SAND2017-9637C

Impact of Linearity and Write Noise of Analog Resistive Memory Devices in a Neural Algorithm Accelerator

Robin B. Jacobs-Gedrim, Sapan Agarwal, Kathrine E. Knisely,
Jim E. Stevens, Michael S. van Heukelom, David R. Hughart,
John Niroula, Conrad D. James, Matthew J. Marinella
Sandia National Laboratories
Albuquerque, New Mexico, 87123
mmarine@sandia.gov

***Abstract*— Resistive memory (ReRAM) shows promise for use as an analog synapse element in energy-efficient neural network algorithm accelerators. A particularly important application is the training of neural networks, as this is the most computationally-intensive procedure in using a neural algorithm. However, training a network with analog ReRAM synapses can significantly reduce the accuracy at the algorithm level. In order to assess this degradation, analog properties of ReRAM devices were measured and hand-written digit recognition accuracy was modeled for the training using backpropagation. Bipolar filamentary devices utilizing three material systems were measured and compared: one oxygen vacancy system, Ta-TaO_x, and two conducting metallization systems, Cu-SiO_2, and Ag/chalcogenide. Analog properties and conductance ranges of the devices are optimized by measuring the response to varying voltage pulse characteristics. Key analog device properties which degrade the accuracy are update linearity and write noise. Write noise may improve as a function of device manufacturing maturity, but write nonlinearity appears relatively consistent among the different device material systems and is found to be the most significant factor affecting accuracy. This suggests that new materials and/or fundamentally different resistive switching mechanisms may be required to improve device linearity and achieve higher algorithm training accuracy.**

Keywords—RRAM, ReRAM, Neuromorphic, Analog, Nanosecond, Radio Frequency, Tantalum Oxide, Chalcogenide, Silicon Dioxide, Copper, Filamentary, Bipolar, MNIST, CrossSim

I. Introduction

A special purpose neural algorithm accelerator based on resistive memory weights shows the potential to achieve orders of magnitude gain in energy efficiency over advanced CMOS-ASIC accelerators and traditional CPU and GPU hardware [1-3]. Several large scale neuromorphic hardware accelerators have been implemented recently, including TrueNorth [4], SpiNNaker [5], Caviar [6], and FACETs [7]. Efforts to capitalize on the energy advantages of incorporating resistive or phase change memory into a fully parallel synaptic array or dot product engine are ongoing [8-10]. The energy scaling advantages of a fully parallel synaptic array would be achieved not only by storing data in resistive memory, but also from performing computation in resistive memory, rather than transferring between a cache and execution unit. For instance, a vector matrix multiply (VMM), one of the most computationally intensive operations in neuromorphic algorithms, can be electronically calculated on a crossbar in a single operation as shown in Figure 1.

Through Kirchkoff's laws, the applied voltages, V_i, are multiplied by the conductances $G_{i,j}$ to give currents $I_j = \sum_0^i G_{i,j}$. Computation on a crossbar array allows the entire vector matrix multiply to be completed in parallel, as compared to the many serial operations necessary that would be needed to access a standard SRAM cache array. In addition to a parallel read, the crossbar can be updated in parallel by the outer product of two vectors by simultaneously applying pulses to the rows and columns, accelerating all the key matrix operations required for a neural network algorithm[11].

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

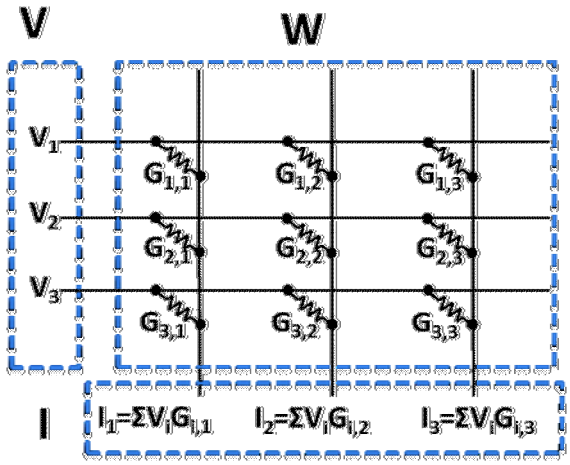


Figure 1: an electronic vector matrix multiply may be implemented on a crossbar given resistors with programmable weights. By applying bias to the left side of the crossbar, Kirkoff's laws give the sum of the weights in the column of the crossbar multiplied by the applied voltage amplitude.

The resistive memory device requirements for implementing an electronic vector matrix multiply on a crossbar depend on attributes including conductance range, write variability, write linearity, and read noise [12]. It is possible to linearize a device with a nonlinear response and reduce or eliminate write variability by applying pulses of varying amplitudes/pulse widths depending on the present device resistance [13-16]. This strategy necessitates addressing individual resistive memory elements, reading their conductance, and applying an update based on that resistance. Unfortunately, this eliminates the parallelism benefits of analog computation [11, 10] as each device is serially written. Consequently, we need devices with multiple resistance states accessible via symmetric voltage pulses of constant magnitude (blind updates). The programming schemes implemented in this work have been limited to approaches that are energy/time efficient on a very large scale neuromorphic architecture with crossbar arrays as large as 1000 x 1000 nodes[12].

Many different material systems and device architectures have been previously studied for resistive memory-based neuromorphic computing and binary memory applications [17-23]. Oxygen vacancy transport-based and metal conductive bridge-based devices have shown promise, andhere a particular device in each category is examined for their respective neural network synapse performance. In particular, we will examine a CMOS-compatible TaO_x-Ta device which operates under a vacancy modulated conductive oxygen filament mechanism [24], a Cu-SiO_2 heterostructure device where Cu filamentary diffusion into the SiO_2 layer modulates the conductivity of the device [25], and a commercially available chalcogenide device [26]. However, the different architectures, electrode designs, and device geometries limit the comparison to a discussion of the performance of a fully fabricated device, rather than material system. The neural network synapse performance of these devices is compared using our neural network algorithm simulator called CrossSim (http://cross-sim.sandia.gov).

II. Experimental Details

TaOx based devices were fabricated in Sandia's CMOS production MESAFab. The TiN-TaOx-Ta-TiN switching stack is deposited using reactive sputtering, using a feedback technique described in [27] . These devices consist of a TiN-TaOx-Ta-TiN stack in a radio frequency ground signal ground electrode and standard crossbar configurations. The reduced TaOx layer was 10 nm thick and the Ta layer was 50 nm thick, and the active region had 1.0 µm X 1.0 µm lateral dimensions.

All measurements were made on a Cascade Microtech manual probe station with an Agilent B1500A Semiconductor Parameter Analyzer mainframe equipped with a B1530 Waveform Generator Fast Measurement Unit. GGB Industries Picoprobe model 40A-GS-250 were employed for ground signal probe configurations. Pulse waveforms were captured on an Agilent CX3300 Device Current Waveform Analyzer.

Analog operation of ReRAM requires precise control of voltage pulse timing. Hence, we used short voltage pulses to characterize analog behavior. It should be noted that standard static current-voltage (I-V) sweeps on a Semiconductor Parameter Analyzer (SPA) are not controlled in time. The SPA spends an arbitrary amount of time sampling at each voltage level while attempting to reduce the noise of the current measurement. This measurement is designed to extract the DC parameters of a MOSFET, and is not sufficient to fully characterize the dynamic analog properties of ReRAM.

In order to conduct high speed pulsed operations which are

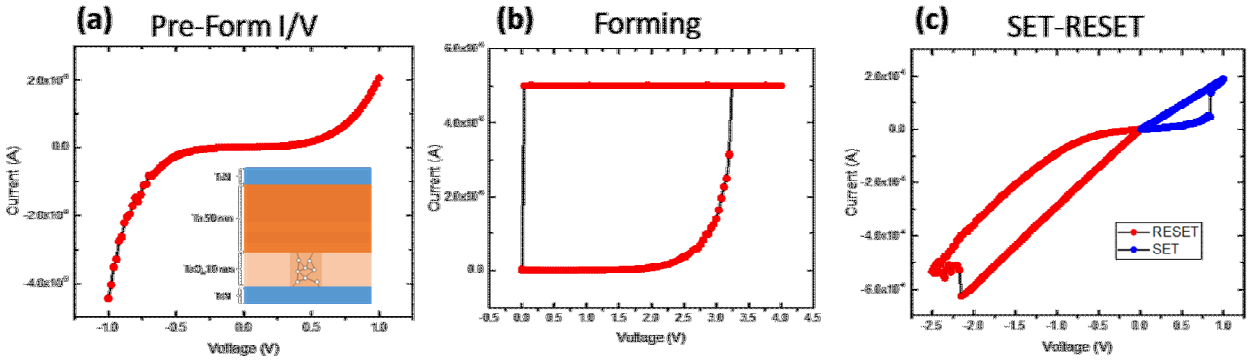


Figure 2: Standard I/V characterizations of TaO_x/Ta devices. (a) shows the pre-forming I/V characteristics of the TiN/TaO_x/Ta/TiN stack, which is nonohmic and asymmetric owing to the nonequal Shottkey barriers at the Ta and TiN interfaces. (b) shows the soft breakdown process known as electroforming when a bias of up to 4 volts is applied to the top electrode of the device. (c) The device can be RESET to a high resistance state after electroforming using negative bias, or SET to a low resistance state by applying positive bias.

equivalent to what will be implemented on a digital controller designed for a neural network algorithm accelerator, a probe station test setup was developed which can achieve 10 ns rise times during read and write. To achieve high frequency pulsed operation on a probe station, RF waveguide electrodes, and ground-signal type probes were employed to maintain pulse fidelity at the device. To measure the analog response of a resistive memory device, the following programming scheme was used. A short write pulse with nominal 10 ns rise/fall times and 10 ns pulse width, with an amplitude of +1 V (partial SET operation) or -1 V (partial RESET operation) is applied. Following each write pulse is a read pulse of 1 ms rise/fall times, 1 ms pulse width, and +100 mV amplitude. Current is read during the middle of this pulse and conductivity is extracted using Ohm's law. The total cycle time is therefore longer than 3 ms, ensuring that there is more than adequate time for thermal effects to dissipate between write pulses. The dissipation of heat from this system may be on the order of nanoseconds, however, so this extended cycle time could be significantly reduced if required.

III. Characterization of Analog Device Properties

The schematic of the TaO_x-Ta device used in this study is shown in the inset of Figure 2(a). A low voltage I-V sweep of the device indicates nonohmic electrical contacts due to the uneven Schottky barriers on the TiN and Ta sides of the TaO_x layer as shown in Figure 2(a). The conventional method for switching a bipolar resistive memory device starts with applying a positive bias sweep to the side of the device with the tantalum active electrode. This bias attracts negatively charged oxygen vacancies towards the Ta metal. At sufficiently high bias, the TaO_x layer goes through a soft breakdown process[28] known as forming, in which the resistance of the layer is reduced by many orders of magnitude as is shown in Figure 2(b) where the device was swept from 0-4 V with a 50 μA compliance. Following the forming step, a negative bias applied to the top electrode may raise the resistance of the TaO_x layer in a process known as RESET and a positive bias decreases the resistance of the layer in a process known as SET. The classic memristor pinched hysteresis loop is evident in the SET and RESET processes shown in Figure 2(c).

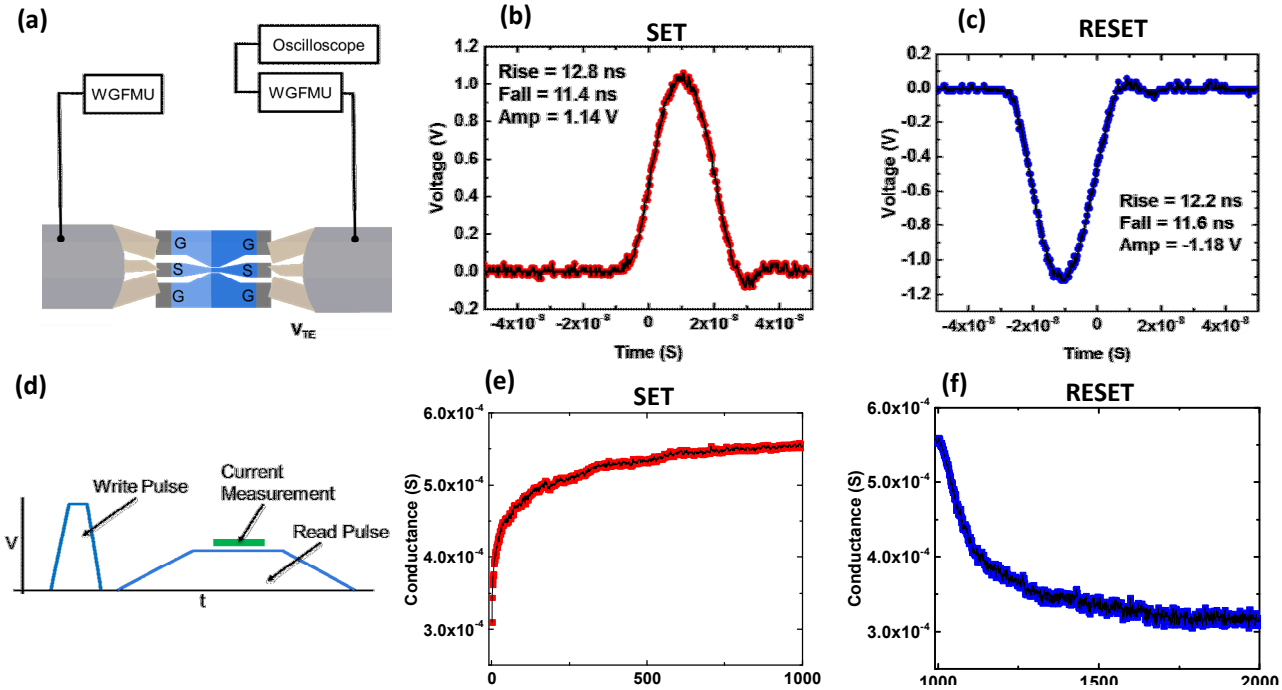


Figure 3: Fast pulsing setup and SET-RESET results. (a) The fast pulsing setup uses an Keysight B1530A module with Resistance Source Units (RSU)s to apply fast pulses and perform pulsed current voltage measurements. Using a ground signal configuration it is possible to achieve write pulses with pulse widths on the order of 10 ns. (b) shows the applied SET pulses, which are nominally 1 V amplitude, 10 ns pulse width and 10 nanosecond edge time. (c) shows the applied RESET pulses, which are nominally -1 V amplitude, 10 ns pulse width and 10 nanosecond edge time. (d) shows the analog device operation scheme, where a short, ±1 V amplitude write pulse is followed by a long + 100 mV read pulse. Current is measured during the read pulse, to observe the change in conductance caused by the write pulse. (e) Application of 1,000 positive SET pulses results in a gradual increase in device conductance. (f) Application of 1,000 negative RESET pulses results in a gradual increase in device conductance.

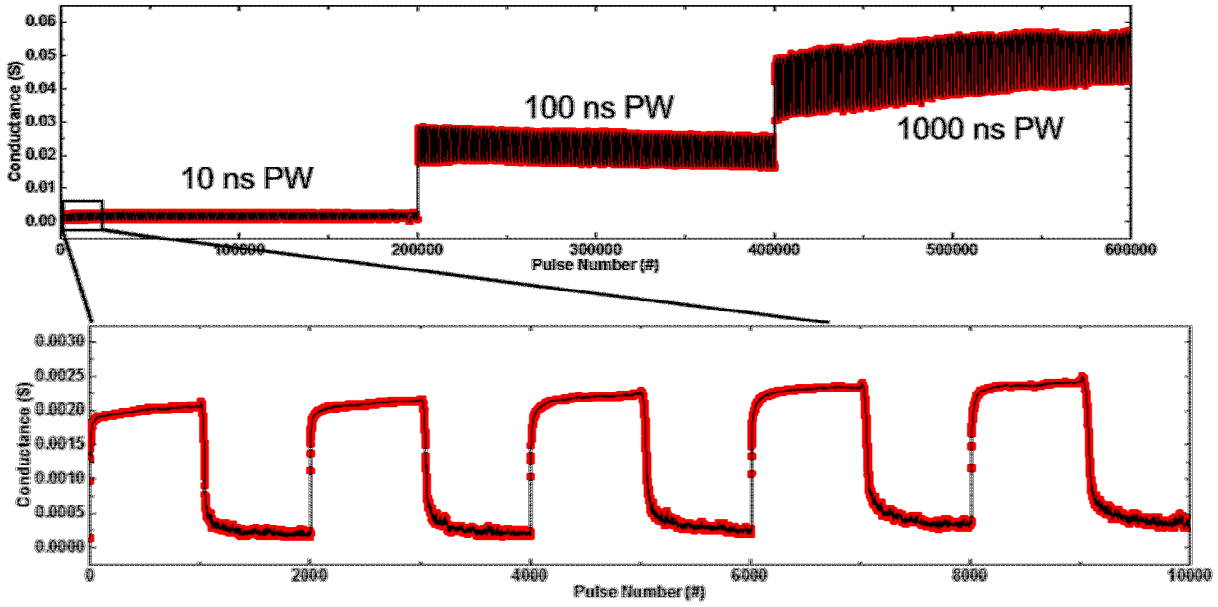


Figure 4: Repeated analog cycling o TaO /Ta device. Application of 100 cycles of 1000 SET+RESET pulses. First 200,000 pulses have 10 ns pulse width and edge time, pulse 200,00(400,000 have 100 ns pulse width and edge time, pulses 400,00(600,000 have 1 µs pulse width and edge time.

An analog programming scheme using nanosecond pulses on the TaO_x-Ta devices is shown in Figure 3(d). A ±1 V write pulse with pulsewidth in the nanosecond time scale is applied followed by a 100 mV read pulse. This cycle is repeated 1000 times for a full SET or RESET operation. The change in conductance on repeated pulsing is shown in Figure 3(e) and 3(f) for 10 ns pulses, indicating a gradual change of device conductance under repetition of SET and RESET write pulses, indicating many analog resistance states may be achieved in these devices. However, the response of the devices to repeated pulsing is nonlinear and asymmetric, a characteristic that will negatively impact neural algorithm performance accuracy. 3(b) and 3(c) show the measured write pulses, indicating a 140 mV overshoot for SET pulses and a -180 mV overshoot for RESET

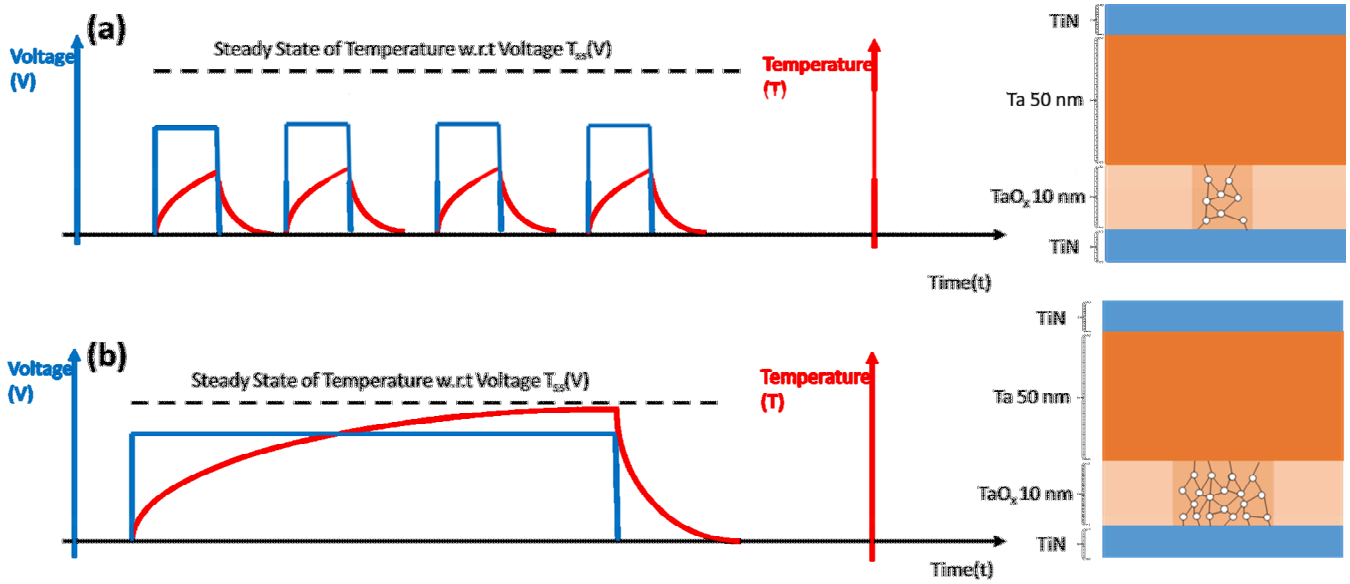


Figure 5: (a) Ultrafast pulses (10 ns or 100 ns) prevent TaO_x/Ta device temperature from fully reaching the steady state temperature. The lower temperature results in a smaller diameter filament, with less permanent vacancies. (b) Longer voltage pulses allow for the thermal buildup to reach a steady state, the higher temperature results in greater ion mobility allowing the filament radius to increase and more permanent vacancies to form, explaining the observed increase in conductivity in both the LRS and HRS state with longer pulses.

pulses, the measured rise times/fall times and amplitude are also shown.

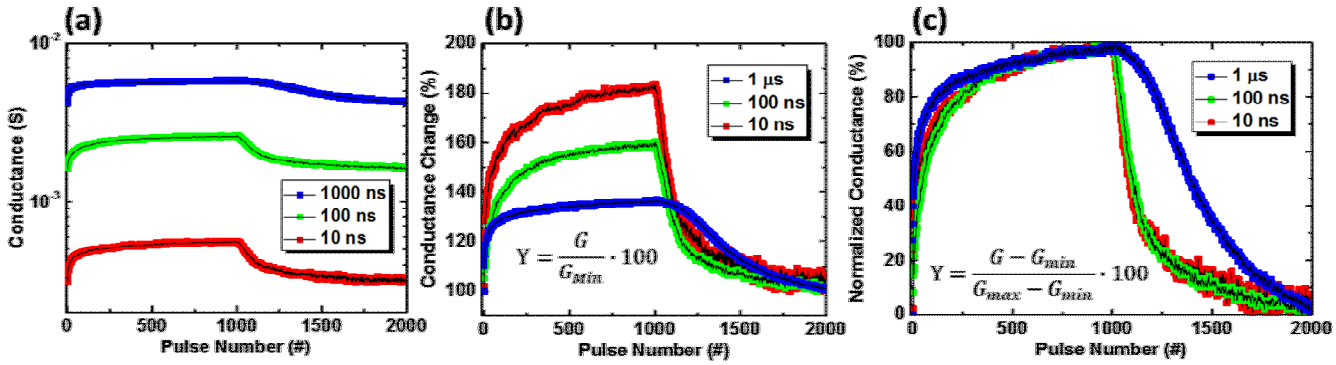


Figure 6: Effect of pulse width and edge time on analog TaO_x/Ta device performance. (a) SET-RESET analog operation of device using pulses with 10 ns, 100 ns and 1 us pulse widths and edge times. The conductance range is shown to increase with pulse width and edge time. (b) The conductance change in percent on the vertical axis of this graph is taken by diving the conductance values in the first plot by the minimum conductance value. The conductance change increases for shorter pulse widths and edge times. (c) Normalized conductance change in percentage showing an insignificant difference in linearity with changing pulse width and edge time.

Repeated cycling under varying pulse widths is shown in Figure 4. For the first 100 cycles (pulses 1-20,000) a 10 ns write pulse width was used, for the second 100 cycles (pulses 20,001-40,000) a write pulse width of 100 ns was used, and for the third 100 cycles (pulses 40,001-60,000) a 1 µs write pulse width was used. Low cycling is variability is observed, especially for 10 ns pulse widths, but some conductance drift is observed for 100 ns and 1 µs pulse widths. A large jump in conductance range is observed with increasing pulse widths.

This conductance range shift is attributed to a widening of the cylindrical filament, or change in the filament conductivity due to greater depletion of oxygen and the formation of permanent vacancy locations as shown in Figure 5[24, 29]. Interestingly, this increase in conductivity occurs during longer pulses, but not during a pulse train with an interval between each pulse with the same total energy as the longer pulse. Therefore, the conductivity change effect observed here cannot be a function of total electrical energy through the device. This effect can be a result of the device temperature not reaching a steady state during ultra-fast (10 ns or 100 ns) pulsing, and the duty cycle is sufficiently low as allow for thermal relaxation. Jeong et al. found the time to reach steady state temperature was ~500 ns for a similar device [30]. Since the pulse timings used in Figure 4 are below the thermal time constant, and the duty cycle is longer than the thermal time constant, the conductance effect may be ascribed to the effect shown in Figure 5.

To further examine the effects of changing pulse width, individual cycles were examined as shown in Figure 6. Figure 6(a) shows that the conductivity range increases with increasing pulse width. Lowering the conductance is important for reducing the required programming energy and enabling the parallel programming of large device arrays. Figure 6(b) shows that shorter pulses have a larger G_{on}/G_{off} ratio. To examine linearity as a function of pulse width, the conductance data from Figure 6(c) was normalized to have the same G_{max} and G_{min} values. This shows that the 10 and 100 ns pulses have nearly identical nonlinearities. The 1 µs pulses have a more linear reset write nonlinearity where the conductance drop is not as abrupt. This could be due to a change in the switching mechanism at high conductance from changing the length of a filament to changing the thickness or number of filaments [32].

IV. Modeling Baseline Tantulum Oxide ReRAM Cells in an Analog Neural Training Accelerator

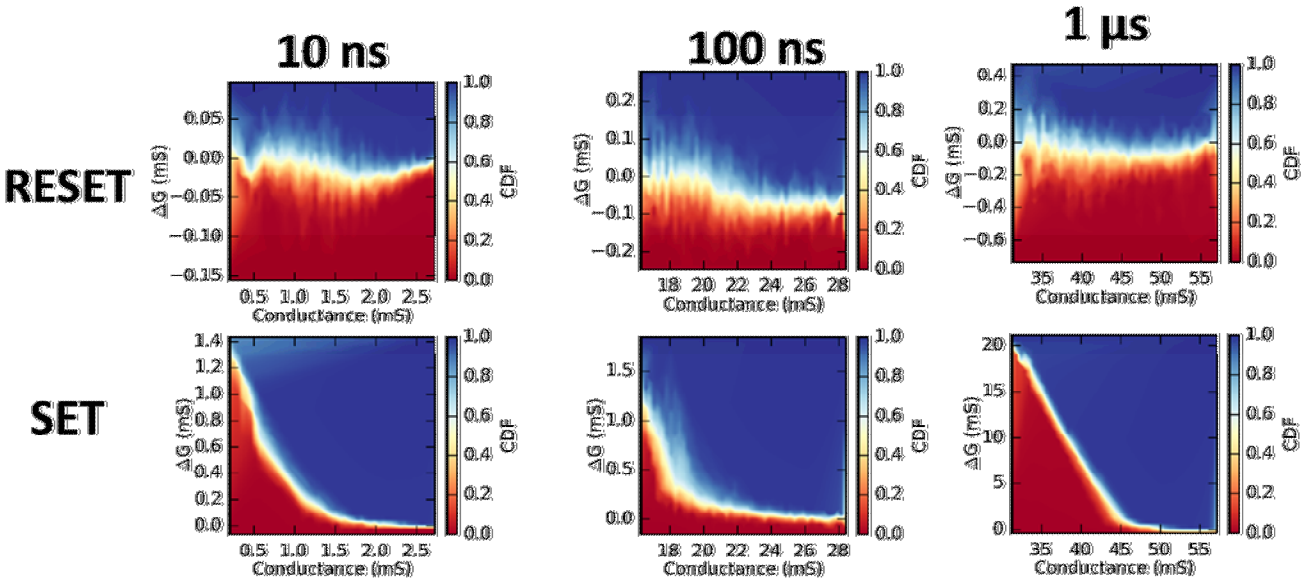
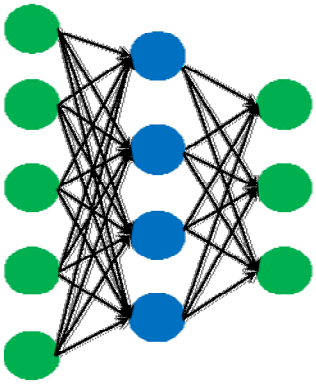


Figure 7: Cumulative distribution function (CDF) with repeated cycling on TaO_x/Ta device. Data from repeated cycling is collected in a cumulative distribution function lookup table for input into CrossSim. Color maps of CDF indicate given a current conductance state, the probability that the next update will increase conductance by a given value.

In this architecture, the ReRAM devices perform key kernels of common algorithms in parallel on an analog crossbar, as described above (Figure 1) [9]. While significant latency and energy efficiency gains are possible with the ReRAM accelerator, the analog operation has a lower accuracy than when executed with a standard double-precision floating point computing system. To quantify this accuracy trade-off, we have developed a CrossSim code suite simulate (http://cross-sim.sandia.gov) executing a neural training algorithm with imperfect devices [12, 33]. Specifically, an MNIST dataset was trained on a network with one hidden layer, of dimensions (784x300x10) [34]. This training set is composed of 60,000 28x28 pixel images of the handwritten single digit numbers "0" through "9". When using a standard double-precision CPU or GPU to train, accuracy of 98% is possible. In order to simulate the response of a device in a neuromorphic algorithm, we create of a lookup table which contains probabilistic change in conductance at a given conductance ($\Delta G/G$) was extracted from the cycling data for a given pulse width. This data is illustrated graphically in the cumulative distribution function (CDF) graphs in Figure 7. These plots are used to predict how a device will respond to an applied pulse as described in the supplementary information of Ref [35].



Data set	# Training Examples	# Test Examples	Network Size
MNIST Large Digits	60,000	10,000	784×300×10

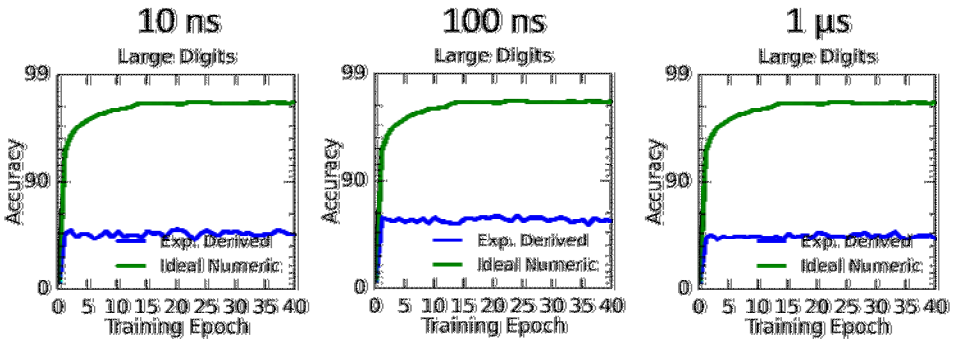


Figure 8: CrossSim Training results on the MNIST data set for the TaO_x/Ta device. The CDF lookup table are in input into a simulation of learning hardware training on the MNIST data set. The example network used has a single hidden layer and training as based on backpropagation of error algorithm. Training accuracies in the 70-80 % range were observed with 10 ns and 100 ns preforming slightly better than training using 1 µs pulses.

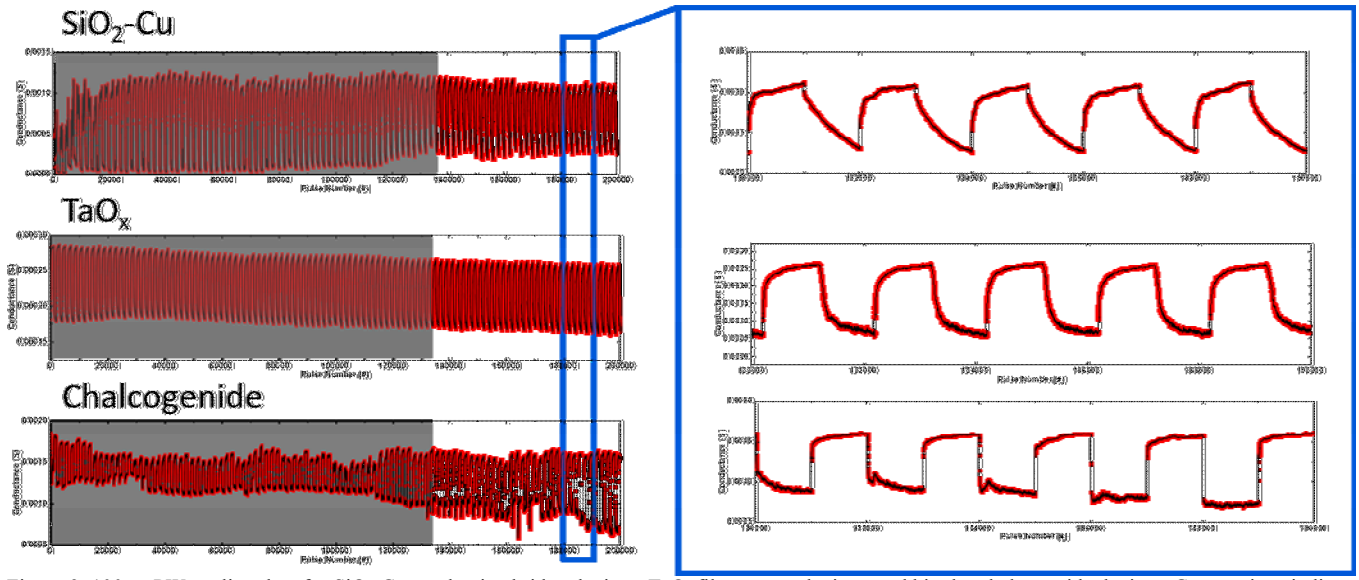


Figure 9. 100 ns PW cycling data for SiO_2-Cu conductive bridge devices, TaO_x filamentary devices, and bipolar chalcogenide devices. Gray regions indicate area trimmed out (cycles 1 to 66) to create a reduced dataset for CrossSim, full cycling is extracted from all 100 cycles. The graphs on the right are an expanded view of a few cycles.

The color indicates the probability that ΔG is less than or equal to the value on the y-axis for a given conductance on the x-axis. Figure 8 illustrates the results of training using the device data in CrossSim on the MNIST, UCI Small images, and File Types dataset. Results are given in Figure 8. Training accuracies in Figure 8 are taken from the highest accuracy during 40 training epochs where each epoch is one forward pass and one backward pass of all training examples.

V. COMPARISON BETWEEN MULTIPLE DEVICES

An overarching goal of this work is to determine the optimal ReRAM cells to use in a neural network algorithm accelerator. Hence, we have assessed both our baseline TaOx-Ta ReRAM devices against both SiO_2-Cu conducting bridge and chalcogenide resistive switching devices. Before discussing the difference between devices, it should be noted that the baseline TaOx-Ta devices are fabricated in Sandia's production fabrication facility, whereas the SiO_2-Cu and chalcogenide devices were fabricated in an academic research environment. Therefore, the differences we observe may attributed to manufacturing process optimization in addition to fundamental materials differences.

In Figure 9, the TaO_x-Ta cycling data is compared against bipolar chalcogenide devices and SiO_2-Cu conductive bridge devices. Each device material and structure tested was unique, and could not support 10 ns pulse widths. Hence, the amplitude of the 100 ns seconds pulses was selected to be the minimum for which a consistent conductance change in the correct conductance change direction was observed for SET and RESET. The minimum pulse amplitude was selected as higher amplitudes were found to negatively impact endurance, and the device must survive a minimum of 20,000 pulses to complete this test. For the chalcogenide device SET was +0.8 V and -0.8V for RESET. For the SiO_2-Cu device the SET voltage was +1.4 V and RESET was -1.6 V. For the TaO_x-Ta device the dataset was the same as used in the remainder of this work: SET voltage was +1 V and RESET was -1 V.

The Ag/chalcogenide and SiO_2-Cu CBRAM devices showed significant cycle-to-cycle variation in conductance ranges (Figure 9) which affected the accuracy considerably as illustrated in Figure 10. The device electrodes geometries are significantly different: the TaO_x-Ta devices have RF ground signal electrodes, the SiO_2-Cu devices are in a crossbar, and the Ag/chalcogenide devices are individually packaged devices. These electrode configurations can significantly change the shape and peak amplitude of the voltage pulse that each device receives, having a large effect on image recognition accuracy after training.

As shown in Figure 10(a) the TaO_x-Ta devices trained to the highest accuracy while the chalcogenide device obtained the lowest accuracy. In order to investigate the cause of this, we show the accuracy that results from using data from only the last 34 cycles in Figure 10(b). Using fewer cycles reduces the noise as the data is more uniform in a smaller range. Interestingly, this actually made the SiO_2-Cu and chalcogenide devices perform worse. This can be attributed to the average write nonlinearity being worse in the smaller cycle range.

Additional insight into the impact of write noise vs. write nonlinearity is gained by separately studying these effects. The CrossSim system allows us to deactivate each model individually. The effects on accuracy due to the individual elements for each device type are plotted in Figure 11. The "linearized" curves assume that updates are performed serially and are calibrated based on the starting conductance. In this case, change in conductance is constant with starting conductance. This ensures the average update is correct and only write noise is present. The "no noise" curves assume that

the updates are nonlinear but have no write noise added in the model.

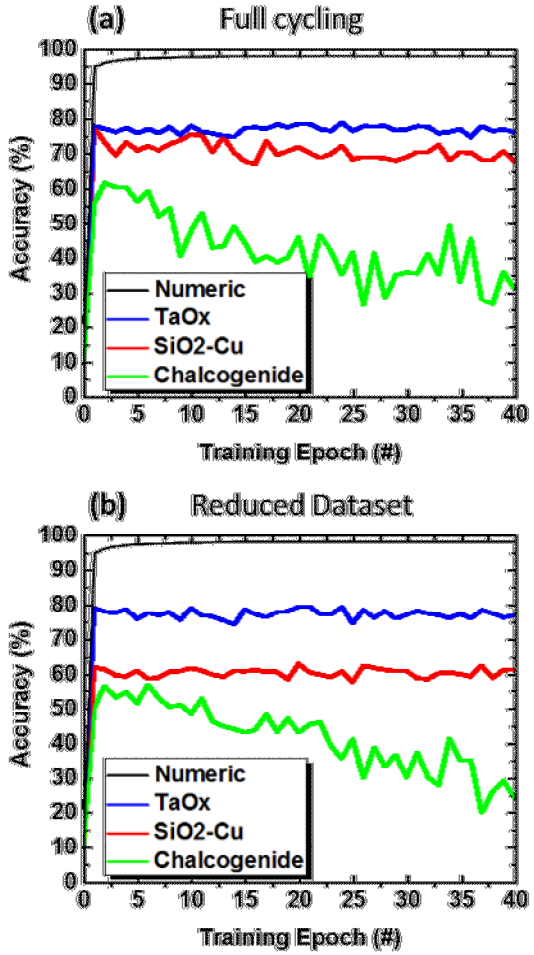


Figure 10: Comparison of recognition accuracy for SiO_2-Cu conductive bridge devices, TaO_x filamentary devices, and bipolar chalcogenide devices compared with ideal numeric accuracy. Devices trained and tested on the MNIST large images data set with 100 ns pulses. (a) Training accuracy for all three devices when the full cycling data is considered. (b)

In Figure 11, we see that for all the devices, eliminating the write nonlinearity greatly improves the accuracy, while eliminating the write noise has a smaller impact. In the case of the chalcogenide device, the write noise is great enough that eliminating it has an impact significantly improves the accuracy. However, the TaO_x-Ta devices have sufficiently low write noise that that nonlinearity dominates. As mentioned previously, the TaOx-Ta devices have the most mature manufacturing processes. Hence, if each device has a very mature manufacturing process, it is reasonable to include that nonlinearity would still limit each of these to accuracy levels similar to the maximum possible for the Ta-TaOx baseline devices. This is consistent with the findings of Burr et al for Phase Change RAM (PCRAM) devices that were fabricated commercially[9]. Hence, we are exploring devices which operate by other physical mechanisms (nonfilamentary switching) such as the lithium ion synaptic transistor for analog computation (LISTA) device[36].

VI. Conclusion

We have presented a method of measuring and modeling the accuracy for a network using analog ReRAM conductance as synaptic weight elements. Analog behavior of bipolar filamentary resistive switching devices based on three material systems was measured, and accuracy was modeled using CrossSim. The TaO_x-Ta ReRAM devices demonstrate the best algorithm accuracy after training compared to the Cu-SiO2 and Ag/chalcogenide and CBRAM. Further examination of the source of error shows that the accuracy degradation in the CBRAM devices is due to write noise. The TaO_x ReRAM cell has the lowest write stochasticity, which is likely due to the more mature manufacturing process. All three devices show significant write nonlinearity, which is inherent in a thermal-feedback based switching mechanism of ReRAM and CBRAM. This suggests that resistive switches based on a thermal feedback mechanism may have a fundamental accuracy limitation due to the highly nonlinear dependence of state change on current state. In the particular case of training the MNIST dataset with the backpropagation algorithm, the limit is consistently in the mid-eightieth percentile. This is consistent with the findings of Burr et al for phase change memory [9]. Hence, physical resistance change mechanisms not based on thermal-feedback (inherent in filamentary systems) may be required in order to achieve high training accuracies.

Acknowledgment

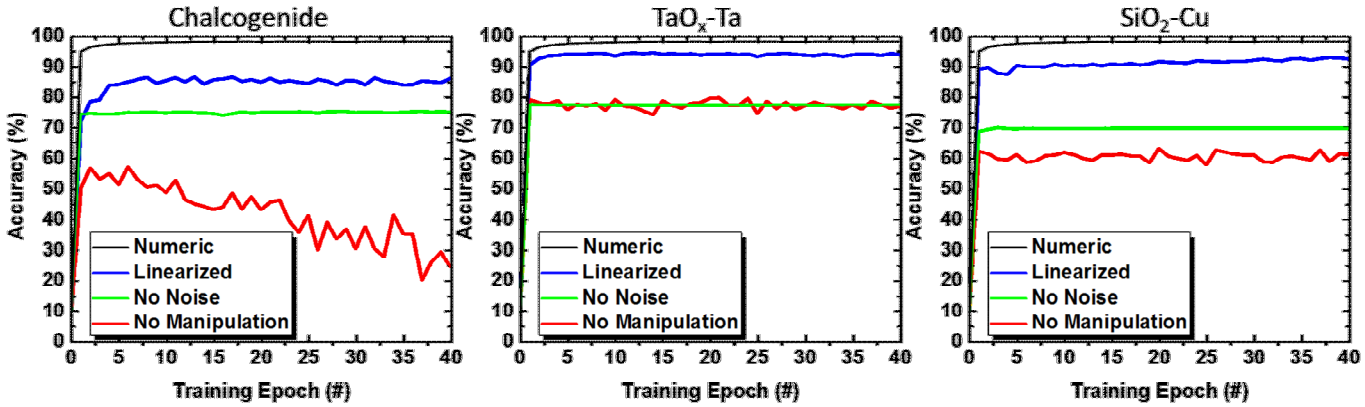


Figure 11: Examination of the effect of device linearity and noise on training accuracy for the three devices in this study. Data extracted from the reduced dataset (cycles 67 to 100) of the 100 ns dataset for each device. Training without data manipulation is shown in red. Training with a dataset where the effect of noise is removed is shown in green, and training with a dataset without the effect of write nonlinearity is shown in blue, the ideal numeric training accuracy is shown in black.

The authors gratefully acknowledge financial support from Sandia National Laboratories' Laboratory Directed Research and Development Program, and specifically the Hardware Acceleration of Adaptive Neural Algorithms (HAANA) Grand Challenge Project. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

REFERENCES

[1] Hasan R, Taha T M, Yakopcic C and Mountain D J 2016 High throughput neural network based embedded streaming multicore processors *2016 IEEE International Conference on Rebooting Computing (ICRC)* 1-8

[2] Kadetotad D, Xu Z, Mohanty A, Chen P-Y, Lin B, Ye J, Vrudhula S, Yu S, Cao Y and Seo J-s 2015 Parallel Architecture with resistive crosspoint array for dictionary learning acceleration *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* **5** 194-204

[3] Chi P, Li S, Xu C, Zhang T, Zhao J, Liu Y, Wang Y and Xie Y 2016 PRIME: a novel processing-in-memory architecture for neural network computation in ReRAM-based main memory *SIGARCH Comput. Archit. News* **44** 27-39

[4] Merolla P A, Arthur J V, Alvarez-Icaza R, Cassidy A S, Sawada J, Akopyan F, Jackson B L, Imam N, Guo C and Nakamura Y 2014 A million spiking-neuron integrated circuit with a scalable communication network and interface *Science* **345** 668-73

[5] Khan M M, Lester D R, Plana L A, Rast A, Jin X, Painkras E and Furber S B 2008 SpiNNaker: mapping neural networks onto a massively-parallel chip multiprocessor *2008 IEEE International Joint Conference on Neural Networks* 2849-56

[6] Serrano-Gotarredona R, Oster M, Lichtsteiner P, Linares-Barranco A, Paz-Vicente R, Gómez-Rodríguez F, Camuñas-Mesa L, Berner R, Rivas-Pérez M and Delbruck T 2009 CAVIAR: A 45k neuron, 5M synapse, 12G connects/s AER hardware sensory–processing–learning–actuating system for high-speed visual object recognition and tracking *IEEE Transactions on Neural Networks* **20** 1417-38

[7] Schemmel J, Briiderle D, Griibl A, Hock M, Meier K and Millner S 2010 A wafer-scale neuromorphic hardware system for large-scale neural modeling *Proceedings of 2010 IEEE International Symposium on Circuits and Systems* 1947-50

[8] Hu M, Strachan J P, Li Z, Grafals E M, Davila N, Graves C, Lam S, Ge N, Williams R S and Yang J 2016 Dot-product engine for neuromorphic computing: programming 1T1M crossbar to accelerate matrix-vector multiplication *Proceedings of DAC* **53**

[9] Burr G W, Shelby R M, Sidler S, Di Nolfo C, Jang J, Boybat I, Shenoy R S, Narayanan P, Virwani K and Giacometti E U 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** 3498-507

[10] Gao L, Wang I-T, Chen P-Y, Vrudhula S, Seo J-s, Cao Y, Hou T-H and Yu S 2015 Fully parallel write/read in resistive synaptic array for accelerating on-chip learning *Nanotechnology* **26** 455204

[11] Agarwal S, Quach T-T, Parekh O, Hsia A H, DeBenedictis E P, James C D, Marinella M J and Aimone J B 2015 Energy scaling advantages of resistive memory crossbar based computation and its application to sparse coding *Frontiers in neuroscience* **9** 484

[12] Agarwal S, Plimpton S J, Hughart D R, Hsia A H, Richter I, Cox J A, James C D and Marinella M J 2016 Resistive memory device requirements for a neural algorithm accelerator *2016 International Joint Conference on Neural Networks (IJCNN)* 929-38

[13] Dodge R K, Ottogalli F, Buda E and Ferraro M 2006 Phase change memory bits reset through a series of pulses of increasing amplitude. U.S. Patent No. 7,099,180.

[14] Park S, Sheri A, Kim J, Noh J, Jang J, Jeon M, Lee B, Lee B, Lee B and Hwang H 2013 Neuromorphic speech systems using advanced ReRAM-based synapse *IEDM Tech Dig* **25** 1-25.6

[15] Alibart F, Gao L, Hoskins B D and Strukov D B 2012 High precision tuning of state for memristive devices by adaptable variation-tolerant algorithm *Nanotechnology* **23** 075201

[16] Yu S, Wu Y, Jeyasingh R, Kuzum D and Wong H-S P 2011 An electronic synapse device based on metal oxide resistive switching memory for neuromorphic computation *IEEE Transactions on Electron Devices* **58** 2729-37

[17] Goux L, Fantini A, Kar G, Chen Y-Y, Jossart N, Degraeve R, Clima S, Govoreanu B, Lorenzo G and Pourtois G 2012 Ultralow sub-500 nA operating current high-performance TiN Al_2O_3 HfO_2 Hf TiN bipolar RRAM achieved through understanding-based stack-engineering: IEEE 159-60

[18] Sidler S, Boybat I, Shelby R M, Narayanan P, Jang J, Fumarola A, Moon K, Leblebici Y, Hwang H and Burr G W 2016 Large-scale neural networks implemented with Non-Volatile Memory as the synaptic weight element: Impact of conductance response *2016 46th European Solid-State Device Research Conference (ESSDERC)* 440-3

[19] Jha R 2013 *Analog and Digital Switching Characteristics of Transition Metal Oxide Based Resistive Random Access Memory (ReRAM) Devices*

[20] Jackson B L, Rajendran B, Corrado G S, Breitwisch M, Burr G W, Cheek R, Gopalakrishnan K, Raoux S, Rettner C T and Padilla A 2013 Nanoscale electronic synapses using phase change devices *ACM Journal on Emerging Technologies in Computing Systems (JETC)* **9** 12

[21] Jo S H, Chang T, Ebong I, Bhadviya B B, Mazumder P and Lu W 2010 Nanoscale memristor device as synapse in neuromorphic systems *Nano letters* **10** 1297-301

[22] Wang Z, Yin M, Zhang T, Cai Y, Wang Y, Yang Y and Huang R 2016 Engineering incremental resistive switching in TaO_x based memristors for brain-inspired computing *Nanoscale* **8** 14015-22

[23] Prezioso M, Merrikh-Bayat F, Hoskins B, Adam G, Likharev K K and Strukov D B 2015 Training and operation of an integrated neuromorphic network based on metal-oxide memristors *Nature* **521** 61-4

[24] Mickel P R, Lohn A J, Choi B J, Yang J J, Zhang M-X, Marinella M J, James C D and Williams R S 2013 A physical model of switching dynamics in tantalum oxide memristive devices *Applied Physics Letters* **102** 223502

[25] Wenhao C, Runchen F, Mehmet B B, Weijie Y, Yago G-V, Hugh J B and Michael N K 2016 A CMOS-compatible electronic synapse device based on $Cu/SiO_2/W$ programmable metallization cells *Nanotechnology* **27** 255202

[26] Oblea A S, Timilsina A, Moore D and Campbell K A 2010 Silver chalcogenide based memristor devices *The 2010 International Joint Conference on Neural Networks (IJCNN)* 1-3

[27] Stevens J E, Lohn A J, Decker S A, Doyle B L, Mickel P R and Marinella M J 2014 Reactive sputtering of substoichiometric Ta_2O_x for resistive memory applications *Journal of Vacuum Science & Technology A: Vacuum, Surfaces, and Films* **32** 021501

[28] O'Dwyer J J 1973 The theory of electrical conduction and breakdown in solid dielectrics *Oxford*: Clarendon Press

[29] Mickel P R, Lohn A J, James C D and Marinella M J 2014 Isothermal switching and detailed filament evolution in memristive systems *Advanced Materials* **26** 4486-90

[30] Jeong Y, Kim S and Lu W D 2015 Utilizing multiple state variables to improve the dynamic range of analog switching in a memristor *Applied Physics Letters* **107** 173105

[31] Lohn A J, Mickel P R and Marinella M J 2014 Modeling of filamentary resistive memory by concentric cylinders with variable conductivity *Applied Physics Letters* **105** 183511

[32] Agarwal S 2017 CrossSim *http://cross-sim.sandia.gov*

[33] LeCun Y, Cortes C and Burges C J 1998 The MNIST database of handwritten digits.

[34] Fuller E J, Gabaly F E, Léonard F, Agarwal S, Plimpton S J, Jacobs-Gedrim R B, James C D, Marinella M J and Talin A A 2017 Li-Ion Synaptic Transistor for Low Power Analog Computing *Advanced Materials* **29** 1604310-n/a

[35] Fuller E J, Gabaly F E, Léonard F, Agarwal S, Plimpton S J, Jacobs-Gedrim R B, James C D, Marinella M J and Talin A A 2017 Li-Ion Synaptic Transistor for Low Power Analog Computing *Advanced Materials* **29** 1604310