AFRL-RI-RS-TR-2019-153

DEVELOPMENT OF A MEMRISTIVE DYNAMIC ADAPTIVE NEURAL NETWORK ARRAY (MRDANNA)

SUNY POLYTECHNIC COLLEGE OF NANOSCALE SCIENCE & ENGINEERING

JULY 2019

FINAL TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

STINFO COPY

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE

■ **AIR FORCE MATERIEL COMMAND** ■ **UNITED STATES AIR FORCE** ■ **ROME, NY 13441**

NOTICE AND SIGNATURE PAGE

This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09. This report is available to the general public, including foreign nations. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RI-RS-TR-2019-153 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE CHIEF ENGINEER:

/ S /	**/ S /**
JOSEPH E. VAN NOSTRAND Work Unit Manager	QING WU Technical Advisor, Computing & Communications Division Information Directorate

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
JULY 2019	FINAL TECHNICAL REPORT	JAN 2016 – JAN 2019

4. TITLE AND SUBTITLE

DEVELOPMENT OF A MEMRISTIVE DYNAMIC ADAPTIVE NEURAL NETWORK ARRAY (MRDANNA)

5a. CONTRACT NUMBER
N/A

5b. GRANT NUMBER
FA8750-16-1-0063

5c. PROGRAM ELEMENT NUMBER
62788F

6. AUTHOR(S)

Nathaniel C. Cady

5d. PROJECT NUMBER
T2NR

5e. TASK NUMBER
NF

5f. WORK UNIT NUMBER
AB

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
SUNY Polytechnic Institute college of Nanoscale Science & Engineering
257 Fuller Road
Albany, NY 12203

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory/RITB
525 Brooks Road
Rome NY 13441-4505

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/RI

11. SPONSOR/MONITOR'S REPORT NUMBER
AFRL-RI-RS-TR-2019-153

12. DISTRIBUTION AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited. This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09

13. SUPPLEMENTARY NOTES

14. ABSTRACT
The objective of this effort was to build affordable, manufacturable, low power, dynamic neuromorphic computing platform for handling spiky, highly variable information/data, as well as develop a low-power, hybrid memristor/CMOS "neuron / synapse" implementation for implementation of a hardware-based dynamic neural network array. The approach was to leverage hybrid CMOS/memristor encryption project to design, fabricate, and test memristor/CMOS "neurons / synapses," and integrate them with existing FPGA implementations for full-scale dynamic neural network array demonstration. In addition, to make a module design kit library available for other circuits and architectures, thus making advancements in this effort useful to future designs that utilize hybrid CMOS/memristor technologies.

15. SUBJECT TERMS

Memristor, FPGA, CMOS, nanoelectronics

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			**JOSEPH E. VAN NOSTRAND**
U	U	U	UU	44	19b. TELEPHONE NUMBER (*Include area code*) **315-330-4920**

Table of Contents

List of Figures / List of Tables

1.0 Summary

This technical report summarizes the R&D efforts for the AFRL project "Fabrication of a memristive dynamic neural network array," locally referred to as "mrDANNA". This research project aimed to enable future generations of computing systems by (1) leveraging an emerging, power-efficient device technology (i.e. the memristor, aka ReRAM) and (2) considering an alternative architectural model (i.e. neuromorphic) that promises to overcome many of the performance limitations of conventional von Neumann systems. The specific neuromorphic architecture on which the proposed mrDANNA is based is the Neuroscience-Inspired Dynamic Architecture (NIDA), developed by researchers at the University of Tennessee, Knoxville (UTK) as an approach to applying neuromorphic principles to a wide variety of applications. Key features of the NIDA architecture include: (1) a spiky representation of data, (2) the ability for the system to adapt during run-time, and (3) a synaptic representation including delay distance as well as weight information. The structure and simplicity of the NIDA architectural model has recently been leveraged in the development of a Dynamic Adaptive Neural Network Array (DANNA), an efficient digital system constructed from a basic element that can be configured to represent either a neuron or a synapse.

For this project we leveraged features of the NIDA architecture in the construction of a mixed-mode, analog/digital memristor-based DANNA (mrDANNA) system. In conjunction with our partners at UTK, we successfully completed the design, tape-out, and fabrication of hybrid CMOS/memristor chips in SUNY Polytechnic Institute's 300mm wafer-scale nanofabrication facilities. In addition to the NIDA-inspired neuromorphic test circuits that were implemented on this platform, the SUNY Poly team also included a reservoir-based Cellular Memristive Output Reservoir (CMOR) test circuit, as well as a variety of one transistor/one memristor (1T1R) cells, and 1T1R arrays (up to 512 x 512 cells in size). The ultimate results of this effort were 1) successful demonstration of integrated CMOS/memristor fabrication and performance testing, 2) implementation of NIDA-inspired mrDANNA test circuits and delivery of test chips to our partners at UTK, 3) implementation and testing of CMOR circuits, and 4) a study on the effects of stochastic memristor behavior on model neuromorphic circuits.

2.0 Introduction

The primary component of the NIDA architecture and corresponding mrDANNA structure is a single programmable element that can be configured as a neuron, synapse or pass-thru function, and be replicated across an "n x m" grid with nearest neighbor or "small world" connectivity. To maximize performance and minimize size and power consumption very simple state-machines and logic circuits are used, avoiding the use of digital signal processors, floating-point units, arithmetic-logic units, memory arrays and other common microprocessor units.

The neuron's function is to accumulate "charge" by adding the weighted inputs to an existing charge level until that level reaches a programmable threshold. Upon reaching this threshold the neuron fires its outputs and resets its charge to a predetermined bias level. A weighted-sum threshold activation function is the primary candidate for the neuron design due to its simplicity and functionality. However, other activation functions can be implemented (such as linear, sigmoid or Gaussian).

The synapse function captures selected features of both axons and synapses found in biological neural networks. Specifically, a NIDA/DANNA synapse contains both the associative delay distance between two neurons and the weight (or strength) of the synaptic connection. Two unique characteristics of the synapse model are: 1) the weight value held by the synapse can automatically potentiate (long-term potentiation, LTP) or depress (long-term depression, LTD) depending on the firing condition of its output neuron and 2) a string of firing events can be stored in a "distance FIFO" to simulate a synapse transmitting a set of firing events down its length (thus constituting temporal storage of events).

An evolutionary optimization (EO) environment has been developed at UTK to configure the neural networks in a DANNA. The EO trains over parameters of the network (weights and delay distances on synapses and thresholds on neurons) as well as the structure (the number and placement of neurons and synapses) and the dynamics of the network. The dynamics of the network are directly embedded in the structure itself (delays in the synapse and charges in the neuron). Most other artificial neural network implementations have a predefined, fixed structure rather than one determined by a dynamic optimization method as in our approach. The EO requires the user to specify the inputs and outputs to the network, and a fitness function that rates how well a network's outputs fare with the given inputs. The EO then generates an initial population of random networks, and gradually evolves the population using mutation and crossover operations, until it generates a network that exceeds a predefined threshold for fitness. These operations occur on the direct representation of the network.

An initial prototype, including global functions and the programming interfaces, was implemented and tested using two different Xilinx FPGAs: the XC7VX690T and the XC7VX485T. A square grid of approximately 2500 elements was placed on a 485T FPGA and approximately 5000 elements on a 690T FPGA. With these sizes of neural network arrays, applications such as handwritten character recognition, image pattern recognition, network traffic analysis and small systems controls automation are supported. The proposed DANNA implementation constructed with nanoscale memristors (this project) extends the number of elements placed on one chip and thus also extends the application possibilities. Specifically, expected improvements in area utilization as compared to an all CMOS DANNA implementation leads to 10x more neuron/synapse elements that can be implemented, in the same area. This increase in the number of available elements in the system translates into an increase in the complexity of applications that can be implemented with a single mrDANNA chip (***Figure 1***).

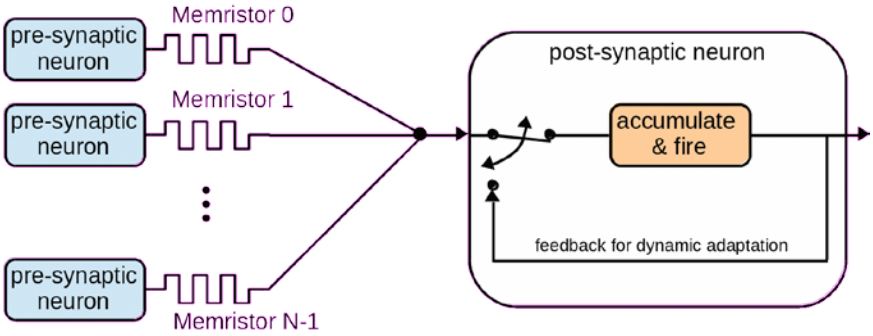

Figure 1. Simplified view of two-stage mrDANNA including several pre-synaptic neurons driving a single post-synaptic neuron through memristor based synapses. The local feedback loop in the neuron enables dynamic adaptation where memristor weights are updated during run time.

To implement the proposed mrDANNA test chip, this project used the hybrid CMOS/Memristor process developed at CNSE under a previous AFRL effort. The process integrates metal-oxide memristors in the metal layers of the 65 nm 10LPe CMOS process from IBM, leading to a seamless CMOS/memristor integration process. The seamless integration of CMOS with memristive technology is a unique feature as compared to related efforts where memristive devices are integrated post-fabrication on an existing CMOS chip. ***Figure 2*** shows the cross-section of a circuit implemented in the CNSE CMOS/Memristor process.

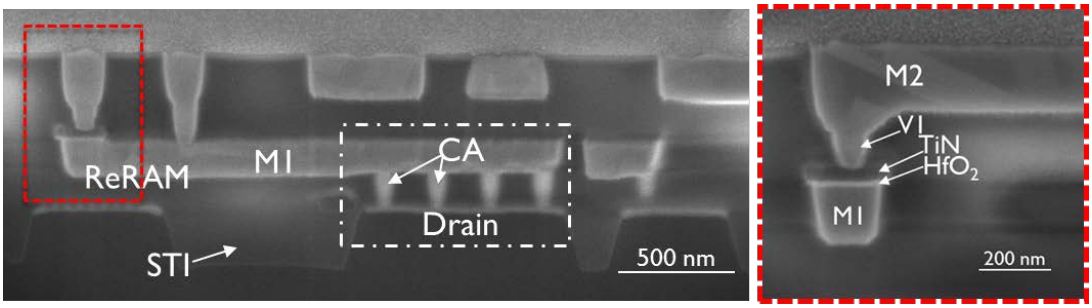

Figure 2. Cross-section of a vertically integrated transistor / memristor (1T1R) device fabricated at CNSE. The bottom electrode of the memristor (ReRAM) device is connected to the drain contact stud (CA) via the M1 line. The figure at right is a cross-section of the memristor element alone, showing the bottom electrode (M1), hafnium oxide dielectric layer, TiN top electrode, and M2 contact.

Under this project, and in collaboration with a team at UTK, we designed and fabricated a CMOS/Memristor mrDANNA evaluation chip to demonstrate the memristor-enabled neurons that can eventually be integrated into a larger, fully functional neuromorphic chip (in a follow-on effort). Each mrDANNA "neuron" was designed to be compatible with rapid configuration, dynamic adaptation, low power operation, and ultimately for use in processing spatio-temporal data in a spiking data format. Memristor-based synapses were designed to enable an analog representation of weighted inputs that can be efficiently summed and accumulated using analog circuitry. The analog operation of key synapse and neuron functions leads to a natural speed-up as compared to a fully digital approach.

3.0 Methods, Assumptions and Procedures

3.1 Fabrication

In this effort, memristor (aka: ReRAM) devices and CMOS were built in-house on 300 mm wafers at the SUNY Poly-technic Institute's Center for Semiconductor Research (CSR). The ReRAM devices were fabricated using a 300mm wafer platform based on the IBM 65nm 10LPe process technology. A custom hybrid CMOS/ReRAM process was developed to allow for a seamless integration of both CMOS and ReRAM devices into one process flow with minimal added costs. ReRAM devices were integrated between metal 1 (M1) and metal 2 (M2); specifically, the intervening via 1 (V1) layer was split to encapsulate the ReRAM device. For the purpose of using a front-end-of-the-line (FEOL) deposition tool for the HfO_2 switching layer (SL) of the ReRAM device, custom CA, M1 and V1 layers were developed using W as the interconnect material. With this approach the resulting W V1 surface can be used as a bottom electrode (BE) for the subsequent ReRAM device stack. As mentioned, HfO_2 is used as the SL with a thickness of 5.8 nm and deposited via atomic layer deposition (ALD). The SL is covered by a 6 nm Ti oxygen scavenger layer (OSL) to yield sub-stoichiometric HfO_x with a gradient of oxygen vacancies from the BE towards the OSL. Due to the rapid oxidation of Ti, a 40 nm TiN film is used to encapsulate the Ti OSL, and serves as the top electrode (TE). Both films are deposited via physical vapor deposition (PVD). After this process is complete, the ReRAM device stack is structured via a reactive ion etch (RIE) process and pads with sizes of 200x200 nm^2 are created on top of 100x100 nm^2 W-V1 BE studs. This creates devices without any edges between the HfO_2 and the surrounding Si_3N_4 that are exposed to the switching dielectric. This fabrication approach is predicted to result in a filament formation within the center of the ReRAM device pad. The devices are connected to a NFET which serves as the on-chip current control during the forming and set process. A transmission electron micrograph of a cross-section of the resulting one transistor / one ReRAM (1T1R) structure can be seen in ***Figure 3***.

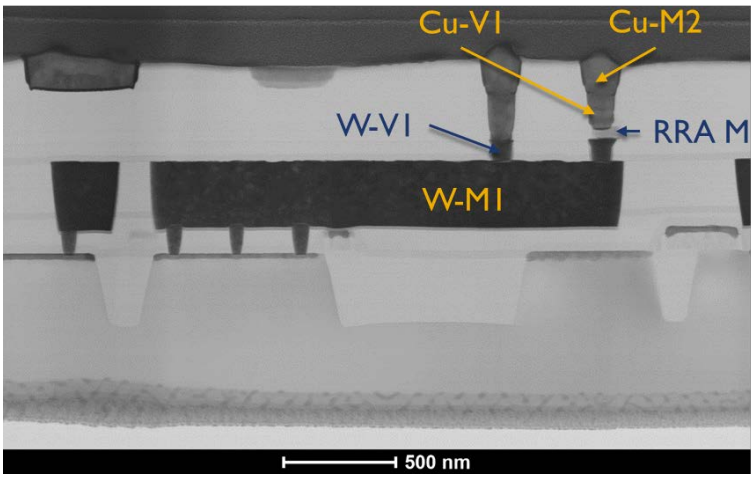

Figure 3. Cross-sectional contrast image of our integrated CMOS/ReRAM circuit taken with a transmission electron microscope. This cross-section shows the W bottom electrode and ReRAM device, as fabricated in a FEOL-compatible process.

3.2 Device Testing

Professor Cady's lab maintains and operates a B1500A semiconductor analyzer connected to a manual probe station capable of handling 150mm wafers or pieces of 300mm wafers. The mainframe is equipped with 4 high-resolution SMU units, a capacitive measurement unit (MFCMU) and waveform generating and fast-measurement unit (WGFMU). ReRAM device characteristics were extracted by using DC-sweep as well as pulsing techniques. In both cases a 1 transistor 1 ReRAM (1T1R) setup was used to limit the current through the ReRAM during the set and forming operation to the saturation current of the transistor which was set by the transistor gate voltage. Mainly two kinds of transistors were used: **1.** an external JFET (Junction gate field effect transistor) which was connected to the system via a discrete Keithley transistor box and **2.** an integrated on-chip transistor which was implemented right underneath our integrated ReRAM. A manual probe station was used to generate preliminary results and longtime endurance measurements. DC-sweeps, as well as a self-developed pulsing software were used in conjunction with the WGFMU enables endurance measurement up to 10^{12} cycles while recording every single cycle.

Two semi-automatic temperature (Suss Microtech) controlled 300mm probe stations, one with the B1500A/B1530A analyzer and another with a Keysight E5270A - 8 channel SMU Parametric Measurement Unit, were also operated. A Keithley Model 707 Switching Matrix setup was used in conjunction with the Keysight E5270A and a 2x12 pin probe card allowing for array testing measurements. An operating GUI was used, created from in-house Python code, that allows for use on all three probe stations.

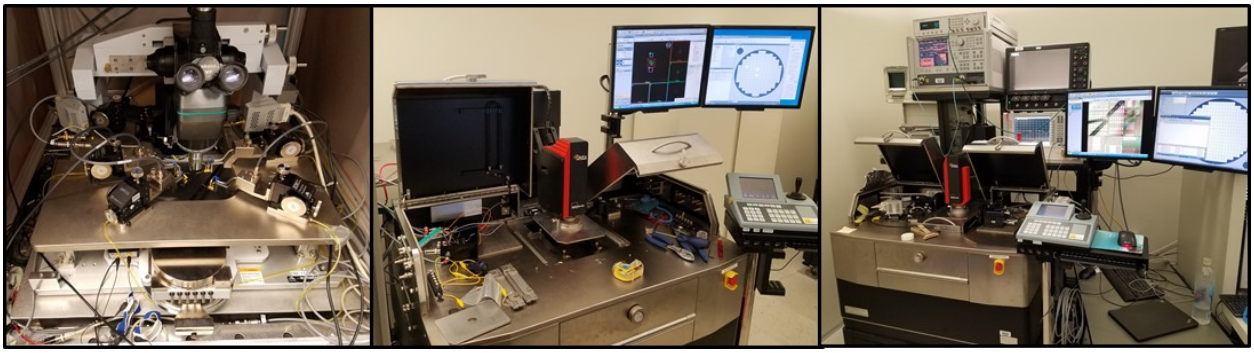

Figure 4. Probe stations used for this effort. Manual probe station with B1500A analyzer (left), Suss Microtech semi-automatic probe station with Keysight E5270A, and Suss Microtech semi-automatic probe station with B1500A/B1530A (right).

Electrical measurements primarily utilized an on-chip NMOS field-effect transistor (NFET) for current control during the forming and set operation, as seen ***Figure 5*** (b) illustrates the pulse form of one switching cycles comprising of a set-read-reset-read stream. The read pulses are necessary due to an increased noise level while reducing the pulse width of the set/reset pulse. To eliminate overshoots during the set/reset pulse a triangular pulse form was deployed. This reduces high frequency components of the pulse itself and thus increases voltage accuracy and limits potential stress to the ReRAM device. The WGFMU setup is used for endurance measurements and to determine switching parameters like forming, set and reset voltages and the dependence of resistance states on different current compliances during the set operation.

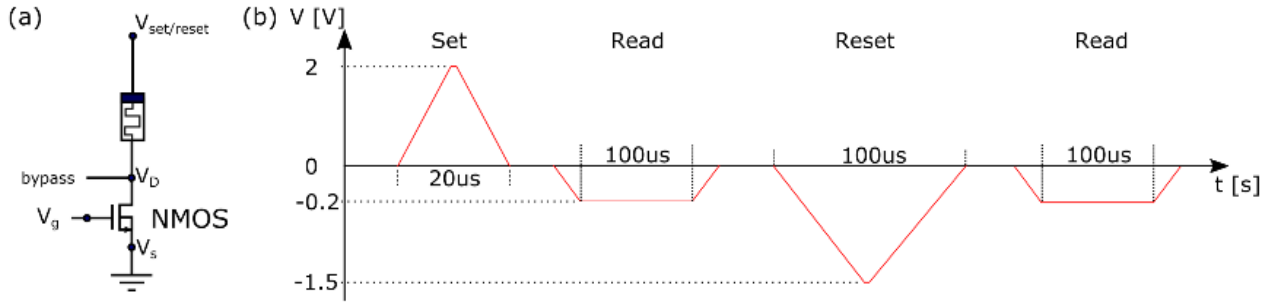

Figure 5. (a) Schematic of a 1T1R structure with an NMOS that acts as the current limiting device during the forming and set operations. A parasitic base diode opens during the reset allowing for a higher current during the reset operation. The bypass connection enables the direct measurement of the transistor. (b) Illustration of a pulse-based switching cycle applied to a 1T1R structure with the B1530A WGFMU.

To enable incremental resistance changes of the ReRAM devices, shorter pulses need to be applied to the 1T1R structure. For this purpose, the B1500A semiconductor analyzer was extended with a digital storage oscilloscope (Keysight DSO 9254A) and a pulse generator (Keysight 81130A), capable of applying pulses with a rise and fall time down to 5 ns and a maximum peak-peak voltage of 3 V. Together with a 50 Ω matched cabling and high frequency probe tips, this allows for an accurate characterization of on-chip 1T1R structures with FWHM pulses of 5 ns and the subsequent resistance read operation.

4.0 Results and Discussion

4.1 Key Accomplishments

4.1.1 CMOS/Memristor Circuit Design

Overall Design and Mask Tapeout

For this task, the SUNY Poly team focused on integration of mrDANNA circuit designs, individual RRAM (memristor) test structures, RRAM memory arrays, and novel reservoir-based neuromorphic circuits into a full mask (reticle) set for fabrication. The mask set (for 65nm CMOS, combined with a nanoscale memristive layer) was fully taped out during the performance period and was submitted to a mask vendor (Toppan) for data preparation and reticle production. The mask set integrated circuit designs from our internal team (at SUNY Poly) as well as our collaborators at UT-Knoxville (PI's - Rose, Plank, Dean). ***Table 1*** provides a detailed account of the circuits that were included on the mask set by the SUNY Poly team. ***Figure 6*** shows the global layout of the chip design and ***Figure 7*** shows the layout of the 512x512 RRAM array.

Table 1. List of individual circuits that are part of the SUNY Poly / UT-Knoxville hybrid memristor/CMOS chip.

ID	Count/Die	Circuit
1	1	512x512 1T1R array (262,144x 100x100nm^2) w/ decoder (9 bit/ 9 word lines)
2	1	512x512 1T1R array (262,144x 100x100nm^2) w/ decoder (9 bit/ 9 word lines) and ESD contact pads
3	4	8x8 1T1R array (100x100nm^2) w/o decoder
4	16	12x12 1R array (100x100nm^2)
5	18	2x2 form and cut 1R structures
6	2	ReRAM time-delay PUF (http://ieeexplore.ieee.org/document/7484314/)
7	1	CMOR (Cellular Memristive Output Reservoir – reservoir computing circ.)
8	1	Full digital DANNA array (UT-Knoxville / Dean)
9	Multiple	mrDANNA “neurons” (various types/configurations – UTK design)
10	100	1T1R (100x100nm^2) rVt NFET 2mA VCM w/ ESD contact pads
11	100	1T1R (100x100nm^2) rVt NFET 2mA VCM w/o ESD contact pads
12	100	1T1R (100x100nm^2) dgx NFET 2mA VCM w/o ESD contact pads
13	100	1T1R (100x100nm^2) dgx NFET 2mA VCM w/ ESD contact pads
14	100	1T1R (100x100nm^2) rVt NFET 2mA ECM w/ ESD contact pads
15	100	1T1R (100x100nm^2) rVt NFET 2mA ECM w/o ESD contact pads
16	100	1T1R (100x100nm^2) dgx NFET 2mA ECM w/o ESD contact pads
17	100	1T1R (100x100nm^2) dgx NFET 2mA ECM w/ ESD contact pads
18	48	RF 1T1R (100x100nm^2) dgx NFET 2mA VCM w/o ESD contact pads
19	16	1T1R (100x100nm^2) capacitive structures (10fF – 50pF)
20	80	1T1R on-chip pulse creation (1ns – 20ps)
21	24	Configurable XOR with pull down/up ReRAM (100x100nm^2)
22	8	1T dgxfet NFET 500uA test structure
23	8	1T dgxfet NFET 1mA test structure
24	8	1T dgxfet NFET 2mA test structure
25	8	1T dgxfet PFET 500uA test structure
26	8	1T dgxfet PFET 1mA test structure
27	8	1T dgxfet PFET 2mA test structure
28	8	1T rVt NFET 500uA test structure
29	8	1T rVt NFET 1mA test structure
30	8	1T rVt NFET 2mA test structure
31	8	1T rVt PFET 500uA test structure
32	8	1T rVt PFET 1mA test structure
33	8	1T rVt PFET 2mA test structure
34	1	Metal-Insulator-Metal capacitive test structures

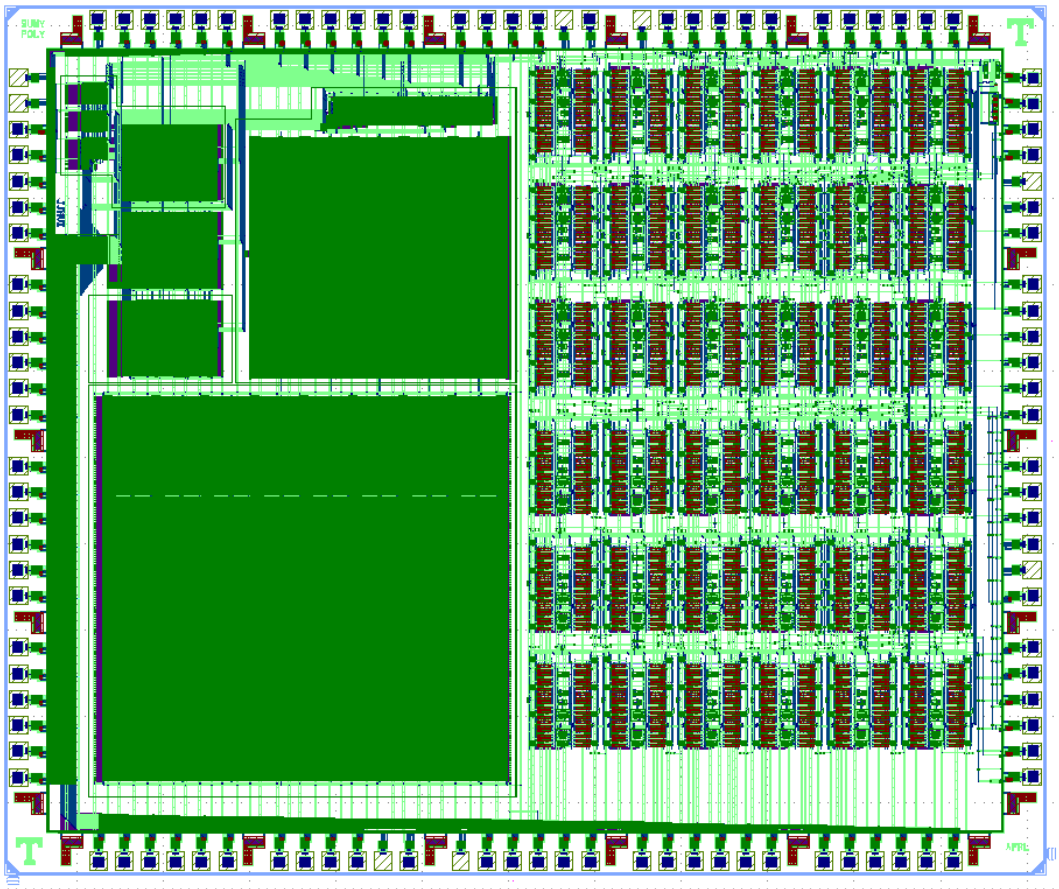

Figure 6. Global layout of the SUNY Poly / UT-Knoxville hybrid memristor/CMOS neuromorphic test circuits designed during this performance period.

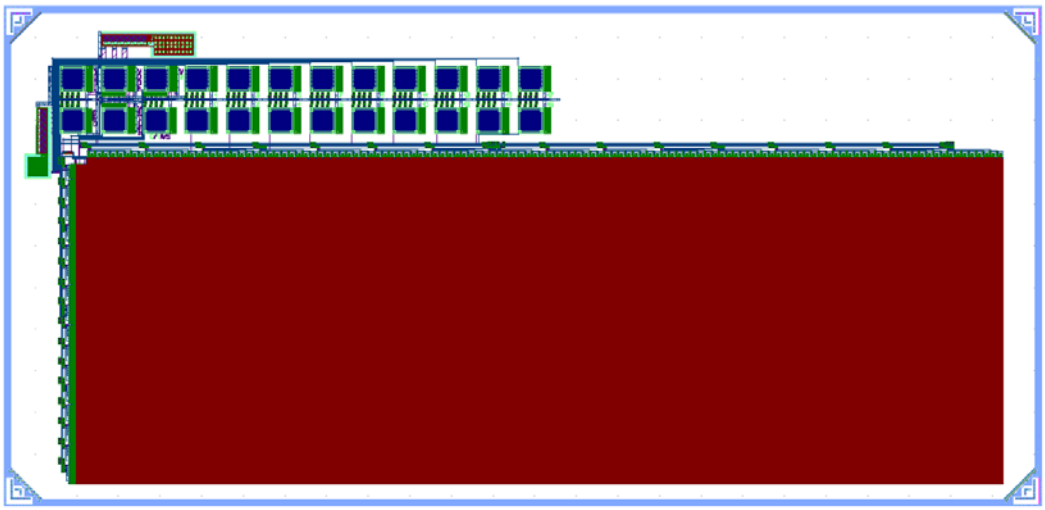

Figure 7. Global layout of the SUNY Poly 512x512 1T1R array with 100x100nm^2 devices using a decoder and ESD contact pads.

The SUNY Poly team performed final "tape out" of the mrDANNA reticle set, along with our partners at UT-Knoxville and our photomask (reticle) vendor, Toppan, Inc. During the tape out process, the various circuit design layouts are assembled into a single dataset, surrounded by a "frame" (which includes various test structures, alignment marks, etc.) and then run through a series of design rule checks (DRC), optical pattern correction (OPC) and layer vs. schematic (LVS) checks. These steps ensure that the designs adhere to the design rules of the process technology (65nm CMOS) and that the as-designed circuits match the optical patterns that have been generated by the reticle vendor (Toppan). This is an iterative process, requiring multiple revisions and corrections of the original layout and OPC to generate the reticle writing data and subsequent patterning of the reticles. During this reporting period, DRC, OPC and LVS were completed for the final tape out, and all reticles were ordered and received from Toppan, Inc. At the time of this reporting, the first 300mm silicon wafers were started in the SUNY Poly 300mm foundry for final chip fabrication.

As described in the summary, above, we performed a complete tape out of the mrDANNA design for a new 300mm reticle set. The reticle vendor for this effort was Toppan, Inc. The SUNY Poly team aggregated all of the design (GDS layout) from our own group and from our collaborators at UT-Knoxville. The aggregated designs were then run through DRC and adjusted accordingly to make sure that all design rules (for the 65nm CMOS process flow with custom RRAM layers) were met. Following DRC, the design files were sent to Toppan, Inc. for final data prep and tape out.

Prior to sending GDS files to Toppan, the SUNY Poly team added sub-reticle frames that included test structures for optical, metrology, and electrical characterization during the CMOS fabrication process flow. The assembled frame + circuit layouts were the subjected to OPC, which is necessary to convert the as-designed layout into optically patternable features that will resolve properly during photolithography. An example of structures before and after OPC are shown in ***Figure 8***.

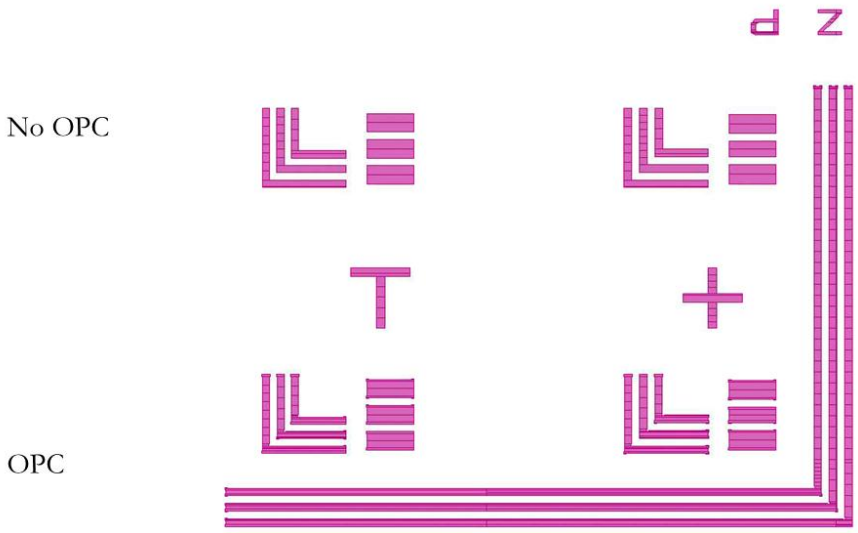

Figure 8. Reticle features before (No OPC) and after (OPC) optical pattern correction. Note the addition of rounded structures to the corners of rectangular features as one example of the pattern correction for photolithography.

Another part of the "data prep" performed by Toppan includes the addition of so-called "fill" features. These are shown in ***Figure 9***. Fill features are additional structures (usually square or rectangular structures) that are placed in between the designed structures/features in a particular layer. The "fill" structures are etched, filled, polished, etc. the same as the other features in that layer, and serve to normalize these processes across the surface of the wafer. In this way, the etch, feature fill, polishing (typically CMP) are more uniform than if the fill structures are not included.

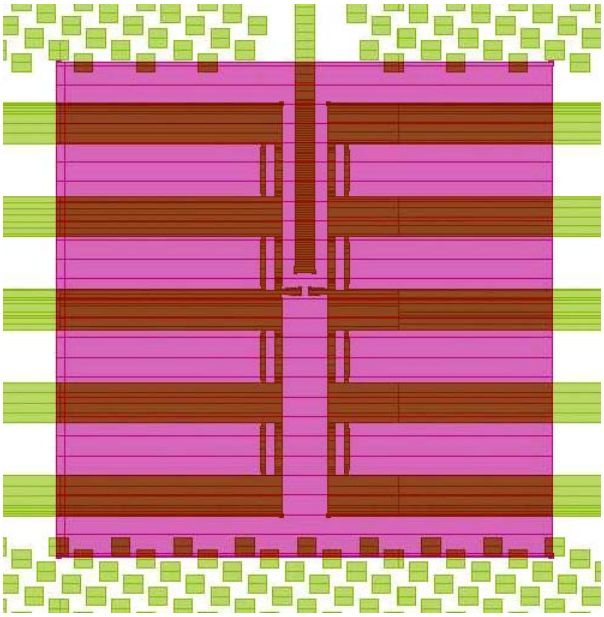

Figure 9. Overlay of two layers within the design. Fill features are the small green squares that surround the designed features (various metal contacts). The large square overlay represents an exclusion zone around the critical structures.

During the tape out process, the SUNY Poly team also compared the as-designed GDS schematics with the data prep that was performed by Toppan for writing the reticles. This process, called layer vs. schematic (or LVS) ensures that designed circuit layout is preserved through the various steps of data preparation, and thus the final reticles and fabrication flow should result in the as-designed circuits.

The unique aspect of the reticle set that was taped out for mrDANNA is the fact that multiple "layers" or "levels" were included per individual reticle. Rather than using the full reticle field, each reticle was divided into quadrants. Each quadrant has its own frame (with alignment marks, fiducials, etc.) that allow for lithographic patterning of an individual quadrant (vs. the entire reticle field). While this reduces the maximum area of a printed layer, it enables us to significantly save on the total cost of the reticle set (approximately 40-50% savings over a full reticle set.

Cellular Memristive Output Reservoir (CMOR)

The SUNY Poly team included a unique neuromorphic circuit design into the final tapeout. This circuit was a so-called "reservoir" computing circuit that leverages both CMOS and memristive (ReRAM) devices. Reservoir computing is a novel approach to neuromorphic computing. In this approach, a "reservoir" consists of many independent computation units, such as neurons which are randomly connected to one another. Inputs applied to the reservoir cause it

to change its behavior and outputs. An output layer which examines the state of the reservoir is then trained independently from the reservoir to recognize outputs from the reservoir which are related to the property being trained (such as an image class).

However, many computational substrates other than neurons have been demonstrated for reservoir computing. One of the seminal papers in reservoir computing used water in a bucket as the literal reservoir; inputs were applied by creating waves in the water, and interference patters on the surface were recorded as the reservoir's output [1]. The basic requirement for a reservoir is that applied inputs can create a wide variety of different states within the reservoir. Additionally, to apply this computing method to temporal tasks, the reservoir must have an 'echo' property in which effects from previous inputs persist in the reservoir for a short time.

Cellular automata (CA) are interesting mathematical constructions which have potential for reservoir computing. A classic example of a CA is Conway's "Game of Life," in which a grid of cells follows a set of rules for each cell to be declared 'living' or 'dead.' One-dimensional CA are a more restricted class of automata which exist in one dimension, with neighbors looped into a ring at each end. For each time step, each cell is determined to be in a 0 or 1 state based on the previous state of itself and its nearest two neighbors. For a one-dimensional binary CA, this gives a possible 2^8 or 256 rules to govern CA behavior. Some of these behaviors have been shown to be very interesting, generating chaotic behaviors and demonstrating Turing-completeness [2].

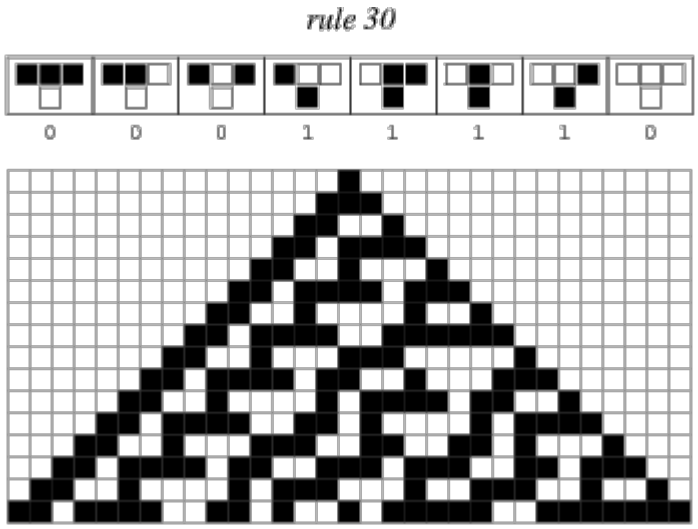

Figure 10. Illustration of a 1-D cellular automata (CA), where the input is applied at the top-center, and the 1D CA evolves with time downwards [3].

CAs have been demonstrated in software as reservoir elements, with several different schemes available for introducing echo-states into the computation [4]. Advantages of using CA as a reservoir system include the simplicity of the active element, and the mathematical tools which can be applied to analyze and improve these systems. Additionally, power consumption is low, as a CMOS implementation will only use active power after inputs are changed and the output layer is being read. Using the IBM 10LPe 65nm Process Design Kit, we designed a CMOS implementation of a 1-D CA reservoir, combined with a trainable, memristive output layer.

Each cell consists of a programmable multiplexer, with its inputs set to the 1-D rule the automata will follow. The selection lines of the multiplexer determine the cell's output, given the previous state of the cell and its two neighbors. In this way, the multiplexer performs the basic

calculation for a cellular automata. Each row of the reservoir consists of a ring of multiplexers which react to the outputs of the previous row, or at the top row, the inputs to the system. These rows are cascaded downwards to create a reservoir system representing *n* time steps, where *n* is the number of rows.

Additionally, each cell is connected to a selection transistor which enables read-access to a memristor. All memristors are connected in parallel, and are enabled or disabled based on the results of the CA reservoir. This forms the reservoir's analog memristive output layer. Measuring the collective resistance value of the parallel memristors can indicate a high or a low state after memristor values have been trained. This is essentially equivalent to the 'kernel trick' in support vector machines, where the CA functions as a non-linear kernel and the memristor output layer is similar to a vector machine.

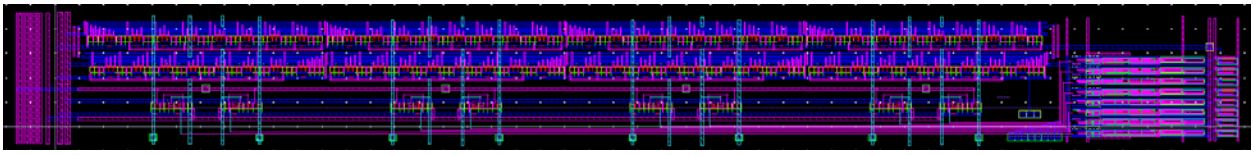

Figure 11. 65nm CMOS layout of a row in a CA reservoir. Ringed multiplexers are center-left, and their inputs are taken from above, their outputs cascaded down to the next row. The outputs also drive the selection of memristors in the output layer (block on the right).

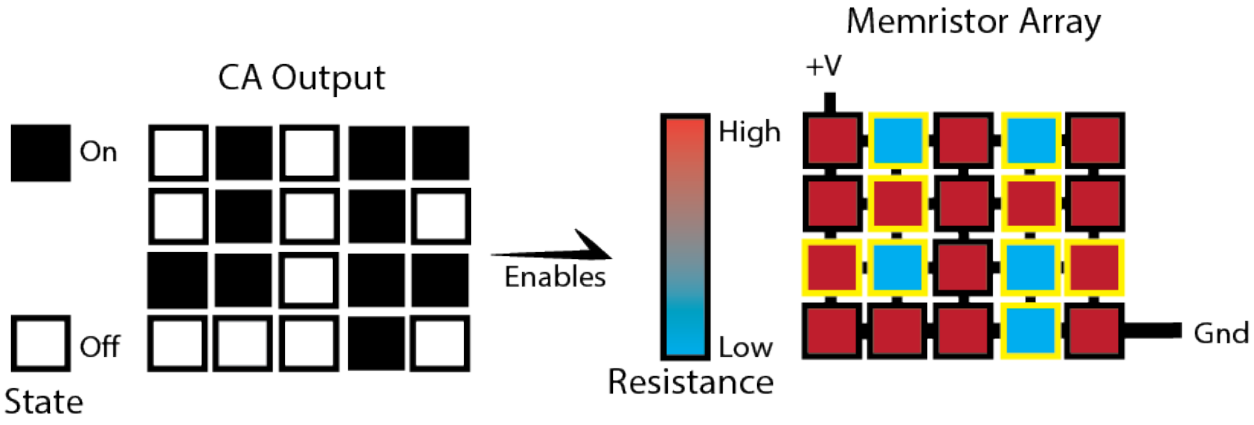

Figure 12. Illustration of how CMOR's CA reservoir controls the memristive output layer. The total resistance given this state will be very low, as 5 low-resistance memristors are enabled in parallel.

Current implementations of this architecture which were added to extra space available on a mask set include 8-bit wide automata, with 7 time steps (rows). Eight copies of this 8x7 structure implement different automata rules for testing. Future designs could include reprogrammable rules for each cells, and larger structures for more complex learning tasks.

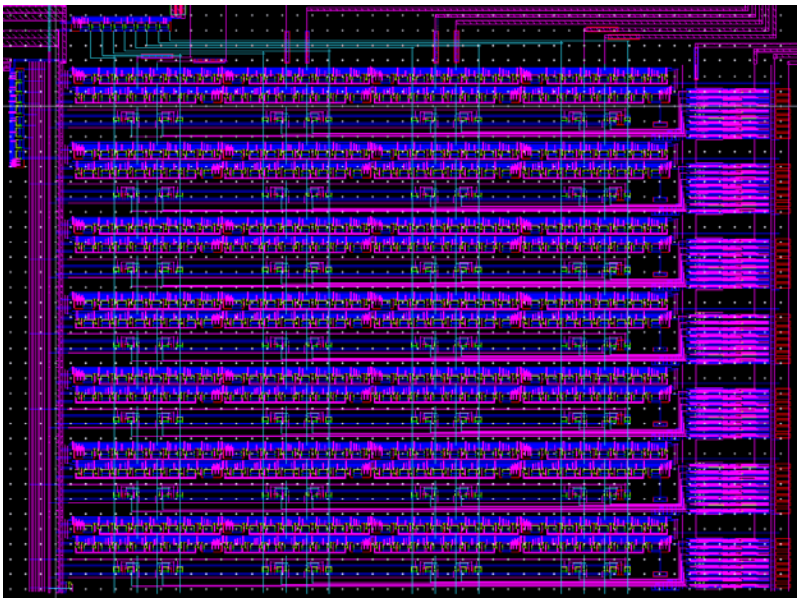

Figure 13. Layout of a full 8x7 CA reservoir structure with memristive output layer.

4.1.2 Chip Testing Results

300mm Wafer Processing Results and Full-wafer Testing

Using the mrDANNA mask set designed under this effort, the SUNY Poly team utilized the SUNY Polytechnic Institute Center for Semiconductor Research (CSR) 300mm wafer scale nanofabrication facility to build functioning CMOS and integrate hafnium oxide ReRAM. The fabrication approach and cross section of the hybrid CMOS/ReRAM stack is described in Section 3 and ***Figure 3***, above. An example 300mm wafer that resulted from this work is shown in ***Figure 14*** and a close up of individual die on this wafer is shown in ***Figure 15***.

During the fabrication process, in-line electrical testing was implemented to measure CMOS performance and yield. Initially PMOS yield was low, but after optimization of implant parameters and gate etch parameters, we were able to yield both PMOS and NMOS within acceptable performance metrics. ***Figure 16*** and ***Figure 17*** show full wafer maps of average transistor current measured from in-line testing of NFETs and PFETs, respectively. These figures demonstrate that we were able to achieve good yield across the wafer, although there was some variation, especially in individual die near the edges of the wafers (which is to be expected, especially in a non-production CMOS foundry). ***Figure 3.*** *Cross-sectional contrast image of our integrated CMOS/ReRAM circuit taken with a transmission electron microscope. This cross-section shows the W bottom electrode and ReRAM device, as fabricated in a FEOL-compatible process.*

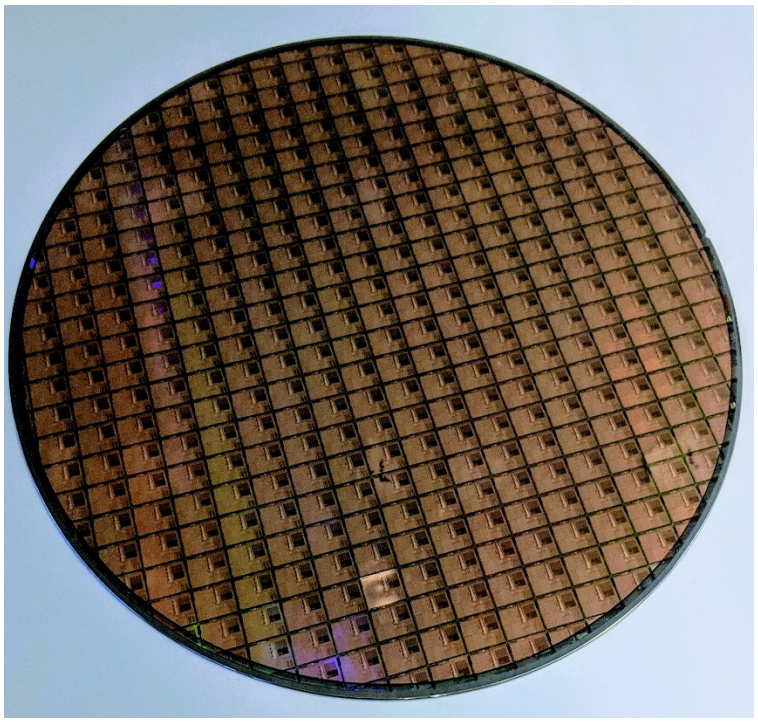

Figure 14. mrDANNA test chips arrayed onto a 300mm Si wafer, fabricated in the SUNY Polytechnic Institute nanofabrication facility.

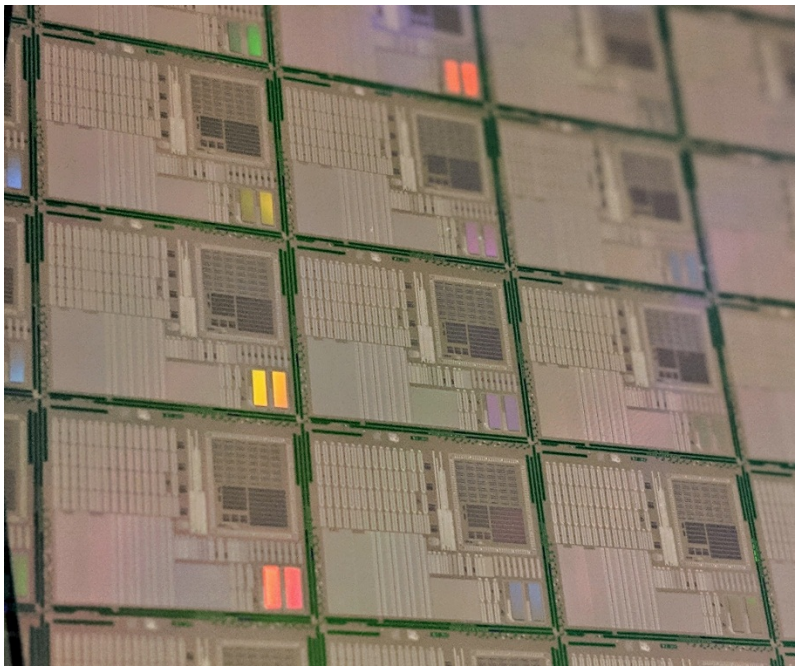

Figure 15. Individual mrDANNA chips fabricated on 300mm Si wafer platform.

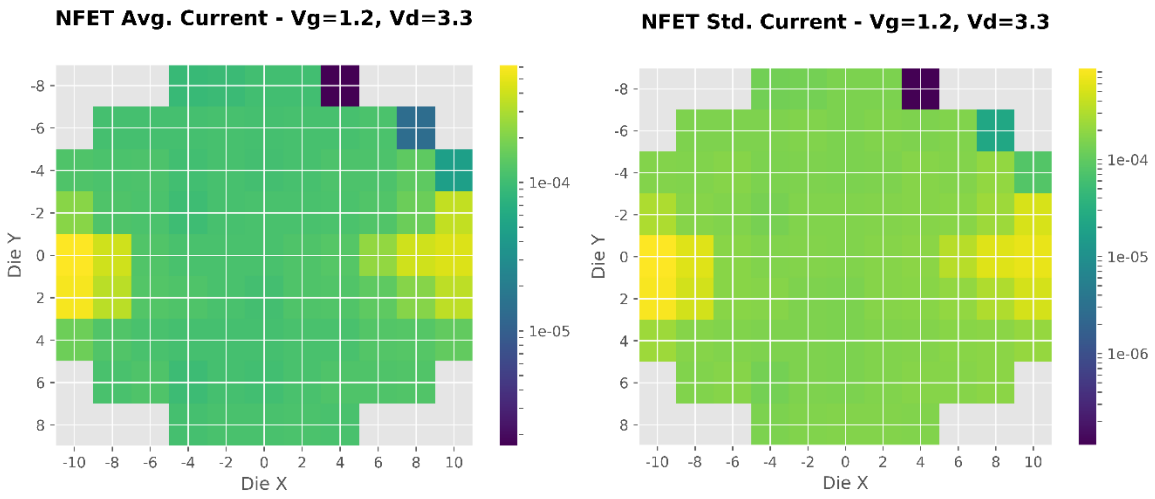

Figure 16. CMOS NFET performance showing average current across a full 300mm Si wafer from the SUNY Poly nanofabrication facility. Average NFET current (left) and standard deviation of measured current (right).

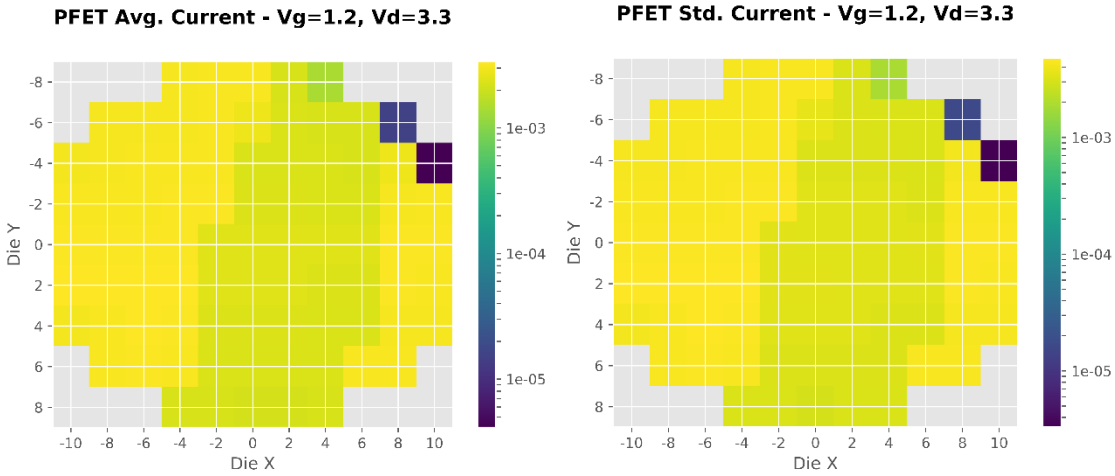

Figure 17. CMOS PFET performance showing average current across a full 300mm Si wafer from the SUNY Poly nanofabrication facility. Average PFET current (left) and standard deviation of measured current (right).

In addition to measuring CMOS performance with in-line testing, integrated ReRAM performance was characterized using full-wafer testing outside of the 300mm cleanrooms at SUNY Poly. After multiple rounds of optimization (including the evaluation of multiple different hafnium oxide film compositions, thicknesses, and oxygen exchange layer (OEL) thicknesses), we were able to achieve extremely high ReRAM yield and performance across the entire wafer. ***Figure 18*** (left) shows that the forming, set and reset voltage for ReRAM (in 1T1R configuration) was very consistent across the wafer (center to edge). The maximum resistance value (HRS) for these devices varied across the wafer, with highest resistance found on die at the center of the wafer (***Figure 18***, right). This is not surprising, since extremely small variations in hafnium oxide

and OEL layer thickness can affect the maximum HRS. For research purposes, die from the center of the wafer are preferable, as they exhibit the largest memory window (approaching 1 MΩ HRS and maintaining < 10 kΩ LRS).

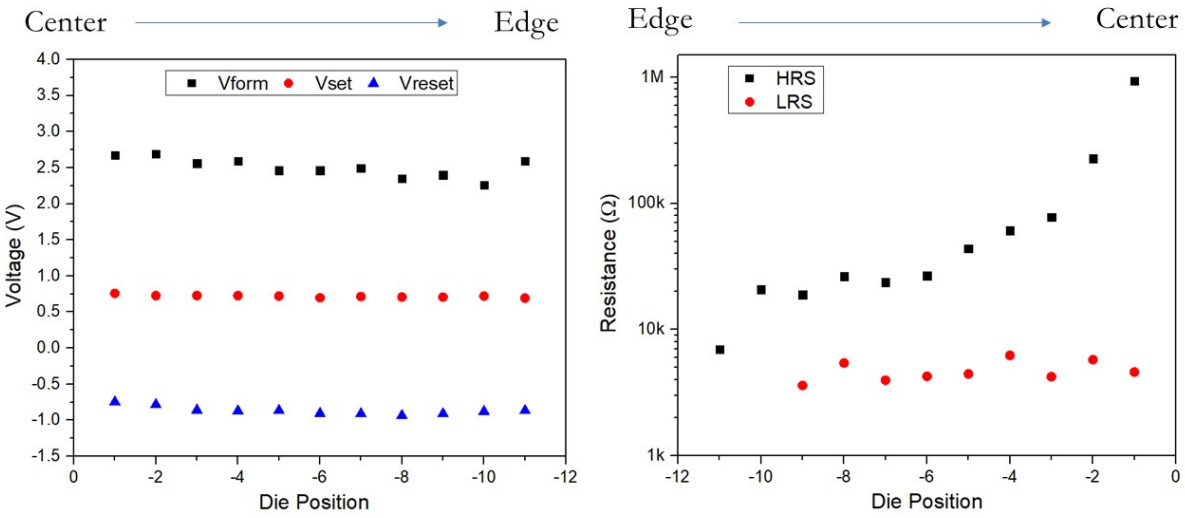

Figure 18. ReRAM performance as a function of die position on a 300mm wafer. Forming, set and reset voltage were highly consistent across the wafer (left), while the maximum HRS increased from edge to center, and LRS was consistent across the wafer (right).

Full wafer mapping of ReRAM performance was also measured, with results shown in ***Figure 19***, ***Figure 20***, and ***Figure 21***. These results demonstrate that we can achieve excellent consistency for forming, set, reset, and HRS across the entire wafer. LRS (as shown in ***Figure 21***) was the most variable, with some devices measuring outside of the expected range that was programmed into the full-wafer test software. We have an ongoing effort to improve ReRAM performance consistency across the entire wafer. That being said, the wafer mapping exercises allow us to characterize representative test structures in individual die, then choose die with the best performance parameters for full-circuit testing.

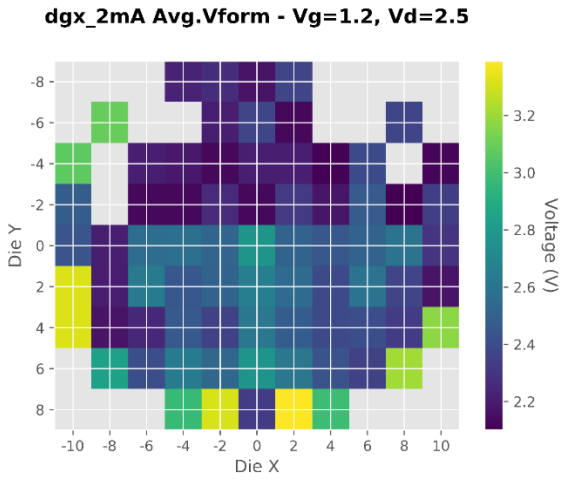

Figure 19. Full 300mm wafer map of mean forming voltage (Vform) for ReRAM devices.

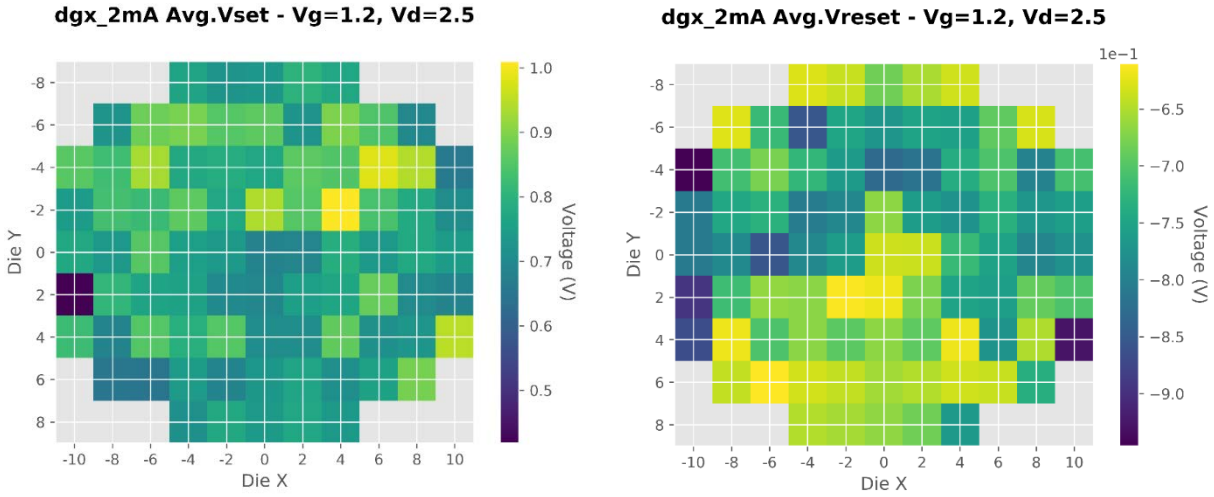

Figure 20. Full 300mm wafer map of mean set voltage (left) and reset voltage (right) for ReRAM devices.

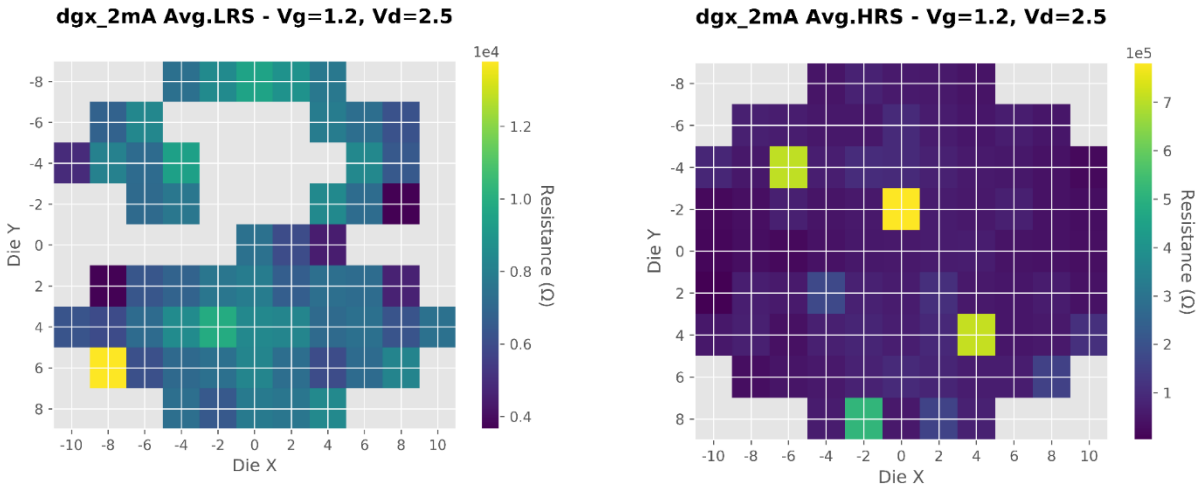

Figure 21. Full 300mm wafer map of mean LRS (left) and HRS (right) for ReRAM devices.

DC characterization of ReRAM devices

The performance of ReRAM devices fabricated using the mask set taped out under this effort was first characterized via DC measurements. ***Figure 22***(a) shows successful switching after the forming event with a current limit set to 100 µA. A total of 22 switching cycles were applied in DC operation. Forming, set and reset voltages were 2.2, 0.7 and -0.7 V, respectively, while the whole reset operation requires a voltage of about -1.3 V to reset to its highest possible resistance. The set current limit was set to 180 µA which was followed by the reset current that equaled the defined set current limit. A total of 22 switching cycles were performed in DC operation and the resulting low resistive state (LRS) and high resistive state (HRS) can be seen in ***Figure 22*** (b) with LRS and HRS values of 4 and 70 kΩ, respectively.

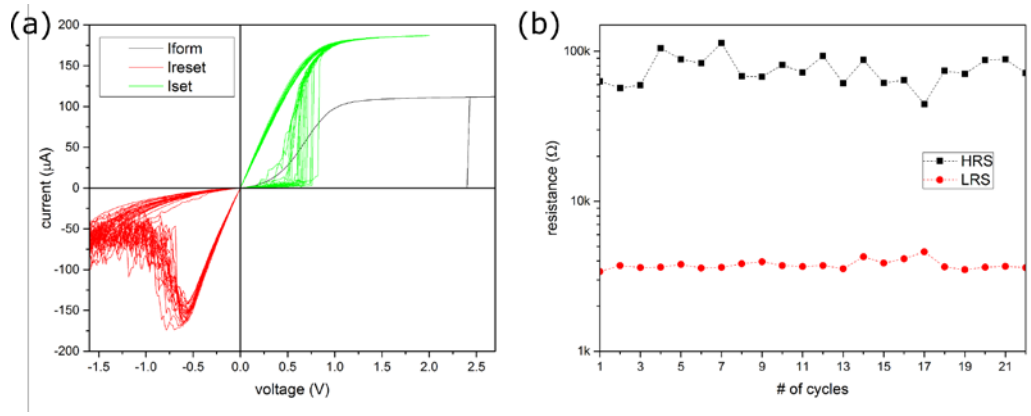

Figure 22. (a) DC switching performance of a ReRAM device formed with a current limit set by the series transistor in the 1T1R structure to 100 µA, (b) shows the achieved resistance states over 22 cycles from (a). A LRS of 4 kΩ and a HRS of 75 kΩ was achieved.

With help of the DC characterization, a change in conduction mechanism from the LRS to the HRS can be observed and quantified. Two equations govern the fitting. First, the LRS can

easily be fitted to an ohmic relationship (Eq. 1). Here R and G represent the resistance and its inverse, the conductance of the ReRAM device. Second, during the switch from the LRS to the HRS, an improved fit to the data can be made using a Schottky-emission model. Equation 2 shows the relationship between the current density, J, with A* being the effective Richardson constant, T is the temperature, q is the elementary charge, Φ_B is the Schottky-barrier height, E is the electric field across the barrier, ε_0 and ε_r are the permittivity and the relative dielectric constant, respectively, and k is the Boltzmann constant. The ReRAM devices in this study showed a linear I-V relationship for the first three reset voltages, namely, -0.7, -0.8 and -0.9 V. The resistance progressively increased from about 4 to 6 kΩ while no decrease in the fitting accuracy was observed (***Figure 23*** (b)). For the Schottky-emission fits, R2 values tended to increase from a minimum of 0.96 to a maximum of 0.995 with increasing reset voltage from -1 to -1.4 V, while the most accurate fit was achieved after a reset voltage of -1.4 V. This tendency is clearly visible in ***Figure 23*** (c) where the regular residual value is plotted versus the fit for the Schottky-emission data. The residual values decrease from a maximum with the -1 V reset to a minimum with the -1.4 V reset. This demonstrates an incremental move to a completely Schottky-emission limited current.

$$I = \frac{V}{R} = V \cdot G \qquad [1]$$

$$J = A \cdot T^2 \exp\left[\frac{-q(\phi_B - \sqrt{\frac{qE}{4\pi\varepsilon_r\varepsilon_0}})}{kT}\right] \qquad [2]$$

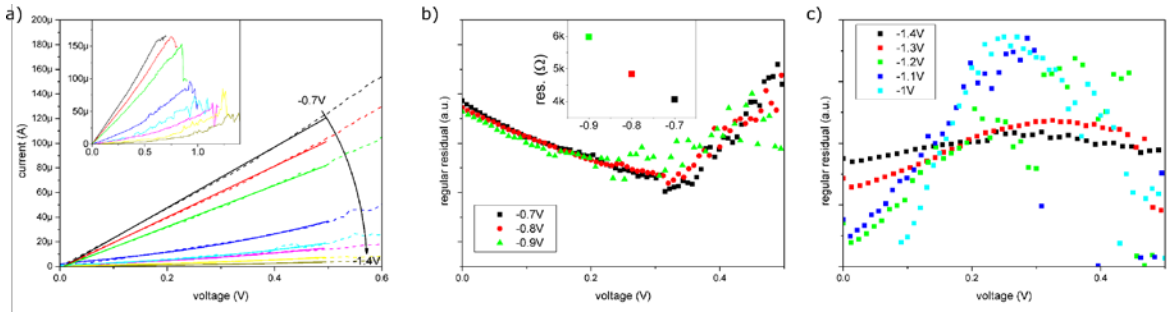

Figure 23. Conduction mechanism change from ReRAM. (a) Shows the fitted curves for reset voltages ranging from -0.7 to -1.4 V. (b) Shows the regular residual for the linear fits and the inset shows the increase in resistance while increasing the reset voltage from -0.7 to 0.9 V. (c) shows the increase in fitting accuracy by the reduction in regular residual from -1 to -1.4 V.

Endurance Testing

An important metric to determine the feasibility for the implementation of a ReRAM device into a neuromorphic circuit is its endurance. ***Figure 24*** shows the capabilities of our ReRAM device to perform reliably over 1.5 billion cycles. An increase in variability can be observed after this point, up to the 10 billion cycles that were characterized. For this effort, a pulse-based switching approach (vs. DC switching) was necessary to keep measurement times down.

The fall and rise times were set to 1 µs while the pulse width was set to 100 ns. The read pulse was 10 µs with rise and fall times of 100 ns to achieve an adequate accuracy in determining the resistance states. Due to the high number of measurement points, only every 50th pulse was measured, resulting in 200 million measurement points in ***Figure 24*** (a) for LRS and HRS. ***Figure 24*** (b) shows the distribution of the LRS and HRS for the first 100 million cycles. The LRS has a tight distribution around 4.9 kΩ with a standard deviation of 185 Ω. The HRS shows a wide distribution ranging from 6.3 kΩ to 4 MΩ with a standard deviation of 76 kΩ and an average of 91 kΩ. A tendency towards higher variability was observed when increasing the reset voltage. This is shown in ***Figure 25***, where the cumulative percentile for the LRS and HRS resistances is show for reset voltages from -1 to -1.3 V, over the course of 10,000 cycles. The set current limit was set to 250 µA, yielding a LRS of about 4 kΩ. A small memory windows between 4.2 kΩ and 5 kΩ is visible. Higher reset voltage tended to increase the average HRS and thus the average R_{off}/R_{on} ratio and it was shown that the minimum R_{off}/R_{on} ratio was not affected.

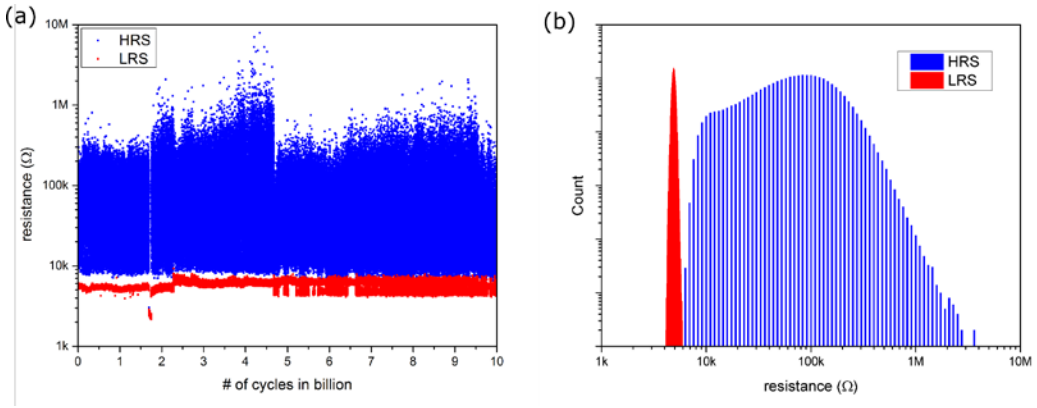

Figure 24. (a) Endurance measurement showing 10 billion cycles. (b) Distribution of resistance states with the LRS showing a tight concentration around 4.9 kΩ and the HRS has a wide distribution starting slightly above the LRS at 6 kΩ and ending at 4 MΩ.

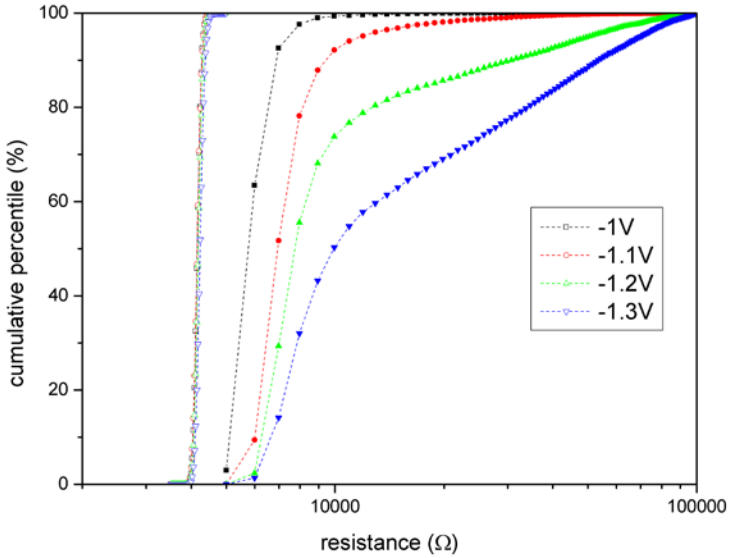

Figure 25. Cumulative percentile plot of device resistance values for increasing reset voltages starting at -1 V and increasing in -0.1 V steps to -1.3 V. The LRS is shown in open symbols while the HRS is shown in solid symbols.

Set and Reset Behavior

To determine the optimal operation for incremental resistance changes the time resolved reset behavior was measured for six different reset voltages, ranging from -0.6 to -1.6 V in 0.2 V steps. In ***Figure 26*** (a) six set operations are shown for the six different reset voltages used. The set operation itself was a 1.5 V pulse with 5 ns rise and fall time and a pulse width of 100 ns. An incremental change in resistance could not be observed, as the measurement resolution of 0.1 ns does not allow for a determination of the detailed set behavior. It needs to be mentioned that the time shift from the voltage pulse to the current response is due to the run-time difference of the pulse. The dip in current during the stationary part of the set voltage pulse is contributed to reflections suppressing the current. The reset operation shows a distinct change in behavior while increasing the reset voltage (***Figure 26*** (b)). For the set operation, the initial delay between the voltage pulse and the current response is due to the run-time difference. All reset pulses had a fall and rise time of 5 ns and a pulse width of 100 ns, which was sufficient to fully reset the ReRAM device within their variability limits. Following the current traces, the -0.6 V reset voltage showed no impact on the resistance state. Increasing this to -0.8 V results in a slow move from its maximum current at -100 µA to about -60 µA over a 30 ns period. Between -1 V and -1.2 V the reset operation was rapid, reducing to ~5 ns, while the final current decreased from -60 µA to -30 µA, without increasing in variability. For reset voltages of -1.4 and -1.6 V the overall progressive reset behavior changed towards an increased variability, which is shown in ***Figure 26***. From this information, an incremental resistance change can be expected near a reset voltage of -0.8 V and set current control of 80 µA.

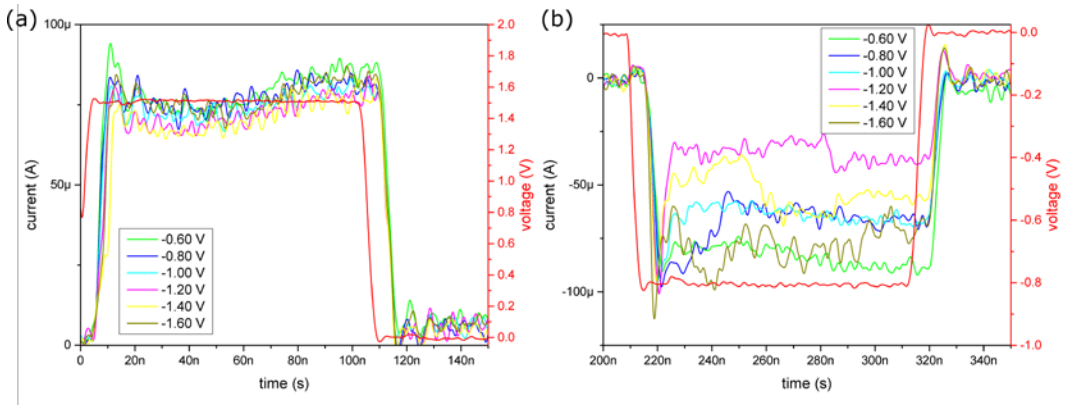

Figure 26. (a) and (b) show the set and reset behavior, respectively, when applying a 100 ns pulse with different reset voltages. A distinct set process cannot be distinguished from (a) while a distinct change of the reset behavior is visible in (b). An increase in reset voltage decreases the reset time but Increases the variability during the reset operation.

Analog Resistance Behavior

For neuromorphic applications, synaptic functionality can be implemented by modulating the relative resistance or conductance of circuit elements such as ReRAM. For this application, binary (high vs. low) resistance states are not enough, rather multiple states, approaching analog behavior, are needed. A pulse-based forming is performed for characterizing the ReRAM behavior with ultra-short pulses in the low nanosecond regime. The 3 V forming pulse is 1 ms long and the current is limited to the set current used during the succeeding analysis. Without pulse forming, the resulting CF increases laterally due to Joule heating causing continuous diffusion during the ramping of the DC voltage. This behavior has been established in other works by observing real-time growth of a filament under voltage stress conditions [1].

Analog Switching Using Compliance Current Control

In previous work we have shown that ReRAM LRS and HRS can be adjusted, in an analog fashion, by changing the peak compliance current used during switching. As shown in ***Figure 27***, adjusting the gate voltage on the control transistor in a 1T1R configuration results in a change in the maximum current driven through the ReRAM element, and incremental changes in the LRS and HRS. Thus, by simply adjusting the Vg, a multitude of ReRAM resistance states can be achieved. While there is significant variability in the HRS, LRS is relatively consistent, and a definitive memory window is maintained throughout.

[1] Q. Liu *et al.*, "Resistive Switching: Real-Time Observation on Dynamic Growth/Dissolution of Conductive Filaments in Oxide-Electrolyte-Based ReRAM (Adv. Mater. 14/2012)," *Adv. Mater.*, vol. 24, pp. 1844–1849, 2012.

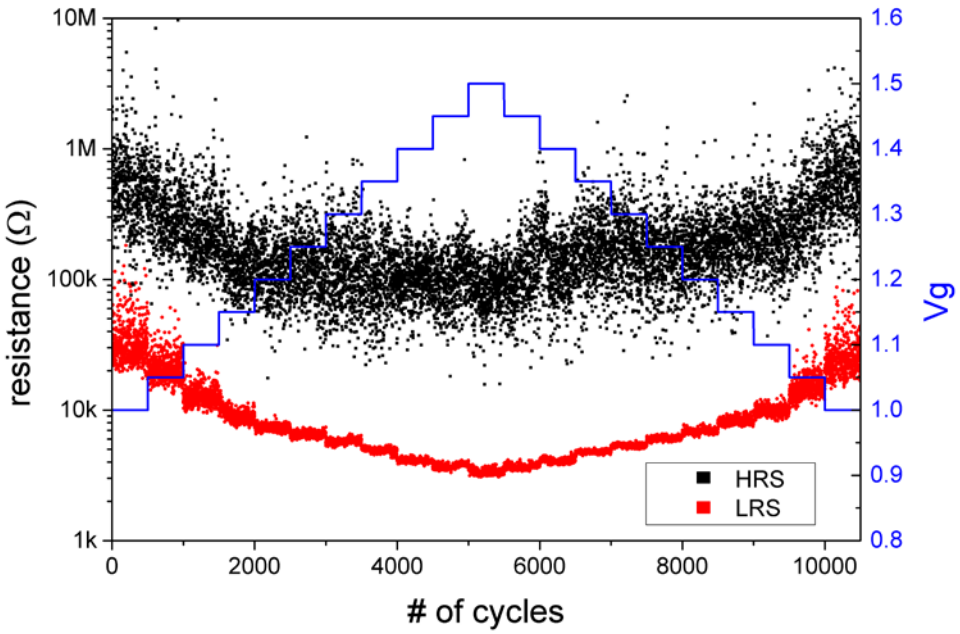

Figure 27. ReRAM resistance levels controlled by the peak current applied during switching. ReRAM devices in a 1T1R configuration were switched for 500 repeated cycles using incremental changes to the gate voltage (Vg) on the 1T1R control transistor. As the Vg increases (blue line), the LRS and HRS was increased accordingly.

Direct Write Approach

The feasibility of achieving an analog resistance behavior for our ReRAM devices was evaluated by applying pulse trains of 10 set/reset cycles with each set/reset pulse stream containing 100 pulses. The results are shown in ***Figure 28***, where the resistance after each set/reset pulse was measured with a 100 ns pulse at -0.2 V, plotted along the x-axis. Set and reset operations were performed with 5 ns FWHM pulses, corresponding to a 5 ns rise and fall time a voltage of 1 and -1 V, respectively. The set current compliance was 350 µA, yielding a final LRS of ~3 kΩ. During the subsequent pulses a change from this LRS up to about 5 kΩ was observed with a tendency for higher variability with progressive reset pulses. A similar tendency towards higher variability was not observed during the subsequent set pulses.

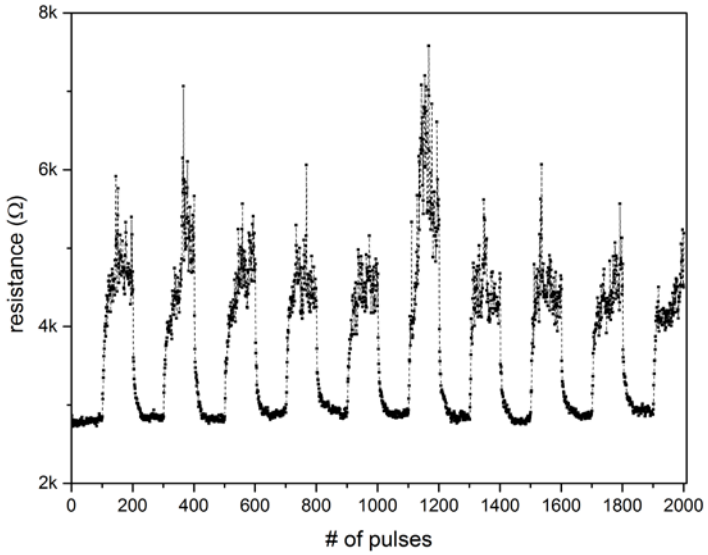

Figure 28. Example of incremental reset pulses applied to a 1T1R configuration. Resistance reads are shown after each 5ns FWHM pulse set/reset pulse. The set and reset voltage was 1 and -1V, respectively. A total of 10 set/reset cycles is shown with each set/reset cycle containing 100 pulses.

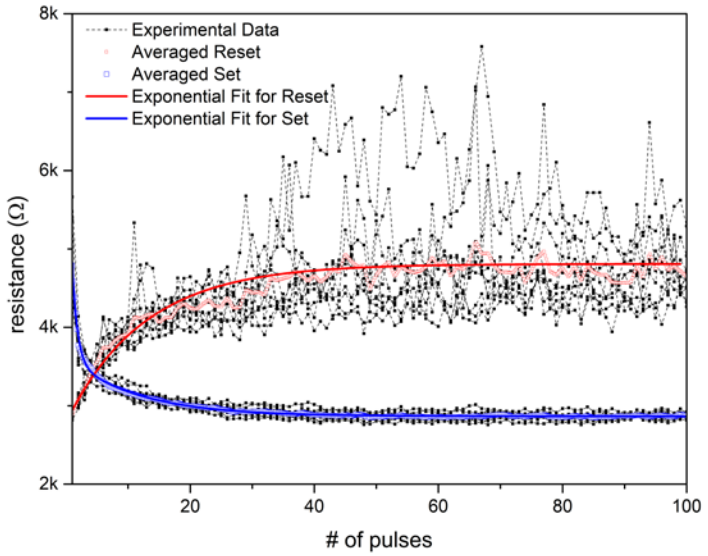

Figure 29. Example of an asymmetric response during the switching of a ReRAM device with a 5 ns pulse. Set and reset voltages were kept constant to 1/-1 V with a set current compliance of 250 µA. A total of 10 set/reset cycles were applied with 100 consecutive pulses for each set/reset operation.

As shown in ***Figure 28***, there was significant variability in the reset portion of the progressive reset pulse cycles. To better study this variability, the average of the 10 switching cycles was calculated for the set and reset operation. ***Figure 29*** shows the evolution of resistance states for 10 cycles, which are stacked atop each other resulting in a stream of 100 pulses for the set and reset

operation (shown with black symbols connected by dashed lines). The average of these resistance values is shown with blue and red symbols for the set and reset operations, respectively. The change in resistance from the LRS to the HRS was successfully modeled with an exponential decay function shown in Equation 3 while the change from the HRS to the LRS required a double exponential function shown in Equation 4. This is a result of the larger response to the applied pulses for set vs. reset, which was previously described by Marchewka et al. [2]. Equations 3 and 4 model a saturation resistance within the first 100 switching cycles. In these equations, R_0 is the saturation resistance, A_1/A_2 the amplitudes, x_0 the offset and t_1/t_2 the decay constants.

$$R_{set} = R_0 + A_1 \exp\left(-\frac{x - x_0}{t_1}\right) + A_2 \exp\left(-\frac{x - x_0}{t_2}\right) \quad [3]$$

$$R_{reset} = R_0 + A_1 \exp\left(-\frac{x - x_0}{t_1}\right) \quad [4]$$

The corresponding fits are shown in ***Figure 29***, where the blue and red lines are the fits for the set and reset operation, respectively.

The optimal incremental switching parameters for our desired applications are that the ReRAM devices would exhibit a linear resistance change (to applied voltage) with minimal stoch asticity. The effect of non-ideal resistance changes (by ReRAM circuit elements on neural network performance was described by Burr et al. in 2015 [8]. Broadly speaking, there are ~5 different performance characteristics that affect the accuracy of a neural network, these include the stochasticity of the resistance change, the non-linear change in resistance, variance in the final resistance level, device failures, and asymmetric changes in resistance during the set and reset operations. As seen from ***Figure 30***, the stochasticity and non-linear resistance change of our devices would probably be the biggest contributors to a reduction in accuracy for a neural network application. Thus, we evaluated the operational parameters of our devices to determine the best way to improve their suitability for synaptic devices in artificial neural networks. The best results were obtained with a forming/set current of 100 µA which is shown in ***Figure 30***. In these graphs, the reset voltage was gradually increased from -0.85 to -1.35 V until an incremental resistance change can be observed. A controllable incremental resistance change could only be observed for 100 µA forming/set current limit.

[2] Marchewka, A., Roesgen, B., Skaja, K., Du, H., Jia, Chun-Lin, Mayer, J., Rana, V., Waser, R., Menzel, S. (2016).Nanoionic Resistive Switching Memories: On the Physical Nature of the Dynamic Reset Process. *Adv. Electron. Mater.*, 2: 1500233. doi: 10.1002/aelm.201500233

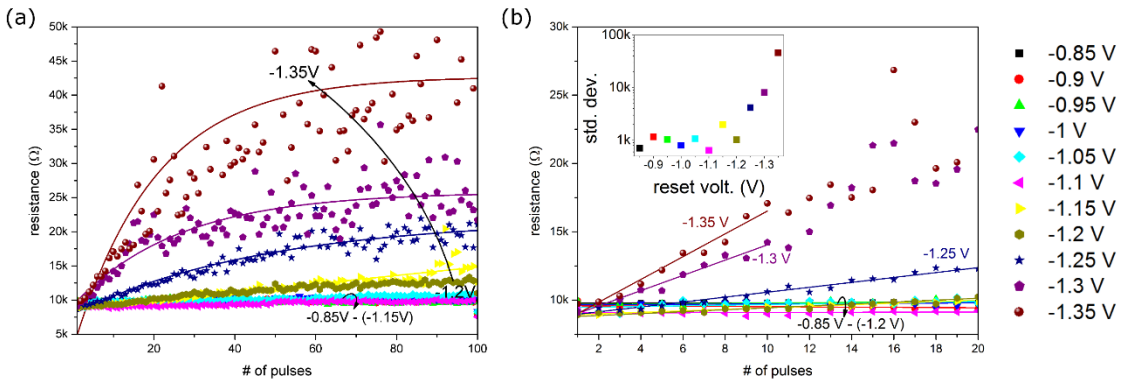

Figure 30. Fitting of the average resistance obtained by applying 10 incremental reset cycles with 100 5 ns FWHM pulses to a ReRAM device. (a) Exponential fit of Equation 4 to the average pulse response values. (b) Linear fit of the first 10 averaged pulse responses.

Figure 31 shows the fitting results, (a) uses an exponential fit shown in Equation 4, while a linear fit was achieved for the first 20 cycles of each iteration shown in (b). The reduced set current has a significant positive effect on the resistance response of consecutive reset pulses. The lower set current reduced the reset voltage necessary to cause significant changes in resistance state (from -1.15 to -1.1 V). The biggest advantage can be seen in how the average resistance state trends with the fitted function. These parameters enable an empirical and predictable model for an implementation into a novel architecture. A good balance between predictable incremental resistance changes and variability was achieved with a -1.2 to -1.3 V reset voltage. The linear fit in ***Figure 31*** (b) shows goodness of fit for ~20 cycles for both -1.20 and -1.25 V pulse voltage levels, with a resistance change of 8.9 to 11.2 kΩ and 9 to 12.5 kΩ, respectively. For -1.3 V reset voltage, a linear tendency can be observed for the first 10 cycles, which was fit to a linear function, as seen in Figure 30(b). The resistance change exceeds that of the lower reset voltages with a change from 9.1 to 14.2 kΩ. As the inset in ***Figure 30*** (b) shows, the variability progressively increases with a higher reset voltage from 1 to 8 kΩ averaged standard deviation. A detailed insight into the resistance response of these parameters is shown in ***Figure 31*** with (a) and (b) showing the linear fit for 30 reset pulses and the exponential fit for 100 reset pulses, respectively, for a reset voltage of -1.2 V. In general, a higher reset voltage increases the observed variability while increasing the maximum achievable R_{off}/R_{on} ratio. Furthermore, it is anticipated that an improved control of the initial LRS levels the starting resistance and could potentially reduce the overall variability during the first linear resistance changes.

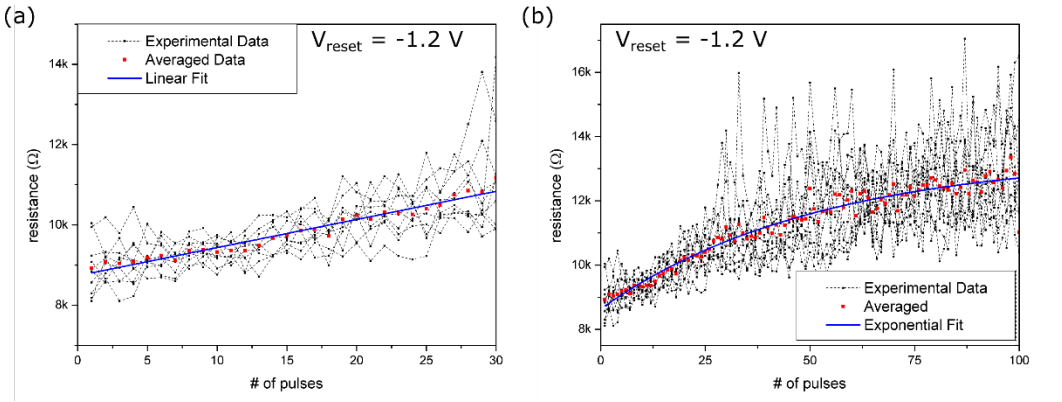

Figure 31. (a) – (b) Example resistance response to 10 cycles of 100 5ns FWHM pulses with a reset voltage of -1.1 V are shown with black dots connected by a dashed line. The red dots represent the averaged resistance values while the blue line corresponds to the linear and exponential fit of the averaged resistance values for (a) and (b), respectively.

Read/Write Verification Approach

An improvement to these initial resistance responses via relatively simple consecutive pulses in the so-called direct write approach was achieved via a write/read verification mechanism. In this case, a resistance goal was defined and reset pulses were applied until the resistance reached this goal value, at this point, a new goal was set. This loop was repeated until the ReRAM device reached 12 kΩ or 50 write/read attempt were performed without reaching the new goal value. At this point, the device was brought back into the LRS with a set operation and a new reset sequence was executed. As with the previous experiment, a controlled resistance change of the set operation was not achieved, hence, only the analog behavior of the reset operation was investigated and will be used in the subsequent simulation.

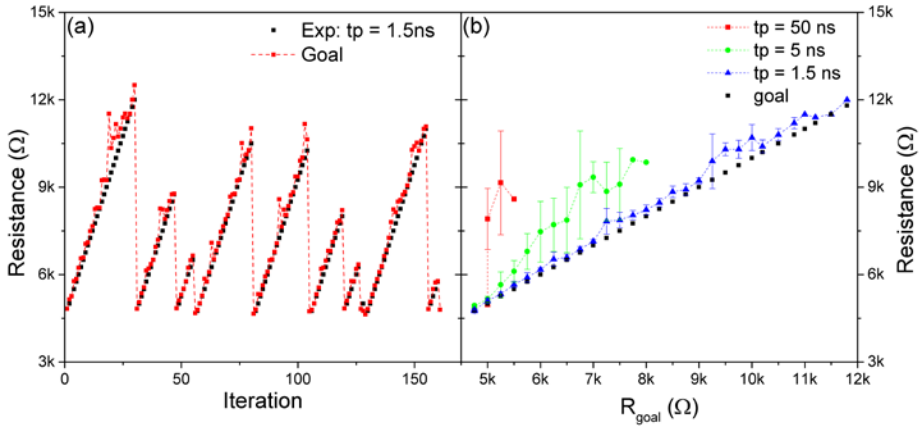

Figure 32. (a) Exemplary response is shown for an analog ReRAM response via a write/read verification approach. (b) The average response for 10 set/reset operations is shown for -1.1 V pulses with a duration of 1.5, 5 and 50 ns. The black symbols represent the goal resistances which are 250 Ω apart.

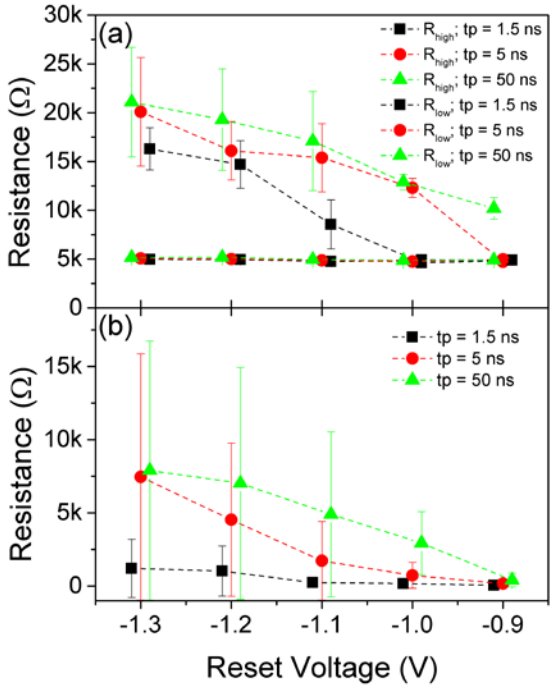

Figure 33. (a) Average low and high resistances while using different pulse conditions. The set operation was kept constant with a 2.5 V pulse amplitude, a duration of 50 µs and a current compliance of 100 µA. The reset voltage and pulse duration was varied from -0.9 to -1.3 V and from 1.5, 5 to 50 ns, respectively.

An example of the resistance response with 10 set/reset operations can be seen in ***Figure 32***(a). After a complete set with a 50 µs long pulse with an amplitude of 2.5 V and a current compliance of 100 µA, a sequence of write/read verification pulses was applied to increase the resistance in 250 **Ω** increments. The set/reset stream shown in ***Figure 32***(a) utilizes a 1.5 ns pulse with a rise/fall time of 80 ps and an amplitude of -1.2 V. The goal resistances are represented by black dots while the measured resistances are shown by red dots. To gain further inside into the statistical response of our devices, a statistical response with pulse widths of 1.5, 5 and 50 ns is shown in ***Figure 32***(b). It is clearly visible, that a more accurate response can be achieved with a reduced pulse width. However, the fluctuations in the R_{off}/R_{on} show a stronger variation with a reduction in pulse width. This can be seen in ***Figure 33***, which gives a broader insight into the pulse response of ReRAM devices, in (a) the reached high and low resistances are shown for reset voltages ranging from -0.9 to -1.3 V and for pulse width of 1.5, 5 and 50 ns. An increase in R_{off}/R_{on} ratio can be observed for increased pulse width as well as decreased reset voltage. This establishes that the response is driven by the energy inserted into the filament and that the heating of the device due to the longer applied current strongly contributes to the switching behavior. In addition, a larger reset voltage contributes to a higher final resistance but increases the variability which is supported by the results shown in ***Figure 34***. This figure shows the statistical deviation of the achieved resistances to the goal values. A clear trend is visible, the lower the reset voltage and pulse width are, the lower the statistical deviation.

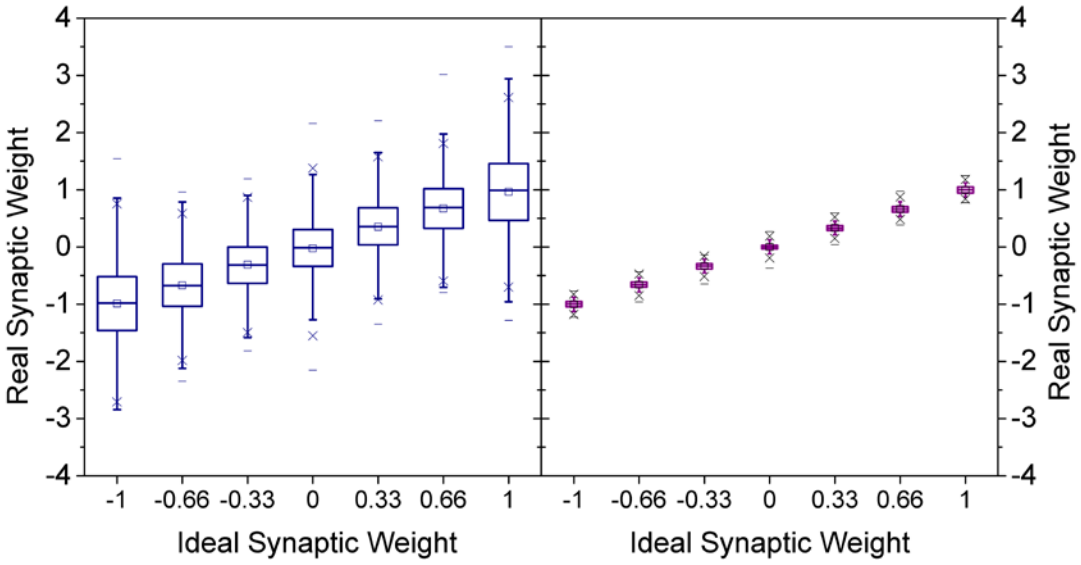

Figure 34. Distributions of the 7 different synaptic weights which can be achieved using a pair of ReRAM devices using 4 resistance levels. (a) Extracted from the best pulsing conditions (V_{reset}=-1.2V, I_{set}=100µA) with a direct write approach with 5 ns FWHM pulse without read verification. (b) Extracted from the best pulsing conditions (V_{reset}=-1.1 V, I_{set}=100µA) with read verification.

CMOR Circuit Testing

The CMOR circuit contains a logic-output element that allows its internal circuitry to be tested. This output relays the digital state of the reservoir at each cellular element. Patterns applied to the 8-bit input can then be checked to ensure they produce the correct reservoir state. Multiple versions of the CMOR circuit implementing different CA rules were included on the mask set (***Figure 35***), and each was tested and confirmed as operating correctly (***Figure 36***).

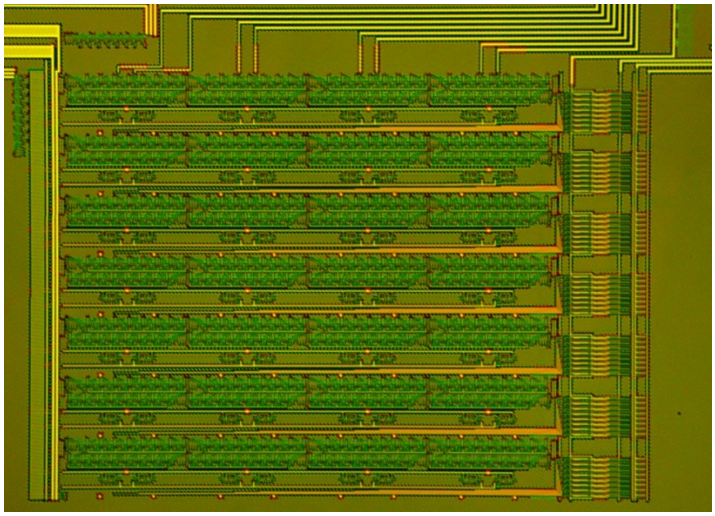

Figure 35. Micrograph of fabricated CMOR circuit.

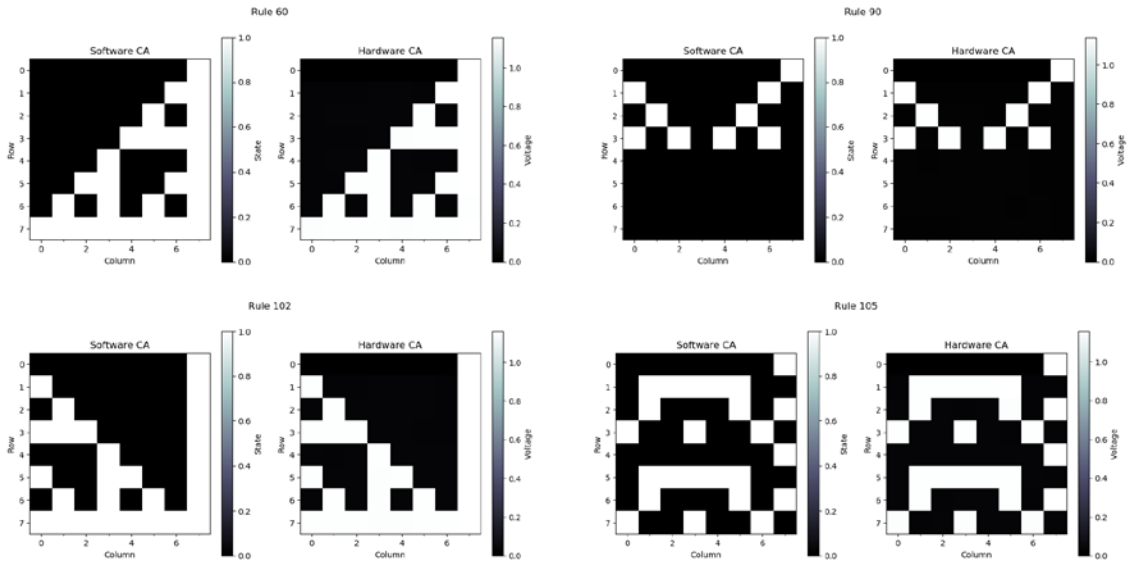

Figure 36. Comparison of the output expected from test patterns as computed in software and using the CMOR circuit. All patterns match exactly, showing that the logic has been implemented correctly.

To test the output layer which provides different resistance levels for internal reservoir states, memristors at three locations were formed into the low-resistance state. Then, by sweeping through all possible reservoir inputs and measuring the output layer's resistance, we observe the number of produced resistance levels. With memristors programmed at this location, we observed 3 distinct resistance levels through the output layer. Having confirmed the basic logical operation of the reservoir and its output layer, further testing can probe into more advanced programming of the output layer to classify non-linear inputs.

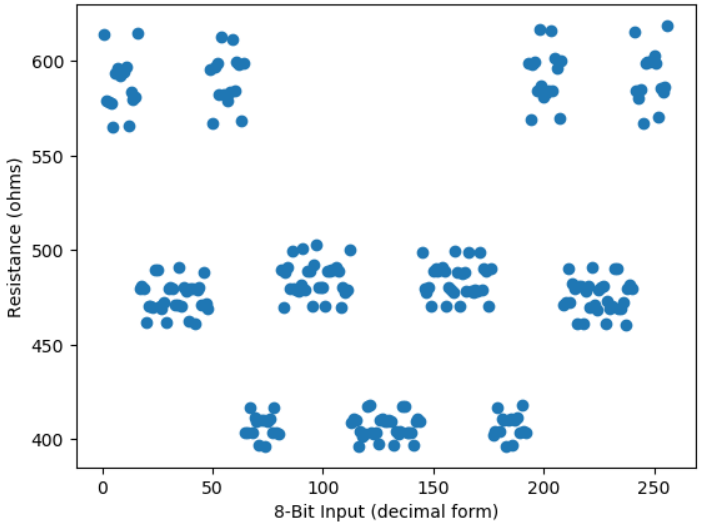

Figure 37. Resistance measured through the memristive output layer after 3 memristors have been set. The activation and deactivation of memristors contributing to the output current creates the collective resistance state. This is modulated as the input changes, enabling and disabling different elements.

4.1.3 Memristive Adaptations of Networks

Having established theoretical and experimental evidence that noisy neurons can be more robust to inaccurately-programmed weights, we wished to investigate whether this robustness is sufficient to withstand the variabilities and limitations expected in real synaptic devices. One of the limitations of synaptic devices is that within their dynamic range, variability in programming will limit the number of distinguishable states that can be reliably achieved. Each of these states may also still contain a variability. Currently, we aimed to achieve 4 reliable resistance levels with our devices.

We use a pulse-based system to gradually reset devices from the LRS up to a desired value, verifying that approximately the correct resistance value is achieved with intermediate read pulses. This allows us to program intermediate states between a 2.75 kΩ LRS and 11.5 kΩ HRS when using our transistor-integrated hafnium-oxide based memristors.

Within our neuromorphic system which utilizes memristors as synapses, a pair of memristors is actually used to represent a single synapse. This is due to the aforementioned asymmetric programming behavior of VCM memristors; to be able to decrement or increment the synaptic weight gradually when devices can only be reset (and not set) gradually, two are used so that one can be used to increment (and the other decrement) weights. Generally, one device is referred to as being 'excitatory' and the other 'inhibitory' (***Figure 38***).

When using 2 devices in this differential-pair configuration, *n* analog device levels yield *(2n-1)* weight levels, giving 7 possible synaptic weight levels from a 4-level memristor. Some of these values can be represented in multiple ways using the pair of memristors (***Figure 38***b), and each of these representations is sampled equally when establishing the distributions for variability at each weight level.

For each representation, a sample of a memristor's resistance at that level is sampled from a normal distribution determined by electrical testing (***Table 1***). These states were selected to provide a coverage across the memristor's achievable dynamic range,and achieve mean programmed values which were evenly spaced as possible; this prevents the achieved weights from being strongly biased from desired values.

Table 2. Distributions of resistance values for the 4 selected resistance states of the HfOx memristor.

Resistance Level	Targeted Value (ohm)	Mean, Programmed Value (ohm)	Std. Dev, Programmed Value (ohm)
1	2750	2790	41.3
2	5000	5610	425
3	7750	8380	479
4	10500	11300	617

The distribution of synaptic weights achieved by these memristor representations was calculated by drawing 1,000 samples at each of the 7 weight levels (***Figure 38***c,d). We find that the pulsed reset/verify programming method is crucial to achieving accurate weights, but with this method reliably distinguishable distributions of weight values can be achieved.

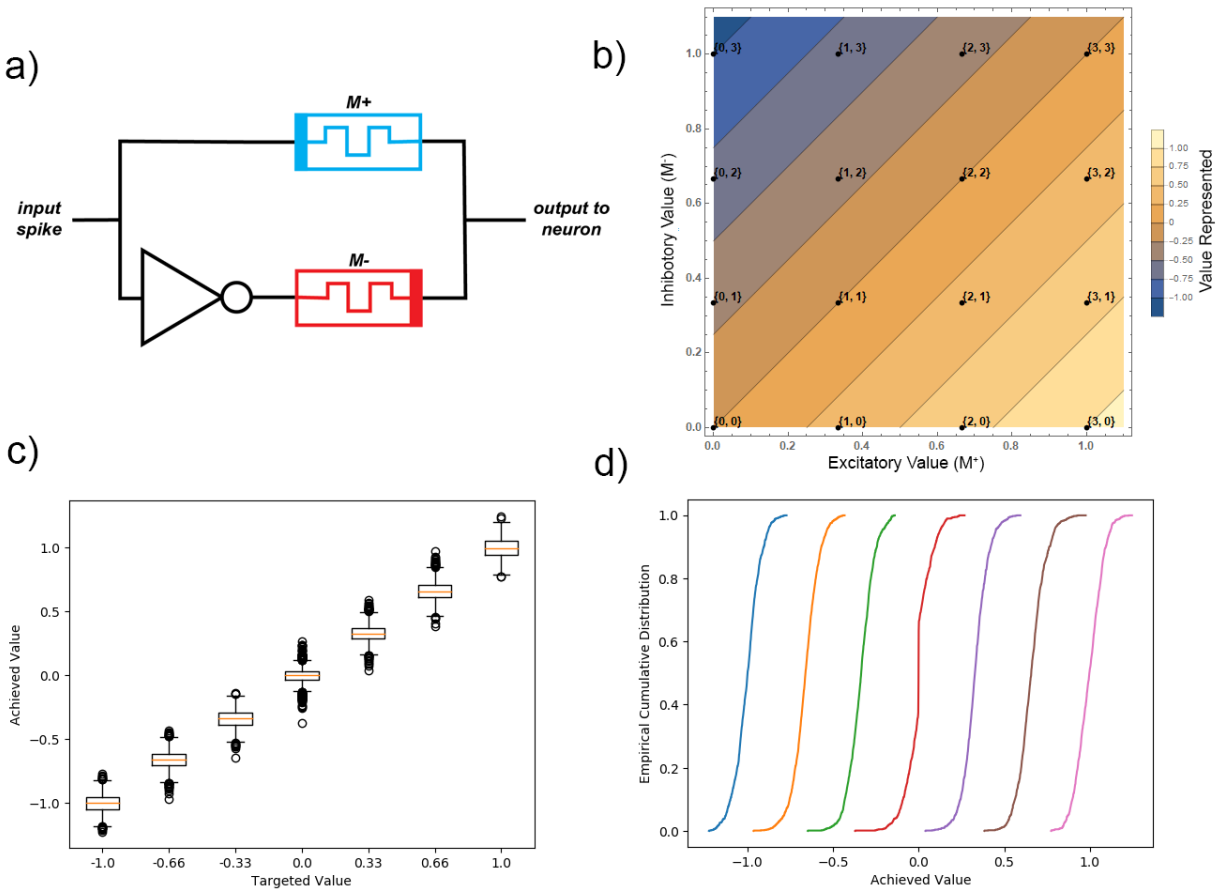

Figure 38. A pair of excitatory/inhibitory (M^+/M^-) memristors (a) is used to represent a synapse so that weights can be evenly modulated up or down even when the synaptic device has asymmetric programming characteristics (such as in VCM ReRAM). In this 'twin synapse,' each value can have multiple representations (b), and the representations for each value given a 4-level device are shown. Assuming an equal probability that each representation is used, the variability of the 7 resulting weight values using pulse-programmed HfOx memristors is shown (c,d).

Spiking Networks under Memristive Variability

From the pole-balancing networks which were evolved to establish the robustness of networks to weight perturbation, we sub-selected two exemplary networks which could also withstand a reduction to 7 synaptic weight levels (4 device levels). These networks maintained full performance when weights were rounded to the nearest representable level, allowing them to be fully adapted by synapses that could perfectly represent these 7 levels. One of these networks utilized perfect integrate and fire (I&F) neurons, and the other used noisy I&F neurons.

The variability of the weights in these networks was gradually raised to the level expected using real HfOx memristors, with network fitness at each step sampled 100 times. This established the range of performance degradation a perfect or noisy I&F network might exhibit when transferred to hardware using memristors as synaptic devices. As expected, the behavior of the network using perfect I&F neurons become much more variable and the median fitness value degrades as weight variability increases (***Figure 39***a). In contrast, the network using noisy I&F

neurons maintains a median performance value at the maximum value even under the full variability levels expected from a real memristor (***Figure 39***b). We believe that this provides evidence that the view of neurons as stochastic computing elements is a powerful perspective which can be used to construct robust networks that can utilize inaccurate or unreliable elements. This provides motivation to consider more stochastic designs for future neuromorphic systems and optimization/simulation frameworks.

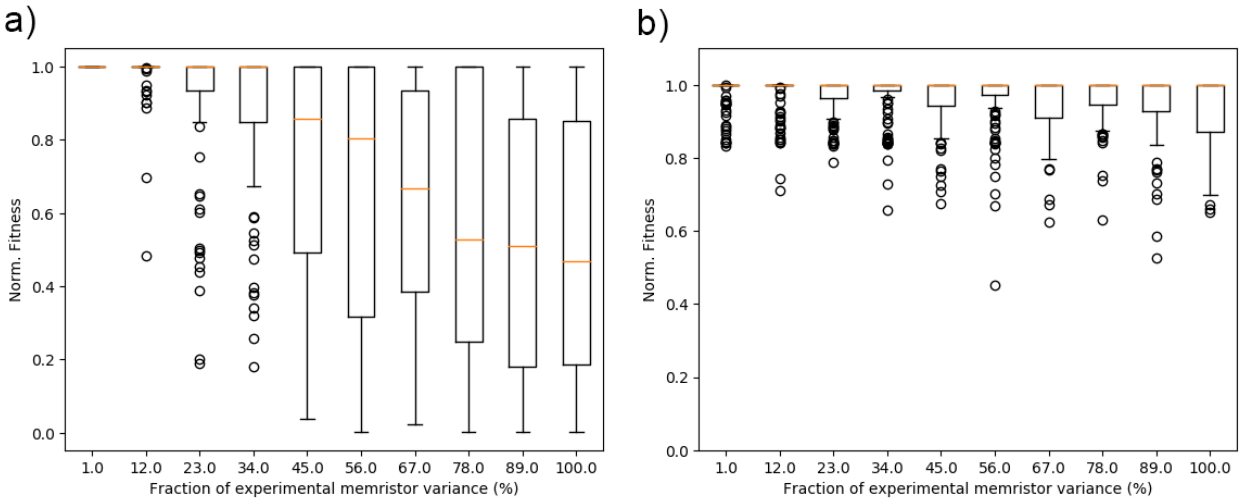

Figure 39. The performance of networks using perfect (a) and noisy (b) I&F neurons carrying out a pole-balancing task as synaptic variability increases to levels expected under memristor implementation.

5.0 Conclusions

In this effort, the goal was to fabricate hybrid CMOS/ReRAM elements and circuits that can be leveraged for spiking neural network approaches to neuromorphic computing applications. The SUNY Poly team has demonstrated a successful design, tape-out and fabrication process to yield such devices and circuits in a 65nm CMOS process, on a 300mm wafer platform. Our fabrication process exhibits good CMOS yield and nearly 100% ReRAM yield across a 300mm wafer. ReRAM devices have strong performance parameters with respect to forming, set, and reset voltages, as well as LRS, HRS and memory window. In this effort and previous efforts, we have demonstrated endurance up to 1E11 switching cycles, which should be more than sufficient for encoding synaptic weights in neuromorphic circuits. Further, we have demonstrated that we can control 1T1R ReRAM cells in an analog fashion, with respect to both HRS and LRS, by controlling either the peak current during switching, or pulse height/width. We have also implemented a variety of methods to control the stochastic nature of ReRAM resistive switching, and can use a read-verify approach to set devices into a specific resistance state. This makes them highly amenable for encoding synaptic weights in neuromorphic circuits / systems.

In addition to fabricating and characterizing hybrid CMOS/ReRAM circuit elements, we also implemented and tested a simple reservoir circuit. Our initial experiments with this circuit show that the underlying CMOS circuitry is functional and that we can encode ReRAM output elements (into various resistance states) with this circuit. Thus, we are primed for follow-on studies of this circuit, including primitive training regimes.

Beyond hardware, we have also performed a variety of simulations to better understand how the inherent variability of ReRAM (especially HRS) can be factored into the ultimate neuromorphic circuit design and training. Our simulations show that intentionally training ReRAM-based neuromorphic circuits with so-called “noisy” integrate and fire neurons can result in robust neural networks that can tolerate variation in ReRAM-encoded synaptic resistance levels, far better than “perfect” non-noisy training neurons.

In summary, the hardware development effort of the mrDANNA project has been successful and we are eager to continue work with our collaborators at UT-Knoxville, to evaluate the evaluation/demonstration circuits that were designed, as well as the fully digital version of dynamic adaptive neural network array (DANNA). At this time we are continuing to process 300mm wafers beyond the metal 3 (M3) layer, up through additional metallization layers that will enable full testing of the digital DANNA array.

6.0 Publications and Patent Applications Resulting from this Project

Peer-Reviewed Journal Articles

1. W. Olin-Ammentorp, **N.C. Cady.** The Motivation, Principles, and State of Neuromorphic Computing. *Submitted to Science Progress, January 2019. Under review.*
2. W. Olin-Ammentorp, **N.C. Cady.** Non-parametric Testing of Discrete Transfer Entropy via a Markov Chain Monte Carlo Method. *Submitted to Journal of Neuro. Methods, December 2018. Under review.*
3. M. Uddin, M.B. Majumder, K. Beckmann*, H. Manem*, Z. Alamgir*, **N.C. Cady,** G.S. Rose. Design considerations for memristive crossbar physical unclonable functions. (2018) *ACM Journal on Emerging Technologies in Computing Systems.* 14(1), 2.
4. K. Beckmann, J. Holt, W. Olin-Ammentorp, Z. Alamgir, J. Van Nostrand, **N.C. Cady.** The effect of reactive ion etch (RIE) process conditions on ReRAM device performance. (2017) *Semiconductor Science and Technology.* 32: 095013
5. Z. Alamgir, J. Holt, K. Beckmann, **N.C. Cady.** The effect of different oxygen exchange layers in TaOx based RRAM devices. (2017) *Semiconductor Science & Technology.* 33: 015014
6. Z. Alamgir, K. Beckmann, J. Holt, **N.C. Cady.** Pulse width and height modulation for multi-level resistance in bi-layer TaOx based RRAM. *Applied Physics Letters.* (2017) 111: 063111 DOI: http://dx.doi.org/10.1063/1.4993058
7. J.S. Holt, K. Beckmann, Z. Alamgir, J. Yang-Scharlotta, **N.C. Cady.** Effect of displacement damage on tantalum oxide resistive memory. (2017) *MRS Advances.* 1-7. DOI:10.1557/adv.2017.422
8. K. Beckmann, H. Manem, **N.C. Cady**. Performance enhancement of a time-delay PUF design by utilizing integrated nanoscale ReRAM devices. (2016) *IEEE Transactions on Emerging Topics in Computing – Special issue "Security of Beyond CMOS Devices: Issues and Opportunities. DOI: 10.1109/TETC.2016.2575448*
9. K. Beckmann, J. Holt, W. Olin-Ammentorp, J. Van Nostrand, **N.C. Cady**. Impact of etch process on hafnium dioxide based nanoscale RRAM devices. (2016) *ECS Transactions.* 75(13): 93-99.
10. K. Beckmann, J. Holt, H. Manem, J. Van Nostrand, **N.C. Cady.** Nanoscale hafnium oxide RRAM devices exhibit pulse dependent behavior and multi-level resistance capability. (2016) *MRS Advances.* 1(49): 3355-3360. *DOI:* http://dx.doi.org/10.1557/adv.2016.377

Peer-Reviewed Conference Proceedings Papers

11. **N.C. Cady,** K. Beckmann, W. Olin-Ammentorp*, J.E. Van Nostrand, G. Chakma, R. Weiss, S. Sayyaparaju, M. Adnan, J. Murray, M.E. Dean, J.S. Plank, G.S. Rose**.** Full

CMOS-Memristor Implementation of a Dynamic Neuromorphic Architecture. *GOMACTECH Conference, Miami, FL. March 2018.*

12. J.S. Plank, G.S. Rose, M. E. Dean, C.D. Schuman, **N.C. Cady**. A Unified Hardware/Software Co-Design Framework for Neuromorphic Computing Devices and Applications. ICRC: IEEE International Conference on Rebooting Computing. November 2017, Washington, DC.
13. S. Amer, S. Sayyaparaju, K. Beckmann*, **N.C. Cady,** G.S. Rose. A Practical Hafnium-Oxide Memristor Model Suitable for Circuit Design and Simulation. ISCAS: International Symposium on Circuits and Systems. May 2017, Baltimore, MD.
14. S. Amer, G.S. Rose, K. Beckmann*, **N.C. Cady.** Design Techniques for in-Field Memristor Forming Circuits. 60th IEEE International Midwest Symposium on Circuits and Systems. August 2017, Boston, MA.
15. M. Uddin, M.B. Majumder, G.S. Rose, K. Beckmann*, H. Manem*, Z. Alamgir*, **N.C. Cady.** Techniques for Improved Reliability in Memristive Crossbar PUF Circuits. 2016 IEEE Computer Society Annual Symposium on VLSI (ISVLSI), Pittsburgh, PA, pp. 212-217. *doi: 10.1109/ISVLSI.2016.33*
16. W. Olin-Ammentorp*, K. Beckmann*, J.E. Van Nostrand, G.S. Rose, M.E. Dean, J.S. Plank, G. Chakma, **N.C. Cady.** Applying Memristors Towards Low-Power, Dynamic Learning for Neuromorphic Applications. *GOMACTECH Conference, Reno, NV March 2017.*
17. G. Chakma, M.E. Dean, G.S. Rose, K. Beckman*, H. Manem*, **N. Cady**, A Hafnium-Oxide Memristive Dynamic Adaptive Neural Network Array. International Workshop on Post-Moore's Era Supercomputing (PMES), *Salt Lake City, UT, November 2016.*
18. Z. Alamgir*, K. Beckmann*, **N.C. Cady,** A. Velasquez, S.K. Jha. Flow-based computing on nanoscale crossbars: design and implementation of full adders. *International Symposium on Circuits and Systems (ISCAS) Conference, May 2016, Montreal, Canada.*
19. **N.C. Cady,** K. Beckmann*, H. Manem*, M.E. Dean, G.S. Rose, J.E. Van Nostrand. Towards Memristive Dynamic Adaptive Neural Network Arrays. *GOMACTEC Conference, March 2016, Orlando, FL.*

Patents

1.“Resistive Random Access Memory Device.” U.S. Provisional Patent Application, November 2018.

2.“Selector Devices for a Memory Cell” U.S. Provisional Patent Application, November 2018.

7.0 List of Acronyms

RRAM: Resistive random access memory

ReRAM: Resistive random access memory

RMD: Resistive memory device

TE: Top electrode

BE: Bottom electrode

HRS: High resistance state

LRS: Low resistance state

I&F: Integrate and fire

Vg: Gate voltage

1T1R: 1 transistor 1 memristor (memory cell containing 1 transistor and 1 memristor)