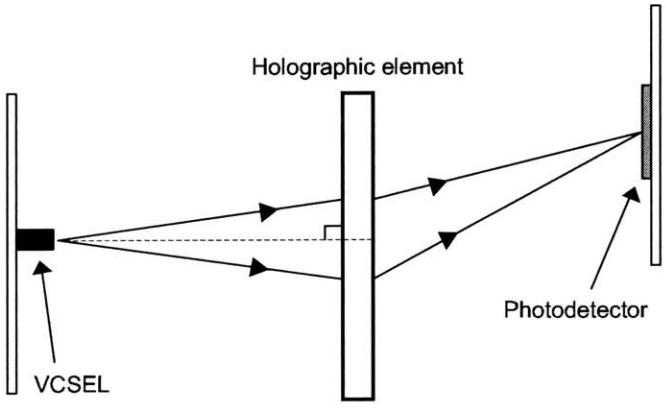

Figure 2-13: Single holographic interconnect

With this information, a complete picture of a single holographic interconnect can be developed. The VCSEL will emit a divergent beam normal to the plane of the hologram. The hologram will then redirect the angle of the beam and focus it onto the plane of the detector. This can be seen in Figure 2-13. Now that a specification

for the holographic interconnect has been developed, a model can be designed to determine exactly how to write holograms like this.

2.3 Modelling Holographic Interconnections

Modelling holographic interconnections begins with understanding the physics and math behind interference theory. From that, it can be understood mathematically how simple plane wave holograms are created. The Bragg condition and how it relates to both writing and reading holograms will be discussed. From there a simple model will be developed to write holograms at one wavelength and read them at a different wavelength. This simple model will be developed further to one where the hologram produced can take the divergent input beam and produce a focused output beam. Volume aspects of the holograms will also be considered.

2.3.1 Interference theory

The first step toward creating holograms is understanding interference. Interference is observed when two or more mutually coherent optical waves are present in the same space. The wave resulting from the interference is the algebraic sum of the individual waves [20].

$$U_{tot}(\mathbf{r}) = U_1(\mathbf{r}) + U_2(\mathbf{r}) \tag{2.1}$$

Where $U(\mathbf{r})$ is a complex, monochromatic wave. When this wave is detected, its intensity is the quantity that will be recorded. We know that

$$I \propto |U|^2 \tag{2.2}$$

where I is the intensity. So the intensity for $U_{tot}(\mathbf{r})$ would be

$$I_{tot} = |U_1 + U_2|^2 = |U_1|^2 + |U_2|^2 + U_1(\mathbf{r})U_2(\mathbf{r})^* + U_1(\mathbf{r})^*U_2(\mathbf{r}) \tag{2.3}$$

This intensity, I_{tot}, is what would be detected or recorded. Notice that the intensity of the resultant wave, I_{tot}, is significantly different from the intensities of the individual waves, $|U_1|^2$ and $|U_2|^2$. The first two terms of the resulting intensity are only magnitudes, they have no complex phase term. These terms are what we would expect to see as the intensities of $U_1(\mathbf{r})$ and $U_2(\mathbf{r})$ individually. The last two terms in the intensity equation, $U_1(\mathbf{r})U_2(\mathbf{r})^*$ and $U_1(\mathbf{r})^*U_2(\mathbf{r})$, are complex waves.

To illustrate what is happening, consider a very simple example where two coherent plane waves at different angles are incident on a recording medium.

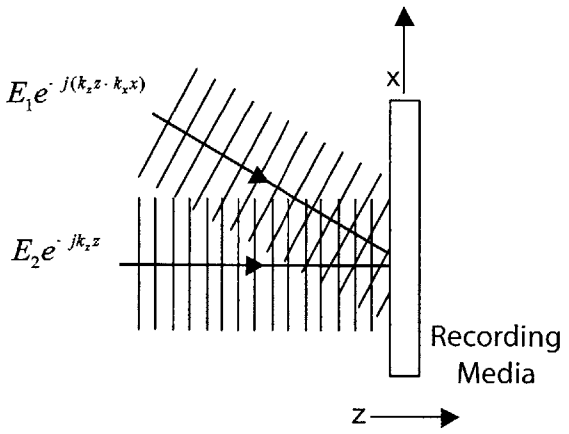

Figure 2-14: Interference pattern of two plane waves being recorded.

As seen in Figure 2-14, the plane waves are of the form

$$\mathbf{E}_1 = \mathrm{E}_1 e^{-j(k_z z - k_x x)} \tag{2.4}$$

and

$$\mathbf{E}_2 = \mathrm{E}_2 e^{-jk_z z} \tag{2.5}$$

At the recording media, the wave resulting from the interference of $\mathbf{E}_1$ and $\mathbf{E}_2$ is

$$\mathbf{E}_{\mathrm{tot}} = E_1 e^{-j(k_z z - k_x x)} + E_2 e^{-jk_z z} \tag{2.6}$$

From Equation 2.3, the resulting intensity at the recording media is

$$I_{tot} = |E_1|^2 + |E_2|^2 + \mathbf{E}_1\mathbf{E}_2^* + \mathbf{E}_1^*\mathbf{E}_2 \tag{2.7}$$

$$I_{tot} = |E_1|^2 + |E_2|^2 + E_1E_2e^{-jk_xx} + E_1E_2e^{+jk_xx} \tag{2.8}$$

$$I_{tot} = |E_1|^2 + |E_2|^2 + E_1E_2\left(e^{-jk_xx} + e^{+jk_xx}\right) \tag{2.9}$$

This is the intensity pattern that will be present at the plane of the recording medium.

2.3.2 Recording holograms

It is at this point that the type of recording media being used must be considered. For the purposes described here, there are two different types of recording media, thin film emulsions and thick film emulsions. A thin film emulsion has a thickness on the order of the period of the interference pattern being recorded. When a thin film emulsion is used, the interference pattern at the surface of the emulsion is what is recorded. It can basically be thought of as a two dimensional surface on which intensity pattern is being recorded. A thick emulsion, on the other hand, has a thickness that is much greater than the period of the interference pattern being recorded. Whereas the thin film emulsion only recorded the interference pattern in a single plane, the thick emulsion records the interference pattern at each plane throughout the volume of the emulsion.

It is also useful to consider the mechanism by which the interference pattern is recorded by the emulsion. As discussed in Section 2.1, photosensitive materials respond to the intensity (amplitude squared) of a wave. Some photosensitive materials, semiconductors for instance, respond to intensity by generating an electrical current. Other photosensitive materials, like photo-polymers and photo-resist, respond to incident light by polymerizing. Photographic emulsions respond by a chemical reaction that involves converting silver halide into silver. The emulsion can then be bleached

making the silver into a transparent substance with a high index of refraction [21]. This means that by exposing a photographic emulsion to light and chemically processing it, the light intensity distribution is recorded as a variation in thickness of a high index of refraction material along the emulsion. If a thick film emulsion is used, high-index of refraction planes will result in the material. Traditionally, high-density photographic emulsions are used for holography. These emulsions may be either thick or thin film emulsions. Photo-polymers which result in an index of refraction modulation will also be considered in later sections. These materials are of particular interest due to their high diffraction efficiency and easy processing.

Returning to Equation 2.9, it is seen that the intensity pattern at the plane of the recording medium consists of a constant term, $|E_1|^2 + |E_2|^2$, and two terms with a phase component, $E_1 E_2 e^{-jk_x x}$ and $E_1 E_2 e^{+jk_x x}$. This is the pattern that will be physically recorded in the holographic recording medium. The transmission function of the hologram will then be

$$t_{holo}(x,z) = A\left\{\left(|E_1|^2 + |E_2|^2\right) + E_1 E_2 e^{-jk_x x} + E_1 E_2 e^{+jk_x x}\right\} \tag{2.10}$$

where $t(x,z)$ is the transmission function and A is a constant that takes into account response of the emulsion and exposure time.

When a thin film emulsion is used, this pattern will be recorded at the surface of the emulsion. If the hologram is then played back with a plane wave identical to $\bar{\mathbf{E}}_2$, the result will be

$$t_{holo}(x,z)\bar{\mathbf{E}}_{\mathbf{2}} = A\left\{E_2\left(|E_1|^2 + |E_2|^2\right)e^{-jk_z z} + E_1|E_2|^2 e^{-j(k_x x - k_z z)} + E_1|E_2|^2 e^{+j(k_x x - k_z z)}\right\} \tag{2.11}$$

Equation 2.11 has three distinct terms. The first term is a plane wave propagating in only the $+z$ direction. The second term is a plane wave propagating in the $+z$ and $-x$ directions. Finally, the third term is a plane wave propagating in the $+z$ and $+x$ directions. This can be rewritten as

$$A\left(|E_1|^2 + |E_2|^2\right)\mathbf{E_2} + A\,|E_2|^2\,\mathbf{E}_1 + A\,|E_2|^2\,\mathbf{E}_1^* \tag{2.12}$$

From Equation 2.12 it can be seen that the first term is an amplitude modulated version of the input wave $\mathbf{E}_2$. The second term is a replica of the original $\mathbf{E}_1$ with a modulated amplitude. Finally, the third term is the conjugate of $\mathbf{E}_1$ also with a modulated amplitude. This is illustrated in Figure 2-15.

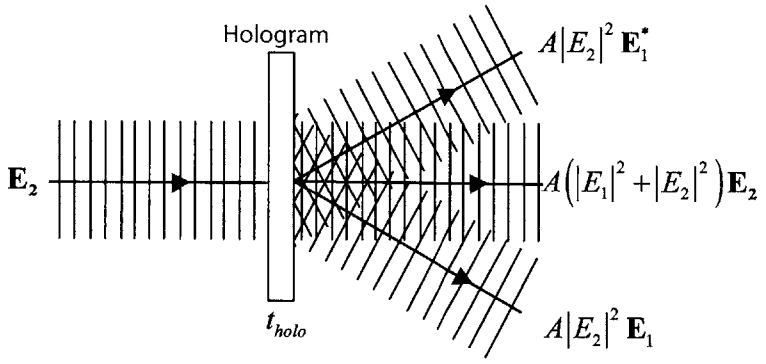

Figure 2-15: Playback of the thin emulsion hologram with wave $\mathbf{E}_2$

In the example shown above, only plane waves were used to write the hologram. If one of the plane wave sources is replaced by a wave with structure from an object, similar results will be seen when the hologram is played back with the reference plane wave. There will be an undiffracted plane wave with an amplitude modulation, a modulated object wave (virtual image), and a conjugated object wave (real image) all propagating in different directions.

While thin film holograms might be useful for some holographic interconnect applications, they are not appropriate for the implementation discussed in the previous section. The reason for this is because when the thin film hologram is played back, three beams result at the output. In a dense, point-to-point interconnection scheme, this will cause massive cross-talk between various nodes. A more desirable situation would be if one beam was input, and a single beam with the correct properties was output.

For this application, thick film holographic emulsions are most appropriate. Thick

film holograms are written in exactly the same way as thin film holograms, the difference is in what is recorded. As discussed earlier, while thin film emulsions only record the interference pattern at the surface, thick film emulsions record the interference pattern at each plane for the entire thickness of the emulsion. Thick film holograms rely heavily on Bragg theory discussed next.

2.3.3 Bragg theory

Before getting into the specifics of how to write holograms with the properties seen in Figure 2-13, it is useful to quickly review interference conditions for writing holograms and the Bragg condition for reading out the holograms. To write holograms, two mutually coherent waves are interfered and the intensity of their interference pattern is recorded in the holographic emulsion. In the case of two mutually coherent plane waves of wavelength λ separated by an angle θ, a sinusoidal interference pattern with a period Λ is produced. The period Λ (also known as the grating spacing) is given by [20]

$$\Lambda = \frac{\lambda}{2\sin(\theta/2)} \tag{2.13}$$

One important point to note is that θ is the angle between the beams inside the emulsion. Generally, the index of refraction of the holographic emulsion is larger than the index of refraction of free space. As the interference pattern is recorded inside the emulsion and the write beams are given external to the emulsion, this difference in index of refraction between free space and the emulsion must be accounted for. This is accomplished by simply applying Snell's Law[1] to the beams external to the emulsion. In this brief review of Bragg theory, it will be assumed that the angles shown are internal angles (i.e. the angle inside the emulsion).

The intensity of this sinusoidal interference pattern with grating spacing (Λ) is what is recorded in the holographic emulsion. If the emulsion is a thin film emulsion this pattern is basically only recorded on the plane of the surface. In this case a

[1] $n_1 sin\theta_1 = n_2 sin\theta_2$

hologram is produced that has three output terms. For the holographic interconnects specified in the previous section, it was desirable to have only a single output term. This can be achieved with thick film holograms. Thick film holograms have an emulsion thickness much greater than the grating spacing (Λ). Thick film holograms allow the interference patterns to be recorded at each plane throughout the thickness of the hologram. When the holographic emulsion is processed, the index of refraction of the material changes as a function of the intensity of the light. Regions that received more optical energy during exposure have a higher index of refraction. This results in the hologram having planes of high index of refraction material. These planes, in turn, can act as partially silvered mirrors to incoming light. Figure 2-16 shows two plane waves interfering inside a thick emulsion and also the emulsion after processing. The slightly angled parallel lines in the processed hologram represent the partial mirror planes.

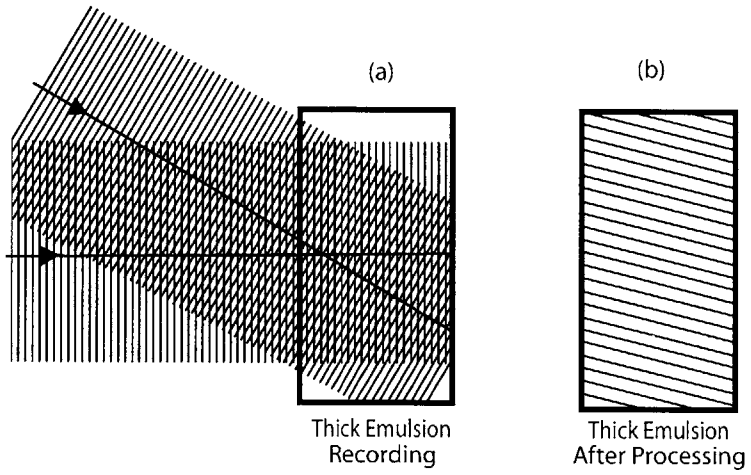

Figure 2-16: (a)Two plane waves interfering inside a thick emulsion while a hologram is being recorded (b) The same thick emulsion after chemical processing

When a beam of light hits one of these partial mirrors, a small portion of the beam will be reflected but most will pass directly through the weak mirror. As an important note for later analysis, the angle of inclination for these partial mirrors is the bisector of the the two recording beams. This can be proven by examining the

interference pattern with respect to the propagation vectors of the waves.

Bragg theory was a result of research into X-ray diffraction to determine crystal structure. In this case, the crystal planes acted as partial reflectors introducing multiple beam interference effects [22]. While developed for the X-ray region of the electromagnetic spectrum, Bragg theory can be just as easily applied to the visible and infrared regions as well.

Because the thick hologram is made up of many thin layers of dielectric media the phenomenon of multiple beam interference will come into play. Looking at Figure 2-17, two of the partial mirrors (n and $n+1$) are seen separated by a distance Λ . Two beams are seen entering at an angle α with respect to the partial mirrors.

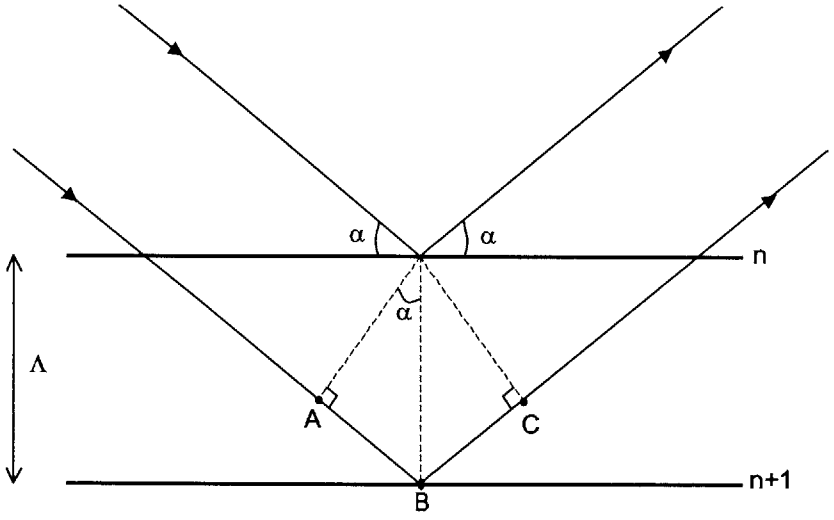

Figure 2-17: The diagram illustrates the derivation of the Bragg Condition

The first beam is reflected off of partial mirror n. The second beam passes through partial mirror n and is subsequently reflected by partial mirror $n+1$. Assuming these two beams are coherent to begin with, they are in phase up until the dashed line at point A. Therefore, the second beam travels a longer distance than the first beam. As a result of travelling a longer distance, the second beam also accumulates more phase. If the two beams are to interfere constructively when they exit the hologram (to obtain the maximum signal), it is necessary that they still be in phase. For this to happen, the extra distance travelled by the second beam must be equal to an integral

number of wavelengths ($m\lambda$) of the beam. This means that for the case where $m = 1$,

$$\overline{AB} + \overline{BC} = \lambda \tag{2.14}$$

Examining the diagram

$$\sin(\alpha) = \frac{\overline{AB}}{\Lambda} = \frac{\overline{BC}}{\Lambda} \tag{2.15}$$

which means that

$$\overline{AB} = \overline{BC} = \Lambda \sin(\alpha) \tag{2.16}$$

Substituting back into Equation 2.14,

$$2\Lambda \sin(\alpha) = \lambda \tag{2.17}$$

Equation 2.17 is known as the Bragg condition. The Bragg condition says that given a hologram with partial mirror (grating) spacing Λ and input beam wavelength λ, the output signal (diffracted beam) will be a maximum when the internal angle with respect to the grating is α.

When the holograms are written, the an interference pattern is produced with the grating spacing given by Equation 2.13. This grating spacing is then used in the Bragg condition (Equation 2.17) to determine the internal read out angle (α) that gives the maximum diffracted signal. While the equation for the interference pattern and the Bragg condition are essentially the same, it is important to think about them separately as conditions may change between writing and reading (i.e. different write and read wavelengths).

2.3.4 Writing and reading holograms with the same wavelengths

When making holograms for holographic interconnections, the input beam angle (θ_{in}) and output beam angle (θ_{out}) along with the write and read wavelengths are usually known. Knowing the desired result (the read geometry), the geometry for writing must be calculated. In the simplest case, the write and read wavelengths will be the same. To develop equations for this simple case assume that a hologram is needed with the geometry shown in Figure 2-18. It must then be determined what the correct writing geometry is to produce a hologram with these characteristics.

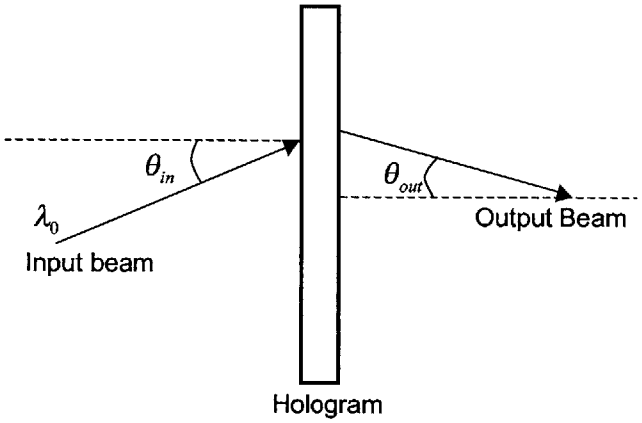

Figure 2-18: Desired readout geometry for hologram

For this basic example assume the hologram is being recorded and played back with the same wavelength (λ_0) of light. In this case, to determine the write geometry the output beam is simply moved the the input side of the hologram such that it intersects with the input beam. These angles external to the emulsion must be converted to angles internal to the emulsion. This is done by applying Snell's Law. The resulting internal angles are denoted with primes.

$$\theta'_{in} = \sin^{-1}\left(\frac{\sin\theta_{in}}{n_e}\right) \tag{2.18}$$

and

$$\theta'_{out} = \sin^{-1}\left(\frac{\sin\theta_{out}}{n_e}\right) \tag{2.19}$$

Where n_e is the index of refraction of the emulsion. This write geometry will produce a hologram with a grating spacing given by

$$\Lambda = \frac{\lambda_0}{2\sin\left(\frac{|\theta'_{in}-\theta'_{out}|}{2}\right)} \tag{2.20}$$

where the normal to the hologram is defined as 0°. To prove that this works, the inclination angle of the grating (partial mirrors) must also be determined. As mentioned before, the angle of inclination for the grating (γ) is the perpendicular bisector of the two internal write beams. This is given by

$$\gamma = \frac{\theta'_{in} + \theta'_{out}}{2} \tag{2.21}$$

The write geometry along with the parameters Λ and γ are illustrated in Figure 2-19.

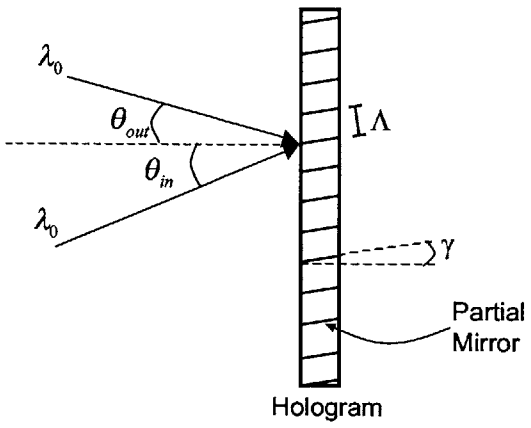

Figure 2-19: Calculated write geometry for hologram

To show that this write geometry produces a hologram with the readout geometry of Figure 2-18 the Bragg condition (Equation 2.17) is used. It is known that the hologram produced has a grating spacing of Λ and a grating inclination of γ. Earlier, α (the Bragg angle) was defined as the internal angle between the read out beam and

grating that produced the maximum output signal. If the hologram works as desired, the input angle with respect to the hologram that produces a maximum output will then be the angle of grating inclination (γ) plus the angle of the input beam with respect to the grating (α). This is illustrated in Figure 2-20 below.

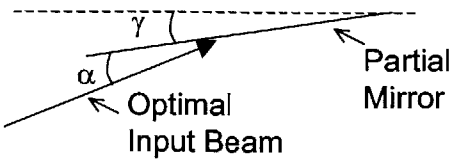

Figure 2-20: Illustration of angles in hologram read out

As seen in the figure above,

$$\theta_{\text{Bragg}} = \gamma + \alpha \tag{2.22}$$

Therefore, if the hologram works, θ'_{in} for the read case will equal θ_{Bragg} for the write case. The Bragg condition is

$$2\Lambda \sin(\alpha) = \lambda_0 \tag{2.23}$$

Substituting for Λ,

$$2\frac{\lambda_0}{2\sin\left(\frac{|\theta'_{in}-\theta'_{out}|}{2}\right)}\sin(\alpha) = \lambda_0 \tag{2.24}$$

Solving for α,

$$\alpha = \frac{|\theta'_{in} - \theta'_{out}|}{2} \tag{2.25}$$

Substituting α and γ_0 into Equation 2.22,

$$\theta_{Bragg} = \frac{\theta'_{in} + \theta'_{out}}{2} + \frac{|\theta'_{in} - \theta'_{out}|}{2} \tag{2.26}$$

Therefore,

$$\theta_{Bragg} = \theta'_{in} \tag{2.27}$$

and the hologram works as predicted.

2.3.5 Writing and reading holograms with different wavelengths

When designing holographic interconnects, operating wavelengths for both reading and writing must be chosen. For reading, semiconductor diode lasers that can be directly integrated on chip are available in a variety of wavelengths from around 800 nm up to about 1500 nm. These devices have been widely exploited in fiber optic and telecom systems. Holographic emulsion materials on the other hand are generally not particularly sensitive to the near infrared radiation produced by the semiconductor lasers. The holographic materials typically have the best sensitivity for the blue-green to red wavelengths. As a result, it is many times necessary to write holograms for holographic interconnection systems at one wavelength and read the holograms with a different wavelength. This difference in read and write wavelengths must be account for when determining the correct writing geometry for a particular hologram.

First off, the hologram will be read out with light of wavelength λ_r and written with light of wavelength λ_w. Assume that the desired read geometry is the same as in Figure 2-18. The only difference is that the hologram will be read out with light of wavelength λ_r. A read geometry is specified by θ_{in}, θ_{out}, and λ_r. These external angles (θ_{in} and θ_{out}) must then be converted to internal angles with respect to the emulsion as previously discussed. This transforms the external read angles to

$$\theta'_{in} = \sin^{-1}\left(\frac{\sin\theta_{in}}{n_e}\right) \tag{2.28}$$

and

$$\theta'_{out} = \sin^{-1}\left(\frac{\sin\theta_{out}}{n_e}\right) \tag{2.29}$$

In order to achieve this read geometry specified by θ_{in}, θ_{out}, and λ_w; a particular grating spacing (Λ) and grating inclination (γ) is needed. These parameters can be calculated directly from the internal read geometry (θ'_{in} and θ'_{out}). Then, with Λ, γ, and the write wavelength (λ_w) the internal write geometry can be calculated. This write geometry inside the emulsion is then transformed to a geometry external to the hologram.

First, the grating spacing (Λ) is calculated just like the previous example. The output beam is moved to the input side of the hologram as in Figure 2-19. The angle between the two beams inside the emulsion (θ'_r) determines the required the grating spacing (Λ) in order for θ_{in} to satisfy the Bragg condition. The angle between the beams (θ'_r) is given by

$$\theta_r = |\theta'_{in} - \theta'_{out}| \tag{2.30}$$

This means that the grating spacing desired is

$$\Lambda_r = \frac{\lambda_r / n_e}{2 \sin\left(\frac{|\theta'_{in} - \theta'_{out}|}{2}\right)} \tag{2.31}$$

In order to produce a hologram with this external read geometry (θ_{in},θ_{out}, and λ_r) it is necessary to write the hologram such that it has a grating spacing of Λ given by Equation 2.31 and a grating inclination of γ given by Equation 2.21. The equation for the grating spacing of the hologram being written

$$\Lambda_w = \frac{\lambda_w / n_e}{2 \sin\left(\frac{\theta'_w}{2}\right)} \tag{2.32}$$

where θ'_w is the angle between the write beams inside the emulsion. If the desired grating spacing (Λ_r) is to be the same as the written grating spacing (Λ_w), Equations 2.31 and 2.32 can be set equal.

$$\frac{\lambda_r}{2 \sin\left(\frac{|\theta'_{in} - \theta'_{out}|}{2}\right)} = \frac{\lambda_w}{2 \sin\left(\frac{\theta'_w}{2}\right)} \tag{2.33}$$

Solving for θ'_w,

$$\theta'_w = 2\sin^{-1}\left\{\frac{\lambda_w}{\lambda_r}\sin\left(\frac{|\theta'_{in} - \theta'_{out}|}{2}\right)\right\} \tag{2.34}$$

Knowing the angle between the write beams inside the emulsion and the grating inclination, the angles of the two write beam inside the emulsion can be calculated.

$$\theta'_{w1} = \gamma - \frac{\theta'_w}{2} = \frac{\theta'_{in} + \theta'_{out}}{2} - \sin^{-1}\left\{\frac{\lambda_w}{\lambda_r}\sin\left(\frac{\theta'_{in} - \theta'_{out}}{2}\right)\right\} \tag{2.35}$$

and

$$\theta'_{w2} = \gamma + \frac{\theta'_w}{2} = \frac{\theta'_{in} + \theta'_{out}}{2} + \sin^{-1}\left\{\frac{\lambda_w}{\lambda_r}\sin\left(\frac{\theta'_{in} - \theta'_{out}}{2}\right)\right\} \tag{2.36}$$

These angles inside the emulsion must then be converted to external angles in free-space. This gives the correct write geometry for the hologram.

$$\theta_{w1} = \sin^{-1}\left[n_e\sin\left(\frac{\theta'_{in} + \theta'_{out}}{2} - \sin^{-1}\left\{\frac{\lambda_w}{\lambda_r}\sin\left(\frac{\theta'_{in} - \theta'_{out}}{2}\right)\right\}\right)\right] \tag{2.37}$$

and

$$\theta_{w2} = \sin^{-1}\left[n_e\sin\left(\frac{\theta'_{in} + \theta'_{out}}{2} + \sin^{-1}\left\{\frac{\lambda_w}{\lambda_r}\sin\left(\frac{\theta'_{in} - \theta'_{out}}{2}\right)\right\}\right)\right] \tag{2.38}$$

where θ_{w1} and θ_{w2} are the angles with respect to the normal of the two beams at wavelength λ_w used to write the hologram. Notice that Equation 2.37 and 2.38 are dependent on n_e. This is especially true when it is considered that θ'_{in} and θ'_{out} are also dependent on n_e.

Generally, the write wavelengths are in the visible spectrum while the read wavelengths are in the near infrared. This generally means that $\lambda_r > \lambda_w$. Likewise, due to the form of the above equation, $\theta_r > \theta_w$.

As discussed previously in this section, when writing holograms, the partial mirror

planes are formed at the angle that is the bisector of the internal write angles. It was also discussed that the read geometry required the angle of the partial mirrors to be γ from Equation 2.21. If both of these statements are true, this means that the bisector of the internal write beams must be at an angle γ with respect to the normal to the hologram. Figure 2-21 illustrates the derivation of the λ_w write geometry. The read geometry with λ_r is shown as a dashed line. The solid line shows the write geometry after it has been transformed to the write wavelength λ_w. The λ_w write beams are also angled such that the bisector of the internal beams is at angle γ with respect to the normal. This write geometry (λ_w, θ_{w1}, θ_{w2}, and γ) will produce a hologram with the desired read geometry (θ_{in}, θ_{out}, λ_r).

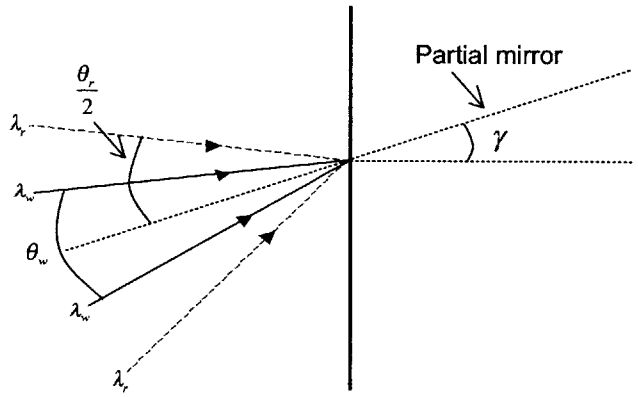

Figure 2-21: Illustration of the derivation of method to write a hologram at wavelength λ_w to be read at wavelength λ_r

2.3.6 How to write a hologram with a divergent beam input and a convergent wave output

The ultimate goal is to determine how to write a hologram with the specifications set forth by the CONNPP. To do this, the hologram must take a divergent beam as an input and output a convergent beam at a given angle. This is illustrated in Figure 2-22. In Figure 2-22, D_{r1} is the distance between the source and the hologram. D_{r2} is the distance in the z direction between the hologram and the plane of the detector. X_{r1} is the distance in the x direction between the divergent source and the

the photodetector. Finally, θ_{rdiv} is the half-angle divergence of the source.

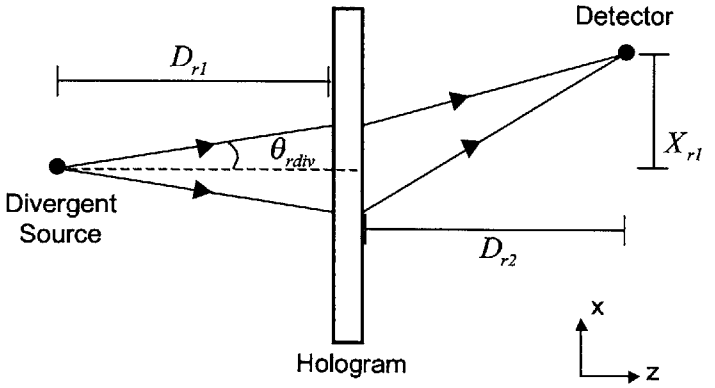

Figure 2-22: Hologram with divergent source as input and convergent beam as output

It can be noted that the divergent wave from the point source is not a plane wave. Its wavefront has a curved surface that gets larger the further it propagates from the source. Fortunately, Fourier optics allows an arbitrary wave to be written as the sum of plane waves with different spatial frequencies [20]. For two optical waves of the same wavelength, spatial frequency is simply a measure of angle. The simple divergent wave can be written as the weighted sum of plane waves with angles from $+\theta_{rdiv}$ to $-\theta_{rdiv}$. This means that the plane waves at $\pm\theta_{rdiv}$ represent the extremes of the read out case shown in Figure 2-22. To simplify the problem, these two beams (at $\pm\theta_{rdiv}$) will be used to develop a write geometry for the hologram.

Once the problem has been reduced to looking at the extreme cases, it can be further broken down by examining each of the two extremes. In Figure 2-23, the $+\theta_{rdiv}$ and $-\theta_{rdiv}$ cases have been separated. In Figure 2-23(a), the $+\theta_{rdiv}$ is seen. From this diagram, it can be seen that

$$\tan(\theta_{out1}) = \frac{X_{r1} - D_{r1}\tan(\theta_{rdiv})}{D_{r2}} \tag{2.39}$$

and hence,

$$\theta_{out1} = \tan^{-1}\left(\frac{X_{r1} - D_{r1}\tan(\theta_{rdiv})}{D_{r2}}\right) \tag{2.40}$$

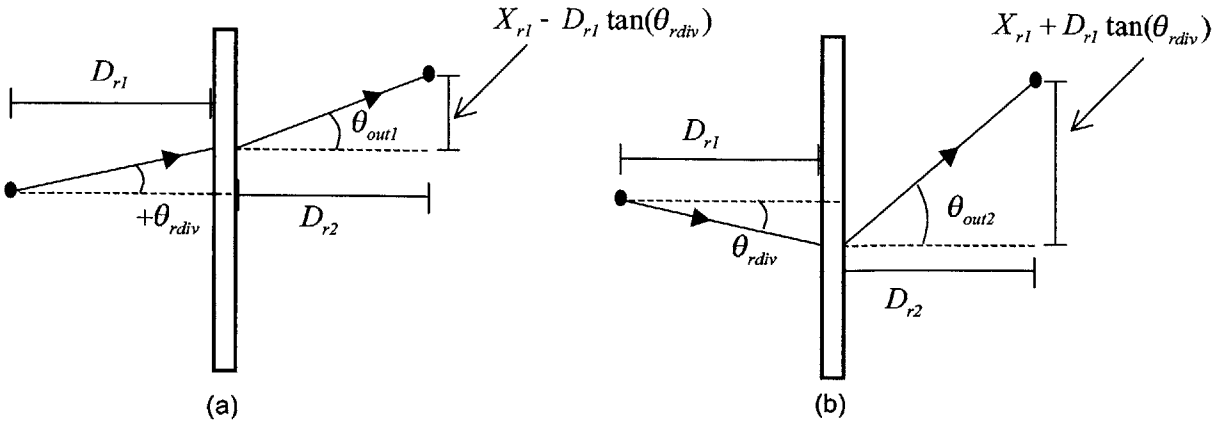

Figure 2-23: Read geometries for the extreme plane waves in the divergent source problem

Analogously, looking at Figure 2-23(b),

$$\theta_{out2} = \tan^{-1}\left(\frac{X_{r1} + D_{r1}\tan(\theta_{rdiv})}{D_{r2}}\right) \tag{2.41}$$

Now, in order to figure out how these holograms should be written, the approach developed in the previous section will be used. First, the angles must be converted to angles internal to the emulsion denoted in the following equations with a prime. Next, the grating angles (γ) and grating spacings (Λ) must be found. From Equation 2.21,

$$\gamma_1 = \frac{\theta'_{rdiv} + \theta'_{out1}}{2} \tag{2.42}$$

and

$$\gamma_2 = \frac{-\theta'_{rdiv} + \theta'_{out2}}{2} \tag{2.43}$$

Now, for Λ,

$$\Lambda_1 = \frac{\lambda_r/n_e}{2\sin\left(\frac{|\theta'_{rdiv} - \theta'_{out1}|}{2}\right)} \tag{2.44}$$

and

$$\Lambda_2 = \frac{\lambda_r / n_e}{2 \sin\left(\frac{\left|-\theta'_{rdiv} - \theta'_{out2}\right|}{2}\right)} \tag{2.45}$$

Next, for the last step of the process, the angle between the write beams (θ_w) is determined using Equations 2.37 and 2.38. The $+\theta_{rdiv}$ will be denoted as case 'a' and the $-\theta_{rdiv}$ as case 'b'. For case 'a',

$$\theta_{wa1} = \sin^{-1}\left[n_e \sin\left(\frac{\theta'_{rdiv} + \theta'_{out1}}{2} - \sin^{-1}\left\{\frac{\lambda_w}{\lambda_r}\sin\left(\frac{\theta'_{rdiv} - \theta'_{out1}}{2}\right)\right\}\right)\right] \tag{2.46}$$

and

$$\theta_{wa2} = \sin^{-1}\left[n_e \sin\left(\frac{\theta'_{rdiv} + \theta'_{out1}}{2} + \sin^{-1}\left\{\frac{\lambda_w}{\lambda_r}\sin\left(\frac{\theta'_{rdiv} - \theta'_{out1}}{2}\right)\right\}\right)\right] \tag{2.47}$$

For the 'b' case,

$$\theta_{wb1} = \sin^{-1}\left[n_e \sin\left(\frac{-\theta'_{rdiv} + \theta'_{out2}}{2} - \sin^{-1}\left\{\frac{\lambda_w}{\lambda_r}\sin\left(\frac{-\theta'_{rdiv} - \theta'_{out2}}{2}\right)\right\}\right)\right] \tag{2.48}$$

and

$$\theta_{wb2} = \sin^{-1}\left[n_e \sin\left(\frac{-\theta'_{rdiv} + \theta'_{out2}}{2} + \sin^{-1}\left\{\frac{\lambda_w}{\lambda_r}\sin\left(\frac{-\theta'_{rdiv} - \theta'_{out2}}{2}\right)\right\}\right)\right] \tag{2.49}$$

This means that the write geometries have the write beams separated by $|\theta_{wa1} - \theta_{wa2}|$ and $|\theta_{wb1} - \theta_{wb2}|$ and the grating inclinations at γ_1 and γ_2 respectively. While the equations are needed, they are not particularly illustrative as to what is happening. For this, an example is necessary.

Assume that the hologram is being read out by an 850nm VCSEL (λ_r) with a half-angle divergence (θ_{rdiv}) of 5°. The VCSEL is placed 1mm behind the hologram (D_{r1}).

The photodetector is sitting 1mm in front of the hologram (D_{r2}) and is displaced by 0.5mm in the x direction (X_{r1}). The hologram is to be written with a 514nm argon ion laser (λ_w).

Using Equations 2.40 and 2.41 the output angles are

$$\theta_{out1} = 22.4^\circ \quad \text{and} \quad \theta_{out2} = 30.43^\circ \tag{2.50}$$

From Equations 2.42 and 2.43, the grating inclinations, γ_1 and γ_2, are

$$\gamma_1 = 9.675^\circ \quad \text{and} \quad \gamma_2 = 8.25^\circ \tag{2.51}$$

Finally, the angles for the 514nm write beams for case 'a' are

$$\theta_{wa1} = 8.36^\circ \quad \text{and} \quad \theta_{wa2} = 18.89^\circ \tag{2.52}$$

for case 'b'

$$\theta_{wb1} = 1.11^\circ \quad \text{and} \quad \theta_{wb2} = 22.27^\circ \tag{2.53}$$

Figure 2-24(a) shows the read geometry using an 850nm divergent source and Figure 2-24(b) shows the write geometry using a 514nm source. Both figures are drawn to scale.

The above diagram helps to give a better understanding of what is happening. Although the problem was broken down into two separate problems to solve it, the solutions can now be combined to yield even more information. When both solutions are put together as in Figure 2-24(b), the first thing to notice is that one of the write beams is a divergent beam from a point source. This point source has been rotated such that the bisector of the divergent beam makes a nonzero angle with respect to the hologram. The other write beam is a slightly convergent beam that would focus on the other side of the emulsion. The point where the beam would focus can be seen on the right hand side of the emulsion.

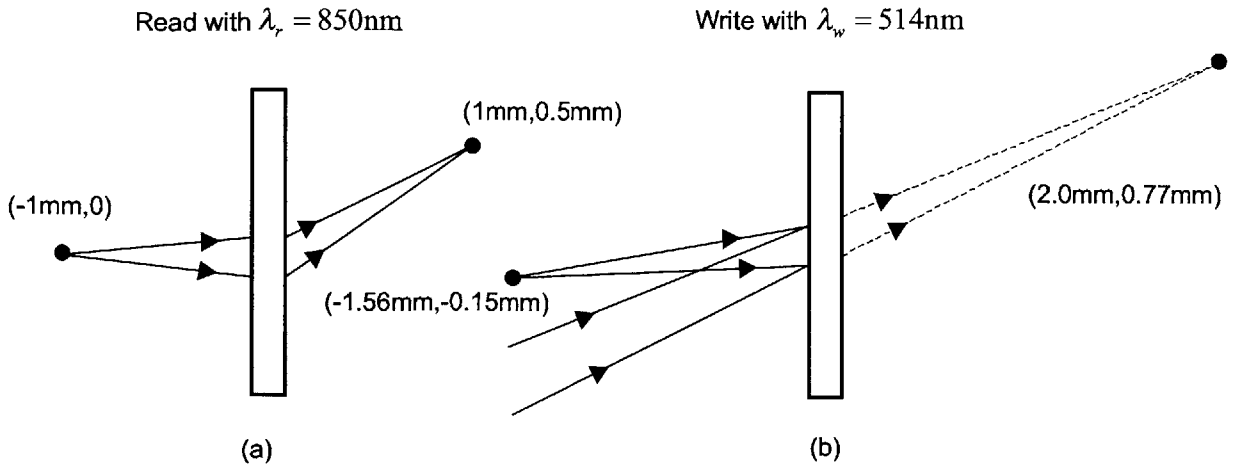

Figure 2-24: (a)Read geometry for example (b) Calculated write geometry for example

2.3.7 How to write a hologram with a divergent beam input and a plane wave output

Might the write geometry calculated above be simplified if the desired output of the hologram was a plane wave instead of a convergent wave? To answer this question, the equations from the previous section can be directly applied to yield a write geometry for a plane wave output.

In order to make the above equations work for a plane wave output, all that has to be done is to set

$$\theta_{out1} = \theta_{out2} \tag{2.54}$$

When the output angles of the hologram are equal, there is a plane wave. To illustrate what happens, assume the same conditions as the previous example (λ_r =850nm, λ_w =514nm,$\theta_{rdiv} = 5°$, and D_{r1} =1mm). In addition, assume that the desired output angle of the plane wave is 15°. Using the equations in the previous section, the angles are calculated and the results are plotted to scale in Figure 2-25.

Unlike the previous example, the write geometry is fundamentally different at the different wavelengths. If the hologram were to be written using λ_r, a divergent

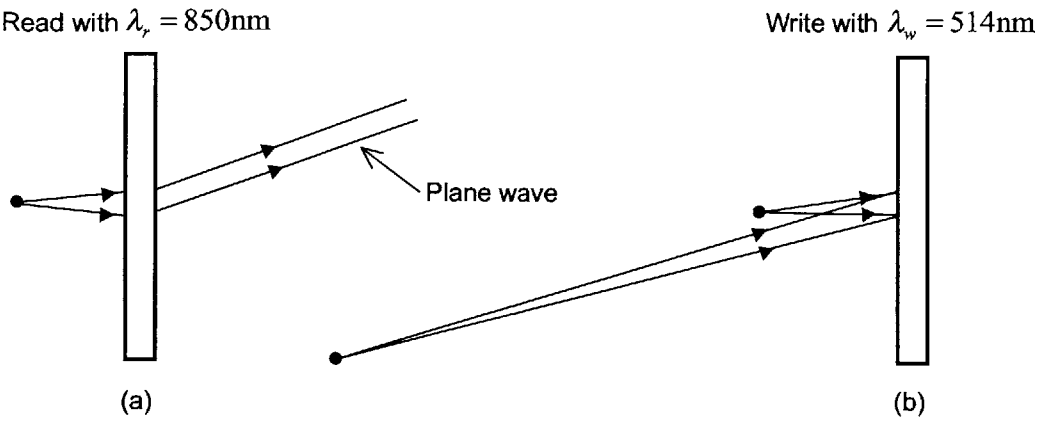

Figure 2-25: (a)Read geometry for plane wave output (b) Calculated write geometry for plane wave output for example

beam and a plane wave would be used. As seen in Figure 2-25, when λ_w is used to write the hologram, two divergent sources are used. This example goes to show that when different read and write wavelengths are used, normal intuition behind hologram geometry does not necessarily hold.

2.3.8 Volume aspects of holograms

As discussed previously, the holograms used for the holographic interconnects here are thick holograms. This means that their thickness is not negligible. In some of the examples, a source distance of 1mm was specified. Some of the photopolymer emulsions tested are as thick as 200μm or one fifth of the source distance. As a result, it is useful to understand what is happening with the fringes inside the hologram.

If two plane waves are used to create the thick hologram, the structure inside the hologram consists of flat planes of high index material as in Figure 2-16(b). When the reference beam and object beam are not plane waves, the structure of the hologram becomes more complicated.

In the case of the holographic interconnects modelled in this section, a divergent and convergent beam are used to write the hologram. For the sake of simplicity and understanding the same wavelength will be used for reading and writing in this

example. In Figure 2-26, a portion of both the read and write geometries is shown. In the diagram, a cross-sectional plane (x_0) of the hologram was chosen. Points were evenly spaced along the cross-section. Rays from the source were then drawn to the points. If the hologram is to work correctly, the rays at each point from the source must be directed to the detector. Finally, by propagating the rays from the detector back (dashed lines) the second write beam (in addition to the source) is shown. In other words, writing a hologram with write beam #1 and write beam #2 will result in a hologram that directs the beams from the source to the detector as desired. Please note that the propagation angles of the beams will change as they enter and exit the emulsion due to the change in index of refraction. This is also illustrated in Figure 2-26.

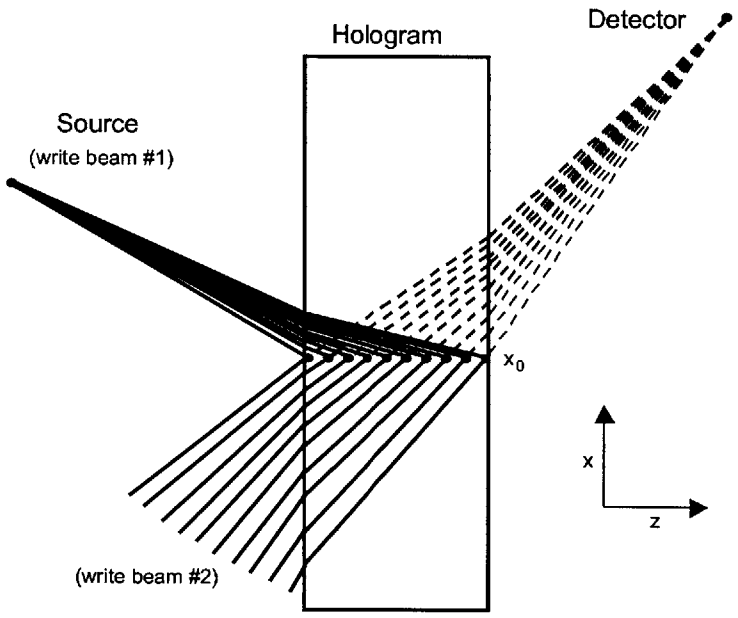

Figure 2-26: Read and write geometries for a cross-sectional plane of the hologram

As can be seen in Figure 2-26, as z increases (with x remaining the same) the angle from the source to cross-sectional plane changes. Likewise, the angle of write beam #2 also changes as the cross-sectional plane is followed through the hologram. Because the source and detector can be at any position, in general it can be said that as z is increased, both the angle between the beams (θ_w) and the inclination

of the grating (γ) will change. Figure 2-27 shows how the grating inclination of the partially silvered mirrors (bisector of the write beams, γ) changes with z for the hologram discussed in Figure 2-26.

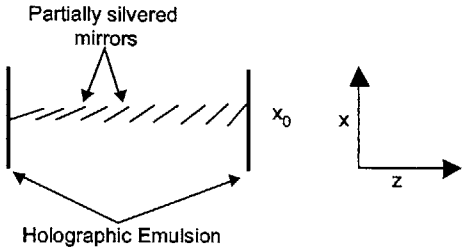

Figure 2-27: Angle of the partially silvered mirror planes as a function of z for the hologram discussed in Figure 2-26

Another interesting aspect of the volumetric nature of the thick holograms is how a single beam propagating through the emulsion is treated. In this case, the angle of the input beam with respect to the hologram, does not change as it advances through the hologram. The output angle needed to hit the detector, however, does change. Once again, the change in index of refraction as the beams propagate into different media is taken into account. This can be seen in Figure 2-28.

In the figure, the single beam can be seen propagating from the source through the hologram. At each point notated on the source beam, the partial mirror structures must be angled such that the reflected beam hits the detector. The reflected beams are projected back (dotted lines) to show the write beam that would produce this result. As in the previous example, this leads to a situation where both the angle between the write beams (θ_w) and the grating inclination angle (γ) changes as the beam propagates through the hologram. Also like the previous example, the partially silvered mirror planes rotate through the hologram. This can be seen in Figure 2-29.

The two previous examples should help illustrate the internal structure of thick film holograms. Earlier in this section, models were developed that showed how to design write geometries from a desired read geometry. If the writing of the hologram has been done according to these write geometries, every point within the hologram will partially reflect an incident beam from the source to the detector. As a result,

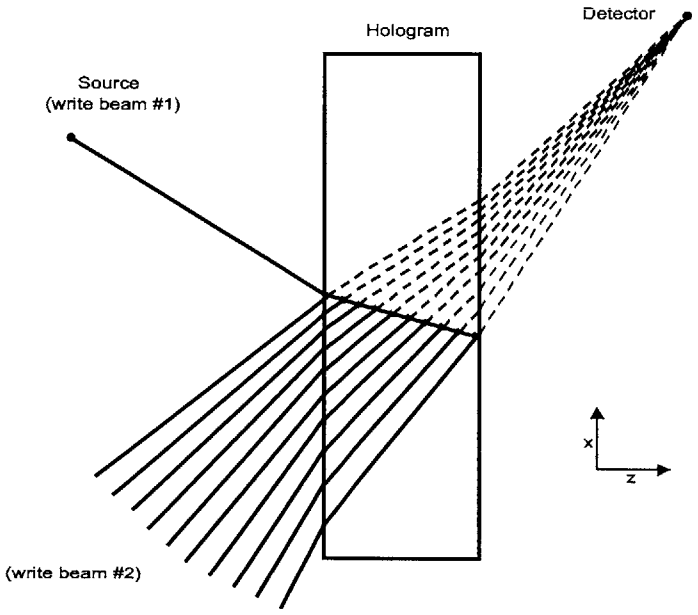

Figure 2-28: Read and write geometries for a single beam propagating through a thick hologram

most of the light from the source will be focused on the detector creating a high-quality holographic interconnect.

These expected geometries are in turn confirmed by Reference [23]. In this source, the case of a divergent and convergent cylindrical beam incident from the same side, interfering in an emulsion is examined. It is shown that this case will produce a grating within the emulsion that are part of ellipses with foci at the two focal points of the beams. This is illustrated in Figure 2-30.

Figure 2-30 shows the situation when both of the foci are aligned with the x-axis. In the case of the holograms for the CONNPP, the divergent source will be on axis, but the convergent source will be off axis. The resulting grating can be seen in Figure 2-31. Notice that this agrees reasonably well with Figure 2-27 (i.e. angle with respect to the normal of the holographic grating increases as x increases).

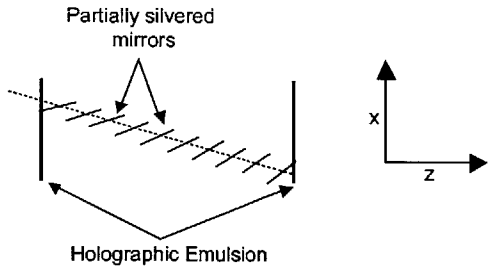

Figure 2-29: Angle of the partially silvered mirror planes for a single beam propagating through the hologram

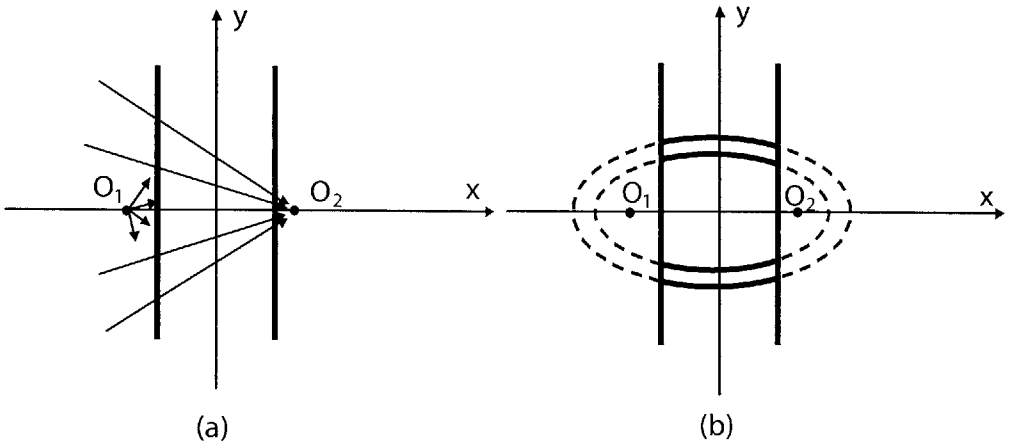

Figure 2-30: (a) A cylindrical divergent beam and cylindrical convergent beam interfering within a holographic emulsion (b)The holographic grating after processing. The grating forms partial ellipses with foci O_1 and O_2.

2.4 Hologram Writing Systems

Several hologram writing systems were designed, built, and tested in order to determine the best way for writing the holographic elements discussed previously. The goal of the system fabrication was to determine the best way to create a production system capable of making holographic element arrays easily. The first system examined was a very simple setup using a microscope objective as the beam shaping element. The next system looked at used a single large lens for both the reference and object beam. The last system investigated was a computer-controlled system using a beam splitter to bring both beams together in close proximity.

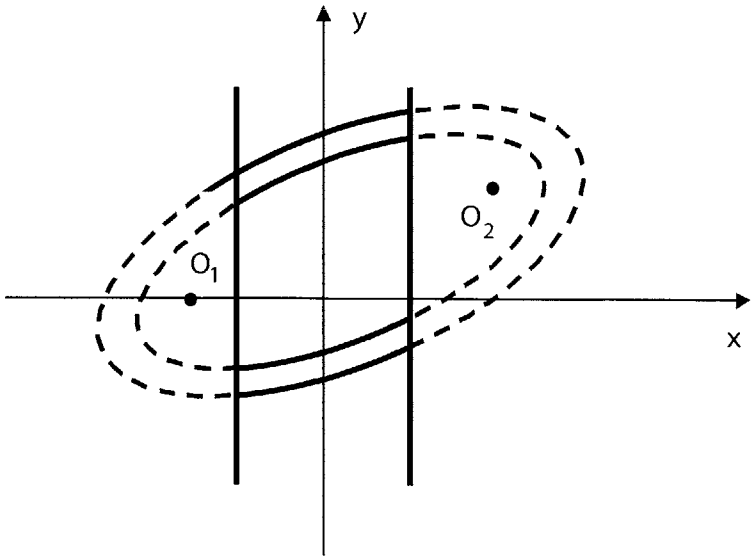

Figure 2-31: Holographic grating formed by a divergent on axis source and a convergent off axis source.

2.4.1 System using a microscope objective

The system using a microscope objective was the first proof-of-concept system built to show that holograms could be made that take a divergent wave as an input, and produce a convergent wave as the output. The system (illustrated in Figure 2-33) begins with a very basic holography setup. The laser source is passed through a microscope objective/spatial filter combination to expand the beam. A convex lens is then used to collimate the beam. Next, the beam is split (approximately 50-50) into the two different interference legs. One of the interference legs goes directly into a microscope objective. The holographic plate is placed slightly past the focus of the beam to a point where it has expanded to the desired hologram size. The other leg is brought in through a shallow convex lens in order to make it slightly convergent. Notice that one of the write beams is convergent and one is divergent. This is exactly the geometry outlined in the last section to produce a convergent output beam upon read. The holographic plate is on a rotation stage to allow the plate to be rotated

with respect to the write beams. This is done because it is much easier to rotate the holographic plate than all of the optical elements associated with each of the write beams. A close-up of the beams at the surface of the hologram can be seen in the right side of Figure 2-32. The divergent beam can be seen as the solid line and the convergent beam can be seen as the dashed line.

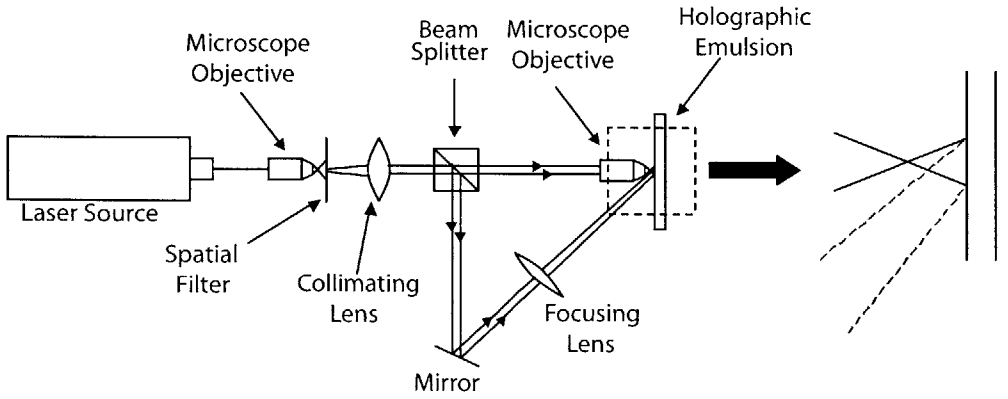

Figure 2-32: Setup for the microscope objective system

The system was built and operated with a 632.8nm Helium-Neon laser as the laser source. As this was a proof of concept setup, the more expensive blue-green sensitive Aprilis photopolymer holographic plates were not used. Instead, cheaper red sensitive PFG-01 silver hallide emulsions from Slavich were used. These silver hallide emulsions are significantly thinner (7-8μm) than the photopolymer emulsions (200μm). This leads to worse diffraction efficiency as a result of less of a Bragg effect within the emulsion. Also, the PFG-01 plates require chemical processing. This can lead to situations where it is difficult to diagnose whether problems are related to the geometry/optics or with the chemical processing of the emulsion. This is a distinct advantage of using the UV cured photopolymer emulsions.

The goal of the system was to write a hologram using 632.8nm light that could be read out at normal incidence using a 850nm divergent source (VCSEL) and produce a convergent beam output. This goal was effectively accomplished. In order to show this, the hologram that was written was read out with an 850nm VCSEL normal to the plane of the hologram. The output of the hologram was then projected onto a

diffuser. The image on the diffuser was then captured with an infrared sensitive CCD camera. This can be seen in Figure 2-33.

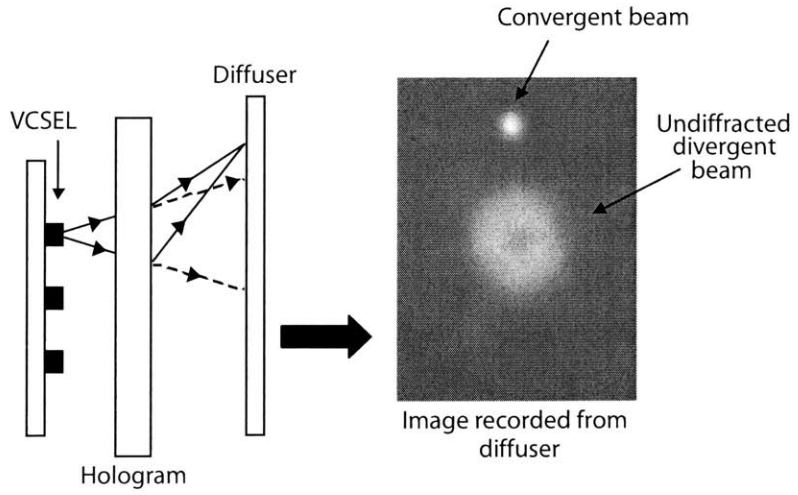

Figure 2-33: Readout of the hologram with an 850nm VCSEL source

Notice in the upper portion of the image recorded from the diffuser that there is small spot with greater intensity than the undiffracted beam. This convergent beam is the desired output of the hologram. The large, fuzzy circular structure in the middle of the image recorded from the diffuser is the undiffracted divergent beam. As was mentioned before, the PFG-01 plates have a relatively low maximum diffraction efficiency ($\sim 50\%$). This explains why there is a significant amount of energy in the undiffracted beam.

Thick holograms put all the light into one of the 1^{st} order beams and the 0^{th} order beam (undiffracted beam). However, because the PFG-01 emulsion is fairly thin, the light is actually diffracted into multiple orders. However, looking at the figure, only the 0^{th} and $+1^{st}$ orders can be seen. This is because the higher orders ($\pm 2^{nd}$, $\pm 3^{rd}$, etc.) don't have enough energy in them to be visible. The -1^{st} order cannot be seen because it is even more divergent than the 0^{th} order and its intensity is below the sensitivity of the CCD camera.

Although this system did accomplish its proof of concept goal, it is fraught with several problems that prevent it from being a production system that can create arrays

of holographic elements easily. The first such problem can be seen by examining Figure 2-32. As can be seen, the angle in the figure between the two separate legs of the write geometry is quite large (almost 45°). In most cases, it would be desirable to have the angle between the two write beams be significantly smaller ($\sim$ 5°-20°). However, as the angle between the write beams is decreased, the convergent beam begins to hit the barrel of the microscope objective. Because one of the goals is to create very small holograms, the microscope objective must be very close to the emulsion. This leads to the problem of the convergent beam hitting the barrel of the microscope objective at small angles. This is illustrated in Figure 2-34.

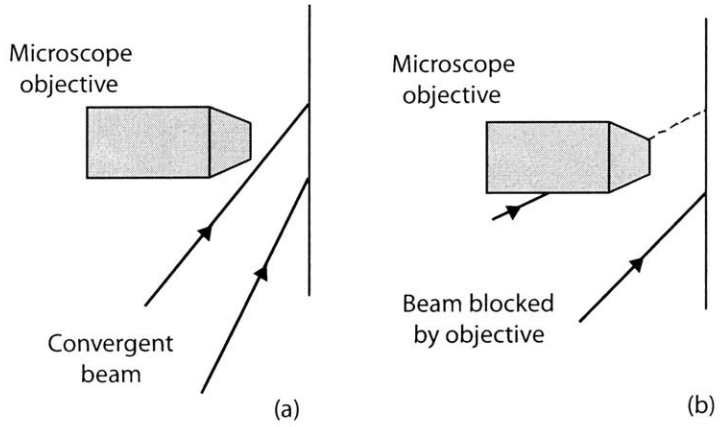

Figure 2-34: (a) Beam at large angle hits emulsion. (b) Beam at smaller angle partially blocked by objective

Another problem with the setup relates to the quality of the microscope objective. In all, about seven different microscope objectives were tested. All showed significant imperfections. These imperfections manifested themselves as high spatial frequency components in the output beam. Usually this is counteracted by a spatial filter to remove these high spatial frequency components. Unfortunately because of the small distances involved in the setup, it was not feasible to integrate a spatial filter into the setup. This resulted in the divergent beam having many high spatial frequency artifacts which can have a significant negative impact on the quality of the holograms produced.

2.4.2 Single lens system

Taking into account some of the issues encountered with the microscope objective setup, a new system (illustrated in Figure 2-35) was designed to overcome these issues. This system is the same as the microscope system up to the beam splitter element. Just as in the microscope system, the beam from the laser source has been expanded and collimated. After the beam splitter, one of the beams goes directly into a large lens and is subsequently focused onto the holographic emulsion. This is the divergent beam of the desired write geometry. The other beam is reflected off of a mirror so that it at the correct angle with respect to the diverging beam. This collimated beam is then put through a concave lens which causes the previously collimated beam to diverge slightly. After that, the now slightly divergent beam is put through the same convex lens as the first beam, but near the edge of the lens. If the focal length of the diverging lens is larger than that of the converging lens, the beam will converge to a point behind the plane of the holographic emulsion.

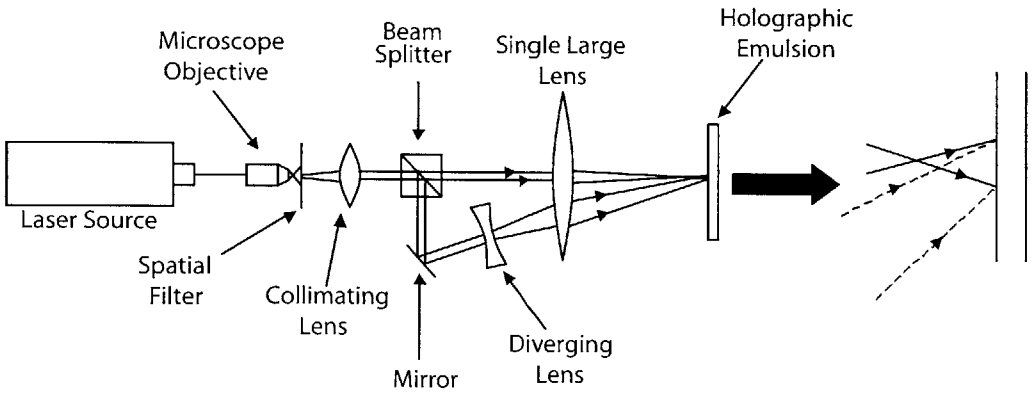

Figure 2-35: Setup for the single lens hologram writing system

In this system, the divergence angle of the beam normal to the large lens can be adjusted by changing the size of the beam entering the lens. If a collimated beam enters the the large lens, it will theoretically focus at a point regardless of its cross-sectional area. Therefore, by increasing the cross-sectional area of the incident beam, the divergence angle of the beam after it crosses the focus will also increase. This can be seen in Figure 2-36.

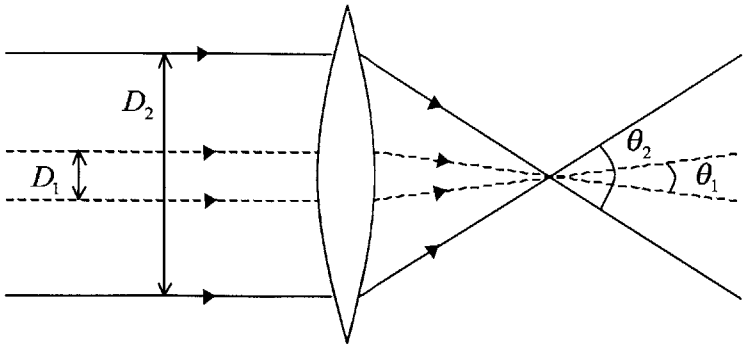

Figure 2-36: How divergence angle is affected by beam size

Another benefit of this system is that it allows for very small holograms to be produced. In the microscope objective setup, in order to make the hologram smaller, the distance between the objective and the emulsion had to be decreased. This in turn required an even larger angle for the second beam which was undesirable. In the single lens setup however, the holographic emulsion can be positioned arbitrarily close to the beam focus in order to make very small holograms. As discussed previously, small holograms are a necessity for the CONNPP.

The problem with this system is that the quality of the holograms that can be written is severely limited by the fact that the second beam does not pass through the center of the lens along with quality of the large lens used in the setup. Because the second beam is put through the large lens at a position that deviates significantly from the center of the lens, the phase of the second beam is not transformed symmetrically about its center. This can explain the non-circular shape of the beam at the plane of the holographic emulsion. A possible way to correct this would be by putting the second beam through the edge of a compensating lens before the large lens. By transforming the beam's phase profile non-symmetrically twice, in opposite directions, this effect could potentially be compensated for. However, the logistics involved in doing this potentially outweigh the benefits.

Because the phase profile of the second beam was transformed non-symmetrically, the holograms that were produced performed very poorly. When attempting to read the holograms with a divergent VCSEL as in Figure 2-33, very low diffraction effi-

ciency was observed (i.e. the undiffracted beam was much brighter than the diffracted beam). A comparison of the hologram output of holograms produced with the microscope objective system and the single lens system can be seen in Figure 2-37.

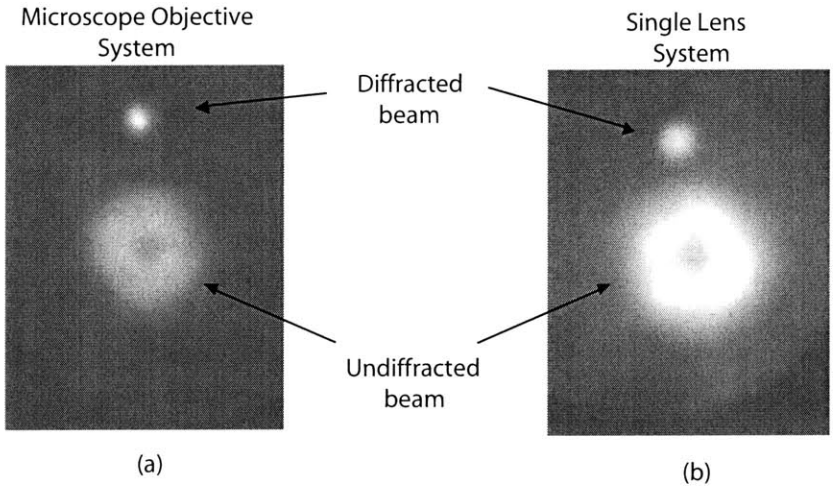

Figure 2-37: (a)Output of microscope objective system hologram (b)Output of single lens system hologram

It is important to notice the difference in the outputs of the holograms produced by the two different systems. The undiffracted beam in the hologram produced by the single lens system is much brighter than the undiffracted beam in the hologram produced by the microscope objective system. As discussed earlier, the goal of the holographic interconnects for the CONNPP is to put as much of the light as possible from the VCSEL into the diffracted beam. While the microscope objective system does have severe limitations in terms of usable geometries, the holograms produced by the system clearly outperform those made with the single lens system.

2.4.3 Computer controlled system with beam splitter

The third and final system examined was originally designed and built by Milos Komarcevic [24]. The system further optimized by Marta Ruiz Llata in the Fall of 2002 [25]. While the first two systems examined were for proof-of-concept, this system was intended to be a production system. As such, it was designed to be

almost completely computer controlled. First, an overview of the optical aspects of the system will be given. Next, the mechanical layout of the system and computer controls will be described. Finally, results will be given showing holograms produced by the system.

Figure 2-38 shows the basic optical setup for the computer controlled beam splitter system. In this system, the laser source first passes through the microscope objective and spatial filter to expand the beam. Immediately after being expanded, the beam is split off into the two separate legs. For visual simplicity, the beam through one of the legs is shown with solid lines and the other is shown with a dashed line.

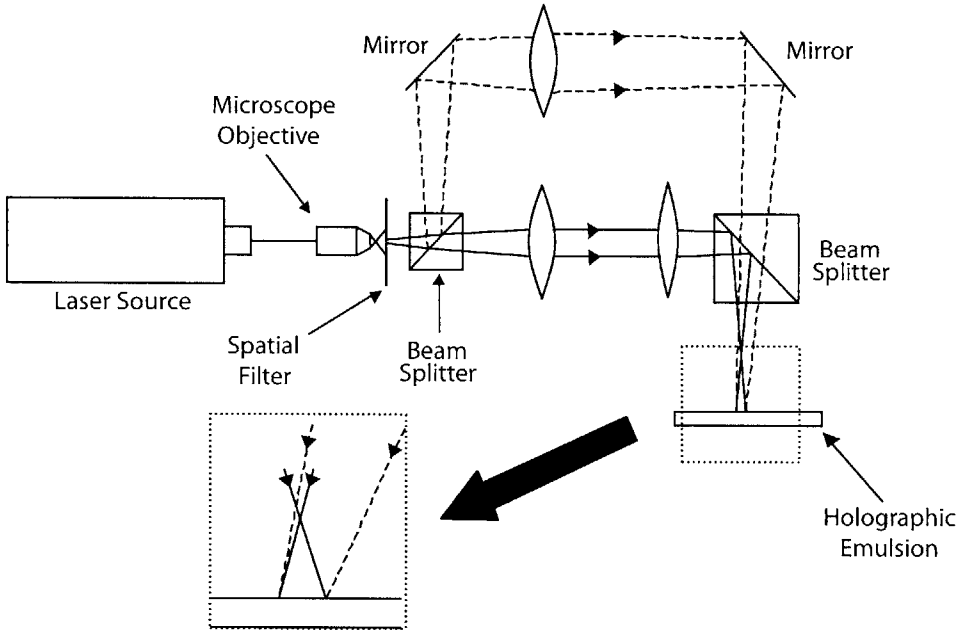

Figure 2-38: The optical setup for the computer controlled beam splitter system

The beam represented by the solid line passes through two lenses to give it the correct divergence characteristics. The beam represented by the dashed line is put through a single lens so that it is convergent at the holographic emulsion. After the lens, the dashed beam is reflected off of a mirror that directs it towards the emulsion at the correct angle for the particular write geometry. Both beams are then put through a beam splitter. A photograph of the actual setup can be seen in Figure 2-39. This allows arbitrary angles between the converging and diverging

beams without having to be concerned about whether the non-normal beam will hit any of the optical elements in the other beam path. As it may be recalled, the non-normal beam hitting the microscope objective barrel was the main problem with the microscope objective system. This was solved in the single lens system by putting both beams through one optical element, however this solution also compromised the quality of the holograms produced. By putting both beams through a single beam splitter, half of the power of each of the beams is wasted thereby increasing the exposure time. However, the benefit of arbitrary angular geometry without worrying about lens corrections or blocking optical elements potentially outweighs the downside of longer exposure time. This can also be compensated for by using a higher power laser.

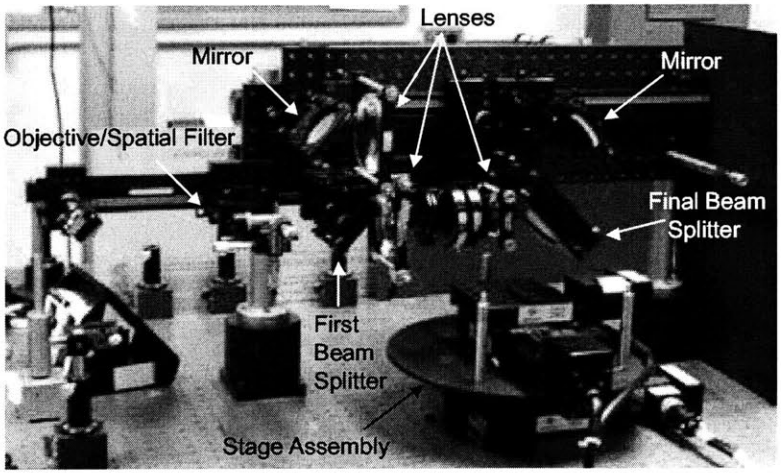

Figure 2-39: A photograph of the computer controlled beam splitter system

Adjustment of the lenses allows the system to be adjusted for arbitrary convergence and divergence angles of the write beams. Likewise, by moving and angling the second mirror in the dashed path, the angle between the beams can be changed. This provides the user with several knobs to setup write geometries determined by the equations from the previous section.

The mechanics of this system are more complicated than the previous two systems. In the other two systems, the beams were confined to a single plane of space that was parallel to the optical table. The original embodiment of the computer controlled

beam splitter system, however, employed a 40 kilogram mechanical rotary stage which required the beams to propagate perpendicular to the plane of the optical table top. This, in turn, required the optical elements to be mounted on an optical breadboard that is vertically suspended over the rotary stage (Figure 2-39). When the system was overhauled by Marta Ruiz Llata, the 40 kilogram rotary stage was eliminated in favor of a somewhat less bulky computer controlled rotary stage.

As was mentioned previously, the goal of this system was to produce a production system capable of writing holograms for the CONNPP. One of the requirements for a system of this type is the the ability to write large arrays of these individual holographic elements. In Figure 1-5, a single neuron view of the hologram plane was given that included nine separate holograms. This set of nine holograms is then repeated in an array structure over the holographic emulsion. An example of the resulting array containing 81 individually written holograms can be seen in Figure 2-40.

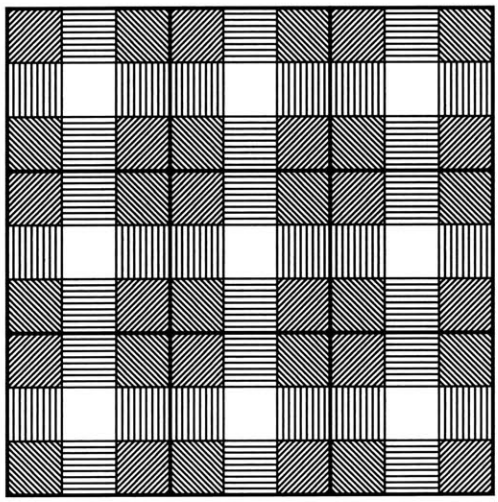

Figure 2-40: An example of a holographic element array

This holographic element array consists of only three basic hologram geometries. On the diagram, these three different holograms are represented by horizontal and vertical lines, angled lines, and circles. By using the rotary stage described above, the system can use these three basic geometries to produce the nine different holograms present in each neuron. In addition to the computer controlled rotary stage, there

are two computer controlled linear stages that allow the holographic emulsion to be translated in the x and y directions. The holographic element array shown in Figure 2-40 is made by stepping through the whole array element by element and exposing the holographic emulsion at each step. A close-up photograph of the stage setup can be seen in Figure 2-41. The system also has a computer controlled shutter that allows the user to set the exposure time for the hologram writing. This is extremely useful in doing characterization to determine the optimum exposure time for a particular holographic material. All three stages and the shutter are then connected to a PC and controlled via LabView.

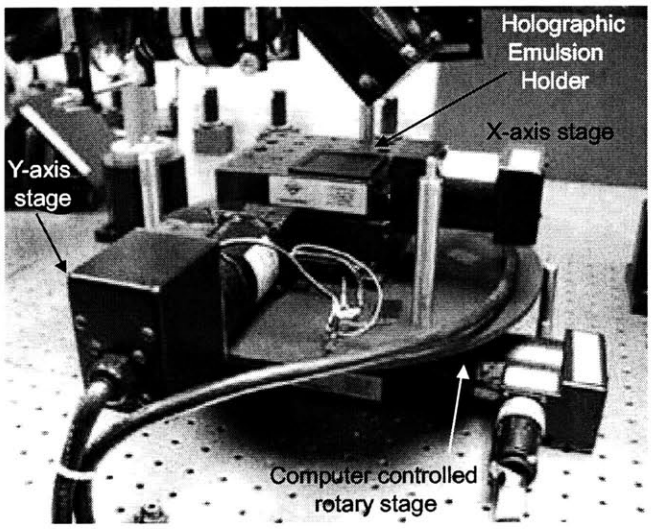

Figure 2-41: A close-up photograph of the computer controlled stage

The builders of the system started by testing it with the relatively inexpensive red-sensitive Slavich PFG-01 holographic emulsions. After testing and refinement of the system, the users began testing a blue-green sensitive photopolymer material made by Aprilis (formerly made by Polaroid). Whereas the Slavich PFG-01 material has a maximum diffraction efficiency of less than 50% with perfect writing and perfect chemical processing, the Aprilis photopolymer material has up to 90% diffraction efficiency when optimized properly. In addition to higher diffraction efficiency, there is also no chemical processing. To fix the hologram, the photopolymer is simply exposed to an ultraviolet light source. This allows the user to concentrate on getting

the geometry correct instead of worrying about the chemical processing.

2.4.4 Model verification using the beam splitter setup

The goal of the experiments performed was to show that a hologram could be written to produce a desired read geometry using the models developed in the last section. The test holograms were written using a 514nm argon ion laser as the write source. The holographic material used was Aprilis photopolymer (formerly Polaroid) with a thickness of 200μm. The desired read out wavelength was 633nm (HeNe). An arbitrary read geometry was chosen and can be seen in Figure 2-42.

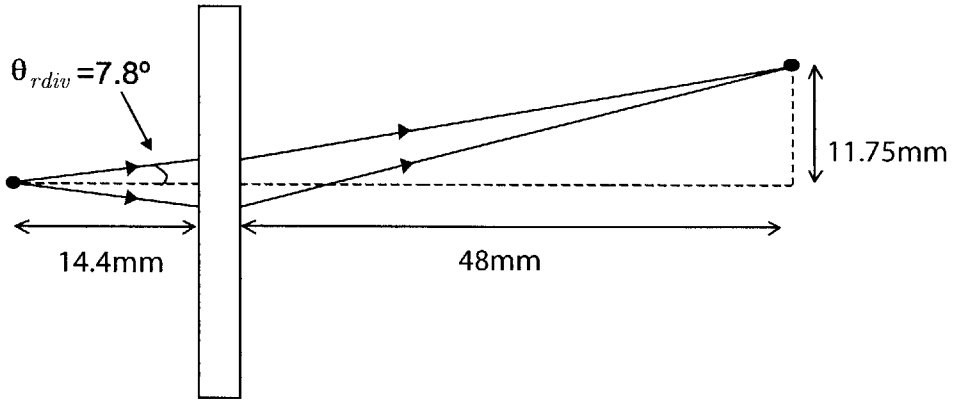

Figure 2-42: Desired read geometry

The write geometry was calculated from the models in the previous section and the Hologram Writer was set up with that write geometry. One of the shortcomings of the Hologram Writer is that because of the vertically suspended optics, it is very difficult to accurately measure distances and angles between the optical elements. Improvements will be made to the system and optical elements will be added specifically to help the user quantify the distances and angles for the write geometry. After the system was set up as close to the calculated write geometry as possible, the hologram was written. After writing, it was exposed to UV light and then characterized. Shown in Figure 2-43 is the output of the hologram using a 633nm HeNe laser.

The read geometry was then measured and compared to the desired read geometry. The results can be seen in Figure 2-44. While the actual read geometry did not

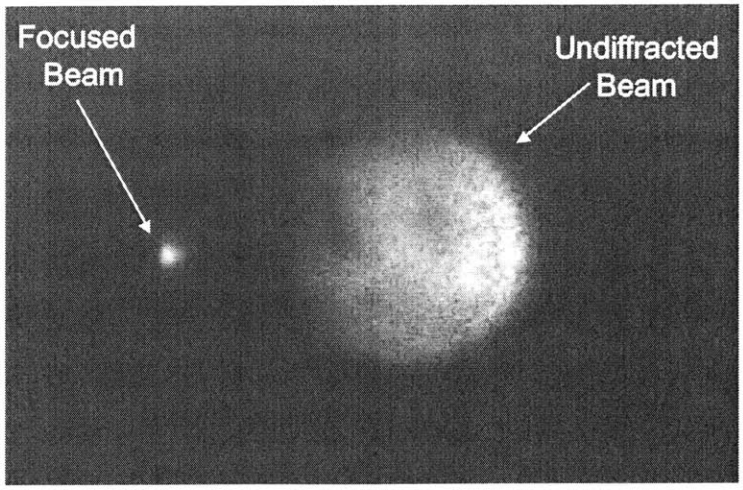

Figure 2-43: Output of hologram

match the desired read geometry exactly, all dimensions were within 10% of their target values. This 10% error can easily be accounted for due to the accuracy of the measurements for the write geometry. As future work, the Hologram Writer system will be upgraded for significantly increased measurement accuracy. The diffraction efficiency for the holograms produced was 35%. Optimization of the key parameters will continue to improve the diffraction efficiency of the holograms produced.

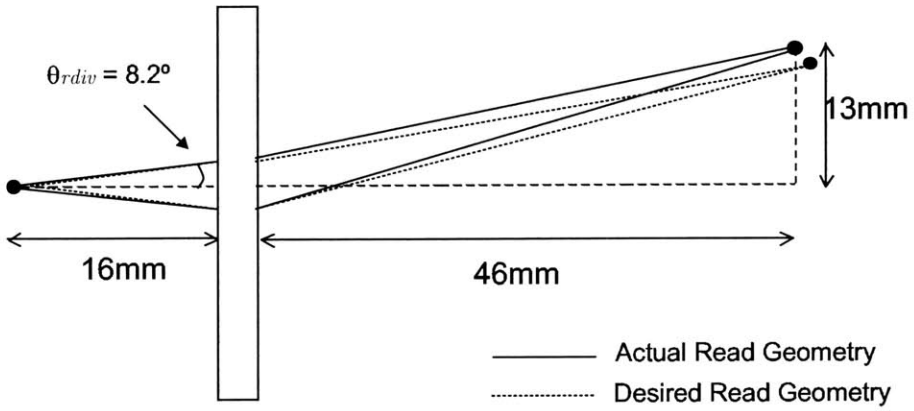

Figure 2-44: Actual read geometry compared to desired read geometry

As is known from the previous sections, the Bragg angle must be matched in order for the waves to constructively interfere and produce a maximum output. Measure-

ments were taken to determine the tolerance of the system to deviations from the Bragg angle. Figure 2-45 shows the normalized diffraction efficiency as a function of deviation from the Bragg angle. A sinc function has been fitted to the data as it is the theoretically predicted form of the curve [20].

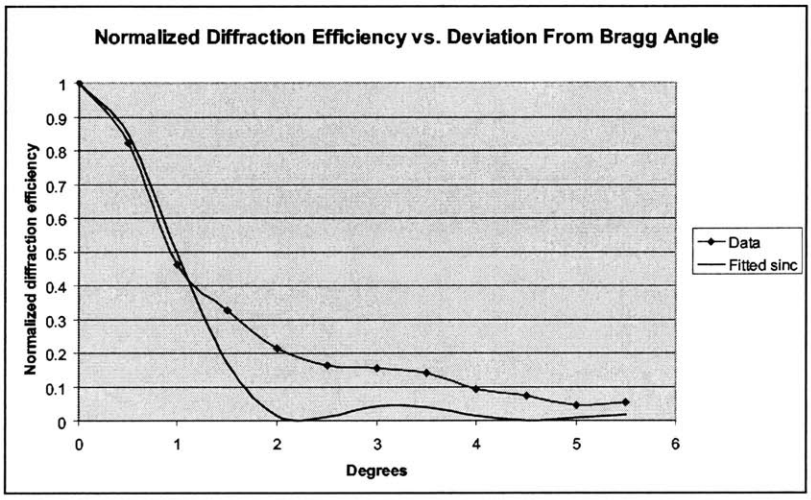

Figure 2-45: Tolerance of hologram to deviation in angle

2.4.5 Suggestions for future work

After working with the three different systems and observing their benefits and deficiencies it would be useful to draw some conclusions that could lead to an easier to use system that reliably produces good quality holograms. Of the three basic systems (microscope objective, single lens, and computer controlled beam splitter) the beam splitter setup offers the greatest flexibility and should be pursued further. It allows the write beams to be set at arbitrary angles without the problems that plagued both the microscope objective system and the single lens system.

The most drastic change that needs to be made to the computer controlled beam splitter system is its orientation. As described earlier, the bulk of the optical elements are mounted on to an optical breadboard vertically suspended over the rotary stage. This potentially leads to many problems. Because the optics and mounts are vertically suspended, they are subject to the forces of gravity without any support from

underneath. As in all holographic writing systems, movements of components of the system by fractions of a wavelength can result in significant degradation of the quality of the holograms produced [21]. Gravity acting on these components in a direction which they are not supported could lead to these movements. In one instance where the setup had not been touched for days, a lens fell out of a mount and broke on the optical table. The fact that this lens fell means that it had been moving for some time.

As may be remembered, the reason for this awkward vertically suspended setup was because of the 40 kilogram rotary stage. Since then it has been replaced with a somewhat smaller rotary stage. The solution, however, is to move to a motorized rotary optical mount such as the New Focus Model 8401 or the Oriel Model 13049. These mounts are usually used for waveplates and polarizers but would also work well for this application. In addition to the motorized rotary mount a motorized vertical translation stage is also needed. Using these two stages in conjunction with one of the existing motorized horizontal translation stages would allow the entire system to be in a single plane parallel to the table. In general, moving to a single parallel plane setup makes the system much easier to work with and align for the user.

In addition to ease of use, this new setup allows for easier initial characterization of holograms produced. The easiest way to test the diffraction efficiency of a hologram is to put it back into the system it was written with and block one of the two write beams. Measuring the output of the hologram (power in diffracted and undiffracted beams) will give an upper limit for the diffraction efficiency of the hologram. This eliminates potential alignment problems that arise when the hologram is put in another system to characterize its maximum diffraction efficiency. Users of the current system are unable to do this because the write beams are stopped directly on the other side of the holographic plate by the base of the rotary stage. In addition, reflections of the write beams from the rotary stage base in the current setup could also potentially be interfering and causing a decrease in diffraction efficiency.

Once the problems due to the vertically suspended optics are solved, effort should be directed towards increasing the accuracy and precision of the setup. The final

version of the CONNPP calls for holograms that are $83\mu m \times 83\mu m$ that need to direct a beam to a detector that may be as small as $25\mu m \times 25\mu m$. With the specifications laid out for the CONNPP if the angle of the output beam is off by even 1°, it will not hit its detector correctly. In the current system, it is extremely difficult to position the beams with such precision. Currently, to determine the angles, several measurements must be taken that incorporate estimations. The magnitude of these estimations is more than enough to cause inaccuracies of more than 1°. The measurements that are needed for the write setup are the distance of the divergent source from the emulsion, the size of the both the divergent and convergent beams at the emulsion, and the position of the convergent beam focus on the opposite side of the emulsion.

An example of how to alter the system to provide one of the needed measurements will be given here. An improvement to measure the distance from the source of the divergent beam to the emulsion will be discussed. Figure 2-46(a) shows the part of the setup containing the divergent beam in the region near the emulsion. The distance D is the measurement that needs to be obtained.

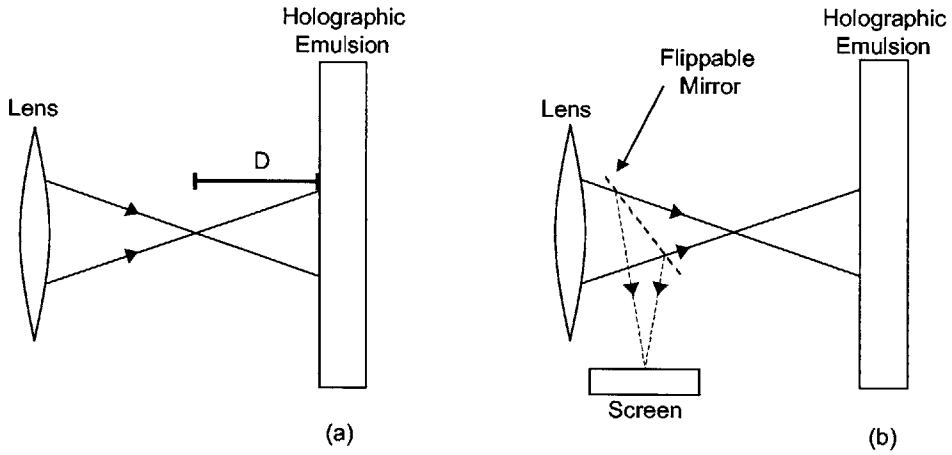

Figure 2-46: (a)Close-up of divergent beam near the emulsion (b)How to measure distance D with a flipper mirror

Figure 2-46(b) shows a possible way to measure the distance D without having to insert rulers or other undesirable objects into the aligned beam path. This solution consists of a flipper mirror (made by New Focus Corporation) and a screen on a ruled

rail system. First, the system must be calibrated so that it is know what ruler mark on the rail system corresponds to the holographic emulsion. This only must be done once when initially setting up the system. By then moving the screen on the rail such that the beam has a minimal spot size and noting the measurement on the ruled rail, the distance from the focus to the emulsion can easily be calculated. Furthermore, by moving the screen to the mark made for the emulsion position, the size of the beam at the emulsion can also be measured with a ruler without having to worry about touching any of the optics. Once the measurements have been taken, the mirror is simply flipped out of the beam path so that the beam propagates to the emulsion.

Another area that bears investigation is using multiplexed holograms. By having each neuron use a single multiplexed hologram and a single VCSEL, the size required for a single neuron could be reduced by the area of eight VCSELs and eight VCSEL drivers. This could easily comprise 40% of the total chip area necessary for a neuron. If the VCSELs are able to produce 1mW of power, this allows over 100μW per channel. As the photodetectors and amplifiers are designed for between 0-10μA, using a single VCSEL per neuron should be feasible from an optical power perspective. This scheme does have architectural ramifications however. In the original CONNPP design, each of the nine lasers was able to independently modulate its power. This allowed each connection from a neuron in plane n to have different optical intensity (i.e. weight) to each of the nine neurons it connects to in plane $n+1$. By using a single VCSEL for each neuron, this makes the intensity, and hence the neural weights, to the nine neurons in plane $n+1$ identical. Further architectural neural network simulations are necessary to determine if multiplexing the holograms at the expense of the connection model is an acceptable solution.

Chapter 3

Devices and Circuits

While the holographic elements discussed in the previous chapter are an integral part of both the holographic interconnect structure, devices and circuits are also serve a crucial role. As the VCSELs used were obtained from a third party without any control over their characteristics or manufacture, they were discussed briefly in the last chapter. The photodetectors and amplifiers for these photodetectors, on the other hand, will be covered in this chapter.

Two full custom chips have been designed thus far for the Compact Optoelectronic Neural Network Processor Project. One of these chips has been fabricated and tested. The other is in the design phase and is expected to be fabricated during the summer of 2003. In this chapter, the design, fabrication, and test of the first chip will be discussed. The key learnings from this chip were then incorporated into the second chip. Its design and simulated performance will also be shown.

3.1 Background

The Compact Optoelectronic Neural Network Processor consists of four separate subsystems. These subsystems are the holograms, the detectors, the logic and electronics, and the emitters. The last chapter focused on the holographic interconnects which dealt with three of these subsystems (holograms, detectors, and emitters) with the emphasis being on the holograms. This chapter focuses on the detectors and the first

part of the electronics block that interfaces the detectors with the rest of the electronics. This basically deals with the conversion of the optical signal to an electric signal along with the conditioning of the subsequent electric signal.

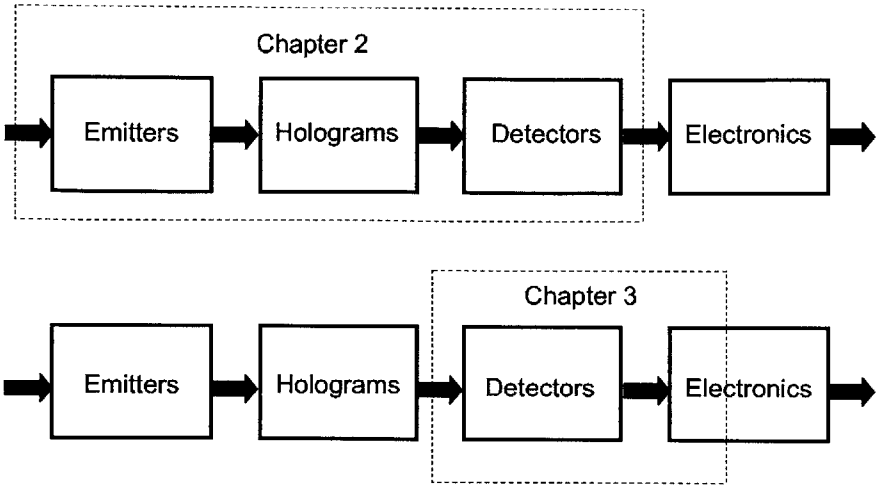

Figure 3-1: Illustration of how the material covered in the chapters fits in the architecture

In the actual system, the light conditioned by the hologram is focused on a photodetector. This photodetector converts a portion of the incident photons into electrons-hole pairs. Once these electrons and holes are created, they are instantaneously subjected to the fields within the photodetector device. If the device is reverse biased, the electron-hole pairs will potentially be ripped apart and collected by the positive and negative terminals of the device. This flow of electrons and holes will result in a current. At this point, the signal has effectively been converted from optical signal to electrical current signal. However, most logic and decision circuitry use voltage inputs rather than current inputs. As a result, the electrical current must be converted to a voltage. This can be done by a transimpedance amplifier. A transimpedance amplifier takes a current as its input and produces a voltage proportional to its input as its output. Generally, the current measured in the photodetector will be very small. So, in addition to converting the current to a voltage, the signal will also need to be amplified. This amplification is usually done after the signal has been converted to a voltage.

In order to design the photodetectors and amplifier circuits, it had to be determined which design parameters to optimize. The first metric to consider is speed. The goal of the CONNPP is not to rival speeds obtained by leading edge microprocessors. Rather, the CONNPP seeks to tackle classes of problems that traditional microprocessors are inherently bad at (i.e. recognition, classification, etc.). As a result, the speed of the devices and circuits was not paramount to their design. After all, the human-brain works with speeds on the order of milliseconds yet it routinely solves problems that computer scientists have been struggling with for 30 years using traditional microprocessors.

The next metric to examine is size. This takes into account both the size of the photodetector and the size of the amplifier circuits. With the CONNPP specification setting the size of a single pixel at $250\mu m \times 250\mu m$, size should be minimized as much as possible. As a result, the amplifier circuits should be made to take up as little chip real estate as possible while still performing their function well. The photodetectors, on the other hand, are a different matter. The size of the photodetectors is related to the input beam size and alignment characteristics of the system. By designing the holograms such that they focus the light on the detector the size of the detectors can be somewhat decreased. As the size of the photodetectors is decreased, however, they become more difficult to align with the beams. For this reason, a more conservative (i.e. larger) size should be chosen in the beginning of the project. As the micro-optic aspects of the system are developed further, it should become apparent what the minimum size of the photodetectors should be.

The last optimization parameter to examine is performance and functionality. For the photodetectors, the performance and functionality metrics are mostly tied to process parameters that cannot be changed. Factors that can be changed such as geometry will be explored. For the amplifier circuits, the most important performance metric is linearity. The linearity of the amplifier circuits can directly affect the maximum bit-level resolution that can be achieved by the system. A secondary consideration for the performance of the amplifiers is voltage swing. A larger voltage swing will potentially allow the electronics following the amplifier circuitry to have

less strict design parameters.

3.2 Photonic Systems Group Chip #1

It is well known that silicon, while cheap, is a less than ideal semiconductor for optoelectronic applications due to its indirect bandgap structure. For this reason, the initial specification for the Compact Optoelectronic Neural Network Processor Project called for the circuits and optoelectronic devices to be fabricated in gallium arsenide (GaAs). Furthermore, the commercial Vitesse HGaAs5 process was decided on due to its availability to the Photonic Systems Group. When the Photonic Systems Group Chip #1 (PSG1) was designed and fabricated, the HGaAs5 was still in the developmental stages. This led to some significant issues which will be discussed later in this section. This section will describe the design and testing of both the photodetectors and the amplifier circuits for the PSG1.

3.2.1 PSG1 Photodetectors

Beginning with the Vitesse HGaAs4 process, p-contacts were introduced to allow a fixed backgate potential. While the p-contacts were introduced to help with issues such as drain-lag and optical cross-talk between devices, they also provide the opportunity for new devices. Up until the HGaAs4 process, metal-semiconductor-metal (MSM) photodetectors were primarily used. While the MSM devices performed well, they required costly extra processing steps. Photodetectors that fit within the standard process flow are much more desirable. As a result, a lateral pin was proposed and designed by Joseph Ahadian [26]. Without p-contacts, these devices were unable to be built in the pre-HGaAs4 processes. The basic structure of the lateral pin photodetector can be seen in Figure 3-2.

Traditionally, pin photodetectors use a heavily doped p and n region with the material between remaining close to intrinsic doping. In many situations this results in the whole intrinsic region being depleted even at zero bias. Unfortunately because of process constraints, having a close-to-intrinsic region between the two contacts was

not an option. In the case of the HGaAs5 process, there is a difference in doping of about 10^3 between the contacts and the p- region. If the device was being engineered from scratch without the constraints of the HGaAs5 process, a doping difference of about 10^{10} would be normal. In the original layout geometry, the spacing between the n-type and p-contact material is 2.0μm. In an unbiased state, the depletion region extends into the p- region by about 0.4μm. A graph of depletion width as a function of reverse bias can be seen in Figure 3-6. As the material parameters do not change with the geometry, the curve in Figure 3-6 holds for all the devices discussed in this section.

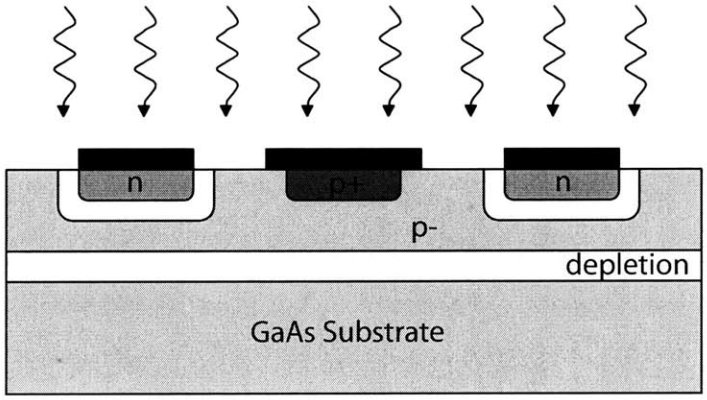

Figure 3-2: Cross-section of lateral pin unit cell

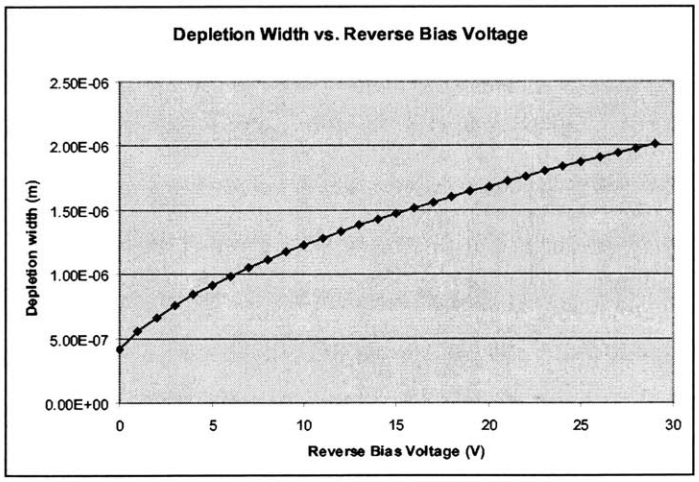

Figure 3-3: Depletion width as a function of reverse bias voltage

As can be seen in the Figure 3-2, there is a p- layer on top of the substrate. This causes a small depletion region between the substrate and p- due to doping mismatch. Next, the source/drain implant (n-type doping) can be seen just below the surface of the wafer. Once again, due to the doping mismatch between the n-type and p-material, there is a natural depletion region formed. The black rectangle on top of the n-type material is the metal contact. As a side note, the metal contacts on both the n-type and p-contact regions do block the incoming light. The light is incident to the surface of the wafer. The light will enter the device everywhere there isn't metal blocking. While Figure 3-2 shows what is happening conceptually, the actual layout is somewhat more complicated. The layout for a $6.4\mu m \times 6.4\mu m$ pin unit cell can be seen in Figure 3-4. This unit cell is then tiled to the desired detector size. All the p-contacts are connected to one another as are the active area n-type regions.

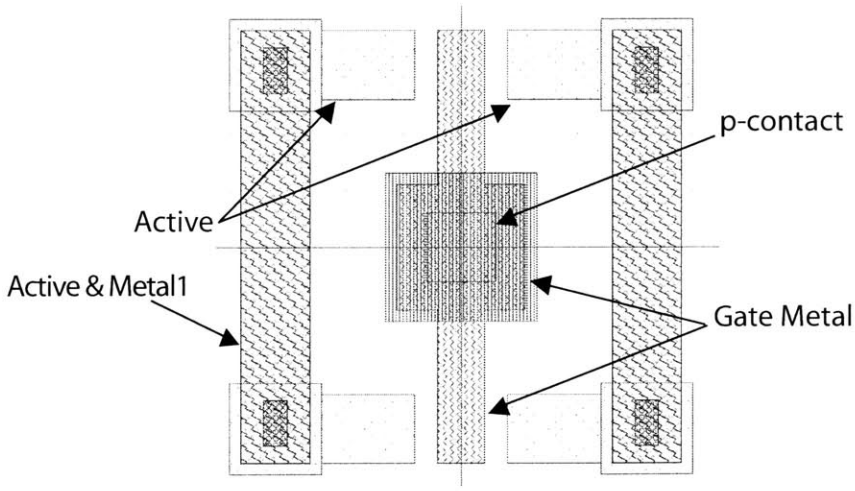

Figure 3-4: Layout of the original lateral pin unit cell

The capacitance for this device can be estimated from the geometry of the device along with the junction capacitance for the process. In a 4V-6V reverse bias condition, the junction capacitance should be approximately $0.05fF/\mu m$ of active area edge. The unit cell in this initial geometry has a active area edge of $20\mu m$. This results in a capacitance of about 1fF/unit cell. In Figure 3-5 the unit cell has been tiled to create a $75\mu m \times 75\mu m$ photodetector. This particular detector is made up of 120 unit cells

for a total capacitance of 120fF.

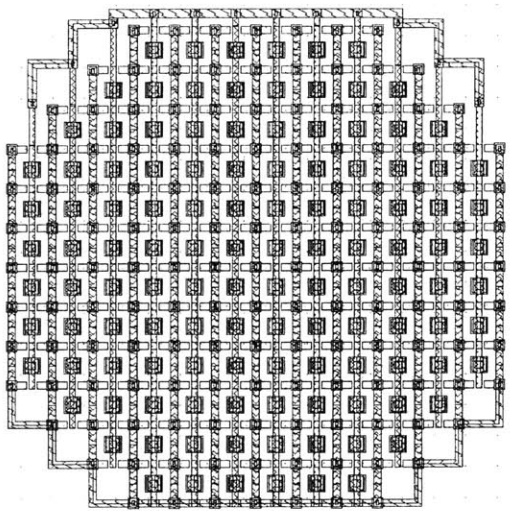

Figure 3-5: Layout of the 75μm$\times$75μm lateral pin

In addition to this original lateral pin geometry, three other lateral pin geometries were also laid out on PSG1. While the original geometry was minimum sized for the HGaAs4 design rules with 2.0μm between the n-type and p-contact regions, a new geometry was created that had minimum spacing for the HGaAs5 design rules. The unit cell has exactly the same structure as the original device, but it has been shrunk. The size of the full detector after the unit cell has been tiled remains the same, 75μm$\times$75μm. The n-type to p-contact spacing in this new device is 0.9μm and the unit cell size is now 3.4μm$\times$3.4μm. This means that the p- region is now fully depleted at a reverse bias voltage of only 5V. The capacitance per unit cell has dropped to 0.5fF/unit cell; however, the total capacitance has risen to about 186fF (50% larger capacitance than the original geometry). The minimum HGaAs5 geometry can be seen in Figure 3-6.

Another useful metric to examine when talking about these devices is fill factor. Fill factor, for the purposes here, is defined as the ratio of area exposed to optical light to total area of the detector. In the case of the two detectors discussed so far, the Metal1 and Gate Metal layers will block light from hitting the detector. The original

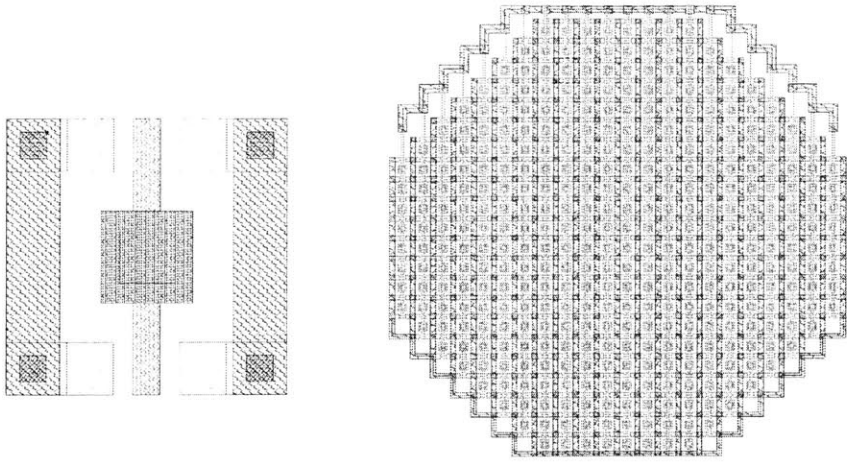

Figure 3-6: (a) Unit cell for the minimum HGaAs5 geometry (b) Full 75μm×75μm detector

detector with 2.0μm n-type to p-contact spacing has a fill factor of 61%. This means that 61% of the light hitting the detector will actually interact with the device. The minimum HGaAs5 geometry detector has a fill factor of 52%. Incrementally, the original geometry detector will see 15% more light than the HGaAs5 minimum detector.

In addition to the two detector structures already discussed, two more pin photodetectors with a simpler geometry were also designed. Instead of using the more complex active area ring with a p-contact in the middle, these detectors simply use strips of metal-1 covered active area and gate covered p-contacts. These strips are then alternated to produce the desired detector size. Two different types of this strip detector were designed. The first uses minimum spacing (0.9μm) between the n-type material and the p-contact. The second device has a larger spacing of 1.5μm between the two strips. An example of the geometry for the strip pin photodetector can be seen in Figure 3-7.

Looking at some of the key metrics for the strip photodetectors, it can be seen that for the minimum strip photodetector capacitance is approximately 151fF and the fill factor is only 35%. For the strip photodetector with 1.5μm spacing, the capacitance

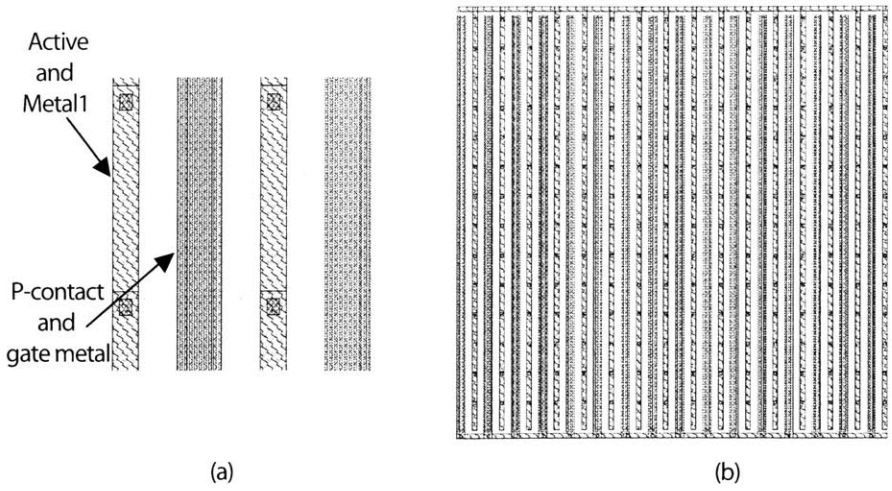

Figure 3-7: (a) Unit cell for the strip pin photodetector (b) Full 72μm×72μm strip detector

is about 115fF and the fill factor is 52%.

3.2.2 PSG1 Amplifiers

As mentioned previously, the PSG1 chip was designed and fabricated in the Vitesse HGaAs5 process. This fact has some significant consequences which bear discussion. Most of the chips designed and fabricated for microprocessors, memories, DSPs, etcetera are fabricated in silicon rather than GaAs. Most of today's silicon based process use CMOS (Complimentary Metal Oxide Semiconductor) technology. CMOS technology uses two flavors of MOSFETs (Metal Oxide Semiconductor Field Effect Tranisistors) to create a low power, high speed logic family. As a result, CMOS and MOSFETs are what most circuit designers are comfortable with. Because of the material parameters, GaAs does not use MOSFETs or CMOS technology. For logic type operation, GaAs can use MESFETs (MEtal Semiconductor Field Effect Transistor) and the DCFL (Direct Coupled FET Logic)logic family [27].

The major difference between MOSFETs and MESFETs is the the structure of their gates. MOSFETs, as their name implies, have their gate separated from the substrate by an extremely thin insulating oxide layer. As a result, the impedance

looking into the gate of the MOSFET is almost infinite. MESFETs, on the other hand, have a metal gate that is directly in contact with the substrate. This results in a Schottky barrier diode formed at the surface of the device. This, in turn, means that the gate is not isolated from the rest of the device and will have leakage current through to the bulk. In addition to this major structural difference between MESFETs and MOSFETs, there are also material parameters such as mobility and bandgap energy that make silicon and GaAs very different to work with.

In the course of designing the first generation amplifier, many different amplifier topologies were examined and explored. In the end, a topology was chosen that was based on the work of Joseph Ahadian and outlined in his Ph.D. thesis [26]. This original topology was modified to meet the requirements for the CONNPP. A block diagram of amplifier topology used for PSG1 can be seen in Figure 3-8.

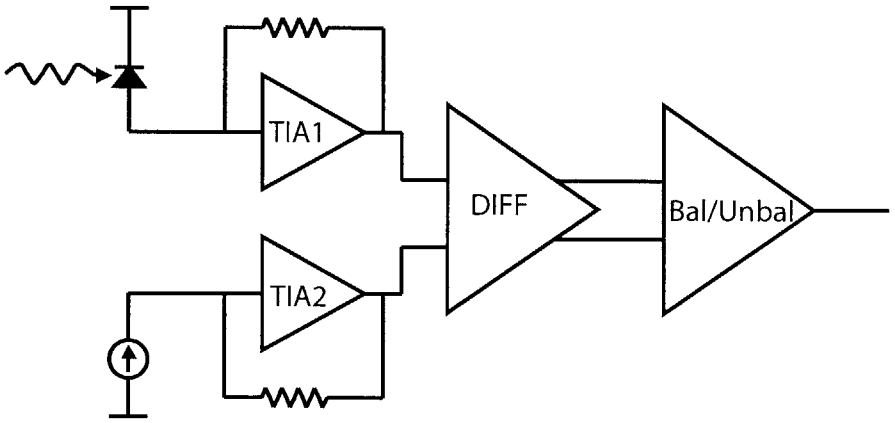

Figure 3-8: Amplifier topology for PSG1

The first stage of the topology consists of two transimpedance amplifiers (TIAs). These TIAs convert the current from a photodetector or other current source to a voltage. The output of the TIAs is then fed into a differential amplifier which amplifies the difference between the two input signals while discarding the common-mode signal. The output of the differential amplifier is itself a differential signal. This differential signal is then input into a balanced-to-unbalanced converter which takes a double-ended input and produces a single-ended output.

The transimpedance amplifier (schematic in Figure 3-9, layout in Figure 3-10)

utilizes feedback to convert the current at the input to a voltage at the output. The basis for the TIA begins with a simple common source amplifier made up of enhancement mode MESFET E1 and the resistor R1. This common source amplifier then has another resistor (R2) added as a feedback element between the input (Iin) and the drain of E1. The diode (D1) has been added to raise the voltage level of the output to correctly bias the differential stage that follows the TIA. This amplifier works by pushing (or pulling) current into the input terminal. The only path for the current to travel is through the transistor E1 to ground. As the current through E1 is being forced to change by the input, the drain voltage of E1 (and hence the voltage at the output) must also change. Also, as the current input modulates minutely, so will the voltage at the output.

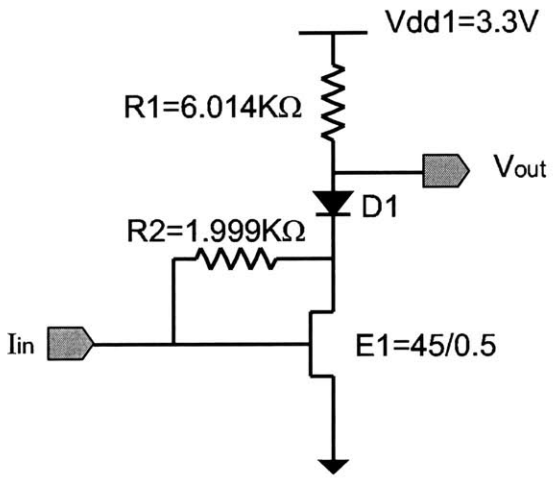

Figure 3-9: Transimpedance amplifier schematic

The choice of the feedback resistor (R2) is a tradeoff between noise performance and bandwidth of the TIA. If the feedback resistor is large, it will cause a large change in voltage at the drain of the transistor giving better noise performance [28]. At the same time, it will also increase the RC time constant at the input and hence will reduce the bandwidth capabilities of the amplifier. Conversely, making the feedback resistor small will make the TIA more susceptible to noise while increasing the maximum frequency at which it can be operated. The RC time constant (τ_{in}) at the input is given by [26]

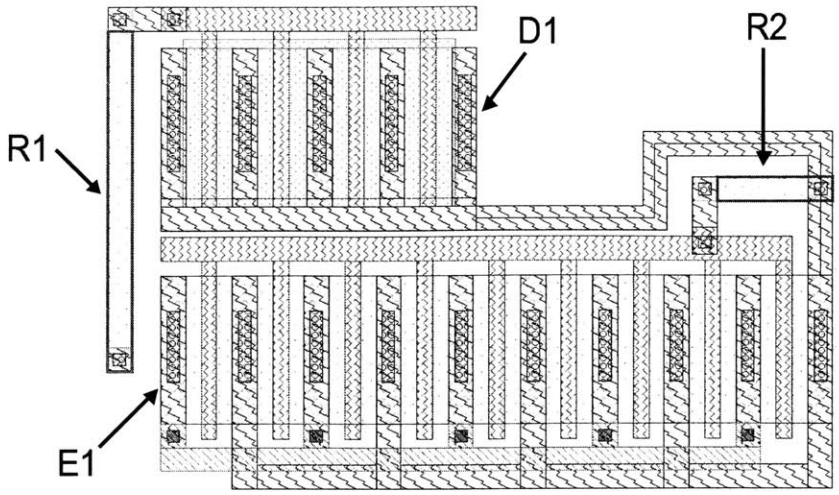

Figure 3-10: Transimpedance amplifier layout

$$\tau_{in} = R_{in}C_{in} = C_{in}\frac{R_2}{(1+G)} \tag{3.1}$$

where C_{in} is the total capacitance between the input node and ground and G is the gain of the common source amplifier without the feedback resistor. It is clear that as R_2 increases so does τ_{in}, reducing the bandwidth.

The transimpedance amplifier designed and shown in Figure 3-9 was simulated using Vitesse's proprietary MESFET models. The results show the amplifier having a transimpedance of about 1800Ω. Therefore, the 10μA input swing produces an 18mV swing at the output of the TIA.

The output voltage signals from the two transimpedance amplifiers are then input into a differential amplifier. This differential stage is made up of two common source amplifiers with their sources coupled together through a common current source. The schematic of this amplifier stage can be seen in Figure 3-11 and the layout in Figure 3-12. When the input signals on the two sides of the differential amplifier are the same, the outputs of the two sides of the differential amplifier are also the same. The current through the left side of the differential amplifier is I_1 and the current through the right side is I_2. The sum of the currents $I_1 + I_2$ flows through E6. As E6

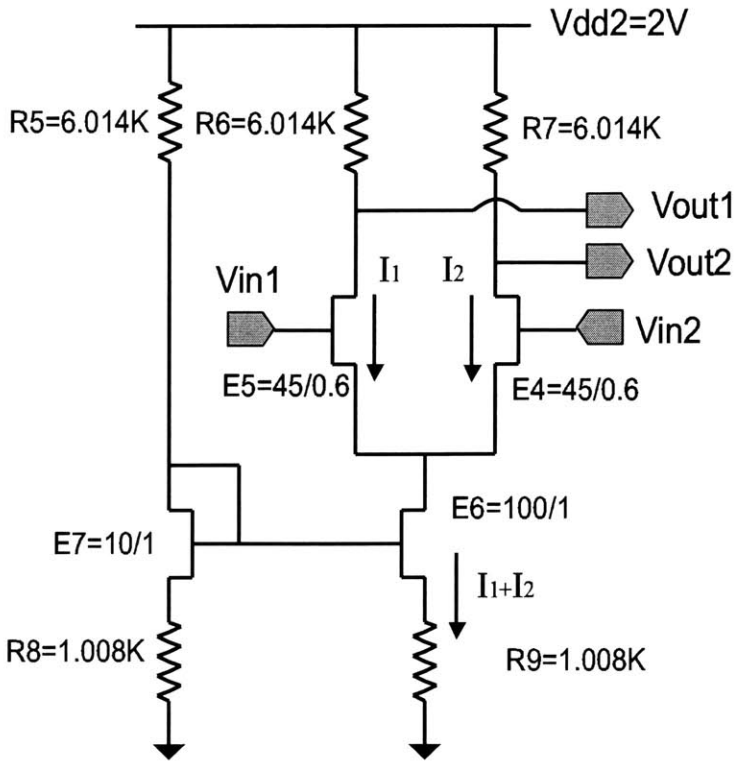

Figure 3-11: Differential amplifier schematic

is a statically biased current source, the current through E6 must remain constant. This mean that if the voltage V_{in1} increases and causes I_1 to increase, the biasing must change for E4 and hence I_2 must decrease. The voltage signals at V_{out1} and V_{out2} will be complementary about a common quiescent voltage.

Once again, the proprietary Vitesse SPICE models were used to simulate the performance of the differential amplifier designed. The differential amplifier was simulated using the 18mV swing output voltage from the TIA simulation as an input to one of the sides of the differential amplifier. The other side of the differential amplifier used the output voltage from the TIA that is held at a constant bias. The results show a differential gain ($(V_{+out} - V_{-out})/V_{in}$) of about 11. For the 18mV input swing on the differential input, a differential output swing of about 200mV is seen.

The third and final stage of the amplifier is a balanced-to-unbalanced converter. This stage is meant to take the complementary double-ended output from the differ-

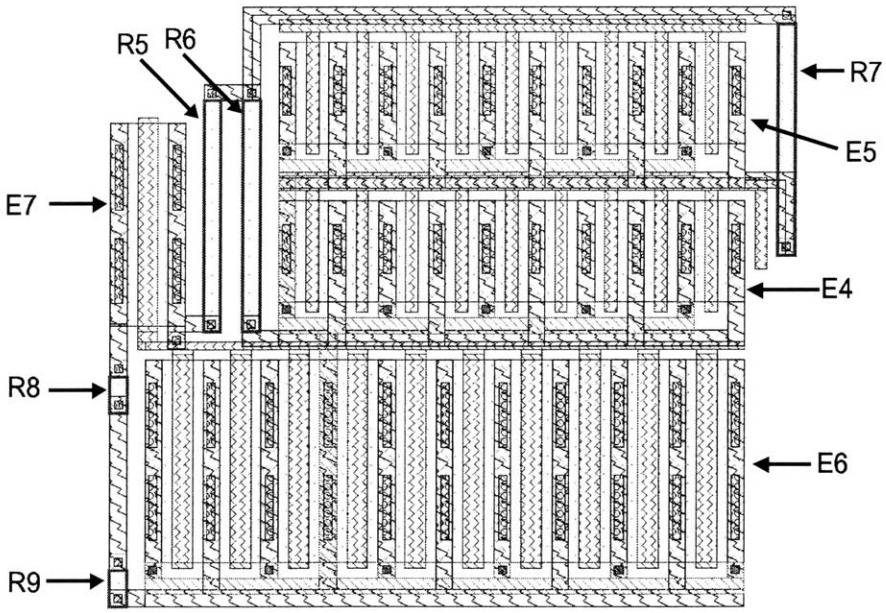

Figure 3-12: Differential amplifier layout

ential stage and convert it to a single-ended output. The circuit accomplishes this through a clever use of a current mirror structure. The circuit schematic can be seen in Figure 3-13(a).

This circuit consists of the output of the differential stage input into two depletion mode FETs (DFETs). The lower half of the circuit is a simple current mirror. The voltage on V_{in1} sets the bias level on DF1 which in turn sets the current level through E8. Because E8 and E9 make up a current mirror, whatever current flows through E8 must also flow through E9 (i.e. $I_1 = I_2$). However, it was stated previously that V_{in1} and V_{in2} are complementary. This means that if V_{in1} decreases, V_{in2} must increase. Starting from equilibrium, assume that $V_{in1} = V_{in2}$ which explicitly implies that $I_1 = I_2$. Now, V_{in1} decreases slightly. This causes I_1 to increase slightly. Because of the current mirror structure, when I_1 increases, I_2 must also increase. Because of the complementary input structure, V_{in2} must increase slightly. This increase in V_{in2} will make the current through E9 want to decrease; however, this current is set by the current mirror structure and must increase. As a result, to meet both of these conditions (increasing V_{in2} and increasing I_2) V_{out} must decrease significantly

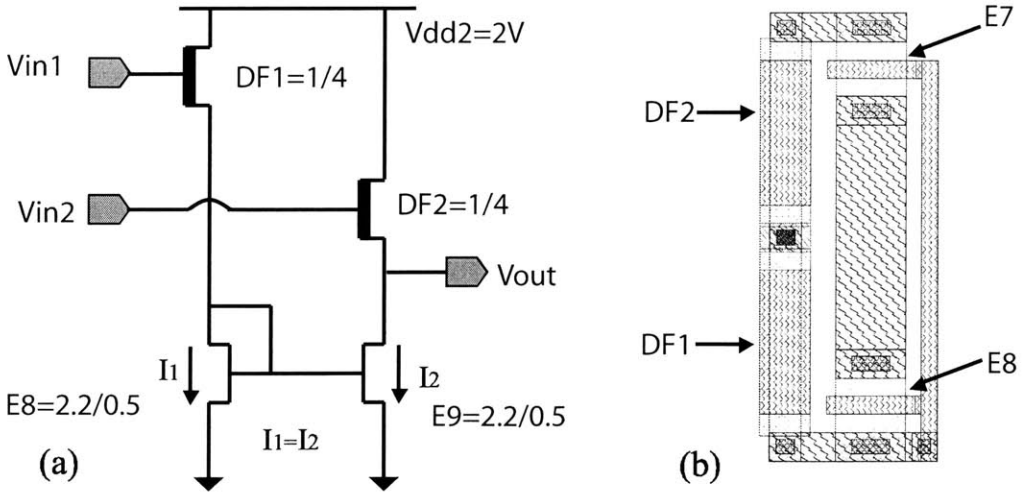

Figure 3-13: Balanced-to-unbalanced circuit schematic

to accommodate these two contradictory forces. By using this structure, the voltage swing on V_{in1} has effectively been effectively negated (by changing it to a current) and then added to the other input, V_{in2}. While the output will be somewhat less than the sum of the signals, it should still produce a single ended output with a magnitude greater than either V_{in1} or V_{in2} (about 150mV instead of the total 200mV for the full differential signal).

The schematic for the complete amplifier circuit with all three stages can be seen in Figure 3-14(a). The layout for the circuit can be seen in Figure 3-14(b). The total size of the circuit is about 46μm$\times 96\mu$m. Notice that there is a metal-3 layer covering the entire circuit. Because the intent is to use this circuit for optoelectronic applications, it is necessary to optically shield the circuit. The uniform metal layer does this effectively. Without the optical shielding, significant numbers of carriers could be created within the substrate and potentially affect the functionality of the underlying circuits. Another structure included in this layout that cannot be seen on the layout of the individual stages is p-contact regions. Along the left hand side, top, and right hand side of the layout can be seen sizable blocks of p-contacts. These p-contacts are contacted to ground in order to hold the substrate at given potential to help keep backgate effects at a minimum.

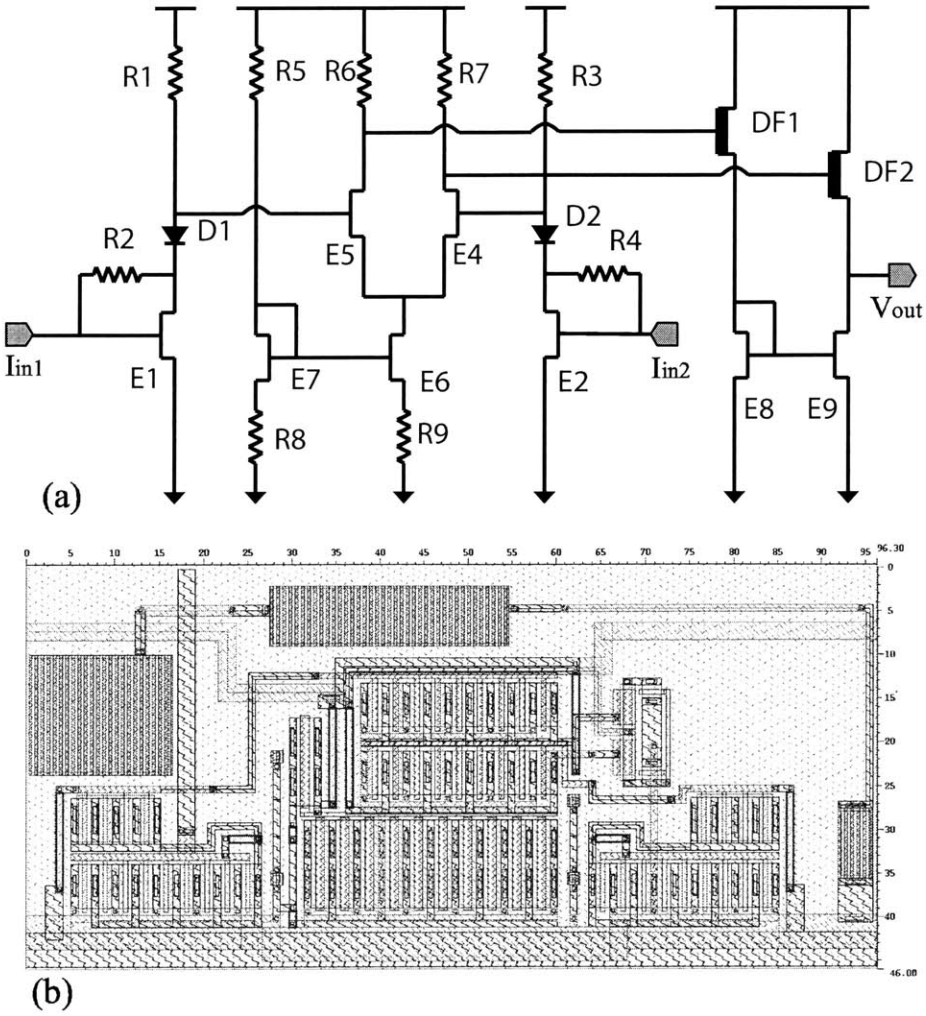

Figure 3-14: Complete 3-stage amplifier circuit schematic and layout

3.2.3 PSG1 testing

The PSG1 was the first chip designed by the Photonic Systems Group in a number of years. As all of the students involved in previous chips had graduated, the group basically had no experience in this area. As a result of this, there were some major challenges faced in the design and test of the chip. These challenges and the learnings gained from these challenges will be discussed here.

The design for the Photonic Systems Group Chip #1 was finished and sent to Vitesse in the Fall of 2001 for fabrication. The finished chips were received from Vitesse in the early Spring of 2002. The chip was fabricated as part of a Vitesse

test chip containing other experimental devices and circuits. The approximately 35 chips were received from Vitesse as bare die without being packaged. The fact that we received bare die instead of packaged parts ended up representing one of the first challenges in testing the PSG1.

After appropriate chip packages were obtained, the chips had to be mounted in the packages. Once mounted in the package, a wire bonder was used to connect the on-chip pads to the lead frame of the package. While the wire bonding was not particularly difficult to do after a couple days of practice, it did lead to some questions of integrity of the connections. The on-chip bond pad sizes were suggested by some of the designers at Vitesse as typical for this type of chip. Unfortunately, it was not considered that almost all of Vitesse's chips (even test chips) are bonded mechanically rather than manually. While an automatic wire bonder would have little difficulty making connections to pads of this size ($75\mu m \times 110\mu m$), it was quite challenging with the manual wire bonder. Making the connections was relatively easy, but it was always questionable if the tail of the wire bonded to the on-chip pad was possibly making contact with other structures on the chip. If this happened, it could possibly lead to a short-circuit, damaged devices, or generally flaky behavior of the circuits.

There are a couple different solutions to the wire bonding problem. The first solution is to increase the size of the on-chip bond pads making them large enough to lay down a bond without question of rogue connections. The other solution is to have the bare die packaged and wire bonded before delivery. Enlarging the bond pads will require more chip real estate which will in turn cost more money to produce. On the other hand, having the chips packaged and bonded will also add to the bottom line cost (usually about $1000). While this trade off could be evaluated for each chip given the number of pads and cost/mm^2 for chip real estate, it seems that for peace of mind sake it would be better to plan on having the chips packaged and bonded by a third party vendor.

The next issue encountered involved the performance of the amplifier circuits. The PSG1 included four amplifier circuits of the type described in the last section connected to the four different types of photodetectors. Additionally, there was also

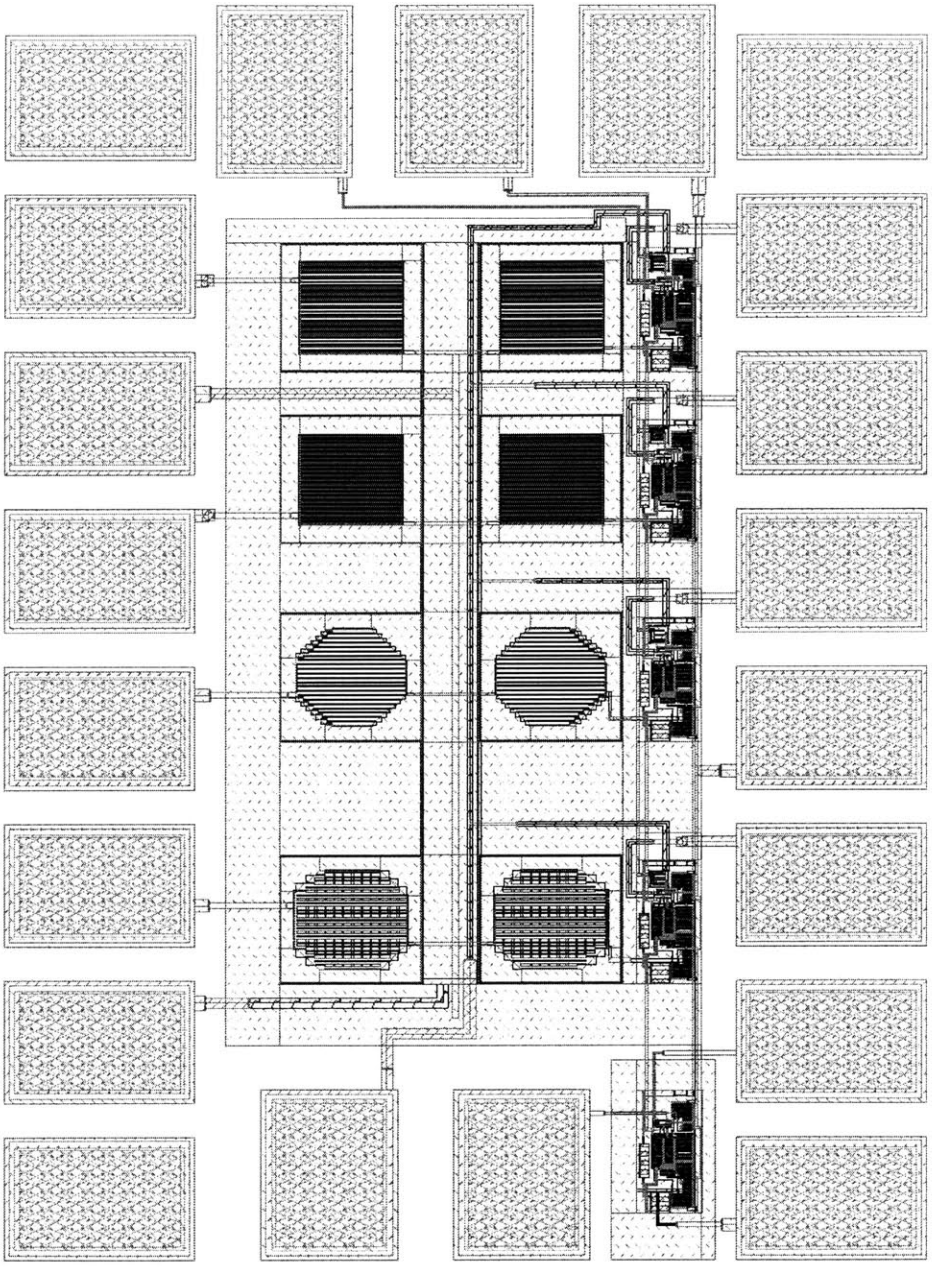

Figure 3-15: Layout of the full PSG1 chip

a single amplifier not connected to any photodetector to allow for electrical-only testing of the circuit. This electrical-only circuit was the first targeted for testing. It was expected that the circuit would produce a variable voltage output linearly proportional to the current difference between the two inputs. When the circuit was biased up as designed and the appropriate current inputs were provided, a static voltage was seen at the output. After trying to adjust all the available variables (V_{dd1}, V_{dd2}, I_{in1}, and I_{in2}) the output voltage still did not show any dynamic change as a function of I_{in1} and I_{in2}.

As a digression, the situation above highlights a design flaw that resulted from inexperience in designing chips. The discussion of the flaw is included here to prevent it from happening again. The 3-stage amplifier was only laid out on PSG1 as a single, monolithic circuit. As a result, the only nodes that were accessible were the two inputs, the output, and the two bias voltages. As a result, the circuit was extremely difficult to troubleshoot. In hindsight, the individual stages of the 3-stage amplifier should have been laid out separately in addition to the single monolithic version. The circuits might not have worked any better as a result, but it would have been much easier to troubleshoot and learn more about specifically where the problem was in the amplifier.

Once it was determined that the 3-stage amplifier circuit was not functioning correctly, it was necessary to determine why it was not functioning correctly. The limited number of internal circuit nodes accessible significantly hindered debug of the circuit. As the output node did not show any change in voltage with changing inputs, the only thing left to look at was the input nodes. As discussed earlier, the circuit functions by taking two input currents (I_1 and I_2). The most telling test that was done was to set the current at both input nodes to 0A and then measure the voltage at these input nodes. It was expected that the voltages measured would be identical. Unfortunately, that was not the case. Shown in Figure 3-16 is a table showing the input voltages measured on four different chips with the input currents set to 0A. The fact that there is over 30mV of mismatch between the inputs in some cases suggests that the inputs are not matched.

	Vin1	Vin2
Chip #1	394mV	378mV
Chip #2	390mV	353mV
Chip #3	383mV	370mV
Chip #4	363mV	329mV

Figure 3-16: Table showing input voltages on four different chips with input currents set to 0A

Once it became clear that the inputs were not matched, it was necessary to determine what was causing the problem. The circuit had been modelled during the design phase at all the available corners with up to 3σ process variation in each direction. While in some cases, performance was not optimal, functionality was always preserved. Through conversations with Vitesse, the resistors became the main focus of the debug effort. As was mentioned previously, the HGaAs5 process was still in development during the period in which the chip was designed and fabricated. After speaking Vitesse about the models used for simulation, it was determined that the models did not take into account process variation of the NRES resistors used in this design. Furthermore, these resistors could vary up to 15% of designed value within the same chip. Using this information, the amplifier circuits were re-simulated at various process corners with various resistor mismatches in the first stage. In doing this, a situation was found that approximately matched what was being seen in the lab. By running a 2σ process corner with a 10% resistor mismatch in the transimpedance stage, the simulation showed a similar level of mismatch of the input voltages to the voltage mismatch measured in the lab. Even without having the appropriate models, this situation could have probably been avoided. In the layout for the amplifier, the two transimpedance amplifiers were put on opposite sides of the differential amplifier

for ease of layout. This meant that the resistors in the different TIAs were separated by about 100μm. If these dual TIAs had been laid out as a single block with the resistors sitting right next to each other, the mismatch would probably not have occurred. Unfortunately, it was unknown at design time that this was a significant issue partially due to inexperience on the part of the designer and partially due to incomplete information from the manufacturer.

This leads to another tangential issue about choosing which process to work within for designing ICs. While the Photonic Systems Group was very fortunate to work with Vitesse Semiconductor using the HGaAs5 process for the PSG1, working with an experimental process also proved somewhat problematic. With models and design parameters continually in a state of flux during development, working in process that is still in development can prove daunting even for a circuit designer or engineer directly employed by the company developing the process. These employees have the luxury of a better flow of information about the development between groups within the company. Being external to the company adds another level of complexity. It makes it quite difficult to ensure sure that the most current design rules and design models are always being used for the design. If the decision is made to use an experimental process again, it is very important to manage the connections and try to be included in the group that has access to all the necessary and up-to-date information.

3.3 Photonic Systems Group Chip #2

One of the inherent problems using GaAs is that it is essentially opaque to light at the wavelengths of interest (i.e. 850nm). This is a result of GaAs being a direct bandgap semiconductor and the absorption curve at the bandgap edge increasing sharply. The fact that GaAs is opaque to light of this wavelength means that for the CONNPP both sides of GaAs chips would need to be processed in order to produce an OEIC that could be cascaded. The emitters would need to be grown or bonded on one side of the chip, and the detectors and electronics would need to be grown, bonded, or

patterned on the opposite side of the chip. Processing both sides of a chip or wafer would require several non-standard process steps some of which would need to be developed specifically for this project. As a result, other alternatives to GaAs have been explored.

One of the most attractive options for the CONNPP is Peregrine's 0.5μm ultra-thin silicon (UTSi) CMOS process. This process consists of a very thin layer of silicon (120nm) on top of a sapphire substrate. This sapphire substrate acts as an insulator also making this a silicon-on-insulator (SOI) process. In addition, sapphire is an optically transparent material for light at a wavelength of 850nm. The thin silicon layer, because of its thickness, is also essentially transparent to the 850nm light. This means that the emitters, detectors, and electronics can potentially all be grown, bonded, or patterned on the same side of the chip or wafer. The difference between the GaAs and UTSi SOS implementations can be seen in Figure 3-17. Due to the benefits gained by using the Peregrine UTSi process, it was chosen for the Photonic Systems Group Chip #2 (PSG2). It was determined that the PSG2 would include amplifiers, photodetectors (if possible), aligned pilar bonding structures, and analog to digital converters. The amplifiers and the photodetector structures will be discussed here.

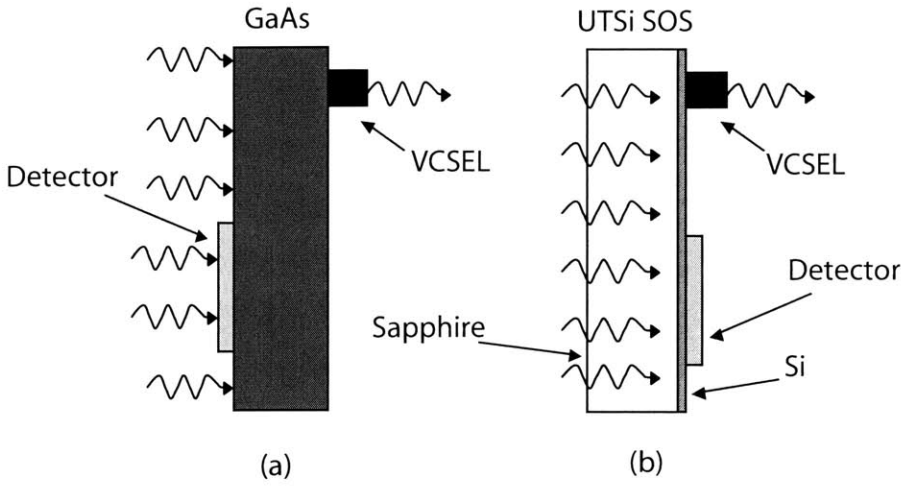

Figure 3-17: (a)GaAs chip with two sided processing (b) UTSi SOS chip with all devices on a single side

3.3.1 PSG2 Amplifier

With lessons learned from PSG1, the next generation amplifier was designed with an eye towards simplicity. Once again, several different topologies were researched, examined, and simulated. With potential revisions in the overall architecture, linearity of the amplifier circuits is of paramount importance. The resolution of the architecture as a whole is potentially limited by the linearity of the amplifiers, detectors, and emitters. Whereas there is only limited design control over the linearity of the detectors and emitters, the amplifiers provide much more control over the linearity through their design. Therefore, they were designed such that they would not be the weakest link in the system in terms of bit-level resolution.

Most of the literature on optical receivers and transimpedance amplifiers is focused on digital communication for high speed optical networks [29–31]. In these cases, the designers are concerned with eye diagrams and obtaining correct logic levels. The CONNPP, however, seeks to optically transmit information in an analog fashion between the planes of neurons. Furthermore, the speed of the CONNPP is not a key metric (at least at this stage of the project). While the amplifiers on the PSG1 chip were adapted from a digital optical interconnect application, the amplifiers on the PSG2 were designed from scratch.

After investigating various transimpedance amplifier topologies, it was determined that an operational amplifier (op-amp) with negative resistive feedback provided the best linearity with the largest voltage swing. A top-level diagram of the amplifier structure can be seen in Figure 3-18.

While there are many varieties of op-amp, a simple two-stage CMOS op-amp topology was chosen for its simplicity and potential small footprint [32]. The first stage of this two-stage CMOS topology consists of a differential PMOS pair with a single-ended output. This single-ended output is then input into a simple common source NMOS amplifier with current supply. A feedback resistor is then connected between the input and the output. A schematic of this amplifier can be seen in Figure 3-19 complete with transistor and resistor sizings.

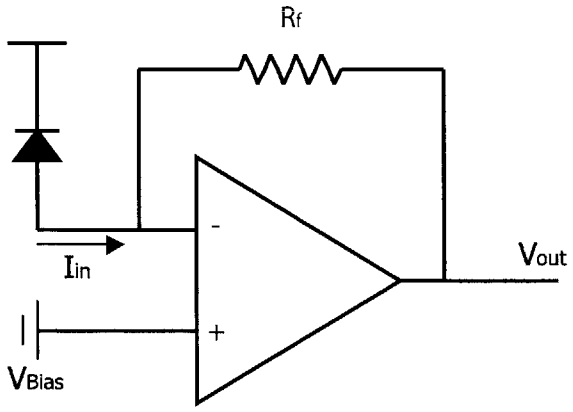

Figure 3-18: Top level diagram of op-amp in transimpedance configuration

It might be noticed that the resistors in this design are unusually large for on-chip applications. One of the benefits of the Peregrine 0.5μm UTSi CMOS process is that it includes a resistor structure with very high resistivity (SN resistor) [33]. This resistor structure exhibits uniformity and variation with temperature similar to other, lower resistivity structures. In the previous design on PSG1, resistors played a crucial role in terms of biasing differential structures and other stages. This inclusion of resistors in the differential structures became a weak point of the design. In the op-amp design used in PSG2 (Figure 3-19), the resistors are only used for a single current supply bias and the resistive feedback element. This eliminates the need for near-perfect matching for resistor structures. The models from Peregrine did include resistor process corners. These corners were simulated and it was shown that the the variation in resistor values affects the operation of the op-amp circuit in a very minor fashion.

The circuit in Figure 3-19 was simulated using the Spectre models obtained from Peregrine Semiconductor in their 0.5μm UTSi CMOS process development kit (PDK). The amplifier was designed to take a 0-10μA input from a photodetector and produce voltage swing on the output. The full range (-100μA-100μA) transfer function for the amplifier can be seen in Figure 3-20.

It can be seen from the full range characteristic that the amplifier has a fairly linear range all the way from -25μA to about 20μA. The best way to characterize

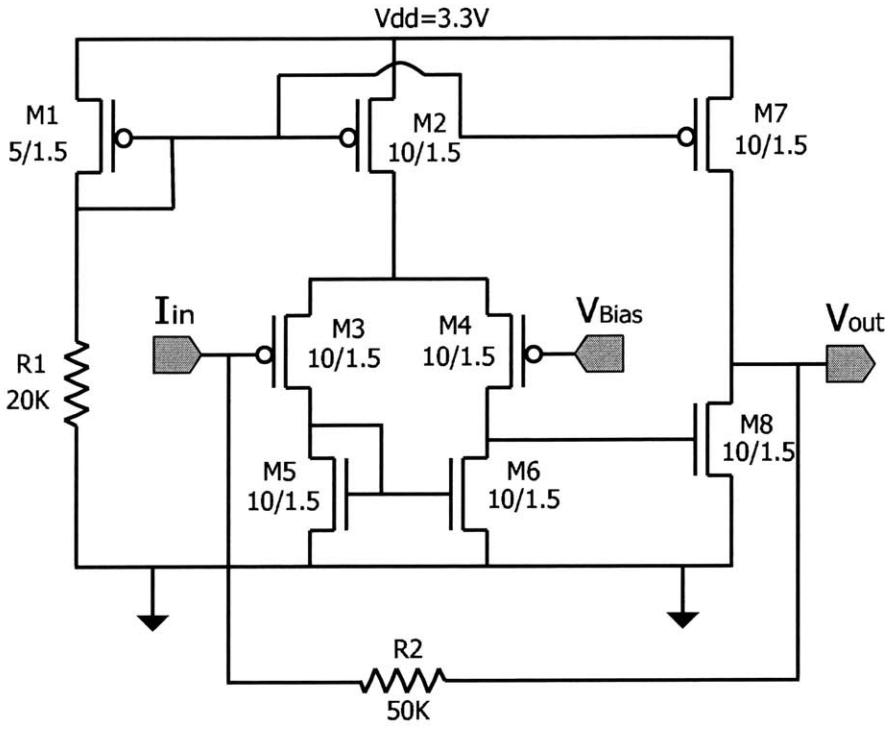

Figure 3-19: Schematic for the two-stage op-amp with negative resistive feedback

the linearity of the amplifier over a given range is to examine the derivative of the transfer function. If the transfer function were perfectly linear in a particular range, it would be expected that the derivative over that range would be a constant. Shown in Figure 3-21 is the derivative of the full range transfer function.

There are a couple of important things to note from Figure 3-21. First, the values of the derivative are very large in magnitude (almost 50kΩ). This transimpedance value is actually the gain of the amplifier. In this case, for every 1A of input swing, 50,000V of output swing will be seen. The target input current swing was about 10μA which, by this convention, corresponds to about 0.5V of voltage swing. Next, as expected, the derivative has a relatively flat section in the range right around 0A. If the target range for the input current is 0-10μA, then it would be desirable to have the absolute flattest part of the derivative curve centered around the middle of that range (5μA). In Figure 3-22, the transimpedance curve has been zoomed in on such that the area of interest (0-10μA) can be seen clearly.

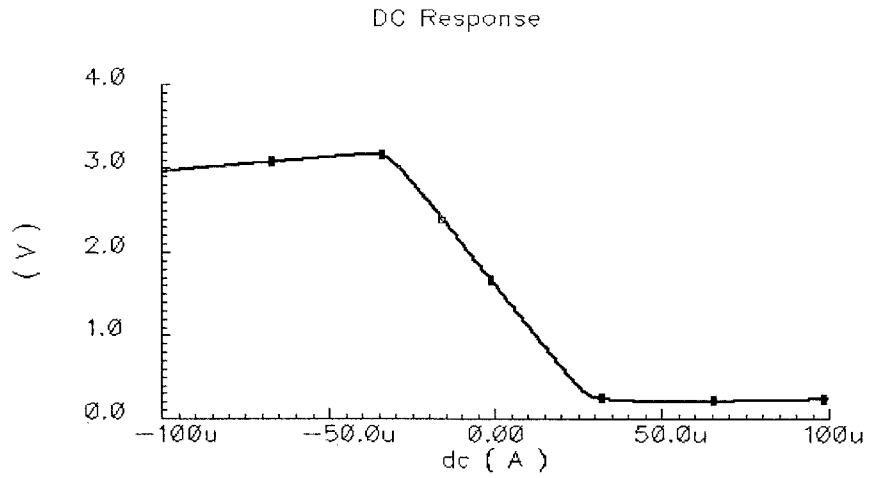

Figure 3-20: Transfer characteristic for PSG2 amplifier (Current input, voltage output)

Notice that the value of the derivative function varies only about 20Ω out of 50,000Ω in the range of interest (0-10μA). This represents a 1 part in 2500 variation of the linearity over the full range of interest. While many of the system's architectural details are yet to be hammered out, it has never been suggested that the system achieve better than 8-bit resolution. This means that the amplifier clearly achieves the maximum 8-bit resolution desired. In addition, the amplifier is currently not the weakest link in terms of bit-level resolution. The optoelectronic devices such as the VCSELs and photodetectors provide much lower bit-level resolution. Improvement of the resolution of these optoelectronic devices should be a major focus of continuing work. While the speed of operation was not a major consideration for the PSG2 amplifier, AC simulation was done to determine the frequency range of operation. The simulations shows that the amplifier has a 3dB cutoff point of greater than 100MHz.

Once the amplifier circuit was designed schematically and simulated, it was then laid out, extracted, and compared to schematic (LVS). The final amplifier structure was about 24μm×45μm. This is about 1/4 of the area used for the PSG1 amplifier. Furthermore, the Peregrine process has a larger minimum feature size (0.5μm) than

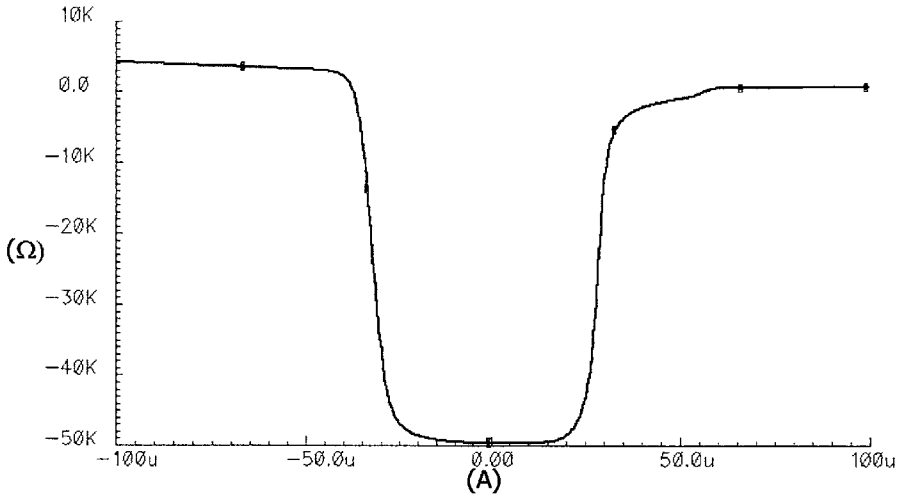

Figure 3-21: Derivative of the transfer function shown in Figure 3-20

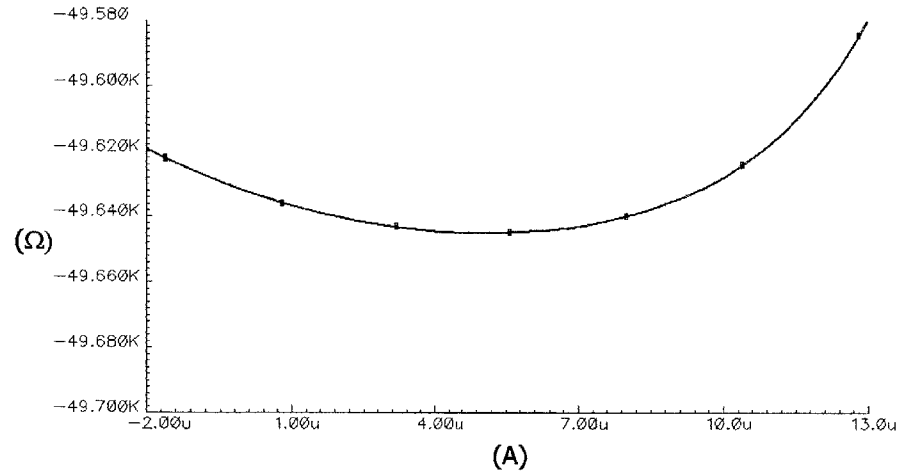

Figure 3-22: Close up of Derivative of the transfer function

the Vitesse HGaAs5 process (0.35μm). This reduction in size is a function of better topology design along with better layout. The layout for the PSG2 amplifier can be seen in Figure 3-23.

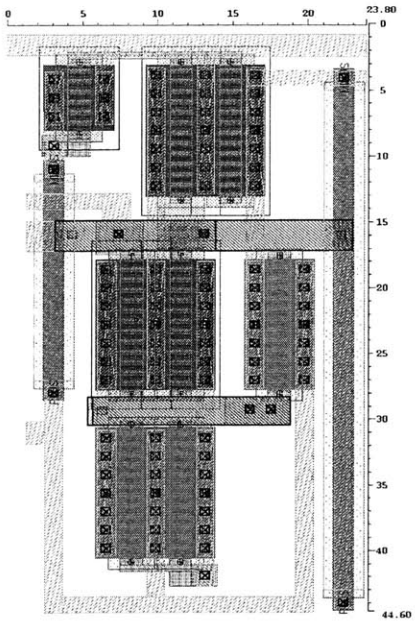

Figure 3-23: Layout of PSG2 transimpedance amplifier

3.3.2 PSG2 Photodetectors

In terms of photodetectors, the intention of the CONNPP using the Peregrine UTSi SOS process was to use aligned-pilar-bonding to mechanically bond GaAs pin photodetectors to the silicon-on-sapphire chip. While this was the eventual goal, it was also realized that it would be useful to produce photodetectors directly in the Peregrine UTSi SOS process for circuit and small system characterization. As a result, an investigation was conducted to determine the feasibility of producing photodetectors in the standard UTSi SOS process.

After researching the topic to understand the specific processing steps used by Peregrine, it became clear that building photodetectors in the standard process was somewhat problematic. The first issue relates to the fact that this is an Ultra-Thin Silicon process meaning that a very thin layer of silicon is deposited on top of a sapphire substrate. For this particular process, the thickness of the silicon layer is about 120nm. At 850nm, silicon has absorption coefficient of

$$\alpha \approx 800\text{cm}^{-1} \tag{3.2}$$

This means that the amount of light that passes through the silicon layer without being absorbed is

$$e^{-\alpha l} = e^{-800(120\times10^{-7})} = 99.04\% \tag{3.3}$$

In other words, for every 1mW of optical power that passes through the silicon layer, about 9.6μW of optical power is absorbed in the silicon layer. This fact by itself does not preclude the design and fabrication of photodetectors. Additionally, the fill factor (percentage of device are exposed to the light) is likely to be around 50% which will reduce the amount of light absorbed by another factor of two. For testing purposes, these are not showstopper issues as the intensity of the light can be increased until sensitivity is seen.

The main issue in building these photodetector devices lies in the process flow itself. Ideally, to build either metal-semiconductor-metal (MSM) photodetectors or lateral pin photodetectors, it is necessary to have a relatively low doped region between the higher doped contact regions. Additionally, this relatively low doped region should be exposed to the incident optical energy the user wishes to detect. An example of a lateral pin of this type can be seen in Figure 3-2.

Most unfortunately, the Peregrine UTSi SOS process does not allow lateral pin structures to be built. The reason for this can be found by examining the isolation structures used to electrically isolate NMOS devices from PMOS devices. The necessary structure for a lateral pin device consists of p+ and n+ regions separated by a low doped or intrinsic region. The problem building this structure is that Peregrine uses a thick field oxidation layer in between the n-well and p-well structures. This field-ox layer penetrates the entire 120nm of the silicon giving complete electrical isolation between the p and n regions. While this is ideal for transistors, it is problematic for optoelectronic devices. This field-ox layer makes it impossible to create a lateral pin structure because any region between a p region and n region is completely filled with field-ox.

There is an exception to this rule for field-ox isolation between p and n regions.

This exception applies to the diode structures in the UTSi SOS process. A drawn diode consists of p region directly abutted to an n region where the junction between the two regions is covered by gate material. By definition, the area under the gate structure remains at a lower doping than the defined p or n regions. While it seems possible that this structure could be used as a lateral pin photodiode, that does not seem to be the case. Photodetectors based on this diode structure were designed by Edward Barkley in 2000. Upon testing the structures, they did not show any optical response. It is postulated that this the lack of optical response is due to the structure of the gate. While the optical energy should be able to penetrate the polysilicon gate, many times a silicide is put on top of the gate structure to facilitate better electrical contact [34]. This silicide will potentially act as a reflector to incoming light. While it has not been able to be confirmed by Peregrine, it appears from the lack of optical response that there is a silicide layer on top of the gate.

In conjunction with Edward Barkley, new photodetector structures were designed in the UTSi SOS process. These devices are structurally much different and use a different principle of operation. In the course of research into the nuts and bolts of the Peregrine UTSi SOS process, it was discovered that there is a drawn layer called SDBLOCK. The function of this layer is to block the source/drain implant in specific areas. The area beneath the SDBLOCK should then have a much lower doping than the source and drain regions abutting it. As a result, this SDBLOCK layer was used to pattern a device that looks much like an NMOS transistor without a gate structure. This device can be seen in Figure 3-24.

The operation of this device is quite unique from anything discussed thus far. As it might be noticed, the regions shown have an npn structure similar to a bipolar junction transistor. There are separate connections to the n+ regions (emitter and collector), but there is no electrical connection to the p region (base). Generally in a BJT, a small current is injected into the base which causes a larger current to flow from the emitter to the collector. In this case, the current is being injected optically into the base as electron-hole pairs (EHPs). Because of the potential difference between the different types of materials along with the bias, the electrons from the EHPs will

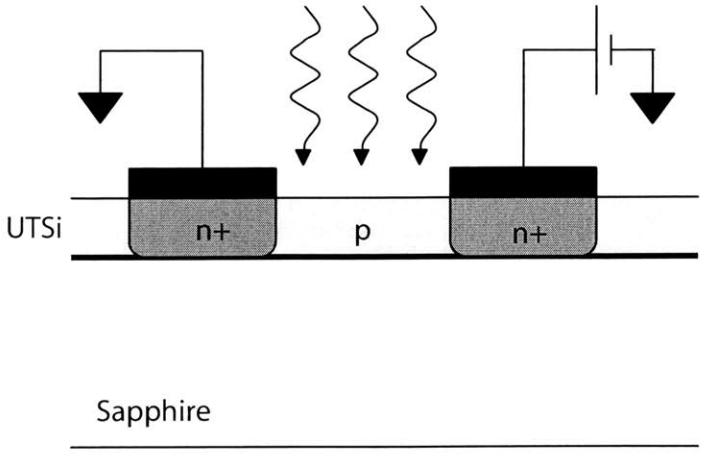

Figure 3-24: UTSi optoelectronic device

be collected by the positively biased terminal. This leaves a buildup of holes in the base region raising the potential of the base region. This phenomenon is referred to as "self-biasing" by Yamamoto et al [35]. Furthermore, it is expected that, becuase of its BJT structure, the photodetector should exhibit gain. This was reported in both [35, 36]. In the devices designed for the PSG2, these source and drain regions were interdigitated and each connected together to produce a single detector with a larger footprint. Two different structures with different source-to-drain spacings were laid out. One was laid out with minimum spacing (1.2μm) and one was laid out with double-minimum spacing (2.4μm). The layout for this detector structure can be seen in Figure 3-25.

3.4 Future Work

As mentioned at the beginning of this chapter, work is continuing on the PSG2 chip. Currently, the amplifiers and photodetectors discussed here have been designed, simulated, laid out, and verified. Some of the other structures slated to be included on the PSG2 chip include Aligned Pilar Bonding sites. These APB sites will allow work to begin on the integration of the optoelectronic devices onto the chip as outlined previously. In addition to the APB structures, work is also ongoing in the area

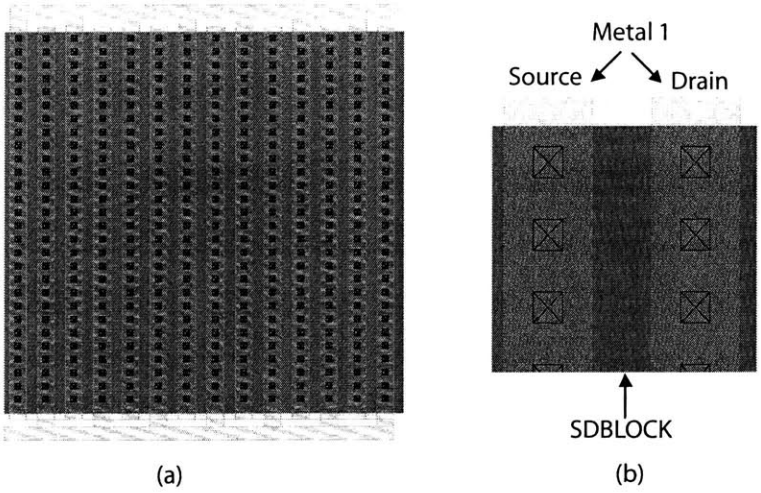

Figure 3-25: Lateral BJT photodetector layout

of circuit design. Currently, a successive approximation analog-to-digital converter (ADC) is being design, simulated, laid out, and verified by Travis Simpkins of the Photonic Systems Group. As this ADC also inherently contains a digital-to-analog converter (DAC), this circuit will provide the facilities to convert data within the chip and the architecture between digital and analog signals. If the CONNP is going to eventually be communicating with a PC, this is an important piece of infrastructure to have regardless of the neural network architecture chosen. As design on the PSG2 chip wraps up, work will begin on the design of the PSG3 chip. This chip will most likely start to implement some of the neural network structure discussed throughout the body of this document including the architecture discussed in Chapter 4. This chip will hopefully have a small array of neurons/pixels that will allow a small-scale proof-of-concept system to be built. This chip will provide learning on some of the key alignment and packaging issues faced by the CONNPP.

Chapter 4

Neural Network Models and Architectures

Over the past twenty years, neural networks have been built in software running on general purpose digital computers, dedicated analog hardware, and dedicated digital hardware. The Compact Optoelectronic Neural Network Processor actually represents a hybrid of all three of these implementations. The current suggested CONNPP architecture will be explained along with a proposal for a major architectural change to the original design. Models will be built to demonstrate the performance differences between the current architecture and the proposed architecture.

4.1 Specification For the CONNPP Hardware Implementation

While it is possible to model large neural networks with a general purpose computer, these implementations do not always produce the fastest results as the hardware is not optimized for the problem. Specialized analog hardware has also been used to replicate the functionality of a neural network. These analog components generally require the use of precise resistive and capacitive elements to accomplish their task. This poses a potential problem as semiconductor processing techniques generally have

difficulty producing resistive and capacitive elements with high precision. With the success and availability of CMOS technology, many researchers have started to look at specialized digital hardware to implement neural network functionality. The CONNPP architecture draws from all three of these areas to create a processor that utilizes the available technology in a very efficient manner.

A block-level diagram of a single CONNPP neuron can be seen in Figure 4-1. These neurons are then tiled in a 2-dimensional array on a VLSI chip. As described in Chapter 1, the detector element and VCSEL array are fabricated such that optical input is accepted on one side of the chip and optical output is produced on the other side of the chip. Because of this novel fabrication technique, the chips can then be cascaded in the third dimension. In conjunction with a holographic element array, the neurons on one chip can communicate optically with the neurons on the next chip.

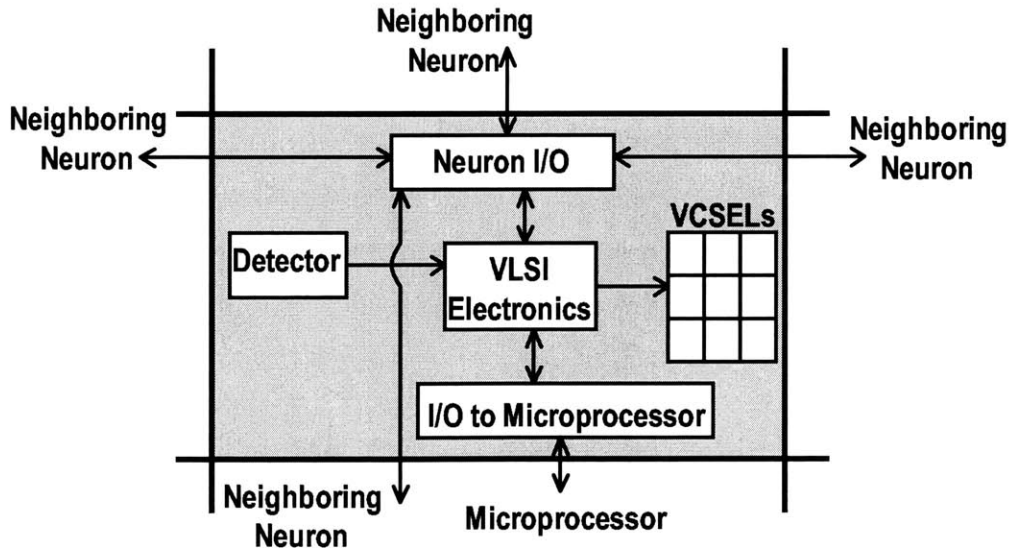

Figure 4-1: Block-level diagram of single neuron in the CONNP

As the system consists of a series of identical elements that have been replicated in 3-dimensions, the workings of the system as a whole can be understood by examining the operation of the single neuron shown in Figure 4-1. The operation of the neuron begins with the detection and conversion of an analog optical signal to a digital

electrical signal by the detector unit (Figure 4-2). The optical signal received comes from the nine nearest-neighbor neurons in the previous layer. As the photodetector senses intensity and the nine beams from the previous layer are incoherent, this detector is basically a summing unit within the neuron. The photodetector converts the light incident on the photodetector to an electrical current. The amplifier circuits discussed in Chapter 3 convert this electrical current into a voltage at a level suitable for analog-to-digital conversion to a digital signal. Finally, this analog voltage is converted into a digital signal by the analog-to-digital converter. This digital signal is then passed on to the VLSI Electronics block.

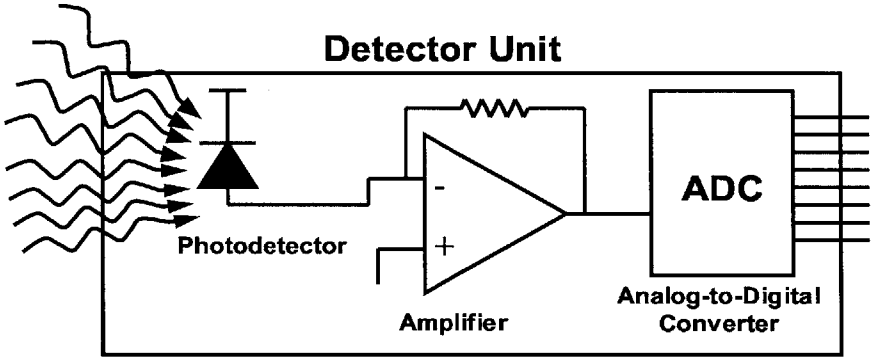

Figure 4-2: The detector unit of a single neuron

The VLSI Electronics block takes the digital signal from the detector unit and passes it along to the block labeled "I/O to Neighbors." As mentioned in previous chapters, the processor has both in-plane connections and between-plane connections. The between-plane connections are the optical interconnects that allow the planes of neurons to communicate with each other. The in-plane connections allow neurons within a single 2-dimensional array to communicate. In this original CONNPP architecture, each neuron is able to communicate with its four nearest-neighbor neurons within the plane of the chip. The "I/O to Neighbors" block sends the digital signal representing the optical input to the four nearest-neighbor neurons. Likewise, this block also receives a digital input from the four nearest-neighbor neurons representing their optical inputs. These signals then are passed back to the VLSI Electronics block.

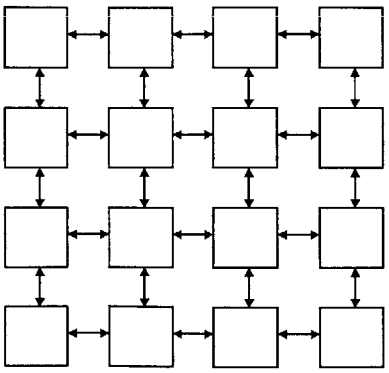

Figure 4-3: 2-dimensional array of neurons with 4-way in-plane nearest-neighbor connections

The VLSI Electronics block proceeds to sum the digital signal from the optical input along with the digital signals from the four nearest-neighbor neurons. This digital sum is then used as the input to the activation function of the neuron. The output of the activation function becomes the output of the VLSI Electronics block and the input to the VCSEL Array block. Each VCSEL in the array connects to a different neuron in the next 2-dimensional array of neurons. Each VCSEL has a current applied which is proportional to the output of the activation function multiplied by the weight between the two neurons. This means that each VCSEL in the array is potentially modulated with a different current. By doing this, the connection weight information is explicitly encoded in the optical signal incident on the photodetector. This allows the optical signals incident on the photodetector to simply be summed as the weighting has already been done.

It may be noticed that the "I/O to Microprocessor" block has not been involved in the operation of the neuron to this point. This block is utilized mainly in the training and programming of the neural network. The CONNP does not have the facilities or infrastructure on-chip to evaluate and recalculate weights. For this task, the neural network processor relies on the host PC running software that calculates the new weights and sends them to back the neural network processor using the "I/O to Microprocessor" block. One advantage of this methodology is that the training algorithm is not implemented in hardware. This means that a variety of different

training algorithms can be used to give the neural network processor greater flexibility.

The other main use of the "I/O to Microprocessor" block is to program the CONNP. Training the network to a small mean squared error is potentially a very time-consuming task. Within the neural network processor, the state of the processor is basically captured by the connection weights between neurons. If after the training session, the weights for all the neurons could be stored on the host PC, they could then be recalled later in order to restore the state. Having the facilities to stream weights in and out of the processor allows this functionality. The neural network can be extensively trained for a particular task, store the weights on the PC, and then later request this list of weights from PC when a similar problem is encountered.

4.2 Proposal For a CONNP With a Global Bus-based Architecture

As discussed in the previous section, there are two different levels of connections within the CONNP. There are in-plane electrical connections between neurons and there are between-plane optoelectronic connections between neurons. The specification described allowed for each neuron to communicate electrically with four neurons within the plane of the chip and nine neurons in the next plane. The optoelectronic connections, as described, clearly allow the processor to achieve 3-dimensional scalability. This 3-dimensional scalability allows the processor to increase its computational power simply by increasing the number of cascaded layers. The in-plane connections, however, seek to increase the computational power of each layer within the neural network. Thus, if the in-plane connection model could be significantly improved, every layer of the cascaded network would benefit.

The between-plane optoelectronic connections within the processor, as specified previously, push the boundaries of the current state-of-the-art technology. It would be desirable if the in-plane electrical connections also pushed these boundaries. While the optoelectronic connections achieve this through the use of novel processing tech-

niques and holographic interconnects, the in-plane electrical connections must accomplish this through novel architectures and circuit design. We propose using a digital bus-based connection model within the plane of the chip to achieve global in-plane interconnects.

Figure 4-4 shows a block-level diagram of a neuron in the new proposed bus-based architecture. As can be seen in the figure, the "I/O to Microprocessor" and "Neuron I/O" blocks have been consolidated into a single block. Previously the "Neuron I/O" block consisted of connections to the four nearest-neighbor neurons within the plane of the chip. This meant that separate infrastructure was potentially needed to communicate with the in-plane neurons and with the microprocessor. In the previous design, a chip-level bus was still required. This chip-level bus connected to every neuron on the chip and allowed the neurons to send and receive weight information during training or programming. In this revised architecture, this chip-level bus is also utilized to allow every neuron on the chip to communicate with every other neuron on the chip.

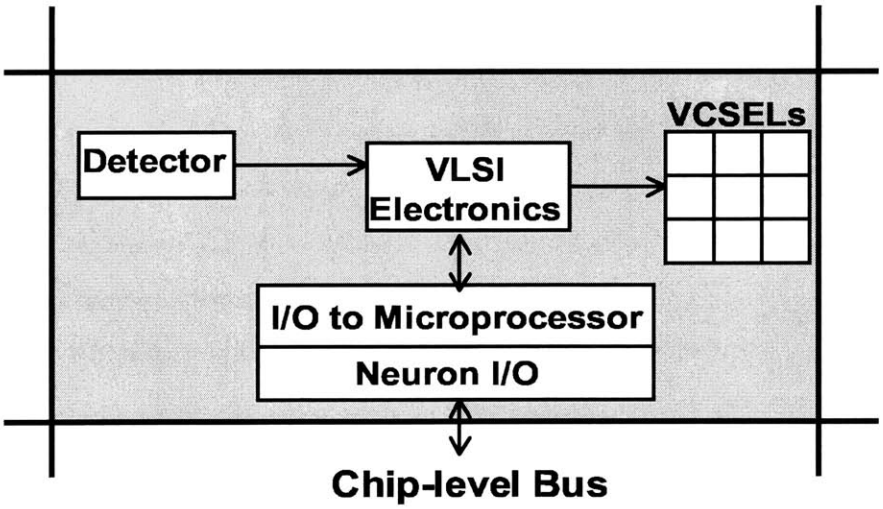

Figure 4-4: Block level diagram of new proposed neuron structure

The operation of the global bus-based architecture is actually very similar to the original CONNP design. First, the optical inputs are received by the detector block of the neuron and converted to a digital signal. This digital signal is then passed

along to the "VLSI Electronics" block. This block holds on to the signal and also forwards it to the I/O block.

This leads to a discussion on the operation of the chip-level bus. In addition to the I/O block within each neuron, there is also a chip-level I/O block that oversees the chip-level bus. This block is responsible for controlling all of the traffic on the chip-level bus. Once the optical inputs have been received by the neurons, the bus control logic takes over. As discussed, the optical input to each neuron was detected and converted to a digital signal. The bus allots each neuron a time slice to broadcast the digital data it received as an optical input. When the neuron is not broadcasting its digital data, it is receiving digital data from other neurons.

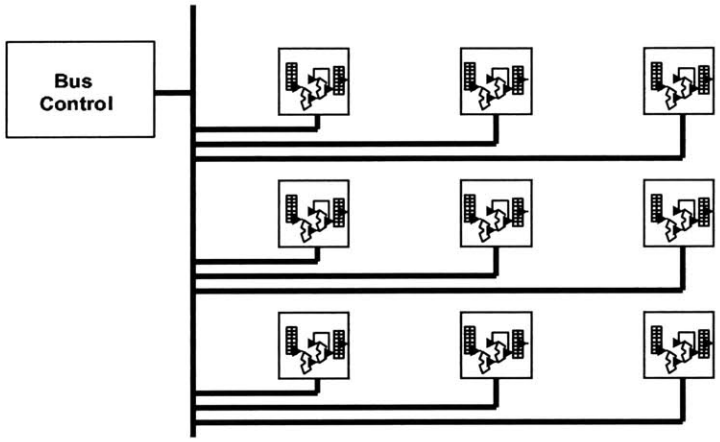

Figure 4-5: Bus based architecture with bus control block and individual neurons

As an example, consider a very small on-chip neural network consisting of four neurons. Each neuron has an address assigned to it ranging from 1 to 4. The bus control logic directs the neurons to broadcast their data in order. During the first clock cycle, neuron 1 broadcasts its data and neurons 2-4 receive the data from neuron 1. In the second cycle, neuron 2 broadcasts its data while neurons 1, 3, and 4 receive the data. This continues until all the neurons have broadcast their data. This scheme allows n neurons to broadcast their data in n cycles.

This clearly leads to a tradeoff between connections and time. In the original implementation of the CONNP, each neuron could communicate with its four nearest-

neighbors in 4 clock cycles (assuming the I/O block can only talk to one neuron at a time). This means that there is a direct tradeoff between number of in-plane connections and number of clock cycles required to complete a single operation. This issue will be examined in more detail in the next section.

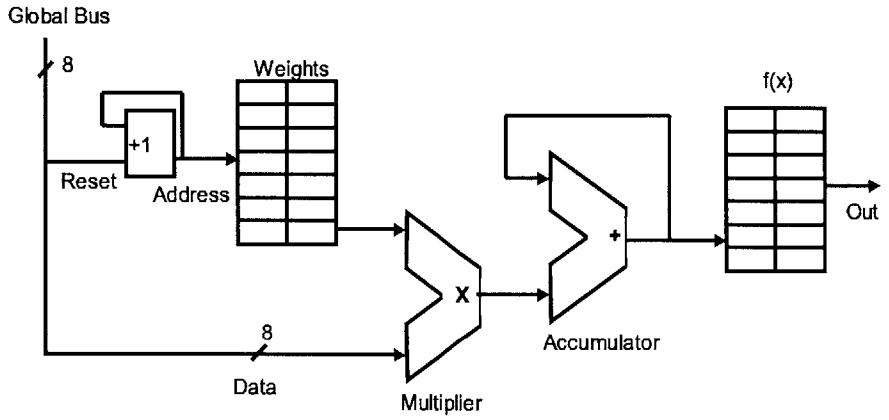

Figure 4-6: Simplified block-level diagram of logic within neuron

Figure 4-6 shows a block-level diagram of the inner workings of the logic within a bus-based neuron. When the broadcast is set to begin, the bus control logic sends a reset signal to the counter (the box labelled "+1"). This counter contains the address of the neuron currently talking on the bus. This address is used to index into the table of weights. The weight retrieved becomes the first operand for the multiplier. The data is received from the neuron currently in control of the bus becomes the second operand for the multiplier. The operands presented to the multiplier are a weight and a data value. This represents one of the $w_i x_i$ terms in the sum

$$\sum_{i=1}^{n} w_i x_i \tag{4.1}$$

The result from the multiplier ($w_i x_i$) becomes the inputs into the adder which is configured with feedback as an accumulator. This accumulator adds each of the $w_i x_i$ elements to a running sum. After all the neurons have broadcast their data, the result of the accumulator is the expression given in Equation 4.1. This expression is then used as an index into a lookup table of activation function $f(x)$. The number of

entries in this table represents the resolution of the activation function. This value,

$$f\left(\sum_{i=1}^{n} w_i x_i\right) \tag{4.2}$$

is the actual output of the neuron. However, instead of firing the VCSELs in proportion to the output of this function, it would be desirable to also encode the weights for those connections. This means that this output value should be multiplied by each of the nine between-plane nearest-neighbor weights to produce nine separate outputs. These nine outputs will then be used to determine the proper current level at which to drive each of the nine VCSELs in the array.

4.3 Neural Network Models and Simulations

While several different architectural models for the CONNPP have been proposed and described, it is only feasible to actually implement one of these architectures in hardware. As a result, models and simulations allow the different proposed architectures to he evaluated according to a particular set of metrics. The purpose of this section is to develop these models and simulations in order to evaluate these different architectures.

The first task in building a model is to determine the best development environment in which to build the model. Many different choices were examined and evaluated for this task. Among the choices were the Stuttgart Neural Network Simulator (SNNS), MATLAB, several Java neural network packages, and writing the models in C or C++ from scratch. The SNNS package seemed to be potentially powerful for large scale neural network simulations, but proved difficult and time-consuming to use from an interface point of view. Building the models and simulations from scratch using C was ruled out as an option because it seemed somewhat foolish to invest a large amount of time in reproducing work that had already been done using commercial products. After examining and evaluating all the available options, MATLAB and its Neural Network Toolbox were chosen as the development environment. The Neural Network Toolbox's ease of use coupled with the power of the underlying

MATLAB environment made it the logical choice.

4.3.1 Background on MATLAB Neural Network Toolbox

Figure 4-7 shows the MATLAB Neural Network Toolbox's internal representation of a single neuron element. In the neuron shown, there are R inputs into the neuron. These R inputs consist of an input value (p_i) and a connection weight ($w_{1,i}$).

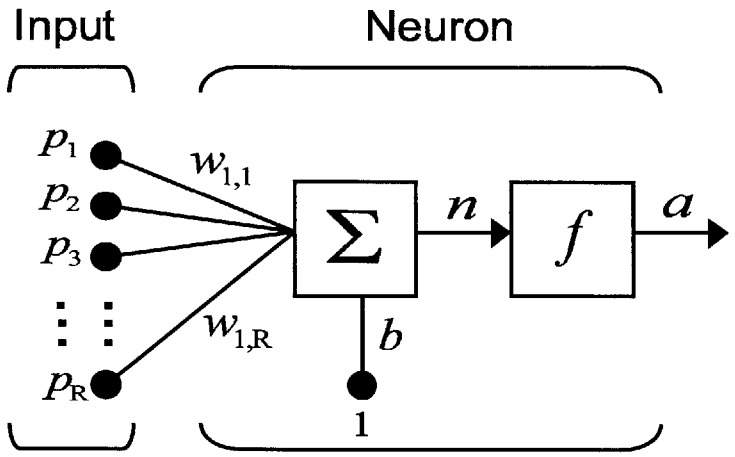

Figure 4-7: MATLAB model of a neuron

The input values are stored in an $R \times 1$ vector $\mathbf{p}$ and the connection weights are stored in a $1 \times R$ single row matrix $\mathbf{W}$. In the summing unit (labelled Σ) each of the input values is multiplied by its corresponding weight element and they are all added together. In addition, a constant input of 1 is multiplied by the weight (b). This provides a dynamic bias for the neuron as the bias (b) can be adjusted through training. The sum of all the input and the bias is labelled as n in the diagram. This value is then fed into the activation function unit of the neuron (labelled f). The activation function unit uses n as the argument for function f and produces the value a. This means that the output of the neuron can be written as

$$a = f(\mathbf{W} \cdot \mathbf{p} + b) \tag{4.3}$$

where $\mathbf{W} \cdot \mathbf{p}$ is the dot product of the single row matrix $\mathbf{W}$ and the vector $\mathbf{p}$.

The neural network toolbox provides a set of activation functions that includes linear functions, logsig, tansig, and the Heaviside function just to name a few.

Once the single neuron structure in MATLAB is understood, it can be extended to create entire layers of neurons. A layer of neurons accepts inputs from a number of sources just like a single neuron. Each neuron in the layer processes the information presented to its inputs and each neuron in the layer then produces its own output. A diagram of this can be seen in Figure 4-8.

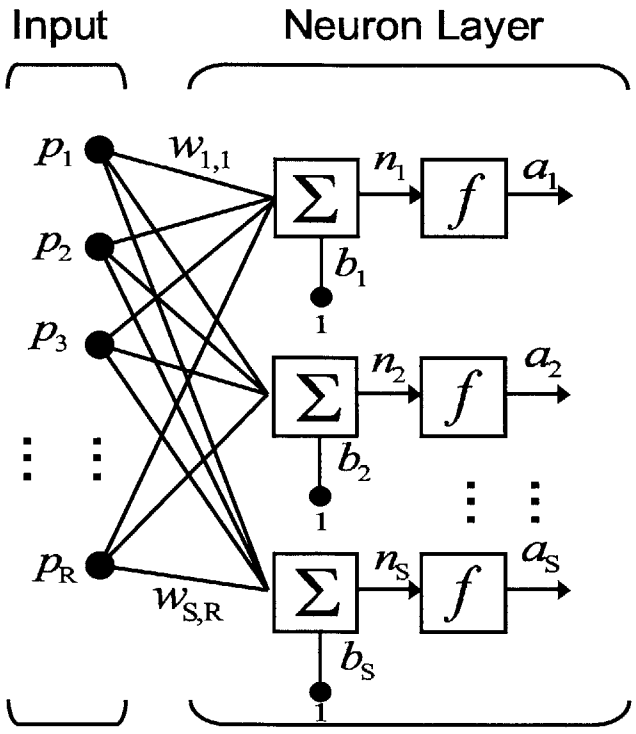

Figure 4-8: MATLAB model of a layer of neurons

The data structures within the layer of neurons changes somewhat from the single neuron model. Assuming there are R inputs to the layer, the input values remain in the form of a $R \times 1$ vector. The connection weights change their form significantly due to the fact that MATLAB assumes that every input has a connection to every neuron in the layer. This is a problematic assumption for the models built to represent the CONNPP. The solution to this problem will be addressed as the CONNPP model is developed. Whereas the weights for a single neuron were in a single row matrix with

dimensions $1 \times R$, for a complete layer of neurons, the weights comprise a matrix with dimensions $S \times R$. Each row of the matrix represents a list of weights for a particular neuron in the layer. Each of the neurons has its own summing unit that it uses to add each of the input values multiplied by the corresponding weight. This produces an $s \times 1$ vector n where n_i represents the output of the summing unit of the i^{th} neuron. The values in this vector are then input into S activation function units to produce an $S \times 1$ vector a representing the output of the layer of neurons. Each value in the vector a corresponds to the output of an individual neuron within the layer. This means that the functionality of the layer of the neurons can be written as

$$\mathbf{a} = \mathbf{f}(\mathbf{W} \cdot \mathbf{p} + \mathbf{b}) \tag{4.4}$$

where all the elements within the equation are matrices. MATLAB simplifier the layer representation shown in Figure 4-8 to reflect the fact that the neuron layer can be represented as matrices and matrix operations. This simplified layer structure can be seen in Figure 4-9

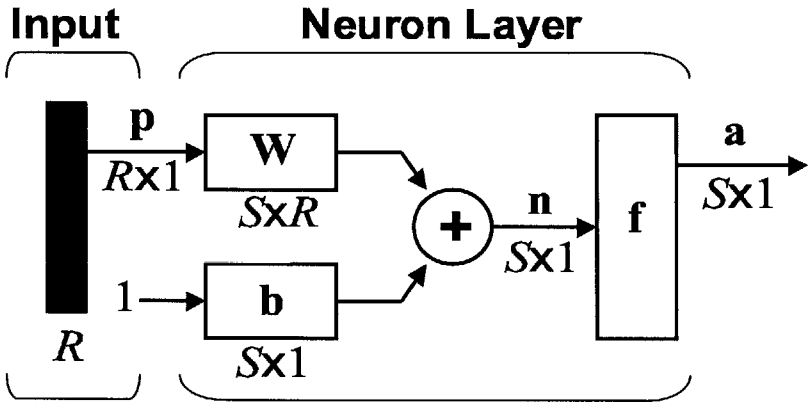

Figure 4-9: Simplifier MATLAB model of a layer of neurons reflecting matrix properties

All of the neurons within a given layer in MATLAB use the same activation function (specified by the user). These layers can then be cascaded such that the outputs of layer i become the inputs to layer $i + 1$. Each layer may have a different activation function if desired. Furthermore, the layers need not contain identical

numbers of neurons.

4.3.2 Model of the CONNPP hardware using MATLAB

To model different proposed hardware implementations, it is necessary to have a good understanding of the operation of the hardware along with the facilities available within MATLAB. From this knowledge, higher-level models can be built up from the low level structures provided by the Neural Network Toolbox.

As discussed previously, the MATLAB Neural Network Toolbox assumes that all layers are globally interconnected. This is far from the case with the CONNPP. In fact, there are many situations where the layers are sparsely connected. For this reason, it is important to develop a method to replicate these various sparsely connected architectures. Recall from the explanation of the Neural Network Toolbox structures that the connection weights between layers are stored as $S \times R$ matrices where S is the number of neurons in the i^{th} layer and R is the number of neurons in the $(i+1)^{th}$ layer. If this is the case, then a set of matrix masks can be created to zero the elements without any connection in the network.

1	2	3	4	5	6
7	8	9	10	11	12
13	14	15	16	17	18
19	20	21	22	23	24
25	26	27	28	29	30
31	32	33	34	35	36

Figure 4-10: A 6×6 layer of neurons with each element numbered

As can be recalled from previous discussions about the CONNPP, the neurons are laid out in 2-dimensional planes. Figure 4-10 shows a layer of 36 neurons arranged

in a 6×6 array with each neuron numbered. It is known that the optoelectronic connections between planes use a 9-way nearest neighbor connections scheme. As an example this means that neuron 8 in layer i would connect to neurons 1, 2, 3, 7, 8, 9, 13, 14, and 15 in layer $(i+1)$.

It can also be recalled that internal to MATLAB, this 36 neuron layer is stored as a 36×1 vector. Because there are 36 inputs from layer i to layer $(i+1)$ along with the fact that layer $(i+1)$ also has 36 inputs, leads to the connection weight matrix between layer i and layer $(i+1)$ having dimensions 36×36. Each row of this 36×36 connection weight matrix represents the weights for a particular neuron in layer $(i+1)$. In this example, row 8 of the connection weight matrix will be examined. In the first of replicating the 9-way nearest neighbor connection scheme a second 36×36 matrix is created. This will be the mask for the connection weight matrix. In order to achieve the 9-way nearest neighbor interconnect scheme between planes, elements 1, 2, 3, 7, 8, 9, 13, 14, and 15 of row 8 of the mask matrix are set to a value of 1. All other elements in row 8 of the mask matrix are set to 0. This process is repeated for all of the 36 neurons and all 36 of the rows of mask matrix. Once the complete mask has been created, an element-wise multiplication is done using the mask matrix and the connection weight matrix. The resulting matrix will be the new connection weight matrix. Any element in the matrix whose corresponding mask matrix value was 1 will remain unchanged. Any element whose corresponding mask matrix value was 0 will also be 0.

It is important to note that these mask matrices must be created and applied between any set of layers where global interconnects are not desired. Furthermore, these masks must be applied after every training cycle. When MATLAB trains the neural network, it calculates a complete new set of weights optimized to minimize the error on the next presentation of a pattern. This potentially changes elements in the connection weight matrix from their desired value of 0 when there is no connection. As a result, it is very important to zero these elements after each training cycle so as not to get misleading results.

Now that it is understood how to implement different connection models between

neural network layers, a model of the CONNPP hardware can be developed. One of the key characteristics of the CONNPP model is that in a single hardware layer, inputs are received optoelectronically, converted to digital format, and than distributed digitally among the neurons in the layer. The converted optoelectronic inputs along with the digital inputs from the other neurons are then used as the input to the activation function of the neuron. In order to model this single hardware layer, two MATLAB neural network layers are required to reproduce the functionality of the single hardware layer. A diagram of the MATLAB single hardware layer model can be seen in Figure 4-11.

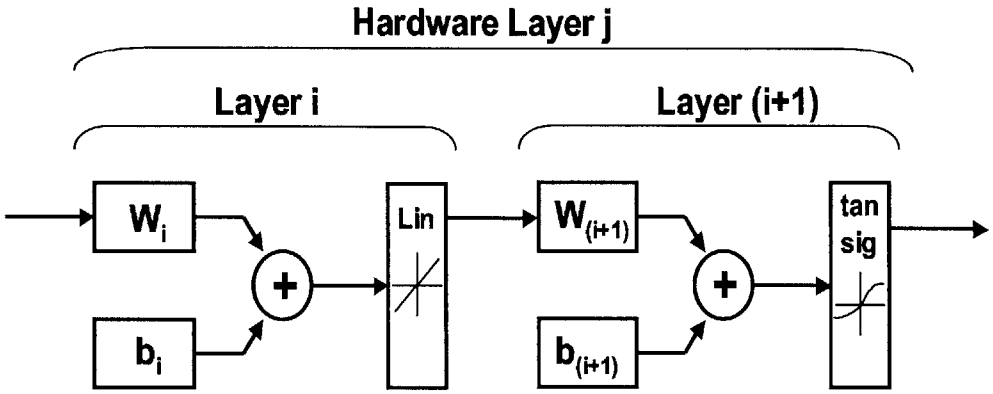

Figure 4-11: MATLAB model of a single CONNPP hardware layer

As can be seen in the figure two layers of neurons are used to replicate the functionality of a single CONNPP hardware layer. Layer i uses the Purelin function as its activation function while layer $(i+1)$ uses a tansig function as its activation function. The purelin function is a linear function with a slope of 1. This means that the sum presented to the activation function of a neuron in layer i will also be the output of the activation function to layer $(i+1)$. W_i is the connection weight matrix between hardware layers $(j-1)$ and j. This represents the optoelectronic connections between the layers of the processor. As such, this first connection weight matrix of the hardware model will always be masked so as to provide 9-way nearest neighbor connections between the hardware layers. For each neuron in layer i, the nine nearest neighbor inputs are multiplied times their corresponding weights, added

together, and then presented as outputs to the next layer. This is analogous to the detector unit of the hardware implementation which optoelectronically sums the nine optical inputs and passes them along to the VLSI Electronics block of the neuron.

The connection weight matrix $W_{(i+1)}$ represents the connections between the neurons within the plane of the chip. This connection weight matrix is basically analogous to the Neuron I/O block in the hardware implementation of the CONNPP neuron. As the goal of simulations conducted here is to evaluate different interconnection models for neurons within the plane of the chip (i.e. 4-way nearest neighbor vs. global), this connection model will be one of the key variables in the simulations. In order to implement different connection models, different mask matrices are simply applied to the connection weight matrix $W_{(i+1)}$ after each training session.

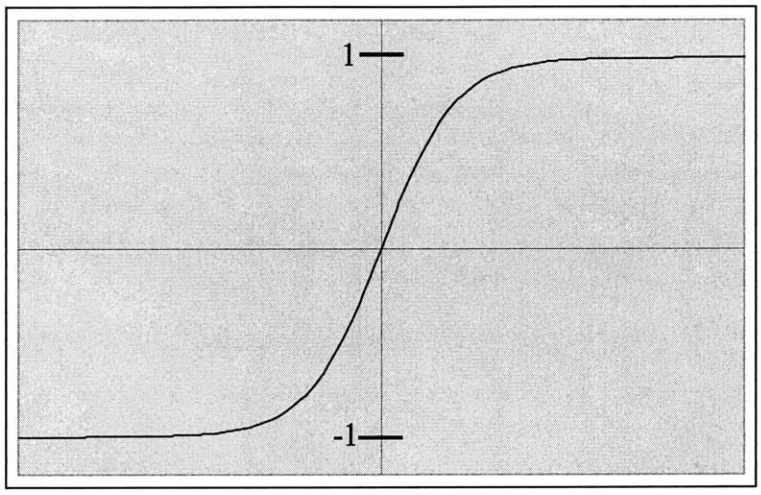

Figure 4-12: Tansig function

Next, the outputs from layer i are multiplied by their corresponding weights from connection weight matrix $W_{(i+1)}$ which distributes the inputs to layer $(i+1)$ according to the connection model being tested. The inputs to each neuron in layer $(i+1)$ are then summed. Finally, this sum is used as the input to the tansig activation function of layer $(i+1)$. This activation function has the form

$$\text{tansig}(n) = \frac{2}{(1+e^{-2n})} - 1 \tag{4.5}$$

The graph of the function can be seen in Figure 4-12. The output of this layer is a vector containing the number of elements equal to the number of neurons in the layer. This represents the output of the hardware layer. In a hardware implementation, this will be the optoelectronic signals from the VCSEL array. This output will act as the input to the next hardware layer. Next, the simulations conducted and the results of those simulations will be discussed.

4.3.3 Simulations and Results

In the last section the models were developed to replicate the functionality of the CONNPP hardware using the MATLAB Neural Network Toolbox. This section will take these models one step further by providing implementation details, simulation conditions, and results of those simulations.

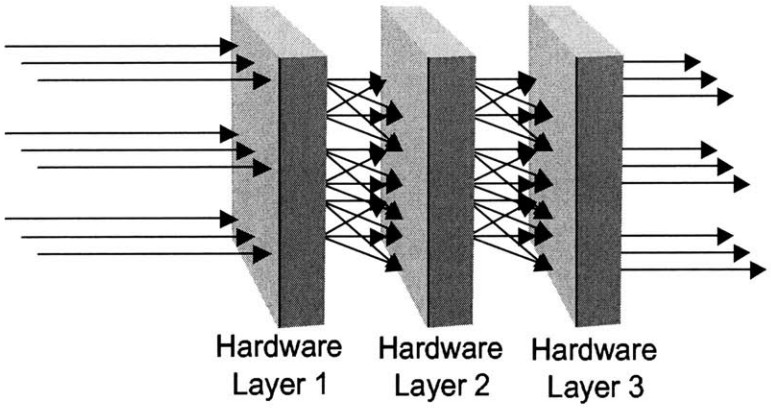

Figure 4-13: Example of the hardware system to be simulated in MATLAB

The goal of the simulations was to provide a quantitative understanding of how different in-plane connection models between neurons affect the computational power and training time of the neural network as a whole. Simulation parameters were chosen such that they both represented potential real world implementations as well as being computationally tractable in the time available. As a result, a simulation was constructed to replicate a CONNP with three hardware layers (shown in Figure 4-13. As can be seen in the figure, the input is presented to Hardware Layer 1. Hardware

Layer 1 and Hardware Layer 2 are connected via the optoelectronic interconnects. Likewise Hardware Layer 2 and Hardware Layer 3 are also connected optoelectronically. The output is then taken from Hardware Layer 3. All three layers have in-plane connections. The in-plane connection model is the key parameter of interest in these simulations.

Four different in-plane connection models were tested. The first in-plane connection model (Referred to as NONE) does not allow the neurons to communicate at all within the plane of the chip. The next connection model simulated was the 4-way nearest neighbor interconnects from the original. As another data point, the nine-way nearest neighbor connection scheme used for the optoelectronic connections between the planes was also tested for in-plane connections. Finally, a global in-plane interconnection model was tested. As mentioned previously, the activation function used for the neurons was a tansig function described in the last section. This is an activation function commonly used for neural networks. The network is a feedforward neural network that implements a backpropagation learning algorithm. The feedforward aspect of the network basically means that the inputs are fed into one layer of the network, propagate forward through the various hidden layers, and an output is produced at the final layer. This means that while a neuron from layer i can talk to a neuron from layer $(i+1)$, a neuron from layer $(i+1)$ cannot talk to a neuron in layer i. The backpropagation property means that in the training of the neural network, the error between the actual output and desired output is fed back through the network so that the weights between the hidden layers are adjusted so as to minimize the error on the next presentation of that pattern.

The goal of the first simulation performed was to look at the effect of the in-plane connection model on training time for a particular neural network. In addition to varying the in-plane connection models, the size of the networks was also varied. Two different sized neural networks were examined. The first neural network had 36 neurons per layer which corresponds to a CONNP with a 6×6 2-dimensional array of neurons in each hardware layer. The second neural network had 144 neurons which represents a CONNP with a 12×12 2-dimensional array of neurons in each hardware

layer. Therefore, it was desirable to determine how the in-plane connection model affects the training time for the two different sized neural networks.

In order to determine this, a series of training patterns needed to be chosen. Each training pattern consists of an input to the neural network and a desired output. Six different training patterns were chosen. As the input to the CONNP is in a two dimensional array, dot-matrix characters were chosen as the patterns. The desired output for the network was to classify the which letter was being presented. Therefore, if the letter 'A' was being presented at the input to the neural network, the desired output would have the output of neuron 1 (from Figure 4-10) produce a maximum output with all the other neurons producing a minimum output. Because two different size networks were being tested (6×6 and 12×12), the training patterns were simply scaled. Figure 4-14 shows the six different training patterns and their corresponding desired outputs.

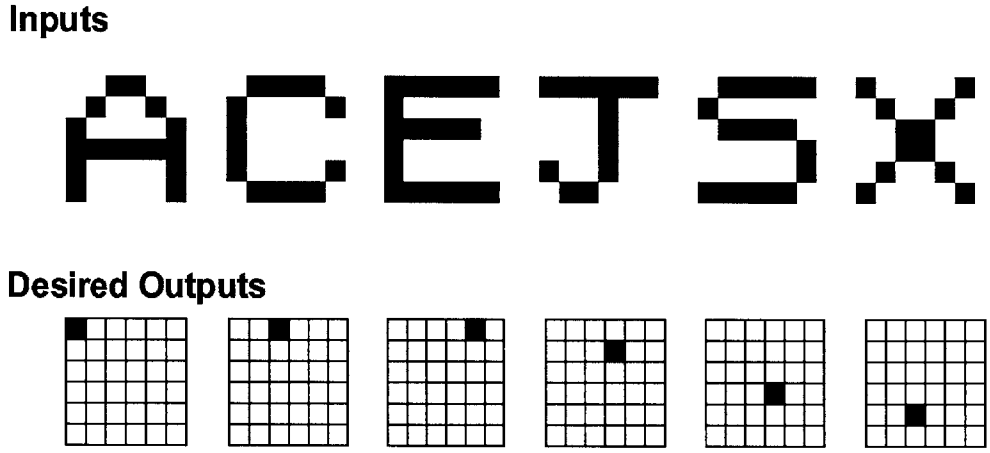

Figure 4-14: Training patterns for the in-plane connection model simulation

The neural network was trained in batch mode. This means that all six of the training patterns were presented to the neural network before the errors were calculated and the weights changed. This training method, in many cases, produces faster convergence of the weights. After each training cycle, a mean squared error (MSE) is calculated by the Neural Network Toolbox. The first set of simulations used these six training patterns for the different networks being tested to find out how many

training cycles it took to produce a MSE of less than 1×10^{-4}. Because the weights are initially random, the number of cycles required to train the neural network to a given level of MSE varies. For this reason, each simulation was run six times so that the average and standard deviation could be calculated. The results of this simulation can be seen in in Figure 4-15.

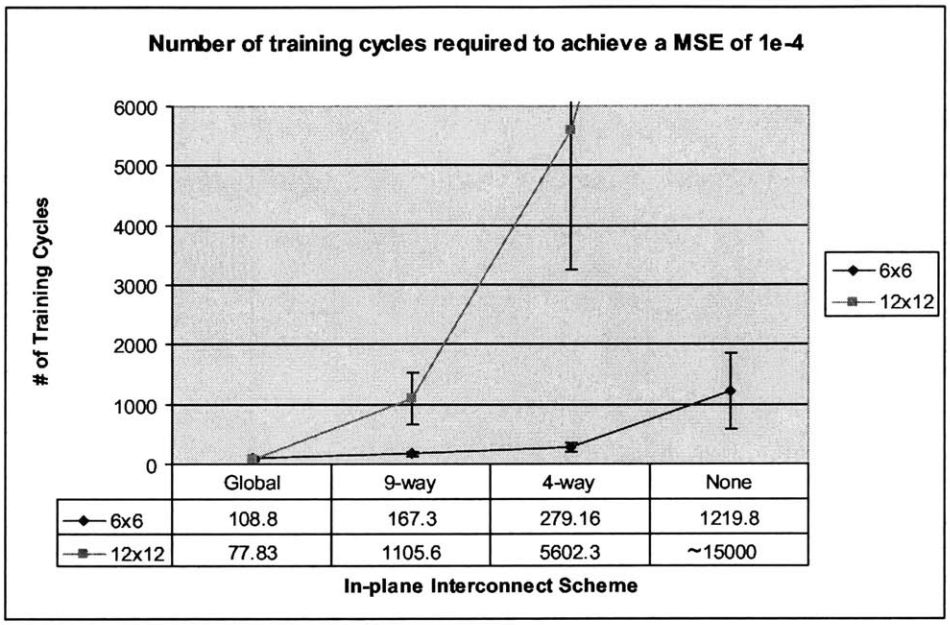

Figure 4-15: Results for training simulation using six training patterns

As would be expected, as the number of in-plane connections decreases, the number of training cycles required to reach a MSE of 1×10^{-4} increases significantly. The original simulations were done only on the 6×6 network. It was then considered that for a 6×6 network, the sparse in-plane interconnect models (4-way and 9-way) were actually reasonably powerful. Consider the fact that for 9-way in-plane interconnects for a 6×6 network, each neuron can talk to 25% of the neurons in-plane. Even the 4-way in-plane interconnects allow for 11% connectivity within the plane. For this reason, the 12×12 neural network was simulated. Using the 12×12 network, a 9-way in-plane connection model corresponds to 6% connectivity and a 4-way in-plane connection model corresponds to a 3% connectivity. As can be seen in the results of the simulations as this connectivity level decreases, the required training time increases.

This is reflected by both varying the connection model and by varying the size of the network.

Focusing on the two proposed connection models (global and 4-way), it can be seen that for the 12×12 network that the neural network using the global in-plane connections trains about 72 times faster than the neural network using the 4-way in-plane connections. At first glance, these results seem very encouraging for the proposed global in-plane connection model. However, it must be considered that the 4-way connection model requires 4 clock cycles to complete the operation while the global connection model required 144 clock cycles to distribute the data globally. This reduces the time advantage of the global connection model from 72 times faster to 2 times faster. At this level, it is not clear that the global connection model offers enough benefits to justify potential increases in real estate and complexity.

Another important metric to consider for the different in-plane connection models is the number of patterns each of the different networks can store. The training patterns used for this simulation consisted of a complete dot-matrix alphabet similar to the patterns shown in Figure 4-14. Likewise, the desired output patterns also consisted of a single element with a maximum value and a the rest of the elements with a minimum value. For this set of simulations, only the 12×12 network was examined. For a given in-plane connection model, each network was trained with an increasing number of training patterns. It was then recorded how many training cycles were required to reach a MSE of 1×10^{-4} for that number of patterns. As an example, when a batch of 3 patterns was presented to the global in-plane connected network, it required about 80 training cycles to train the neural network to a MSE of 1×10^{-4}. However, for the 4-way in-plane connected network, it took 1107 cycles to store 3 patterns to a MSE of 1×10^{-4}. The complete results can be seen in Figure 4-16.

If a cutoff of 5000 training cycles is used, it is shown that while both 9-way and 4-way in-plane connections fail at storing 9 patterns, the neural network using global in-plane connections has absolutely no trouble storing all 26 letters. Also of note is that the 9-way and 4-way connection models perform very similarly. In addition, the connection model with no in-plane connections is only able to store 2 patterns

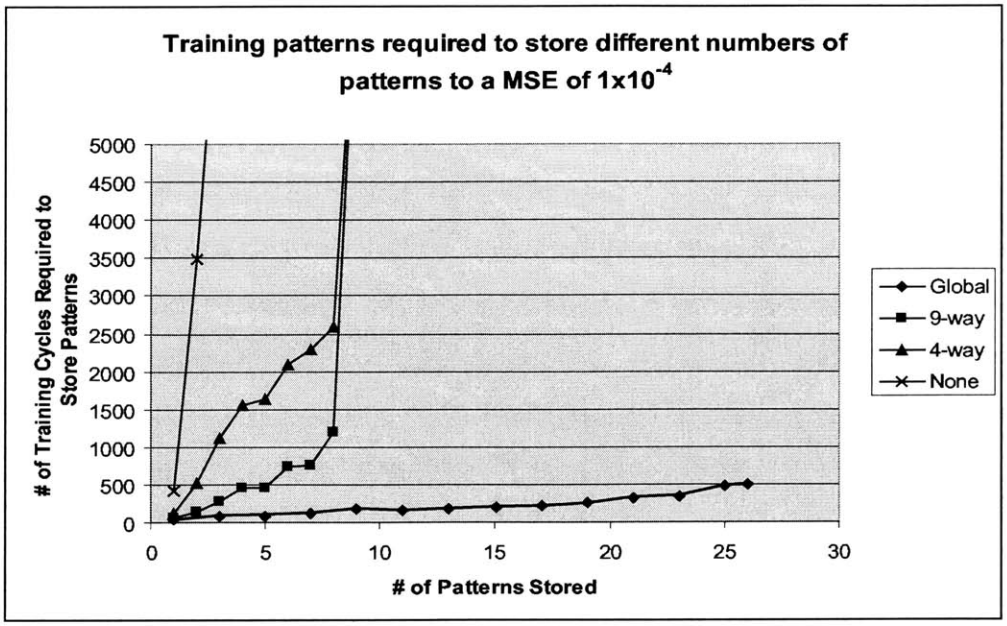

Figure 4-16: Training cycles required to store different numbers of patterns to a MSE of 1×10^{-4}

before it reaches the 5000 training cycle threshold. While it is difficult to quantify "computational power" for neural networks, the fact that the neural network using global in-plane connections is able to store significantly more patterns than the more sparsely connected networks makes it a much more desirable architecture. This is why the number of patterns that a neural network can store is one of the key metrics used for evaluation. This result, showing the global in-plane connection model outperforms the 4-way connection model, justifies significant further research and investigation into the feasibility of the global in-plane connection model.

4.4 Suggestions For Future Work

As the work on modelling and implementing the various in-plane connection architectures is in the early stages, there are several areas that require a significant amount of research. First, the models used for simulation need to be reworked in order to closer replicate the actual hardware being built. One of the first ways this can be

done is by discretizing the the weights stored in the network. As was described previously, both proposed architectures rely on the fact that weights are stored in digital form within the neurons. MATLAB, on the other, assumes analog weights by using high precision values for the connection weights. By passing the weights through an analog-to-digital converter after each training cycle within MATLAB, the digital nature of these architectures can be replicated. This will hopefully give a good understanding of the consequences of moving from an analog architecture to a digital architecture.

Another potential issue that needs to be understood is the storage of the weights in the global in-plane connection model. For global interconnects, the number of connections required increases as n^2 where n is the number of neurons. This means that for a true globally connected in-plane neural network like the one described, each neuron must store n weight values. While the CONNPP does intend to increase the number of neurons per layer as the project progresses, it does not intend to increase the physical size of the neurons. As a result, it is necessary to examine alternate methods to storing all n^2 weights. One way of accomplishing this is by storing a subset of the weights that are determined to be the "most important" by some predetermined algorithm. This potentially allows most of the functionality to be be retained while reducing the physical size requirements. Further research is needed to understand the impact of this architectural change on the performance of the CONNP.

The amount of chip real estate needed to implement each of the different architectures is also a major consideration that needs more investigation. While the global in-plane connections appear to produce a neural network with better performance (by some metrics), it is unclear how much physical space is needed to implement a single neuron. Likewise, this physical space requirement is also unknown for the originally proposed 4-way connection model. As the available chip space is a bounded resource, it is necessary to take this into consideration when looking at performance. One way to potentially quantify the space requirements is to use hardware level functional models built using a hardware description language such as Verilog or VHDL. By factoring in the size requirements for the various in-plane connection models along with

other performance based metrics, the best model can then be chosen and implemented in physical hardware to produce a performance optimized Compact Optoelectronic Neural Network Processor.

Bibliography

[1] C. Warde and C. Fonstad. Compact optoelectronic neural co-processor project. In *Proceedings of the ICOLA '02*, pages B07–B11, Jakarta, October 2002.

[2] S. Atkins. The fuss about the bus: PC 100 and the move to 100MHz SDRAM modules. *TechNet*, 1999.

[3] D. A. B. Miller and H. M. Ozaktas. Limit to the bit-rate capacity of electrical interconnects from the aspect ratio of the system architecture. *Journal of Parallel and Distributed Computing*, 41(1):42–52, 1997.

[4] Jan M. Rabaey. *Digital Integrated Circuits*, chapter 9. Prentice Hall, 1996.

[5] D. A. B. Miller. Rationale and challenges for optical interconnects to electronic chips. *Proceedings of the IEEE*, 88(6):728–749, 2000.

[6] Yi Ding. U.S. Patent 6,351,576, February 2002.

[7] J. W. Goodman, F. J. Leonberger, S. Kung, and R. A. Athale. Optical interconnections for VLSI systems. *Proceedings of the IEEE*, 72(7):850–866, 1984.

[8] D. Psaltis and N. Farhat. Optical information processing based on an associative model of neural nets with thresholding and feedback. *Optics Letters*, 10(2):98–100, 1985.

[9] J. H. Hong, S. Campbell, and P. Yeh. Optical pattern classifier with perceptron learning. *Applied Optics*, 29(20):3019–3025, 1990.

[10] K. Wagner and D. Psaltis. Multilayer optical learning networks. *Applied Optics*, 26(23):5061–5076, 1987.

[11] Donald Crankshaw. Aligned GaAs pilar bonding. S.m. thesis, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, June 1998.

[12] Wojciech Giziewicz. Optoelectronic integration using aligned metal-to-semiconductor bonding. S.M. Thesis, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, June-August 2000.

[13] Hao Wang. *Monolithic Integration of 1.55 Micron Photodetectors with GaAs Electronics for High Speed Optical Communications.* PhD thesis, Massachusetts Institute of Technology, Department of Materials Science and Engineering, August 1998.

[14] A. Rahman and A. Fan. Three-dimensional integration: analysis, modelling and technology development. *MIT Microsystems Technology Labs Research Report*, 2000.

[15] Russel Beale and Tom Jackson. *Neural Computing: an introduction.* Institute of Physics Publishing, 1990.

[16] L. Zhang, M. G. Robinson, and K. M. Johnson. Optical implementation of a scond-order neural network. *Optics Letters*, 16(1):45–47, 1991.

[17] A. D. McAulay, J. Wang, and X. Xu. Optical perceptron learning for binary classification with spatial light rebroadcasters. *Applied Optics*, 32(8):1346–1353, 1993.

[18] Eugene Hecht. *Optics.* Addison Wesley, 2002.

[19] T. E. Sale. *Vertical Cavity Surface Emitting Lasers.* Research Studies Press LTD., 1995.

[20] Bahaa E. A. Saleh and Malvin C. Teich. *Fundamentals of Photonics.* John Wiley and Sons, Inc., 1991.

[21] Graham Saxby. *Practical Holography.* Prentice Hall, 1988.

[22] Charles Kittel. *Introduction to Solid-State Physics.* Wiley, 7th edition, 1996.

[23] L. Solymar and D. J. Cooke. *Volume Holography and Volume Gratings.* Academic Press, 1981.

[24] Milos Komarcevic. Production of holographic optical interconnection elements. Master's of Engineering, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, September 2000.

[25] Marta Ruiz Llata. Final report September - November 2002. Final report of work done at MIT in the Fall 2002.

[26] Joseph F. Ahadian. *Development of a Monolithic Very Large Scale Integration Optoelectronic Integrated Circuit Technology.* PhD thesis, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, February 2000.

[27] Omar Wing. *Gallium Arsenide Digital Circuits.* McGraw-Hill Publishing Company, 1990.

[28] Paul R. Gray, Paul J. Hurst, Stephen H. Lewis, and Robert G. Meyer. *Analysis and Design of Analog Integrated Circuits.* Wiley, 4th edition, 2001.

[29] R. Li, J. D. Schaub, S. M. Csutak, and J. C. Campbell. A high-speed monolithic silicon photoreceiver fabricated on SOI. *IEEE Photonics Technology Letters*, 12(8):1046–1048, 2000.

[30] T. K. Woodward and A. V. Krishnamoorthy. 1 Gbit/s integrated optical detectors and receivers in commercial CMOS technologies. *IEEE Journal of Selected Topics in Quantum Electronics*, 5(2):146–156, 1999.

[31] M. Ingels and M. S. J. Steyaert. A 1 Gb/s 0.7-μm CMOS optical receiver with full rail-to-rail output swing. *IEEE Journal of Solid-State Circuits*, 34(7):971–977, 1999.

[32] Roger T. Howe and Charles G. Sodini. *Microelectronics: an integrated approach.* Prentice Hall, 1997.

[33] Peregrine Semiconductor, San Diego, CA. *Peregrine UTSi 0.5um Design Manual.*

[34] Ben G. Streetman. *Solid-State Electronic Devices.* Prentice Hall, 4th edition, 1995.

[35] H. Yamamoto, K. Taniguchi, and C. Hamaguchi. High-sensitiviy SOI MOS photodetector with self-amplification. *Japanese Journal of Applied Physics Part 1*, 35(2):1382–1386, 1996.

[36] W. Zhang, M. Chan, S. K. H. Fung, and P. K. Ko. Performance of a CMOS compatible lateral bipolar photodetecor on SOI substrate. *IEEE Electron Device Letters*, 19(11):435–437, 1998.