Conductive Bridge Random Access Memory (CBRAM) as an Analog Synapse for Neuromorphic Computing

by

Priyanka Apsangi

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved July 2022 by the
Graduate Supervisory Committee:

Hugh Barnaby, Chair
Michael Kozicki
Ivan Sanchez Esqueda
Matthew Marinella

ARIZONA STATE UNIVERSITY

August 2022

ABSTRACT

The Deep Neural Network (DNN) is one type of a neuromorphic computing approach that has gained substantial interest today. To achieve continuous improvement in accuracy, the depth, and the size of the deep neural network needs to significantly increase. As the scale of the neural network increases, it poses a severe challenge to its hardware implementation with conventional Computer Processing Unit (CPU) and Graphic Processing Unit (GPU) from the perspective of power, computation, and memory. To address this challenge, domain specific specialized digital neural network accelerators based on Field Programmable Gate Array (FPGAs) and Application Specific Integrated Circuits (ASICs) have been developed. However, limitations still exist in terms of on-chip memory capacity, and off-chip memory access. As an alternative, Resistive Random Access Memories (RRAMs), have been proposed to store weights on chip with higher density and enabling fast analog computation with low power consumption. Conductive Bridge Random Access Memories (CBRAMs) is a subset of RRAMs, whose conductance states is defined by the existence and modulation of a conductive metal filament. Ag-Chalcogenide based Conductive Bridge RAM (CBRAM) devices have demonstrated multiple resistive states making them potential candidates for use as analog synapses in neuromorphic hardware. In this work the use of Ag-$Ge_{30}Se_{70}$ device as an analog synaptic device has been explored. Ag-$Ge_{30}Se_{70}$ CBRAM crossbar array was fabricated. The fabricated crossbar devices were subjected to different pulsing schemes and conductance linearity response was analyzed. An improved linear response of the devices from a non-linearity factor of 6.65 to 1 for potentiation and -2.25 to -0.95 for depression with non-identical pulse application is observed. The effect of improved linearity was quantified by simulating the devices in an artificial neural network. Simulations for area, latency, and power consumption of the CBRAM device in a neural accelerator was conducted. Further, the changes caused by Total Ionizing Dose (TID) in the conductance of the analog response of Ag-$Ge_{30}Se_{70}$ Conductive Bridge Random Access Memory (CBRAM)-based synapses are studied. The effect of irradiation was further analyzed by simulating the devices in an artificial neural network. Material characterization was performed to understand the change in conductance observed due to TID.

To my grandparents

ACKNOWLEDGMENTS

I would like to express my sincerest gratitude to my advisor Dr. Hugh Barnaby for his support of my research. I would also like to thank Dr. Michael Kozicki for his guidance and constant encouragement throughout this work. I would like to extend my gratitude to Dr. Barnaby and Dr. Kozicki for their invaluable mentorship and influencing me as a person and a researcher during formative years of my career in an immensely positive way.

I am grateful to Dr. Ivan Sanchez Esqueda for always being interested in my research and providing a fresh perspective. I am thankful to Dr. Matthew Marinella for agreeing to be my committee member and giving me valuable advice for my thesis. I am thankful to Dr. Yago Gonzalez-Velo who has been a great mentor and has advised me throughout my research career at ASU. I am grateful to my colleagues, collaborators and friends Dr. Jennifer Taggart, Dr. Mehmet Balaban, Dr. Ninad Chamele, Arshey Patadia, Smitha Sekhar Swain, Dr. Kiraneswar Muthuseenu, Dr Aymeric Privat and Xu Han and many others who supported my research and enabled completion of this work.

I would like to thank the staff at Center for Solid State Electronics Research and The Eyring Materials Center at Arizona State University for their assistance and guidance with nanofabrication and characterization. I am also thankful for the staff at School of Electrical, Computer and Energy Engineering for their support and advice throughout my graduate education. I would like to thank various funding agencies, Defense Threat Reduction Agency, and Northern Arizona University for their financial support throughout my graduate research. I would also like to thank ON Semiconductor and Western Digital for providing me internship opportunities during my studies.

I want to thank my dearest friend and my husband Ishaan Puranik for his constant support and love. I am grateful to my sister Shreya Apsangi for always being there for me. Lastly and most importantly, I would like to thank my parents and grandparents for their love, support, and sacrifices to which I owe everything.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

CHAPTER 1

INTRODUCTION

1.1.Overview of Deep Neural Networks (DNN)

Deep Neural Network (DNNs), one class of Artificial Neural Networks (ANNs), have recently had a major impact on how smart devices interpret, respond, and interact to large amounts and different types of information. DNN technologies are being developed for sophisticated cognitive applications that utilize interfaces more intuitive than the traditional keyboard and mouse or even the popular approach of touchscreen. The most famous, recent examples of DNN applications are Virtual Personal Assistants like the Apple's Siri, Microsoft's Cortana, and Google's Assistant [1]-[3]. These applications use efficient deep learning algorithms that are used on daily basis by millions of people. Nowadays almost every high-technology industry is influenced by the field of deep learning. Some examples include automotive, defense, manufacturing, and gaming industries [4]. DNNs have been shown to be well suited for performing a broad range of classification and decision-making tasks such as speech recognition [5], image processing [6], and machine translations [7].

Deep Neural Networks (DNNs) are neural networks inspired by the biological neural networks in the human brain. They consist of a collection of nodes, called artificial neurons, and connections between the nodes, called artificial synapses. The neurons are organized in layers. Typically, a DNN consists of an input layer (layer 0) and an output layer. Any layers between the two layers are called hidden layers. Figure 1 represents a 2-layer DNN, where the black dots represent neurons and the lines connecting the dots are synapses.

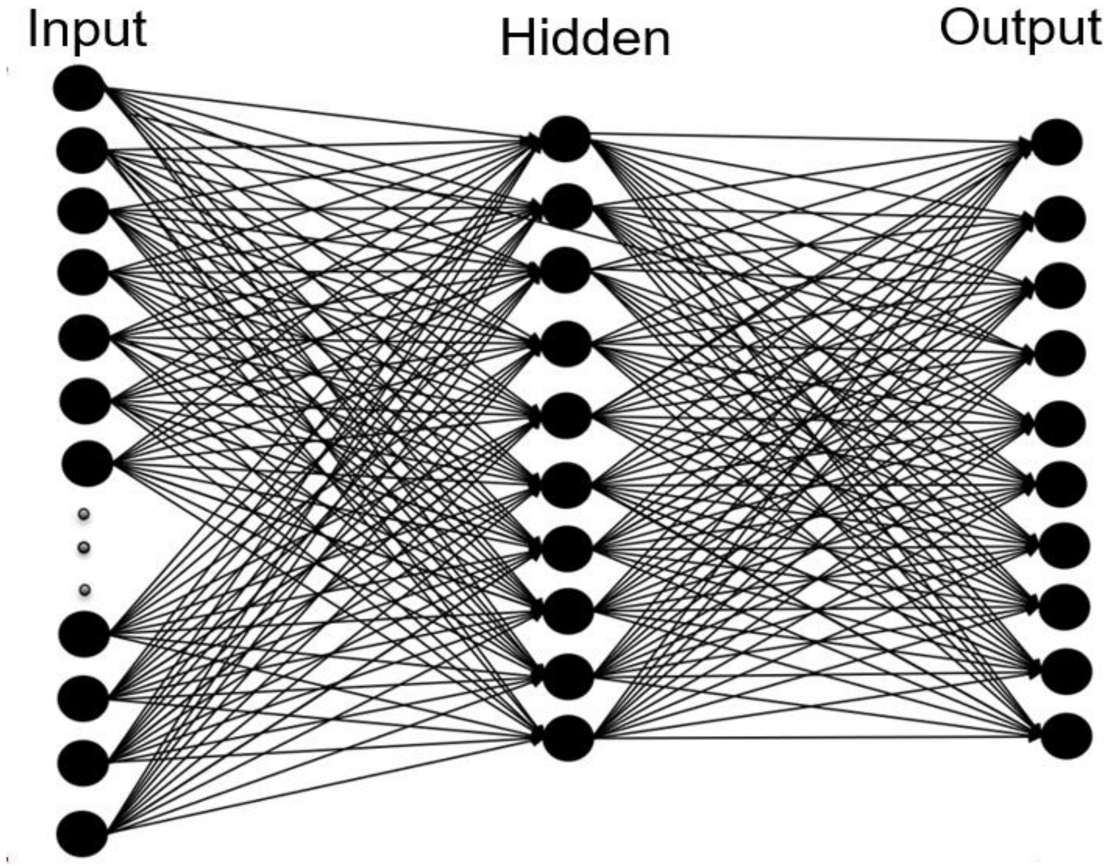


Figure 1. A simple deep neural network (DNNs).

For a fully connected DNN system, each neuron in a layer receives signal x_j from all neurons in the previous layer. The signal is then weighted by a scaling factor known as synaptic weight w_j. Mathematically, it can be represented as-

$$y = \sum_{j=0}^{j=n} x_j w_j \tag{1.1}$$

The weighted output signal, y, is then passed through a non-linear activation function as shown in Figure 2. The non-linear function, also known as squashing function, is typically a sigmoid function, but can also be a hyperbolic tangent [8] or a ReLU (rectified linear unit) [9] function. Mathematically, the output for a sigmoid non-linear activation function can be expressed as

$$Output = \frac{1}{1+ e^{-y}} \quad . \tag{1.2}$$

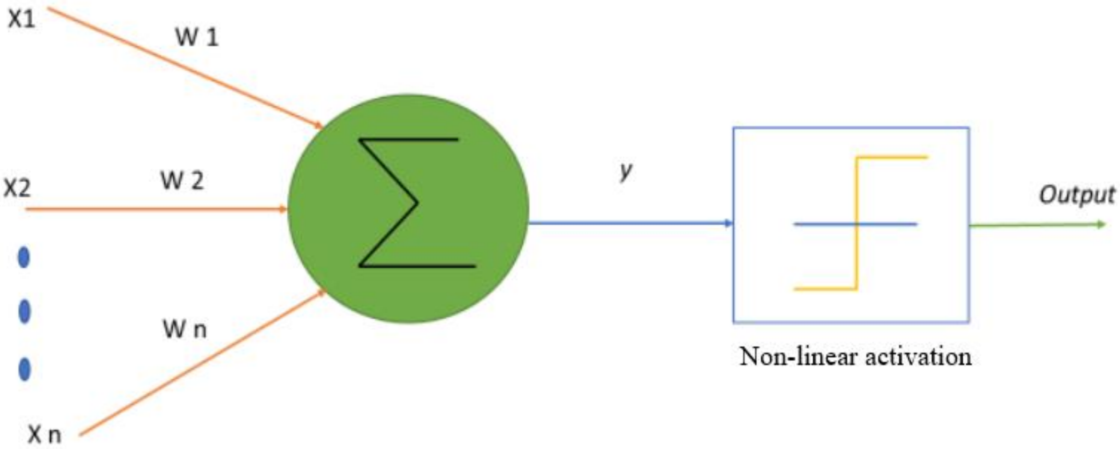


Figure 2. Model of a neuron

1.2. Backpropagation Algorithm

One of the most popular algorithms used to train a neural network for supervised learning has been the backpropagation algorithm [10], [11],[12]. For training the neural network, the number of input (layer 0) neurons is typically chosen to match the size of incoming data set (e.g., pixel of an image or an audio sequence etc.). In simple terms, the meaning of training is optimization. We need to optimize the neural network to minimize the error between the target output and the output produced after forward propagation of the input signal. The training typically begins with random weight values assigned to the synapses. The backpropagation algorithm adjusts the synaptic weights with gradient descent to minimize the sum squared error between the network’s targeted values and the values at the given iteration step. Consider the 2-layer network in Figure 3 showing a hidden layer sandwiched between with the input and output layers.

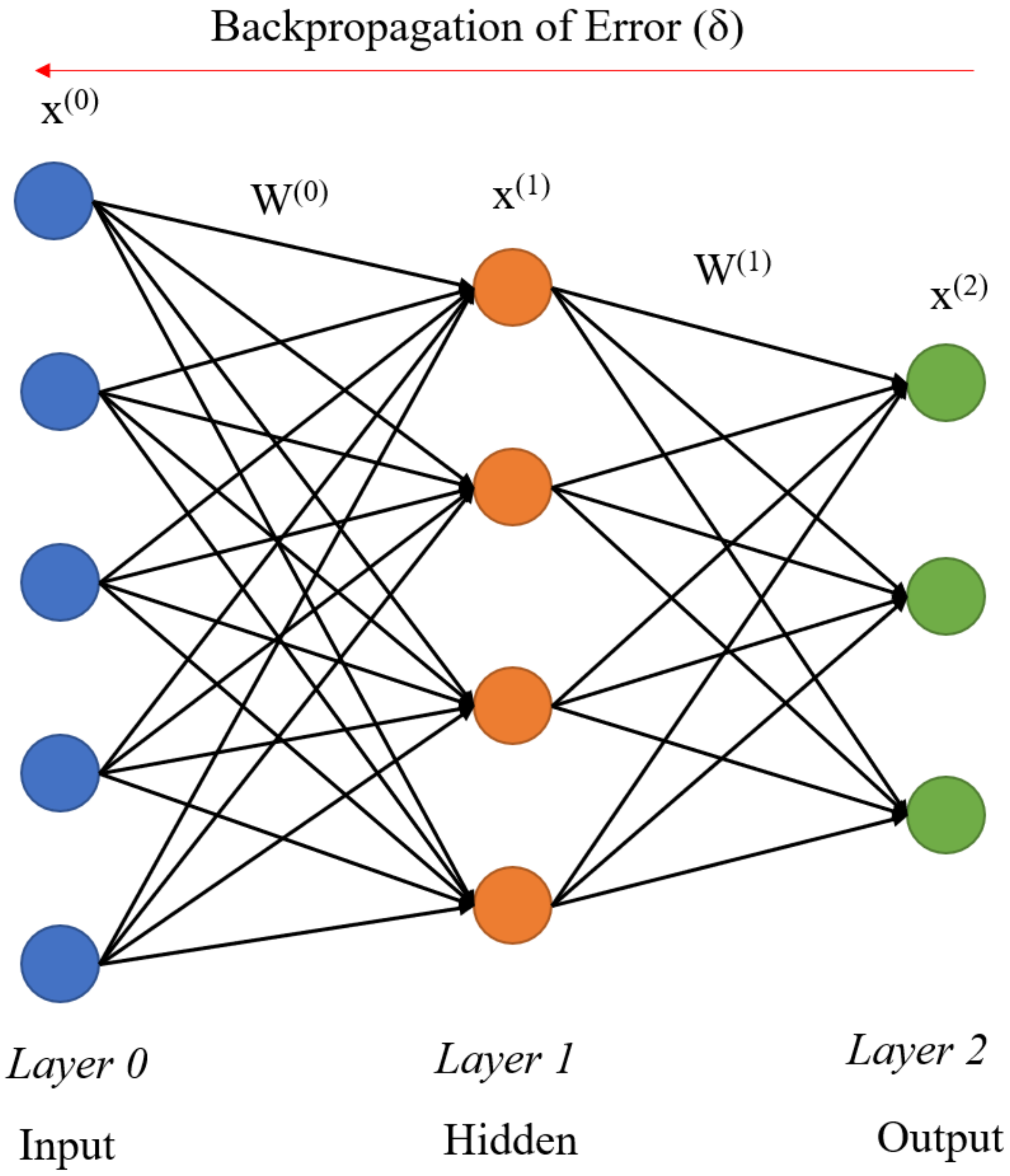


Figure 3. A multi-layer perceptron neural network, showing a single hidden layer. A feed forward network propagates activations to produce an output and the error is backward propagated to determine the weight adjustments.

In the network illustrated Figure 3, $W^{(1)}$ represents the weight set between the hidden and output layers, $W^{(0)}$denotes the weights between the input and hidden layers. The activation functions are denoted as f. Signals at the input layer neurons, x $^{(0)}$, are mapped converted to the hidden layer values, x$^{(1)}$, by the general activation function

$$x^{(1)} = f\left(W^{(0)} x^{(0)}\right), \tag{1.3}$$

where $W^{(0)}$ is a M (column) x N (row) matrix of synaptic weights connecting layer 0 with M neurons to layer 1 with N neurons. Similarly, this can be extended to the output layer 2, where the activation function can be mathematically represented as-

$$x^{\,(2)} = f\left(W^{(1)} x^{(1)}\right). \tag{1.4}$$

The non-linear activation function, $f,$ ensures that the multiple layers in the network do not collapse trivially into an equivalent single-layer network. The propagation of input data through the network which produces an output prediction is termed as inference step. The neural network is trained by to make accurate predictions by iteratively updating weights $W^{(i)}$ so that the outputs approach the target values over repeated training steps. Backpropagating the error is the most widely used algorithm for training. For a given training example set, first an inference step is carried out. Next, an error value, δ, is obtained at the output layer (2) which is a usually a mean squared function that can be mathematically represented as

$$\delta^2 = x^{(2)} - y \;\; , \tag{1.5}$$

where y is the initially obtained output after first inference step. Similarly, the error at the hidden layer (1) can be calculated as –

$$\delta^1 = \; (W^{(1)})^{\,T} \, \delta^2 \odot f'\left(W^{(1)} x^{(1)}\right), \tag{1.6}$$

where f' is the derivative of f and ⊙ is the element - wise dot product . One can observed that the above expression involves the multiplication of the transpose of the weight matrix with an error value. This operation is referred toa s the Matrix Vector Multiplication (MVM). There is no error value associated with the layer (0) because it corresponds to the given input signals.

Once error vectors are obtained at each layer in the network, these values are used to update the weights. The update applied to each weight matrix in the output layer (2) is given by –

$$\varDelta W^{(1)} = -\eta\delta^{(2)}(x^{(2)})^T, \tag{1.7}$$

Similarly for the hidden layer (1), the weight update to be applied is given by –

$$\varDelta W^{(0)} = -\eta\delta^{(1)}(x^{(1)})^T, \tag{1.8}$$

where η is the learning rate. The optimal update to the weights is termed as the outer product update of the two vectors.

The weight update is carried out after each training step. This type of gradient descent is known as stochastic gradient descent. To speed up training, various methods such as batch gradient descent or mini batch gradient descent are utilized. For batch gradient descent, the entire training dataset is feedforwarded through the network and the gradient is calculated for the entire training dataset. Mini batch gradient descent is a type of gradient descent in which only part of training dataset is utilized for calculating the gradient descent. Parameters such as the number of hidden layers and the learning rates are typically fixed during training. The goal of the designer is to have a network which is well adapted to the test set.

CHAPTER 2

HARDWARE FOR NEUROMORPHIC COMPUTING

The recent advances in the field of machine learning have led to a tremendous increase in the depth and the size of the neural network. Neural networks developed for computer vision can now have on the order of 10^5 - 10^{10} weights [14], [15]. The training of such large-scale networks becomes difficult due to the demands placed on computational resources. It is also very challenging to deploy inference applications for such networks on the edge devices (such as mobile phones) which lack the required infrastructure like- the storage, computational capacity as well as the power requirement. Progress made in algorithms which have made the neural networks compact without losing accuracy. One such approach is quantization; we reduce the need for high bit precision for the weights and activation functions. Quantization not only reduces the size of the network but also makes significant gains in component density and memory capacity [16]. Such advances still suffer from the need to process large amounts of data [17]. Hence it becomes necessary to develop hardware innovations to achieve large and high-performance neural networks. The use of conventional computing architecture like CPU which is specialized for handling only one or few complicated instructions at one time, is not very well suited for tasks like training or inference for large-scale neural networks.

The use of conventional computing architecture like CPU which is specialized for handling only one or few complicated instructions at one time, is not very well suited for tasks like training or inference for large-scale neural networks. The use of Graphic Processing Units (GPUs) with its multiple cores that compute in parallel has helped to improve for the neural network processing throughput [18]. But despite having significant advantages over a CPU, the GPU too suffers from memory transfer which is a bottleneck in processing large data sets [19].

2.1. Digital Neuromorphic Architecture

The difficulties posed using conventional CPUs or GPUs for data-centric applications has led to research and development of customized domain-specific hardware for machine learning. Field Programmable Gate Arrays (FPGA's) is one suggested platform for deep learning networks [20]-[24]. The reconfigurability advantage of the FPGA enables greater flexibility in hardware and algorithm design. However, multiple FPGA's might need to be implemented to support this reconfigurability which lead to increase in area and power consumption. The above shortcomings have been balanced out using arrays of FPGA's interconnected by high-speed reconfigurable networks [25]. One such example is the Microsoft's Brainwave project that uses multiple FPGAs to perform inference with relatively low latency [26].

Another example in digital neuromorphic architecture is Google's Tensor Processing Units (TPUs) [27]. Figure 4 shows a representative system diagram of Google's TPU. The TPU v1 is an inference accelerator which contains at its core a 256 X 256 block of 8-bit multiply-accumulate (MAC) units that perform individual multiplication and addition operations. As depicted in Figure 4 the block is operated as systolic array where the input data is fed into the array horizontally and VMM partial sums are accumulated from the top to downward approach. A systolic array, a term used in parallel computation, is a homogeneous network of tightly coupled data processing units (DPUs) called cells or nodes. Each node or DPU independently computes a partial result as a function of the data received from its upstream neighbors, stores the result within itself and passes it downstream [28]. For every clock cycle, each MAC unit receives an input and partial sum from its neighbors, performs a computation and then forwards the updated partial sum to the next unit. This architecture provides the advantage of higher compute density and reduces the number of intermediate reads and writes to a buffer unit during VMM operation [29].

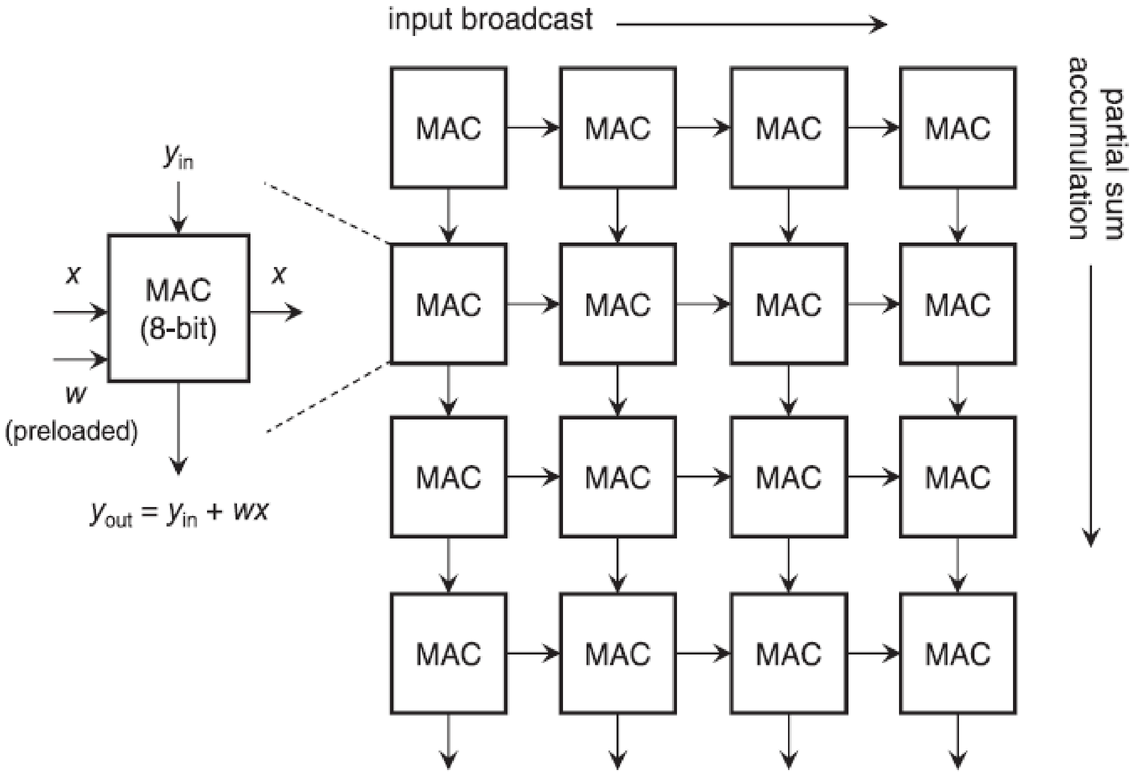


Figure 4. A systolic array of 8-bit MAC units in the Google TPU v1 [29] [13].

The main bottleneck for digital neuromorphic hardware is the off-chip memory access. One way the issue is addressed is to place the memory bank optimally close to the processing element to initiate quick transfer of intermediate results. Examples are DaDianNao [30] and Neurocube [31].

By performing computation inside modified digital memory array, the problem of the memory access bottleneck can be addressed. This is done by exploiting the internal bandwidth of dynamic random access memory (DRAM) and static random access memory (SRAM) technology. The most prominent example of digital process in memory (PIM) is Ambit, which implements a logically complete set of operations within a DRAM array with minimal additional area overhead [32]. Ambit exploits the analog operation of DRAM technology to perform the bitwise operations completely inside DRAM, thus using the full internal DRAM bandwidth. Ambit constitutes two main components – (a). It uses simultaneous activation of three DRAM rows that share the same set of sense amplifiers that enables the system to perform the bitwise AND and OR operations. (b). It uses the inverters present inside the sense amplifiers, with modest changes to perform bitwise NOT operations. Although the bitwise operations in Ambit are logically complete, many cycles are needed to execute multi-bit arithmetic operations – hundreds of cycles to compute multiplications with three bits or more [32]. The disadvantage is the latency which is introduced due to the need to have multiple cycles to execute multi-bit operations. Another disadvantage associated with DRAM is that it suffers from scalability limitations.

2.2. Analog Neuromorphic Architecture

Analog neuromorphic accelerators can significantly improve the energy and latency costs associated with data movement by embedding the neural network computations directly inside the memory elements. This type of analog architecture is inherently parallel and can perform the neural network basic operations like the VMM, MVM and outer product update as discussed in Chapter 1. Therefore, this architecture is efficient for enabling both inference and training. A generalized conductance/resistance-based memory array circuit is depicted in Figure 5 can be realized using various types of NVM devices. The vector matrix multiplication (VMM) operation is performed within the array by activating all the rows (or word lines) simultaneously with voltage V_i on row *i*th. The total current is collected by the *j*th column. Mathematically represented as –

$$I_j = \sum_{i=0}^{N_c-1} G_{ij} V_i \qquad\qquad 0 < j < N_c - 1 \qquad\qquad (2.1)$$

Here, G_{ij} is the conductance of the memory element at the array position (i, j) and N_c represents the number of rows. With the above equation, the implementation of a vector dot product is realized using the Ohm's law and the summation is executed by using the Kirchhoff's current law down each column. This crossbar architecture can execute the VMM operation in a single step. The transpose MVM operations can be executed driving the columns and reading the currents on the rows. The outer product update which constitutes an important aspect of training a neural network is carried out by applying programming pulses simultaneously to all the rows and columns of a crossbar which is executed in a parallel fashion [13],[33].

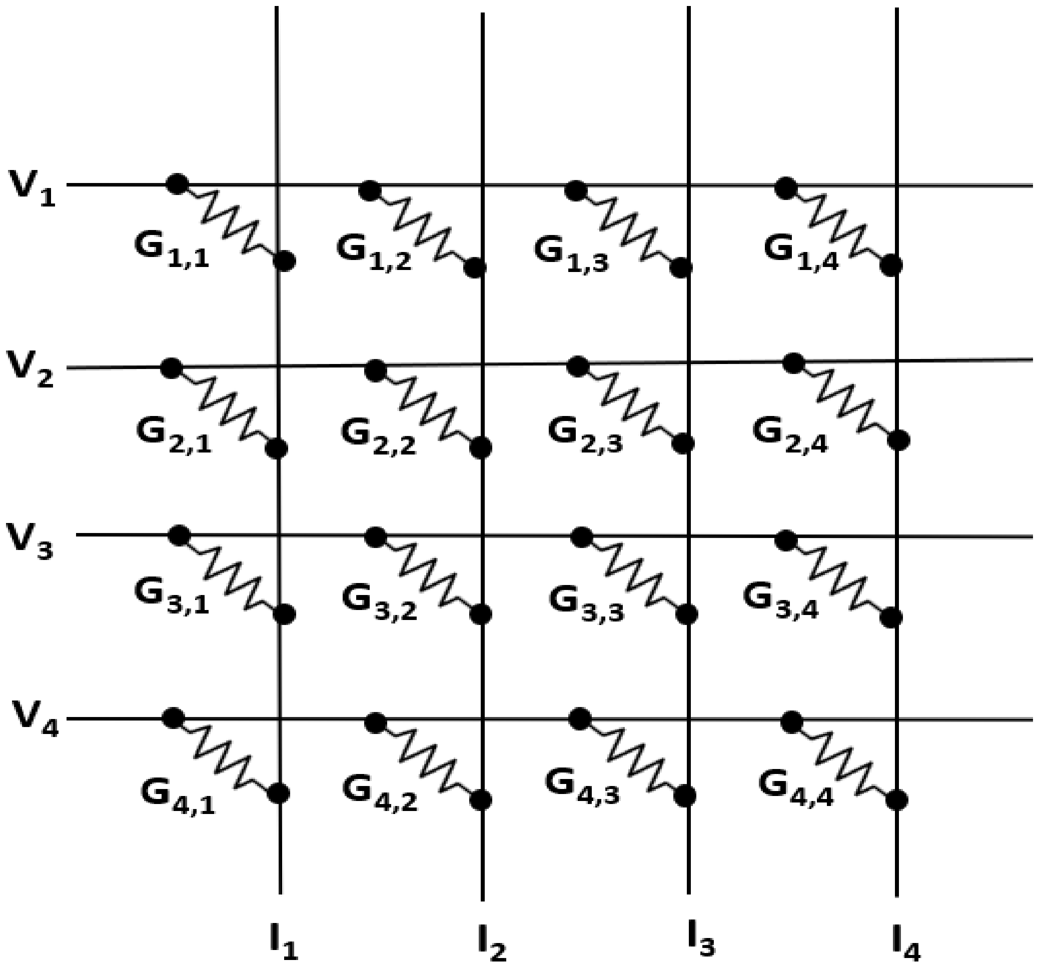


Figure 5. A basic representation of resistive array architecture performing dot product and summation in parallel.

The resistive crossbar array has more favorable energy scaling compared to digital accelerators. The latency of a crossbar array is dependent on the time it takes to charge each row of the array. As the number of columns increase for an array, the resistance and capacitance of the rows increase as well hence scaling the latency by O (N^2) for an N X N array [31]. The resistive crossbar array has much more favorable energy scaling as compared to the digital accelerators. For VMM computation on the digital accelerators (i.e., DRAM arrays), requires charging one row at a time and then one column at a time for each cell. Hence, computing a VMM using a SRAM or DRAM memory is energy intensive as it scales N^3 while for an analog memory it scales as N^2 [32]. It is interesting to note that the in case of analog accelerators, the area and energy consumption is dominated by the peripheral circuitry and not the crossbar itself [31].

2.3. Desirable performance metrics for analog synaptic devices

a. Device dimensions

One of the key requirements for synaptic devices is the potential scalability of the devices to nanometer regime. To integrate large scale neural networks, synaptic devices with small device footprints is necessary. Today synaptic devices with scalability down to sub-10 -nm regime is preferred. Among different synaptic devices, Phase change memory (PCM) and Resistive RAM (RRAM) devices have shown the capability to scale to sub-10nm regime [33], [34] for digital memory applications. However, here we can argue that a tradeoff can exist between the analog synaptic characteristics and scalability.

b. Multilevel conductance states

The synaptic plasticity showed by a biological synapse shows multilevel states. Pattern recognition and learning are implemented using neuro inspired algorithm and require multilevel conductance states for accuracy and fast learning. In brief we can say that the higher the level of conductance the better the learning accuracy and robustness of the system. Ideally a device to be implemented as a synaptic device in neural hardware should show >100 conductance states [35]-[37].

c. Dynamic Range

Dynamic Range is the ratio of maximum conductance and minimum conductance. An ideal synaptic device should show a dynamic range of 100 or greater [38]. This is important from the point of view of mapping the weights in a neural network. The weights in the algorithm are normalized. For example, if the dynamic range of a given device is 1000, therefore, the lowest weight can be represented as 0.001. For online training it is important that there are many available weights for accurate training.

d. Symmetry and Linearity in Weight Update

The linearity refers to linear behavior between conductance and increasing pulse number. This linear and symmetric behavior is important from the point of view weight mapping in algorithms and to improve the accuracy in pattern recognition. This linear and symmetric behavior is an ideal characteristic. The real devices show asymmetry and nonlinear weight update characteristics. For all most all present RRAM devices, there is an asymmetry between potentiation and depression characteristics which impede their application for pattern recognition and online training. The present devices show an increase in conductance for initial pulse train but tend to saturate at the end. Recent results have shown that this nonlinearity/asymmetry has caused the learning accuracy loss in the neural networks [39], [40].

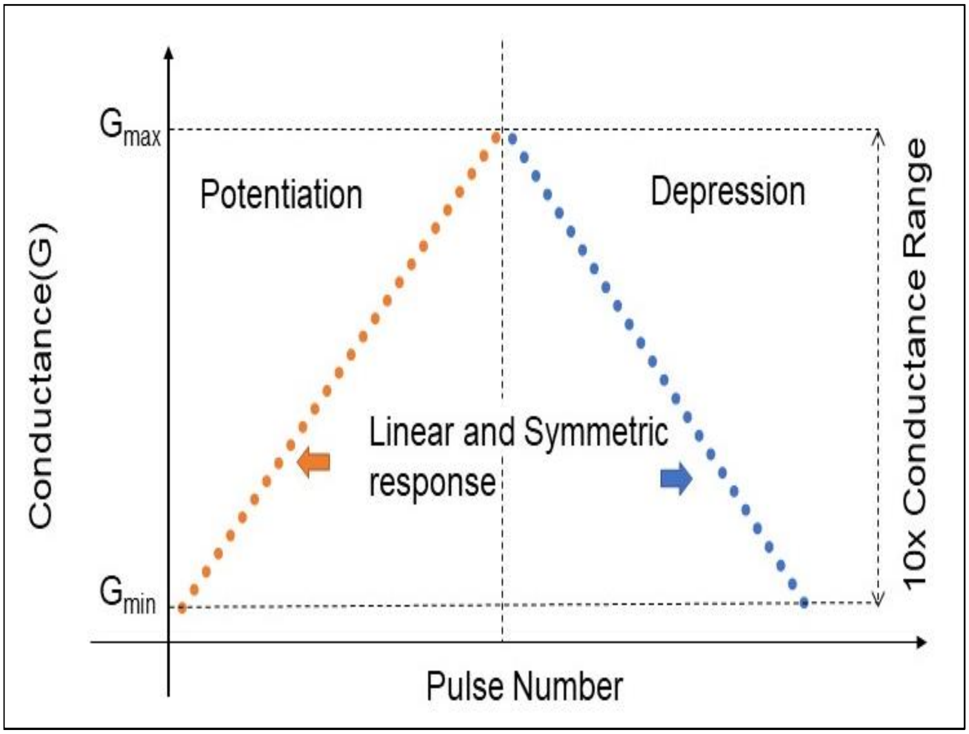


Figure 6. An ideal linear response of synaptic device.

e. Energy Consumption

For an ideal synaptic device, the energy consumption should be < f J per synaptic event if we want to implement the synaptic characteristics shown by a biological synapse. Most RRAM devices show a programming energy around 100 fJ~10 pJ, while most PCM devices may have even higher programming energy of 10~100 pJ. To get close to a biological synapse, engineering efforts need to be implemented in the current RRAM devices to reduce energy consumption and have faster programming speeds of several ns.

f. Data retention

The data retention of RRAM devices should be somewhere around 10 years at an operating temperature of about 85. The number of cycling endurances is very application dependent, relying on how many weight updates are required in the training processes. For a relatively simple task (i.e., the MNIST handwritten digit recognition [41]), 60000 training images with 50 training epochs give a maximum weight update possibility to be 3 × 106 updates. Not every synapse is updated in training in each cycle, thus the endurance of ~ 104 cycles is enough for training MNIST data set [42]. However, considering more challenging tasks (i.e., ImageNet Challenge [43]), much higher endurance may be required.

Table 1. Summary of performance metrics for analog synaptic devices [44].

PARAMETERS	TARGETS
Device dimension	<10nm
Multilevel conductance	>100
Dynamic range	>100
Non-Linearity	0/0
On-state Resistance (R_{ON})	100 kΩ -1 MΩ
Write Voltage	0.5 V-1V
Retention	>10 years
Endurance	$>10^9$

2.4. Conductive bridge random access memories (CBRAMs)

Resistive random access memories (RRAMs) have emerged as the most popular choice for the analog neuromorphic architectures. RRAM devices have shown interesting properties like scalability to sub lithographic limit, multilevel conductance characteristics, low power operation, data retention and endurance. The RRAM devices can be fabricated in the X - bar architecture which helps in the implementation of matrix multiplication functions. The conductance states of the RRAM devices are used to store the synaptic weights within the neural network. The computing power efficiency of an x-point RRAM array has been estimated to be 31000 times than the state-of-the-art microprocessor, with ultra-low power requirement and real time pattern recognition capability [44].

RRAMs are simple metal-insulator-metal structures that show a controllable resistance change. The resistance change of these devices relies on the movement of ions within the material sandwiched between the electrodes. Depending on the applied electric potential, the devices can be set/programmed into low resistance state (LRS) or reset to a high resistance state (HRS). RRAM are filamentary in nature and divided into two main categories: anion based also known as OxRAM and cation based CBRAM devices.

Conductive Bridge RAM (CBRAM) is a subclass of RRAM. These are cation -based devices that are based on redox reactions leading to the formation of positively charged ions i.e., cations. The migration of such metallic cations from the anode to the cathode enables the growth of a metallic filament with nanometric dimensions [45]. The anode is made up of an electrochemically active metal which can be easily oxidized or reduced (e.g., Cu or Ag) and the cathode is made of a relatively inert metal (e.g., Ni, W, Pt). The layer sandwiched between the electrodes is a solid-state electrolyte in which the ions can drift/diffuse through it. Examples of solid-state electrolytes used are germanium selenides (Ge_xSe_y), germanium sulfides (Ge_xS_y), SiO_2, Al_2O_3 implemented successfully with Ag or Cu anodes to form ReRAM cells [46]-[52].

In this work, we focus on the Ag-$Ge_{30}Se_{70}$ CBRAM devices. The growth conductive filament across the Ag-$Ge_{30}Se_{70}$ CBRAM devices is illustrated in Figure 7.

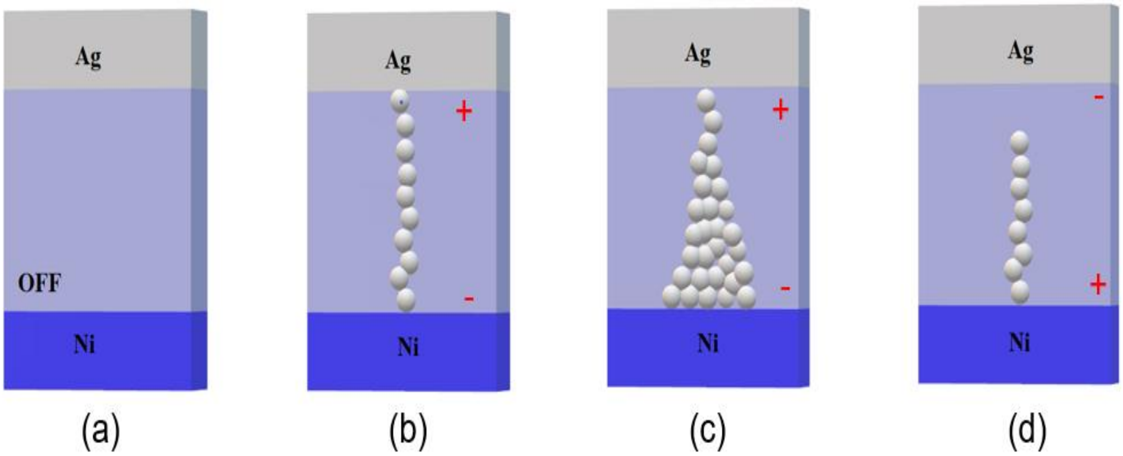


Figure 7. Schematic representation of mechanism of resistive switching in Ag-$Ge_{30}Se_{70}$ CBRAM devices. (a). Pristine condition –OFF state (b). On application of positive bias, redox reaction takes place leading to formation of conductive filament. (c). On continuous application of positive bias, the conductive filament begins to thicken. (d). Reverse bias leads to removal of the filament – Erase operation.

Referring to Figure 2.4, the device is in a high resistance state (HRS) initially. When a positive voltage is applied to the anode, the metal on the anode i.e., Ag gets oxidized (i.e., it loses electron) and Ag^+ cations are generated. The cations start drifting towards to the cathode under the influence of an electric field. As the positive potential on the device is increased, the more Ag^+ ions are generated, and they drift through the solid-state electrolyte. These Ag^+ cations transport across the electrolyte layer and are reduced to Ag at the cathode leading to the formation of a conductive filament. Once the filament bridges the anode and the cathode, the devices can be modulated over a range of "low resistance states" (LRS) by increasing or decreasing the average cross-sectional area (width) of the filament. On application of negative bias on the anode, the Ag which constitutes the filament gets oxidized within the filament on the cathode side. The Ag^+ ions formed start drifting toward the anode under the influence of the electric field. This process continues till the metallic filament is completely broken and the device resistance switches back to HRS. In this work, the use of Ag-$Ge_{30}Se_{70}$ CBRAM devices as an analog synapse for crossbar array type architecture is explored.

2.5. Dissertation Outline

Chapter 1 gives an overview of Deep Neural Networks (DNN). The computational fundamentals of deep learning i.e., inference and training the deep neural networks with backpropagation algorithm is presented. Chapter 2 begins with the needs of a deep learning hardware architectures. It provides a brief review of the digital architectures as well as analog based resistive crossbar-array based architectures specialized for deep neural networks. The requirements of an ideal resistive synaptic device for analog crossbar array architecture are also discussed. An introduction to Resistive Random Access Memories (RRAMs), a popular choice for analog resistive devices, with focus on Ag-$Ge_{30}Se_{70}$ Conductive Bridge RAM (CBRAM) is presented.

Chapter 3 discusses the fabrication of Ag-$Ge_{30}Se_{70}$ CBRAM crossbar arrays. Chapter 4 presents the DC IV characteristics of the devices. The analog response of the fabricated crossbar array to two different pulsing schemes - constant and increasing amplitude is studied. Beyond this analysis, we model, using Cross-Sim [53], an open-source software, developed by Sandia National Laboratories, how various degrees of non-linearity affect the classification accuracy of a neural accelerator that utilizes these RRAM synapses.

Using Neurosim [54], the circuit level performance of the device to the different pulsing scheme is compared.

In Chapter 5, the impact of Total Ionizing Dose (TID)on the conductance of Ag-$Ge_{30}Se_{70}$ CBRAM devices developed for use as *analog* synapses in DNN designs is analyzed. Material characterization is performed to better understand the effect the TID on the analog conductance response of these devices. Finally, in Chapter 6 the summary of this work and conclusion is offered.

CHAPTER 3

DEVICE FABRICATION

3.1. CBRAM ARRAY DEVICE FABRICATION

The NVM devices for the in-memory computing that are studied in this thesis are conductive-bridging RAM (CBRAM). These devices were fabricated at the NanoFab class 100 cleanroom facility at the at Arizona State University NanoFab. The CBRAM fabrication process is described in this chapter.

For the initial processing step, depicted in Figure 8, a Si wafer is coated with a 170 nm layer of low-pressure chemical vapor deposited (LPCVD) Si_3N_4. This layer serves as a passivation layer between the Si substrate and the CBRAM structure.

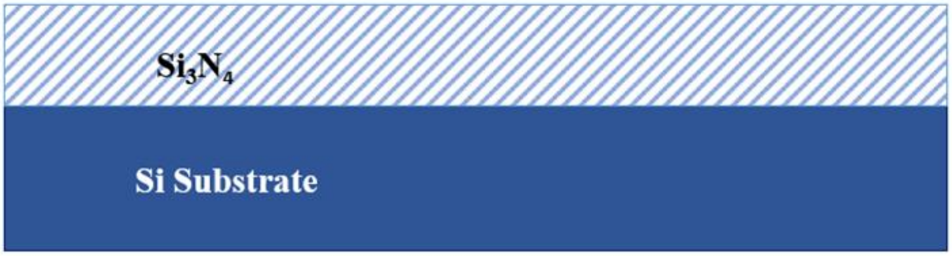


Figure 8. LPCVD deposition of Si3N4.

In the second step, a nickel cathode is created by evaporating 65nm Ni using a Lesker PVD electron-beam (e-beam) evaporator. The chamber pressure prior to deposition was 10^7 Torr. The deposition rate for nickel was 0.5 Ǻ/s to ensure smooth and even deposition of the metal.

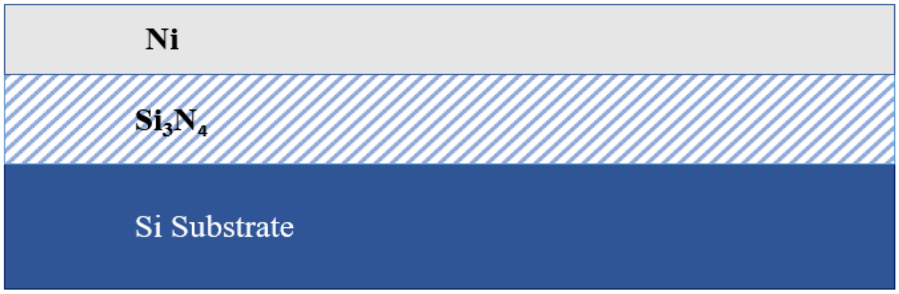
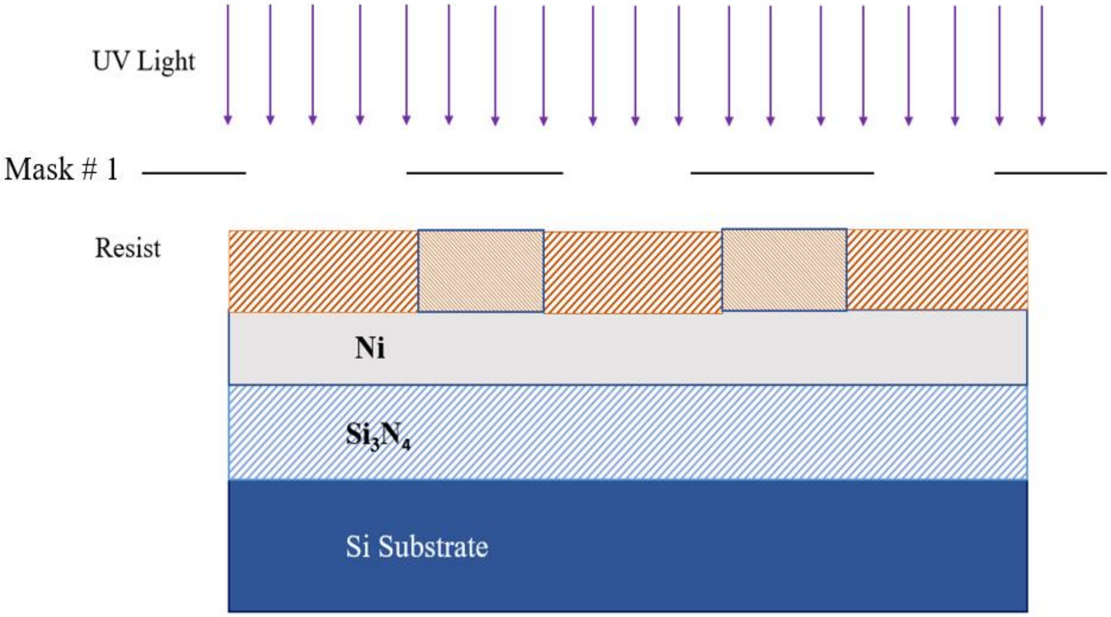


Figure 9. Deposition of Cathode (Ni) metal.

To pattern the nickel electrode, hexamethyldisilazane (HMDS) is spin coated at 3500 rpm for 30sec and baked at 100 °C for 60 s to facilitate photoresist adhesion. AZ 3312 was then spin coated on the substrate at 3500 rpm for 30 sec and baked at 100 °C for 60 seconds. The photoresist coated wafer is then patterned using EVG 620 lithography aligner set to 45mJ/cm ^{2} to create Ni cathode bars as shown in Figure 10, which is a top view of the crossbar tile. After exposure, the resist is developed for 60 seconds using Az 300 MIF developer. The resist is hard baked at 110 °C post developing to create an etch mask. The exposed nickel is chemically etched using Nickel Etchant TBF (20 % Nitric Acid solution) for 4 minutes. The remaining resist is removed by soaking the wafer in acetone solution followed by methanol and isopropyl alcohol rinse.



. Figure 10. Mask # 1 creating Nickel cathode bar.

To create isolation between the CBRAM devices, 100nm of SiO_2 is deposited using an Oxford Enhanced Chemical Vapor Deposition (PECVD) tool. To create the device vias, a double layer resist recipe was used to pattern the SiO_2 layer. First, HMDS is spun on at 4000 rpm for 30 sec followed by a bake at 120 °C for 1 minute. Next, OCG 825 is spun at 4000 rpm for 30sec followed by a bake at 105 °C for 1 minute. The first resist layer is followed by AZ 3312 spun on at 4000 rpm for 30 sec and baked at 95 °C for 1 minute. Mask # 2 was used to create device vias for etching. The resist is developed at 45 mJ / cm^{-2} and developed in AZ 300 MIF for 1 minute. The device vias are etched using anisotropic reactive ion etching (RIE) for 4 mins 30 secs.

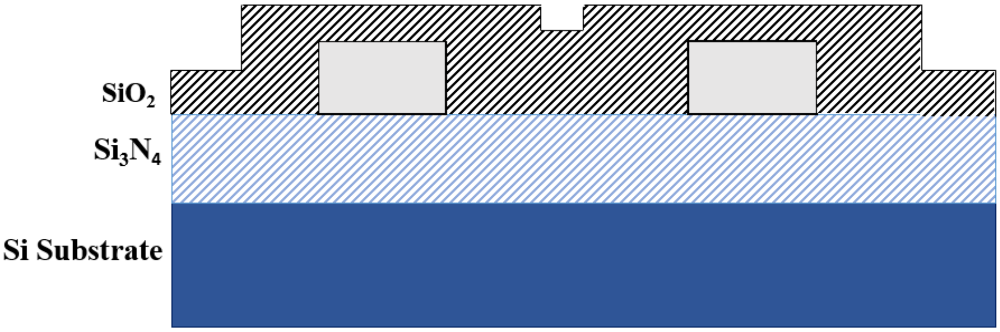


Figure 11. Ni layer chemically etched to create cathode bar followed by oxide deposition.

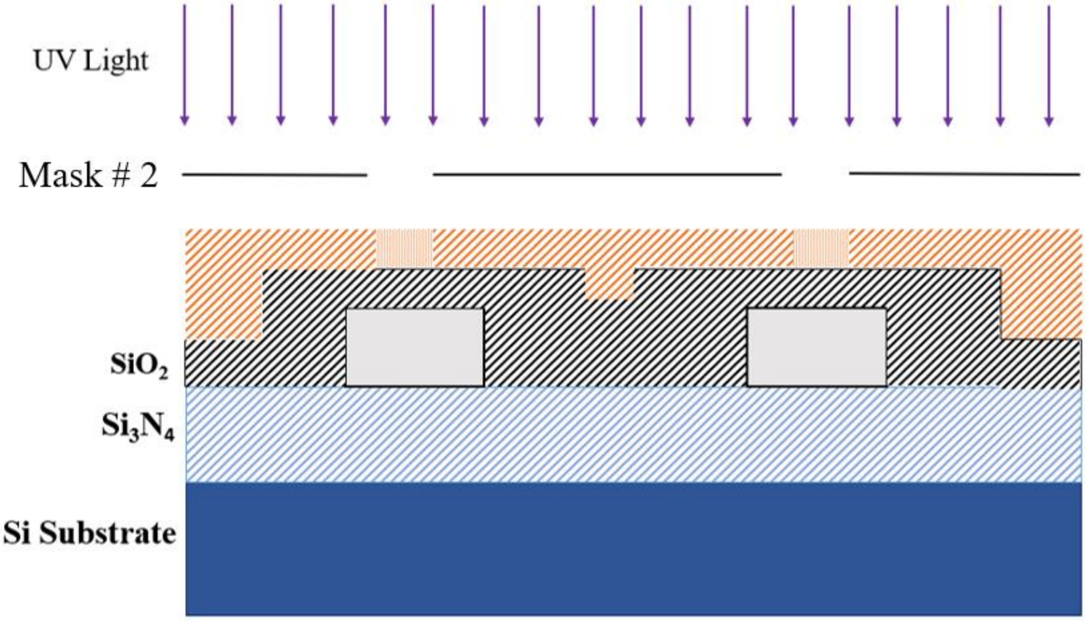


Figure 12. Mask # 2 used to etch via through SiO2.

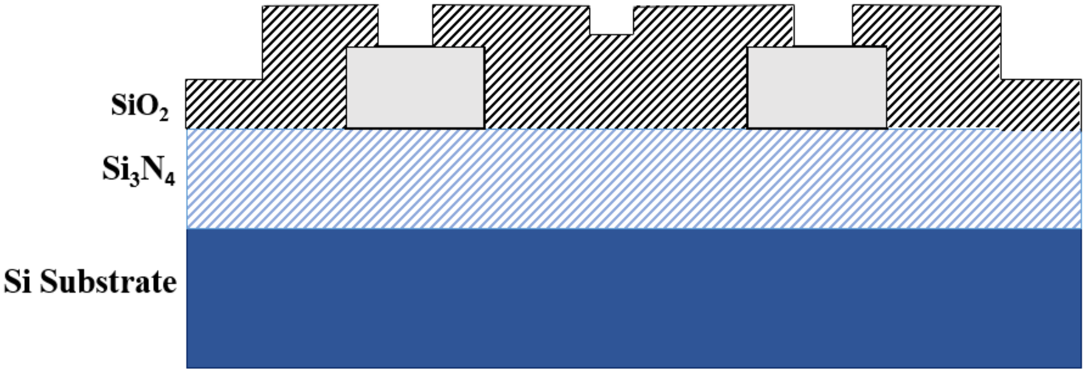


Figure 13. Dry etching used to create vias through SiO2 to Ni.

Double layer resist is also used as a lift -off layer. The chalcogenide lift-off mask i.e., Mask # 3 is used to pattern the resist with UV exposure to 45 mJ/cm^{-2} and it is developed for 1 minute using AZ 300 MIF. Next, the wafer is placed in a Cressington 308R thermal evaporator where 60nm of $Ge_{30}Se_{70}$ is thermally evaporated followed by 30nm Ag at a deposition rate of 0.75 Å/s. The chalcogenide layer ($Ge_{30}Se_{70}$) layer us photodoped with Ag by exposing the wafer to UV light for 1 hour to a dose of 5.3 J/cm^2. The photo-doping process is used to drive the Ag into the chalcogenide layer to attain a saturation concentration of 33 .at% concentration [55], [56]. After the doping process, the wafer is placed back into the Cressington evaporator where additional 35nm of Ag was deposited to form the top active metal electrode layer. Next, the resist is thermally cracked by placing the wafer on a thermal hot plate at 150 °C. The thermal shock breaks the resist film and allows a clean lift-off. The wafer is then placed in acetone to remove the metal as well as the resist layers.

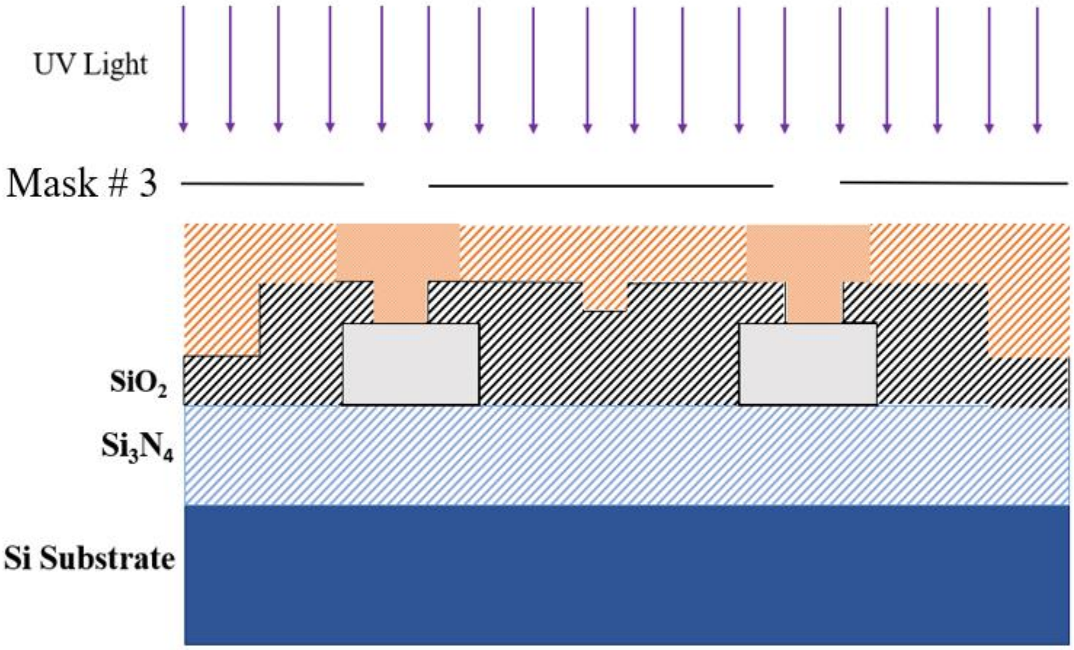


Figure 14. Mask # 3 used to for chalcogenide -Ag lift off layer.

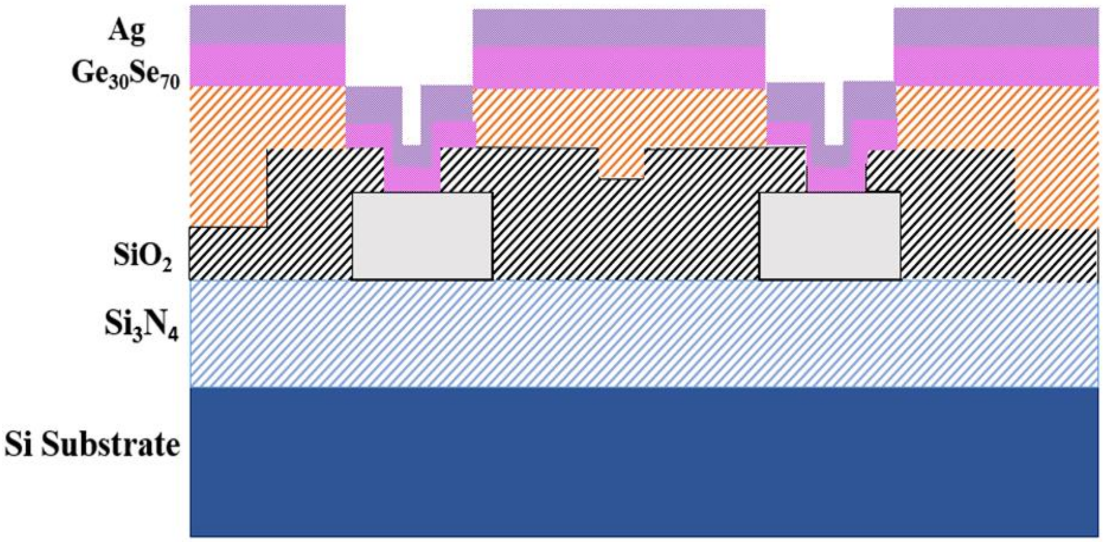


Figure 15. Active metal electrode deposition.

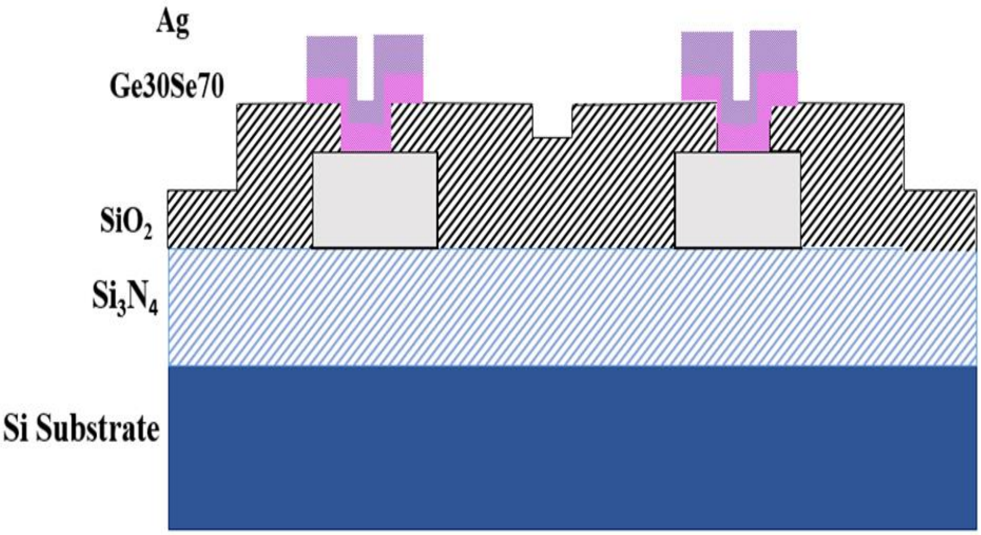


Figure 16. Lift-off patterning of top active metal electrode layer.

The final step is to create a top Al crossbar electrode. To create the Al metal contact, double layer lift-off resist recipe is used. Mask # 4 is used to pattern the resist with UV exposure to 45 mJ/cm^{-2} and it is developed for 1 minute using AZ 300 MIF. The wafer is then placed in the Lesker PVD sputtering machine and 250 nm of Al is deposited at a deposition rate of 1.5 Ǻ/sec. After deposition the resist is thermally cracked to allow clean lift off and placed in acetone. Figure 19 depicts the final device structure.

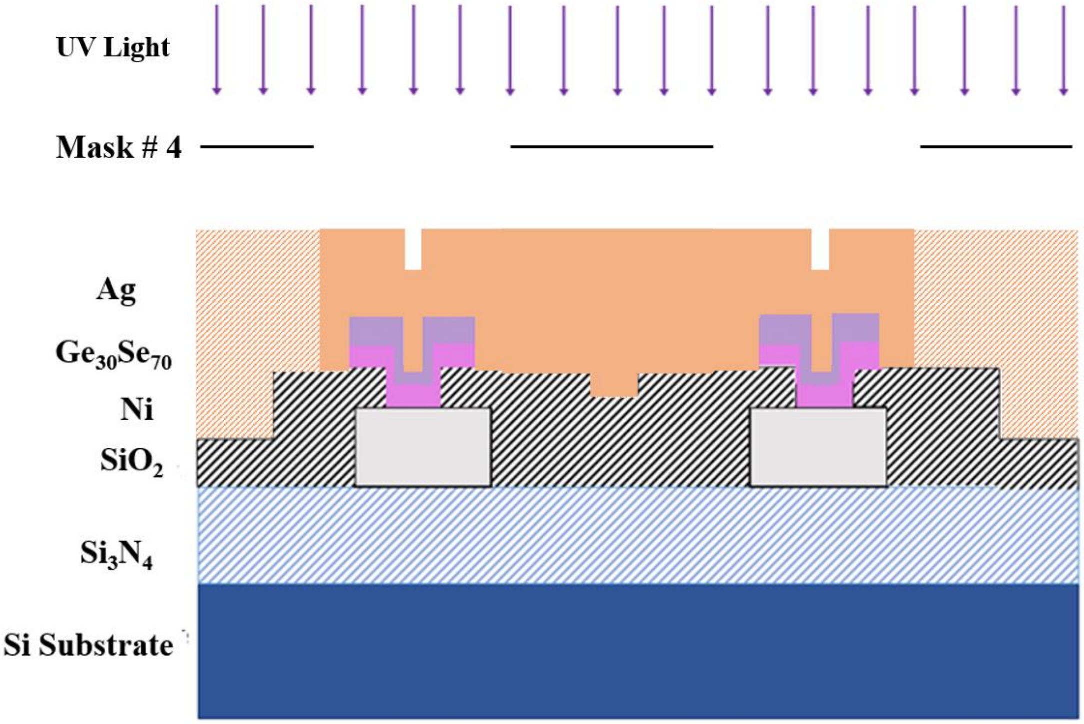


Figure 17. Mask #4 used for patterning the top contact metal electrode.

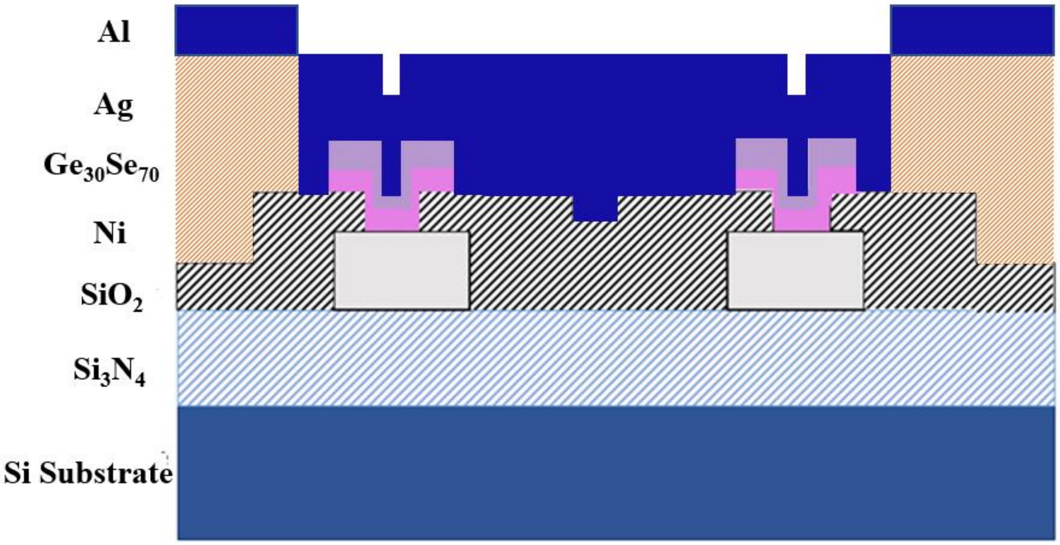


Figure 18. Contact metal (Al) deposition.

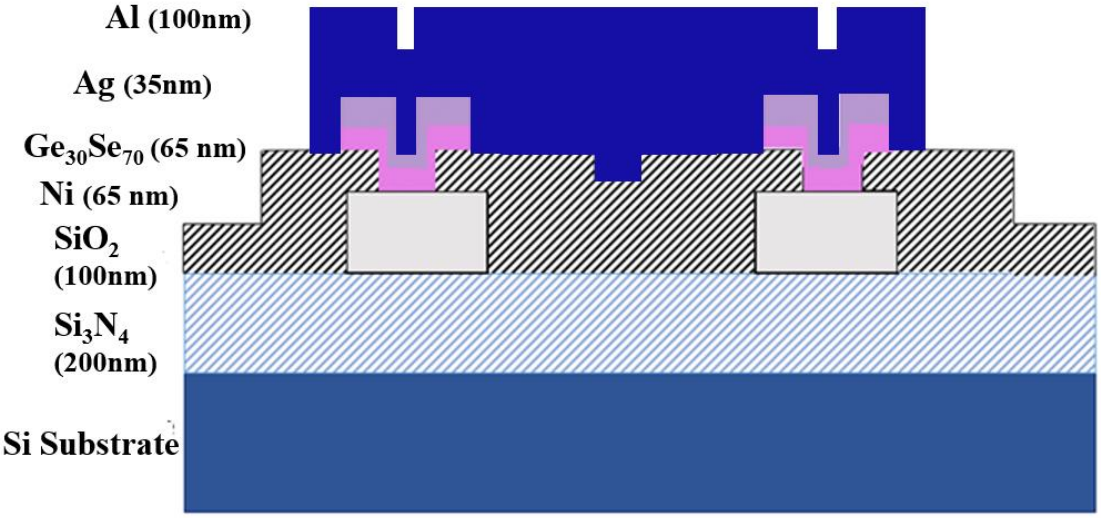


Figure 19. Final device structure.

CHAPTER 4

STUDY OF CBRAM AS AN ANALOG SYNAPSE

4.1. Introduction

The main advantage of the analog crossbar array architecture is that it can perform the inference and training efficiently. The most popular choice for an analog crossbar accelerator has been the Resistive RAM (RRAM) devices. However, for neuromorphic computing applications, these devices must meet several stringent requirements. For inference applications it is necessary that the devices have at least two conductance states, low cycle-cycle variability, low read noise as well as good retention. To accelerate training, the RRAM elements should be able to switch between a wide range of analog conductance states [13]. One desirable device characteristic for training is a gradual and symmetric response to different programming pulses which are used for weight update [57], [58],[59]. In this chapter, we investigate the response of Ag-$Ge_{30}Se_{70}$ CBRAM crossbar devices to different programming pulse schemes. We analyse the important synaptic properties like conductance linearity response, device-to-device variability, and conductance on/off ratio of Ag-$Ge_{30}Se_{70}$ CBRAM. In addition to understanding the device synaptic behaviour of these devices, it is necessary to understand the circuit level performance of these devices in an analog neural accelerator. We demonstrate the circuit level performance of Ag-$Ge_{30}Se_{70}$ CBRAM analog accelerator using NeuroSim [54],[60]. NeuroSim is an open-source software to evaluate   Cross-Sim, an open-source software, developed by Sandia National Laboratory, is used to analyse the classification accuracy of the fabricated devices for two different datasets [53].

4.2. DC I-V Characterization

The current-voltage (I-V) characteristic of one fabricated CBRAM synapse is shown Figure 20. The figure shows the characteristic hysteresis DC switching behaviour over three voltage sweep cycles. It is important to note that Ag-$Ge_{30}Se_{70}$ CBRAM devices are forming-free devices since Ag is introduced in the switching layer by the process of photo dissolution [61]. The DC sweep was performed using Agilent 4156 C analyser. Bipolar switching behaviour was observed reproducibly over 50 cycles. It should be noted that these large-scale DC characterizations which toggle the CBRAM cell between HRS and LRS are performed to assess general operational integrity not the incremental LRS switching required for use as a synaptic element. Cumulative distributions for all HRS and LRS measurements are shown in Figure 4.2. The distributions show a minimum HRS to LRS ratio above 20, well above the target minimum of 10 [38].

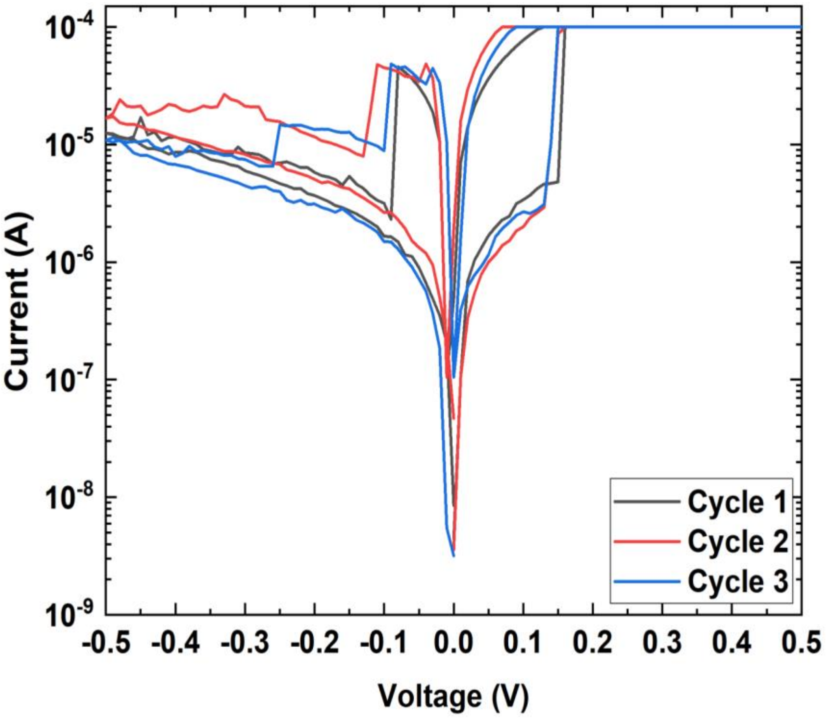


Figure 20.I-V characteristics of Ag-$Ge_{30}Se_{70}$ CBRAM device cycled at 100µA compliance.

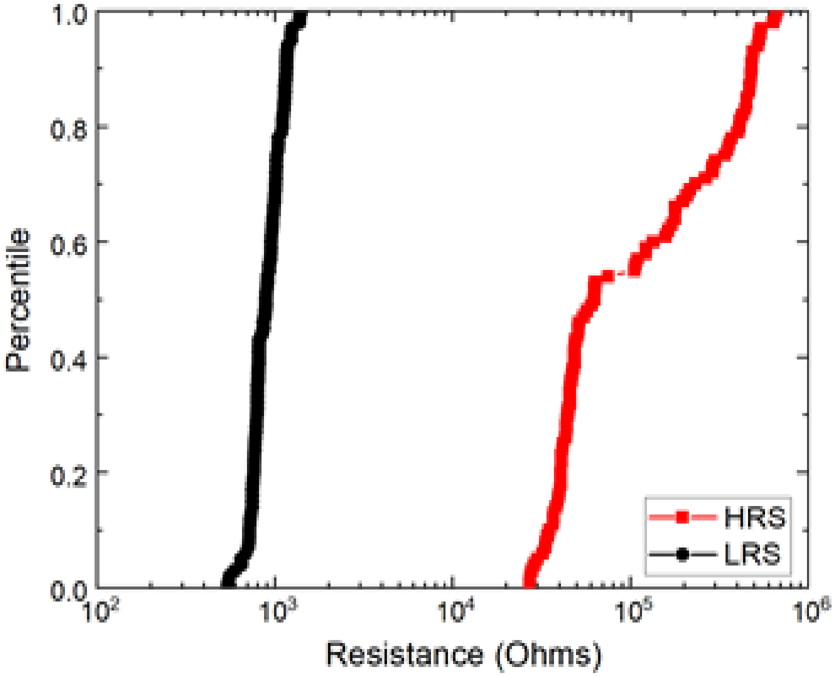


Figure 21. High Resistance State (HRS) and Low Resistance State (LRS) distribution of the device extracted @0.03V for 50 cycles.

Another important requirement of the synaptic device is to have multilevel conductance characteristics. The multilevel conductance characteristics of crossbar Ag-$Ge_{30}Se_{70}$ CBRAM is demonstrated in Figure 22 by increasing the current compliance limit for each consecutive DC cycle. The compliance current is the maximum allowable current through the device. A higher the compliance current causes more ion flow through the device leading to a thicker conductive filament and lower resistance states. This characteristic of the CBRAM is well established in earlier works [62], [63].

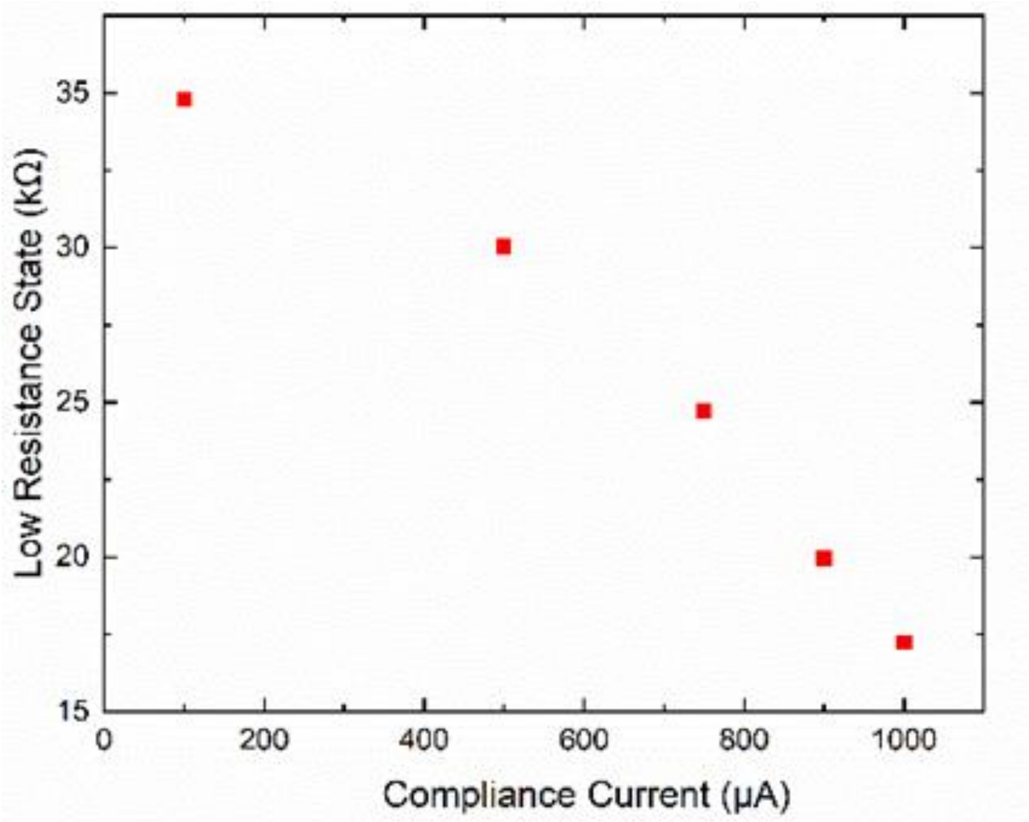


Figure 22. Multi-level conductance demonstration of Ag-Ge30Se70 CBRAM with increasing compliance current.

4.3. Constant Pulse Programming

Multilevel conductance switching is demonstrated by applying voltage pulses to top electrode (Ag anode) while the bottom electrode (Ni cathode) is grounded. The measurements are performed using the Keithley 4200 SCS parameter analyser with a built-in 4225 PMU module. To achieve gradual switching behaviour, we applied 100 consecutive SET pulses (0.35V for 100ns) and 50 consecutive RESET pulses (-0.25V for 100ns). Each write and erase pulse was followed by read voltage of 30 mV to extract conductivity. The change in conductance of the device is plotted as a function of pulse number in Figure 23 for 20 cycles of SET/RESET operations. For these input pulse parameters, the results indicate a highly non-linear response in device conductance, where the incremental change in conductance is greatest for the early SET/RESET signals in each sequence, saturating quickly to a to a maximum/minimum level. We observe a 10x change in conductivity for the 100ns pulse width.

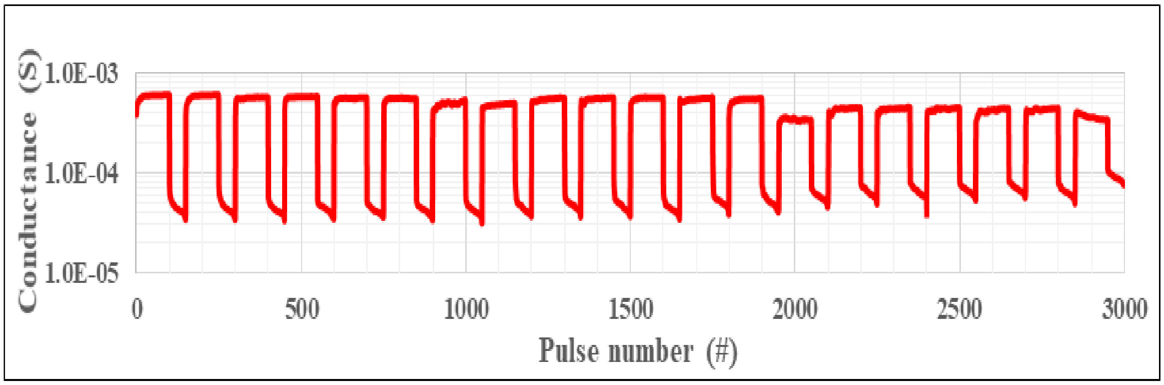


Figure 23. The potentiation and depression cycling of Ag – $Ge_{30}Se_{70}$ CBRAM with P.W. = 100 ns.

The effect of pulse width on the switching response is observed in Figure 24. For these data, the SET/RESET pulses have a longer pulse width of 100µs. An increase in maximum and minimum conductance range is observed for higher pulse width, which can be attributed to the greater widening of the conductive filament during the longer voltage stress time.

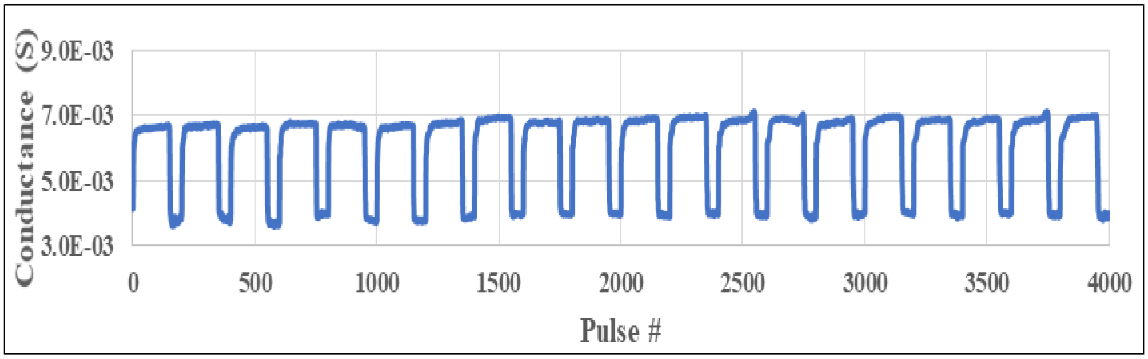


Figure 24. Repeated pulse cycling Ag-Ge30Se70 CBRAM with P.W =100 µs.

Figure 25 plots the response for both pulse widths over one cycle and shows that for shorter pulses, i.e., 100ns, we observe a larger G_{MAX} /G_{MIN} ratio and lower conductance levels. When considering the use of these devices for parallel VMM programming lower conductance levels are preferred since this will reduce the energy required for both VMM programming and inference [64].

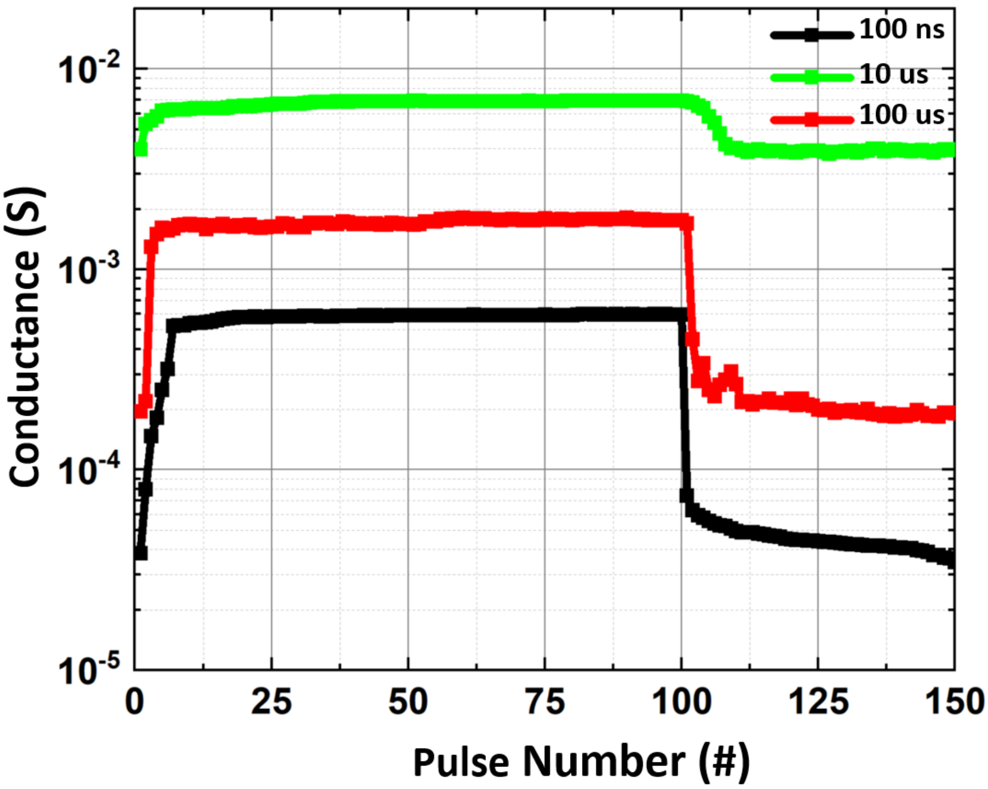


Figure 25. The analog response of the Ag-Ge30Se70 CBRAM device to different pulse-widths.

The conductance update response for both pulse widths was compared by normalizing the conductance range as depicted in Figure 26. The conductance update for the potentiation is identical for both the pulse-widths. The depression behaviour for 100ns pulse width seems to be better. A detailed study of the linearity response is provided in the following subsections.

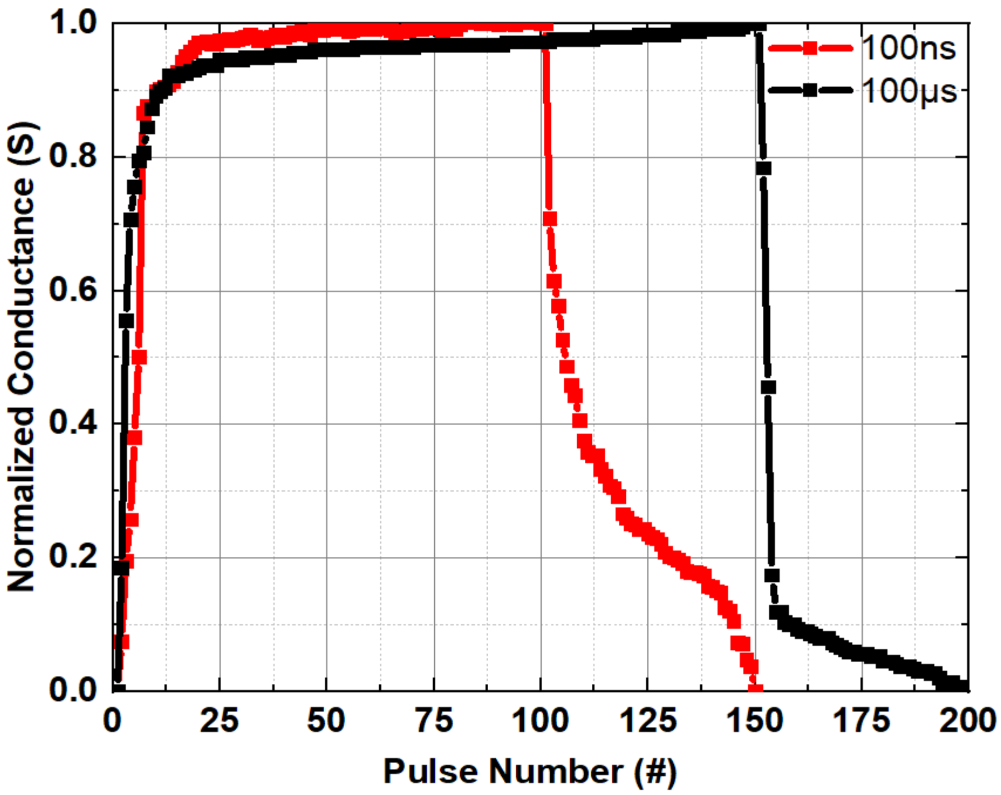


Figure 26. Normalized conductance of the device to different pulse widths.

4.4. Variable Pulse Programming

An increasing amplitude pulses were applied to the anode of the device. Increasing amplitude voltage pulses from 0.2 V to 1.5 V (pulse width = 100 µs and 50 mV step) for potentiation and -0.2 V to -1.2 V (pulse width = 100 µs and 50 mV step) for depression were applied. Each pulse was accompanied by a read pulse of 50 mV with a pulse width of 100 µs. Figure 27 shows the potentiation and depression characteristics of the device. Figure 28 displays the response of the devices to multiple pulse cycles. Unlike the characteristics above, these devices are shown to exhibit good linearity over a suitable conductance range. This linear response is realized by careful control of the thickness of the metallic filament. One should however note that using varying amplitude pulses eliminates the advantage of parallelism since it necessitates addressing each analog resistive element in a crossbar, individually.

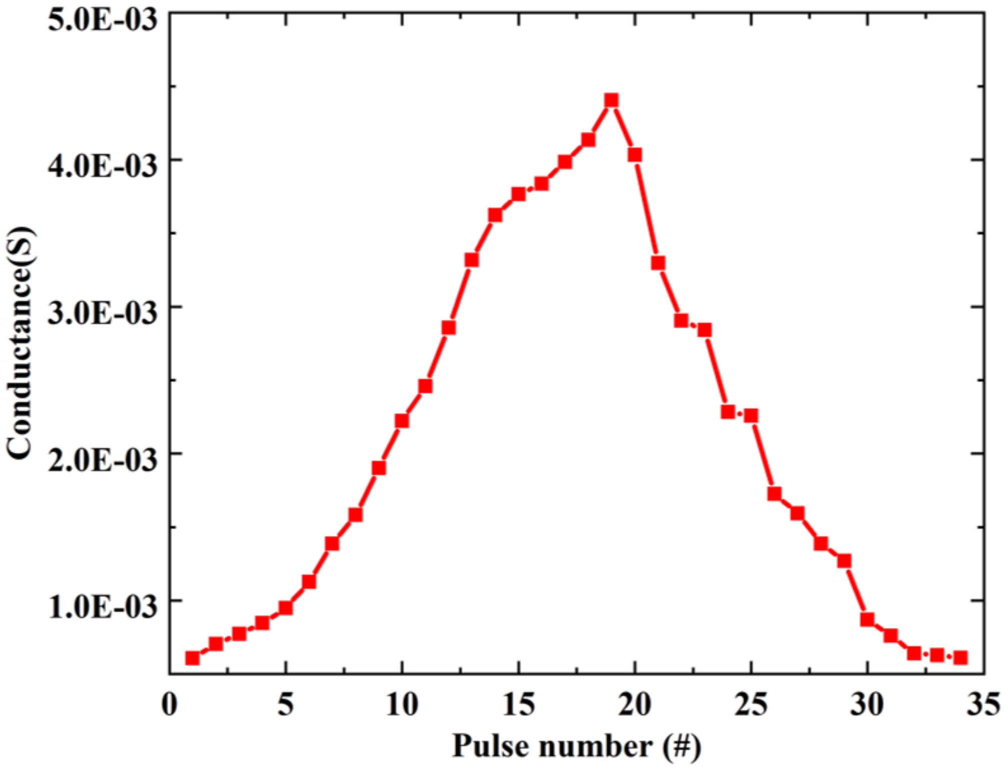


Figure 27. Conductance response of Ag-$Ge_{30}Se_{70}$ CBRAM to incremental amplitude pulses.

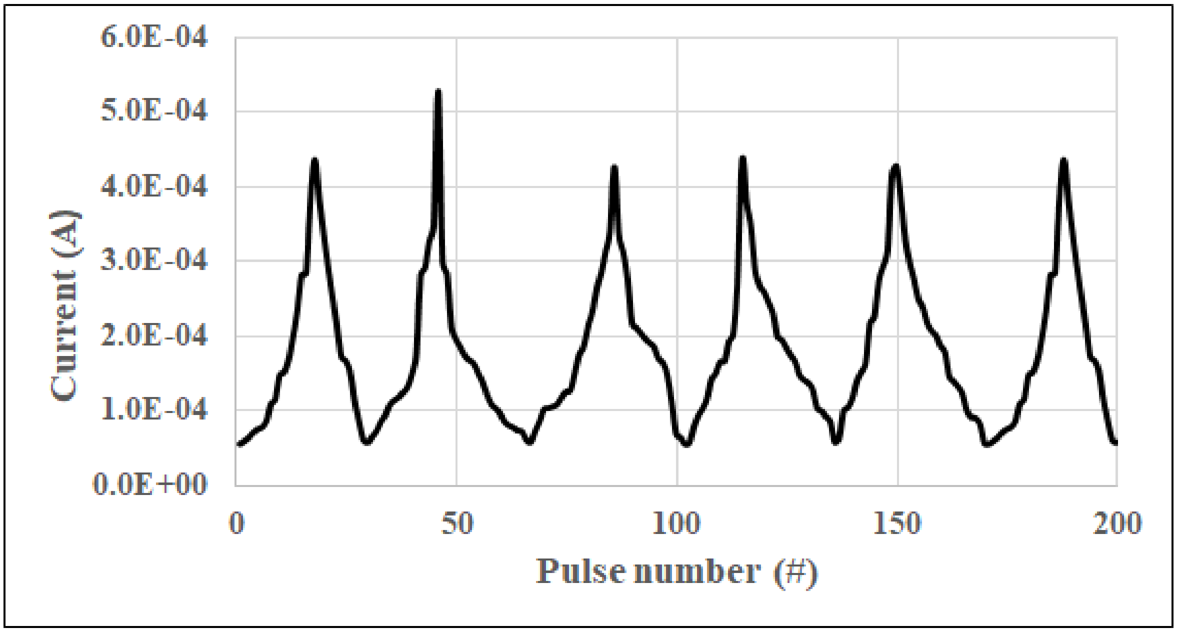


Figure 28. Potentiation and depression cycling to increasing amplitude pulses.

4.5. Study of important synaptic properties for Ag-$Ge_{30}Se_{70}$ CBRAM device

a. Non-Linearity Response Study

The linearity metric refers to the linear behavior of the conductance vs. pulse number curve. The linearity metric is important in the weight mapping algorithms. Linear and symmetric behavior is an ideal characteristic. Actual devices typically show some degree of asymmetry and nonlinearity in the weight update characteristics. Asymmetry between potentiation and depression characteristics can degrade pattern recognition and online training by the neural network. The Ag-$Ge_{30}Se_{70}$ CBRAM devices studied in this show an increase in conductance for initial pulse train but tend to saturate at the end. Recent results have shown that this nonlinearity/asymmetry has caused the learning accuracy loss in the neural networks [65].

To mathematically extract the non-linearity factor, the potentiation and depression response of the device to the 2-pulse scheme response are fit to the following equations [54]:

$$G_{LTP} = B \left(1 - e^{\left(-\frac{P}{A}\right)}\right) + G_{min} \tag{4.1}$$

$$G_{LTD} = B \left(1 - e^{\left(P - \frac{Pmax}{A}\right)}\right) + G_{max} \tag{4.2}$$

$$B = \frac{G_{max} - G_{min}}{1 - e^{-\frac{Pmax}{A}}} \quad , \tag{4.3}$$

where, *G*, *P*, and *A* are the conductance value, pulse number, and nonlinear behaviour of weight update, respectively. G_{max} and P_{max} are the maximum conductance and pulse widths, respectively, obtained from the experimental data. The different nonlinearity factors for potentiation and depression are obtained by simulating the curves using MATLAB with equations (1) and (2) respectively [54]. The non-linearity factor *A* for the potentiation and depression are found to be 6.65 and -2.5, respectively, for the constant amplitude (P.W.= 100 ns) as shown in Figure 29.

Similarly, for the increasing amplitude pulses, the non-linearity factors the potentiation and depression was found to be 1 and -0.96 respectively as depicted in Figure 30.

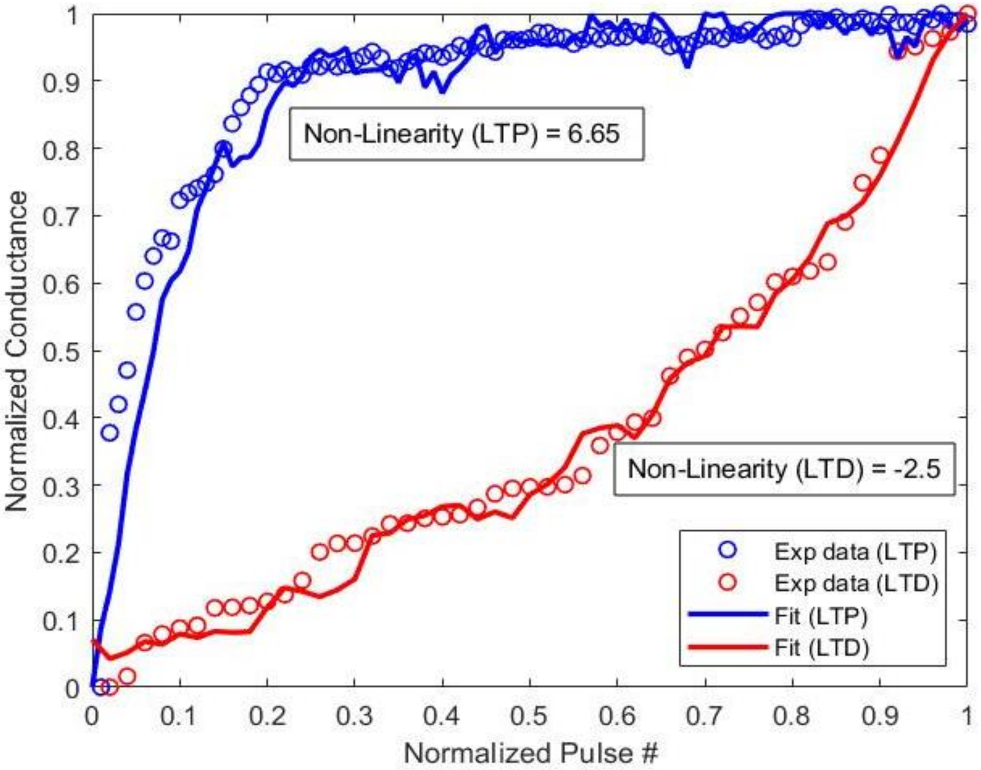


Figure 29. Non-linearity fitting factors -Constant Amplitude pulses (100 ns).

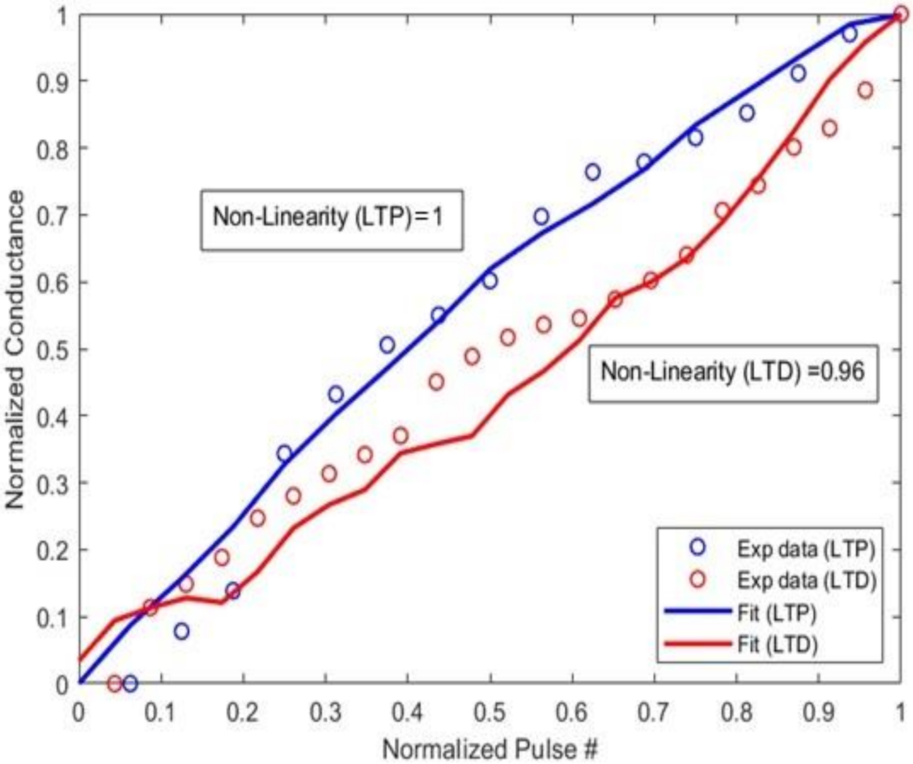


Figure 30. Non-linearity fitting factors - Increasing Amplitude pulses (100 μs).

b. Device - Device Variation

The operation of $Ag\text{-}Ge_{30}Se_{70}$ CBRAM devices involves the drift and diffusion of metal ions to form a filament and hence these devices are prone to variation in conductance update from device-to-device across the crossbar array. Different devices will tend to have different non-linearity baselines [66]. The device-device variability is studied by measuring the weight update behaviour of 10 devices in the crossbar array. Here the variability is calculated as the deviation in the non-linearity factor for different devices. The device-device variability is seen to be ±1.05 as shown in Figure 31.

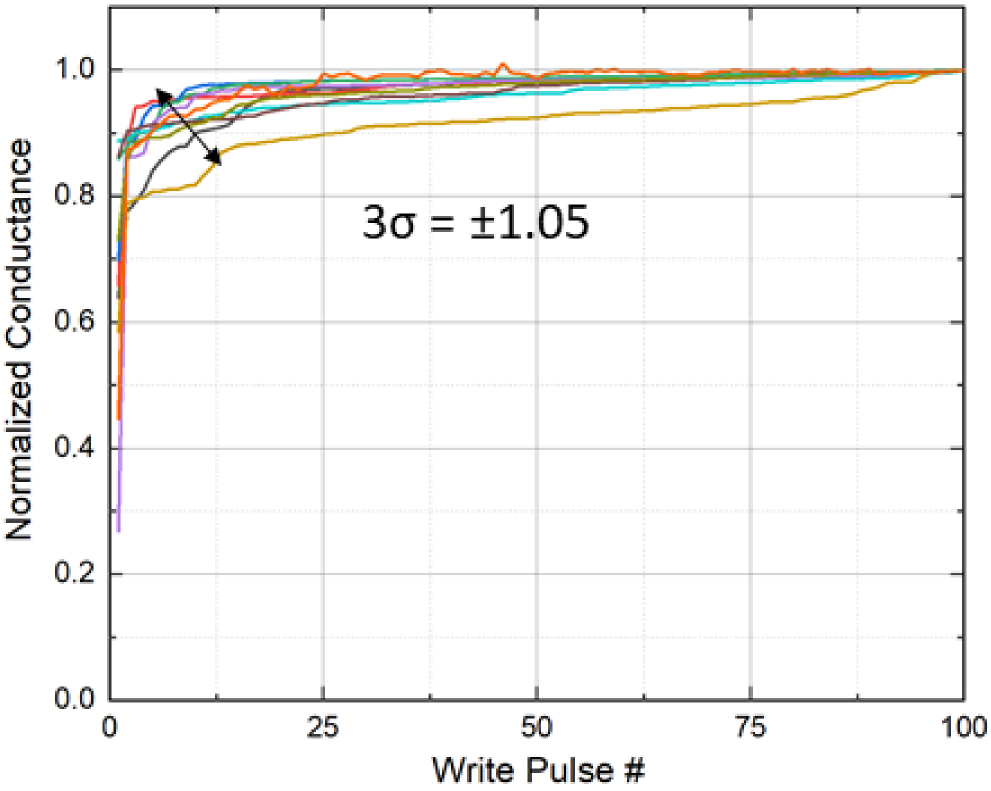


Figure 31. Device to device weight update variation for $Ag\text{-}Ge_{30}Se_{70}$ CBRAM devices.

c. On/Off Conduction Ratio

Ideally the weight values in the training algorithms are represented as normalized conductance, ranging from 0 to 1. However, the minimum conductance can be zero only if the ratio of maximum and minimum conductance i.e. On/Off ratio approached infinity which is practically not feasible. The learning accuracy has been observed to degrade with a small on/off conductance ratio [66] because for calculations on a small weight update value will lead to significant distortions. The on/off conduction ratio for Ag-$Ge_{30}Se_{70}$ CBRAM devices for identical pulse application i.e., Figure 32, is 16.

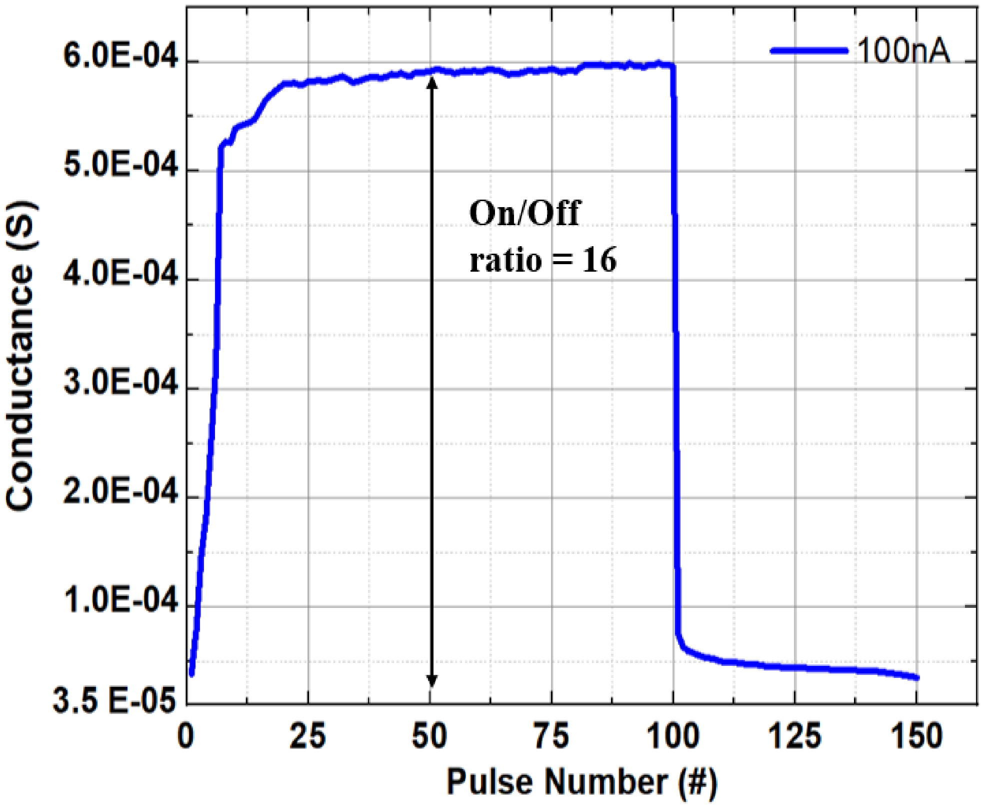


Figure 32. Conductance on/off ratio for Ag-$Ge_{30}Se_{70}$ CBRAM device for constant amplitude pulses (100ns).

4.7. Modelling Effect of Linearity on Pattern Recognition Accuracy

The impact of linearity and weight update behavior on system response to different pulse schemes is studied by modelling these devices in an analog neural training accelerator. An artificial neural network was simulated using the device properties of the Ag-$Ge_{30}Se_{70}$ CBRAM. The CrossSim simulator was used to perform the supervised learning [31], [53],[67]. A three-layer neural networks depicted in Figure 33 were trained using a backpropagation algorithm. This algorithm is a computationally intensive algorithm that uses the two important kernels: vector matrix multiply (VMM) and outer product update. For this work, we have used two datasets: a small image version (8 X 8 pixels) of handwritten digits from the 'Optical Recognition Handwritten Digits' dataset [68] and a large image version (28 X 28 pixels) of handwritten digits from the MNIST dataset [69]. A two-layer neural network consisting of 64 input, 36 hidden and 10 output nodes are used for small image dataset. The larger MNIST dataset was trained on a network of dimensions 784x300x10. A crossbar array is simulated using the analog synaptic properties. The crossbar a part of the neural core performs the required kernel operations. To perform the vector matrix multiplications, the conductance states of the synaptic devices are programmed by applying input voltages or input pulse lengths to each row and the corresponding output vector is read in the form of current. Parallel read, multiplication and summation operations are performed in this single step. This helps in reducing the total energy of vector-matrix multiply (VMM) and highlights the main advantage of using analog resistive memories for these operations.

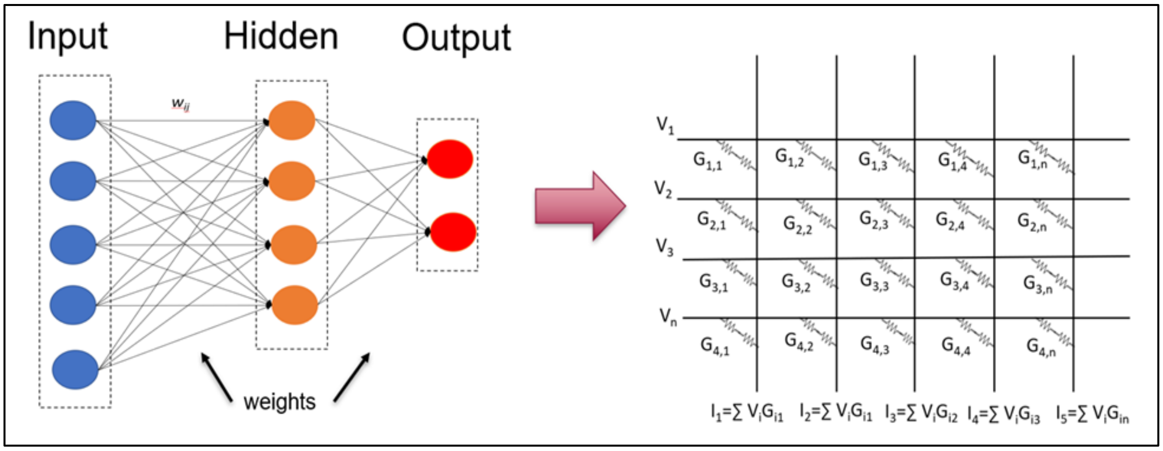


Figure 33. Mapping the neural network layers to crossbar array.

Data set	**#Training Examples**	**#Test Examples**	**Network size**
UCI Small images	3823	1797	64X36X10
MNIST large images	60000	10000	784X300X10

To simulate the response of the Ag-$Ge_{30}Se_{70}$ CBRAM devices in the neural network, a lookup table is created containing $\Delta G/G$ response of the devices to multiple potentiation and depression cycles at a given pulse widths using data from Figures 23 and 28 [70]. For small image dataset, we observe the training accuracy of 85% for constant amplitude pulsing as depicted in Figure 34(a) while the training accuracy for the same dataset with increasing pulse amplitude was found to be 92% as seen in Figure 34(b). For large digits the accuracy was seen to be improved from 79% to 87% for increasing pulse amplitude as seen in Figure 35a & 35b. The ideal training accuracy using a double-precision CPU or GPU is 98%.

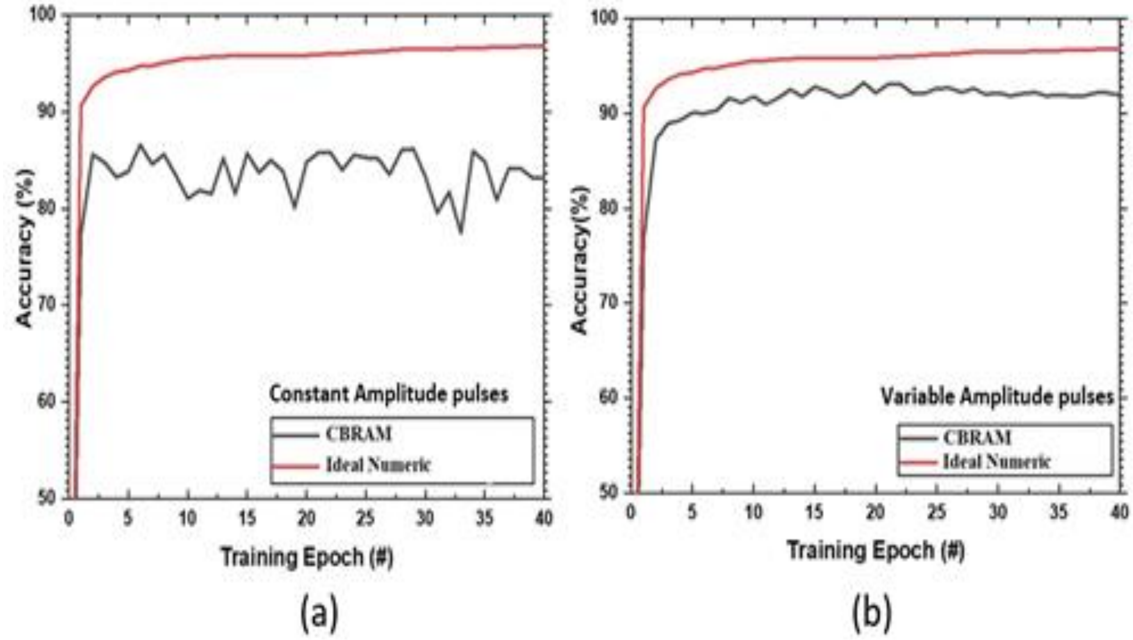


Figure 34. Classification accuracy for small image dataset.

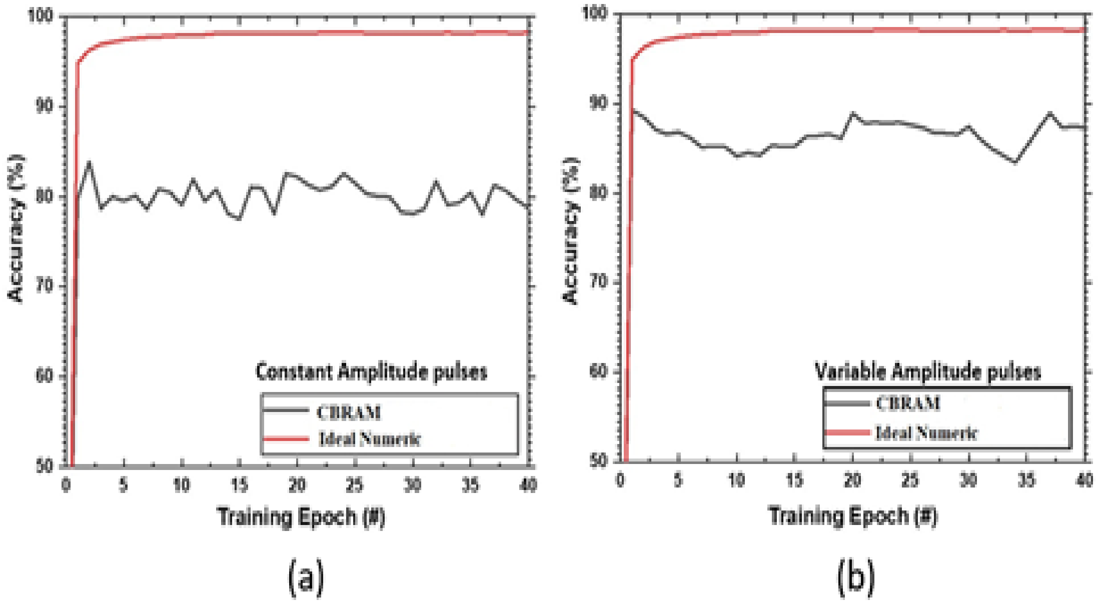


Figure 35. Classification accuracy for large image MNIST dataset.

4.6. Circuit level performance of Ag-$Ge_{30}Se_{70}$ CBRAM in a Crossbar Array

To benchmark the circuit level performance of Ag-$Ge_{30}Se_{70}$ CBRAM device we use NeuroSim, a neuro-inspired architecture at circuit level, to estimate the area, latency, dynamic energy of a neuromorphic hardware accelerator. NeuroSim utilizes the circuit/device model parameters on a small but realistic neuro-inspired architecture, enabling fast early-stage design exploration [54], [60]. The overview of neuromorphic hardware design utilized by NeuroSim module is depicted in Figure 36.

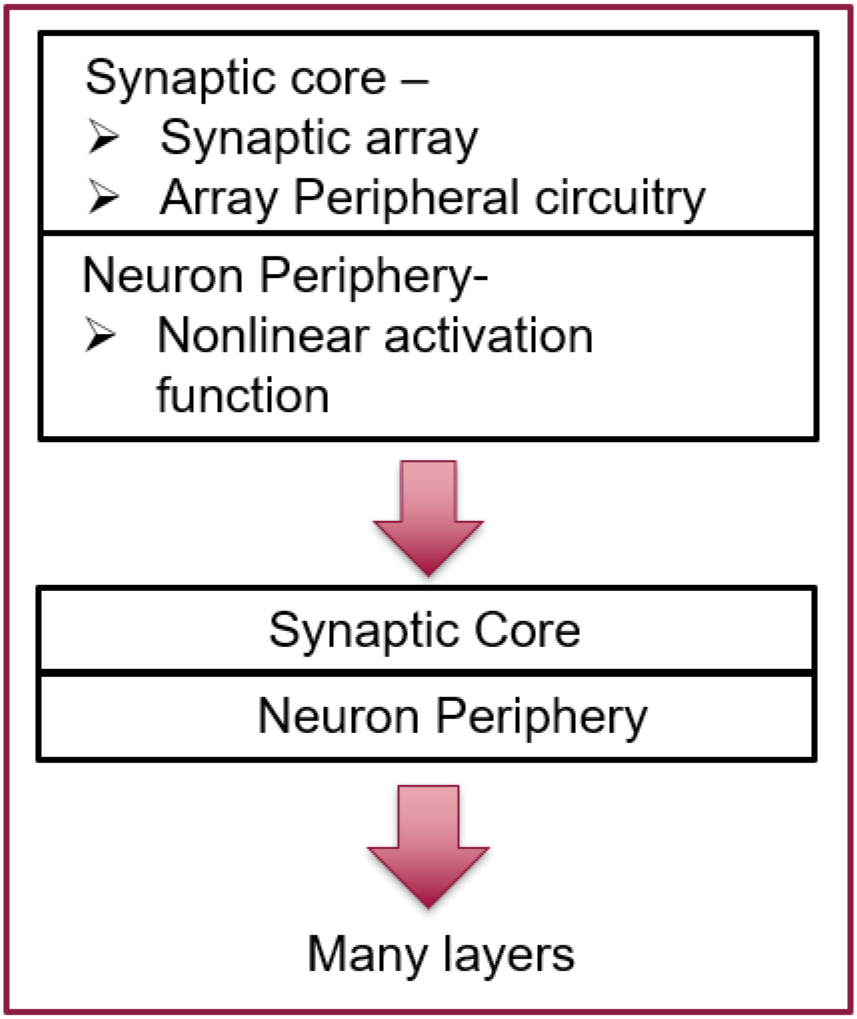


Figure 36. An overview high level neuromorphic accelerator design used in NeuroSim to implement neural network. [Adapted from [54]].

To estimate the circuit level performance of Ag-$Ge_{30}Se_{70}$ CBRAM, a 2-layer MLP network with 400 input neurons, 100 hidden neurons and 10 output neurons is mapped to the analog eNVM crossbar array architecture. Each synaptic layer consists of synaptic core and neuron periphery. The synaptic core consists of a crossbar array architecture as depicted in Figure 37.

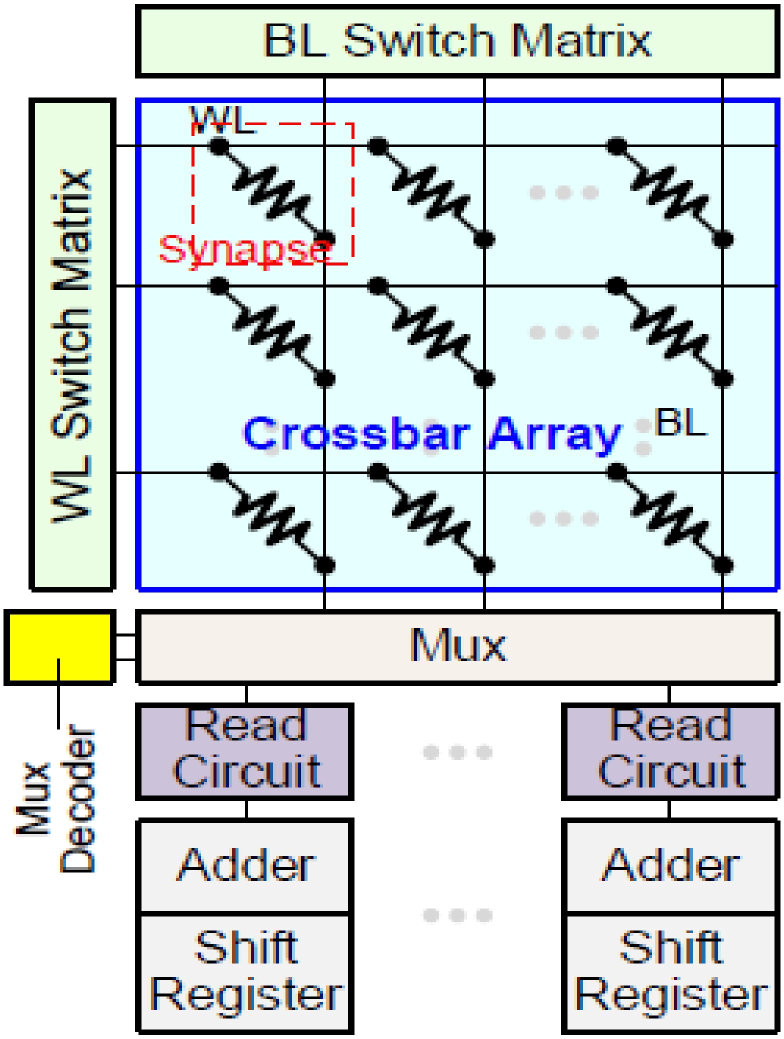


Figure 37. Circuit block diagram of analog eNVM synaptic core. © 2017 IEEE

In this type of array structure, each CBRAM device is located at the cross point of a word line (WL) and a bit line (BL). This is the most compact and simplest form of array structure. This layout can achieve the highest integration density of $4F^2$/cell. The input vector can be coded in read voltage signals and the weighted sum operation (i.e., MVM) can be performed in a parallel fashion. A word line (WL) switch matrix can be used to switch multiple rows or columns at a time thus enabling parallel read voltage input for the weighted sum operation or selecting a rows or columns to perform weight update scheme that require different voltage biases. Adder and shift registers are used to perform the shift and the addition on all the bits cycles to obtain the final weighted results. To convert the analog weighted sum currents to digital outputs, read circuitry is used [71]. The read circuit uses the principle of integrate-and-fire neuron model. The neuron peripheral circuitry performs the non-linear activation function on the digital outputs and is also responsible for the communication between different synaptic cores [54].

A list of the simulation parameters is provided in Table 2. The parameters are extracted from the synaptic response of the Ag-$Ge_{30}Se_{70}$ CBRAM to constant amplitude pulses.

Table 2.Simulation circuit and device parameters -Constant Amplitude Pulses

Parameters	**Values**
Technology node (F)	32 nm
Clock frequency	2 GHz
Assigned height of cell	2 F
Assigned width of cell	2 F
Crossbar eNVM cell area	4 F^2 (F = tech node)
Max. Conductance	6 E-4
Min. Conductance	3.8 E-5
Write Voltage	0.35V / -0.25V
Read Voltage	30 mV
Write pulse width (P.W.)	100 ns
Read pulse width (P.W.)	100 µs
Non-linearity	6.65 / -2.5

4.6.1. Simulation Results

The simulated results for NeuroSim for constant amplitude pulses are depicted in Table 3. The total area is calculated as the summation of the synaptic core area and the total neuron peripheral circuit area. The latency in a crossbar array is caused by wire parasitics (*R* & *C*) and the peripheral circuitry. The power consumption is mainly due to the static energy consumption in an eNVM synaptic array. The power consumption is directly proportional to the conductance of the cell, write voltage used for conductance increase and decrease, number of applied write pulses and the pulse widths used [54].

Table 3.Represents the circuit level performance of Ag-Ge30Se70 CBRAM device at 32nm.

Area (μm^2)	Read Energy (J)	Write Energy (J)	Read Latency (s)	Write Latency (s)
1.4028E5	1.0176E-01	5.0520E-02	1.5796E-02	4.5121

One should note that when using variable amplitude scheme that yields highly linear weight update behavior, a read-before-write verify step is necessary to determine the current conductance state. After the conductance state is read, a correct pulse amplitude is applied. This increases the peripheral circuit complexity along with increasing the latency and energy consumption. Figure 38 depicts the crossbar array circuitry used for non-identical pulse application.

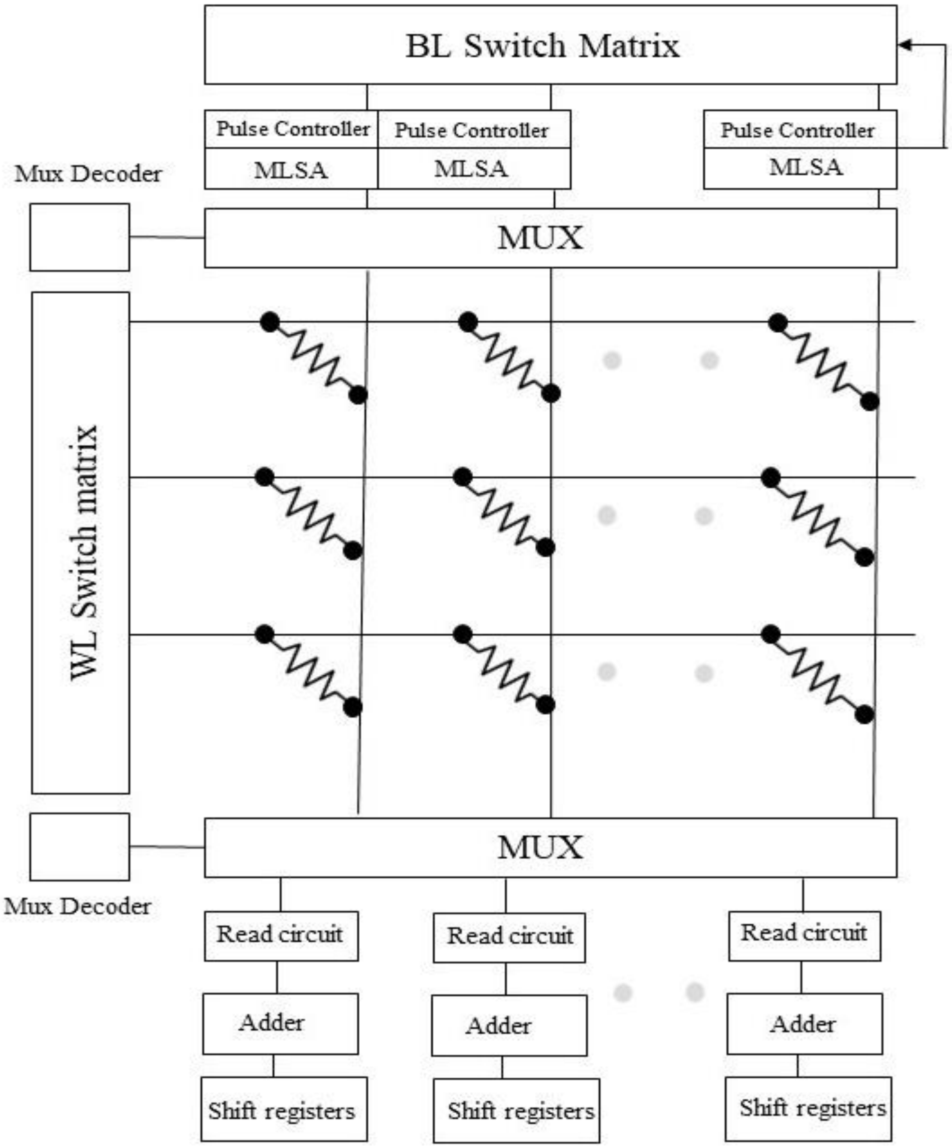
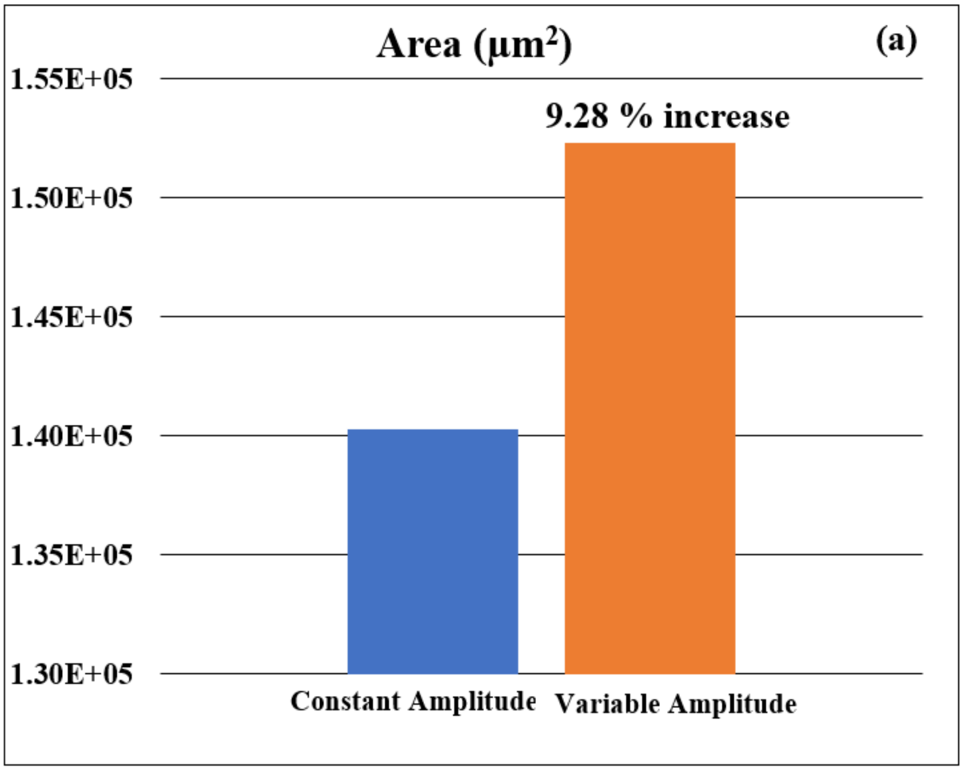


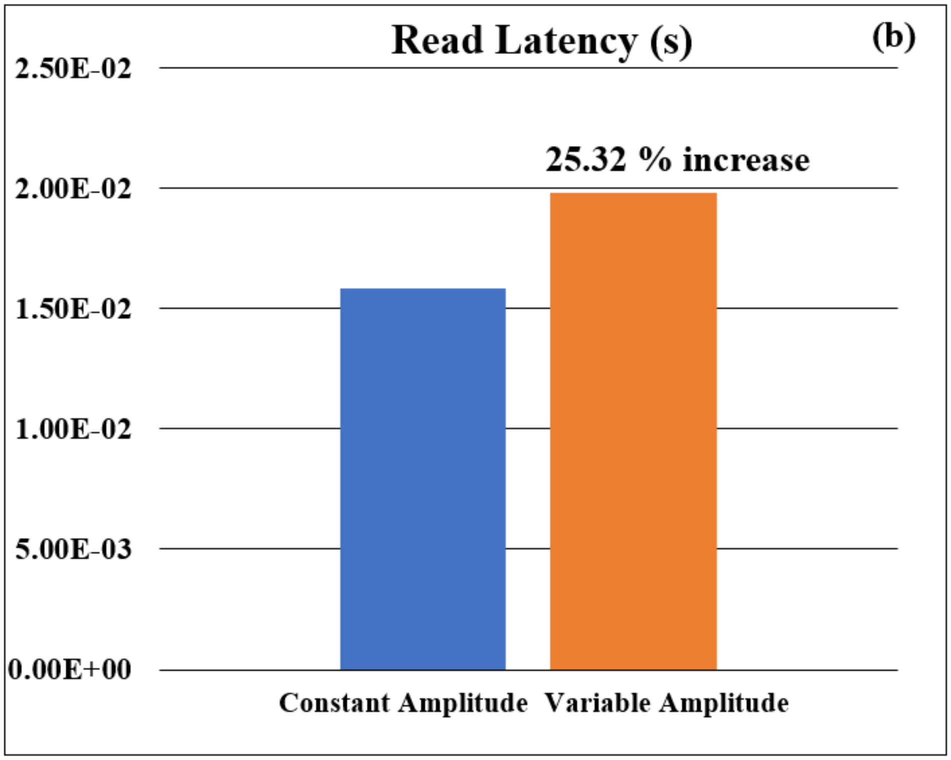
Figure 38. Block diagram of Ag-Ge30Se70 CBRAM crossbar array with peripheral circuitry used for variable amplitude pulse application [72].

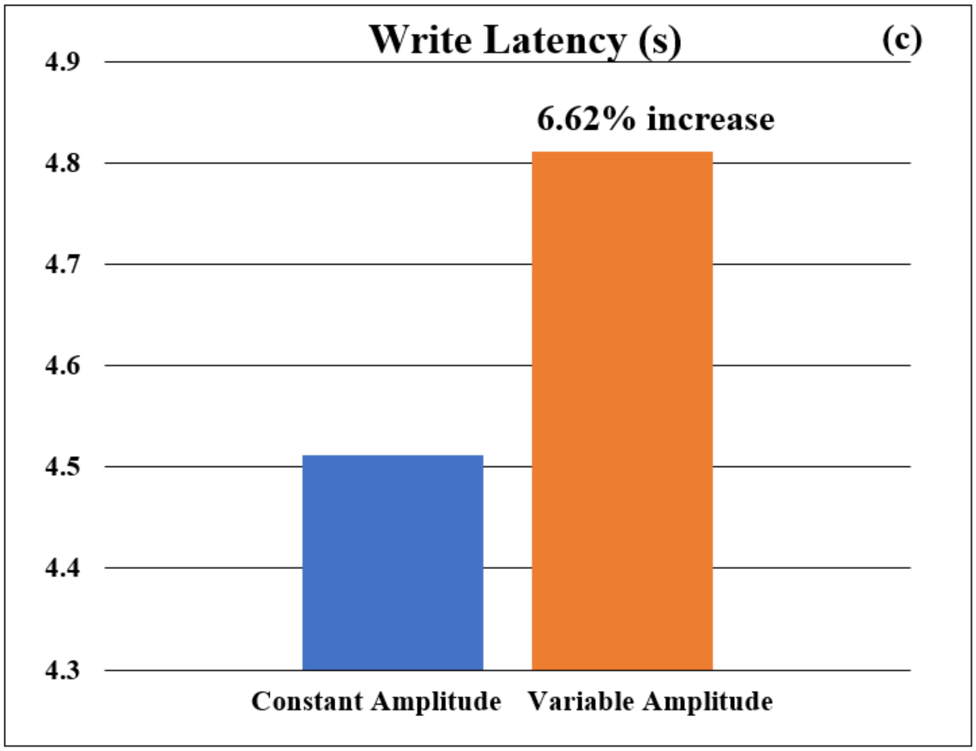
Additional circuitry like pulse controller and multilevel sense amplifier are added to the crossbar array to feed back the current state to choose appropriate pulse amplitude. Circuit simulation was performed using the variable pulse amplitude response of the Ag-$Ge_{30}Se_{70}$ CBRAM devices as depicted in Figure 4.9. The simulated results are shown in Table 4.

Table 4.Represents the circuit level performance of Ag-$Ge_{30}Se_{70}$ CBRAM device at 32nm with Variable pulse amplitude.

Area (μm^2)	Read Energy (J)	Write Energy (J)	Read Latency (s)	Write Latency (s)
1.53297E5	5.2522E-01	4.3830E-01	1.9796E-02	4.8107




Read Latency (s)
(b)
2.50E-02
2.00E-02
1.50E-02
1.00E-02
5.00E-03
0.00E+00
25.32 % increase
Constant Amplitude
Variable Amplitude


Write Latency (s)
(c)
4.9
4.8
4.7
4.6
4.5
4.4
4.3
6.62% increase
Constant Amplitude
Variable Amplitude

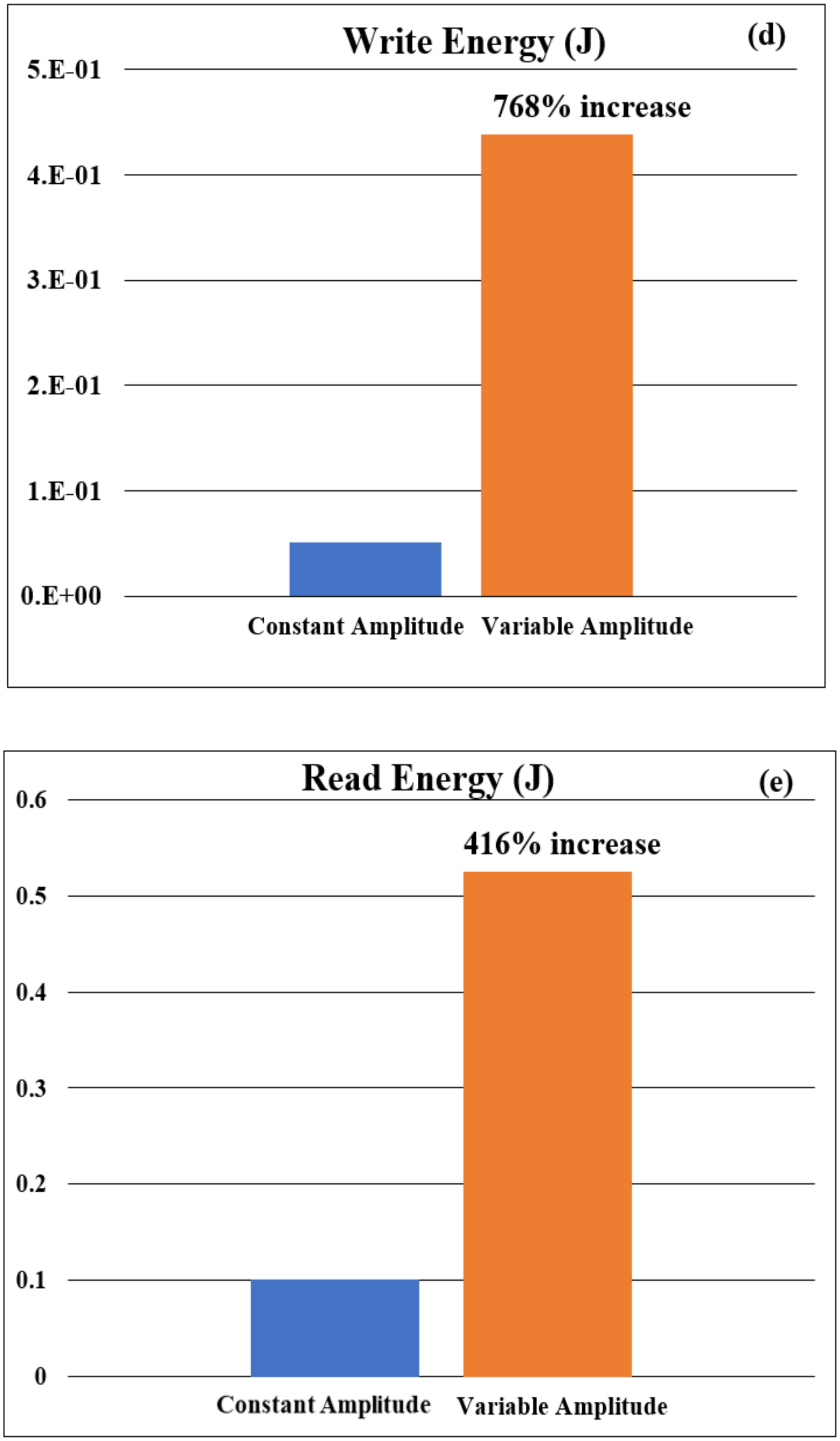


Figure 39. Comparing the circuit level performance of Constant amplitude and Variable amplitude pulses.

Comparing the performances, we observe that there is 9.28% area overhead, 6.62 % write latency, and 768% write energy, 426 % read energy penalty is introduced when using variable amplitude pulses as compared to constant amplitude pulses.

The overhead energy and latency problem can be resolved by using combination of resistors in series with RRAM device as suggested in some previous works [73], [74]. K. Moon *et.al* suggest using a voltage divider arrangement in which a fixed resistor is connected in series with the CBRAM device as shown in Figure 40 [73]. However, for potentiation, i.e., increasing the CBRAM conductance (i.e., reducing the resistance), it will not be possible to use a fixed voltage divider to realize an increase in the pulse voltage across R_{CBRAM}. To increase the V_{CBRAM} pulse while R_{CBRAM} switches to a lower resistor a more sophisticated design is required.

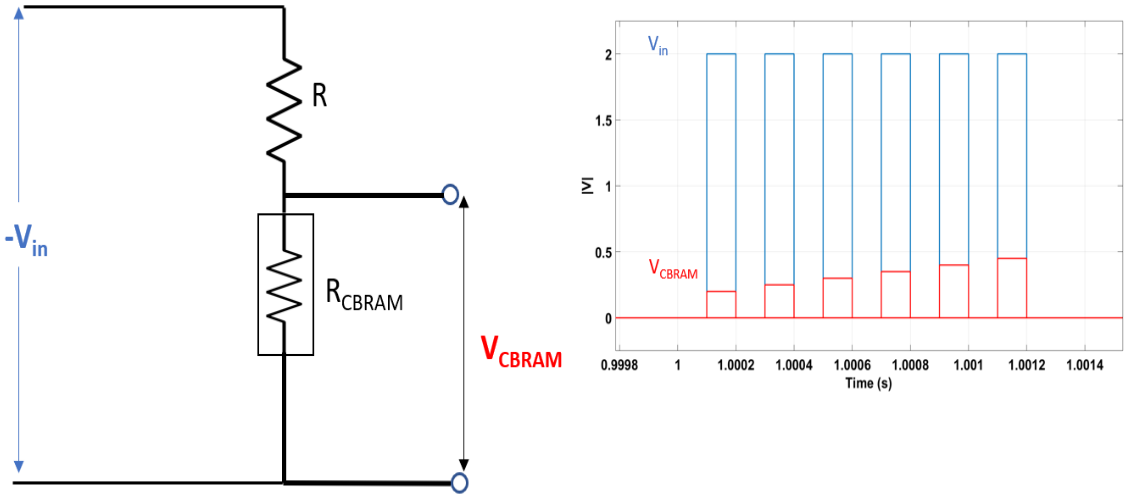


Figure 40. Two resistor circuit to implement increasing amplitude effect.

One proposed design is shown in Figure 41. In this design, the constant voltage pulse input, V_{IN}, toggles a custom 4-bit mixed signal counter, CNTR1, which selects eight different bias currents to CBRAM device. When the 4th bit of CNTR1 is set low, a second

4-bit counter, CNTR2, is enabled. The first 3 output bits of CNTR2 combine with the CNTR1 currents to add another four monotonically increasing currents to the sequence (the last three 3-bit outputs of CNTR 2 can be neglected as the CBRAM device will have reached a level close to its minimum value by that step). The 4th bit of CNTR2 is fed back to the enable switch of CNTR1. When the 4th bit on CNTR2 is toggled, CNTR1 is disabled, and the pulse sequence terminated. The currents are set by biasing a select combination of p-channel transistors in saturation mode. The counter outputs are toggled between VDD (off-state) and VDD-V_{ref} (saturation), where V_{ref} is a tunable voltage slightly larger than the p-channel threshold voltage, which ensures the p-channel transistors are biased in saturation. When the input voltage is low, an n-channel transistor in parallel with R_{CBRAM} is activated to shunt current to ground, which debiases the R_{CBRAM} during that half cycle of the input pulse.

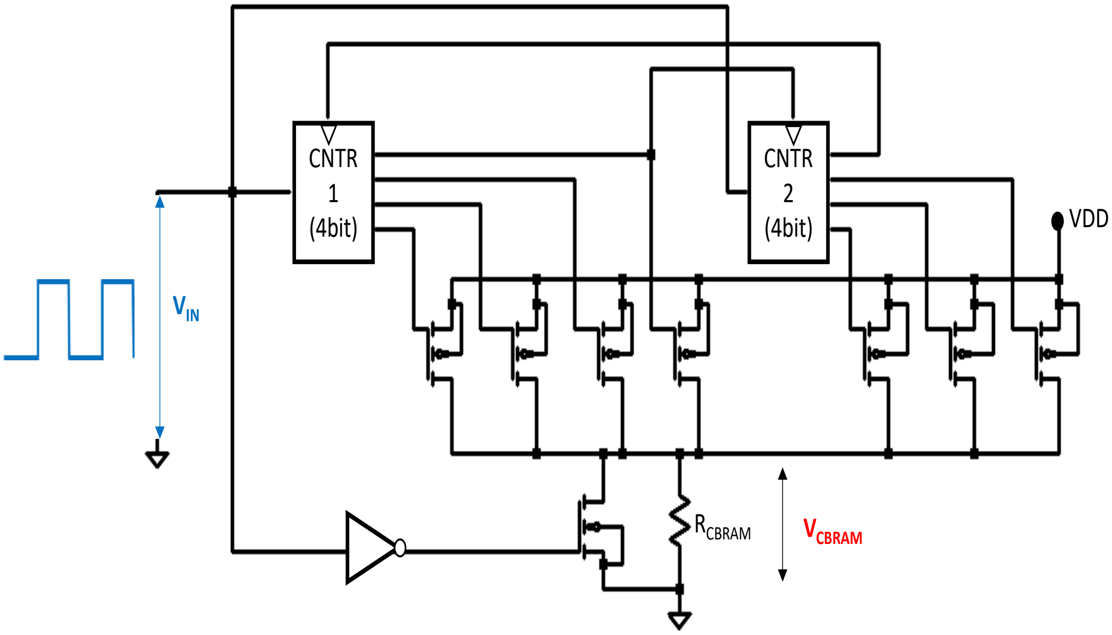


Figure 41. Switched current design for potentiation.

The increasing current magnitude sequence selected by the counter is determined as follows. To achieve a V_{CBRAM} that increases approximately linearly as R_{CBRAM} is reduced, the following condition must be satisfied:

$$V_{CBRAM,i} = R_{CBRAM,i} I_i \approx M R_{CBRAM,i} + B. \tag{4.4}$$

In Eq. 1, I_i is the selected current at pulse number i. M is a negative value and B is positive value set to match the targeted (V_{CBRAM}, R_{CBRAM}) order pairs, e.g., (0.2V,2kΩ) and (1.36V, 730Ω). By rearranging Eq. 1, the targeted I_i is therefore,

$$I_i \approx M + \frac{B}{R_i} \tag{4.5}$$

Figure 4.26 shows the simulated response of R_{CBRAM} and V_{CBRAM} with the input voltage pulse sequence for V_{IN}. It should be noted that using switch programming, the hardware of the potentiation circuit (Figure 42) can be reconfigured for use in R_{CBRAM} depression.

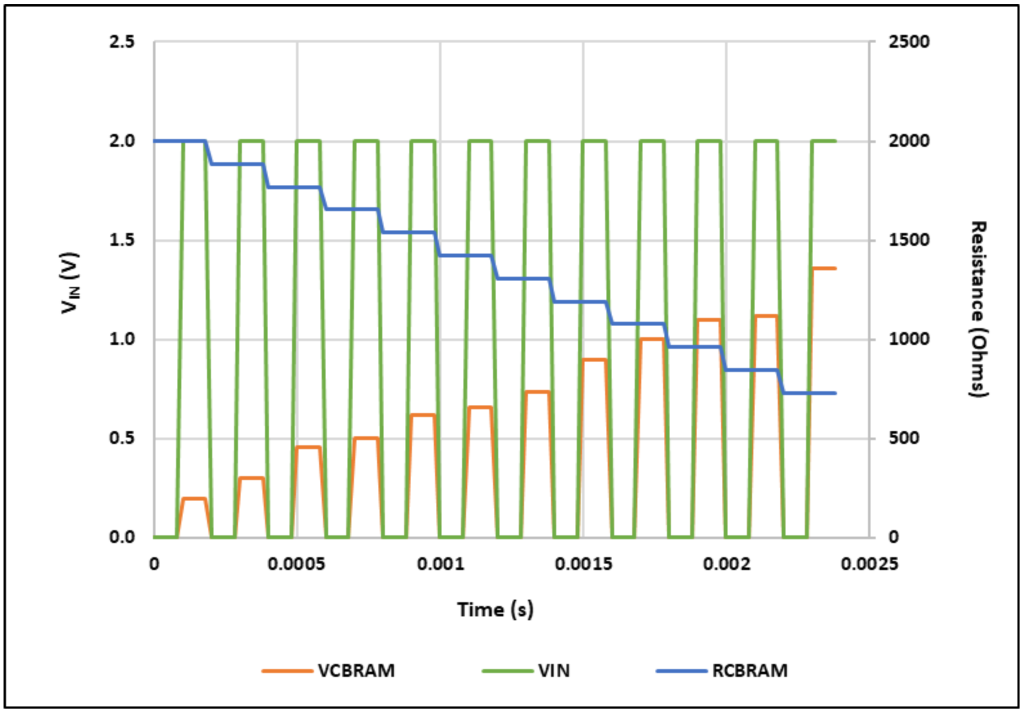


Figure 42. Reducing Resistance and increasing VCBRAM for input voltage pulse sequence.

CHAPTER 5

IMPACT OF TID ON THE ANALOG CONDUCTANCE AND ACCURACY OF CBRAM-BASED NEURAL ACCELERATOR

5.1. Introduction

Apart from acting as storage class memories as this thesis has shown, resistive random-access memories (RRAM) can be used train deep neural networks (DNN) [75]. Training a DNN is computationally intensive and difficult to implement in edge devices such as spacecrafts, rovers, or satellites. A special purpose accelerator based on analog RRAM crossbars, has shown to potentially enable real-time training on embedded systems [31],[32]. The conductance states of the RRAM devices are used to store the synaptic weights within the neural network. Vector-matrix multiplication can be performed within the crossbar array by driving the rows of the array with appropriate input activation functions. Conductive Bridge RAM (CBRAM), one class of RRAM, have demonstrated synaptic capability for neuromorphic applications in [76], [77]. CBRAM devices fabricated with Ag-$Ge_{30}Se_{70}$ material stacks and operating as digital (binary) memory elements have shown excellent tolerance to total ionizing dose (TID) up to 10 Mrad [78], [79]. With such high radiation tolerance, these devices could potentially be used in a variety of systems for space applications.

Though the above-mentioned studies have indicated high TID tolerance levels, these results do not completely rule out the radiation sensitivity of the same CBRAM devices in an analog neural accelerator. The operation of CBRAMs acting as digital devices might not be affected by small radiation-induced changes in conductance but when the same device is used in an analog accelerator a small change in conductance might bring about a large degradation in training accuracy. Indeed, in a neural accelerator the RRAM device is often changed by a small fraction of the entire conductance range during every neural weight update. Hence it becomes crucial to study the impact of cumulation radiation dose on the analog device characteristics of RRAM, specifically CBRAM in this thesis, and the impact this would have on training a neuromorphic computing accelerator. Only one study to date has investigated the effect of TID exposure on the operation of CBRAM used as a neuromorphic synapse [80]. This chapter explores the effect of TID on the specific analog behavior of Ag-$Ge_{30}Se_{70}$ CBRAM devices. The experimental results are further used to model the training accuracy on an analog accelerator before and after irradiation.

5.2. Experimental Setup

CBRAM tiles were diced from the wafer and crossbar rows and columns were wire-bonded in a pin grid array (PGA) package, as shown in Figure 43. One package was used as a control test chip and the other was exposed to ionizing radiation. Control and exposed CBRAM devices were subjected to identical DC cycling using an Agilent 4156C parameter analyzer to verify functionality. The devices were swept from 0V to 0.5V and back to 0V in 10mV steps for programming and from 0V to -0.5V to erase the devices. The I-V characteristics of the control and the exposed CBRAM devices are depicted in Figure 44 & Figure 45 respectively. For this experiment, two devices were tested on both the control and exposed test chips.

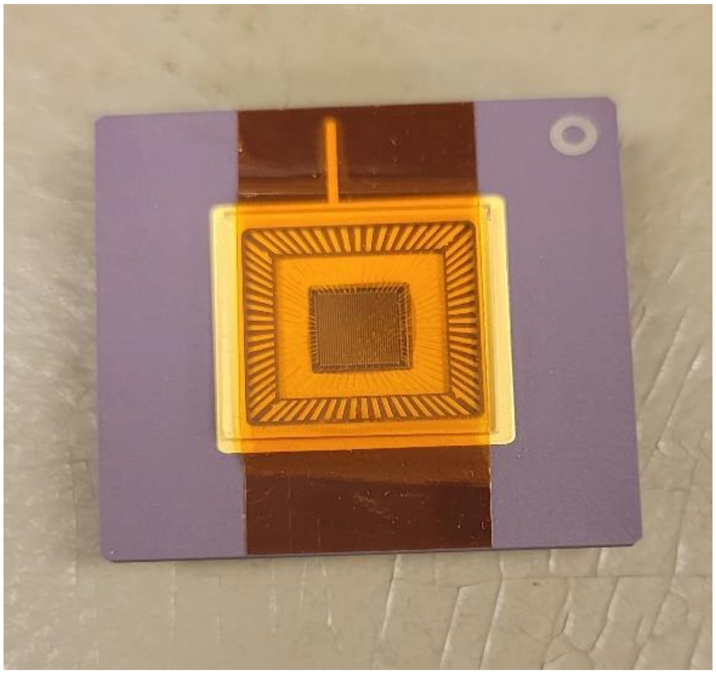


Figure 43. Grid array package used for radiation testing.

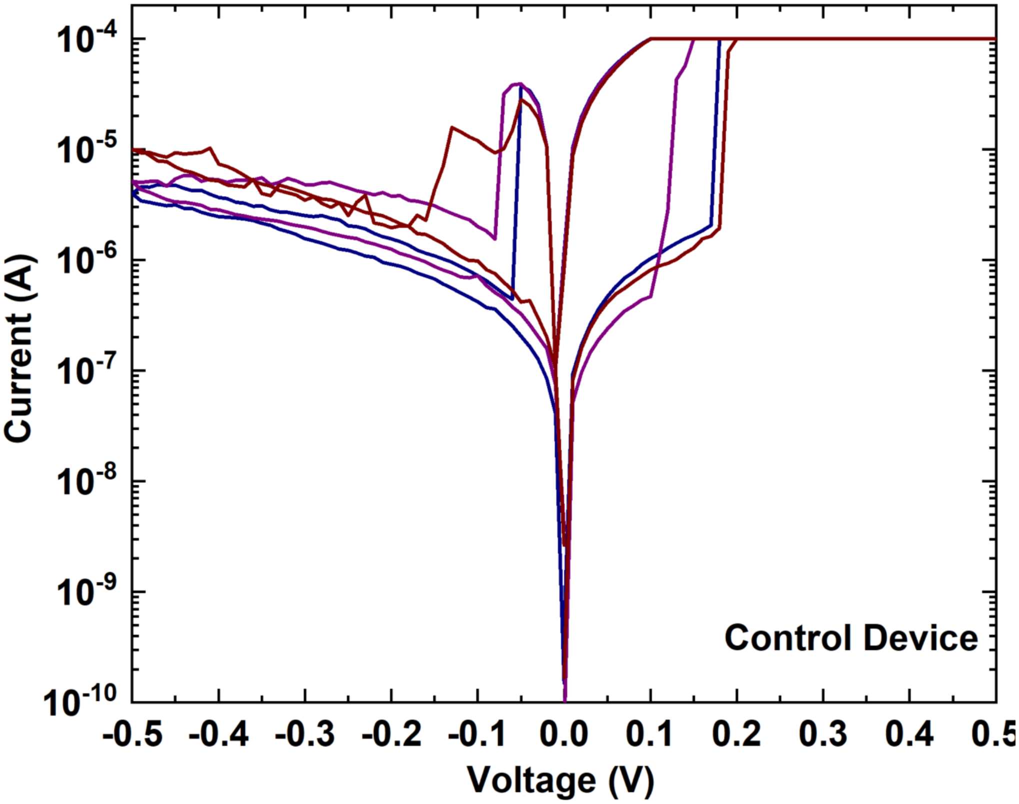


Figure 44. I-V characteristics of Control sample prior to irradiation.

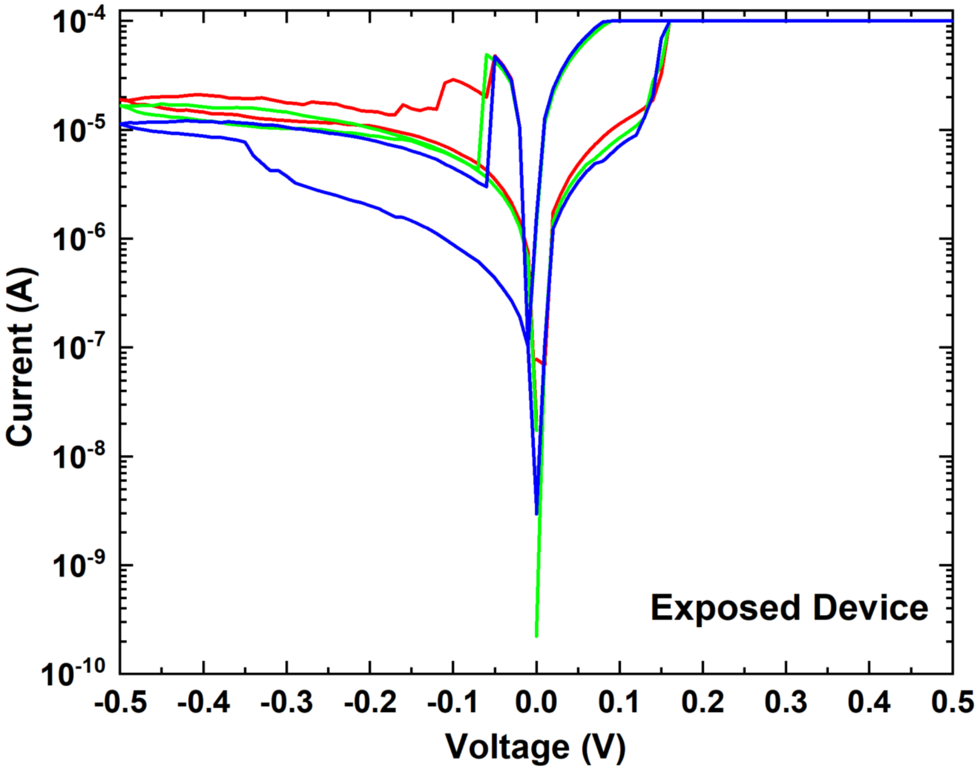


Figure 45. I-V characteristics of exposed device prior to irradiation.

The exposed chip was irradiated with ^{60}Co gamma-rays in the Gamma Cell 220 irradiator at ASU. The package was placed into a socket which was connected to a test board. All device terminals were grounded while the test board was placed inside the irradiation chamber. Figure 36 is a picture of the experimental setup. The dose level inside the chamber was 185.91 rad ($Ge_{30}Se_{70}$)/min. The devices were removed from the chamber after a fixed dose and tested at room temperature. A summary of the dose step and exposure time is shown Table 5. The devices were tested for TID of up to 1Mrad ($Ge_{30}Se_{70}$).

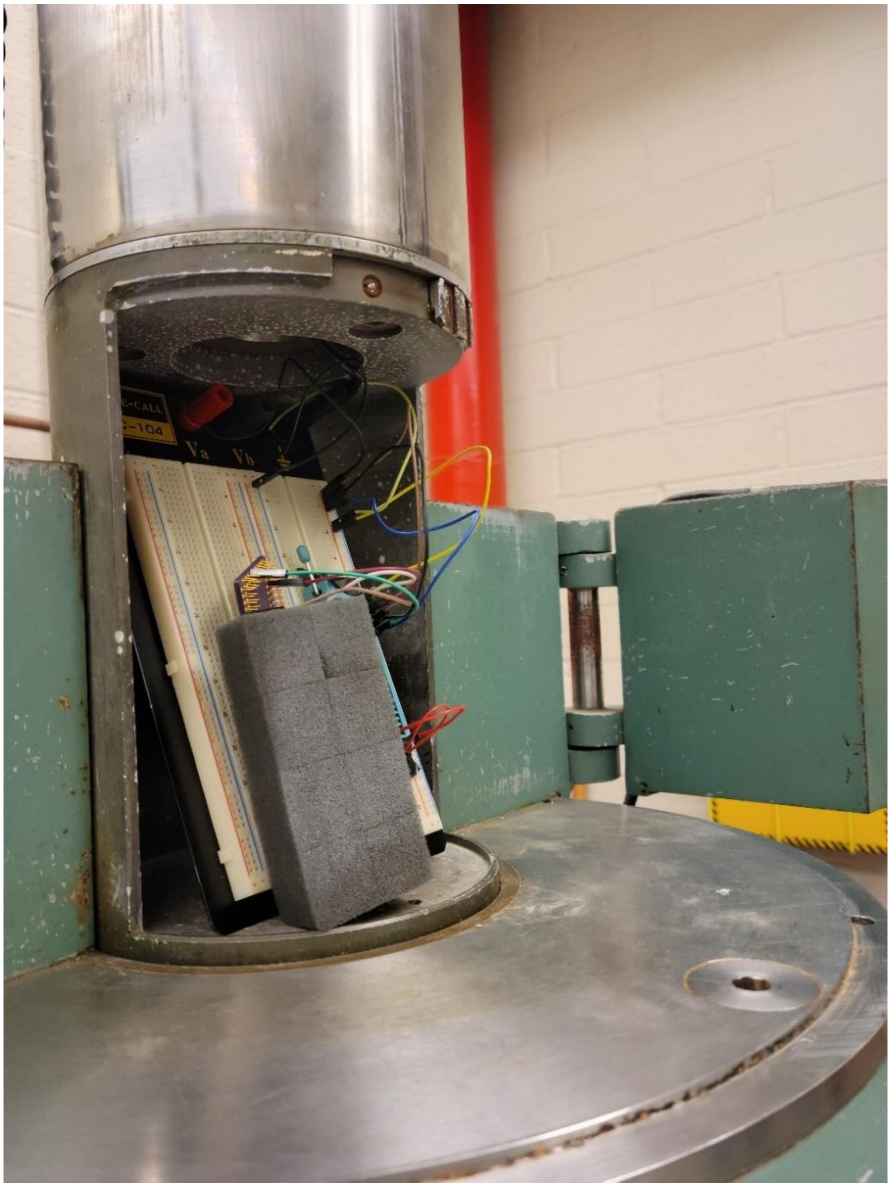


Figure 46. Experimental test setup for irradiation experiments.

Table 5.Dose step and exposure times.

STEP STRESS LEVELS (krad ($Ge_{30}Se_{70}$))	EXPOSURE TIME (hrs.)
50	4.56
100	4.56
250	13.29
500	22.41
700	17.93
1000	26.89

After irradiation, the test and control devices were subjected to pulse testing. Pulse measurements were performed using the Keithley 4200 SCS parameter analyzer with built-in 4225 PMU module. To achieve gradual switching behavior, we applied 100 consecutive SET (potentiation) pulses (0.35V for 100µs) and 50 consecutive RESET (depression) pulses (-0.25V for 100µs). Each write and erase pulse was followed by read voltage of 30 mV to extract conductivity. At each dose step, 20 cycles of potentiation and depression were performed.

6.3. Experimental Results

The pulse responses of the control and the test devices prior to irradiation are shown in Figures. 47 and 48, respectively. The pulse amplitude and the pulse width were chosen to obtain a smooth monotonic change in conductance for both potentiation and depression. The exposed CBRAM devices were tested at the six TID levels listed in Table 5. The programming responses at each TID level except 1Mrad ($Ge_{30}Se_{70}$) are shown in Figures. 49-53. The maximum conductance level for the control sample obtained at the same time as the step stress measurements on exposed devices is indicated by the red line on Figures. 48-53. For clear observability of the conductance change profile vs. pulse number, only 5 of the 20 potentiation/depression cycles are shown. Exposed responses after 1Mrad ($Ge_{30}Se_{70}$) are not provided since those devices failed to switch at that dose level.

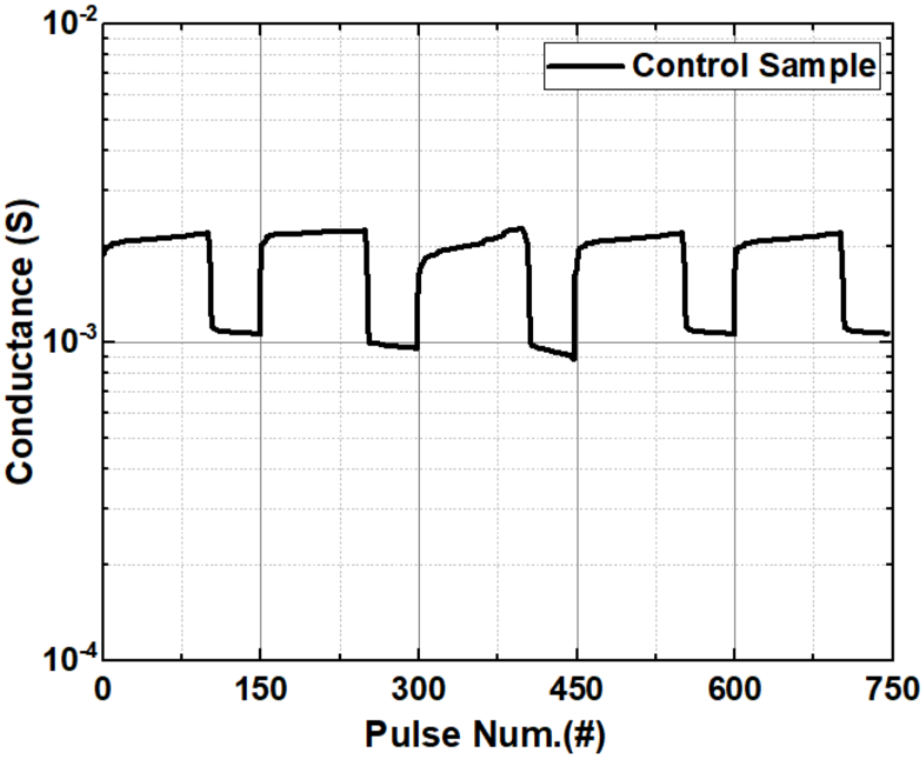


Figure 47. Potentiation and depression characteristics of the device of the control sample.

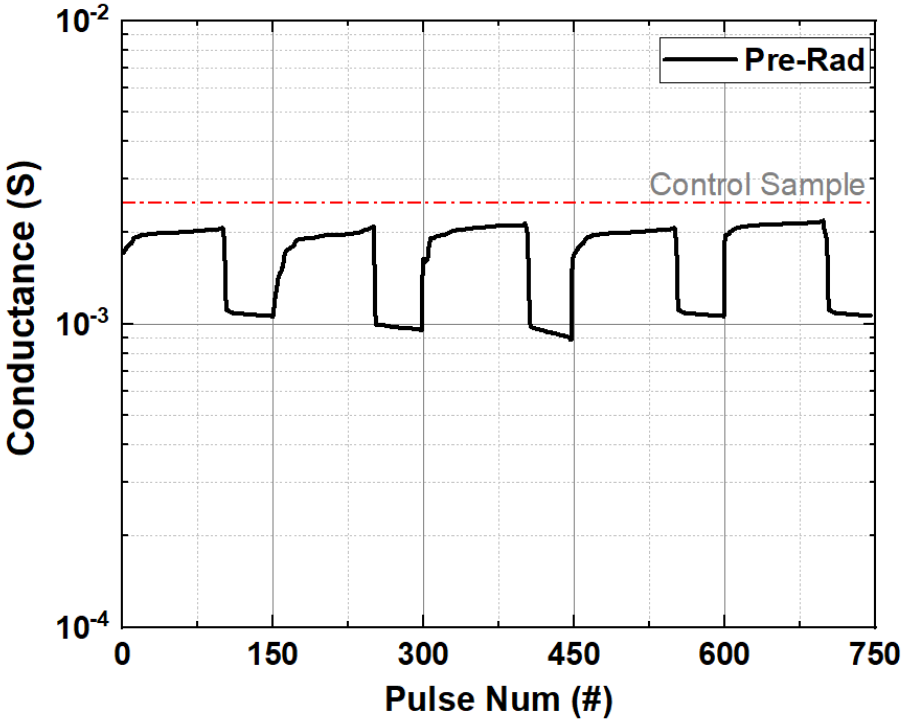


Figure 48. Potentiation and depression characteristics of the pre-rad device.

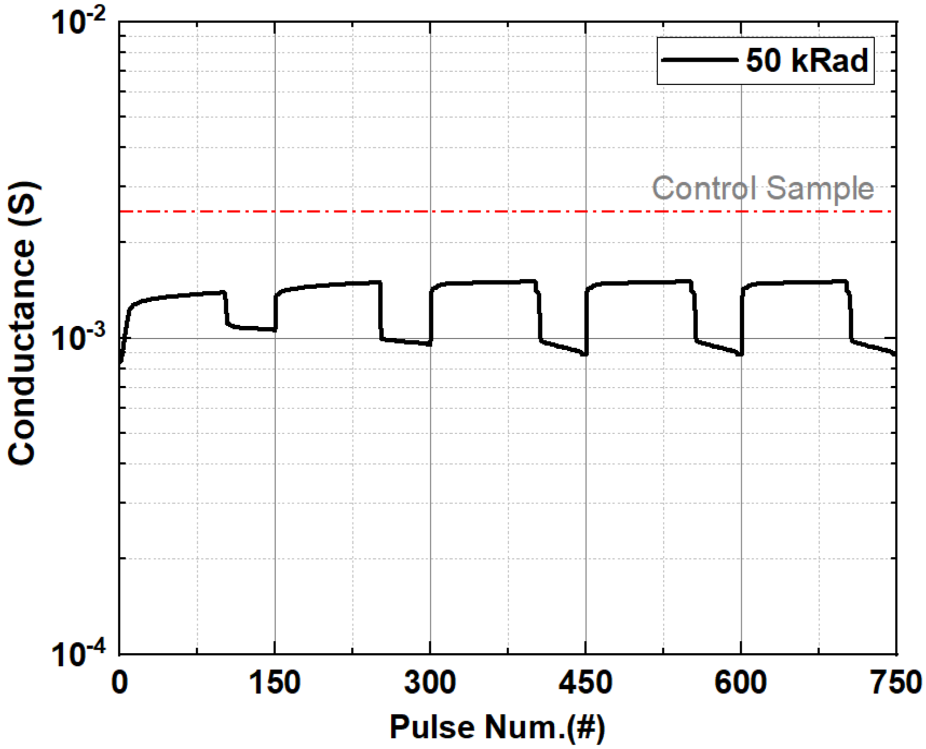


Figure 49. Potentiation and depression characteristics of the device at 50kRad.

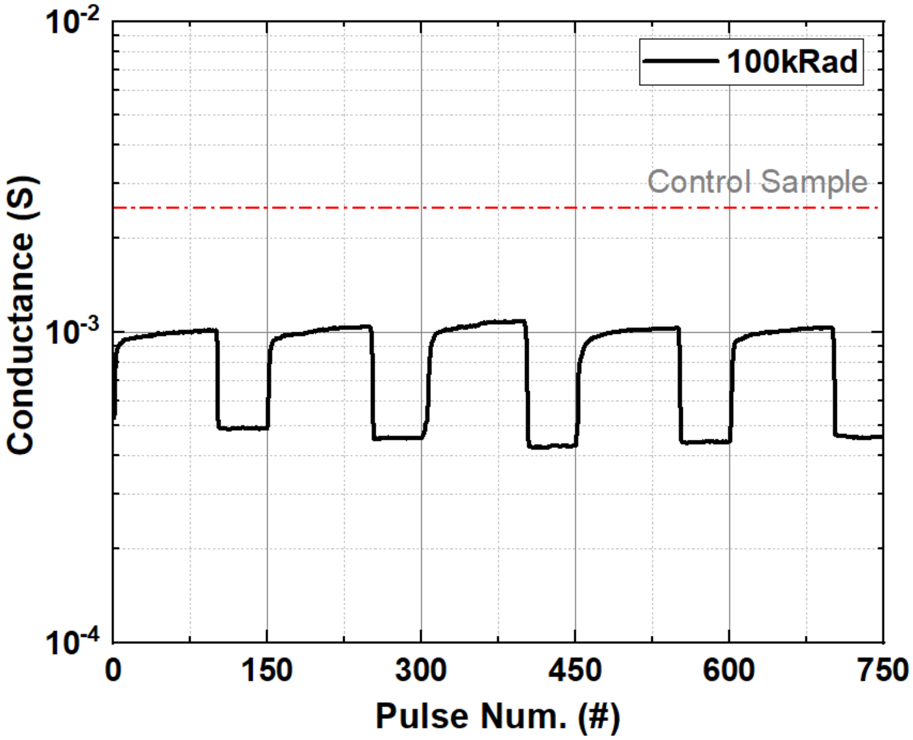


Figure 50. Potentiation and depression characteristics of the device at 100kRad.

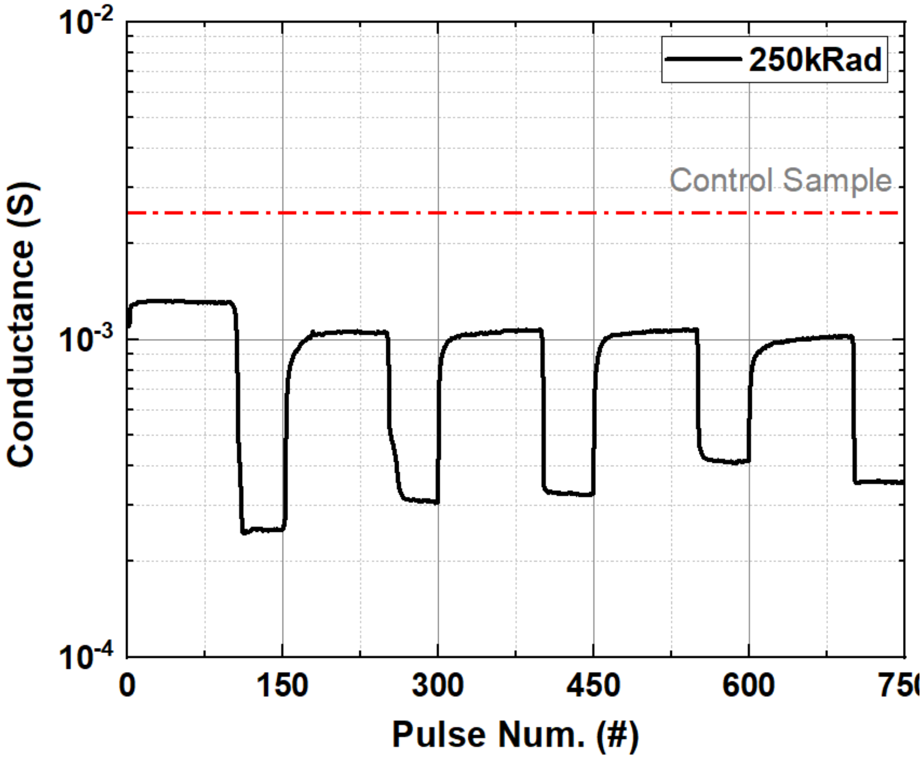


Figure 51. Potentiation and depression characteristics of the device at 250kRad.

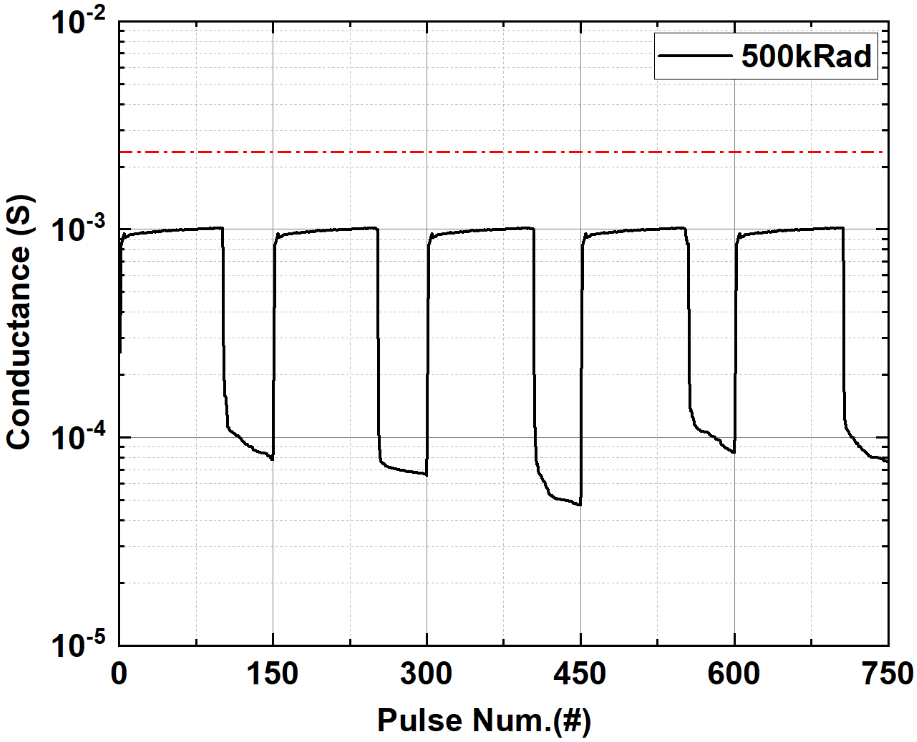


Figure 52. Potentiation and depression characteristics of the device at 500kRad.

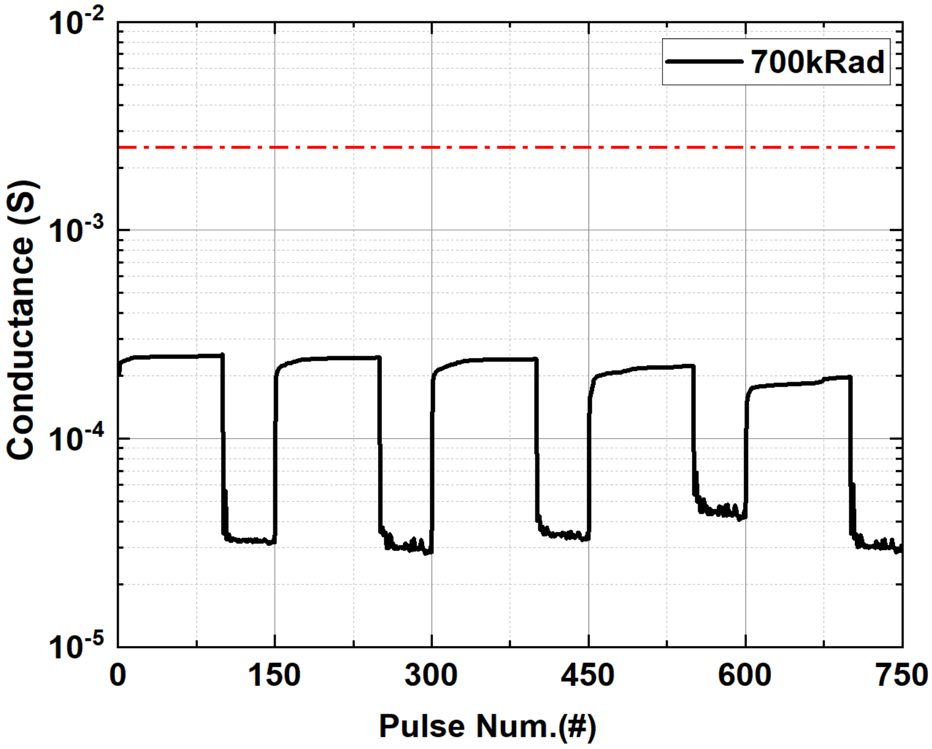


Figure 53. Potentiation and depression characteristics of the device at 700kRad.

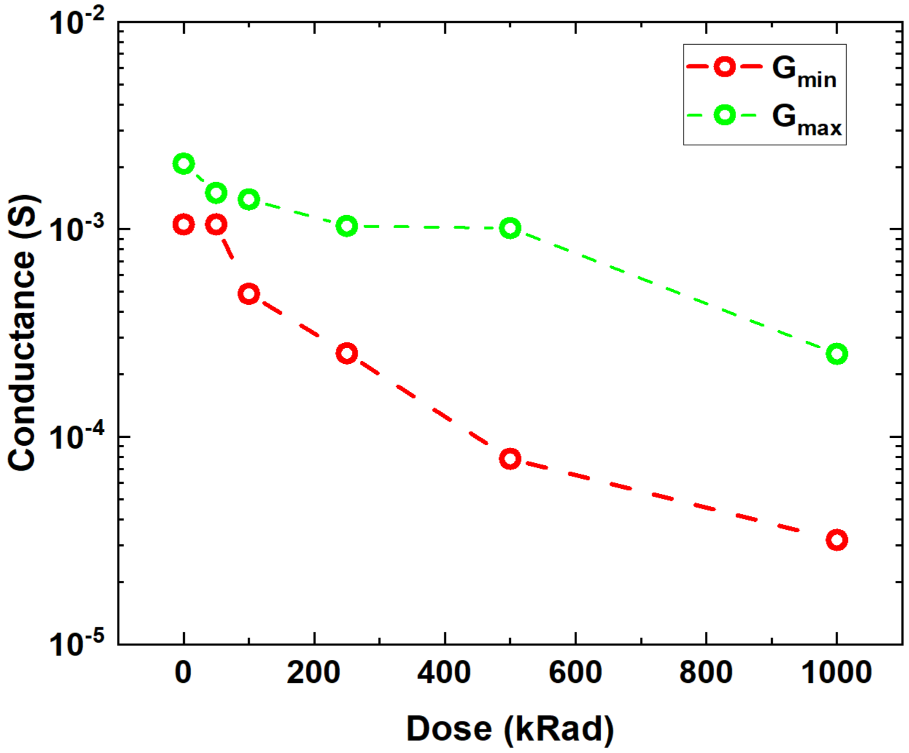


Figure 54. Effect of Dose step on the G_{min} and G_{max} of Ag-$Ge_{30}Se_{70}$ device.

As is typical for many RRAM technologies, even prior to irradiation, the pulsed conductance response is highly non-linear. As the figures show, the conductance change per pulse is large at the beginning of each pulse cycle, then the rate of change slows substantially until the conductance converges to a maximum or minimum value. While linear potentiation/depression characteristics are the most ideal for analog synapses, and indeed, some devices can show this (see Figure 27 in previous chapter), recent RRAM-based accelerator studies have presented algorithms for achieving good accuracy even with non-linear characteristics. What is critical to training accuracy is that the range and cycle-to-cycle variation of the conductance behavior does not change significantly over time and stress. It is evident from the post-irradiation data that the range of conductance change is altered by TID. For potentiation (thickening of the conductive filament) the maximum conductance achieved decreases monotonically with increasing TID. This effect becomes pronounced for TID levels of 100 krad ($Ge_{30}Se_{70}$) and above. The depression programming (thinning of the conductive filament) is also affected with increasing TID levels. We observe that the minimum conductance is monotonically decreasing with increasing TID as well.

5.4. Modelling Radiation Effects on CBRAM-based Neural Accelerator

The TID impact on Ag-$Ge_{30}Se_{70}$ CBRAM devices is further studied by modelling these devices in an analog neural training accelerator. Here we use the open-source CrossSim simulator developed at Sandia National Laboratories to perform the training [81]. An artificial two-layer neural network, shown in Figure. 54, was simulated using the pre- and post-rad data shown in Figures. 48-53. The neural network was trained using a backpropagation algorithm. This algorithm uses three important kernels: vector matrix multiply (VMM), matrix vector multiply (MVM) and outer product update to train the neural network. For this study, we use a small version (8 X 8 pixels) of handwritten digits from the ‘Optical Recognition Handwritten Digits’ dataset [68]. This two-layer neural network consisting of 64 input, 36 hidden and 10 output nodes for small image dataset was used for training.

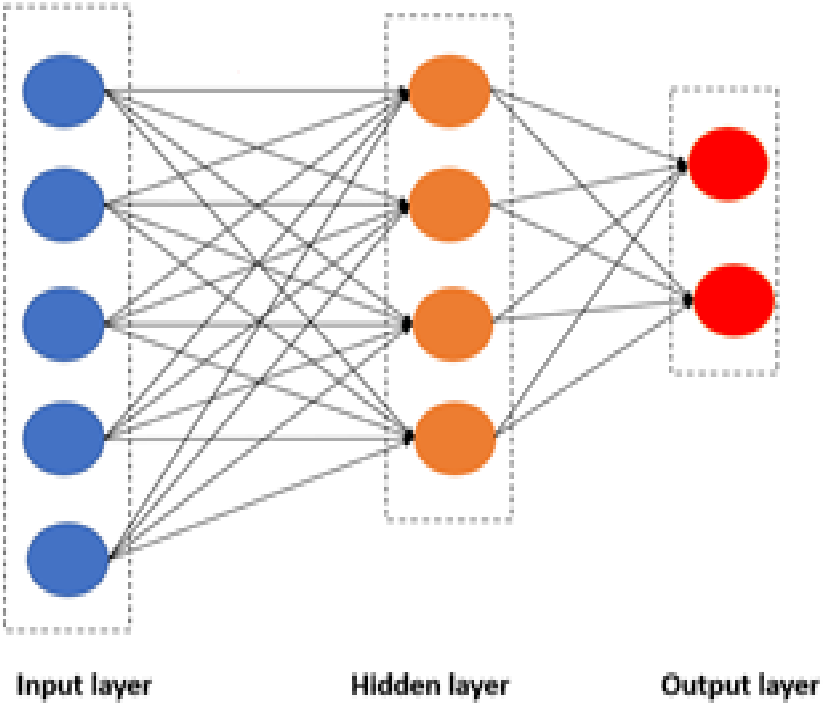


Figure 54. Simulated 2-layer neural network.

To simulate the response of the Ag-$Ge_{30}Se_{70}$ CBRAM devices in the neural network, a look up table is created containing $\Delta G/G$ response of the devices to multiple potentiation and depression cycles at a given TID level [70]. The training accuracy of the neural network was seen to degrade monotonically from 86 % prior to irradiation to 15 % at the 700 krad ($Ge_{30}Se_{70}$) dose level. A plot of classification accuracy versus the neural network iteration (epoch) for the test device at various TID levels is shown in Figure. 55.

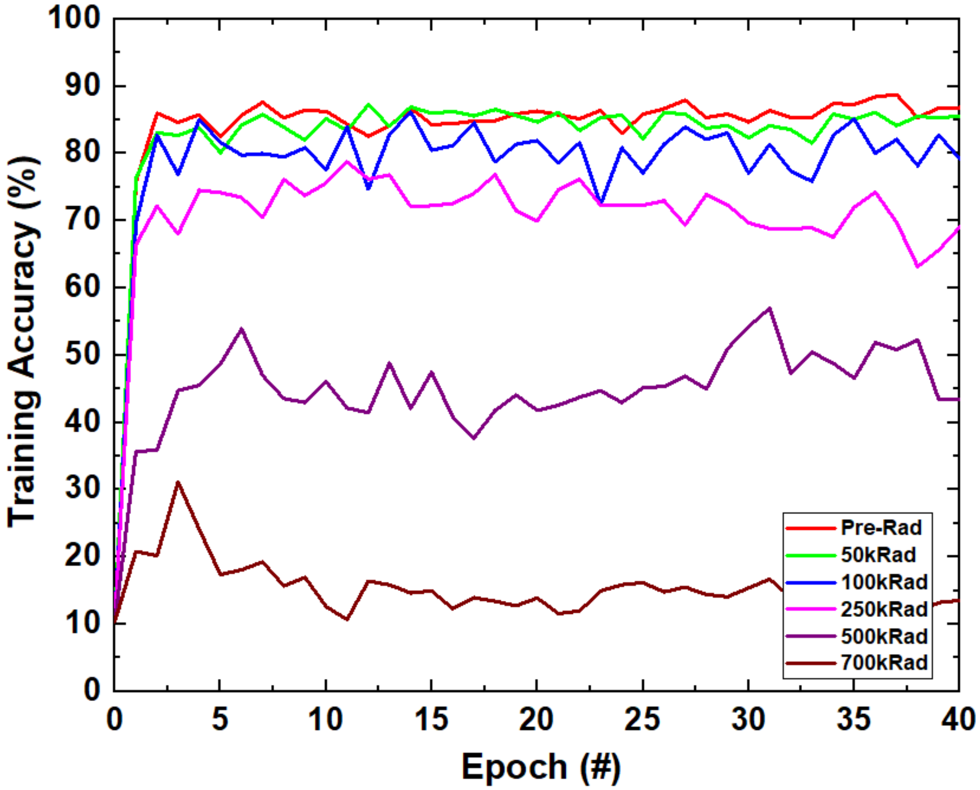


Figure 55. Training accuracy of Ag-Ge30Se70 CBRAM devices at different TID levels.

6.5. Characterization of Ag-$Ge_{30}Se_{70}$ CBRAM devices after exposure to gamma rays

To investigate the monotonic reduction in conductance after irradiation, Field Emission Scanning Electron Microscopy (FESEM) was performed on the samples. FESEM is a scanning electron microscope used for detailed image analysis of devices and circuits. It can magnify images up to 500,000X and can resolve features down to 2nm. Figure. 56(a) shows the FESEM image of a biased un-irradiated sample and Figure 56(b) shows the image of a biased irradiated sample at 1MRad. The white dots in the post-irradiation image reveal TID-induced crystallization across the tile, which is more pronounced above the active device. Previous studies have revealed that such crystallization can result in resistance state locking in CBRAM cells [82] and [83]. In our case, the devices seem to be converging to a high resistance, locked state.

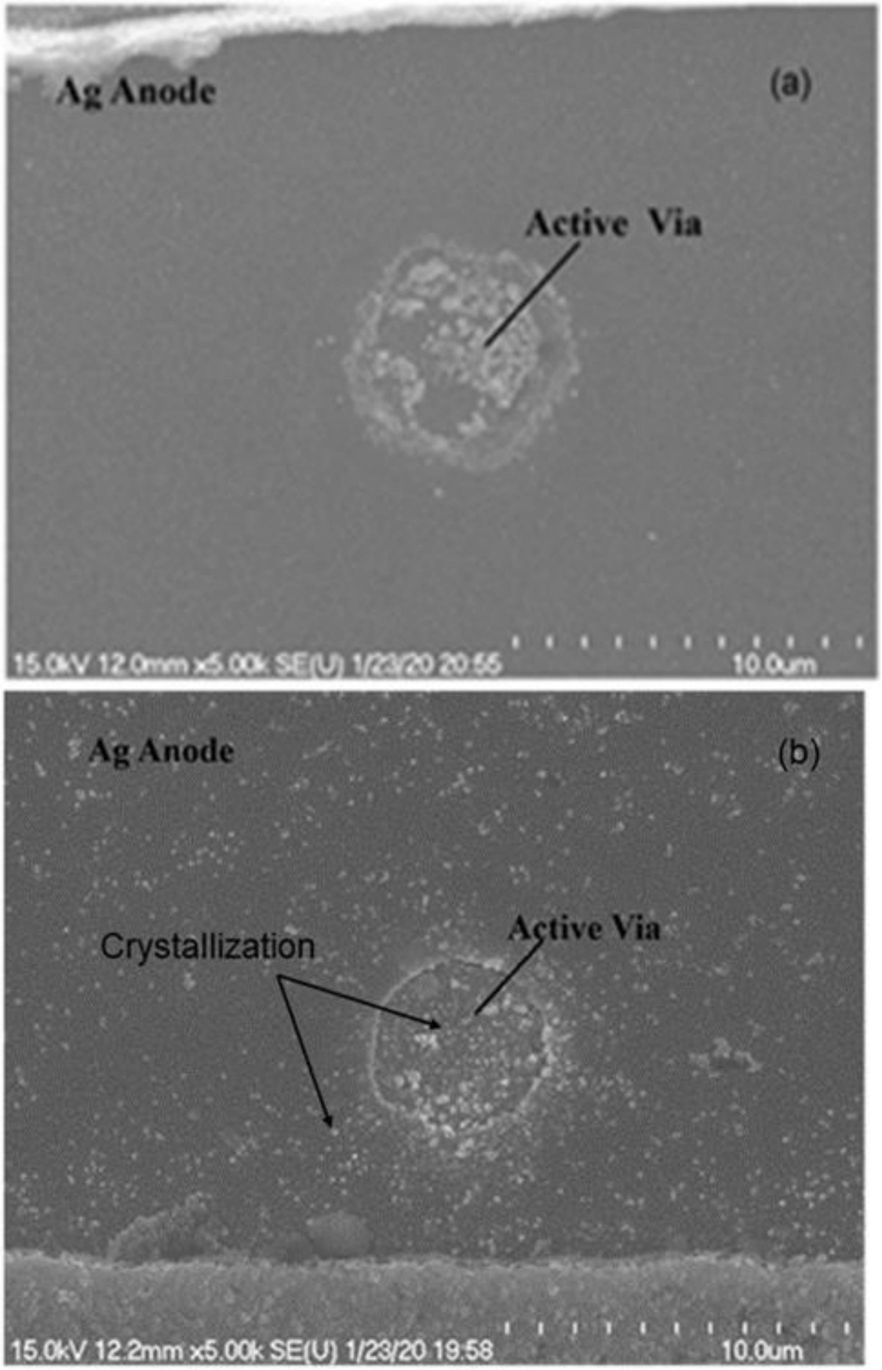

Ag Anode
(a)
Active Via
15.0kV 12.0mm x5.00k SE(U) 1/23/20 20:55
10.0um
Ag Anode
(b)
Crystallization
Active Via
15.0kV 12.2mm x5.00k SE(U) 1/23/20 19:58
10.0um

Figure 56. FESEM images (a) biased pre-radiated and (b) biased irradiated device.

It has been observed in previous studies that for pristine Se-rich films, Ag introduced by photodoping is seen to form Ag_2Se crystalline phases throughout the chalcogenide glass [84], [85]. These Ag_2Se nanocrystals exist in two phases – *α*- Ag_2Se and *β*- Ag_2Se. The α phase of Ag_2Se is a body centered cubic with high ionic conductivity while the β phase of Ag_2Se is an orthorhombic structure with low ionic conductivity [86], [87], [88]. In previous XRD studies [86], it was observed that at a high TID level of 4.5 Mrad, creation of β-Ag_2Se occurs. The reduction in ionic conductivity of Ag-$Ge_{30}Se_{70}$ material can be attributed to the increasing percentage of β- Ag_2Se as depicted in Figure 57.

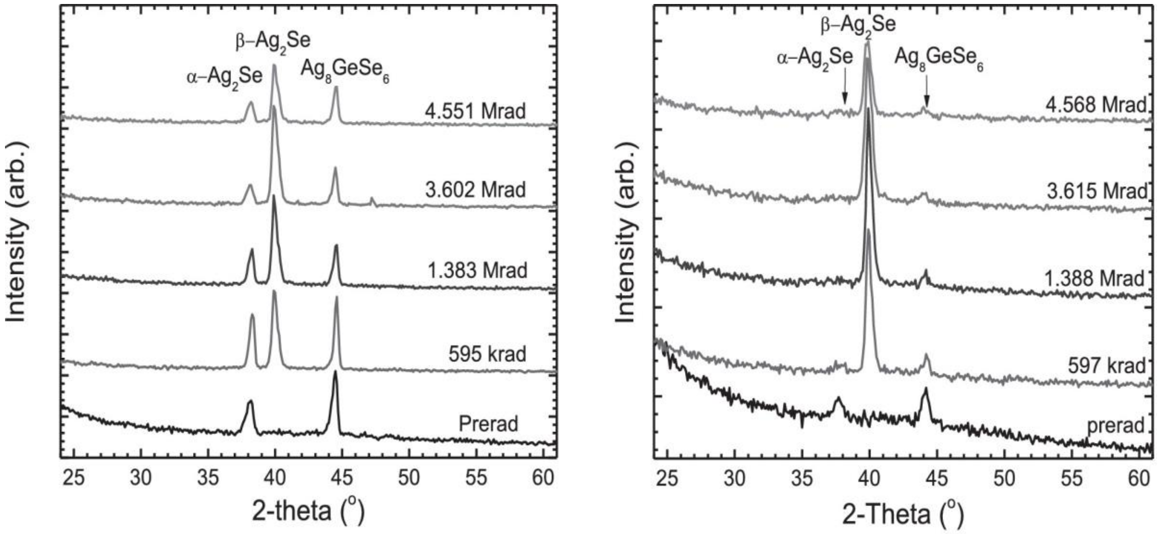


Figure 57. XRD results depicting the formation of *β* - Ag_2Se with increasing TID levels [15] Copyright © 2014 AIP Publishing LLC.

The formation of crystallites on top of the electrode reduces the availability of active Ag, which is necessary for making contact to and for broadening the existing filament. EDS mapping results of Ag-$Ge_{30}Se_{70}$ PCM devices shown in [89] confirm that the clusters or crystals appearing on the top of the devices are silver-rich coming from the depleted Ag regions. Hence maximum conductance is not reached for the same train of applied pulses after irradiation. In addition to reducing the available active metal, another deleterious impact of the buildup of these phases is that they increase the resistivity of the top electrode, which leads to a large voltage drop at the contact. Therefore, the pulse amplitude is not large enough to bring about the required change in the conductance. Another potential cause of reduced conductivity is the oxidation of Ge to GeO_2 in a radiation environment as was previously studied in [90]. Due to oxidation, the conductivity of the switching layer is reduced. Hence the switching layer becomes more insulative and is not able to transport Ag+ ions to the filament. Under these post-irradiated conditions, the neuromorphic pulse response of the device is severely impacted due to the formation of different Ag phases as well as the reduction in the conductivity of the switching layer.

CHAPTER 6

CONCLUSION

In conclusion, we have demonstrated the response of Ag-$Ge_{30}Se_{70}$ CBRAM to two different pulse schemes-constant amplitude and increasing amplitude pulse scheme. Also, the effect of varying pulse width was studied. The device shows a 10x increase in conductance for shorter pulse width but for longer pulse width we observe an increase in conductance range attributed to widening of the filament. A more linear and symmetric synaptic response of the device to increasing amplitude pulse response was observed. Further, an artificial neural network was simulated based on the measured device properties to demonstrate supervised learning abilities. The training accuracy was found to improve from 80% to 92% for small image dataset and from 79% to 87% for large image dataset using the increasing pulse amplitude scheme. The area, latency, energy consumption of the crossbar array architecture using Ag-$Ge_{30}Se_{70}$ devices was simulated and compared for the two different pulse schemes. To address the issue of overhead latency and energy consumption associated with using variable amplitude scheme, a new circuit design approach is proposed. The impact of total ionizing dose (TID) on the analog conductance behavior for the fabricated devices was performed. The devices were exposed to ^{60}Co gamma rays up to radiation levels of 1 Mrad ($Ge_{30}Se_{70}$). The maximum and minimum conductance ranges were impacted by radiation exposure. Up to a TID level of 100 krad ($Ge_{30}Se_{70}$), the minimum and maximum conductance show some reduction, but above 100 krad ($Ge_{30}Se_{70}$), we observe a large decrease in the conductance range. FESEM imaging of irradiated samples indicated the formation of Ag phases that has led to a reduction of maximum and minimum conductance for the device. From the simulation of a neural accelerator using the pre- and post-rad data, it was seen that the training accuracy is significantly degraded by ionizing radiation.

REFERENCES

1. Von Neumann, J. (1993). First Draft of a Report on the EDVAC. *IEEE Annals of the History of Computing*, *15*(4), 27-75.

2. Pandiyan, D., & Wu, C. J. (2014, October). Quantifying the energy cost of data movement for emerging smart phone workloads on mobile platforms. In *2014 IEEE International Symposium on Workload Characterization (IISWC)* (pp. 171-180). IEEE.

3. Kestor, G., Gioiosa, R., Kerbyson, D. J., & Hoisie, A. (2013, September). Quantifying the energy cost of data movement in scientific applications. In *2013 IEEE international symposium on workload characterization (IISWC)* (pp. 56-65). IEEE.

4. Nvidia. (2016). *Accelerating AI with GPUs: A New Computing Mode.* https://blogs.nvidia.com/blog/2016/01/12/accelerating-ai-artificialintelligence-gpus/

5. Deng, L., Hinton, G., & Kingsbury, B. (2013, May). New types of deep neural network learning for speech recognition and related applications: An overview. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 8599-8603). IEEE.

6. Egmont-Petersen, M., de Ridder, D., & Handels, H. (2002). Image processing with neural networks—a review. *Pattern recognition*, *35*(10), 2279-2301.

7. Egmont-Petersen, M., de Ridder, D., & Handels, H. (2002). Image processing with neural networks—a review. *Pattern recognition*, *35*(10), 2279-2301.

8. Harrington, P. D. B. (1993). Sigmoid transfer functions in backpropagation neural networks. Analytical Chemistry, 65(15), 2167-2168.

9. Hara, K., Saito, D., & Shouno, H. (2015, July). Analysis of function of rectified linear unit used in deep learning. In 2015 international joint conference on neural networks (IJCNN) (pp. 1-8). IEEE.

10. Leung, H., & Haykin, S. (1991). The complex backpropagation algorithm. IEEE Transactions on signal processing, 39(9), 2101-2104.

11. Rojas, R. (1996). The backpropagation algorithm. In Neural networks (pp. 149-182). Springer, Berlin, Heidelberg.

12. Leonard, J., & Kramer, M. A. (1990). Improvement of the backpropagation algorithm for training neural networks. Computers & Chemical Engineering, 14(3), 337-341.

13. Xiao, T. P., Bennett, C. H., Feinberg, B., Agarwal, S., & Marinella, M. J. (2020). Analog architectures for neural network acceleration based on non-volatile memory. *Applied Physics Reviews*, *7*(3), 031301.

14. LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278-2324.

15. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, *115*(3), 211-252.

16. Jouppi, N. P., Young, C., Patil, N., Patterson, D., Agrawal, G., Bajwa, R., ... & Yoon, D. H. (2017, June). In-datacenter performance analysis of a tensor processing unit. In Proceedings of the 44th annual international symposium on computer architecture (pp. 1-12).

17. Chen, Y., Yang, T. J., Emer, J., & Sze, V. (2018). Understanding the limitations of existing energy-efficient design approaches for deep neural networks. Energy, 2(L1), L3.

18. Coates, A., Huval, B., Wang, T., Wu, D., Catanzaro, B., & Andrew, N. (2013, May). Deep learning with COTS HPC systems. In International conference on machine learning (pp. 1337-1345). PMLR.

19. Raina, R., Madhavan, A., & Ng, A. Y. (2009, June). Large-scale deep unsupervised learning using graphics processors. In Proceedings of the 26th annual international conference on machine learning (pp. 873-880).

20. Farabet, C., Poulet, C., Han, J. Y., & LeCun, Y. (2009, August). Cnp: An fpga-based processor for convolutional networks. In 2009 International Conference on Field Programmable Logic and Applications (pp. 32-37). IEEE.

21. Farabet, C., Martini, B., Corda, B., Akselrod, P., Culurciello, E., & LeCun, Y. (2011, June). Neuflow: A runtime reconfigurable dataflow processor for vision. In Cvpr 2011 Workshops (pp. 109-116). IEEE.

22. Zhang, C., Li, P., Sun, G., Guan, Y., Xiao, B., & Cong, J. (2015, February). Optimizing fpga-based accelerator design for deep convolutional neural networks. In Proceedings of the 2015 ACM/SIGDA international symposium on field-programmable gate arrays (pp. 161-170).

23. Chakradhar, S., Sankaradas, M., Jakkula, V., & Cadambi, S. (2010, June). A dynamically configurable coprocessor for convolutional neural networks. In Proceedings of the 37th annual international symposium on Computer architecture (pp. 247-257).

24. Wei, X., Yu, C. H., Zhang, P., Chen, Y., Wang, Y., Hu, H., ... & Cong, J. (2017, June). Automated systolic array architecture synthesis for high throughput CNN inference on FPGAs. In Proceedings of the 54th Annual Design Automation Conference 2017 (pp. 1-6).

25. Putnam, A., Caulfield, A. M., Chung, E. S., Chiou, D., Constantinides, K., Demme, J., ... & Burger, D. (2014, June). A reconfigurable fabric for accelerating large-scale datacenter services. In 2014 ACM/IEEE 41st International Symposium on Computer Architecture (ISCA) (pp. 13-24). IEEE.

26. Chung, E., Fowers, J., Ovtcharov, K., Papamichael, M., Caulfield, A., Massengill, T., ... & Burger, D. (2018). Serving dnns in real time at datacenter scale with project brainwave. iEEE Micro, 38(2), 8-20.

27. Jouppi, N., Young, C., Patil, N., & Patterson, D. (2018). Motivation for and evaluation of the first tensor processing unit. IEEE Micro, 38(3), 10-19.

28. Chen, Y., Luo, T., Liu, S., Zhang, S., He, L., Wang, J., ... & Temam, O. (2014, December). Dadiannao: A machine-learning supercomputer. In *2014 47th Annual IEEE/ACM International Symposium on Microarchitecture* (pp. 609-622). IEEE.

29. Kim, D., Kung, J., Chai, S., Yalamanchili, S., & Mukhopadhyay, S. (2016). Neurocube: A programmable digital neuromorphic architecture with high-density 3D memory. *ACM SIGARCH Computer Architecture News, 44*(3), 380-392.

30. Seshadri, V., Lee, D., Mullins, T., Hassan, H., Boroumand, A., Kim, J., ... & Mowry, T. C. (2017, October). Ambit: In-memory accelerator for bulk bitwise operations using commodity DRAM technology. In *2017 50th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)* (pp. 273-287). IEEE.

31. Marinella, M. J., Agarwal, S., Hsia, A., Richter, I., Jacobs-Gedrim, R., Niroula, J., ... & James, C. D. (2018). Multiscale co-design analysis of energy, latency, area, and accuracy of a ReRAM analog neural training accelerator. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, *8*(1), 86-101.

32. Agarwal, S., Quach, T. T., Parekh, O., Hsia, A. H., DeBenedictis, E. P., James, C. D., ... & Aimone, J. B. (2016). Energy scaling advantages of resistive memory crossbar based computation and its application to sparse coding. *Frontiers in neuroscience*, *9*, 484.

33. Wong, H. S. P., Raoux, S., Kim, S., Liang, J., Reifenberg, J. P., Rajendran, B., ... & Goodson, K. E. (2010). Phase change memory. *Proceedings of the IEEE*, *98*(12), 2201-2227.

34. Waser, R., Dittmann, R., Staikov, G., & Szot, K. (2009). Redox‐based resistive switching memories–nanoionic mechanisms, prospects, and challenges. *Advanced materials*, *21*(25‐26), 2632-2663.

35. Yu, S., Li, Z., Chen, P. Y., Wu, H., Gao, B., Wang, D., ... & Qian, H. (2016, December). Binary neural network with 16 Mb RRAM macro chip for classification and online training. In *2016 IEEE International Electron Devices Meeting (IEDM)* (pp. 16-2). IEEE.

36. Yu, S., Gao, B., Fang, Z., Yu, H., Kang, J., & Wong, H. S. P. (2013). Stochastic learning in oxide binary synaptic device for neuromorphic computing. *Frontiers in neuroscience*, *7*, 186.

37. Suri, M., Bichler, O., Querlioz, D., Palma, G., Vianello, E., Vuillaume, D., ... & DeSalvo, B. (2012, December). CBRAM devices as binary synapses for low-power stochastic neuromorphic systems: auditory (cochlea) and visual (retina) cognitive processing applications. In *2012 International Electron Devices Meeting* (pp. 10-3). IEEE.

38. Yu, S. (2018). Neuro-inspired computing with emerging nonvolatile memorys. *Proceedings of the IEEE*, *106*(2), 260-285.

39. Yu, S. (2018). Neuro-inspired computing with emerging nonvolatile memorys. *Proceedings of the IEEE*, *106*(2), 260-285.

40. Burr, G. W., Shelby, R. M., Sidler, S., Di Nolfo, C., Jang, J., Boybat, I., ... & Hwang, H. (2015). Experimental demonstration and tolerancing of a large-scale neural network (165 000 synapses) using phase-change memory as the synaptic weight element. *IEEE Transactions on Electron Devices*, *62*(11), 3498-3507.

41. Deng, L. (2012). The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, *29*(6), 141-142.

42. Yu, S., Li, Z., Chen, P. Y., Wu, H., Gao, B., Wang, D., ... & Qian, H. (2016, December). Binary neural network with 16 Mb RRAM macro chip for

classification and online training. In *2016 IEEE International Electron Devices Meeting (IEDM)* (pp. 16-2). IEEE.

43. Competition, I. L. S. V. R. (2012). Available online: http://www. image-net. org/challenges. *LSVRC/(accessed on 27 December 2016)*.

44. Moon, K., Lim, S., Park, J., Sung, C., Oh, S., Woo, J., ... & Hwang, H. (2019). RRAM-based synapse devices for neuromorphic systems. *Faraday discussions*, *213*, 421-451.

45. Kozicki, Michael N., and Hugh J. Barnaby. "Conductive bridging random access memory—materials, devices and applications." *Semiconductor Science and Technology* 31.11 (2016): 113001

46. Kund, M., Beitel, G., Pinnow, C. U., Rohr, T., Schumann, J., Symanczyk, R., ... & Muller, G. (2005, December). Conductive bridging RAM (CBRAM): An emerging non-volatile memory technology scalable to sub 20nm. In *IEEE InternationalElectron Devices Meeting, 2005. IEDM Technical Digest.* (pp. 754-757). IEEE.

47. Gopalan, C., Ma, Y., Gallo, T., Wang, J., Runnion, E., Saenz, J., ... & Hollmer, S. (2011). Demonstration of conductive bridging random access memory (CBRAM) in logic CMOS process. *Solid-State Electronics*, *58*(1), 54-61.

48. Kozicki, M. N., Gopalan, C., Balakrishnan, M., Park, M., & Mitkova, M. (2004, November). Nonvolatile memory based on solid electrolytes. In *Proceedings. 2004 IEEE Computational Systems Bioinformatics Conference* (pp. 10-17). IEEE.

49. Kozicki, M. N., Balakrishnan, M., Gopalan, C., Ratnakumar, C., & Mitkova, M. (2005, November). Programmable metallization cell memory based on Ag-Ge-S and Cu-Ge-S solid electrolytes. In *Symposium Non-Volatile Memory Technology 2005.* (pp. 7-pp). IEEE.

50. Kamalanathan, D., Akhavan, A., & Kozicki, M. N. (2011). Low voltage cycling of programmable metallization cell memory devices. *Nanotechnology*, *22*(25), 254017.

51. Schindler, C., Thermadam, S. C. P., Waser, R., & Kozicki, M. N. (2007). Bipolar and unipolar resistive switching in Cu-Doped $\hbox {SiO} _ {2} $. *IEEE Transactions on Electron Devices*, *54*(10), 2762-2768.

52. Celano, U., Goux, L., Belmonte, A., Opsomer, K., Franquet, A., Schulze, A., ... & Vandervorst, W. (2014). Three-dimensional observation of the conductive

filament in nanoscaled resistive memory devices. *Nano letters*, *14*(5), 2401-2406.

53. Plimpton, S. J., Agarwal, S., Schiek, R., & Richter, I. (2016). *CrossSim* (No. CrossSim; 005389MLTPL00). Sandia National Lab.(SNL-NM), Albuquerque, NM (United States).

54. Chen, P. Y., Peng, X., & Yu, S. (2018). NeuroSim: A circuit-level macro model for benchmarking neuro-inspired architectures in online learning. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, *37*(12), 3067-3080.

55. . Mitkova, M., and M. N. Kozicki. "Silver incorporation in Ge–Se glasses used in programmable metallization cell devices." *Journal of non-crystalline solids* 299 (2002): 1023-1027.

56. Kolobov, A. V., and S. R. Elliott. "Photodoping of amorphous chalcogenides by metals." *Advances in Physics* 40.5 (1991): 625-684.

57. Burr, G. W., Shelby, R. M., Sebastian, A., Kim, S., Kim, S., Sidler, S., ... & Leblebici, Y. (2017). Neuromorphic computing using non-volatile memory. Advances in Physics: X, 2(1), 89-124.

58. S. Yu, X. Sun, X. Peng and S. Huang, "Compute-in-Memory with Emerging Nonvolatile-Memories: Challenges and Prospects," 2020 IEEE Custom Integrated Circuits Conference (CICC), 2020, pp. 1-4, doi: 10.1109/CICC48029.2020.9075887.

59. Zhang, Y., Wang, Z., Zhu, J., Yang, Y., Rao, M., Song, W., ... & Joshua Yang, J. (2020). Brain-inspired computing with memristors: Challenges in devices, circuits, and systems. Applied Physics Reviews, 7(1), 011308.

60. Chen, P. Y., Peng, X., & Yu, S. (2017, December). NeuroSim+: An integrated device-to-algorithm framework for benchmarking synaptic devices and array architectures. In 2017 IEEE International Electron Devices Meeting (IEDM) (pp. 6-1). IEEE.

61. Kozicki, Michael N., and Hugh J. Barnaby. "Conductive bridging random access memory—materials, devices and applications." *Semiconductor Science and Technology* 31.11 (2016): 113001.

62. Mahalanabis, D., et al. "Incremental resistance programming of programmable metallization cells for use as electronic synapses." *Solid-state electronics* 100 (2014): 39-44.

63. Chen, Wenhao, et al. "A CMOS-compatible electronic synapse device based on Cu/SiO2/W programmable metallization cells." *Nanotechnology* 27.25 (2016): 255202

64. Jacobs-Gedrim, Robin B., et al. "Impact of linearity and write noise of analog resistive memory devices in a neural algorithm accelerator." *2017 IEEE International Conference on Rebooting Computing (ICRC)*. IEEE, 2017.

65. Wu W et al 2018 A methodology to improve linearity of analog RRAM for neuromorphic computing 2018 IEEE Symp. on VLSI Technology (Piscataway, NJ: IEEE) pp 103–4

66. Pai-Yu Chen, S., Binbin Lin, I-Ting Wang, Tuo-Hung Hou, Jieping Ye, Vrudhula, Jae-Sun Seo, Yu Cao, and Shimeng Yu. "Mitigating Effects of Non-ideal Synaptic Device Characteristics for On-chip Learning." *2015 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)* (2015): 194-99. Web.

67. Agarwal, Sapan, et al. "Resistive memory device requirements for a neural algorithm accelerator." 2016 International Joint Conference on Neural Networks (IJCNN). IEEE, 2016.

68. Bache K and Lichman M 2016 UCI Machine Learning Repository: Data Sets. University of California at Irvine http://archive.ics.uci.edu/ml

69. LeCun Y, Cortes C and Burges C J 2016 The MNIST database of handwritten digits (http://yann.lecun.com/ exdb/mnist)

70. Fuller, Elliot J., et al. "Li-ion synaptic transistor for low power analog computing." *Advanced Materials* 29.4 (2017): 1604310.

71. Kadetotad, D., Chen, P. Y., Cao, Y., Yu, S., & Seo, J. S. (2017). Peripheral circuit design considerations of neuro-inspired architectures. In Neuro-inspired Computing Using Resistive Synaptic Devices (pp. 167-182). Springer, Cham.

72. Wang, Panni, and Shimeng Yu. "Ferroelectric devices and circuits for neuro-inspired computing." *MRS Communications* 10.4 (2020): 538-548.

73. Moon, Kibong, et al. "Improved conductance linearity and conductance ratio of 1T2R synapse device for neuromorphic systems." *IEEE Electron Device Letters* 38.8 (2017): 1023-1026.

74. Lee, Daeseok, et al. "Oxide based nanoscale analog synapse device for neural signal recognition system." *2015 IEEE International Electron Devices Meeting (IEDM)*. IEEE, 2015.

75. Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, pp. 436–444, May 2015.

76. B. Swaroop, W.C. West, G. Martinez, M. N. Kozicki and L. A. Akers, " Programmable current mode Hebbian learning neural network using programmable metallization cell," IEEE International Symposium on Circuits and Systems (ISCAS), vol. 3, pp. 33-36, 1998.

77. P. Apsangi, H. J. Barnaby, M. N. Kozicki, Y. Gonzalez-Velo and J. L. Taggart, "Effect of conductance linearity of Ag-chalcogenide CBRAM synaptic devices on the pattern recognition accuracy of an analog neural training accelerator.," Neuromorphic Computing and Engineering, vol. 2(2), p. 021002, 2022.

78. Y. Gonzalez-Velo, H. J. Barnaby, M. N. Kozicki, P. Dandamundi, A. Chandran, K.E. Holbert, M. Mitkova and M. Ailavajhala, "Total-ionizing-dose effects on the resistance switching characteristics of chalcogenide programmable metallization cells.," IEEE Transactions on Nuclear science, vol. 60(6), pp. 4563-4569, 2013.

79. W. Chen, H. J. Barnaby, M. N. Kozicki, A. H. Edwards, Y. Gonzalez-Velo, R. Fang, K. E. Holbert, S. Yu and W. Yu, "A Study of Gamma-Ray Exposure of Cu–SiO_2 Programmable Metallization Cells," IEEE Transactions on Nuclear Science 62, vol. no. 6, pp. 2404-2411, 2015.

80. Taggart, J. L. et al. (2018) In Situ Synaptic Programming of CBRAM in an Ionizing Radiation Environment. *IEEE transactions on nuclear science*. [Online] 65 (1), 192–199.

81. "CrossSim: Crossbar Simulator," Sandia National Laboratories, 2016. [Online]. Available: https://cross-sim.sandia.gov. [Accessed Dec 2021].

82. Y. Gonzalez-Velo, A. Patadia, H. J. Barnaby, and M. N. Kozicki, "Impact of radiation induced crystallization on programmable metallization cell electrical characteristics and reliability.," Faraday Discussions, vol. 213, pp. 53-66, 2019.

83. J. L. Taggart, R. Fang, Y. Gonzalez-Velo, H. J. Barnaby, M. N. Kozicki, J. L. Pacheco, E. S. Bielejec, E. S. McLain, N. Chamele, A. Mahmud and M. Mitkova, "Resistance state locking in CBRAM cells due to displacement damage effects," IEEE Transactions on Nuclear Science 64, vol. 8, pp. 2300-2306, 2017.

84. Mitkova, M., Wang, Y., & Boolchand, P. (1999). Dual chemical role of Ag as an additive in chalcogenide glasses. *Physical Review Letters*, *83*(19), 3848.

85. Miyatani, S. Y. (1981). Ionic conductivity in silver chalcogenides. *Journal of the Physical Society of Japan*, *50*(10), 3415-3418.

86. M. S. Ailavajhala, Y. Gonzalez-Velo, C. Poweleit, H. Barnaby, M. N. Kozicki, K. Holbert, D. P. Butt, and M. Mitkova, “Gamma radiation induced effects in floppy and rigid Ge-containing chalcogenide thin films,” J. Appl. Phys., vol. 115, no. 4, p. 043502, Jan. 2014

87. S.-y. Miyatani, “Ionic conductivity in silver chalcogenides,” J. Phys. Soc. of Japan, vol. 50, no. 10, pp. 3415–3418, Oct. 1981.

88. F. Kirchhoff, J. M. Holender, and M. J. Gillan, “Structure, dynamics and electronic structure of liquid Ag-Se alloys investigated by ab initio simulation,” Phys. Rev. B, vol. 54, no. 1, p. 14, Feb. 1996.

89. Gonzalez-Velo Y, Mahmud A, Chen W, Taggart J, Barnaby H, Kozicki M N, Ailavajhala M, Holbert K and Mitkova M 2016 Radiation hardening by process of CBRAM resistance switching cells IEEE Trans. Nucl. Sci. 63 2145–51

90. M. Ailavajhala, Y. Gonzalez-Velo, C. D. Poweleit, H. J. Barnaby, M. N. Kozicki, D. P. Butt and M. Mitkova, "New functionality of chalcogenide glasses for radiation sensing of nuclear wastes.," Journal of Hazardous Materials, vol. 269, pp. 68-73, 2014.