2017

SkyNet: Memristor-based 3D IC for Artificial Neural Networks

Sachin Bhat
University of Massachusetts Amherst

SKYNET: A MEMRISTOR-BASED 3D IC FOR ARTIFICIAL NEURAL NETWORKS

A Thesis Presented

by

SACHIN BHAT

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

MASTER OF SCIENCE IN ELECTRICAL AND COMPUTER ENGINEERING

SEPTEMBER 2017

Electrical and Computer Engineering

SKYNET: A MEMRISTOR-BASED 3D IC FOR ARTIFICIAL NEURAL NETWORKS

A Thesis Presented

by

SACHIN BHAT

Approved as to style and content by:

____
Csaba Andras Moritz, Chair

____
Daniel Holcomb, Member

____
Zlatan Aksamija, Member

____
Christopher V. Hollot, Department Head
Electrical and Computer Engineering

ACKNOWLEDGMENTS

I take this opportunity to express gratitude to my advisor Prof. Csaba Andras Moritz. He has been a constant source of inspiration throughout my Master's study. I would like to thank my Master's Thesis committee consisting of Prof. Daniel Holcomb, Prof. Zlatan Aksamija for their valuable feedback. I would like to thank my colleagues in research and friends Sourabh Kulkarni, Jiajun Shi and Mingyu Li for their valuable feedback and suggestions.

Finally, I would like to express my gratefulness to my parents who have sacrificed many things to help me get a decent education. They have been very supportive right from my childhood and I owe whatever success I have had in life, to them.

ABSTRACT

SKYNET: A MEMRISTOR-BASED 3D IC FOR ARTIFICIAL NEURAL NETWORKS

SEPTEMBER 2017

SACHIN BHAT, B.E, SIDDAGANAGA INSTITUTE OF TECHNOLOGY

M.S.E.C.E., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Csaba Andras Moritz

Hardware implementations of artificial neural networks (ANNs) have become feasible due to the advent of persistent 2-terminal devices such as memristor, phase change memory, MTJs, etc. Hybrid memristor crossbar/CMOS systems have been studied extensively and demonstrated experimentally. In these circuits, memristors located at each cross point in a crossbar are, however, stacked on top of CMOS circuits using back end of line processing (BOEL), limiting scaling. Each neuron’s functionality is spread across layers of CMOS and memristor crossbar and thus cannot support the required connectivity to implement large-scale multi-layered ANNs.

This work proposes a new fine-grained 3D integrated circuit technology for ANNs that is one of the first IC technologies for this purpose. Synaptic weights implemented with devices are incorporated in a uniform vertical nanowire template co-locating the memory and computation requirements of ANNs within each neuron. Novel 3D routing features are used for interconnections in all three dimensions between the devices enabling high connectivity without the need for special pins or metal vias. To demonstrate the proof of concept of this fabric, classification of binary images using a perceptron-based feed forward neural network is shown. Bottom-up evaluations for the proposed fabric considering 3D implementation of fabric components reveal up to 19x density, 1.2x power benefits when compared to 16nm hybrid memristor/CMOS technology.

TABLE OF CONTENTS

Page

CHAPTER

LIST OF TABLES

Table **Page**

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The field of Artificial Neural Networks (ANNs) has attracted increasing attention in recent years. ANNs are preferred computation models for a wide variety of information processing applications such as computer vision, pattern recognition, process control, signal processing among others which are inefficient when algorithmic approaches of conventional rule-based programming are used. ANNs are biologically inspired abstract computation models made up of densely interconnected parallel processing units called neurons, typically organized in layers. These processing units take several inputs weighted by the synaptic weights, which are integrated and mapped to outputs based on a non-linear function called the activation function.

ANNs have a highly parallel architecture, dense connectivity, and distributed memory and computation. Hence, implementing neural networks on traditional von-Neumann-based computers is very inefficient because of their inherent difference in architecture. Therefore, several hardware implementations have been proposed with analog CMOS [1], digital CMOS [2], and hybrid memristor/CMOS [3] [4], which take advantage of their inherent parallelism and perform orders of magnitude faster than their software counterparts. Recently, the hybrid memristor crossbar/CMOS systems have received widespread attention. Memristors are novel nanoscale devices with multi-state persistent memory, which makes them suitable candidates for modeling key features of synaptic weights. Analog or digital circuits using CMOS technology address decoding circuits, activation function, and other supporting features as part of the neuron functionality. In these implementations, synaptic weights are mapped to a global memristor crossbar array integrated on top of CMOS circuits with communication achieved either through area distributed interfaces [3] or Through-silicon Vias (TSVs) [5]. A typical implementation of a hybrid memristor/CMOS system is shown in Figure 1.

Conceptually, in ANNs, the synaptic weights and the neurons are co-localized and spatially distributed. Synaptic weights grow quadratically with the number of neurons. However, the heterogeneity of the coarse-grained stacked hybrid memristor/CMOS technology introduces connectivity and scalability bottlenecks, which limit their ability to implement neural networks. Furthermore, CMOS logic doesn’t scale as well as the denser memristor crossbar arrays and hence, to implement large-scale neural networks, multi-chip systems are required which also causes inter-chip communication overhead [6]. The key to efficient implementation of ANNs is to restrict the communications to local data transfers. As synaptic weights are mapped to a global memristor crossbar array, area distributed interface or TSVs are required for communication between the synaptic weights and neurons, and decoding circuitry for addressing, which leads to additional performance and power overhead [4]. Hence, currently, there is no integrated circuit technology for implementing large-scale neural networks.

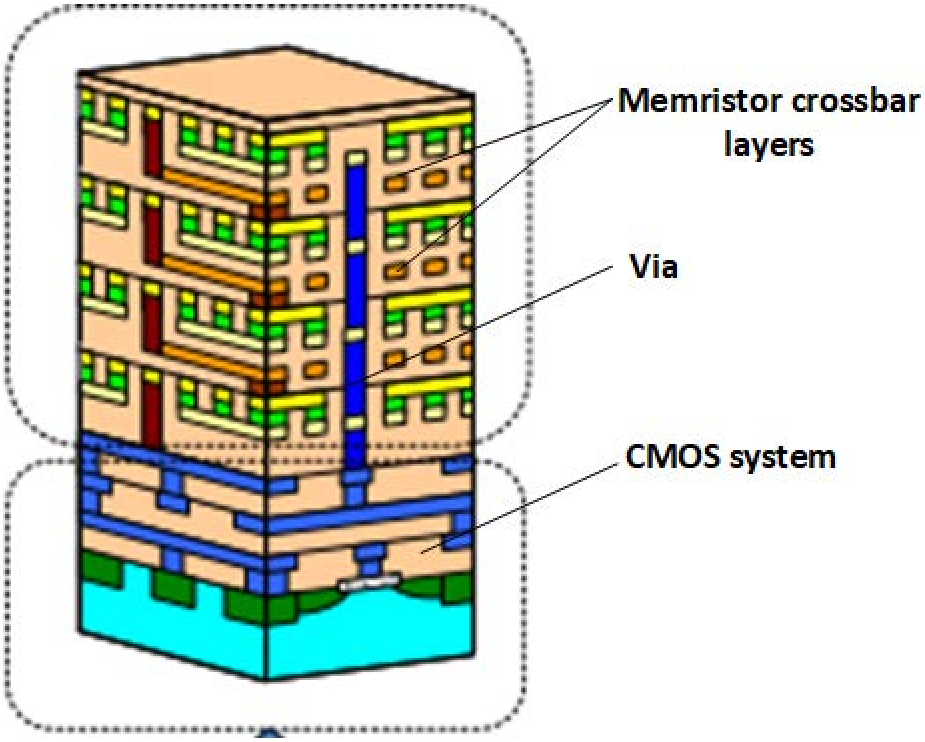


Figure 1: Typical implementation of the hybrid memristor/CMOS systems [3]

To overcome these aforementioned challenges, a new fine-grained 3D ASIC technology to implement artificial neural networks for cognitive computing applications, SkyNet, is proposed. This technology which builds on uniform vertical nanowire templates meets several criteria for addressing ANN requirements as: (i) it enables dense 3D vertical integration of synaptic weights, neurons and interconnect in a fabric-centric mindset; (ii) it allows for 3D spatial co-distribution of synaptic weights and neurons across the fabric thus mitigating the need for stacked hybrid architecture; and (iii) achieves high local connectivity between synaptic weights and neurons by utilizing novel 3D interconnect and routing features.

CHAPTER 2

ARTIFICIAL NEURAL NETWORKS – BACKGROUND

Artificial neural networks try to model the information processing capabilities of nervous systems. They are classified to be one of the major models of computation. ANNs are characterized by passively parallel and redundant non-linear processing units called as neurons. ANNs are characterized by an activation function and interconnection of these neurons defines the functionality of the network. Figure 2 shows an abstract model of a neuron with 'n' inputs. The inputs to the neuron can be any real values, with each input having a weight associated with it. Strengths or weights associated with the neurons are called synaptic weights. The inputs are multiplied with their corresponding synaptic weights and integrated at the neuron. The integrated weighted inputs are fed to the activation function, which maps it to a real value. Synaptic weights are used to store the knowledge acquired by the network and they can be adapted to attain a desired objective. ANNs can be made to learn to perform a certain task by adjusting their synaptic weights, which can increase or decrease the strength of the signals sent to other neurons. Thus, the ANNs can learn to perform without being explicitly programmed.

As mentioned previously, the functionality of the neural networks is defined by the interconnection of neurons. Hence, they can be categorized into various types based on interconnections. If the neurons are organized in multiple layers where input from previous layers feeds to the next layer without any feedback, the, this type of network is called Feedforward neural network. If the output of one layer is fed to the input of its previous layer, then these are termed as Recurrent Neural Networks (RNNs). Different types of ANNs differ mainly due to their activation function and interconnection of nodes.

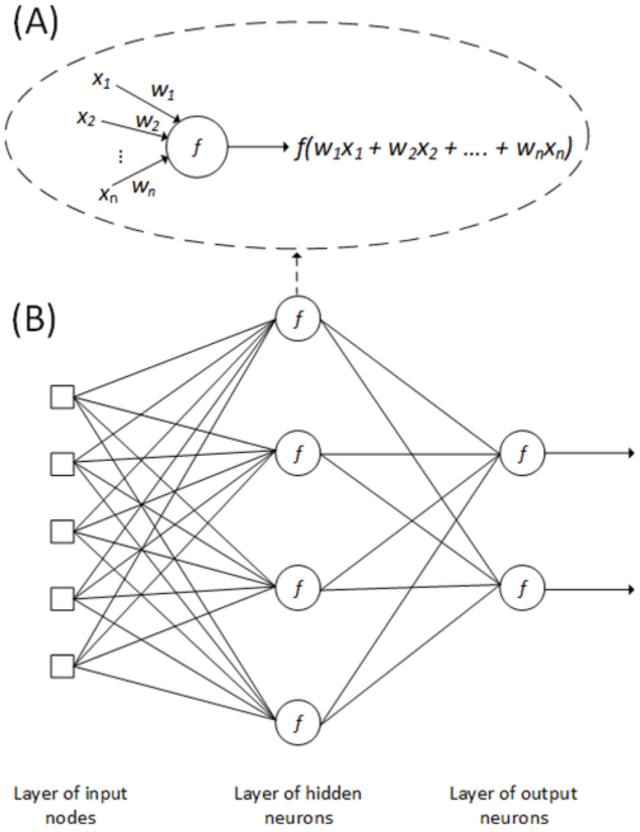


Figure 2: (A) A general model of a neuron; (B) Feedforward neural network

2.1 Neuron

A neuron is the most basic information processing unit that is fundamental to the operation of a neural network. It consists of a set of synapses which are characterized by a strength of its own. Each input signal is multiplied with a specific synaptic weight attached to it. The synapse can have any range that includes positive and negative numbers. It also has an additional weight called the bias which can sway the output of the neuron in either direction depending on whether it is positive or negative. The sum of the dot products of the synaptic weights and the inputs is fed to the activation function which acts as the limiting function. It limits the amplitude range of the output signal to some finite value depending dot products. There are different types of activation functions available for use such as threshold function, sigmoid function etc. The choice of a particular activation function depends on the application. If the output of the activation function of a neuron is a binary value, then such a neuron is called a Perceptron. A single perceptron can act as a binary classifier which can classify linearly separable patterns. A feedforward multilayer perceptron with more than one layer can be used to approximate any continuous function according to universal approximation theorem. Hence, neural networks promise a great potential for information processing applications if hardware realization can be achieved. The figure of an abstract neuron is shown is Figure 3.

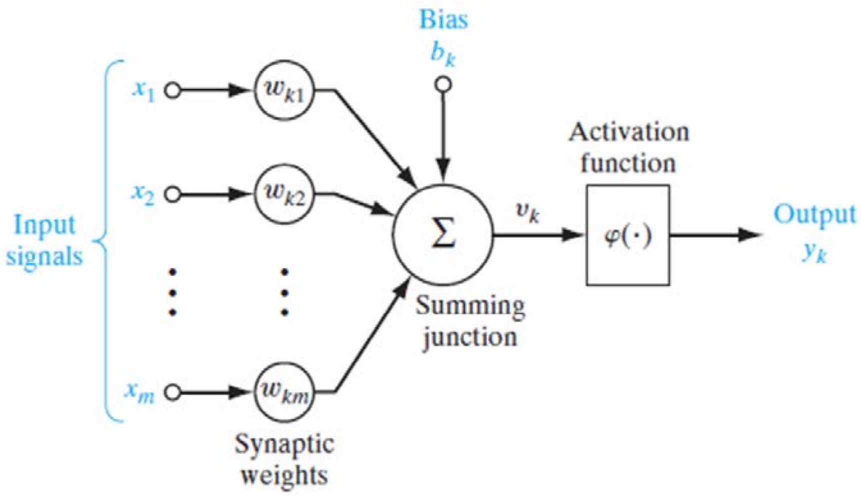


Figure 3: An Abstract Neuron

CHAPTER 3

OVERVIEW OF THE SKYNET FABRIC

3.1 Motivation

In hybrid memristor/CMOS systems, connectivity between the memristor crossbar arrays and underlying CMOS circuits are engineered as an after-thought and is a compromise. As ANNs scale in size, number of synapses and connections grow quadratically which quickly becomes impractical to wire. SkyNet follows a fabric-centric mindset where the active and passive devices, circuit framework, and connectivity are carefully engineered together towards a 3D organization. Its manufacturability requirement follows the same mindset as other 3D IC fabrics (SkyBridge [7] and Skybridge-3D-CMOS [8] [9] [10]). The fabric uses a regular array of uniform pre-doped vertical nanowires as a template which is then functionalized with vertical junctionless transistors, memristors, 3D routing structures such as bridges, co-axial routing structures, Interlayer-Connection (ILC), etc., through material deposition techniques.

3.2 Core Components

3.2.1 Vertical Nanowires

An array of dual-doped regular vertical nanowires are the fundamental building blocks of the SkyNet fabric. All the devices and components of the fabric are formed on a uniform nanowire template. Forming the vertical nanowires precedes all the manufacturing steps. Heavily doped p-type and n-type substrates are vertically stacked and bonded together using molecular bonding techniques [11]. A layer of silicon dioxide named as

Interlayer Dielectric (ILD) provides the isolation between the n-type and p-type doped silicon layers. More layers with different doping profiles can be stacked by performing the process iteratively.

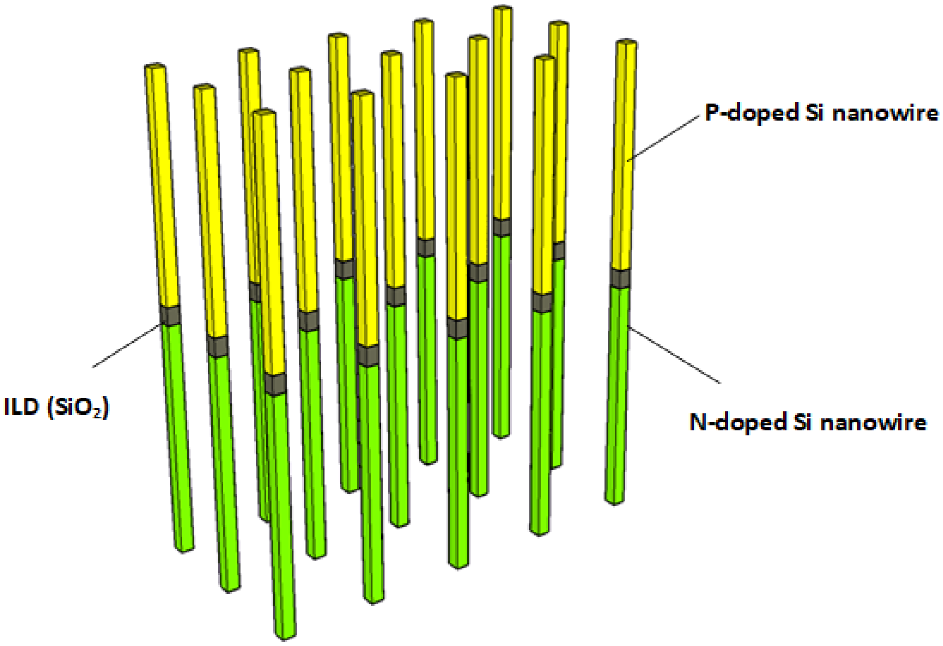
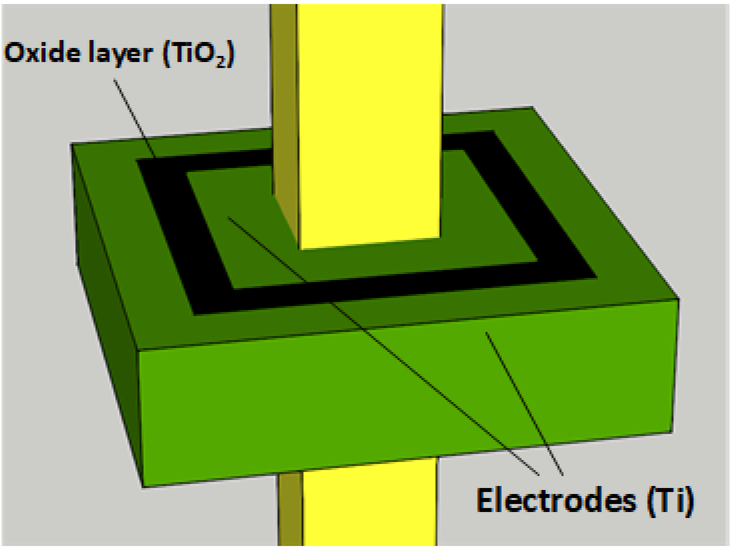


Figure 4: Vertical Nanowires

3.2.2 Memristors

Memristors or Memristive devices are promising candidates for implementing synaptic weights because of their analog memory functionality and persistence. They are passive two-terminal devices whose internal resistance depends on the history of the applied voltage and current. Upon excitation by a bipolar periodic stimulus, they exhibit a pinched hysteresis in the current-voltage domain. Memristive devices typically consist of a transition metal oxide layer sandwiched between two electrodes. The resistive switching behavior is attributed to the formation and rupture of conductive filaments that aid the current flow through the oxide layer. Over the years, memristors with several different oxide materials have been proposed such as titanium dioxide [12] and hafnium dioxide [13], to name few.



The proposed fabric uses titanium dioxide memristive devices for synaptic weight implementation. Figure 5 shows the memristor device design. Memristors are distributed throughout the fabric along with other fabric components with fine granularity unlike stacked architectures in hybrid memristor/CMOS systems. The titanium oxide based memristors have an intrinsic rectifying property due to their highly non-linear switching dynamics, and hence external select devices such as transistors or diodes are not required for their operation [14]. Since the memristors can be deposited with material deposition techniques, the manufacturing requirements for them do not depart from that of the other SkyNet components. They have similar feature size as fabric components; as small as 10x10 nm^2 has been experimentally demonstrated [13]. Since the silicon nanowires are heavily doped, the inner electrode forms an ohmic contact. This kind of structure is like the memristors with asymmetric electrodes experimentally demonstrated [15].

3.2.3 Vertical Gate All around Transistors

Vertical Gate-All-Around (V-GAA) junctionless p-type and n-type transistors are the active devices in the proposed fabric. Figure 6 and Figure 7 show the vertical junctionless transistors in SkyNet. The transistors are used to realize the functionality of the neurons and other peripheral circuitry in the fabric. These transistors have uniform doping across source, channel, and drain regions. Hence, these junctionless transistors don't possess abrupt junctions which reduces doping complexities. The work function difference between the gate electrode and the heavily doped silicon nanowires modulates the electrical behavior of these transistors. The work function difference depletes the charge carriers in the channel if no gate voltage is applied. However, when an appropriate gate voltage is applied, the channel becomes conductive. Since the channel length is dictated by the thickness of material deposition instead of lithography accuracy, it allows for scaling the channel length beyond the lithography limitations. For the n-type transistors, Titanium Nitride is the gate material while Tungsten Nitride is the gate material of choice for p-type transistors. Hafnium Dioxide provides the isolation between the gate and the channel. Because of their structural simplicity, these transistors can be stacked on the vertical nanowires to form 3D neuron circuits thus achieving very high density. These types of transistors have been well researched [17] and experimentally demonstrated by our group [18].

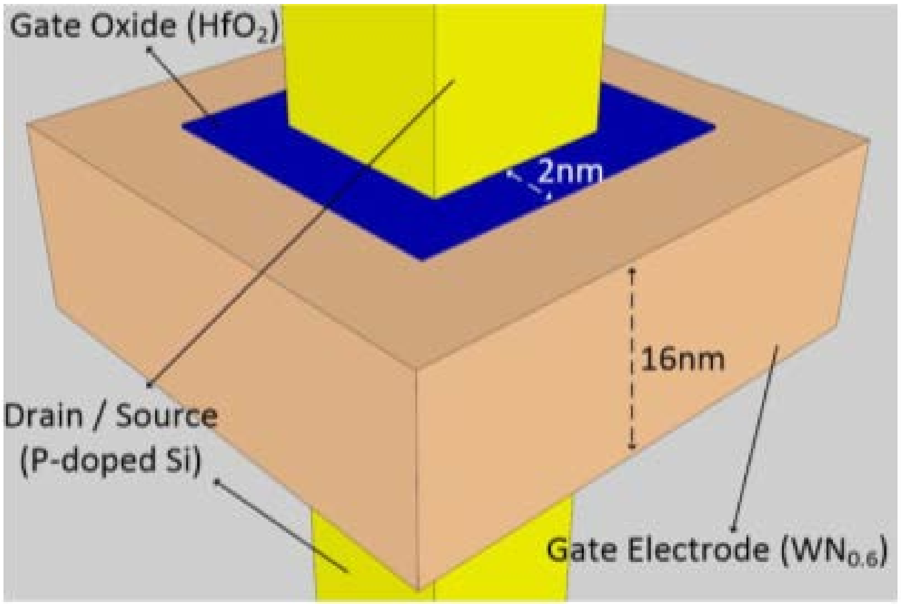


Figure 5: P-type Junctionless Transistor in SkyNet

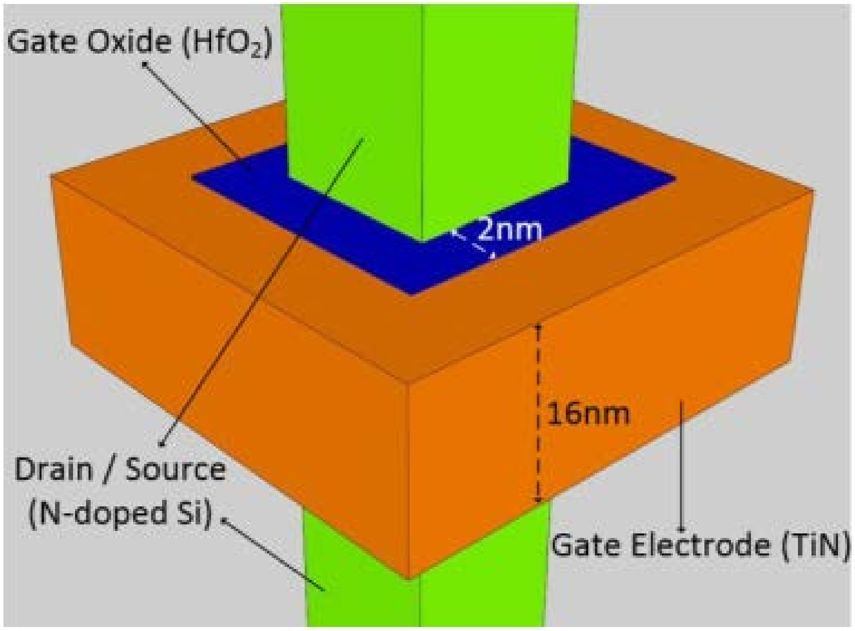


Figure 6: N-type Junctionless Transistor in SkyNet

3.2.4 Contacts

Contacts are required to connect doped silicon nanowires with other components of the fabric. Hence, good ohmic contact is necessary. Titanium is the contact electrode material for n-doped silicon while nickel is chosen as the contact material for the p-doped silicon. The work function difference between the contacts and the silicon must be like have a good ohmic contact. Contacts designs are shown in.

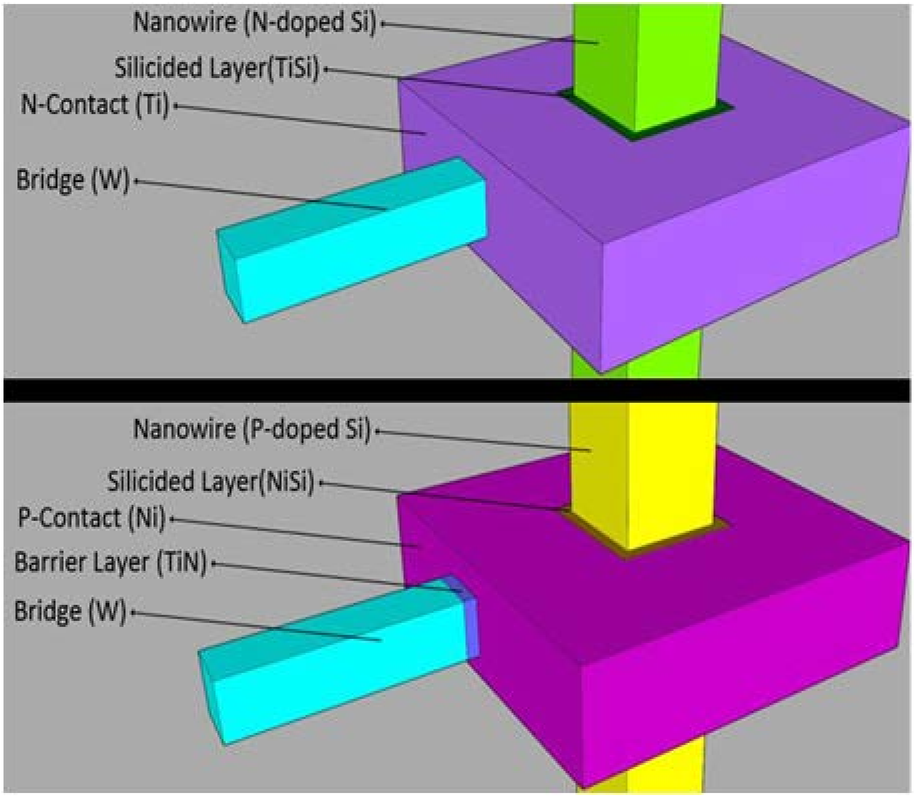


Figure 7: N-type and P-type Contacts

3.2.5 3D Connectivity features in SkyNet

The functionality of the ANNs depends on the interconnection of the neurons in the network. In hybrid memristor/CMOS systems, metal vias are used for connecting CMOS neurons with the memristor crossbar arrays. This is sufficient for very small-scale ANNs. However, for large-scale ANNs, the wiring requirement explodes with the number of synaptic weights. Hence, to efficiently implement ANNs, a good interconnection framework is necessary. The proposed fabric supports many kinds of interconnect structures to accommodate this connectivity without routing congestions. (i) Bridges are metal wires used for horizontal routing of signals between nanowires; (ii) The heavily doped nanowires can be used for vertical routing of signals; (iii) Co-axial routing structures can be used for vertical routing in addition to the nanowires; and (vi) ILC is for connecting n-type and p-type nanowires when implementing circuits with the vertical GAA transistors.

3.2.5.1 Bridges and Coaxial Routing

Bridges and Co-axial routing structures are the two kinds of routing structures in the SkyNet fabric. Bridges provide connection between the adjacent nanowires while coaxial structure offers connectivity in the vertical direction. A single nanowire can accommodate multiple bridges at different heights to improve connectivity. The coaxial routing structure can at most have two metal layers separated by insulator for isolation. This is in addition to the doped silicon nanowire which also provides the vertical connectivity. Tungsten is used for routing structures due to its good electric characteristics.

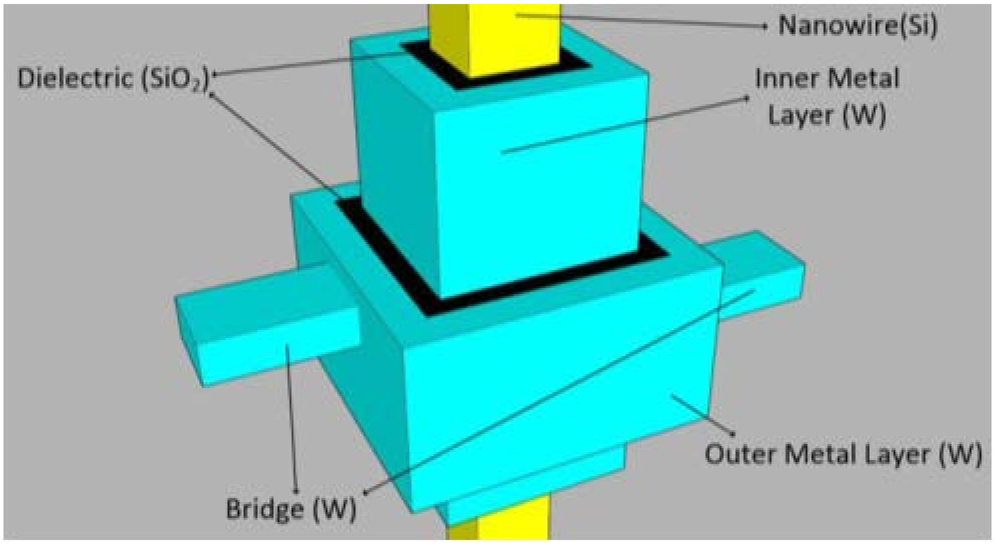
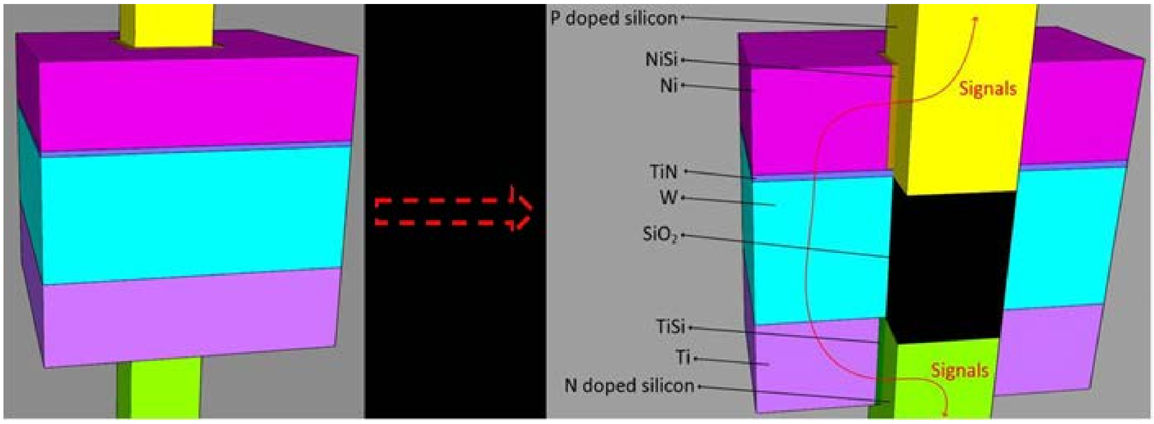


Figure 8: Bridges and Coaxial Routing

3.2.5.2 Interlayer Connection

The N-type and P-type nanowires are isolated from each other by the interlayer dielectric. Hence, a connection is required between the active devices on p-type and n-type nanowires. In addition to this, the connection is required to enable the vertical signal routing to bypass the interlayer dielectric. Figure 10 shows the structure of interlayer connection. The materials are carefully chosen to ensure good ohmic contact between them.



CHAPTER 4

PERCEPTRON IN SKYNET

This chapter shows the complete implementation of a Perceptron using various components of SkyNet outlined in the previous chapter. It starts with the hardware implementation of perceptron, later discussing about individual components.

4.1 Hardware Implementation of a Perceptron

Perceptron is the most basic processing unit of an ANN. There are three important factors which requires consideration for developing a hardware solution for the neural networks, they are, (i) encoding of signals used in the network; (ii) implementation of weights; (iii) the integration and output function (neuron functionality). Figure shows an analog implementation model of a perceptron. The inputs to the perceptron is encoded as voltages (Vj). For example, if the perceptron is used to identify/classify an image, then the brightness of the pixels encoded as voltage values. Memristors are used as synaptic weights and a differential amplifier performs as the activation function. There are other alternative implementations but analog implementation offers higher implementation density on Silicon and requires less power.

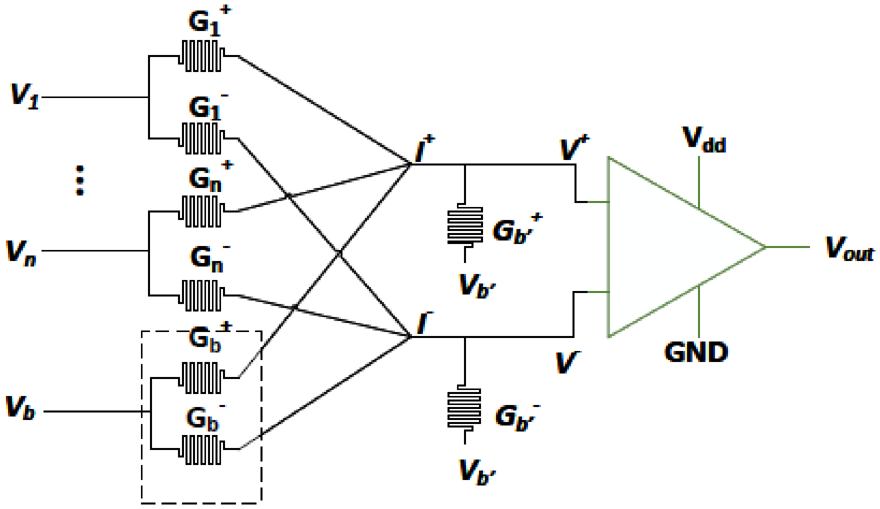


Figure 9: Analog implementation of the Perceptron

4.1.1 Synaptic Weights in SkyNet

Memristors are suitable candidates for the implementation of synaptic weights. When the input voltages are applied to the memristors, the currents generated are the dot product between the input voltages and the resistances of the memristors according to Ohm's law. The currents can then be summed at a common point which follows according to Kirchhoff laws. The voltages corresponding to the summed currents can then be fed to a differential amplifier.

The synaptic weights in a neural network can be positive or negative. Since memristors can only represent positive conductances, negative weights cannot be implemented with a single memristor. Hence, each weight w is implemented as a differential pair of memristor conductances G = G+ - G-. The currents corresponding to conductances G+ and G- are summed separately and converted to equivalent voltages before being fed to the differential amplifier. Although operational amplifiers in virtual ground mode are typically used for converting currents into voltages, they consume a lot of energy and area. For this work, a technique shown in [18] is used; in this, the voltage drops across grounded memristors is fed to the differential amplifier, and by choosing memristors with appropriate conductances, the inputs to the differential amplifier can be swayed one way or the other, to classify a set of linearly separable input patterns.

Figure 12 shows the implementation of weight with a pair of memristors stacked on the nanowire. The memristors with conductances G+ are implemented on the p-type nanowires whereas memristors with conductance G- are implemented on the n-type nanowires. Since the currents corresponding to conductances G+ and G- are summed separately, they are isolated from each other in SkyNet through the interlayer dielectric between the p-type and n-type nanowires. This isolation effectively reduces the footprint of the memristor array. Since sneak path currents are directly proportional to this footprint, this isolation reduces the sneak path currents substantially. The currents from the p-type and n-type nanowires can easily be summed using the SkyNet routing structures such as bridges and co-axial routing. Although the figure shows a pair of memristors on the dual-doped nanowire, many pairs of memristors can be stacked to achieve high synaptic weight density. In contrast to the hybrid memristor/CMOS systems, the proposed fabric doesn’t impose any restrictions on the placement of memristors along with the other fabric components, and hence high density and homogeneous distribution of synaptic weights and neurons is possible.

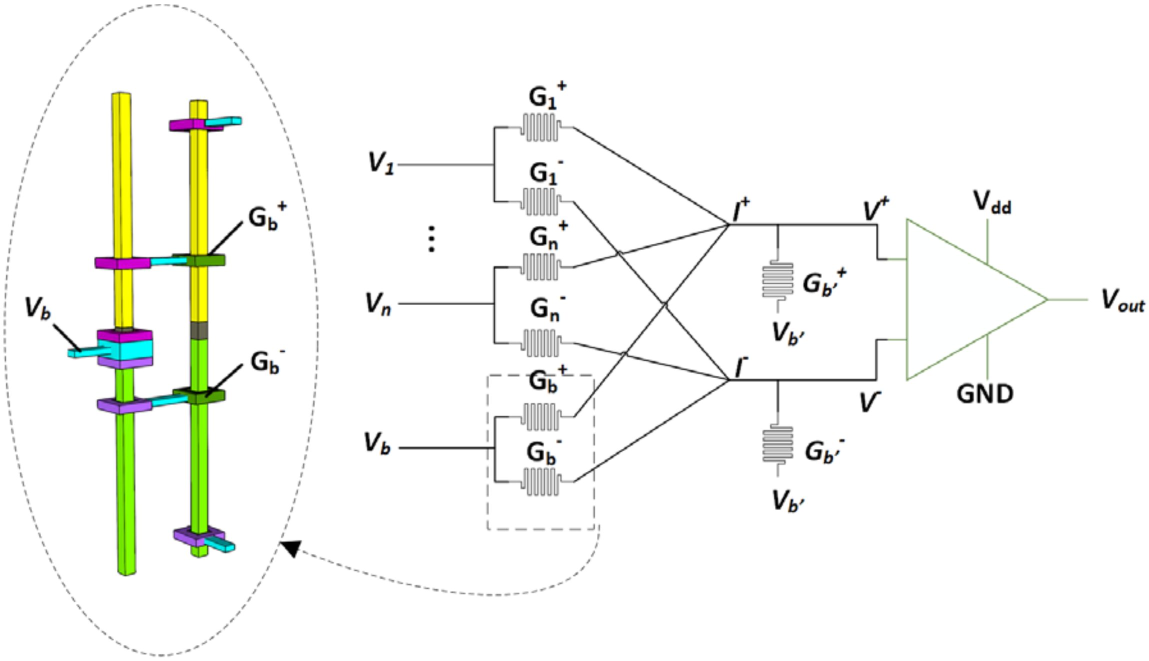


Figure 10: Synaptic weights implementation in SkyNet

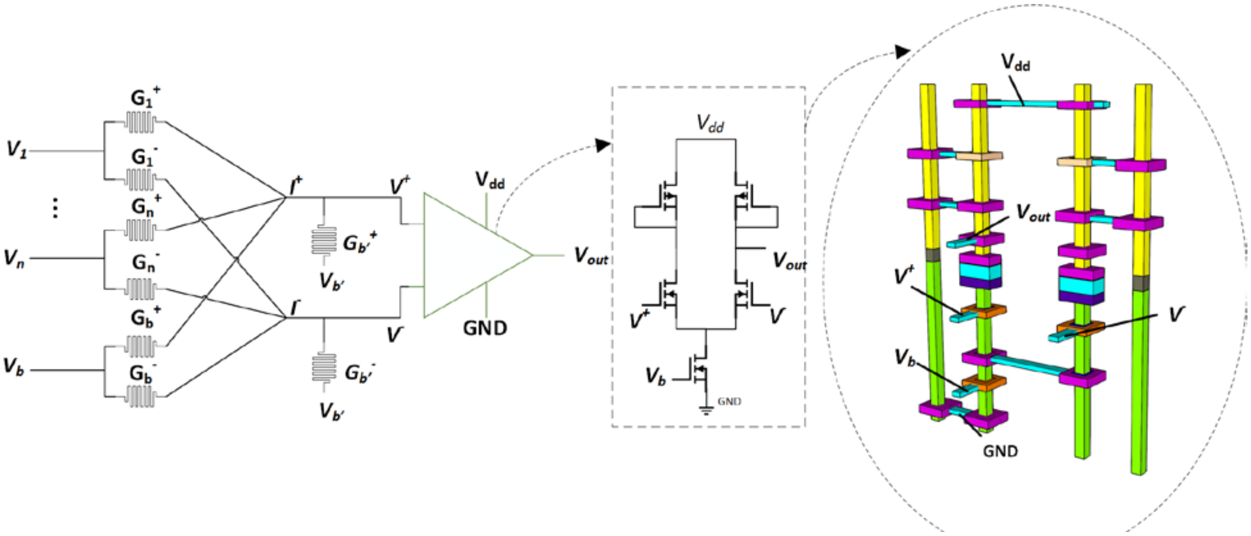


Figure 11: Circuit Implementation of Differential amplifier in 3D

4.1.2 Differential Amplifier

A differential amplifier is chosen to implement the activation function. If the difference between the one input and the other is positive, then it outputs a logic high otherwise a logic low. It is to be noted that the transfer characteristic of the differential amplifier closely resembles that of the sigmoid function. The figure shows the implementation of a differential amplifier using vertical p and n-type junctionless transistors. The circuit schematic is shown Figure 13. The p-type transistors are used as current source loads while the n-type transistors are as input differential transistors. ILC connects the p-type and n-type nanowires. The benefits of the 3D integration are obvious from the figures. The entire differential amplifier can be realized by using only four nanowires as shown in Figure 13. In hybrid memristor/CMOS systems, the neuron functionality is implemented in the 2D CMOS layer resulting in a large neuron footprint.

4.1.3 Read/Write Support Functionality

The conductances of the memristors must be changed according to the type of pattern that is to be classified. Two phases of operation, read phase and write phase need to be supported. During the read phase, the conductances of the memristors must be sensed without disturbing their state; for memristors with non-linear switching dynamics, this is accomplished using $V = V_{rd}$. During the write phase, their conductance must be changed. A common scheme for this is to apply V_{wr} on one terminal and $-V_{wr}$ on the other terminal of the memristor. This results in a total voltage drop of $2V_{wr}$ across the memristors, sufficient since it is greater than the threshold voltage of a memristor.

Supporting this scheme requires additional circuitry. The V_{rd}, V_{wr} and $-V_{wr}$ signals must be multiplexed so that both the read and write schemes can be supported. The circuit schematic for such a scheme is shown in Figure 14. Read and write control ($V_{RD\text{-}CTRL}$ and $V_{WR\text{-}CTRL}$) signals enable and disable the transmission gate-based switches depending on the type of operation. During the read phase, the $V_{RD\text{-}CTRL}$ signal enables the switches such that the memristors can be read simultaneously. Write operation is sequential, where $V_{WR\text{-}CTRL}$ signals are enabled sequentially depending on the memristor that needs to be written. Figure 14 shows the implementation of the read/write circuitry in SkyNet. Co-axial routing structures are used to supply the control signals to n and p-type vertical junctionless transistors. ILC is used to short the terminals of the p-type and n-type transistors, which are connected to the memristors through the bridges. This results in a very compact implementation vs. state-of-the-art.

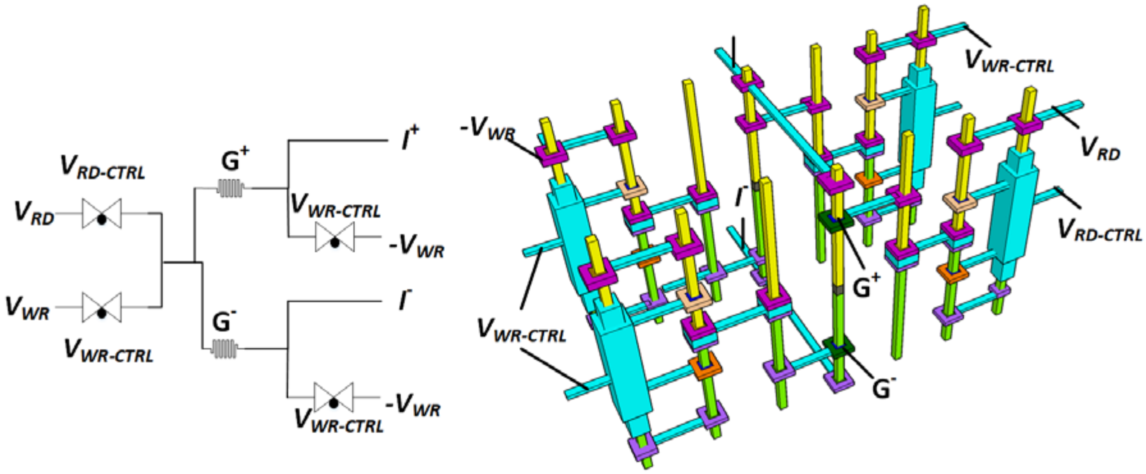


Figure 12: Read/Write Circuit Support implementation in SkyNet

4.2 Perceptron in SkyNet

Figure 15 shows the complete implementation of a perceptron along with the read/write support for memristors in SkyNet. V-GAA transistors and memristors are distributed throughout the fabric resulting in a compact implementation of a perceptron. While p and n-type nanowires sum the currents from the memristors in the vertical direction, metal bridges are used for summing the currents from all the nanowires in the horizontal direction. The voltage drop across the grounded memristors is then fed to the differential amplifier for classification.

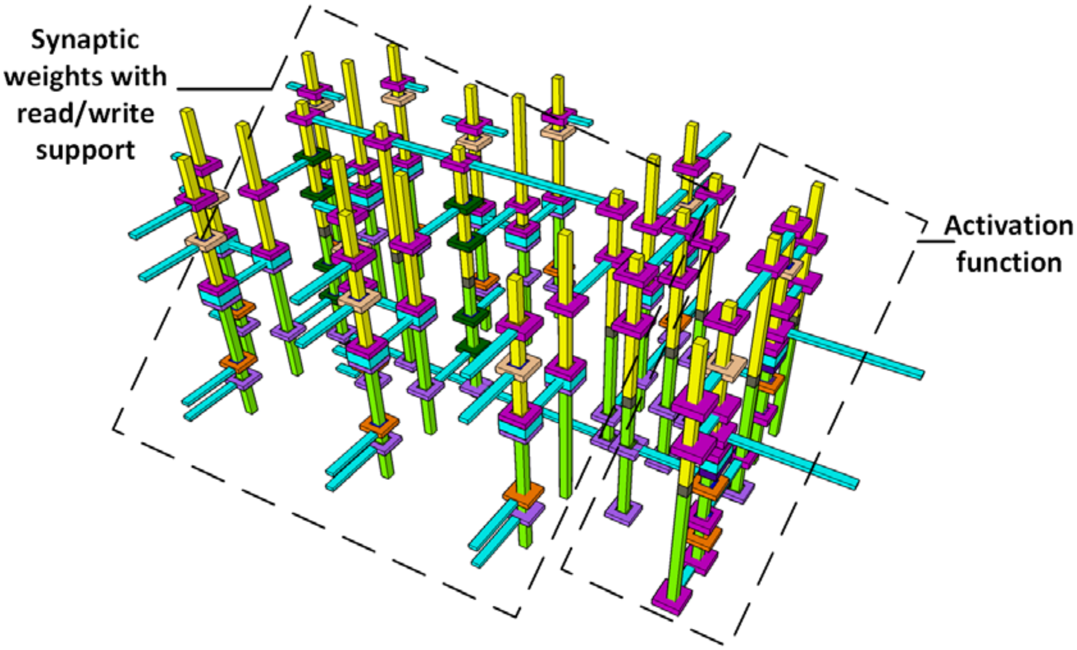


Figure 13: Perceptron in SkyNet

CHAPTER 5

EVALUATION METHODOLOGY

In this chapter evaluation methodology of the SkyNet fabric will be described. The device and material level evaluation of the vertical junctionless transistors is described by previous works carried in our group [22]. The chapter starts with memristor model description. Later, the circuit and layout design methodology is described. The chapter ends with the description of the CMOS baseline used.

5.1 Memristor Model

As mentioned earlier, titanium oxide memristors are considered for this work. Verilog-A VTEAM[20] memristor model compatible with HSPICE was chosen to model them. It is a general model for voltage controlled memristors and is used to fit the experimental results of titanium dioxide memristors demonstrated [21]. For these devices, due to their high non-linear switching dynamics, the memristor conductances can be read with Vrd ≈ 0.8V without disturbing the state of the memristors. For all memristors considered in this work, G_{max} = 100µS and G_{min} = 10µS. The synaptic weights can be set from – G_{max} + G_{min} to + G_{max} -G_{min} because of the differential representation.

5.2 Circuit and Layout Design

The TCAD device and process simulation data were used to create behavioral models for HSPICE simulation. The schematics of the circuit are designed in physical-level layout in 3D using the design rules described in. Area footprint is calculated based on the number of nanowires and nanowire pitch. A HSPICE netlist is built to describe the layout with all the

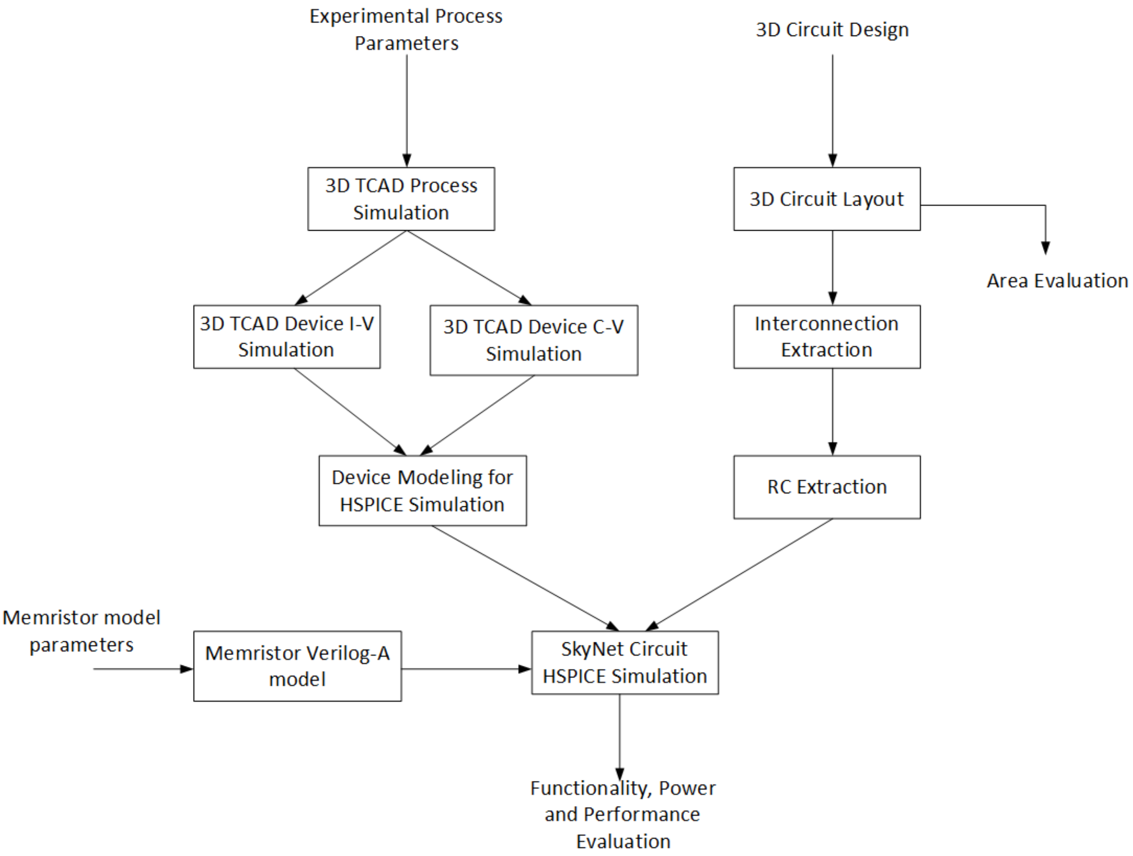

Experimental Process Parameters
3D Circuit Design
3D TCAD Process Simulation
3D Circuit Layout
Area Evaluation
3D TCAD Device I-V Simulation
3D TCAD Device C-V Simulation
Interconnection Extraction
Device Modeling for HSPICE Simulation
RC Extraction
Memristor model parameters
Memristor Verilog-A model
SkyNet Circuit HSPICE Simulation
Functionality, Power and Performance Evaluation

Figure 14: SkyNet Fabric Evaluation Methodology

transistor, memristor and interconnection models. The inputs are applied to the netlist for functional verification. Due to absence of CAD tools for SkyNet fabric, RC extraction is manually done by considering the layout to measure parasitic resistances and capacitances. The resistance and capacitance of the interconnect are modeled using Predictive Technology Model [23]. 3D layouts that were manually built using the 3D design rules in [7].

5.3 Hybrid Memristor/CMOS Baseline Evaluation

To serve as the baseline for the SkyNet fabric evaluation results, identical benchmarking needs to be done for hybrid CMOS technology. The hybrid memristor/CMOS system considered here consists of a layer of memristor crossbar array stacked on top 45nm CMOS substrate. The memristor model used for the SkyNet fabric is also used here. The schematic and layout for the CMOS circuits are drawn manually using the Cadence Virtuoso. The parasitics are extracted using Calibre tool from Mentor Graphics. Afterwards, the scaling factors are used to achieve the results in 16nm CMOS technology.

CHAPTER 6

EVALUATION AND RESULTS

For proof of concept of the SkyNet fabric, a single layer perceptron network capable of classification of binary images is implemented. In this chapter, the results for such an implementation will be shown. Finally, the chapter ends with the discussion of the proposed evaluation.

6.1 Single-layer Perceptron

A single-layer perceptron is a feedforward neural network, which is capable of classification of linearly separable patterns. To validate correct functionality, we implement a single-layer perceptron with 3 perceptron, which can classify binary images of 3x3 pixels. We completed detailed simulation including a physical layer of such an implementation. The functional scheme is shown in Figure 17. It consists of 10 inputs, 32 synaptic weights, and three output perceptron to classify three different input patterns ‘X’, ‘T’ and ‘+’. Inputs corresponding to pixels are encoded using voltages V_1 to V_9. The black pixels were encoded with 0V while the white pixels with 0.8V. Since such patterns are linearly separable, there exists a set of synaptic weights $w_{i,j}$ which enable successful classification. The synaptic weights for such classification were calculated using the perceptron learning rule.

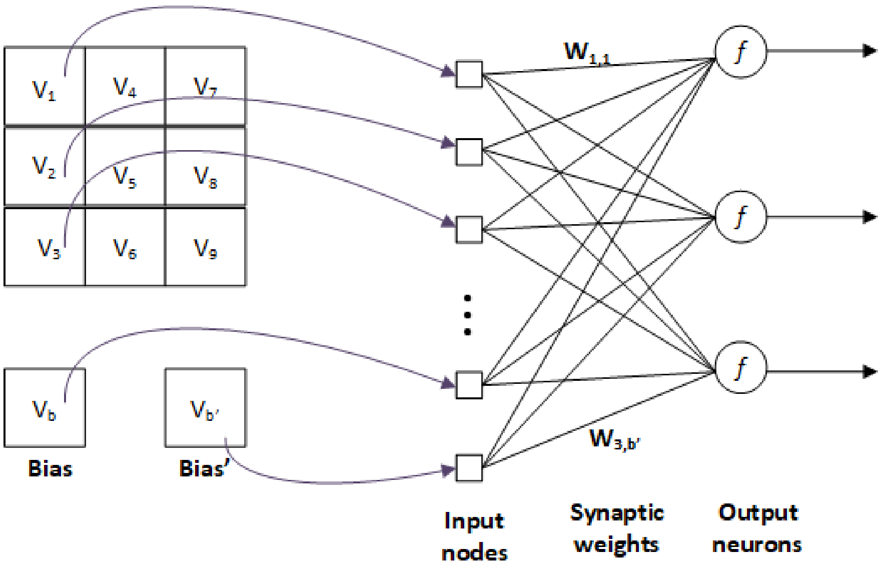


Figure 15: 3x3 binary image pixels encoded as voltages and single-layer perceptron used to classify the image

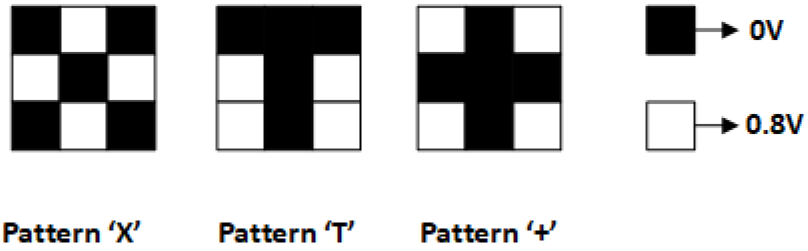


Figure 16: Test Patterns used

Table 1: Benchmarking results for the single-layer perceptron

Single-layer perceptron	Area (um^2)	Power (uW)	Latency (ps)
SkyNet	0.21	5.325	9.49
Hybrid System	4.59	13.845	17.085

Table 1 shows the single-layer perceptron benchmarking results vs. the hybrid memristor/CMOS 16nm, which also was completed. The proposed SkyNet design has 21x density benefits, 2.6x improvement in latency and 1.8x power efficiency over the hybrid stacked version. These density benefits are substantial even at this small ANN. Larger designs would benefit increasingly from the connectivity in this fabric vs. state-of-the-art hybrid schemes due to the higher routing demand in the stacked CMOS version that has no dedicated resources for connecting the neurons between hidden layers in an ANN.

6.2 Multi-layer Perceptron

The multi-layer perceptron (MLP) is a feedforward neural network with one or more hidden layers between input and output layer of neurons. MLP networks are used for non-linearly separable pattern classification. They can approximate any continuous function to any given accuracy, provided sufficiently many hidden units are available. Supervised learning techniques such as backpropagation is used for training the weights in MLP networks. This involves minimizing the mean-squared error between the inputs and outputs iteratively using the gradient descent optimization algorithm. The errors are calculated iteratively using the chain rule to calculate the gradients layer by layer.

A MLP network to classify binary images of 4x4 pixels into four different classes is implemented. It consists of 18 inputs including with bias inputs, 228 synaptic weights (228 pairs of memristors), 10 hidden layer neurons and 4 output layer neurons with sigmoid activation function to classify noisy versions of patterns ‘H’, ‘L’, ‘+’ and ‘O’. Inputs corresponding to pixels were encoded using voltages V_1 to V_{16}. The black pixels were encoded with 0V while the white pixels were encoded with 0.8V. Such patterns with noises are not linearly separable and hence, an MLP network with multiple layers is required to classify such patterns. The MLP perceptron was trained and weights were computed using the backpropagation algorithm in Matlab software running on an external computer. This kind of weight importing requires the least amount of on-chip hardware overhead but however doesn’t consider the defects of the memristors. The ‘real’ weights were then normalized and converted to equivalent memristor conductances in the range of 10uS – 100uS. Such kind of training is called as ex-situ training.

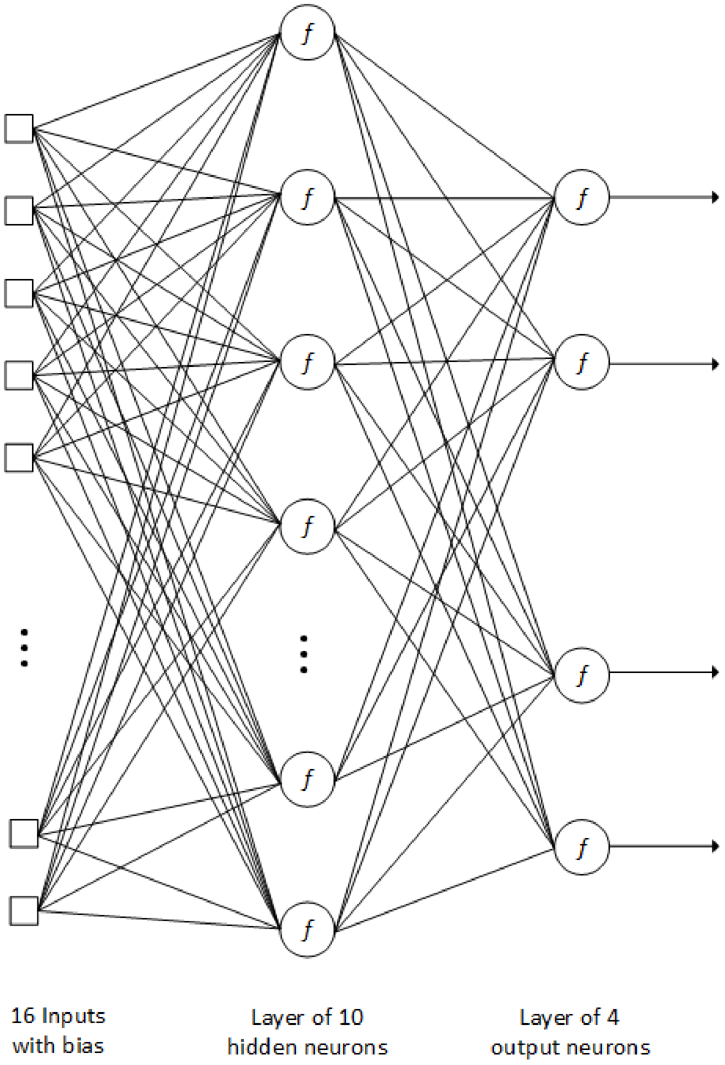
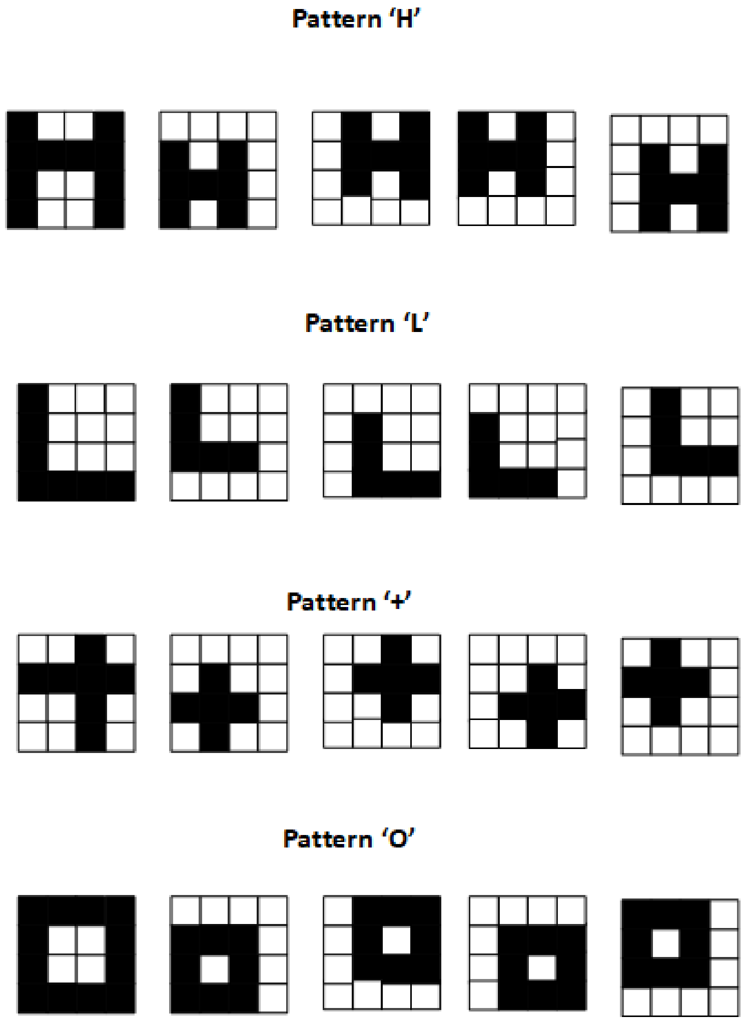


Figure 17: MLP network for classifying 4x4 binary images


Pattern 'H'

Pattern 'L'

Pattern '+'

Pattern 'O'

Figure 18: 4x4 binary images of various patterns for training the MLP network

Table 2: Benchmarking results for the multi-layer perceptron

Single-layer perceptron	Area (um^2)	Power (uW)	Latency (ps)
SkyNet	1.04	85.6	35.6
Hybrid System	19.58	106.95	31.9

Table 2 shows the MLP network benchmarking results of SkyNet vs. the hybrid memristor/CMOS 16nm. The SkyNet achieves significant improvement in density over the hybrid system because of the 3-D stacked implementation and routing structures. The

SkyNet consumes less power because of the low power junctionless devices and smaller interconnections overhead when compared to the hybrid system. The performance is slightly worse than the hybrid system because of the low performance junctionless transistors in SkyNet. The intrinsic delay of the transistors because of their structures and junctionless operation results in slightly worse performance. The junctionless transistors have weaker device driving capability which makes slower when driving large capacitive loads. In the MLP network, the transistors in the hidden layer drive the memristors and output neurons resulting in slow operation. However, the performance of the SkyNet is still orders of magnitude better than the software implementations of the neural networks because of fewer layers of abstraction. The area density and power are two most critical aspects for any hardware implementation of neural network and SkyNet achieves good results in both.

6.3 Impact of Weight Resolution on Pattern Recognition Accuracy of MLP

Theoretically, memristors should be able to switch to any resistance between the specified high-resistive (HRS) and low-resistive states (LRS). The conductance of the memristor can be arbitrarily programmed by controlling the amount of charge flowing or flux across the memristor. Practically, memristors can only switch between finite number of states. Thus, this system can’t provide the same precision and resolution as that of a digital computer. However, it has been shown that for neural networks to converge within a reasonable recognition accuracy, low precision devices are sufficient. In this section, the impact of memristor precision on recognition accuracy of the MLP network is studied.

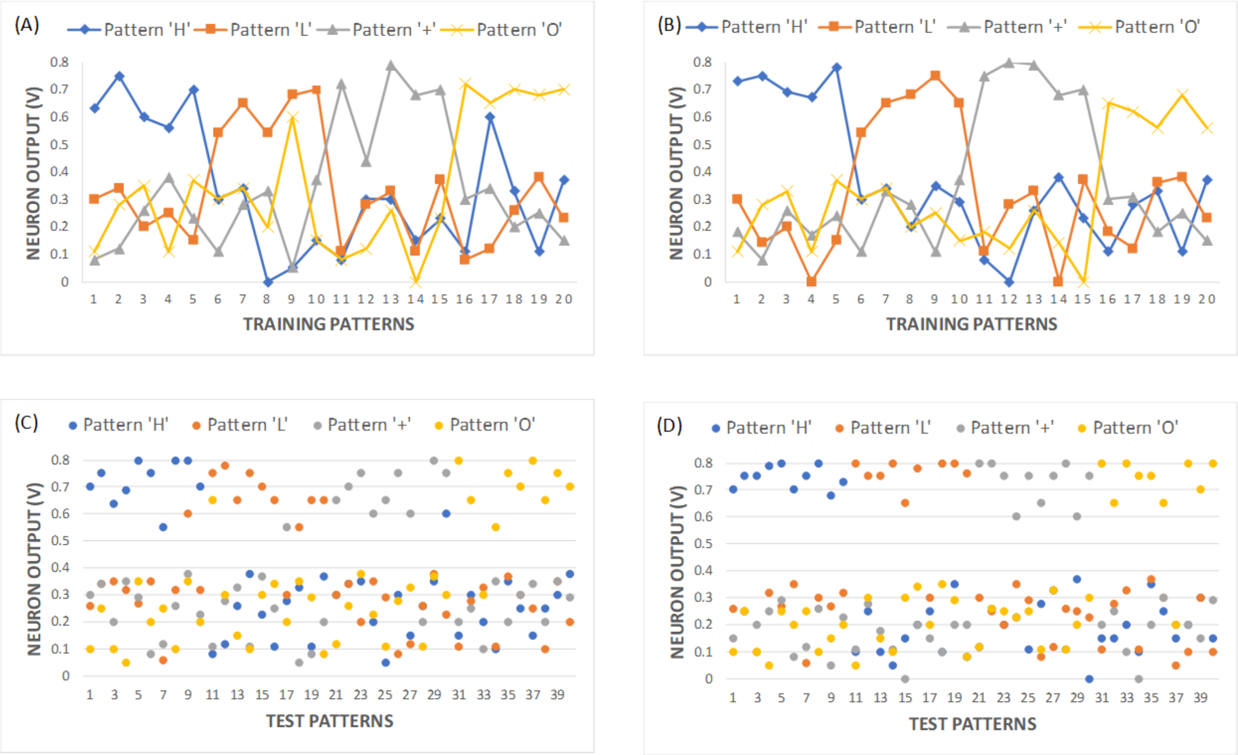


Figure 19: Pattern recognition accuracy results for different weight resolutions

Figure 21 shows the Pattern recognition accuracy results for four different cases. Two weight resolutions of 100µS (10kΩ) and 200µS (5kΩ) is considered for this work. A resolution of 100µS provides slightly more than 4-bit precision for the ranges of conductances considered while a resolution of 200µS provides 5-bit precision. The differential weights can only be these values or integer multiples of them. The x-axis of the plots represents the patterns fed to the MLP network whereas the y-axis represents the neuron output of four output neurons of MLP network. If an output neuron outputs a value above 0.4 V, then it signifies the successful classification of the applied input pattern. In

Figure 21(A), MLP network for the classification of trained input patterns with weights having a resolution of 100µS is studied. The plot demonstrates that a 10µ resolution is too coarse to be acceptable as the misclassification rate is 10% for the training patterns. Usually MLP networks perform better for input patterns that they were trained for. In Figure 21(B), classification of trained input patterns with weights having a finer resolution of 200µS is plotted. In this case, the MLP network classifies all the patterns without any errors as expected. Then the MLP network is used to classify 40 additional test patterns with one-bit flipped from the original trained patterns. Since these patterns were not used for training, the MLP network tries to classify these patterns based on what it saw during the training. Figure 21(C) shows the classification accuracy for 100µS weight resolution. Here, the misclassification rate goes to 17.5% (7 patterns wrongly classified out of 40) for the resolution of 100µS which is not unacceptable. Figure 21(D) shows the classification accuracy for weight resolution of 200µS. However, the for the weight resolution of 200k, the misclassification rate is 5% (2 patterns wrongly classified out of 40) which is acceptable given that the MLP network is classifying images it hasn't seen before. This work shows that a 5-bit precision is sufficient to classify patterns in a reasonably large neural network. If memristors with higher on/off ratio are considered, then an 8-bit precision should be possible which can further improve the convergence of the MLP network.

CHAPTER 7

CONCLUSION

In this dissertation, a new 3D ASIC technology for ANNs, SkyNet, has been proposed and evaluated. SkyNet achieves 3D fine vertical integration of various components of ANN such as synaptic weights, neurons and interconnects. SkyNet enables the co-location of the synaptic weights and the neurons which is essential for any implementation of ANN.

As part of the fabric, various components are introduced, and their use in the implementation of ANNs is demonstrated. Dual-doped vertical nanowires are the building blocks of the fabric around which various components are deposited with material deposition techniques. The core devices of the SkyNet are vertical memristors, vertical junctionless p-type and n-type transistors, other components include various 3D routing structures such as bridges, contacts, interlayer-layer connection. All these components contribute to the realization of various aspects of ANNs such as synaptic weights, neurons and interconnections. The fabric allows for co-localization of synaptic weights and neurons, which is not possible with the hybrid memristor/CMOS approach. SkyNet is then used to implement a Perceptron, one of the first ANNs to be demonstrated. A comprehensive evaluation methodology is developed to evaluate the fabric. Single-layer and multi-layer perceptron is used for evaluating the fabric.

When compared to the hybrid memristor/CMOS system, SkyNet shows benefits in all aspects including performance, power and area for the ANNs considered. Finally, a study to show the impact of weight resolution on the recognition accuracy of the MLP network is studied. A comprehensive analysis on all evaluation results is applied to understand the benefits and shortcomings of SkyNet. Bottom-up evaluations for the proposed fabric considering 3D implementation of fabric components reveal up to 19x density, 1.2x power benefits when compared to 16nm hybrid memristor/CMOS technology.

BIBLIOGRAPHY

[1] Mead, Carver. "Neuromorphic electronic systems." Proceedings of the IEEE 78.10 (1990): 1629-1636.

[2] F. Akopyan et al., “TrueNorth: Design and Tool Flow of a 65 mW 1 Million Neuron Programmable Neurosynaptic Chip,” IEEE Trans. Comput. Des. Integr. Circuits Syst., vol. 34, no. 10, pp. 1537–1557, 2015.

[3] Strukov, Dmitri B., et al. "Hybrid CMOS/memristor circuits." Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on. IEEE, 2010.

[4] Kim, Kuk-Hwan, et al. "A functional hybrid memristor crossbar-array/CMOS system for data storage and neuromorphic applications." Nano Letters 12.1 (2011): 389-395.

[5] Sacchetto, Davide, et al. "Resistive programmable through-silicon vias for reconfigurable 3-D fabrics." IEEE Transactions on Nanotechnology 11.1 (2012): 8-11.

[6] Zamarreño-Ramos, Carlos, et al. "Multicasting mesh AER: a scalable assembly approach for reconfigurable neuromorphic structured AER systems. application to ConvNets." IEEE transactions on biomedical circuits and systems 7.1 (2013): 82-102.

[7] M. Rahman, S. Khasanvis, J. Shi, M. Li, C. A. Moritz. "Skybridge: 3D Integrated Circuit Technology Alternative to CMOS." Available Online: http://arxiv.org/abs/1404.0607.

[8] Shi, Jiajun, et al. "NP-Dynamic Skybridge: A Fine-grained 3D IC Technology with NP-Dynamic Logic." IEEE Transactions on Emerging Topics in Computing (2017).

[9] Li, Mingyu, et al. "Skybridge-3D-CMOS: A Vertically-Composed Fine-Grained 3D CMOS Integrated Circuit Technology." VLSI (ISVLSI), 2016 IEEE Computer Society Annual Symposium on. IEEE, 2016.

[10] Li, Mingyu, et al. "Skybridge-3D-CMOS: A Vertically-Composed Fine-Grained 3D CMOS Integrated Circuit Technology." IEEE Transactions on Nanotechnology, In press, 2017.

[11] Batude, P., et al. "Advances in 3D CMOS sequential integration." Electron Devices Meeting (IEDM), 2009 IEEE International. IEEE, 2009.

[12] Strukov, Dmitri B., et al. "The missing memristor found." nature 453.7191 (2008): 80-83.

[13] Govoreanu, B., et al. "10× 10nm 2 Hf/HfO x crossbar resistive RAM with excellent performance, reliability and low-energy operation." Electron Devices Meeting (IEDM), 2011 IEEE International. IEEE, 2011.

[14] Yang, J. Joshua, et al. "Engineering nonlinearity into memristors for passive crossbar applications." Appl. Phys. Lett 100.11 (2012): 113501.

[15] Williamson, Adam, et al. "Synaptic behavior and STDP of asymmetric nanoscale memristors in biohybrid systems." Nanoscale 5.16 (2013): 7297-7303.

[16] Lee, Chi-Woo, Aryan Afzalian, Nima Dehdashti Akhavan, Ran Yan, Isabelle Ferain, and Jean-Pierre Colinge. "Junctionless multigate field-effect transistor." Applied Physics Letters 94, no. 5 (2009): 053511.

[17] Rahman, Mostafizur, et al. "Experimental prototyping of beyond-CMOS nanowire computing fabrics." Nanoscale Architectures (NANOARCH), 2013 IEEE/ACM International Symposium on. IEEE, 2013.

[18] Yakopcic, Chris, et al. "SPICE analysis of dense memristor crossbars for low power neuromorphic processor designs." Aerospace and Electronics Conference (NAECON), 2015 National. IEEE, 2015.

[19] Zidan, Mohammed Affan, et al. "Memristor-based memory: The sneak paths problem and solutions." Microelectronics Journal 44.2 (2013): 176-183.

[20] Kvatinsky, Shahar, et al. "VTEAM: A general model for voltage-controlled memristors." IEEE Transactions on Circuits and Systems II: Express Briefs 62.8 (2015): 786-790.

[21] Alibart, Fabien, Elham Zamanidoost, and Dmitri B. Strukov. "Pattern classification by memristive crossbar circuits using ex-situ and in-situ training." Nature Communications 4 (2013).

[22] Shi, Jiajun, et al. "Architecting NP-Dynamic Skybridge." Nanoscale Architectures (NANOARCH), 2015 IEEE/ACM International Symposium on. IEEE, 2015.

[23] Arizona State University. PTM-MG device models for 16nm node, <www.ptm.asu.edu>.

[24] Rosenblatt, Frank. "Principles of neurodynamics." (1962).