Effect of conductance linearity of Ag-chalcogenide CBRAM synaptic devices on the pattern recognition accuracy of an analog neural training accelerator

Priyanka Apsangi, Hugh Barnaby, Michael Kozicki, Yago Gonzalez-Velo and Jennifer Taggart
School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ 85287-5706, United States of America

E-mail: papsangi@asu.edu

Keywords: ANN, RRAM, DNN, analog, synapse

Abstract
Pattern recognition using deep neural networks (DNN) has been implemented using resistive RAM (RRAM) devices. To achieve high classification accuracy in pattern recognition with DNN systems, a linear, symmetric weight update as well as multi-level conductance (MLC) behavior of the analog synapse is required. 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, we analyze the conductance linearity response of these devices to different pulsing schemes. We have demonstrated 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. The effect of improved linearity was quantified by simulating the devices in an artificial neural network. The classification accuracy of two-layer neural network was seen to be improved from 85% to 92% for small digit MNIST dataset.

1. Introduction
In recent years, the advancement of semiconductor technologies coupled with the critical need for 'big' data processing has prompted renewed interest in artificial intelligence and neuromorphic computing [1]. Neuromorphic computers have shown significantly improved efficacy over traditional computing architectures in large scale visual, auditory as well as classification tasks [2]. A deep neural network (DNN) is one type of a neuromorphic computing system garnering substantial interest today, particularly for applications requiring data pattern recognition and machine learning. In order to achieve continuous improvements in accuracy, the depth and size of the deep neural network needs to significantly increase. However, the need for larger scaled neural networks poses a significant challenges in hardware implementation, particularly with respect to the traditional metrics of power, performance and density [3].

Many DNN algorithms like backpropagation and sparse coding execute iterative vector-matrix multiplications (VMMs). Various research groups have tried to implement these matrix operations with conventional analog/digital CMOS circuitry [4, 5]. However, many CMOS implementations are not well suited to meet targeted specifications for power consumption, latency, or functional area [6]. To overcome such limitations, hybrid DNN architectures that combine CMOS with resistive RAM (RRAM) devices have been proposed. RRAM devices have shown better scalability to sub-lithographic limits, multilevel conductance characteristics, low power operation, good data retention and endurance. RRAM devices can be fabricated in a crossbar architecture which helps in the implementation of high density VMM functions. For example the computing power efficacy of a crossbar RRAM array has been estimated to be 31 000 times the state-of-the-art microprocessor [3, 7].

RRAM arrays can be used as synaptic weight elements in DNN VMM. In the configuration that we are considering here, the output of one column in the vector-matrix operation is an output current ($I_j$), which is the sum of the products of row voltage ($V_i$) and the synaptic weight matrix elements ($G_{ij}$). Mathematically, ...
Figure 1. An ideal linear response of synaptic device.

Figure 2. Schematic representation of the CBRAM cell. (a) Filament formation on application of positive bias representing the LRS. (b) Filament dissolution on application of negative bias representing the high resistance state.

This is expressed as

\[ I_j = \sum V_i \times G_{ij}. \]  

(1)

This VMM operation is directly applicable to the inference carried out on a neural accelerator. The in situ training operation on a neural accelerator involves a more complex and energy intensive operations—VMM, matrix vector multiply (MVM) and outer product update [8].

The synaptic weight corresponds to the conductance across a two-terminal RRAM device. Conductance can be modified with an applied voltage pulse. The increase in conductance is called potentiation and a decrease in conductance is called depression. For training, the RRAM devices should have more analog conductance states for precise weight tuning, small variability in read and write operation and high endurance so that the system can be trained efficiently and accurately [9]. An efficient in situ learning neural network can be potentiated or depressed by a pulse, independent of its initial resistance state. Therefore, it is desired that the change in RRAM conductance be a highly linear and symmetric function of the number of voltage pulses applied, e.g., linear analog response as depicted in figure 1 [10]. Earlier works show that the dependence of conductance change to the input signal is directly related to the learning accuracy of a neural systems [11, 12].

Conductive bridge random access memory (CBRAM) is one type of RRAM device technology [13–16]. The memory states of the CBRAM are defined by the formation and modulation of a conductive metal filament as depicted in figure 2 [17]. The operation of these devices is determined by metal ion transport processes through an ion conducting electrolyte and chemical redox (reduction/oxidation) reactions. Upon the application of a positive bias, oxidation takes place at the CBRAM anode, where metal ions are formed. These ions transport across the electrolyte layer and are reduced at the cathode. As the redox reaction continues, a metallic conductive filament is formed across the electrolyte layer. Once the filament bridges the anode and the cathode, the devices can be modulated over a range of ‘low resistance states’ (LRSs) by increasing or decreasing the average cross-sectional area (width) of the filament; conductance being to first order proportional to filament width. Unfortunately, the incremental change in filament width is typically dependent on the conductance state [18], which introduces inherent non-linearities in the CBRAM response to pulse number. However, it is possible to control these non-linear mechanisms by exploiting the dependence of filament growth and dissolution on the amplitude and pulse width of the programming voltage [18].
3. Results and discussions

3.1. Constant pulse programming

The current–voltage ($I$–$V$) characteristic of one fabricated CBRAM synapse is shown figure 4. The figure shows the characteristic hysteresis DC switching behavior 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 [17]. The DC sweep was performed using Agilent 4156 C analyser. Bipolar switching behavior 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 5. The distributions show a minimum HRS to LRS ratio above 20, well above the target minimum of 10 [2].

To work as an analog synapse, CBRAM devices must demonstrate the capacity to switch incrementally into multiple LRSs. Previous works [19, 20] on Ag–Ge$_{30}$Se$_{70}$ CBRAM devices have demonstrated the multilevel conductance characteristics as a function of compliance current. 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 and the in-built 4225 PMU module. To achieve gradual switching behavior, we applied 100 consecutive SET pulses (0.35 V for 100 ns) and 50 consecutive RESET pulses ($-0.25$ V for 100 ns). Each write and erase pulse was followed by read voltage of 30 mV to extract conductivity. The change
Figure 4. $I-V$ characteristics of Ag–Ge$_{30}$Se$_{70}$ CBRAM device cycled at 100 $\mu$A compliance.

Figure 5. High resistance state (HRS) and low resistance state (LRS) distribution of the device extracted @0.03 V for 50 cycles.

Figure 6. The potentiation and depression cycling of Ag–Ge$_{30}$Se$_{70}$ CBRAM with P.W. = 100 ns.

in conductance of the device is plotted as a function of pulse number in figure 6 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 $10 \times$ change in conductivity for the 100 ns pulse width.

The effect of pulse width on the switching response is observed in figure 7. For these data, the SET/RESET pulses have a longer pulse width of 100 $\mu$s. 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 8 plots the response for both pulse widths over one cycles and shows that for shorter pulses, i.e., 100 ns, 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 [6].

The conductance update response for both pulse widths was compared by normalizing the conductance range as depicted in figure 9. The conductance update for the potentiation is nearly identical for both the pulse-widths. The depression behavior for 100 ns pulse width seems to be better. A detailed study of the linearity response is provided in the following subsections.
3.2. 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 10 shows the potentiation and depression characteristics of the device. Figure 11 depicts the response of the devices to multiple pulse cycles. The devices have shown to exhibit good linearity response over a suitable conductance range. This linearity response of the device is exploited by varying 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.

This problem can be resolved by using combination of resistors in series with RRAM device as suggested in some previous works [24, 25]. 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 12 [24]. However, for potentiation, i.e., increasing the CBRAM conductance (i.e., reducing the resistance), it will not be possible to use a fixed
Figure 10. Conductance response of Ag–Ge$_{30}$Se$_{70}$ CBRAM to incremental amplitude pulses.

Figure 11. Potentiation and depression cycling to increasing amplitude pulses.

Figure 12. Two resistor circuit to implement increasing amplitude effect.

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.

One proposed design is shown in figure 13. 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 three 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.

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:

$$V_{CBRAM,i} = R_{CBRAM,i}I_i \approx MR_{CBRAM,i} + B.$$  

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

\[
I_i \approx M + \frac{B}{R_i}.
\]  

(3)

Figure 14 shows the simulated response of \( R_{\text{CBRAM}} \) and \( V_{\text{CBRAM}} \) with the input voltage pulse sequence for \( V_{\text{IN}} \). It should be noted that using switch programming, the hardware of the potentiation circuit (figure 13) can be reconfigured for use in \( R_{\text{CBRAM}} \) depression.

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 [26]

\[
G_{\text{LTP}} = B \left( 1 - e^{-\frac{P}{A}} \right) + G_{\text{min}}
\]  

(4)

\[
G_{\text{LTD}} = B \left( 1 - e^{-\frac{P_{\text{max}}}{A}} \right) + G_{\text{max}}
\]  

(5)

\[
B = \frac{G_{\text{max}} - G_{\text{min}}}{1 - e^{-\frac{P_{\text{max}}}{A}}}
\]  

(6)

where, \( G \), \( P \), and \( A \) are the conductance value, pulse number, and nonlinear behavior of weight update, respectively. \( G_{\text{max}} \) and \( P_{\text{max}} \) 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 [27]. The non-linearities 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 15. 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 16.

The impact of linearity and weight update behavior on system response to different pulse schemes is studied by modeling these devices in an analog neural training accelerator. An artificial neural network was simulated using the device properties of Ag–Ge\textsubscript{30}Se\textsubscript{70} CBRAM. The CrossSim simulator was used to perform the supervised learning [28, 29]. A three-layer neural network depicted in figure 14 was trained using the 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 \((8 \times 8 \text{ pixels})\) of handwritten digits from the ‘optical recognition handwritten digits’ dataset [30] and a large image version \((28 \times 28 \text{ pixels})\) of handwritten digits from MNIST dataset [31]. A
two-layer neural network consisting of 64 input, 36 hidden and 10 output nodes for small image dataset. Similarly, an MNIST dataset was trained on a network of dimensions ($784 \times 300 \times 10$). 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 operation is 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 17).

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 pulse widths [32]. For small image dataset, we observe the training accuracy of 85% for constant amplitude pulsing as depicted in figure 18(a) while the training accuracy for the same dataset with increasing pulse amplitude was found to be 92% as seen in figure 18(b). For large digits the accuracy was seen to be improved from 79%
to 87% for increasing pulse amplitude as seen in figures 19(a) and (b). The ideal training accuracy using a double-precision CPU or GPU is 98%.

4. Conclusion

In conclusion, we have demonstrated the response of Ag–Ge30Se70 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 10× 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.

Acknowledgments

This work was funded in part by DTRA under Grant No. HDTRA1-17-1-0038. The authors would like to thank Jacob Calkins of DTRA for his support of this work and NanoFab facility at Arizona State University.

Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

ORCID iDs

Priyanka Apsangi https://orcid.org/0000-0002-9684-9329
References

[1] Waser R, Dittmann R, Menzel S and Noll T 2019 Introduction to new memory paradigms: memristive phenomena and neuromorphic applications Faraday Discuss. 213 11–27
[2] Yu S 2018 Neuro-inspired computing with emerging nonvolatile memories Proc. IEEE 106 260–85
[3] Moon K, Lim S, Park J, Sung C, Oh S, Woo J, Lee J and Hwang H 2019 RRAM-based synapse devices for neuromorphic systems Faraday Discuss. 213 421–51
[4] Ananthanarayanan R et al 2009 The cat is out of the bag: cortical simulations with 109 neurons, 1013 synapses Proc. Conf. on High Performance Computing Networking, Storage and Analysis
[5] Merolla P A et al 2014 A million spiking-neuron integrated circuit with a scalable communication network and interface Science 345 688–93
[6] Jacobs-Gedrim R B et al 2017 Impact of linearity and write noise of analog resistive memory devices in a neural algorithm accelerator 2017 IEEE Int. Conf. on Rebooting Computing (ICRC) (IEEE)
[7] Golmen T and Vlasov Y 2016 Acceleration of deep neural network training with resistive cross-point devices: design considerations Front. Neurosci. 10 333
[8] Jacobs-Gedrim R B et al 2018 Analog high resistance bilayer RRAM device for hardware acceleration of neuromorphic computation J. Appl. Phys. 124 202101
[9] Xiao T P, Bennett C H, Feinberg B, Agarwal S and Marinella M J 2020 Analog architectures for neural network acceleration based on non-volatile memory Appl. Phys. Rev. 7 033101
[10] Ielmini D 2018 Brain-inspired computing with resistive switching memory (RRAM): devices, synapses and neural networks Microelectron. Eng. 190 44–53
[11] Woo J and Yu S 2018 Resistive memory-based analog synapse: the pursuit for linear and symmetric weight update IEEE Nanotechnol. Mag. 12 36–44
[12] Woo J, Moon K, Song J, Kwak M, Park J and Hwang H 2016 Optimized programming scheme enabling linear potentiation in filamentary HfO2 RRAM synapse for neuromorphic systems IEEE Trans. Electron Devices 63 5064–7
[13] Kozicki M N, Park M and Mitkova. M 2005 Nanoscale memory elements based on solid-state electrolytes IEEE Trans. Nanotechnol. 4 331–8
[14] Kund M et al 2005 Conductive bridging RAM (CBRAM): an emerging non-volatile memory technology scalable to sub 20 nm IEEE Int. Electron Devices Meeting 2005. IEDM Technical Digest (IEEE)
[15] Waser R 2008 Electrochemical and thermochemical memories 2008 IEEE Int. Electron Devices Meeting (IEEE)
[16] Symanczyk R et al 2003 Electrical characterization of solid state ionic memory elements Proc. NVMTS Tech. Dig. pp 1–17
[17] Kozicki M N and Barnaby H J 2016 Conductive bridging random access memory—materials, devices and applications Semiconductor Sci. Technol. 31 113001
[18] Russo U, Kamalanathan D, Ielmini D, Lacaia A L and Kozicki M N 2009 Study of multilevel programming in programmable metallization cell (PMC) memory IEEE Trans. Electron Devices 56 1040–7
[19] Mahalanabis D, Barnaby H J, Gonzalez-Velo Y, Kozicki M N, Vrudhula S and Dandamudi P 2014 Incremental resistance programming of programmable metallization cells for use as electronic synapses Solid State Electron. 100 39–44
[20] Chen W, Fang R, Balaban M B, Yu W, Gonzalez-Velo Y, Barnaby H J and Kozicki M N 2016 A CMOS-compatible electronic synapse device based on Cu/SiO2/W programmable metallization cells Nanotechnology 27 255202
[21] Sandia National Laboratories: CrossSim: Crossbar Simulator 2019 https://cross-sim.sandia.gov (accessed July 2019)
[22] Mitkova M and Kozicki M N 2002 Silver incorporation in Ge–Se glasses used in programmable metallization cell (PMC) memory IEEE Trans. Electron Devices 63 5064–7
[23] Mahalanabis D, Barnaby H J, Gonzalez-Velo Y, Kozicki M N, Vrudhula S and Dandamudi P 2014 Incremental resistance programming of programmable metallization cells for use as electronic synapses Solid State Electron. 100 39–44
[24] Chen W, Fang R, Balaban M B, Yu W, Gonzalez-Velo Y, Barnaby H J and Kozicki M N 2016 A CMOS-compatible electronic synapse device based on Cu/SiO2/W programmable metallization cells Nanotechnology 27 255202
[25] Sandia National Laboratories: CrossSim: Crossbar Simulator 2019 https://cross-sim.sandia.gov (accessed July 2019)
[26] Mitkova M and Kozicki M N 2002 Silver incorporation in Ge–Se glasses used in programmable metallization cell devices J. Non-Cryst. Solids 299 1023–7
[27] Kolobov A V and Elliott S R 1991 Photodoping of amorphous chalcogenides by metals Adv. Phys. 40 625–84
[28] Moon K, Kwak M, Park J, Lee D and Hwang H 2017 Improved conductance linearity and conductance ratio of TiZr synapse device for neuromorphic systems IEEE Electron Device Lett. 38 1023–6
[29] Lee D et al 2015 Oxide based nanoscale analog synapse device for neural signal recognition system 2015 IEEE Int. Electron Devices Meeting (IEDM) (IEEE)
[30] 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
[31] Chen P.Y, Peng X and Yu S 2018 NeuroSim: a circuit-level macro model for benchmarking neuro-inspired architectures in online learning IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst. 37 3067–80
[32] Agarwal S et al 2016 Resistive memory device requirements for a neural algorithm accelerator 2016 Int. Joint Conf. on Neural Networks (IJCNN) (IEEE)
[33] Marinella M J, Agarwal S, Hsia A, Richter I, Jacobs-Gedrim R, Niroula J, Plimpton S J, Ipak E and James C D 2018 Multiscale co-design analysis of energy, latency, area, and accuracy of a ReRAM analog neural training accelerator IEEE J. Emerg. Sel. Topics Circuits Syst. 8 86–101
[34] Bache K and Lichman M 2016 UCI Machine Learning Repository: Data Sets (University of California at Irvine) http://archive.ics.uci.edu/ml
[35] LeCun Y, Cortes C and Burges C J 2016 The MNIST database of handwritten digits http://yann.lecun.com/exdb/mnist
[36] Fuller E J et al 2017 Li-ion synaptic transistor for low power analog computing Adv. Mater. 29 1604310