Design of Optoelectronic Activation, Local Memory and Weighting Circuits for Compact Integrated Optoelectronic Neural (COIN) Co-processor

by

Wegene Haile Tadele

B.S, Massachusetts Institute of Technology (2014)

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

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2015

© Massachusetts Institute of Technology 2015. All rights reserved.

Signature redacted

Author .........................

Department of Electrical Engineering and Computer Science

Signature redacted

July 27, 2015

Certified by...

Cardinal Warde
Professor
Thesis Supervisor

Accepted by ........

Dr. Christopher J. Terman
Chairman, Department Committee on Graduate Theses
DISCLAIMER NOTICE

Due to the condition of the original material, there are unavoidable flaws in this reproduction. We have made every effort possible to provide you with the best copy available.

Thank you.

Despite pagination irregularities, this is the most complete copy available.

* Thesis is missing pages 7-8
* Pages 51-52 contains text that runs off the edge of the page.
This Master of Engineering thesis has been examined by a Committee of the Department of Electrical Engineering and Computer Science (EECS) as follows:

Dr. Christopher J. Terman ........

Chairman, Thesis Committee
Professor of EECS

Professor Cardinal Warde.........

Thesis Supervisor
Professor of EECS
Design of Optoelectronic Activation, Local Memory and Weighting Circuits for Compact Integrated Optoelectronic Neural (COIN) Co-processor

by

Wegene Haile Tadele

Submitted to the Department of Electrical Engineering and Computer Science on July 27, 2015, in partial fulfillment of the requirements for the degree of Master of Engineering in Electrical Engineering and Computer Science

Abstract

The Compact Integrated Optoelectronic Neural (COIN) Co-processor, a prototype of artificial neural network implemented in hybrid optics and optoelectronic hardware, aims to implement a multi-layer neural network algorithm by performing parallel and efficient neural computations. In this thesis, we design and implement optoelectronic thresholding (activation), weighting and memory circuits for the COIN processor. The first version involved the design of fixed thresholding and weighting functions. The second version incorporated a local capacitive memory element as well as variable weighting schemes. The third version introduces an additional flexibility for variable thresholding by changing the bias voltages of control transistors. A 9x9 array of proof of concept printed circuit board (PCB) with an area of 4.5x4.5 in\(^2\) and total power consumption of 1.37W was designed and tested for version-I optoelectronic neuron architecture. A spice simulation was performed for the last two versions for integrated circuit (IC) implementation. The work developed in this thesis provides some guidance on the design of optoelectronic neural activation function for the realization of the embodiment of the fully integrated COIN Co-processor to be built in the future.

Thesis Supervisor: Cardinal Warde
Title: Professor
Acknowledgments

I would not be here without the support of so many people. First and foremost, I would like to thank my thesis supervisor Prof. Cardinal Warde. His continued support during my undergraduate and graduate research work at MIT has helped me grow both personally and professionally. Next, I would like to thank William Herrington for his invaluable advice and insight during the course of this research.

My sincere gratitude goes to my grandmother, Kuri Woldeyes, in my hometown in the South Central Ethiopian countryside. Without her unconditional support and sacrifice, nothing else would have been possible. My parents Damenech Kifle and Tadele Haile have always been my inspirations. My mother’s wise advice and my father’s encouragement to use my full potential have played an indispensable role in helping me accomplish my goals and ambitions. I would also like to thank my siblings Zenebe, Dawit, Abel, Mentesinot, Edelawit and Samrawit for their moral support and boosting my confidence. My utmost gratitude goes to the Perry Family especially to Barbara, Joan and Jimbo Perry for their loving parental support and helping me grow personally and spiritually. I would also like to thank Park and Whitley Morris for supporting me financially during my undergraduate education.

Furthermore, I would like to thank my valued friends Abiy Tassisa and Iyob Gebremariam. Their continued support both in and out of school has played a substantial role in my academic and professional endeavors. Moreover, I would like to thank everyone whose names are not mentioned here but have been part of my success journey directly or indirectly. Last but not least, I would like to thank my amazing girlfriend Natalie Strong for her encouragements and tirelessly reviewing my thesis.
DISCLAIMER NOTICE

MISSING PAGE(S)

#7
DISCLAIMER NOTICE

MISSING PAGE(S)

#8
# Contents

## 1 Introduction

1.1 Background on Neural Computing .......................... 22  
   1.1.1 Artificial Neuron Model .......................... 24  
   1.1.2 Learning in ANNs and Training Algorithms ............ 26  

1.2 The COIN Co-Processor and Hardware Implementation of Artificial  
    Neural Networks ........................................ 27  

1.3 Thesis Summary ........................................... 30

## 2 The COIN System

2.1 Background on Compact Optoelectronic Integrated Neural (COIN) Co- 
    processor .................................................. 33  

2.2 COIN Component Considerations ............................ 34  
   2.2.1 Photodetector Considerations ........................ 35  
   2.2.2 The Threshold Circuit ................................ 38  
   2.2.3 Driver circuits and weight storage elements ............ 38  
   2.2.4 Output Light Sources .................................. 38  
   2.2.5 Optical Interconnects for COIN ......................... 40  

2.3 Previous Work on Neural Activation Functions and Weighting Circuits 41  
   2.3.1 Optoelectronic Neuron Circuit with Variable Synaptic Weights 41  
   2.3.2 Digital-Analog Hybrid Synapse Chips for Electronic Neural Net- 
        works ...................................................... 43

## 3 Mathematical Models of Different Types of Activation functions 47
3.1 Generalized Thresholding Function ........................................ 48
3.2 The Sigmoidal Function ...................................................... 49
3.3 The Arc Tangent Activation Function ...................................... 50
3.4 The Hyperbolic Tangent Activation Function ............................ 51
3.5 Comparison of Neural Activation Functions ............................. 52

4 Activation Function Circuit for COIN: Version-I ....................... 55
  4.1 Implementation of Optoelectronic Neuron for Version-I .............. 58
    4.1.1 LED Characterization and Selection .............................. 65
    4.1.2 MOSFET and LED Modeling and Characterization ............... 67
    4.1.3 Phototransistor Modeling and Characterization ............... 69
  4.2 Version-I Optoelectronic Neuron Circuit ............................ 71
  4.3 The 9x9 Array of Optoelectronic Neuron Implementation in Discrete
    Components .............................................................. 73
    4.3.1 Result .............................................................. 74
    4.3.2 Discussion ........................................................ 75

5 Improved Circuit Architectures for COIN: Versions-II and III ....... 77
  5.1 COIN Model for the improved circuit Architecture .................... 78
  5.2 Low-Level Block Diagram of a Pixel ................................... 79
  5.3 Neural Weighting and Associative Memory Circuits ................... 80
  5.4 Optoelectronic Neuron Model for Version-II .......................... 82
  5.5 Spice Simulation Results for version-II .............................. 83
  5.6 Optoelectronic neuron model for version-III ......................... 85
    5.6.1 Subthreshold MOSFET Operation .................................. 85
    5.6.2 Differential Pair .................................................. 86
    5.6.3 Current Mirrors ................................................... 87
    5.6.4 Operational Transconductance Amplifier(OTA) .................. 88
    5.6.5 The Optoelectronic Neuron Model of Version III ............... 90
  5.7 Spice Simulation Results for version-III ............................. 90
  5.8 Recommendation for Integrated Circuit (IC) Layout ................ 92
5.9 Limitations on Circuit Design and Neuron Density

6 Conclusions
6.1 Summary of Contributions
6.2 Applications
6.3 Discussion of Future Work
List of Figures

1-1  Structure of Biological Neurons: Presynaptic and postsynaptic neurons[61]  23
1-2  A three layer network: stimulation is applied to the inputs of the first layer, and signals propagate through the hidden layer to the output layer. Each connection between neurons has a unique weighting value.  25
1-3  Artificial Model of Neuron ................................................................. 25
1-4  Schematic architecture of a multilayer optical neural network. Each circle represents a neuron while a filled circle means that the neuron is in the "ON" state ................................................................. 29

2-1  A Block Diagram of a three layer COIN Co-processor system. Adopted from the work of Travis Simpkins[56]. Each layer consists of arrays of photodetectors, optical interconnection elements and thresholding circuits ................................................................. 34
2-2  A Diagram of a COIN neuron ................................................................. 35
2-3  Lump-sum equivalent-circuit model of a photodiode ............................... 36
2-4  Phototransistor Equivalent Circuit .......................................................... 37
2-5  Optoelectronic neuron circuit and synaptic connection \( (T_{ij}\) cell). a) Basic neuron circuit. Photocurrent \( I_{ij}\) flows from MOSFET Q1 to PD for excitatory connection, and flows in inverse direction for inhibitory connection. b) Photo-currents for excitatory and inhibitory connections. Bipolar \( T_{ij}\) values were obtained by following combination. Adapted from [66] ................................................................. 42
2-6  .............................................................................................................. 42
2-7 Primitive optoelectronic CAM circuit with 4 x 4 network. Adapted from [66] .......................... 43

2-8 Simplified block diagram of hybrid 32x32x7-bit synapse chip. Adapted from [38] .................. 44

2-9 Block diagram of 7-bit multiplying DAC. Adapted from [38]. .................. 44

2-10 Transfer characteristics of a 7-bit synapse for weight values of 0 +/-1, 3, 7, 15, 31, 63 (a) with long channel transistors and (b) with external 10MΩ for voltage to current conversion. Adapted from [38]. .................. 45

2-11 Electronic circuit neuron model and synapses. Where V_i is the output of the neuron i, T_ij is the synaptic weight from neuron j to neuron i, R_f and C_f are the feedback capacitance and resistance of the neuron, and I_i is the external input current. R_F and C_f should be chosen to ensure stability against oscillations. Adapted from [38]. .................. 45

3-1 Plots of generalized thresholding function plotted at three biases and rise speeds. a) Generalized activation function responses at three different bias points of B=0,1, and 3. b) Responses at three different rise speed of V= 2, 3.33 and 2.5. Only one of these parameters are varied for a given plot. .................. 49

3-2 a) Plots for different values of the rise speed(\( \frac{C}{M} \)). b) The sigmoidal function for three different bias values (B). The bias value in the optoelectronic activation corresponds to the turn on voltage of the transistor 50

3-3 a) A plot of arc-tangent function for three different lateral shift (zero crossing) values. The zero crossing values specified by B are 2, 3 and 4. b) The response of arc-tangent thresholding function at fixed bias point B for three different values of rise speed (C). The value of B is 3 and the rise speeds for each curve are specified in the equations 51
3-4  a) The plot of hyperbolic tangent function three different lateral shift values (B). The lateral shift (B) corresponds to the zero crossing or the thresholding point in this function. b) A plot of hyperbolic tangent function for three different values of rise speed (C).

3-5  a) Plots of the three thresholding functions; sigmoidal, hyperbolic tangent and arc-tangent functions. The sigmoid function goes from 0 to 1 and the arc-tangent and hyperbolic tangent functions go from -1 to 1. b) The derivatives of the three functions which is the same as the error. We can compare the convergence speed of the error associated with each function from the plot.

4-1  Block Diagram of one pixel of the COIN Co-processor for Version-I Optoelectronic neuron circuit: Interconnects driven by one light source per neuron and modulated by a Spatial Light Modulator (SLM).

4-2  A signal flow diagram between layers of neurons. Adopted from [56].

4-3  The inputs and outputs of a single neuron are represented by the central circle. Each rectangle associated with some weight represents a synapse, such that a neuron and a succeeding synapse comprises one pixel of the proposed COIN Co-processor. The preceding synapses are also associated with pixels in the preceding neuron layer.

4-4  An Optoelectronic Neuron Circuit: VDD is supply voltage, VB1 and VB2 are bias voltages for transistors M1 and M3 respectively. The voltages are generated such that the two transistors operate in triode region. These two transistors can also be replaced with a fixed resistor.

4-5  Photodarlington transistor pair. PT1 = Phototransistor1 and PT2 = Phototransistor2.

4-6  MOS resistor operated in triode region.

4-7  (a) LED characterization circuit, and (b) a plot of LED current vs. Relative output intensity.
4-8  (a) MOSFET and LED characterization circuit, and (b) a plot of gate voltage ($V_{Gate}$) vs. Relative output intensity ........................................ 67

4-9  (a) Relative output intensity ($I_{out}$) vs. gate voltage ($V_{Gate}$) (b) Output LED current ($i_{out}$) vs. gate voltage ($V_{Gate}$) (c) Relative output intensity ($I_{out}$) vs. output LED current ($i_{out}$) ........................................ 68

4-10 (a) Phototransistor characterization circuitry setup. (b) Relative input intensity vs. Phototransistor emitter current. (c) The non-saturated section of plot (b). The polynomial fit was done only for the non-saturated section of plot (b) since that is the only region of interest for this application. ........................................ 70

4-11 An Optoelectronic Neuron Test Circuit. The input light source was placed 4.5 cm away from the phototransistor and at the angle of maximum-beam divergence of the LED ........................................ 72

4-12 Relative output intensity ($I_{out}$) vs. Emitter current of the phototransistor ($i_{in}$). Raw data is shown in dotted lines and the sigmoidal fit is shown in solid lines. The sigmoidal fit was achieved using a function of the form $y = \frac{a}{1+e^{k(x-x_0)}}$ ........................................ 73

4-13 9x9 Optoelectronic Neuron Array PCB front. The front plane consists of arrays of output LEDs and MOSFETs ........................................ 74

4-14 9x9 Optoelectronic Neuron Array PCB back showing the phototransistors array and resistors. ........................................ 74

4-15 Input of 5. 5 is presented as an input to the 2-D array of optoelectronic neurons. The brightness is adjusted with the polarizers ........................................ 75

4-16 Output of 5. The output is generated by activating the corresponding neurons in the 2-D array of optoelectronic neurons ........................................ 75

5-1 COIN Block diagram of Versions-II and III electronic circuits: Interconnects driven by individual sources and modulated by controllable driver circuits ........................................ 78
5-2 Block Diagram of one pixel of the COIN Co-processor for Version-I Optoelectronic neuron circuit: Interconnects driven by one light source per neuron and modulated by a Spatial Light Modulator (SLM) .......................... 79

5-3 A Pixel. $w_i$ corresponds to weights stored in local memory elements. $C_i$ corresponds to the optical interconnection element used for inter-plane communication. .......................................................... 80

5-4 A Basic short term memory circuit implemented using sample-and-hold circuit .......................................................... 82

5-5 Version-II Optoelectronic neuron model with capacitive local memory element. Weights are periodically updated from the stored weight elements. $i_{outx}$ and $I_{outx}$ correspond to the output current and relative output intensity of synapse x respectively. $i_{in}$ is the photogenerated current of the pixel from external illumination. ............................ 83

5-6 Version-II optoelectronic neuron: A schematic with three synapses used to obtain simulation plots .......................... 84

5-7 Plots for various weights. The three synapses have three different weights that are controlled by the voltage divider at the gate of the output MOSFET. .......................................................... 84

5-8 Differential Pair .......................................................... 87

5-9 Current Mirrors .......................................................... 88

5-10 Operational Transconductance Amplifier (OTA) ................. 89

5-11 OTA transfer curve for $I_{PD} = 10 \mu A$, $\frac{1}{k_s} = 1.3$ ........................ 90

5-12 Version-III Optoelectronic Neuron Model with capacitive local memory element. It is comprised of the activation circuit which contains the mirror legs and the input transistors, the current mirror for distributing current among all the synapses, the analog memory in order to enable local memory storage and the output circuit which is regulated by the analog memory element to output a weighted light intensity through the LED ....................................................... 91
5-13 Version III optoelectronic neuron: A schematic circuit for spice simulation using the 0.25 µ process technology ... 91

5-14 Spice Simulation results for variable thresholding circuit. The threshold value is controlled by the relative gate voltages of the OTA. Four plots are given for various combinations of V2 and V3 ... 92

5-15 Simulation for various weights. Not only the threshold value but also the weights can be adjusted in this circuit. This will enable on chip learning in an integrated analog neural processors ... 92
List of Tables
Chapter 1

Introduction

In recent years, fueled in part by the rapid development of integrated circuits, faster digital computers and advances in optics, there has been a growing effort to imitate the capabilities of the brain both in hardware and software. Although we do not know how even the simplest brain works, we have some clues on how information propagates from one neuron to another. With the simple clues we have so far, we aim to use artificial neural networks (ANNs) to mimic and simulate some of the ubiquitous processing capabilities of the human brain. The human brain is composed of loosely coupled, globally asynchronous, massively parallel, noisy and unreliable components [24], [49], [52] that can carry out robust and energy efficient computation within a matter of milliseconds. These networks of neurons and synaptic interconnections work in unison to solve specific classes of problems such as pattern recognition, speech recognition, classifications and many other computations that human brain can do but digital computers can't efficiently and reliably.

Software versions of artificial neural networks have been around for longer than half a century, and large ones have recently made breakthroughs in speech and feature recognition from companies such as Facebook and Google. But such simulated neural networks require large number of conventional digital computers. A recent Google experiment in which a large number of neural networks learned to recognize cats from YouTube images run on 16,000 processors over three days [35]. Physically building a
neural network on a chip could make this type of information processing much faster, power efficient and most importantly practical. For example, such neuro-inspired chips might allow mobile robots to become much smarter and give them the power to recognize object features and various patterns in real time.

Nothing is more suited for such an application than an optoelectronic implementation of hardware neural networks due to its innate parallelism [39], [40], high density inter-connectivity [34] and direct image processing ability. The Compact Optoelectronic Integrated Neural (COIN) Co-processor which has been a major area of research for more than a decade in the Photonics Systems Lab at MIT aims to come up with an optimal solution for bottleneck problems associated with software based neural network systems that run on thousands of digital computers. The COIN system is a three dimensional device consisting of light emitting diode arrays, photodetector arrays, electronic activation function and optical interconnects for neuron-to-neuron communication. For the most recent embodiment of the COIN Co-processor, we plan to incorporate a local memory element to realize on-chip learning and large scale integration of the system for various applications.

This chapter focuses on background information on artificial neurons models, training algorithms and the motivation for the development of optoelectronic neural processors. It will conclude by introducing the compact integrated optoelectronic neural (COIN) Co-processor and a brief outline of the remainder of the thesis.

1.1 Background on Neural Computing

The human brain consists of billions of neural cells that process (receive, decide and propagate) information. Each neural cell works like a simple non-linear processing elements with a local memory and computation unit. The computation strength of a simple element is almost null but the massive interaction between all the neural cells results in the most powerful machine-the brain. The key components in neural signal
processing are the dendrites (weighting and delay elements), the soma (integration unit), the axon hillock (thresholding unit), the axon (output nodes) and the synapse (inter-neuronal junctions).

**Synapses** are junctions that allow electrochemical communication between the axons and the dendrites. These inter-neuronal interconnections can be either excitatory or inhibitory depending on whether they enhance or diminish the probability of firing a postsynaptic neuron.

**Dendrites**: are the root like fibers that emanate from the cell body or soma. They receive messages from other neurons by grabbing onto neurotransmitters in the synaptic cleft.

**Cell body or Soma**: contains the processing unit of the cell. It processes the incoming activation signals and informs the cell to fire or not to fire an action potential.

**Axons**: are wire like fibers that extend from the cell body to the axon terminal. They serves as a transmission line for sending activation signals to other neurons.

**Axon hillock**: is a site in the neuron located at the interconnection of the axon and the soma where incoming information is summed.

Signals from connected neurons are collected by the dendrites and these signals get summed spatially or temporally by the soma. The majority of neurons encode their outputs as a series of spikes or action potentials. When sufficient input is re-
ceived such as the summed value exceeds the threshold, the neuron generates action potential or spikes. The action potential is propagated along the axon to other neurons or to structures outside the nervous system e.g. muscles. If sufficient input is not received such that the summed signal value is less than the threshold value, the inputs quickly decay and no action potential is created. In neural processing, timing is very important-input signals must arrive simultaneously, strong input signals generate more action potentials per unit time.

1.1.1 Artificial Neuron Model

Biological neural networks have inspired the design of artificial neural networks. An artificial neuron is a mathematical function conceived as a model of a real biological neuron. Ideally, we would like to imitate the human brain's function. However, because of technological limitations, we should settle for a much simpler model. The most obvious approach to create an artificial neuron is to design a small electronic device which has a transfer function similar to a biological neuron, and then connect each neuron to many other neurons, using RLC networks to mimic the dendrites, axons and synapses. This type of electronic model is still rather complex to implement and we may have difficulty training the network to do anything useful. Further constraints are needed to make the design more manageable. First, we change the connectivity between the neurons so that they are in distinct layers, such that each neuron in one layer is connected to every neuron in the subsequent layer. Second, we define that signals flow in only one direction across the network and we simplify the neuron and synapse design to behave as analog comparators being driven by other neurons through passive elements such as resistors. This will create a feed-forward neural network model that may actually be practical to build and use.

Referring to Figures 1-2 and 1-3, this simplified model of the real neurons contains a set of input connections that brings in signals from other neurons, a synapse that connects neurons of different layers and a weight unit for each synapse. The summing
Figure 1-2: A three layer network: stimulation is applied to the inputs of the first layer, and signals propagate through the hidden layer to the output layer. Each connection between neurons has a unique weighting value.

Figure 1-3: Artificial Model of Neuron

junction integrates the inputs and then applies a non-linear activation (thresholding) function to make an output decision. The output of the activation function is then broadcasted to some or all of the neurons in the subsequent layers. In order to use the network, we apply the input values to the inputs of the first layer, allow the signals
to propagate through the hidden layers of the network and read the output values at the output layer.

Fig-1-3 represents the mathematical model of a neuron. Signals enter the synapse on the left, are either enhanced or diminished depending on the weights associated with the synapse and are then summed in the neuron. The neuron then applies a simple activation function \((f(\cdot))\) to the sum and outputs \((y_k)\) the result of which propagates to further neurons via other synapse. The operation of this simple neuron can be modeled as follows in the following equation:

\[
y(x, w) = f\left(\sum_{i=1}^{k} x_i w_i\right)
\]

where \(x_i\) and \(w_i\) are the inputs and the weights respectively, and \(f(\cdot)\) is the thresholding function of the neuron.

1.1.2 Learning in ANNs and Training Algorithms

Artificial Neural Networks (ANNs) learn through training. The real intelligence of the networks exist in the values of the weights between neurons. Learning in biological systems involves adjustments to the synaptic connections (weights) that exist between neurons. Similarly, we need a method of adjusting the weights between neurons in artificial neural networks to solve a particular task. One of the most commonly used weight adaptation algorithms for multilayer perceptron neural network (MLPNN) is known as the back propagation algorithm (BPA). BPA is a learning procedure for networks of neuron-like units by repeatedly adjusting the weights of the connections in the network to minimize the difference between the actual output of the network and the desired output vector[10]. BPA is the best example of a parametric method for training supervised multi-layer perceptron (MLP) neural network for classification and pattern recognition [6], [7], [50]. BPA, like other supervised multi-layer feed forward neural network training algorithms has the ability to learn weights and input biases.
BPA uses gradient descent to iteratively minimize the gradient of the error function. It always shows the direction of the steepest ascent of the error function. Therefore, we can start with a random weight vector and follow the negative gradient with some learning rate $\eta$.

$$w_{t+1} = w_t + \delta w_t, \quad \delta w_t = -\eta \frac{\partial E(w)^T}{\partial w} \bigg|_{w_t}$$  \hspace{1cm} (1.2)

It should be noted that this will only find the local minimum. Therefore, the gradient descent procedure should be repeated for several random initial states. Gradient descent based BPA is a powerful method to manage systems that use data to adjust the network weights and thresholds for minimizing the errors in its prediction on the training data set. In Simpkins’s [56] work a gradient descent algorithm was used to train a five layer COIN co-processor.

### 1.2 The COIN Co-Processor and Hardware Implementation of Artificial Neural Networks

There are some direct analogies between biological neural systems and analog neural processors. This includes amplification, exponentiation, thresholding, compression and integration. The parallels between these two worlds run from the level of device physics to circuit architecture. Artificial analog neurons are realized through analog components like adders, multipliers, current to voltage converters, memory circuits and thresholding elements.

In the last few decades, most efforts to develop ANN’s have focused on the area of mathematical considerations and are mainly implemented in software [29]. However, in recent years researchers have made efforts to create neuro-inspired processor chips to process information in hardware. One of such processors include the recent state-of-the-art IBM’s custom made ”brain-like” chip called TrueNorth. TrueNorth
comes packed with 256 million neurosynaptic cores, 64 billion neurons and 16 Trillion synapse [47]. One of the reasons for the growing interest in hardware based ANN's is because they can be faster compared to the software ones when very large number of interconnections are required as in the case of the IBM TrueNorth.

Most hardware implementation of neural networks is based on very large scale integration (VLSI) technology which is functionally highly versatile but mostly limited to two dimensions. Although the planar nature of VLSI is not necessarily a restriction for neural implementations but the lack of a third dimension limits the number of synaptic densities that can be accomplished due to interconnection delays and packaging of input/output pins [65], [68]. The use of optical techniques to implement interconnections in neural network systems appears to be a very attractive approach [1],[17],[21]. Moreover, the dynamic modification of interconnections as required by the training algorithm is difficult to implement electrically. The key advantages over electronic interconnections are the massive parallelism [41], speed and cross-talk free interconnections.

Hybrid opto-electronic neural systems such as the COIN processor combine the technological advantage of both domains, with functionally rich local electrical processing and global optical interconnects. With optical signals as inputs, COIN performs the required computation using electronic circuits and the processed result is presented as another optical signal which can be interconnected optically through free space or optical elements to subsequent processing units. Because optical signals do not interfere with each other in the same way as electrical signals, we can accomplish parallel computation and eliminate interconnection delays and I/O issues [19], [27], [28]. Consequently, the performance of an integrated processor is greatly enhanced. Thus, optics is well suited for this kind of system because the interconnections can be implemented via the third plane by utilizing the parallelism of optical signals [26], [27].

A schematic of a three layer optoelectronic neural network architecture is given in Fig. 1-4. The neurons are arranged as 2-dimensional arrays and each array corre-
sponds to one neuron layer in the network. Although there is always one input layer and one output layer, the number of hidden layers can be changed depending on the complexity of the task. Since the interconnections between different neuron layers are implemented by optical signals, each pixel must be able to detect incoming optical signals and produce optical outputs. As a result, each neuron is generally comprised of three components: optical receivers (i.e. photodetectors), non-linear electronic elements (activation functions) and optical output transmitters (e.g OLED). The most commonly used activation function in neural networks is the sigmoid thresholding function. Other types of neuron activation functions include the tanh function, the saturated linear function, etc. A description of different types of activation functions and their performance metrics will be discussed in detail in chapter-3.

Figure 1-4: Schematic architecture of a multilayer optical neural network. Each circle represents a neuron while a filled circle means that the neuron is in the "ON" state.

Input signals detected by the optical receivers are mapped to the optical output.
transmitters through the non-linear electronic elements as:

\[ y_j = G f(\sum_{i=1}^{N} w_{ij} x_i) \]  \hspace{1cm} (1.3)

where \( y_j \) is the optical output from the \( j \)th neuron in the current layer, \( G \) is the opto-electronic gain, \( f(.) \) is the non-linear neuron response function, \( N \) is the total number of input neuron in the previous layer, \( w_{ij} \) represents the interconnection weight between the two neuron layers, \( x_i \) is the input from \( i \)th neuron in the previous layer.

A neural network can be trained to solve a specific task by adjusting the interconnection weights using an algorithm such as the gradient descent learning algorithm[3]. In addition to the easy implementation of optical I/O's and the benefits of densely packed interconnections, an optical neural network can also be used to implement the weights in the form of holographic memory [11], [31], [32]. Furthermore, the interconnection weights can be adjusted in real-time using dynamic holograms recorded in photo-refractive crystals. Therefore, the training of an optical neural network can be done relatively easily by updating the weight according to the comparison result between the output of the neural net and the desired response.

1.3 Thesis Summary

The remainder of this thesis consists of three parts encompassing the body of the work completed as part of the development process of the COIN Co-processor. Chapter 2 provides background information on COIN system. Furthermore, previous work regarding the circuit implementations of activation functions are discussed. In chapter 3, the detailed mathematical models of different types of neural activation functions and their performance metrics will be explored.

Next, the major contribution of this thesis is presented in chapters 4 and chapter 5. Three versions of the optoelectronic thresholding function will be explored in these chapters. The first version used the original COIN processor model to come up with a memory-less, fixed weighting and fixed thresholding circuits to perform
optoelectronic thresholding of the input optical signal. The original COIN processor used spatial light modulators (SLMs) to adjust the weights between neurons. Version-II and version-III focus on the most recent development of the COIN Co-processor architecture and will have both analog weighting and memory elements. The integration of the weighting circuits into the thresholding circuits eliminates the presence of the bulky SLMs in between neuron layers. Version-III is different from version-II because of the additional flexibility in changing the thresholding voltage of the activation function by applying different bias voltages to shift the turn on points. Furthermore, version-III requires IC implementation since some of the transistors need to be operated in sub-threshold regimes. Version-III is based on the optoelectronic implementations of the hyperbolic tangent functions in analog CMOS circuits.

Finally, chapter 6 presents the conclusions, the applications of ANNs (particularly the COIN processor) and future plans.
Chapter 2

The COIN System

In this chapter, a description of the COIN Co-processor and the major components that makeup the architecture of the system will be discussed thoroughly. Furthermore, previous work on hardware, particularly, mixed-signal and optoelectronic circuit implementations of the thresholding functions will be presented.

2.1 Background on Compact Optoelectronic Integrated Neural (COIN) Co-processor

The work presented in this thesis is part of the development of the COIN Co-processor in the Photonics Systems lab at MIT. The COIN Co-processor is a pixellated, rugged, parallel, feed-forward and integrated optoelectronic implementation of an artificial neural processor. The basic architecture of the COIN co-processor is shown in Fig. 2-1 consists of 3 layers each of which contains an array of N-by-N pixels. Information enters optically through one end of the device, propagates through several layers of individual processing elements, and exits the opposite end of the device. Each layer consists of 2-D arrays of photodetectors which do summation of the input light signals, threshold electronics which perform thresholding, optical interconnect elements which steer optical signals between layers, and a controller which supervises the system. These layers of networks are cascaded to form a multi-layer processor and in the
work of [56] a five layer network was implemented and trained. The full embodiment of this system would allow computers to solve a new class of problems and complex functions such as pattern recognition and approximation at a faster speed and lower power expenditure.

This architecture will be slightly modified to incorporate weighting circuits and associative (local) memory elements in versions-II and III. Designs of the electronic processing circuits are discussed in chapter 5.

Figure 2-1: A Block Diagram of a three layer COIN Co-processor system. Adopted from the work of Travis Simpkins[56]. Each layer consists of arrays of photodetectors, optical interconnection elements and thresholding circuits

2.2 COIN Component Considerations

In this section, the core components that are required for the realization of a single layer of the COIN co-processor will be presented. This includes a photodetector, threshold circuit, driver and weight circuits, optical output sources and optical interconnection elements. The pros and cons of various detectors and optical output sources (such as transmitters and emitters) are also explored briefly.
2.2.1 Photodetector Considerations

The photodetector is an optical signal receiver which converts the received optical signal to an electrical signal (current or voltage). The three important performance parameters for the selection of the COIN photodetectors are high sensitivity, linearity and dynamic range. The magnitude of the dynamic range which is the ratio of the maximum and the minimum detectable power is one of the most important measures of the performance of the detector. The maximum detectable power is limited by a non-linear response (saturation in this case) whereas the minimum detectable power is determined by the noise level of the photodetector. Furthermore, the photodetector should exhibit high linearity over a wide range of dynamic range. The linearity is extremely important because it determines the maximum amount of light intensity that can be collected and summed without saturating the phototransistor. High responsivity and high quantum efficiency with minimum equivalent noise are also relevant. Quantum photodetection can be accomplished by using photodiodes, phototransistors or photoresistors [10], [20], [36], [64].

Photodiodes are semiconductor devices with a p-n junction or p-i-n structure and
they generate a photocurrent by absorbing light in a depletion region[53], [5]. Photodiodes are very compact, fast, highly linear and exhibit high quantum efficiency (i.e generates nearly one electron per incident photon [22], [42]) and high dynamic range if they are operated in combination with suitable electronics. The ideal photodiode can be modeled as a current source in parallel with a semiconductor diode. The current source corresponds to the current flow caused by the light-generated drift current whereas the diode corresponds to the behavior of the photodiode junction in the absence of incident light. An actual photodiode will have a junction capacitance ($C_j$) that is generated due to the depletion layer that exists between the two conductive layers and a finite shunt resistance ($R_{SH}$)[59] as shown in Fig. 2-3. A photodiode acts like a photo-controlled current source in parallel with a semiconductor diode and is governed by the standard diode equation.

\[ I = I_{\text{photo}} + I_{dk}(e^{\frac{qV_D}{kT}} - 1) \]  

(2.1)

where $I$ is the device current, $I_{\text{photo}}$ is the photo-generated current, $I_{dk}$ is the dark or the "non-light" reverse current, $V_D$ is the forward bias voltage, $q$ is charge of an electron, $k$ is Boltzmann's constant and $T$ is the temperature in degrees Kelvin. The photocurrent can be amplified and fed into a readout device. $I_{dk}$ is the diode's reverse leakage current and flows even in the absence of incident light. As we see from the standard diode equation, this dark current increases with increasing reverse voltage ($V_D$) bias and temperature. It approximately doubles every 10°C [55], [67].

Figure 2-3: Lump-sum equivalent-circuit model of a photodiode
Phototransistors are photodiode-amplifier combinations integrated within a single silicon chip which exhibit more sensitivity compared to photodiodes. These combinations are designed to overcome a major limitation of photodiodes: unity gain. The phototransistor can be viewed as a photodiode whose output current is fed into the base of a conventional signal transistor as depicted in Fig. 2-4. Based on this model of the phototransistor it displays some of the characteristics of both types of devices. The typical gain of a phototransistor can range from $100 - 100000$. The structure of phototransistor is very similar to that of a photodiode. In fact, the collector-base junction of the phototransistor can be used as a photodiode with fairly good results. The major structural difference is that the phototransistor has two junctions compared with one for the photodiode.

Photoresistors which are based on certain types of semiconductors can also be used as photodetectors. They are cheaper than photodiodes but they are slow, less sensitive and exhibit a strongly non-linear response. The photoresistivity of photoresistors is highly sensitive to temperature and thus they are unsuitable for applications requiring precise measurement of sensitivity to light.
2.2.2 The Threshold Circuit

Optical networks always require a non-linear functions that are difficult to implement optically. As a result, optical neural networks leave this non-linear function to be implemented electronically. The use of optoelectronic circuits provide the flexibility of implementing complex neuron response functions and of fine tuning the properties of the neurons as is required by the neural network algorithm that is being implemented. The threshold circuit is a light in-light out device that performs intensity thresholding. Based on the sum of the input intensities from nine neighboring previous layer nodes, it decides whether to generate a high or a low output intensity. Each neuron in array would receive signals from previous layer neighboring arrays, compute a non-linear function based on these input values and transmit signals to next layer neuron arrays. The threshold circuit is analogous to the squashing function in artificial intelligence(AI). However, the threshold circuit in this case employs optoelectronic circuit devices as opposed to software algorithms to accomplish light intensity summation and output decisions.

2.2.3 Driver circuits and weight storage elements

The driver circuits are often used to regulate the current flowing through the output light sources and can easily be accomplished with a simple MOS resistor. The weights are stored in a short term analog memory such as capacitors. These weights are uploaded to the gates of driver circuits periodically during the learning process to regulate the relative output intensity of the synapse circuits.

2.2.4 Output Light Sources

The output light sources act as the interface between the electronic domain and the optical domain of the COIN Co-processor. The ultimate goal of the COIN project is to develop a compact and integrated optoelectronic neural processors. Thus, the optical output light sources are extremely important in determining the transmission efficiency and the size of the system. There are a few candidate optical transmitters for
the integrated optoelectronic neural networks. These include semiconductor LEDs, laser LEDs, optical modulators and Organic LEDs (OLEDs).

The Light Emitting Diode (LED)

The light emitting diode (LED) possesses a p-n junction which emits light when a forward bias voltage is applied. Electrons in the donor level below the conduction band of the junction move to the acceptor level above the valence band and lose energy in the process. This energy is emitted as radiant energy in the infrared or the visible region. LEDs are commonly made of GaAs doped with one or more impurities. They are relatively easier to integrate into a chip compared to optical modulators. Furthermore, the internal quantum efficiency of a double heterojunction semiconductor LED can be very close to 100%[54]. However, the fact that LED’s emit light in all direction makes it very difficult to collect the optical output. Consequently, LED’s generally suffer from low external efficiency. As a result, higher driving current is needed which results in higher power consumption. Furthermore, since the output of the LED comes from spontaneous emission, it has a wide-band spectrum and is therefore temporally incoherent. This incoherence makes the recording of holographic interconnections impossible[48] and limits the application of LED-based neuron arrays to non-adaptive systems.

Organic LED (OLED)

OLED which stands for organic light-emitting diode is made with organic compound that lights up when a voltage greater than the minimum turn-on voltage is applied across the two terminals. Although LEDs have high optical output and a wide bandwidth in the order of MHz, the production is expensive compared to OLEDs. One of the major advantages of OLEDs compared to LEDs is that a large area of its arrays can be produced at lower cost. Compared to incandescent and fluorescent light sources, white OLEDs can emit light that is brighter, more uniform with higher power efficiency[33]. Furthermore, OLEDs can be made to be extremely small, thin, flexible and most importantly can be spun on to a chip on a wafer. The MEng thesis
work of [51] explored the applications of OLEDs for COIN system and compared its performance to Vertical Cavity Surface Emitting Lasers (VCSELS). The work of [51] and others indicate that OLEDs have the potential to emerge as alternative light sources for integrated optoelectronic neural processors.

**Diode Lasers**

Laser diodes are the other candidates for on chip output light sources. Its bandwidth is much narrower compared to LED’s and therefore has high degree of coherence. The light is emitted in one direction and therefore has very high efficiency. However, laser diodes require more complicated material preparation and processing techniques[37]. In addition to that, lasers have a threshold input power before coherent light is emitted, and therefore will have a higher power consumption per neuron than LEDs.

**Optical modulators**

Optical modulators are also possible alternatives for output light sources. The optoelectronic gain can be increased easily by increasing the power level of the laser source without affecting the power consumption of the neuron chip. As a result, we can build arrays with high density due to the lower consumption of modulator based neuron light sources. By using single laser as the off-chip light source, we can achieve both temporal and spatial coherence which is required for the effective implementation of the holographic interconnections in adaptive neural networks[57]. However, the alignment of the optical system is more complicated due to the addition of off-chip light sources.

**2.2.5 Optical Interconnects for COIN**

The optical interconnection element provides a free space communication between different layers of neurons. It directs the output of a synapse to the correct set of neurons in the subsequent layer. Bill Herrington’s PhD work[23] focused on developing this medium for the COIN Co-processor.
2.3 Previous Work on Neural Activation Functions and Weighting Circuits

Various efforts were made previously to create neuro-inspired local (associative) memory elements and activation circuits in both optics and electronics domains[18]. For many reasons that were discussed in the previous chapter most optical weighting and thresholding implementations were found to be bulky and non-adaptive for hardware based neural networks. Consequently, the majority of the discussion on previous work will be focused on optoelectronic implementations of local memory and activation functions in digital, analog or a combination of both domains along with optical emitters and receivers.

2.3.1 Optoelectronic Neuron Circuit with Variable Synaptic Weights

In this journal letter[66], a basic optoelectronic circuit of artificial neuron was realized by controlling the synaptic weights electronically. An addressable memory unit was constructed using an optical interconnection element between the emitters and the detectors. The basic neuron circuit in this work consists of three parts which are the $T_{ij}$ cells, a current summing node and an output node, as shown in Fig.2-6. The $T_{ij}$ cell acts as the synaptic connection and connection weights and polarity are controlled by the gate voltages $V_{g1}$ and $V_{g2}$ of NMOS FETs of Q1 and Q2. For excitatory neural connection $V_{g1}$ is high and $V_{g2}$ is low; thus Q1 is "on" and Q2 is "off". For inhibitory neural connections, $V_{g1}$ is low and $V_{g2}$ is high; thus Q1 is "on" and Q2 is "on". When both $V_{g1}$ and $V_{g2}$ are low no connection is formed and thus both Q1 and Q2 are "off". Furthermore, the artificial neuron could have many $T_{ij}$ cells like the human brain collecting signals broadcasted from neighboring synapses.

The photocurrent $I_{ij}$ is generated using a bidirectional photodiode(PD) when the optical outputs from other neurons are received. The photocurrent flows from Q1 to PD in the excitatory connection and from PD to Q2 in the inhibitory connection. The
Figure 2-5: Optoelectronic neuron circuit and synaptic connection ($T_{ij}$ cell). a) Basic neuron circuit. Photocurrent $I_{ij}$ flows from MOSFET Q1 to PD for excitatory connection, and flows in inverse direction for inhibitory connection. b) Photo-currents for excitatory and inhibitory connections. Bipolar $T_{ij}$ values were obtained by following combination. Adapted from [66]

<table>
<thead>
<tr>
<th>$T_{ij}$</th>
<th>$V_{g1}$</th>
<th>$V_{g2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
<td>V</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>-1</td>
<td>0</td>
<td>5</td>
</tr>
</tbody>
</table>

Figure 2-6:

Photocurrents from neighboring $T_{ij}$ cells were summed in the current summing portion of the circuit. When the summed result is above the threshold value, the GaAs LED located in the output circuitry emitted optical light to the other neurons. When the summed value is less than the threshold value, no optical output light was emitted to other neurons. Finally, a 4x4 array with a content addressable memory (CAM) was implemented as a proof of concept for a large scale optoelectronic network as shown in Fig. 2-7.
2.3.2 Digital-Analog Hybrid Synapse Chips for Electronic Neural Networks

In this work [38], a 32X32 programmable array of synapse chip was fabricated for parallel implementation of neural networks. The synapses are based on a hybrid digital-analog design which utilized on-chip 7-bit data latches to store quantized weights and a two quadrant multiplier DAC’s to compute weighted outputs. As depicted in the block diagram given in Fig. 2-8, the chip consists of an address/data de-multiplexer, row and column address decoders, 64 analog input/output lines and 1024 cells arranged in the form of a 32x32 cross-bar matrix.

A circuit diagram of the 7-bit DAC is shown in Fig.2-9. The DAC consists of a current input circuit, a set of binary weighted current sources and a current steering circuit. The current in the input circuit is mirrored by the binary-weighted current sources for all synapses along the column. Two versions of the chip were fabricated. In one version, a single long-channel FET was used to convert the synapse input voltage into current. In the second version, an external resistor was employed to perform input voltage to current conversion when high linearity was required.
responses of both versions of the chips are shown in Fig. 2-10

In this chip, two optimization problems, namely, the one-to-one assignment problem and the traveling salesman problem were implemented. In order to test the performance of the network for speed and quality, a 64-neuron breadboard system used along with operational amplifiers operating as a current summing amplifier as depicted in Fig. 2-11.

One of the biggest problems related to this architecture in addition to reduced network density and lower speed is the stability issue associated with the opera-
Figure 2-10: Transfer characteristics of a 7-bit synapse for weight values of 0 +/-1, 3, 7, 15, 31, 63 (a) with long channel transistors and (b) with external 10MΩ for voltage to current conversion. Adapted from [38].

Figure 2-11: Electronic circuit neuron model and synapses. Where $V_i$ is the output of the neuron i, $T_{ij}$ is the synaptic weight from neuron j to neuron i, $R_f$ and $C_f$ are the feedback capacitance and resistance of the neuron, and $I_i$ is the external input current. $R_F$ and $C_F$ should be chosen to ensure stability against oscillations. Adapted from [38].

ational amplifier. As a result, this architecture is not a viable choice for processor architectures such as COIN. Other previous related work includes Travis’s electronic thresholding circuit[56] and a comparator based activation circuit designed for artificial vision systems[63]
Chapter 3

Mathematical Models of Different Types of Activation functions

The objective of this chapter is to develop the mathematical models of neural activation functions needed to accomplish bistability in neural networks. Artificial neural networks have developed rapidly in the past few decades both in theory and in hardware implementations. Developing the mathematics behind will give us a flexibility in designing and implementing the various types of neural activation functions in optoelectronic hardware. Activation functions of neural networking can be continuous, i.e sigmoidal functions, piece-wise linear or isosceles triangular functions. Continuous functions are the preferred squashing functions for neural networks for their non-linearity and the requirement of derivative based parameter methods such as back-propagation [9]. One example of thresholding functions is the well known sigmoidal function. Sigmoidal output functions are good squashing functions for unbounded activation[15]. Other examples include the hyperbolic tangent function and the arc-tangent function which can be used to accomplish both positive and negative unbounded activation.

1Thresholding function is also known as the activation function, squashing function or weighting function. In some literature, it is also called the neural transfer function. They all refer to the same term and can be used interchangeably in this thesis.
First, we will explore the properties of the generalized activation function such as their mean square derivation and derivatives. Then we will manipulate the generalized thresholding function as needed to derive the three other types of functions. The reason for exploring these different types of neural thresholding functions is to find an intuitive and simple optoelectronic implementation techniques. The implementation of these functions in optoelectronic circuits will be explored in chapters 4 and 5.

### 3.1 Generalized Thresholding Function

The most general form of the activation function is given by the following mathematical equation.

\[
I_o(I_I) = C + \frac{K - C}{(1 + De^{-B(I_I - F)})^{(\frac{1}{v})}} \tag{3.1}
\]

where \(I_o\) and \(I_I\) are the output and the input intensities of the activation function respectively, \(K\) is the upper asymptote, \(C\) is the lower asymptote, \(B\) is the rise speed, \(F\) is the horizontal shift (electrical bias), \(D\) is the initial output intensity which depends on the initial intensity \(I_o(0)\) and \(v\) is a constant which affects the output near the maximum asymptote speed occurs (the mid point). For example if \(C=-1, D=0.5, F=3, B=2\) and \(v=0.5\), then the equation will look like

\[
I_o(I_I) = -1 + \frac{2}{(1 + 0.5e^{-2(I_I - 3)})\left(\frac{1}{0.5}\right)} \tag{3.2}
\]

The plots depicted in Fig. 3-1 show the dynamics of the generalized thresholding function for various constants.

Furthermore, one of the most important characteristics of Artificial Neural Networks (ANNs) is their ability to learn. While traditional logic based processors are programmed with explicit instructions on how to perform particular tasks, ANNs are trained. ANNs are trained with different input/output data sets until acceptable error level has been achieved. After that, the networks can be presented with inputs.
that have unknown outputs and trained easily. Thus, the error presented by the thresholding function is important in accomplishing acceptable levels of error during training. In this case, the error can be calculated by taking the derivative of the generalized thresholding function with respect to the input signal ($I_I$).

Other specific thresholding functions can be derived from the generalized squashing function i.e hyperbolic tangent functions, arc-tangent functions and the most commonly used sigmoidal function.

### 3.2 The Sigmoidal Function

A sigmoid function produces a curve with an "S" shape. The most commonly used sigmoidal function for neural networking takes the following form:

$$I_c(I_I) = \frac{1}{1 + e^{-\frac{C(I_I-B)}{M}}}$$

where B represents a threshold which shifts the function laterally, M is a parameter which determines the slope of the function. The range of this function is between 0 and 1. If M approaches 0, the function becomes a unit function. By adjusting the rise speed, $\frac{C}{M}$, and the bias, B, we should be able to choose a sigmoidal function to meet specific design requirements. The rise speed determines the slope of the function.
Figure 3-2: a) Plots for different values of the rise speed \( \frac{C}{M} \). b) The sigmoidal function for three different bias values \( B \). The bias value in the optoelectronic activation corresponds to the turn on voltage of the transistor which corresponds to the gain in hardware implemented systems. The value of the rise speed is an important parameter in determining the responsiveness of the system such as the brightness of the output LED for a given input light intensity. As it can be seen from Fig-3-2, the sigmoid squashing function can be used only for positive activation. The error corresponding to the sigmoid function can be evaluated by taking the derivative with respect to the input as shown in Equation-3.4

\[
I'(I) = \frac{Ce^{\frac{C(I-B)}{M}}}{1 + e^{\frac{-C(I-B)}{M}}}
\]  

(3.4)

The sigmoid activation function type is implemented in an optoelectronic circuit in chapters 4 and 5 (versions I and II designs of the optoelectronic neuron. Version-II design also incorporates local memory and weighting circuits). Furthermore, a 9x9 neurons array of this activation function is implemented in a printed circuit board (PCB) which will be discussed in the forthcoming chapters.

### 3.3 The Arc Tangent Activation Function

Other possible class of neural activation function is the arc-tangent activation function. The input-output relationship for the arc-tangent function is given as follows:

\[
I_\theta(I) = \text{arctan}(C(I - B))
\]  

(3.5)
where \( C \) is the rise speed and \( B \) is the bias value. The error corresponding to the derivative of this function is given as follows.

\[
\dot{I}_{o} = \frac{C(X - B)}{C(X - B)^2 + 1} \tag{3.6}
\]

Figure 3-3: a) A plot of arc-tangent function for three different lateral shift (zero crossing) values. The zero crossing values specified by \( B \) are 2, 3 and 4. b) The response of arc-tangent thresholding function at fixed bias point \( B \) for three different values of rise speed (\( C \)). The value of \( B \) is 3 and the rise speeds for each curve are specified in the equations.

### 3.4 The Hyperbolic Tangent Activation Function

The hyperbolic tangent function is another useful function for the activation of artificial neural networks. The formula is given by:

\[
I_o(I_t) = \frac{A[1 - e^{-C(I_t - B)}]}{[1 + e^{-C(I_t - B)}]} = \tanh(C(I_t - B)) \tag{3.7}
\]

where \( C \) corresponds to the rise speed and \( B \) corresponds to the lateral shift or the zero crossing point. The derivative (error) of the hyperbolic tangent function is given as follows:

\[
\dot{I}_o(I_t) = \frac{B}{2A}[A + I_o(I_t)][A - I_o(I_t)] \tag{3.8}
\]

This type of activation function was used to implement the optoelectronic neuron.
Figure 3-4: a) The plot of hyperbolic tangent function three different lateral shift values (B). The lateral shift (B) corresponds to the zero crossing or the thresholding point in this function. b) A plot of hyperbolic tangent function for three different values of rise speed (C).

discussed in chapter-5 (version-III design of the optoelectronic neuron). Both the hyperbolic tangent and arc tangent function have similar characteristics to the sigmoid function except the difference in the output range.

### 3.5 Comparison of Neural Activation Functions

All these functions are bounded, continuous, monotonic and continuously differentiable. These functions however have different output ranges. In addition to this difference, the three functions have different error values which is very important in determining the convergence speed of the activation function during training. For input intensity values between -1.5 and 1.5, the derivatives of the hyperbolic tangent and the arc-tangent functions change faster than the derivative of the sigmoidal function. Compared to the derivative of the arc-tangent function, the derivative of the hyperbolic tangent function is faster for the same ranges of inputs. Therefore, the correction of error in the hyperbolic tangent activation function is faster and thus the convergence is faster. However, in hardware implementation, the hyperbolic tangent function is a bit more resource intensive than its sigmoid counterpart.
Figure 3-5: a) Plots of the three thresholding functions; sigmoidal, hyperbolic tangent and arc-tangent functions. The sigmoid function goes from 0 to 1 and the arc-tangent and hyperbolic tangent functions go from -1 to 1. b) The derivatives of the three functions which is the same as the error. We can compare the convergence speed of the error associated with each function from the plot.
Chapter 4

Activation Function Circuit for COIN: Version-I

This chapter focuses on the version-I design of the optoelectronic neuron activation function. This version of the electronic circuits for the COIN processor is based on the block diagram given in Fig. 5-2 and it implements fixed thresholding.

![Figure 4-1: Block Diagram of one pixel of the COIN Co-processor for Version-I Optoelectronic neuron circuit: Interconnects driven by one light source per neuron and modulated by a Spatial Light Modulator (SLM)](image)

The implementation of the unit pixel for the proposed architecture is shown in Fig. 5-2, which also includes a spatial light modulator (optical weighting element) that is supervised by a PC controller or a microcontroller, and an optical interconnection
element. The pixel consists of ten synapses, a photodetector, an activation circuit, a spatial light modulator (SLM) and an LED.

Depicted in Fig. 4-2 is a simplified signal flow diagram between layers implemented with all optoelectronic neurons. Between each layer there is a spatial light modulator that performs weighting of the output optical signal and an optical interconnection element that guides the output of each neuron to subsequent layers of neurons. As shown in the data flow diagram in Fig. 4-2, each neuron receives inputs from nine neurons in the previous plane and transmits its outputs to nine neurons in the succeeding plane. All neurons implement a sigmoidal type of activation function which increases computational power and inter-plane communications are done optically. Each neuron consists of a photodetector which sums the incident optical signals from previous plane and produces a photocurrent proportional to the amount of incident light that falls on it. An electronic circuit then converts this photocurrent to a voltage which can be used to drive output optical sources (i.e.LED). The output is then distributed to an array of synapses in the next layer. Each synapse serves as a link to a neuron in the next layer and can weight the signal passing through it with spatial light modulators appropriately.

Figure 4-2: A signal flow diagram between layers of neurons. Adopted from [56]
The proposed COIN architecture is implemented as shown in Fig. 4-3. Each neuron in the COIN is identical such that the neural processor consists of identical optoelectronic circuits. The output of the neuron in each pixel is given by

\[ x' = \sum_{i}^{k} f(x_i)w_i \]  \hspace{1cm} (4.1)

where \( w_i \) and \( x_i \) are the weights and inputs to a given pixel, \( f \) is the activation function, and \( x' \) is the output (which becomes the input to the next layer). The sum is taken over the \( k \) inputs to a given neuron. By defining the synapse output as

\[ y_i = \sum_{i=1}^{k} f(x_i)w_i \]  \hspace{1cm} (4.2)

(4.1) can be rewritten as

\[ x' = \sum_{i}^{k} f(x_i)w_i \]  \hspace{1cm} (4.3)

![Figure 4-3: The inputs and outputs of a single neuron are represented by the central circle. Each rectangle associated with some weight represents a synapse, such that a neuron and a succeeding synapse comprises one pixel of the proposed COIN Co-processor. The preceding synapses are also associated with pixels in the preceding neuron layer.](image)

Equations (4.1) and (4.2) provide the mathematical basis for the synapse and the neuron respectively. The next section will discuss the design and implementation of a synapse and a neuron at the transistor level. Furthermore, characterization results of the individual components will be presented and discussed.
4.1 Implementation of Optoelectronic Neuron for Version-I

This section will present the design and characterization of the optoelectronic circuits required to implement the proposed architecture. Fig. 4-4 shows the low-level block diagram of the circuit implementation of an optoelectronic neuron. The input-output characteristics of the optoelectronic neuron is approximated by an activation function whose output remains zero until the input exceeds the predetermined threshold value. Beyond this threshold value, the output saturates. This input-output characteristic can easily be implemented using electronic components but since the neuron has to have optical input and optical output, additional elements, such as photodetectors and output light sources must be incorporated to obtain a complete optoelectronic neuron. The gate voltage on the M2 is controlled by the input circuit which consists of a photodetector acting as an optical input port and a biasing transistor M1. The switching characteristics of the neuron circuit is determined by the I-V characteristics of the photodetector and M1. The input light intensity is summed and converted to current ($i_{in}$) by the photodetector which is proportional to the total input light intensity. This current, $i_{in}$, is then converted to voltage by M1 to regulate the switching characteristics of M2. Finally, the LED in the output stage converts the current back to light and propagates it to the next layers of neurons. M3 which is a voltage controlled MOS resistor is used to protect the LED from burning out from over-current injection.

The Optical Input Detector of Fig. 4-4 is composed of a darlington phototransistor pair which can also be replaced with a photodiode. However, the phototransistor is more advantageous compared to the photodiode for a few reasons. Firstly, phototransistors have a high current capacity (collector current can be in the range of milliamperes, whereas the current in photodiodes is usually in the range of microamperes). Secondly, phototransistors combine the properties of photodetection and amplification in one device. With the injection of optically generated carriers in the base, it can be used to produce amplification which is further enhanced by the
Figure 4-4: An Optoelectronic Neuron Circuit: VDD is supply voltage, VB1 and VB2 are bias voltages for transistors M1 and M3 respectively. The voltages are generated such that the two transistors operate in triode region. These two transistors can also be replaced with a fixed resistor photodarlington pair. Furthermore, they exhibit a strongly non-linear response which is highly desired for this type of activation function.

We know that the collector current of a BJT is given by

\[ I_C = \alpha I_E + I_{CO} \]  

(4.4)

where \( I_C \) = collector current, \( I_E \) = emitter current, \( \alpha = \frac{I_C}{I_E} \), \( I_{CO} \) = collector to base reverse saturation current. \( I_{CO} \) is produced due to EHPs generated as a result of thermal energy present in the atmosphere. This implies that \( I_{CO} \) will vary if the transistor is subjected to optical illumination, because of EHP production due to the incident optical energy. Now, let’s reconsider (4.4). Due to the additional EHP
current, this equation can be modified as

\[ I_C = \frac{\beta}{\beta + 1} (I_C + I_B) + I_{CO} \]  \hspace{1cm} (4.5)

Rearranging (4.5), we get

\[ I_C = \beta I_B + (1 + \beta)I_{CO} \]  \hspace{1cm} (4.6)

If we write,

\[ I_{CEO} = (1 + \beta)I_{CO} \]  \hspace{1cm} (4.7)

where \( I_{CEO} \) = collector-to-emitter reverse saturation current with base open, \( \beta \) base-to-collector current gain and \( I_B \) = base current, then (4.6) can be rewritten as

\[ I_C = \beta I_B + I_{CEO} \]  \hspace{1cm} (4.8)

As we can see from Fig. 4-4, for the phototransistor, there is no base connection instead the base-emitter junction when illuminated with light will produce EHPs and this produces the base current. Consequently, the collector current \( I_C \) in the case of the phototransistor is not the emitter-injected component \( aI_E \) but is entirely due to the optically generated leakage component \( I_{CEO} \). Thus, the collector current can be expressed as

\[ I_{CEO} = (1 + \beta)I_{CO} \]  \hspace{1cm} (4.9)

where \( I_B \) = current generated in the base as a result of optical irradiation.

The photodarlington transistor pair is a combination of a common-emitter stage driving another common emitter transistor stage such that there is only one emitter, one collector and one base for the compound structure as shown in Fig. 4-5. For the pair transistors, the emitter current of the first transistor current is the base current of the second transistor. Therefore, the emitter current of the second transistor is given by

\[ I_{E2} = (1 + \beta^2)I_{B2} \]  \hspace{1cm} (4.10)
Figure 4-5: Photodarlington transistor pair. PT1= Phototransistor1 and PT2= Phototransistor2

Similarly, the emitter current of the first transistor is given by

\[ I_{E1} = (1 + \beta 1)I_{B1} \quad (4.11) \]

Finally, combining (4.10) and (4.11), the emitter current of the PT2 in terms of the base current of PT1 is:

\[ I_{E2} = (1 + \beta 1)(1 + \beta 2)I_{B1} = \beta^2 I_{B} \quad (4.12) \]

If the transistors are matched, we have \( \beta 1 = \beta 2 = \beta \) and \( I_{B} \) = photogenerated base current of PT1. (4.12) is modified by replacing the emitter current \( I_{E2} \) with the reverse saturation current \( I_{CEO} \) to produce the equation

\[ I_{CEO} = (1 + \beta 1)(1 + \beta 2)I_{B1} = \beta^2 I_{B} \quad (4.13) \]

Thus, the photodarlington pair produces more current gain than a simple phototransistor. However, due to large carrier storage the switching time of the photodarlington transistor is significantly slower than the transistor/photodiode counterpart [14].
**The Activation Circuit** of Fig. 4-4 performs the required non-linear mapping from the detector to the output light sources. Transistor M2 will not be turned on until the gate voltage, which directly depends on the product of the gate resistance $R_{GS}$ and the photogenerated current, $i_{in}$, has exceeded the threshold voltage. After which, the transistor amplifies the signal received to produce an output current that drives the LED. This process continues until the transistor saturates which also causes the neuron to saturate. Transistor M2 acts as the driving element for the output light source and provides electrical voltage gain. Its non-linearity along with the exponential characteristic of the photodetector provides the required sigmoid activation function for the neuron. The three operating regimes for this transistor are given in (4.14)

\[
I_{DS} = 0 \text{ if } V_{GS} < V_T \text{ cut Off} \\
I_{DS} = \mu C_{ox} \frac{W}{L} ((V_{GS} - V_T)V_{DS} - \frac{V_{DS}}{2})(1 + \lambda V_{DS}) \text{ triode} \\
I_{DS} = \mu C_{ox} \frac{W}{2L} (V_{GS} - V_T)^2(1 + \lambda V_{DS}) \text{ saturation} \\
\]

where $\mu C_{ox}$ is the physical parameter of the device, $\frac{W}{L}$ the ratio of the channel width to the channel length, $V_{GS}$ is the gate-to-source voltage, $V_{DS}$ is the drain-to-source voltage, $V_{DS}$ is the drain-to-source voltage and $V_{TH}$ is the threshold voltage of the MOSFET. The MOSFET is assumed to be off below the threshold voltage according to the conventional MOSFET theory.

\[
g_m = \sqrt{2I_{DS}\mu C_{ox} \frac{W}{L}} = \mu C_{ox}(V_{GS} - V_T) = \frac{2I_{DS}}{V_{GS} - V_T} \\
r_o = \frac{1}{\lambda I_{DS}}
\]

where $g_m$ is the transconductance and $r_o$ is output resistance of the MOSFET looking
down from the drain-end. For a common source configuration without back-gate effect (bulk is connected to the source) the gain, the input impedance and the output impedance are calculated as follows.

\[ R_{in} = \infty \]  
\[ R_o = r_o || R_{Load} \]  
\[ A_V = -g_m R_o \] Voltage Gain

Transistors M1 and M3 are basically resistors implemented in MOS transistors. MOS transistors based resistors are better because they occupy a smaller area and also can pass a large current with a huge resistance without a large drop in voltage compared to ordinary resistors [2]. The MOSFET has to be operated in a triode region in order to be used as a resistor. Fig. 4-6 shows the biasing structure of the transistor. In the triode region, the drain current is related to the MOS terminal voltages as

\[ I_{DS} = \mu C_{ox} \frac{W}{L} ((V_{GS} - V_{TH})V_{DS} - 0.5V_{DS}) \]  

In order it to work as a resistor the current should be proportional to the drain-to-source voltage. In (4.26), non-linearity is introduced due to the square dependence of the current on the drain-to-source voltage. This second order term can be neglected.
if \((V_{GS} - V_{TH})V_{DS} \gg 0.5V_{DS}\). With this condition satisfied, the current equation can be re-expressed as

\[
I_{DS} = \mu C_{ox} \frac{W}{L} (V_{GS} - V_{TH}) V_{DS}
\]

(4.25)

giving an effective resistance of

\[
R = \frac{V_{DS}}{I_{DS}} = \frac{1}{\mu C_{ox} \frac{W}{L} (V_{GS} - V_{TH})}
\]

(4.26)

This function is a voltage controlled resistor where the controlled voltage is the gate-to-source voltage. For a bigger over-drive voltage\((V_{GS} - V_{TH})\), MOS resistor retains its linearity over a wide range of the drain-to-source voltage. The only disadvantage of the MOS resistor is that it requires a gate bias voltage which can be problematic when a large number of neurons are required. These two transistors were replaced by fixed value resistors for the 9x9 array of neurons implemented in a PCB as discussed in section 4.3.

The Sources and Driver Circuit of Fig. 4-4 is composed of the output light source usually LED and a driver circuit consisted of either a MOS resistor or a fixed resistor to regulate the current flowing through the LED. It forms a link from electrical domain to optical domain where the LED produces an illumination to be broadcasted to subsequent stages.

The structure of this type of logic processing optoelectronic device is simple, has good reliability, and most importantly, it is compatible with current very large scale electronics (VLSI). It is capable of carrying out optoelectronic hybrid integration. Furthermore, the only electrical connection would be the power and the global biases for the two MOSFETs. As a result, there is promising hope for applications in optoelectronic computing as well as hybrid hardware neural processors.

A proof of concept PCB was designed and tested for the optoelectronic neuron.
discussed above. In the following subsections, lab characterization results and derivations of equations for each component will be presented based on lab measurement results. Both the data collected from the lab and least square fitted curves will be presented side-by-side for comparison. Then, we will combine the individual equations and results to evaluate constants related to the sigmoid function of the optoelectronic neuron. Finally, we will present the PCB design and measurement results for the 9x9 array of optoelectronic neurons.

4.1.1 LED Characterization and Selection

The multilayer COIN co-processor will have large number of neurons and interconnection elements. Consequently, a large number of light sources and drivers will be required. There are a number of constraints that will determine the selection of the light emitting source and the drivers but the two important factors are physical size and electrical characteristics. In terms of size, they should be compact and small enough to be integrated into a neuron. Regarding the electrical characteristics, the light sources should be energy efficient to allow low power operation and the driving elements should be electronically programmable to allow changing connection weights.

In this characterization, an off-the-shelf, 5MM white LED was used to make measurements. The experimental setup is shown in Fig. 4-9a. A potentiometer resistor, $R_{pot}$, was used to control the amount of current flowing through the LED. The relative output intensity was measured using ThorLabs S130A Silicon Power Meter Optical Head[58] light detector which was placed 4.5cm away from the LED. This experiment was performed in a completely dark room to minimize the effect of ambient light from other sources. Furthermore, extra care was taken to avoid the saturation of the detector. The saturation power of the detector was set to 500mW at a wavelength of 530nm which is much bigger than all measured powers. The detector measures the power that falls on the active sensor area. In order to calculate the relative intensity, the power must be divided by the size of the active sensor area which is 10mm x 10mm. Since the LED light source has non-constant beam divergence, it is hardly
possible to get all the light onto the active area. But the region of maximum beam divergence and its surroundings were focused onto the photodetector for all experiments to measure the power. The collected data is plotted as shown in Fig. 4-9b. Furthermore, this data was fit into a second order polynomial equation using Matlab's LED Characterization Curve Fitting Toolbox™. The relationship between the relative output intensity and the input current is given by

\[ I_{\text{out}}(i) = -1.518 \times i^2 + 0.277 \times i - 4.578 \times 10^{-5} \quad \text{second order polynomial fit (4.27)} \]

where \( I_{\text{out}} \) is the relative output intensity of the LED and \( i \) is the current flowing through the LED. This equation is valid only for range of values given in this lab measurement. As it can be inferred from the plot in Fig. 4-9, the relative output intensity increases proportionally with the forward input current. However, the relative intensity vs. input current curve falls off towards the upper limit of the forward input current.

In conclusion, although this experiment was performed using ordinary off-the-shelf LED components, for the full and integrated embodiment of the COIN Co-
processor we plan to use other types of light sources such as OLEDs (organic light emitting diodes) since they have high quantum efficiency, low operating voltage and visible luminescence besides their ability to be integrated with silicon substrates easily [12],[13], [62].

4.1.2 MOSFET and LED Modeling and Characterization

In this subsection, a combination of LED and MOSFET is characterized. The gate voltage of the MOSFET was varied from 0 to VDD to measure the current and the relative output intensity of the LED.

![MOSFET and LED characterization circuit](image)

Figure 4-8: (a) MOSFET and LED characterization circuit, and (b) a plot of gate voltage ($V_{Gate}$) vs. Relative output intensity

The plots of the measured results for the relative output intensity and output LED current as a function of gate voltage ($V_{Gate}$) resemble the sigmoid activation function of the form $f(x) = \frac{a}{1+e^{-b(x-c)}}$ which was discussed in chapter ???. Where a, b and c
are real constants. Thus, the data was fit into a sigmoid curve to extract the values of the real constants using Matlab’s Curve Fitting Toolbox.

\[
I_{out}(V_{Gate}) = \frac{0.0082}{1 + e^{-10.48(V_{Gate}-2.534)}} \quad \text{sigmoid fit of intensity as a function of gate voltage} \\
\tag{4.28}
\]

\[
i_{out}(V_{Gate}) = \frac{0.0038}{1 + e^{-10.59(V_{Gate}-2.56)}} \quad \text{sigmoid fit of output LED current as a function of gate voltage} \\
\tag{4.29}
\]

By combining the LED and the MOSFET, we are able to form an electrical-optical link where the non-linear electrical component was able to respond to a varying input.
parameter in a neural activation function fashion. The non-linear response was then fed into an LED and a thresholded optical intensity was obtained as an output. This output intensity can be weighted and distributed to subsequent layers of neurons using the optical interconnection element. Since the input is also in the form of light, we have to design a simple light input-current output circuit using phototransistors. The current can be converted to voltage with a resistor and fed into the MOSFET for decision and thresholding.

4.1.3 Phototransistor Modeling and Characterization

In this subsection, lab characterization results for the phototransistor that is used to detect and sum incoming lights from previous layer neurons and convert the sum into current with a current gain($A_i$) is presented. The light generated by the LED affects the base region of the phototransistor to generate a current proportional to the the intensity at the base of the phototransistor. The configuration of the phototransistor is common collector, thus the output current is just the buffered version of the input current formed in the base region. Again, $R_{opt}$ is a potentiometer resistor that is used to vary the current flowing through the LED. The phototransistor used in this experiment was a discrete, 5mm, BPW17N high radiant silicon NPN phototransistor.

The phototransistor transforms light received from the input LED into current which would be used to drive current based devices such as BJT or gate voltage based devices such as MOSFET, MESFET or JFET. Examination of the relative input intensity and the phototransistor emitter current plot shows that there is a tendency for the phototransistor’s current to saturate at a higher input illumination. We were careful not to saturate the phototransistor. Otherwise, it will not be able to perform proper summation of the input intensity illuminations from previous layer neurons in the actual COIN processor once it is saturated. The saturation current can be increased by using the photodarlington pair instead of a single phototransistor. The minimum and maximum input intensities are determined by the sensitivity/noise floor and the saturation currents of the phototransistor respectively. According to the
collected data, the phototransistor shows a linear behavior up to a relatively low photocurrent $I_{max}$ ($15\mu A$), followed by a very rapid saturation. For the COIN processor since there are 9 synaptic inputs from previous layer networks, the maximum intensity produced by each synapse needs to be less than $\frac{1}{9}I_{max}$ and the minimum intensity needs to be greater than the noise current of the phototransistor.

Moreover, to calculate the intensity that is falling on the phototransistor we have to know the area of the phototransistor where the light falls on. In this case, the phototransistor is overfilled and the area of the detector region can be calculated as given in [44]. Important information for the calculation of the area such as the half angle, radius of the sphere and diameter of the phototransistor base circles are
obtained from the data-sheet for BPW17N[55]. Furthermore, the measured data was fitted into a second order polynomial using the Matlab Curve Fitting toolbox. The emitter current of the phototransistor \( i_{out} \) is related to the input intensity \( I_{in} \) as follows.

\[
i_{out} = -1.646I_{in}^2 + 0.00945I_{in} + 5.112 \times 10^{-7}
\]

second order polynomial fit \( (4.30) \)

The second order polynomial fit was done only for the non-saturated region of the phototransistor since that is the only region of interest for this application. Now that we are able to convert an incoming optical input into electrical signal, we can combine the circuit discussed in this section and in section 4.1.2 to realize the optoelectronic neuron that performs thresholding and weighting of an input illumination to produce optical output.

### 4.2 Version-I Optoelectronic Neuron Circuit

The circuit components discussed in the previous sections are assembled to form a light in-light out optoelectronic neuron. The version-I design given in Fig. 4-4 has been slightly modified by replacing M1 and M3 with fixed resistors \( R_{GS} \) and \( R_{DS} \) respectively. Replacing these two transistors with fixed resistors eliminates the need for two biasing voltages per neuron.

The performance of this unit neuron circuit was tested in the lab. The two quantities that are easily measurable with multimeter and powermeter are the emitter current of the phototransistor and the current through the output LED, and the output intensity respectively. The measured emitter current of the phototransistor and the relative output intensity of the optoelectronic neuron can be fitted into a sigmoidal function of the form discussed in chapter-3.

\[
I_{out}(i_{in}) = \frac{1}{1 + e^{-\alpha(f(V(i_{in}))}}
\]

\( (4.31) \)
Figure 4-11: An Optoelectronic Neuron Test Circuit. The input light source was placed 4.5 cm away from the phototransistor and at the angle of maximum-beam divergence of the LED.

where $\alpha$ represents the gain of the MOSFET, $I_{out}$ the relative output intensity of the output LED, $i_{in}$ the emitter current of the phototransistor and $f(V(i_{in}))$ the sigmoidally transformed input light signal by the non-linear electronic device.

The sigmoidal fit was achieved using a function of the form $y = \frac{a}{1+e^{k(x-x_0)}}$. The following are the values of the constants obtained after performing sigmoidal fit using the Matlab Curve Fitting Toolbox. $x_0 = 0.024$, $k = -340.38$, $a = .0048$. Thus, the final sigmoid equation for the optoelectronic neuron shown in Fig. 4-11 is given as follows

$$y = \frac{0.0048}{1 + e^{-340.38(x-0.0237)}} \quad (4.32)$$

$y$ represents the relative output intensity and $x$ represents the emitter current of the phototransistor.
Figure 4-12: Relative output intensity ($I_{out}$) vs. Emitter current of the phototransistor ($I_{in}$). Raw data is shown in dotted lines and the sigmoidal fit is shown in solid lines. The sigmoidal fit was achieved using a function of the form $y = \frac{a}{1 + e^{k(x-x_0)}}$

4.3 The 9x9 Array of Optoelectronic Neuron Implementation in Discrete Components

A 9-by-9 array of optoelectronic neurons was designed and tested to understand the performance of the fixed threshold activation function using off-the-shelf discrete components. The unit neuron discussed in the previous subsection in Fig. 4-11 was used to build the 9X9 array of neurons. The back and front picture of the 2-D optoelectronic neuron array is shown in Figures. 4-13 and 4-14 respectively. The PCB is double sided and it is designed to house the phototransistors in the back plane and all the other components such as the resistors, the MOSFETS and the LEDs in the front plane. By doing so, we are able to isolate the input layer from the output layer and thus lights generated from the output LEDs will not affect the
response of the neurons.

Figure 4-13: 9x9 Optoelectronic Neuron Array PCB front. The front plane consists of arrays of output LEDs and MOSFETs

Figure 4-14: 9x9 Optoelectronic Neuron Array PCB back showing the phototransistors array and resistors.

4.3.1 Result

The PCB fabricated is a 9x9 array of optoelectronic neurons for neural network applications. This PCB contains 81 neurons arranged in a square board size of 4.5x4.5 \( \text{in}^2 \). The power consumed by a single neuron is about 17mW and the total power consumption for the 81 neurons, when all of them are fully turned on, is about 1.377W (81*17mW). The performance of the optoelectronic neuron array was tested with various inputs such as numbers and letters. The number 5 was presented as one of the inputs to the neurons. The input was generated by shining an unpolarized light onto a transparent film and projecting it onto the back of the PCB (to the phototransistor arrays). Two thin-film linear polarizers were placed in front of the transparent film to vary the brightness. The brightness can be adjusted as required by changing the polarization angles of the polarizers with respect to each other (the relationship between the input and the output intensities of the polarizers can be
calculated by applying the famous Law of Malus [30]). Since the phototransistors have a very low saturation intensity, the input to a given neuron needs to be lower than the saturation intensity of the phototransistor. Using the polarizers allows us to adjust the intensity falling on the phototransistors manually. Depicted in Fig. 4-15 is the maximum input intensity when the polarizers are in-phase and shown in Fig. 4-16 is the output to that input. Furthermore, the LEDs used in this experiment are not uniform and matched. As a result, we see varying intensities for the same level of input illumination.

Figure 4-15: Input of 5. 5 is presented as an input to the 2-D array of optoelectronic neurons. The brightness is adjusted with the polarizers

Figure 4-16: Output of 5. The output is generated by activating the corresponding neurons in the 2-D array of optoelectronic neurons

4.3.2 Discussion

In this proof of concept PCB, the size of the array is big and as a result it is not ideal for integrated circuit implementation. But this PCB along with the optical interconnection element, a PC/microcontroller for weight storage as well as other input and output display devices can be used to build a shoe-size multi-layer and feed-
forward optoelectronic neural processor. In addition to this, the power consumption for this PCB is high because we are using big and inefficient off-the-shelf discrete components with high voltage supplies. Nevertheless, this system has a potentially large advantage overall with regards to electronic and all-optics implementations in order to scale up the number of pixel neurons. The issues mentioned above will be mitigated when we fabricate the integrated circuit (IC) version of the neurons with lower supply voltage and compatible CMOS process. This work will serve as a precursor for the integrated version of the neurons.
Chapter 5

Improved Circuit Architectures for COIN: Versions-II and III

In the previous chapter, we discussed one possible architecture of the COIN processor and the implementation of an optoelectronic neuron circuit for the proposed architecture. This chapter will present alternative COIN processor architectures that can lead to an improved version of the optoelectronic neuron circuit. Versions-II and III of the electronic circuits of COIN will be discussed in this chapter. These versions of the optoelectronic implementations eliminate the need for the use of the SLM which acts as the interconnection weight matrix sandwiched between layers of networks. Version-II is an improvement of version-I, which was discussed in chapter-4, with the addition of local storage element. Version-III embodies a new topology of activation circuit that can be used to adjust the the thresholding voltage of the activation function. The SLMs which can be controlled using electrodes can be used to implement the weights optically. The elimination of the SLM from the COIN system has two advantages; reduced optical cross-talk and reduced overall size of the system.

Furthermore, local storage of synaptic weights is a central issue in all implementations of learning and adaptation of neural networks. In the new versions of the optoelectronic neuron activation circuits we plan to incorporate the local memory circuit to enable on chip learning and adaptations. Each storage cell should be very
dense since each synapse requires a storage element.

Based on the unique computation and communication requirements of the neural processor, the primary goals of the circuit designs are functionality, size and power consumption. Speed is much less of a concern because neural computing achieves its performance through massive parallelism rather execution time. Since our neural network is composed of simple neurons, maintaining a relatively simple circuit topology is also desirable.

5.1 COIN Model for the improved circuit Architecture

![COIN Block diagram of Versions-II and III electronic circuits: Interconnects driven by individual sources and modulated by controllable driver circuits](image)

The basic difference between the two COIN Co-processor architectures shown in Figures 5-1 and 5-2 is that the first architecture (shown in Fig. 5-2 and explored in chapter 4) uses the bulky optical SLM to perform weight adjustments and the second architecture (to be explored in this chapter and shown in Fig. 5-1) uses electrical circuit
to adjust weights. Furthermore, a single LED is required per pixel for the first design as opposed to one LED per synapse for the second architecture. Thus, nine times more LEDs are required for the second architecture but the SLM has been eliminated. The choice of the COIN design architecture has to be decided based on the required size, integreability, manufacturability and power consumption level.

5.2 Low-Level Block Diagram of a Pixel

A pixel consists of a neuron and nine synapses as shown in Fig. 5-3. Each synapse has a local (associative) memory element where the weights are stored in the form of voltages in capacitors. The phototransistor collects the input signals from previous stage neurons and converts the summed intensity into a photocurrent which subsequently gets thresholded by the activation function. In version-I, a single LED is used per neuron and the output of the LED was distributed to nine synapses using SLMs. In versions-II and III, every synapse has its own output light source or LED. Previously, SLM was used to modulate and distribute the output of a single neuron to 9 synapses. Now the distribution is done in the electrical domain in the form of current instead of optical intensity and is distributed to synapse using current mirrors. The output
of the LEDs (synapses) is the thresholded and weighted version of the input light intensity. The synapses are coupled to subsequent stage neurons through the optical interconnection elements.

Figure 5-3: A Pixel. $w_i$ corresponds to weights stored in local memory elements. $C_i$ corresponds to the optical interconnection element used for inter-plane communication.

### 5.3 Neural Weighting and Associative Memory Circuits

The most important specific problem faced in the analog implementation of neural networks is the storage of synaptic weights. In a Von Neumann computer, the storage and computation are divided between the machine’s main memory and its central processing unit. To do a specific task, computers carry out a set of instructions sequentially transporting data regularly from a storage memory to the CPU where computation is performed. Since the memory and the CPU are physically separated
data needs to be transferred regularly and thus consuming tremendous amount of power per computation. However, the circuits in the brain pack the memory, computation and communication parts into smaller modules that can process information locally but can communicate with each other easily and quickly which is a very energy efficient process.

The goal of this section is to come up with a circuit element that can remember a certain voltage and output that voltage for a brief period of time. There are several methods to realize weights electronically: digital, analog and fixed weights. Digital techniques use addressable registers to store and modify the weights. Analog techniques for creating modifiable weight connections include variable resistors, field emitting transistors (FETs) gate voltage control and capacitive storage methods. Fixed weights inhibit weight modification by allowing the values of the weights be set once. Weight adjustments can also be realized optically using non-linear optics and spatial light modulators (SLMs).

In hardware based neural processors, information about synapse weight must be stored within a chip and analog memories seem to be the best for this purpose[45]. In the electronic circuit of the COIN, each synapse is accompanied by a local short term analog memory. This memory is needed to hold on silicon information about the synapse during the learning process. In CMOS, analog weights can be conveniently stored in the form of charge or voltage on a capacitor. The stored weight charge is preserved when brought in contact with the gate of an MOS transistor, which serves as a buffer between weight storage and the implementation of the synaptic function. Fig. 5-4 shows the most basic way to store the synaptic weight as a voltage across a capacitor C. The value $V_{in}$ to be stored is sampled by means of a transistor operated as a switch controlled by the gate voltage $V_G$. The storage time is limited by the leakage current of the reverse-biased diffusion-to-bulk junction which discharges the storage capacitor C. The leakage current can be canceled if the local substrate potential $V_B$ would be exactly equal to the stored voltage V, provided the junction
does not receive any light. This is a good way to realize the local memory unit for the COIN through a voltage stored across a capacitor as the gate voltage of a MOS transistor. The transistor has a conductance that is proportional to the amount of voltage retained at the gate and the brightness of the output light source (i.e. LED) will directly depend on the conductance of the MOSFET.

![Diagram](image)

Figure 5-4: A Basic short term memory circuit implemented using sample-and-hold circuit

This memory circuit can be modeled as a simple RC circuit where the $R_{ON}$ is given by

$$\frac{1}{\mu C_{ox} \frac{w}{L} (V_{GS} - V_T)} = \frac{1}{\mu C_{ox} \frac{w}{L} (V_G - V_{in} - V_T)}$$

(5.1)

Thus, the acquisition or the tracking time constant is given by $\tau = R_{ON} C$. There is also a channel injection charge given by $q_{ch} = WLCOX(V_{GS} - V_T)$. When the MOSFET is turned off, the channel charge $q_{ch}$ must disappear, a portion of which may be injected on C depending on the fall time of $V_G$. Also, direct coupling of $V_G$ to C through the overlap capacitance causes additional error. A complimentary or dummy switch can be used to increase the output to the full input range and to mitigate the charge injection.

### 5.4 Optoelectronic Neuron Model for Version-II

Version-II implements fixed threshold activation function, weighting circuits and local memory elements in analog CMOS. A single pixel is composed of a phototransistor, an activation circuit, an analog memory and an output circuit. The circuit model for this version is very similar to that of version-I except for the additional analog...
weighting and memory circuits. The analog memory is implemented using a switch that is controlled globally and a memory capacitor \( C_{\text{memory}} \).

![Diagram](image)

Figure 5-5: Version-II Optoelectronic neuron model with capacitive local memory element. Weights are periodically updated from the stored weight elements. \( i_{\text{outx}} \) and \( J_{\text{outx}} \) correspond to the output current and relative output intensity of synapse x respectively. \( i_{\text{in}} \) is the photogenerated current of the pixel from external illumination.

### 5.5 Spice Simulation Results for version-II

A Spice simulation result for version-II optoelectronic neuron is presented below. Shown in Fig. 5-6 is the schematics of the neuron and 3-synapses circuits. A 0.25um TSMC CMOS process technology model is used to simulate the circuit in LTSpice and cadence. The sizes of the transistors are designed to reduce the power consumption and the area of the neuron. Regarding the size, the area of the 9x9 arrays of neurons should be less than 300\( \mu \)m X 300\( \mu \)m in order it to be compatible with the optical interconnection element designed by Herrington’s PhD thesis [23]. For this simulation the weight is realized through resistor dividers located at the gate of the output transistors. In the actual implementation, a capacitive memory element like the one
shown in 5-5 will be used. The timing of the memory switch will be regulated by a global memory element which has information about the weight of each neuron in the network. Shown in Fig. 5-7 is the plots of the output current vs. photogenerated input current for the three synapses. The amplitude of the output current is regulated by the voltage established by the voltage divider at the gate of the output MOSFET as shown in Fig. 5-6.

Figure 5-6: Version-II optoelectronic neuron: A schematic with three synapses used to obtain simulation plots

Figure 5-7: Plots for various weights. The three synapses have three different weights that are controlled by the voltage divider at the gate of the output MOSFET.
5.6 Optoelectronic neuron model for version-III

This section discusses the analog circuits that are used as the building blocks in the implementation of version-III optoelectronic neuron model for the COIN processor. Version-III implements a tunable activation function, weighting circuits and local memory elements in CMOS. First, the principle of subthreshold physics of CMOS transistors are discussed. Then, the circuit diagrams essential for the building blocks such as differential pair, current mirrors and operational transconductance amplifiers are discussed in detail. Finally, spice simulation result in 0.25um TSMC process technology for the complete optoelectronic neuron model is presented.

5.6.1 Subthreshold MOSFET Operation

In traditional analog CMOS circuits, the transistors are usually biased in strong inversion region, where the drain current is given by

\[ I_{DS} = \frac{W}{L} \mu C_{ox}(V_{GS} - V_{TH})^2(1 + \lambda V_{DS}) \]  

(5.2)

where \( \mu C_{ox} \) is the transistor's physical parameter, \( \frac{W}{L} \) is the ratio of the transistor channel width to channel length, \( V_{GS} \) is gate-to-source voltage, \( V_{TH} \) is the threshold voltage of the transistor, \( \lambda \) is a the inverse of the Early Voltage of the transistor, and \( V_{DS} \) is the drain-to-source voltage.

Subthreshold region also know as weak inversion region is present when the transistor is operated below its threshold voltage, where digital circuits are sometimes approximated as being off. In this region, the current levels are very small and the drain current varies exponentially with the gate voltage. The source referenced equation for this region where the bulk of the transistor is tied to the source is given by

\[ I_{DS} = k_x \frac{W}{L} e^{\frac{k_x V_{GS}}{\phi_t}} (1 - e^{\frac{-V_{DS}}{\phi_t}}) \]  

(5.3)

where \( k_x \) and \( K_s \) are fabrication dependent parameters and \( \phi_t \) is thermal voltage.
which is approximately 26 milliVolts at room temperature. For drain-to-source voltage greater than a few thermal voltage, the drain current is essentially independent of the drain-to-source voltage. Thus, Equation 5.4 can be simplified to

\[ I_{DS} = k_x \frac{W}{L} e^{\frac{V_{GS}}{\phi_t}} \]  

(5.4)

For the implementation of version-III optoelectronic neurons some of the transistors are biased in sub-threshold region to take advantage of the exponential relationship between voltage and current which resembles the actual biological neuron model in the brain.

5.6.2 Differential Pair

The most important circuit for this neuron is the differential pair shown in Fig. 5-8. Assuming large enough drain-to-source voltages, the currents \( I_1 \) and \( I_2 \) can be expressed as

\[ I_1 = k_x \frac{W_1}{L_1} e^{\frac{k_x(V_{G1}-V_s)}{\phi_t}} \]  

(5.5)

\[ I_2 = k_x \frac{W_2}{L_2} e^{\frac{k_x(V_{G2}-V_s)}{\phi_t}} \]  

(5.6)

The two MOS transistors, M1 and M2, and the phototransistor are interconnected in such a way that the currents flowing through all of them meet at a common point. This means, the sum of these two currents must be equal to the photocurrent \( I_{PD} \) through the photodetector transistor:

\[ I_{PD} = I_1 + I_2 = k_x e^{\frac{-k_xV_s}{\phi_t}} \left( \frac{W_1}{L_1} e^{\frac{k_xV_{G1}}{\phi_t}} + \frac{W_2}{L_2} e^{\frac{k_xV_{G2}}{\phi_t}} \right) \]  

(5.7)

Solving this equation for the node voltage \( V_s \) yields:

\[ e^{\frac{-k_xV_s}{\phi_t}} = \frac{I_{PD}}{k_x} \left( \frac{1}{\frac{W_1}{L_1} e^{\frac{k_xV_{G1}}{\phi_t}} + \frac{W_2}{L_2} e^{\frac{k_xV_{G2}}{\phi_t}}} \right) \]  

(5.8)
Substituting Equation 5.8 into Equations 5.5 and 5.6 yields expressions for the drain currents to the differential pair

\[
I_1 = I_{PD} \frac{W_1}{L_1} e^{\frac{k_s V_{G1}}{\phi_t}} \left( \frac{W_1}{L_1} e^{\frac{k_s V_{G1}}{\phi_t}} + \frac{W_2}{L_2} e^{\frac{k_s V_{G2}}{\phi_t}} \right) \tag{5.9}
\]

\[
I_2 = I_{PD} \frac{W_2}{L_2} e^{\frac{k_s V_{G2}}{\phi_t}} \left( \frac{W_1}{L_1} e^{\frac{k_s V_{G1}}{\phi_t}} + \frac{W_2}{L_2} e^{\frac{k_s V_{G2}}{\phi_t}} \right) \tag{5.10}
\]

This means, if \( V_{G1} \) is more positive than \( V_{G2} \) by many \( k_s/\phi_t \), the voltage \( V_s \) transistor M2 off so that all the current is steered into M1 and \( I_1 \approx I_{PD} \). The vice versa is also true when \( V_{G2} \) is more positive than \( V_{G1} \) such that \( I_2 \approx I_{PD} \).

### 5.6.3 Current Mirrors

The circuits shown in Fig. 5-9 are NMOS and PMOS current mirrors. In the NMOS current mirror, transistor M1 is diode connected such that \( V_{DS1} = V_{GS1} \) and it is also connected to the gate of transistor M2. Since both M1 and M2 share the same source, they also have the same \( V_{GS} \). If the transistors are matched well such that they have the same threshold voltage(\( V_{TH} \)) and same device physical parameters such as same
Figure 5-9: Current Mirrors

\[ \mu_1 C_{ox1} = \mu_2 C_{ox2} \text{ and matched drain-to-source voltages such as } V_{DS1} = V_{DS2}, \text{ then} \]

\[ I_{out} = I_{ref} \left( \frac{W_2}{L_2} \right) \left( \frac{W_1}{L_1} \right) \] (5.11)

Thus, the current mirror provides both mirroring and scaling of the reference current \(I_{ref}\) by the aspect ratios of the two transistors.

### 5.6.4 Operational Transconductance Amplifier (OTA)

The schematic of a simple operational transconductance amplifier (OTA) is shown in Fig. 5-10. The circuit consists of a differential pair loaded with NMOS current mirror. If the current mirrors have the same aspect ratios, then the current \(I_1\) is mirrored with unity gain to the other leg of the current mirror. The output current of the OTA is proportional to the difference between the two drain currents. This difference is

\[ I_1 - I_2 = I_{PD} \left( \frac{W_1}{L_1} e^{\frac{k_s V_{G1}}{\phi_t}} - \frac{W_2}{L_2} e^{\frac{k_s V_{G2}}{\phi_t}} \right) \] (5.12)

If the two pairs have the same \(\frac{W}{L}\) ratio, then

\[ I_{out} = I_1 - I_2 = I_{PD} \left( e^{\frac{k_s V_{G1}}{\phi_t}} - e^{\frac{k_s V_{G2}}{\phi_t}} \right) \] (5.13)
This equation, with some additional algebraic steps, can be further rearranged to a tanh(x) function and the result will take the following form.

\[ I_{out} = I_{PD} \tanh\left( \frac{1}{2} \frac{k_s (V_{G1} - V_{G2})}{\phi_t} \right) \]  

(5.14)

This function goes through the origin with unity slope, becomes +1 for large positive input arguments and becomes -1 for large negative input arguments. Furthermore, this equation shows that a potential difference \( V_{G1} - V_{G2} \) produces a non-linear result \( I_1 - I_2 \). The hyperbolic tangent function is just the symmetric sigmoid used in neural network training algorithms. The current out of this amplifier is plotted as a function of \( V_{G1} - V_{G2} \) in Fig. 5-11 for \( I_{PD} = 10\mu A \) and \( \frac{1}{k_s} = 1.3 \). The curve is very close to tanh as expected. The output current is linearly dependent on the input voltage difference for small voltage differences \( V_{G1} - V_{G2} \) and for big voltage differences, it begins to saturate to positive or negative \( I_{PD} \).

The transconductance \( g_m \) of the OTA which is the slope of the tanh at origin is given
and is proportional to the bias current $I_{PD}$. Other performance parameters such as the gain of the OTA, the voltage swing and output resistance can also be derived for this circuit. The discussion above illustrates how analog circuits and optoelectronic devices can be used to implement the main computations for hardware based neural network models. This includes integration of the optical input signal using a phototransistor, weighting of the signal with gate voltage controlled analog MOS devices and computation of the non-linear output.

### 5.6.5 The Optoelectronic Neuron Model of Version III

The tanh(x) function based activation circuit, also know as version-III, neuron diagram is given in Fig. 5-12. This optoelectronic neuron comprises of the OTA which includes the differential pair with the phototransistor current source, the current mirror, the analog memory circuit and the output light source.

### 5.7 Spice Simulation Results for version-III

Presented below is the spice simulation results for version III optoelectronic neuron model. The simulation was done using the 250nm TSMC CMOS process technology. Two plots are given below to show on chip learning capabilities by changing
Figure 5-12: Version-III Optoelectronic Neuron Model with capacitive local memory element. It is comprised of the activation circuit which contains the mirror legs and the input transistors, the current mirror for distributing current among all the synapses, the analog memory in order to enable local memory storage and the output circuit which is regulated by the analog memory element to output a weighted light intensity through the LED.

The weights dynamically and shifting the threshold voltages by adjusting the input transistors gate voltages relative to each other. Both of these behaviors are depicted in Figures 5-14 and 5-15.

Figure 5-13: Version III optoelectronic neuron: A schematic circuit for spice simulation using the 0.25μ process technology.
Figure 5-14: Spice Simulation results for variable thresholding circuit. The threshold value is controlled by the relative gate voltages of the OTA. Four plots are given for various combinations of V2 and V3.

Figure 5-15: Simulation for various weights. Not only the threshold value but also the weights can be adjusted in this circuit. This will enable on-chip learning in an integrated analog neural processors.

5.8 Recommendation for Integrated Circuit (IC) Layout

Unfortunately, due to time constraints, I was not able to do the layout in Cadence but below is a recommendation for laying out the components of the optoelectronic neurons. The simulation was done using the TSMC 0.25um CMOS process technology. The IC layout can also be performed using the same process technology in Cadence. Most of the circuits discussed in versions II and III optoelectronic neurons consist of OTA, current mirrors, driver circuits, capacitors, phototransistors and LEDs. In general this building blocks excluding the phototransistor and the LED can be laid out.
in cadence. Further care should be taken to ensure that the transistors in differential pairs and current mirrors are as matched as possible, using layout techniques such as interdigitation/interleaving. This layout technique will minimize the effect of process variations on the parameters of the transistors. But in general slight mismatches are tolerable and also make the circuits more life-like. The memory capacitors should also be laid out carefully to minimize errors that would arise due mismatches and charge leakage in the programming switches. A common centroid method can be used to layout the memory capacitors for a given pixel so that the variations in process are averaged out.

5.9 Limitations on Circuit Design and Neuron Density

The computation power of an optoelectronic neural network depends on the number of neurons on each layer and the number of interconnection an optical system can support. For a given optoelectronic system, the size of the neuron circuit is limited by the finite dimension of the optical lenses and the interconnection elements. Therefore, the higher the neuron density is on an array, the more neurons we can fit into the optical system, which means a better performance for the neural network. The optical interconnect elements and the electronic integrated circuit elements occupy a big area which places a geometric limit on the maximum neuron density that can be achieved. Furthermore, the required metal lines for transferring signal and voltages must be taken into account. For a density higher than $10^4$ neurons/cm$^2$, each pixel has an area of less than $100 \times 100 \mu m^2$. Therefore, we must use the simplest circuit that performs the required functionality to reduce the pixel size and increase the neuron density. This will not only save the occupied area of each pixel but also the power consumption of each neuron.
Chapter 6

Conclusions

A summary of the completed work will be presented along with recommendations for future work.

6.1 Summary of Contributions

The work developed in this thesis provides some guidance on the design of optoelectronic neural activation function for the realization of the embodiment of the fully integrated COIN Co-processor to be built in the future. The thesis designed, developed and tested compact, power-efficient and integrable optoelectronic neuron circuitry for the COIN Co-processor. Optoelectronic implementations of the fundamental building blocks-the synapse, the neuron and the local weight memory-that are necessary to implement an optoelectronic neural network were designed. All the operations- multiplications, activation and addition-required by the synapse and the neuron were realized through optical devices and MOS transistors. Three versions of optoelectronic activation functions for two architectures of the COIN Co-processors have been designed that emulate the electrical characteristics of actual neurons. First, a 9x9 array of simple and non-linear sigmoid type of analog circuits was designed, tested and implemented in PCB. After that, in version-II of the optoelectronic neuron a local memory element was incorporated to enable on chip-learning and real-time neuron weight adjustments. Finally, a new and tunable threshold value CMOS op-
toelectronic neuron has been proposed. A spice simulation was performed using the TSMC 0.25µm CMOS process technology for the proposed neuron circuits.

6.2 Applications

Hardware based artificial neural networks find applications in numerous areas. One of the most recent application of these types of ANNs is in portable electronic nose (E-nose) also known as the artificial olfactory system. A three layer (with one hidden layer) analog MLPNN circuit was designed and fabricated in 0.18µm standard CMOS process to identify fruit odors of bananas, lemons, and lychees with 91.7% accuracy [43]. The E noise can be used to monitor air quality, food quality, hazardous gas detection and health care[8]

Previously, the Photonics systems Group (PSG) collaborated with other researchers from Europe to develop an Artificial Vision Systems[63]. A prototype unit cell was built using a discrete optical, electronic and optoelectronic device to demonstrate the implementation of basic neural network algorithm used in artificial vision systems.

Furthermore, most recently Travis Simpkins[56] designed and trained a five layer 12x12 neuron array of COIN system for face recognition applications. If the COIN Co-processor system is fully realized, it will have countless applications. To mention a few:

- Pattern recognition from noisy and fragmented input images
- Image segmentation for medical applications
- Fusion of multi-spectral-multipolarization images for contrast enhancement
- Adaptive optical wave front phase computation
- Pattern recognition of human features from large quantities of parallel biometric data
- Multisensor signal processing (e.g. for robotic guidance and control)
- Simulation for DNA computation and modeling
- Multimedia processing, symbolic processing, and encryption breaking
- Probabilistic reasoning, estimation, extrapolation and interpolation.

### 6.3 Discussion of Future Work

In the short run, we plan to create a working prototype of a multi-layer COIN system using off-the-shelf electronic components as well as the optical interconnection element (developed by William Herrington[23]) that can fit in a shoe box and can be trained to recognize faces.

Mid-term plans include IC fabrication of the optoelectronic neuron arrays and integration of the optical interconnect element with them to finally realize the fully integrated hardware version of the COIN processor. These two components can be integrated with other devices that are required to form a complete system for a five layer integrated and feed-forward optoelectronic neural processor. Based on lessons gained from the mid-term goal outcomes, we will work on making the system space efficient, low-power and marketable product for the applications mentioned above which leads us to the long term plans.

For the long run we have two major milestones; firstly, we want to redesign the COIN system itself to be more robust. For example, so far we have been able to design and train a neural net that can perform a feed-forward network involving communications between one layer to the next layers of neurons. Incorporating multi-directional communication such as bidirectional broadcasting (both feed-forward and feed-back) as well as co-planar communications with neighboring neurons within the same plane will improve the robustness of the system. Secondly, we plan to expand
to incorporate memristive based neural system [25] following the recent resurgence of these devices for neural net applications.
Bibliography


[42] OSI Optoelectronics. Photodiode characteristics and applications. *Application Notes*, N/A.


