AFRL-RI-RS-TR-2023-004

FABRICATION OF EFFICIENT RECONFIGURABLE NEUROMORPHIC SYSTEMS

THE RESEARCH FOUNDATION FOR STATE UNIVERSITY OF NEW YORK ON BEHALF OF SUNY POLYTECHNIC INSTITUTE

COLLEGES OF NANOSCALE SCIENCE & ENGINEERING

JANUARY 2023

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

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

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-2023-004 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE CHIEF ENGINEER:

/ S /
JOSEPH E. VANNOSTRAND
Work Unit Manager

/ S /
GREGORY J. HADYNSKI
Assistant Technical Advisor
Computing & Communications Division
Information Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED START DATE	END DATE
JANUARY 2023	FINAL TECHNICAL REPORT	FEBRUARY 2019	AUGUST 2022

4. TITLE AND SUBTITLE
FABRICATION OF EFFICIENT RECONFIGURABLE NEUROMORPHIC SYSTEMS

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
	FA8750-19-1-0014	61102F
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**
		R2R5

6. AUTHOR(S)
Nathaniel Cady

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
The Research Foundation for State University of New York on behalf of SUNY
Polytechnic Institute Colleges of Nanoscale Science & Engineering
257 Fuller Road
Albany NY 12203

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER(S)
Air Force Research Laboratory/RITB 525 Brooks Road Rome NY 13441-4505	AFRL/RI	AFRL-RI-RS-TR-2023-004

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 goal of this effort was to fabricate hybrid CMOS/RRAM chips that could be used for a variety of Air Force specific applications, using a low-power, reconfigurable framework. Working with collaborators at UT-Knoxville and a number of other institutions, the SUNY Poly team was able to compile a design containing an unique neuromorphic, reconfigurable test chip (UT-Knoxville design), a variety of hybrid CMOS/RRAM demonstration circuits (from UT-Austin, ASU), and a neuromorphic computing demonstrator chip (from UT-San Antonio). In addition, the SUNY Poly team designed multiple memory arrays for demonstration of compute in memory operations and analog information encoding. To this end, the SUNY Poly team also developed optimized RRAM testing parameters to 1) maintain high device yield and endurance, 2) enable a large device memory window, and 3) perform analog information encoding using sub-nanosecond pulses. Together, this work provides a framework for ongoing testing and evaluation of hybrid CMOS/RRAM circuits by the SUNY Poly team and collaborators. Full testing of the circuits and chips fabricated under this effort is underway at the partnering institutions and will continue under the OUSD "NeuroPipe" ARAP program.

15. SUBJECT TERMS
Memristor, FPGA, CMOS, nanoelectronics

16. SECURITY CLASSIFICATION OF: a. REPORT	b. ABSTRACT	C. THIS PAGE	17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
U	U	U	SAR	40

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
JOSEPH E. VANNOSTRAND	**N/A**

Table of Contents

1.0 SUMMARY **1**
2.0 INTRODUCTION **2**
3.0 BACKGROUND **3**
4.0 TECHNICAL OVERVIEW OF PROPOSED "RAVENS" SYSTEM **5**
4.1 TECHNICAL APPROACH UTILIZED 6
5.0 METHODS, ASSUMPTIONS AND PROCEDURES **7**
5.1 FABRICATION 7
5.2 DEVICE TESTING 8
6.0 RESULTS AND DISCUSSION **11**
6.1 KEY ACCOMPLISHMENTS 11
6.1.1 Optimization of memristor performance and validation of hybrid CMOS/memristor hardware *11*
6.1.2 Fabrication of fully reconfigurable neuromorphic hardware using hybrid CMOS/memristor architecture *23*
7.0 CONCLUSIONS **31**
8.0 PUBLICATIONS AND PATENT APPLICATIONS RESULTING FROM THIS PROJECT **32**
9.0 LIST OF ACRONYMS **34**

Approved for Public Release; Distribution Unlimited.

List of Figures

Figure 1. The NIDA/DANNA/mrDANNA approach can utilize a simulated training environment to generate a sparse, lightweight neural network, which has been successfully implemented for autonomous control of robotic vehicles. 4
Figure 2. System level view of the proposed reconfigurable neuromorphic array, complete with spiky neuromorphic cores (nCores) and dot product engines (DPEs). 6
Figure 3. Global view of the neuromorphic application structure to be implemented in RAVENS. 6
Figure 4. Cross-sectional contrast image of our integrated CMOS/RRAM circuit taken with a transmission electron microscope. This cross-section shows the W bottom electrode and RRAM device, as fabricated in a FEOL-compatible process, as well as an inset showing the elemental mapping of the key layers of the RRAM device (based on EDS mapping in the TEM). 8
Figure 5. 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). 9
Figure 6. (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. 9
Figure 7. a) TEM micrograph of 1T1R integration of HfO_2 RRAM devices b) cross-section of RRAM stack, the inset displays EDS map of different elements. Yellow, blue and purple represent Hf, O, and Ti respectively. 12
Figure 8. Schematic of typical pulse switching cycle to study switching endurance of RRAM devices. 12
Figure 9. Resistance with switching variation for different a) LRS and b) HRS resistance values for 100+ tested RRAM cells. Memory window with normalized switching variation for c) LRS and d) HRS resistances. 13
Figure 10. Cumulative frequency plot of a) LRS and b) HRS values with maximum operating current I_{max}. Impact of I_{max} on c) median LRS/HRS resistances d) memory window, and e) LRS/HRS switching variation. 14
Figure 11. a) Switching endurance of fabricated RRAM cell for different RESET pulse width t_{RESET} and amplitude V_{neg}. Impact of t_{RESET} on b) memory window, and c) HRS switching variation for different V_{neg} values. 15
Figure 12. a) Impact of SET pulse width t_{SET} on a) LRS, b) HRS, c) memory window, and d) HRS switching variation for different SET amplitude V_{pos} values. 16
Figure 13. Impact of V_{pos} on a) memory window, b) LRS, and c) HRS switching variation for different RESET amplitude V_{neg} values. Switching endurance of a RRAM cell for d) "unbalanced", and e) "balanced" SET/RESET. 17
Figure 14. a) Switching endurance up to 1B cycle, and b) corresponding HRS and LRS cumulative frequency plot of optimized 1T1R cell. c) Data retention for two different HRS and LRS resistance levels at 373K. 18
Figure 15. Full 300 mm wafer statistics for a) memory window, b) cycle-to-cycle HRS switching variability, c) cell-to-cell HRS switching variability d) switching yield, and e) memory window vs. normalized HRS switching variation using optimized operation conditions: I_{max} = 100 μA, V_{neg} = -1.6 V, t_{RESET} = 100 ns, V_{pos} = 2 V, t_{SET} = 100 ns. Inset in Fig. 9e) shows pre-optimization plot for memory window with HRS switching variation. 19
Figure 16. Potentiation and depression of a hafnium oxide based 1T1R cell using a series of uniform 300 ps pulses (100 uA operating current and constant set/reset voltage). 20

Approved for Public Release; Distribution Unlimited.

Figure 17. Application of variable pulse height (set or reset voltage) results in improved linearity and symmetry of conductance change through the potentiation and depression cycle for hafnium oxide 1T1R cells. 21
Figure 18. Image of AFRL-RI probe card contact with SUNY Polytechnic 8x8 1T1R array pads (performed at AFRL-RI, Rome, NY). 22
Figure 19. Successful encoding of 1T1R array with high and low resistance states using a SUNY Polytechnic 8x8 1T1R array at AFRL-RI. Raw resistance data is shown at left, while a threshold-adjusted image of the array data is shown at right. 22
Figure 20. (A) Successful linear sweep based switching of HfOx 1T1R devices (from the SUNY Polytechnic team) made by AFRL-RX (Dayton, OH). (B) Example of a 1T1R cell prepared for in situ electrical switching in transmission electron microscope (TEM) at AFRL-RX to observe conductive filament formation / rupture during the electroform/set/reset processes 23
Figure 21. (A) Full-field reticle view of the tape-out with the UTK "RAVENS" processor, as well as designs from other collaborating groups. (B) Image of full 300mm RAVENS wafer and (C) close-up image of an individual RAVENS die. 25
Figure 22. Labeled sub-die map of the RAVENS design, showing the location of various designs on the reticle. 26
Figure 23. Sub-designs from collaborating research groups. 26
Figure 24. SUNY Polytechnic designs from the RAVENS tape-out. Designs include memory arrays up to 1MB (with decoder circuit), neurons for demonstrating RRAM-based learning behavior, a 64x64 1T1R array for performing vector matrix operations, a 12x12 RRAM array (no transistors), and finally novel 8x8 1T1R arrays with deep well transistors for enabling RRAM switching with either polarity 27
Figure 25. 30keV STEM image from the center of the first fully build RAVENS test chip. Visible are the transistors (bottom layer), four metallization layers (M1-M4) and BA. The two upper most layers (BB, LB) did not fit into the frame. 28
Figure 26. 30keV STEM image from the center of the first fully build RAVENS test chip. Visible are three ReRAM elements embedded between the tungsten M1 and copper M2. In addition, the M1 lines are connected to dummy poly silicon lines via CA. 29
Figure 27. Exemplary inline test data for a 5x minimum width low threshold voltage NFET. Top left, top right, and bottom show on current, off current and threshold voltage, respectively 30
Figure 28. Initial edge-center-edge pulse-based analysis of 1T1R test structures. 11 dies with each 5 devices for a total of 55 devices were measured. The devices were 10000 times cycled after an initial forming pulse. A rise and fall time of 10 µs was with +5, +2 and -1.5 V for the forming, set and reset voltages. The current during the forming and set event was limited to 100 µA. 30

Approved for Public Release; Distribution Unlimited.

1.0 Summary

This technical report summarizes the R&D efforts for the AFRL project "FABRICATION OF EFFICIENT RECONFIGURABLE NEUROMORPHIC SYSTEMS," with the primary outcome being fabrication of a neuromorphic test chip named "RAVENS". The ultimate objective of this work was to develop approaches that lead to low-power, reconfigurable, high-efficiency brain-inspired computing hardware for embedded control applications, and other applications dealing with dynamic and spatio-temporal data streams. The application domains that frame this work are specifically related to resource constrained systems with limited size, weight, and power (SWaP), necessitating lightweight approaches for implementing brain-inspired, neuromorphic networks. This was achieved by leveraging our ongoing work on memristive dynamic adaptive neural network arrays (mrDANNA). Building on this effort, we selected optimally performing neuron/synapse configurations (CMOS/memristor circuits) to design and fabricate a fully functional memristive neuromorphic processor capable of efficiently implementing command/control, navigation and avoidance, as well as other spatio-temporal data processing applications. Memristors (or "memory resistors") are nanoscale electronic switching devices, leveraged here in combination with more conventional CMOS technology to maximize circuit density and reduce overall energy consumption. For this effort, the SUNY Polytechnic Institute team (this report) focused on optimizing CMOS/memristor circuit performance, aggregating designs from our partner institution, the University of Tennessee – Knoxville, and fabricating the novel memristive-enhanced neuromorphic processor design and associated test circuits using the SUNY Polytechnic Institute / Albany Nanotech 300mm wafer scale foundry.

2.0 Introduction

In recent years, the semiconductor industry has begun to experience a significant slowdown in the performance improvements gained from technology scaling. While this is due in part to the impending end of Moore's Law scaling, power consumption has also become a critical limiting factor for the level of performance achievable. The research proposed here aims to enable future generations of computing systems by (1) leveraging an emerging, power-efficient device technology (i.e. the memristor) and (2) considering an alternative architectural model (i.e. neuromorphic) that promises to overcome many of the performance limitations of conventional von Neumann systems. It is worth noting that neuromorphic or neuro-inspired computer architectures are particularly worthwhile given the increasing number of big data problems requiring techniques and systems that can filter knowledge from an abundance of data. Furthermore, it is important to also consider such neuromorphic systems through the lens of the ubiquitous computer systems, including the internet of things (IoT) paradigm and deployable, autonomous systems. Thus, the proposed Reconfigurable and Very Efficient Neuromorphic System (RAVENS) platform aims to be an energy-efficient neuromorphic architecture specifically tailored for control and other spatio-temporal applications commonly implemented with resource constrained computer systems.

The ultimate objective of this work was to develop approaches that lead to low-power, reconfigurable, high-efficiency brain-inspired computing hardware for embedded control applications, and other applications dealing with dynamic and spatio-temporal data streams. The application domains that frame this work are specifically related to resource constrained systems with limited size, weight, and power (SWaP), necessitating lightweight approaches for implementing brain-inspired, neuromorphic networks, achieved by leveraging our ongoing work on memristive dynamic adaptive neural network arrays (mrDANNA).[1] Building on this effort, we planned to optimize neuron/synapse configurations (CMOS/memristor circuits) and device performance, and to work with our collaborators at UT-Knoxville to design and fabricate a fully functional memristive neuromorphic processor capable of efficiently implementing command/control, navigation and avoidance, as well as other spatio-temporal data processing applications. Memristors (or "memory resistors") are nanoscale electronic switching devices, leveraged here in combination with more conventional CMOS technology to maximize circuit density and reduce overall energy consumption.

[1] G. Chakma, M.M. Adnan, A.R. Wyer, R. Weiss, C.D. Schuman, and G.S. Rose, "Memristive Mixed-Signal Neuromorphic Systems: Energy-Efficient Learning at the Circuit-Level," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems (JETCAS)*, available online, November 2017. (DOI: 10.1109/JETCAS.2017.2777181)

3.0 Background

The specific neuromorphic approach on which the proposed work is based is the Neuroscience-Inspired Dynamic Architecture (NIDA)[2,3,4], 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) dynamic run-time adaptation/learning, 3) a tendency toward recurrent neural networks, and 4) a synaptic representation that includes delay and weight information. This work builds off of our prior efforts to develop a NIDA-based memristive dynamic adaptive neural network array (mrDANNA), which in turn, builds upon established digital (FPGA) and software-based dynamic neural networks based on the our dynamic neural network array (DANNA) design. These efficient neural network arrays are distinct from deep learning neural networks, in that they are more efficient (using fewer neurons, fewer synapses, smaller total architecture), are better suited for command/control, dynamic datasets, and spatio-temporal data[5]. Existing hardware-based approaches to neuromorphic computing use power-intensive (eg. FPGAs and graphics accelerators) or custom all-CMOS chips (eg. IBM's TrueNorth chip) that are limited in their ability to dynamically respond or undergo dynamic training/optimization.

In the work performed under this program, we took the concept of a "purpose-built" reservoir and realized it in hardware. Specifically, we aimed to develop a CMOS/memristive system (Fig. 1) with an array of spiky, reconfigurable neuromorphic cores that can be configured to implement a mrDANNA network for a given neuromorphic application. This spiky array could then interface with classifier stages built from either CMOS/memristive or all-CMOS dot product engines (DPEs), providing the ability to interact with convolutional stages around the mrDANNA core. This arrangement of classifiers surrounding the core is not different from reservoir computing except in the proposed system the core is trained via EONS (Evolutionary Optimization for Neuromorphic Systems) for a particular application. Further, we will include online training techniques based on spike time-dependent plasticity (STDP) such that the system adapts to its environment in real-time. Altogether, the proposed system is expected to realize efficient networks (RNNs) using area and energy efficient hardware, specifically nanoscale memristors. Thus, the proposed system provides a suitable framework for implementing neuromorphic applications for resource constrained embedded systems.

Neuromorphic-specific hardware is increasingly seen as the 1) best way to reduce the need for intensive data transmission during command/control operations, 2) provide autonomous operation, or a sub-set of autonomous functions, and 3) reduce the need for human interaction/oversight of deployed systems. What is currently unavailable, however, is neuromorphic hardware that meets the strict SWaP requirements to make it feasible for field-based and flight-based systems (robotics, UAV, forward controllers, satellite, aircraft, etc.). The proposed work would fill this gap – providing low-power, highly efficient, small size neuromorphic hardware that is distinctly optimized for handling evolving, dynamic data streams.

---

[2] C. D. Schuman and J. D. Birdwell, "Variable structure dynamic artificial neural networks," Biologically Inspired Cognitive Architectures, vol. 6, no. 0, pp. 126–130, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2212683X13000364

[3] Schuman, C.D., Birdwell, J.D, "Dynamic artificial neural networks with affective systems," PLoS ONE, vol. 8, no. 11, p. e80455, 11 2013.

[4] C. D. Schuman, J. D. Birdwell, and M. E. Dean, "Spatiotemporal classification using neuroscience-inspired dynamic architectures," Procedia Computer Science, vol. 41, pp. 89—97, 2014.

[5] C.D. Schuman, T.E. Potok, S.R. Young, R. Patton, G. Perdue, G. Chakma, A. Wyer, G.S. Rose. "Neuromorphic Computing for Temporal Scientific Data Classification," in *Proceedings of Neuromorphic Computing Symposium: Architectures, Models, and Applications* (*NCAMA*), Knoxville, TN, July 2017.

As an example of previous work by the UT-Knoxville team, **Figure 1** shows a simulated environment for training a NIDA/DANNA network, implementing this network, and then using the network for autonomous navigation of a robotic vehicle[6].

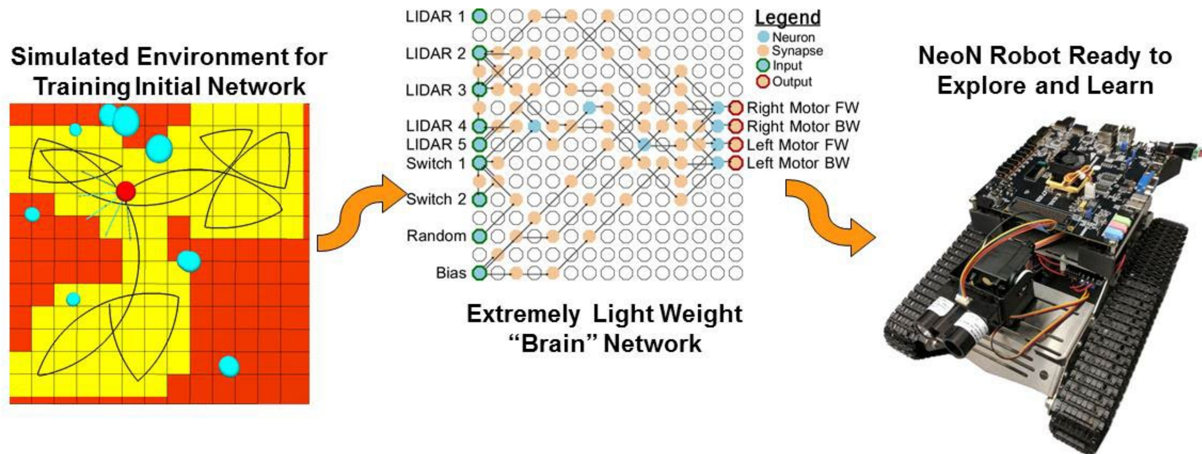


Figure 1. *The NIDA/DANNA/mrDANNA approach can utilize a simulated training environment to generate a sparse, lightweight neural network, which has been successfully implemented for autonomous control of robotic vehicles.*

[6] J.P. Mitchell, G. Bruer, M.E. Dean, J.S. Plank, G.S. Rose, and C.D. Schuman, "NeoN: Neuromorphic Control for Autonomous Robotic Navigation," in *Proceedings of the International Symposium on Robotics and Intelligent Sensors*, Ottawa, Canada, October 2017.

4.0 Technical Overview of Proposed "RAVENS" System

Under this project, we proposed to design and fabricate a CMOS/memristor based neuromorphic system to demonstrate the feasibility of low-power, reconfigurable, high-efficiency, brain-inspired computing hardware for embedded control and dynamic data-stream applications. The nickname we proposed for this system was the Reconfigurable and Very Efficient Neuromorphic System (aka: RAVENS). This project builds on previous efforts by the SUNY Poly (PI-Cady) and UT-Knoxville (PIs-Rose, Dean, Plank) teams in which we developed the mrDANNA architecture and hardware, which is enabling as a unique dynamic adaptive neural network array (DANNA), which is fundamentally different than deep neural networks (DNNs). DNN style systems require 1,000's to 10,000's of nodes/neurons to solve spatio-temporal problems. Notably, the DANNA and mrDANNA architectures (via prior AFRL support) have been established as a baseline for reconfigurable, spiky recurrent neuromorphic networks in hardware. What we propose in this effort is to fully implement the mrDANNA approach, which includes unique CMOS/memristor "neurons", dense crossbar-based memristors, I/O layers, and an interconnect fabric. The proposed system level architecture for RAVENS is shown in **Figure 2**, while an example of the application flow for this approach is shown in **Figure 3**.

The reconfigurable neuromorphic array consists of "spiky" neuromorphic cores (nCores – based on our previous mrDANNA architecture) and a number of dot product engines (DPEs). This overall architecture leverages our mrDANNA neuromorphic computing approach in which neural plasticity is enabled via multi-level memristive synapses, to enable online training and adaptability. The low power memristive synapses and DPEs, along with event-driven computation will enable significant SWaP savings. Furthermore, the proposed architecture combines feed-forward neural networks with high recurrent "spiky" cores to leverage the functionality of each technology.

The classifier or DPE layers around the spiky cores can also act as interfaces to other brain-inspired systems, specifically DNN style architectures. The ability to interface with other systems enables the development of applications that benefit from both the spatio-temporal nature of our spiky mrDANNA-based cores but also the strong classification capabilities of DNNs. Alternatively, the spike-digital conversion layers can be leveraged to interface directly with any of a variety of possible sensor systems. Thus, the proposed system aimed to be flexible in terms of both how neural networks are implemented in physical space and also the kinds of applications that can be realized.

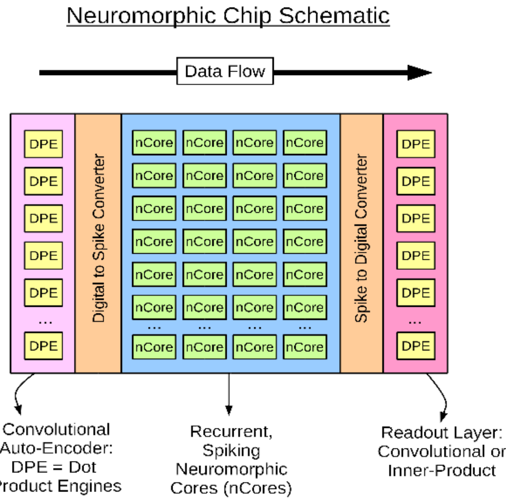


Figure 2. *System level view of the proposed reconfigurable neuromorphic array, complete with spiky neuromorphic cores (nCores) and dot product engines (DPEs).*

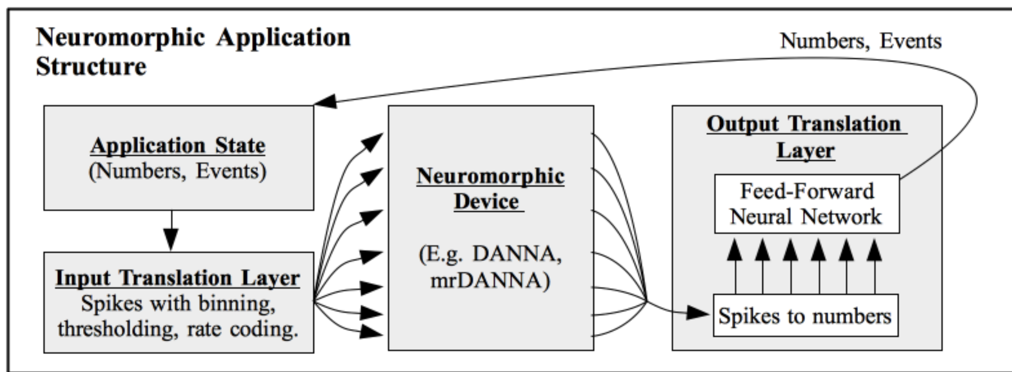


Figure 3. *Global view of the neuromorphic application structure to be implemented in RAVENS.*

4.1 Technical Approach Utilized

For this effort there were two primary tasks that were proposed for the SUNY Polytechnic Institute team. This project was a collaboration between SUNY Polytechnic Institute and the University of Tennessee-Knoxville. The UT-Knoxville team focused on the design, simulation and software development for the proposed RAVENS chip, while the SUNY Polytechnic team was focused on the optimization of memristor (RRAM) performance in hardware, and fabrication of the hardware design provided by UT-Knoxville. The following tasks were completed:

Task 1 Optimize memristor performance and validate hybrid CMOS/memristor hardware.

Task 2 Fabricate fully reconfigurable neuromorphic hardware using hybrid CMOS/memristor architecture.

5.0 Methods, Assumptions and Procedures

5.1 Fabrication

In this effort, memristor (aka: RRAM) devices and CMOS were built in-house on 300 mm wafers at the SUNY Polytechnic Institute's Center for Semiconductor Research (CSR). The RRAM devices were fabricated using a 300mm wafer platform based on the IBM 65nm 10LPe process technology. A custom hybrid CMOS/RRAM process was developed to allow for a seamless integration of both CMOS and RRAM devices into one process flow with minimal added costs. RRAM devices were integrated between metal 1 (M1) and metal 2 (M2); specifically, the intervening via 1 (V1) layer was split to encapsulate the RRAM device. For the purpose of using a front-end-of-the-line (FEOL) deposition tool for the HfO_2 switching layer (SL) of the RRAM device, custom CA, M1 and V1 layers were developed using W as the interconnect material and a TiN cap to serve as the bottom electrode. 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 RRAM 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 RRAM 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 RRAM (1T1R) structure can be seen in **Figure 4.**

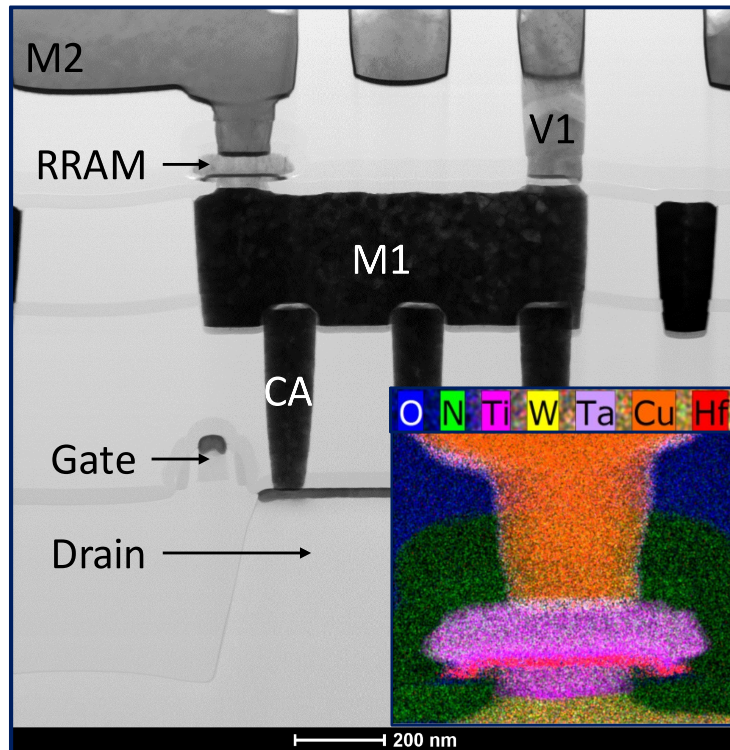


Figure 4. *Cross-sectional contrast image of our integrated CMOS/RRAM circuit taken with a transmission electron microscope. This cross-section shows the W bottom electrode and RRAM device, as fabricated in a FEOL-compatible process, as well as an inset showing the elemental mapping of the key layers of the RRAM device (based on EDS mapping in the TEM).*

5.2 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). RRAM device characteristics were extracted by using DC-sweep as well as pulsing techniques. In both cases a 1 transistor 1 RRAM (1T1R) setup was used to limit the current through the RRAM 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 RRAM. 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.

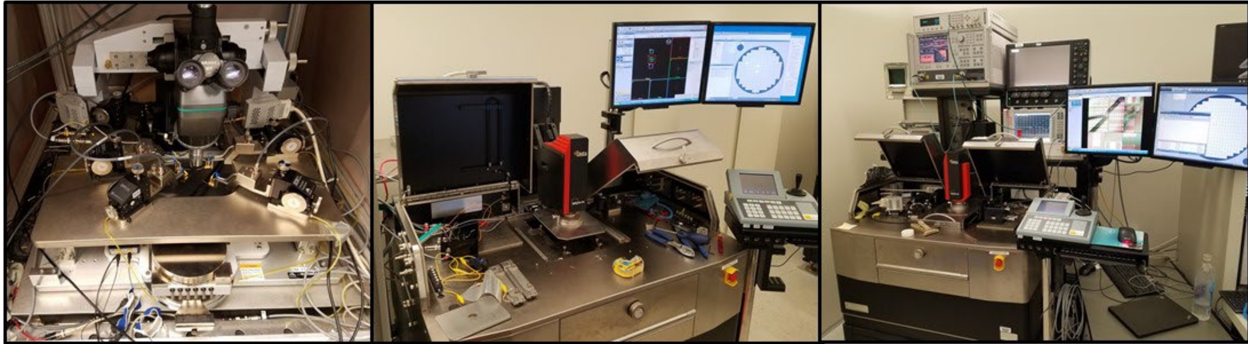


Figure 5. *Probe stations used for this effort. Manual probe station with B1500A analyzer (left), Suss Microtech semi-automatic probe station with Keysight E5270A, and Suss Microtech semi-automatic probe station with B1500A/B1530A (right).*

Electrical measurements primarily utilized an on-chip NMOS field-effect transistor (NFET) for current control during the forming and set operation, as seen **Figure 6**(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 RRAM 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.

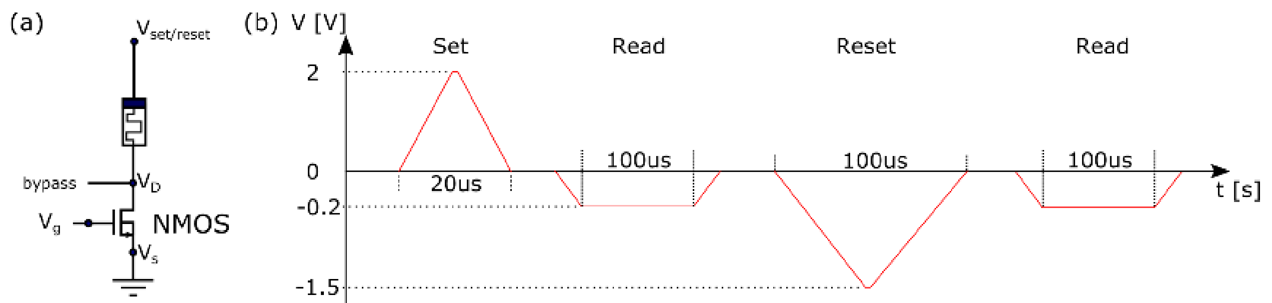


Figure 6. *(a) Schematic of a 1T1R structure with an NMOS that acts as the current limiting device during the forming and set operations. A parasitic base diode opens during the reset allowing for a higher current during the reset operation. The bypass connection enables the direct measurement of the 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 RRAM 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.

Approved for Public Release; Distribution Unlimited.

6.0 Results and Discussion

6.1 Key Accomplishments

The following sections outline the key accomplishments under this project. The first section, "Optimization of memristor performance and validation of hybrid CMOS/memristor hardware" is an excerpt from our 2021 publication that was a result of this work and provides a concise summary of the outcomes.[7] The second section, "Fabrication of fully reconfigurable neuromorphic hardware using hybrid CMOS/memristor architecture" describes our accomplishments in aggregating design information from our collaborating partners at UT-Knoxville and the results of our fabrication efforts and initial testing efforts with this hardware.

6.1.1 Optimization of memristor performance and validation of hybrid CMOS/memristor hardware

6.1.1.1 Optimization of memristor (RRAM) performance for binary switching and reliability

In this portion of the work, we investigated switching reliability of 65nm CMOS integrated 100 nm x 100 nm HfO_2 RRAM devices as a function of various RRAM operation parameters. An in-depth analysis studying the impact of various pulse conditions, operating current on RRAM switching were performed, which enabled switching optimization in terms of R_{off}/R_{on} ratio, switching variability and yield for devices across a full 300mm wafer.

As described in the methods section, above, CMOS/RRAM structures were implemented on a 300 mm wafer platform utilizing 65nm CMOS technology, in which the RRAM device stack was incorporated between the metal 1 (M1) and metal 2 (M2) metallization layers, using a custom designed front-end-of-the-line (FEOL) compatible M1 process flow. The RRAM devices use a 100 nm x 100 nm TiN bottom electrode (BE) fabricated via a subtractive integration on top of tungsten (W) M1 layer. This TiN BE acts as via 0 (V0) layer on which the subsequent RRAM device stack was patterned. The BE deposition is followed by the conformal atomic layer deposition (ALD) of ~6 nm HfO_2 RRAM switching layer, physical vapor deposition (PVD) of ~6 nm Ti oxygen exchange layer (OEL), and ~40 nm in-situ PVD deposition of TiN top electrode (TE) layer. A metal-halide based precursor was used for ALD of HfO_2 switching layer since this precursor offers significant switching improvement over metalloorganic precursor for HfO_2 deposition.[8] The entire RRAM device stack was then lithographically patterned using custom designed reactive ion etch (RIE) process. **Figure 7a** shows the transmission electron microscope (TEM) cross-sectional image of the 1T1R structure. The TEM cross-section of the fabricated RRAM device stack (TiN/HfO_2/Ti/TiN) with energy dispersive X-ray spectroscopy (EDS) elemental composition map is shown in **Figure 7b.**

[7] J. Hazra, M. Liehr, K. Beckmann, M. Abedin, S. Rafiq, **N.C. Cady.** Optimization of Switching Metrics for CMOS Integrated HfO_2 based Bipolar RRAM Devices on 300 mm Wafer Platform. (2021) *IEEE International Memory Workshop (IMW) 2021.* 1-4. DOI: 10.1109/IMW51353.2021.9439618

[8] Hazra, Jubin, et al. "Improving the Memory Window/Resistance Variability Trade-Off for 65nm CMOS Integrated HfO 2 Based Nanoscale RRAM Devices." *2019 IEEE International Integrated Reliability Workshop (IIRW)*, pp. 1-4, 2019.

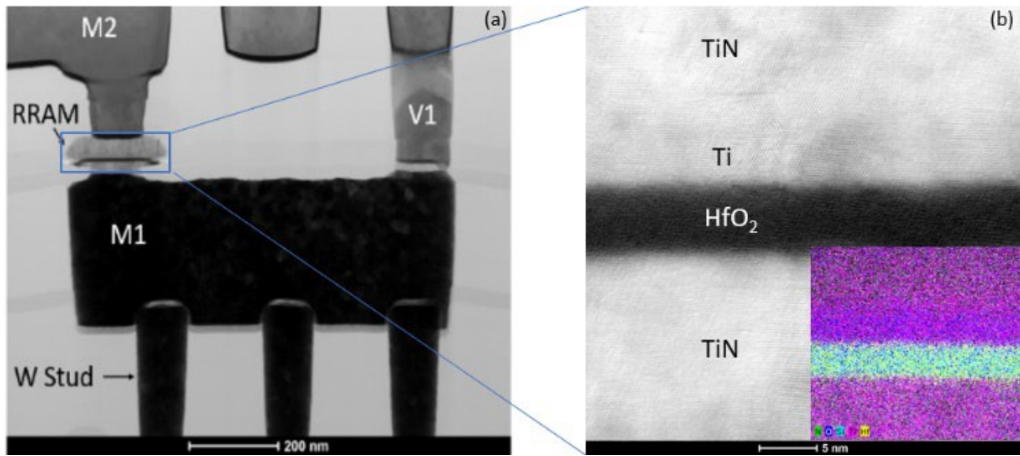


Figure 7. *a) TEM micrograph of 1T1R integration of HfO_2 RRAM devices b) cross-section of RRAM stack, the inset displays EDS map of different elements. Yellow, blue and purple represent Hf, O, and Ti respectively.*

Figure 8 shows a typical switching pulse consisting of μs length SET/RESET pulses for switching the devices in on- and off-state, respectively. A read pulse of -0.2 V was applied after each SET/RESET pulse to read out the LRS and HRS values. The maximum operating current during the forming/set operation of the RRAM cells was controlled by gate voltage modulation of the integrated NMOS transistor.

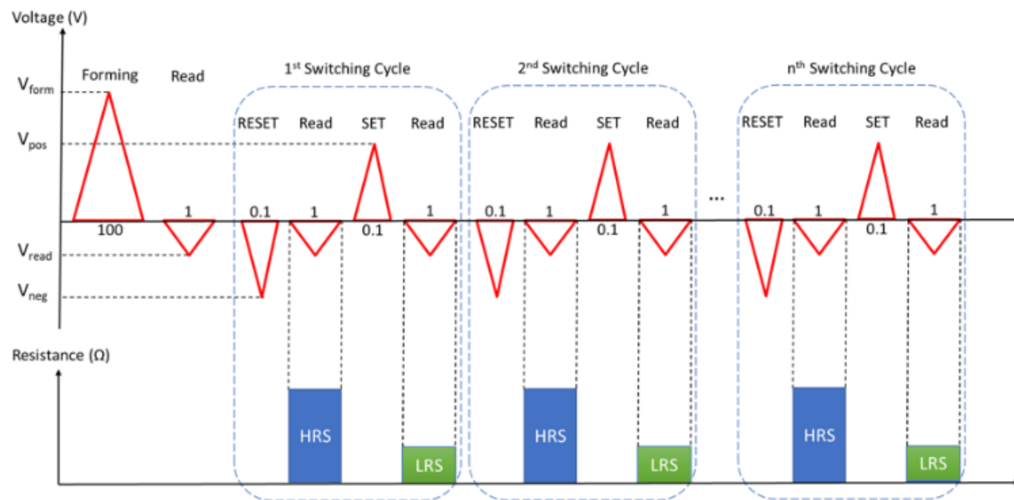


Figure 8. *Schematic of typical pulse switching cycle to study switching endurance of RRAM* devices.

In order to understand inter-relationship of different switching metrics we plotted resistance state (HRS & LRS) and memory window (MW) variation for 100+ tested 1T1R cells. We observed

HRS switching variation showed much stronger relationship with HRS resistance (**Figure 9b**) as compared to LRS variation (**Figure 9a**). This is likely due to the metallic nature of filament (Ohmic conduction) in LRS regime and negligible impact of individual defects/traps states on current transport as observed in HRS regime which triggers resistance variability[9]. Hence, the MW-resistance variability trade-off becomes critical for HRS as shown in **Figure 9d**, where the majority of high MW RRAM cells show significantly large HRS switching variability. This trade-off is much lower for LRS resistances, where the majority of 1T1R cells with large memory window showed lower switching variation ($\sigma_R/R < 0.1$) as depicted in Fig. 3c. Hence, device switching optimization was needed to achieve the desired "target region" (MW> 30, $\sigma_R/R < 0.5$).

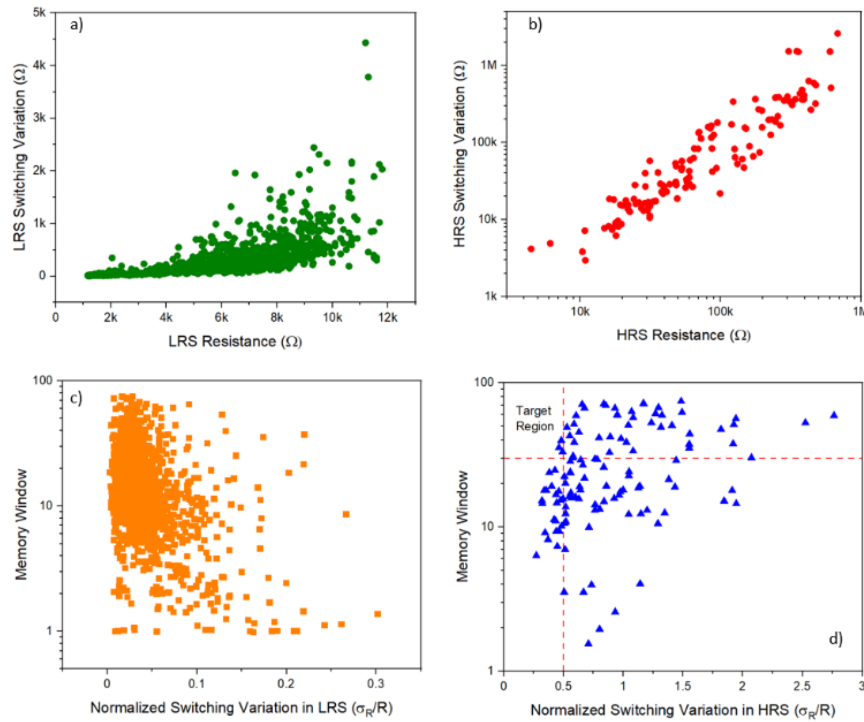


Figure 9. *Resistance with switching variation for different a) LRS and b) HRS resistance values for 100+ tested RRAM cells. Memory window with normalized switching variation for c) LRS and d) HRS resistances.*

The switching performance of fabricated RRAM devices was first studied with a wide range of maximum operating current I_{max} values. A greater I_{max} likely results in a larger effective CF cross-section[10], reduces LRS as illustrated in cumulative frequency plot in **Figure 10a**, and median LRS plot in Fig. 4c. With $I_{max} < 890$ µA, the HRS distribution (**Figure 10b**) remains unaffected with change in I_{max}. In Fig. 4c, "SET failure" was observed for very low I_{max} (~1 µA) when the device failed to switch to LRS due to absence of enough oxygen vacancies in depletion gap region. For $I_{max} > 1$ mA, the devices showed "RESET failure" due to possible accumulation

[9] Puglisi, Francesco M., et al. "An empirical model for RRAM resistance in low-and high-resistance states." *IEEE Electron Device Letters* 34.3 pp. 387-389, 2013.

[10] Bersuker, G., et al. "Metal-oxide resistive random access memory (RRAM) technology: Material and operation details and ramifications." *Advances in Non-Volatile Memory and Storage Technology*. Woodhead Publishing, pp. 35-102, 2019.

of too many oxygen vacancies, and hence the inability to undergo gap formation by RESET pulses. Both failure conditions result in collapse of memory window as shown in **Figure 10d**. However, a very stable programming window (**Figure 10d**) and switching variability (**Figure 10e**) were achieved for a wide range of I_{max} (100 µA- 800 µA). For low switching energy considerations, I_{max} of 100 µA was chosen for subsequent optimization analysis.

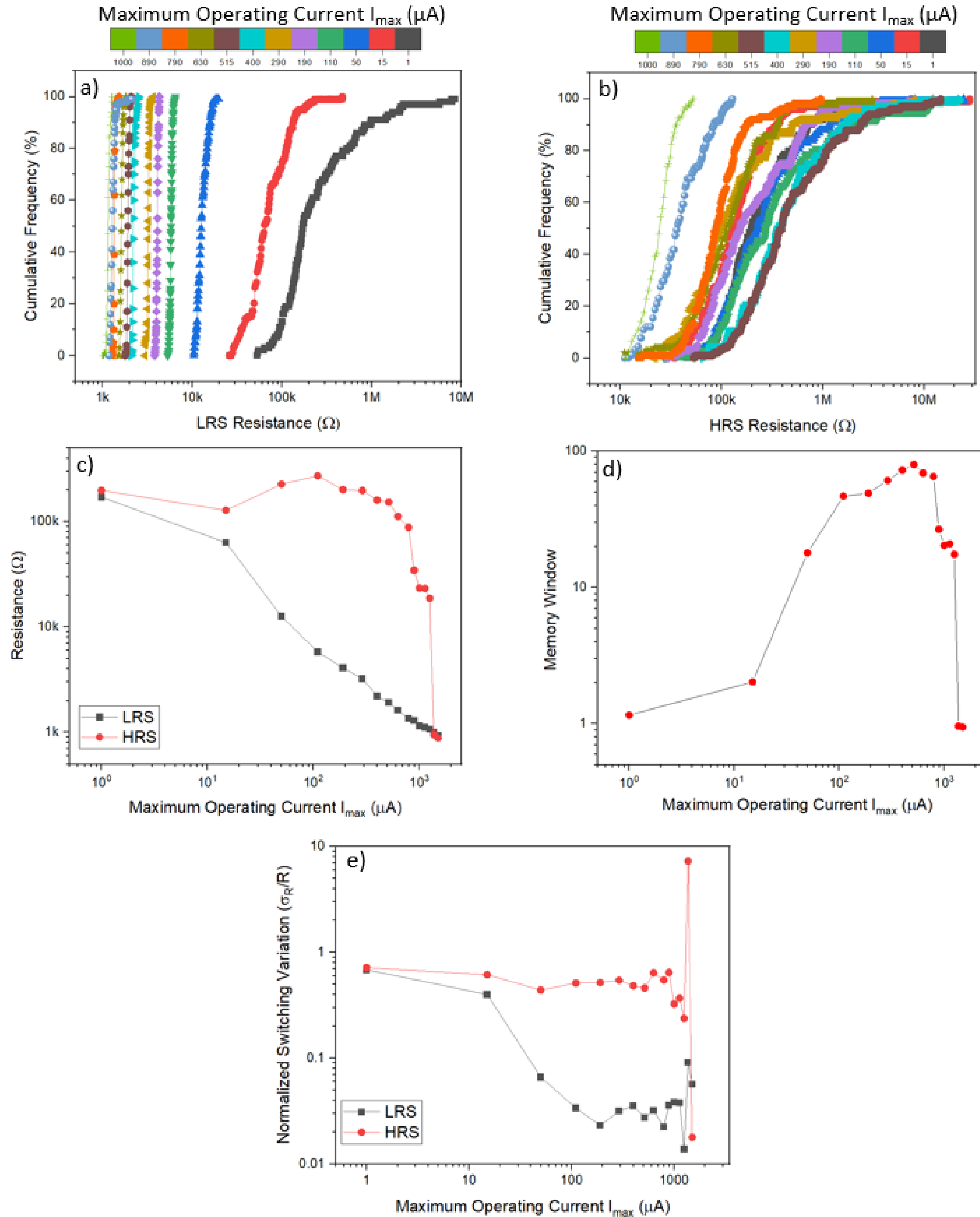


Figure 10. *Cumulative frequency plot of a) LRS and b) HRS values with maximum operating current I_{max}. Impact of I_{max} on c) median LRS/HRS resistances d) memory window, and e) LRS/HRS switching variation.*

Figure 11a shows the impact of RESET pulse width tRESET (500 ns – 1 ms) for different pulse amplitude Vneg values (-1 V to -1.8 V). For Vneg -1 to -1.4 V, a gradual increase in HRS distribution was observed with increase in tRESET. However, for Vneg ≥ -1.6 V, a saturation region is reached where the HRS cannot increase further with an increse in tRESET. At, Vneg = -1.8 V and tRESET = 1 ms, the RESET pulse has too much total power for reliable device operation and causes device failure. As expected, the RESET amplitude/pulse width variation had minimal impact on the LRS. The MW showed a similar trend compared to the HRS where HRS/LRS resistance ratio saturated for Vneg = -1.6 V for all tRESET values. Lower HRS switching variability was observed with lower tRESET (≤ 1µs) for all Vneg values as illustrated in Figure 11c. This is most likely due to less Joule heating, which could reduce the stochasticity of independent defects in insulating gap region at lower RESET pulse width[11]. Hence, I_{max} = 100 µA, V_{neg} = -1.6 V, t_{RESET} = 100 ns were chosen for further optimization steps.

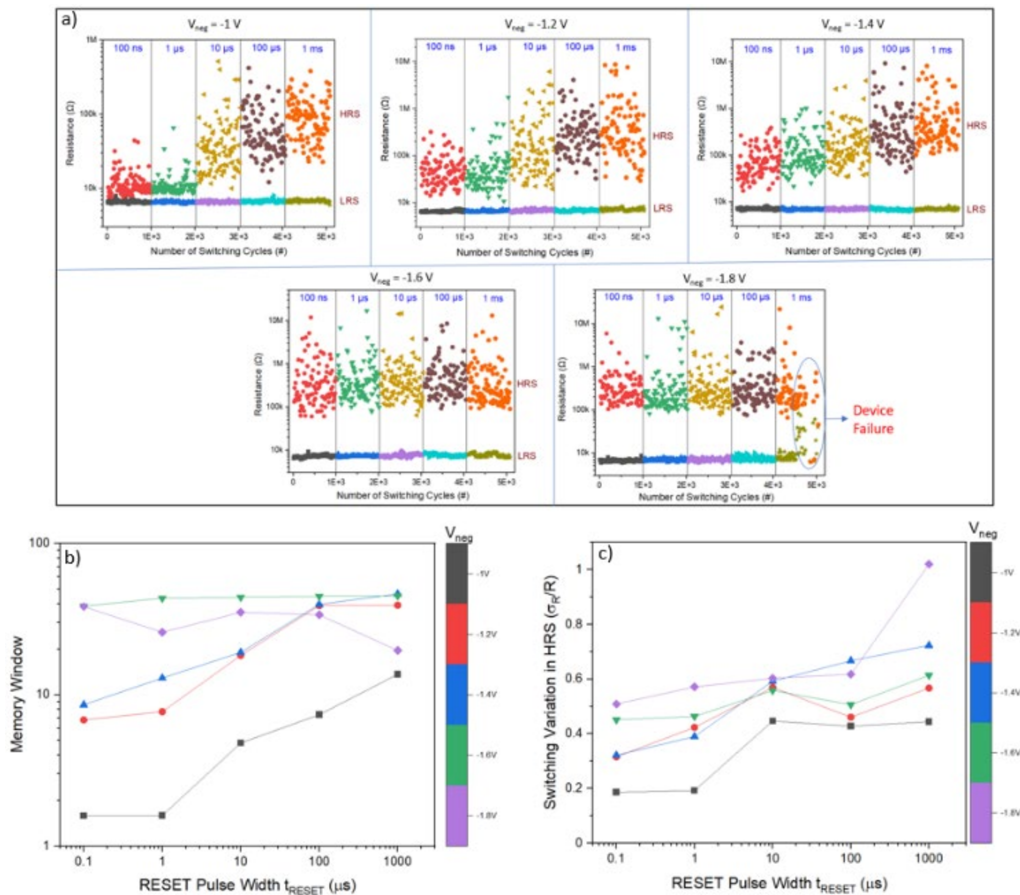


Figure 11. *a) Switching endurance of fabricated RRAM cell for different RESET pulse width t_{RESET} and amplitude V_{neg}. Impact of t_{RESET} on b) memory window, and c) HRS switching variation for different V_{neg} values.*

[11] Nminibapiel, David M., et al. "Characteristics of resistive memory read fluctuations in endurance cycling." *IEEE Electron Device Letters* 38.3, pp. 326-329, 2017.

A similar study was done with respect to the SET pulse width t_{SET} and pulse amplitude V_{pos}. As shown in **Figure 12a**, LRS resistance decreases with increase in t_{SET} due to a higher injection of oxygen vacancies into the gap region. Notably, contrary to the RESET process, no saturation was observed in the LRS modulation by t_{SET} for any V_{pos} values. Maximum HRS resistances and memory windows were achieved for V_{pos} of 2 V and 2.5 V by optimum SET/RESET balance (**Figures 12b & 12c**). Too high (3 V) or too low (1 V and 1.5 V) V_{pos} negatively impacts the balance, and hence, lowers HRS and HRS/LRS ratio. Similar to the reset, 100 ns t_{SET} showed the lowest HRS switching variability (**Figure 12d**).

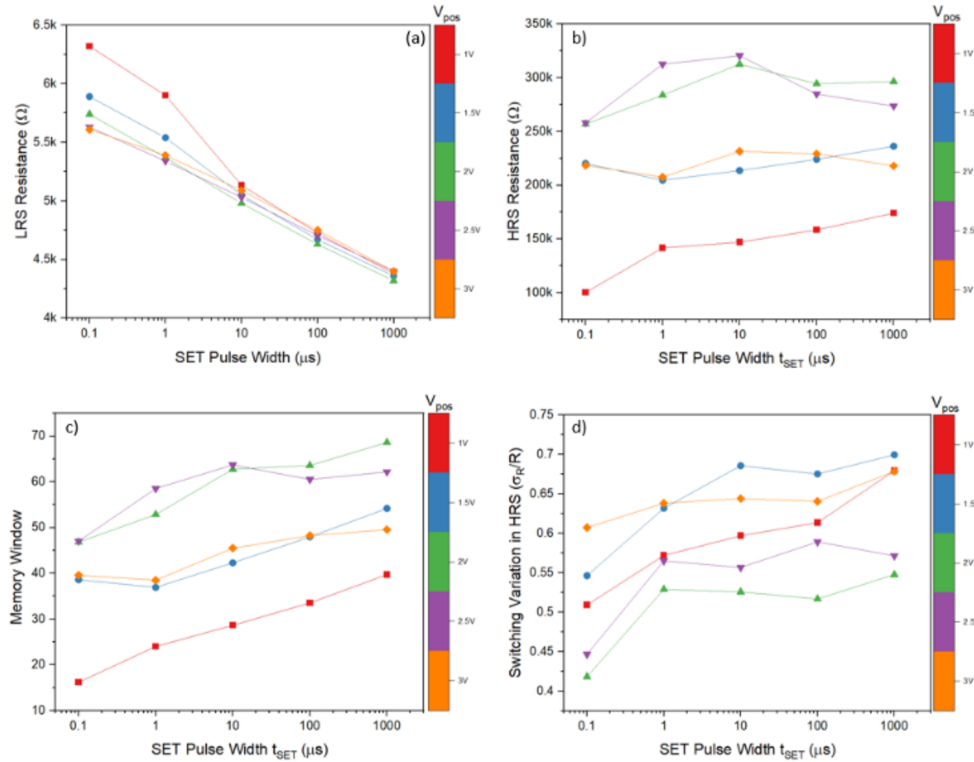


Figure 12. *a) Impact of SET pulse width t_{SET} on a) LRS, b) HRS, c) memory window, and d) HRS switching variation for different SET amplitude V_{pos} values.*

To understand the SET/RESET balance in depth, a switching study was conducted varying both V_{pos} and V_{neg} keeping t_{SET} and t_{RESET} at 100 ns for low switching variability. Shorter pulse times are also preferred for designing low power memory cells and reliable devices[12]. It should be noted that minimum pulse width of 100 ns used for the analysis were limited by the experiment instrument capability. As depicted in **Figure 13a**, V_{neg} of -1.6 V with V_{pos} = 2 V and 2.5 V showed the largest memory window (> 45) by optimally balancing the SET/RESET process. Since, operated at lower pulse width (100ns), both LRS and HRS showed reasonable switching variation unless operated at V_{neg} = -1.8 V, where the device is pushed towards the "failure" regime and the switching performance degrades (Fig. 7b & 7c). **Figures 13d** and **13e** show

[12] Y. Y. Chen *et al.*, "Balancing SET/RESET Pulse for >10^{10} Endurance in HfO_2/Hf 1T1R Bipolar RRAM," in *IEEE Transactions on Electron Devices*, vol. 59, no. 12, pp. 3243-3249, 2012.

switching endurance of a typical "failed" and "yielded" RRAM cell for "unbalanced" and "balanced" SET/RESET conditions respectively enforcing the need of optimum RRAM operation parameters for superior switching performance of 1T1R cells. Henceforth, I_{max} = 100 μA, V_{neg} = -1.6 V, t_{RESET} = 100 ns, V_{pos} = 2 V, t_{SET} = 100 ns were chosen as operating conditions for statistically significant full wafer and long cycling switching analysis.

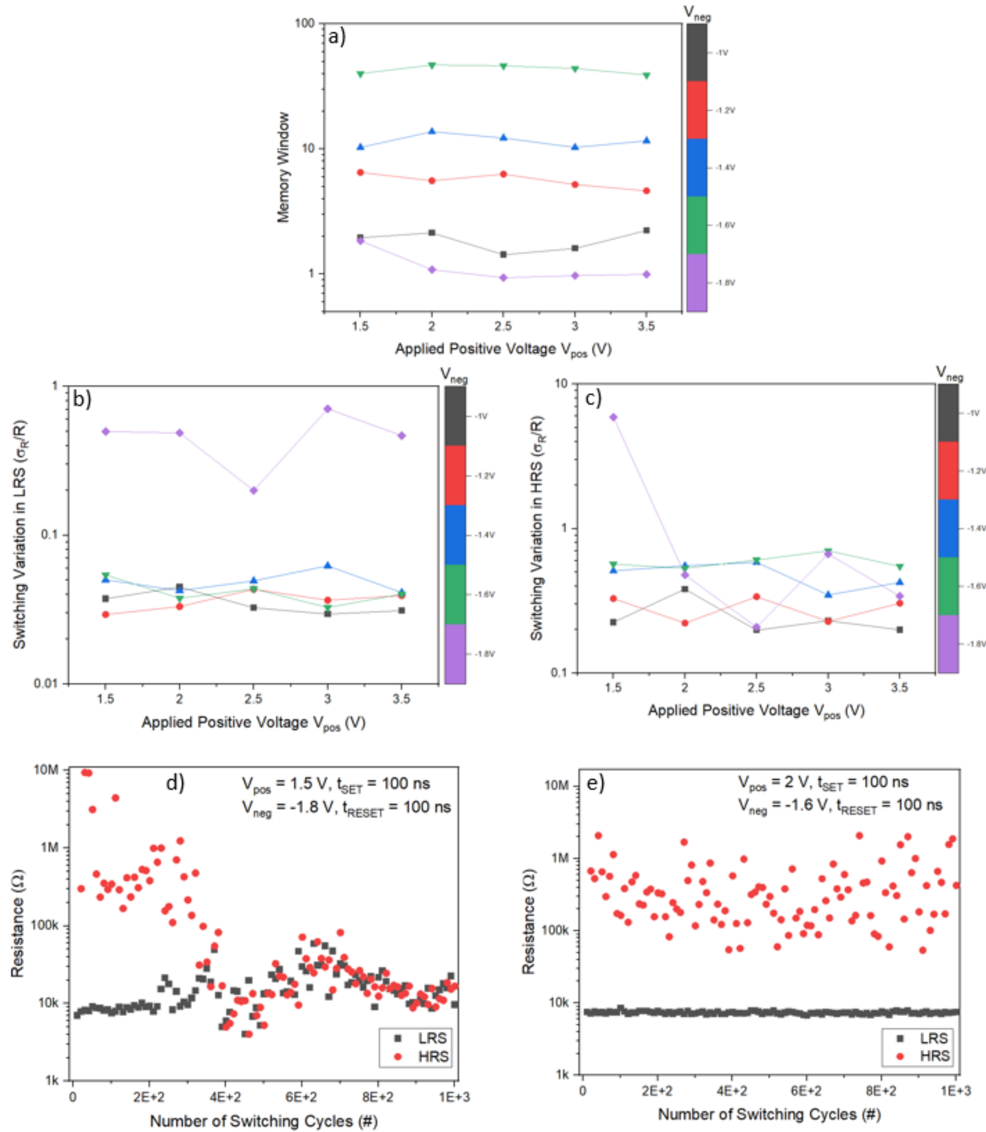


Figure 13. *Impact of V_{pos} on a) memory window, b) LRS, and c) HRS switching variation for different RESET amplitude V_{neg} values. Switching endurance of a RRAM cell for d) "unbalanced", and e) "balanced" SET/RESET.*

Figure 14a shows excellent long-term switching endurance characteristics (up to 1B cycles) for a typical optimized RRAM cell. The device showed an impressive R_{off}/R_{on} ratio of 47.5 with reasonable switching variation of 0.43 for HRS and 0.12 for LRS (**Figure 14b**). The device also

showed excellent data retention performance at 373K with highly stable programming window and no visible HRS/LRS degradation until 10^5 seconds (**Figure 14c**).

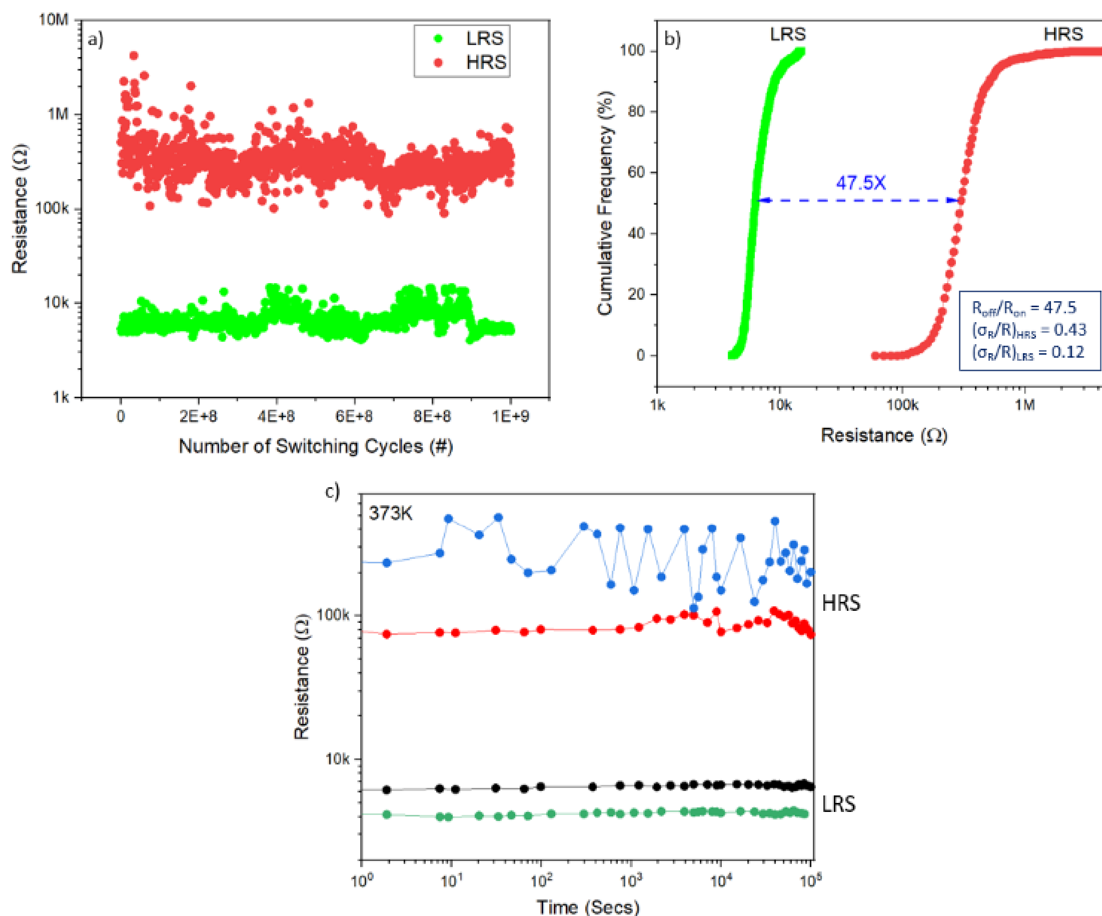


Figure 14. *a) Switching endurance up to 1B cycle, and b) corresponding HRS and LRS cumulative frequency plot of optimized 1T1R cell. c) Data retention for two different HRS and LRS resistance levels at 373K.*

Finally, we present full 300 mm wafer switching statistics of optimized RRAM switching conditions. The optimized scheme showed excellent memory window (~33), low cycle-to-cycle (~0.45) and cell-to-cell HRS switching variation (~0.3) and an impressive switching yield (~90%) across a full wafer as illustrated in **Figure 15a – 15d**. The MW / HRS resistance variability trade-off was significantly improved post-optimization with the majority of 1T1R cells showing switching performance in/around the target region (checked area: MW> 30, σ_R/R< 0.5) as compared to pre-optimization performance shown in inset of **Figure 15e**.

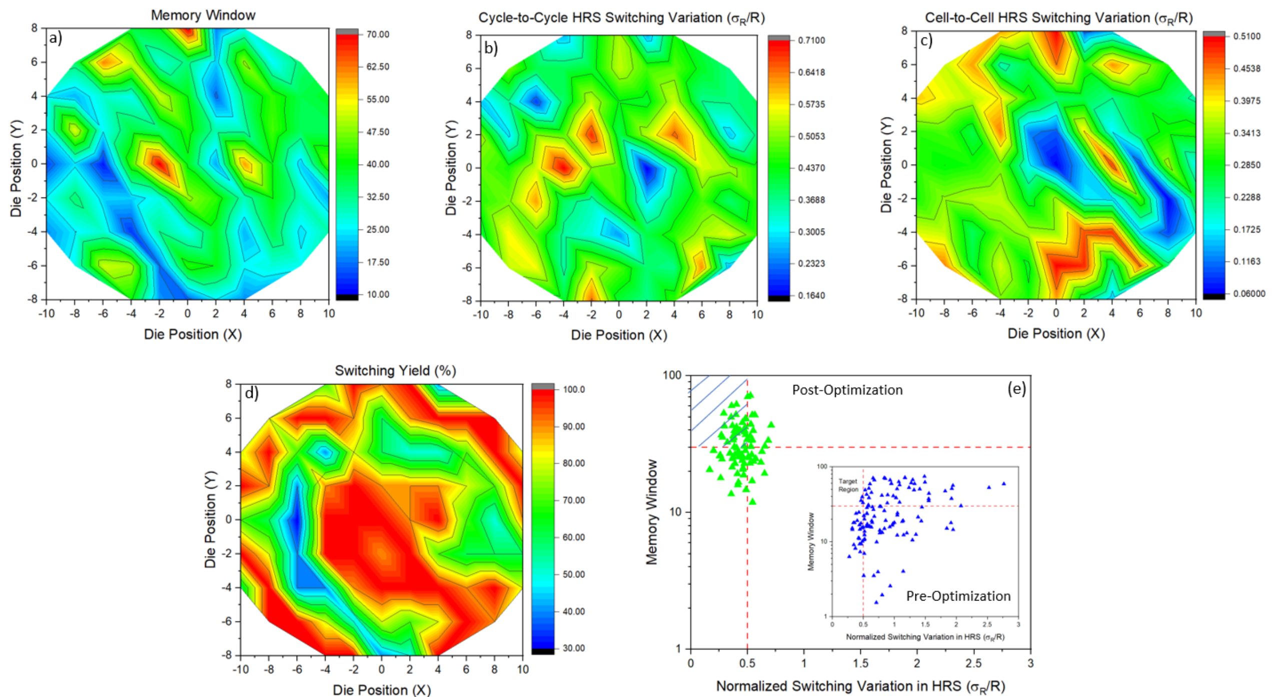
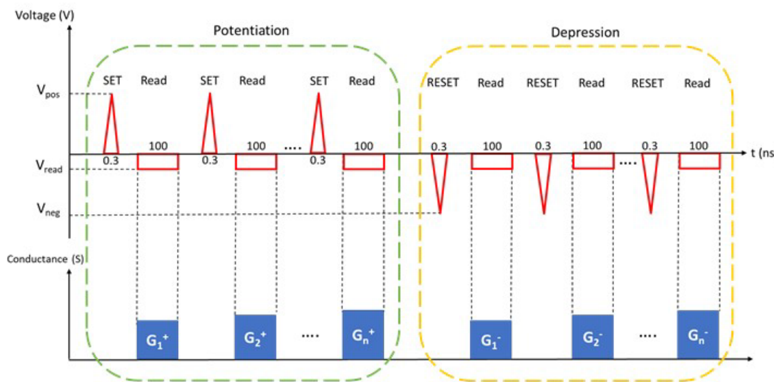


Figure 15. *Full 300 mm wafer statistics for a) memory window, b) cycle-to-cycle HRS switching variability, c) cell-to-cell HRS switching variability d) switching yield, and e) memory window vs. normalized HRS switching variation using optimized operation conditions: I_{max} = 100 µA, V_{neg} = -1.6 V, t_{RESET} = 100 ns, V_{pos} = 2 V, t_{SET} = 100 ns. Inset in Fig. 9e) shows pre-optimization plot for memory window with HRS switching variation.*

In summary, we demonstrated excellent 1T1R performance metrics with respect to memory window, switching variability and switching yield on a full 300mm wafer scale via an optimal balance of programming current, SET/RESET pulse width and amplitude conditions. The optimization approach helped to improve the trade-off between the memory window and switching variation. It provides critical insight into the operation of our HfO_2 RRAM cells to enable optimum device switching. The devices showed excellent endurance (>10^9) and data retention performance at elevated temperature (10^5 s at 373K). Demonstration of superior switching metrics for CMOS integrated 100 nm x 100 nm RRAM cells at 300mm wafer scale establishes the promising potential of RRAM devices as possible replacement of NAND flash memory by meeting high volume manufacturing requirements of the semiconductor industry.

6.1.1.2 Optimization of memristor (RRAM) performance for analog switching

The SUNY Poly team also focused on analog switching performance for RRAM, to achieve ideal functionality for neuromorphic computing and AI hardware applications. We focused on the linear potentiation (increase) and depression (decrease) in RRAM resistance (shown here as the inverse, conductance), which could be used for analog memory and/or synaptic weight adjustment. As shown in **Figure 16**, applying a series of short (300 ps) pulses to our hafnium oxide 1T1R devices results in a non-linear potentiation and depression cycle.



- Successive SET/RESET pulses for Depression Potentiation Cycle (DPC) behavior of RRAM cells.
- Gradual Conductance Modulation.

- 100 µA operating current limit.
- DPC consists of 200 +1V SET & 200 -1V RESET pulses.
- Using short (300 ps) pulses, the conductance states can be gradually increased/decreased.

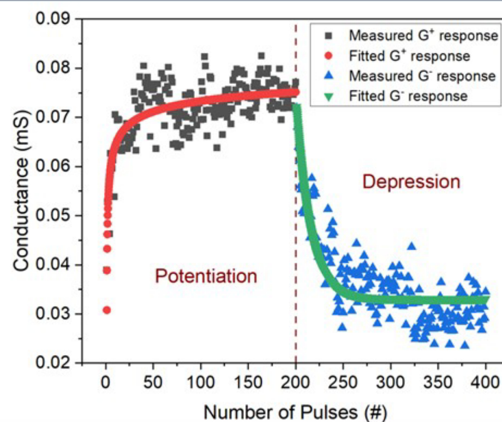


Figure 16. *Potentiation and depression of a hafnium oxide based 1T1R cell using a series of uniform 300 ps pulses (100 uA operating current and constant set/reset voltage).*

To deliver linear potentiation and depression cycles, we utilized a variable pulse height (set or reset voltage) of the same pulse width (300 ps), as shown in **Figure 16**. In this experiment, pulse width was kept at 300 ps, while the set voltage was varied from 0.75 – 1.05 V and the reset voltage was varied from -0.8 V to -1.1 V. This resulted in a near-linear change in conductance for portions of the potentiation and depression cycle, and an overall improvement in the linearity and symmetry of the conductance change throughout the entire conductance range for these devices (**Figure 17**).

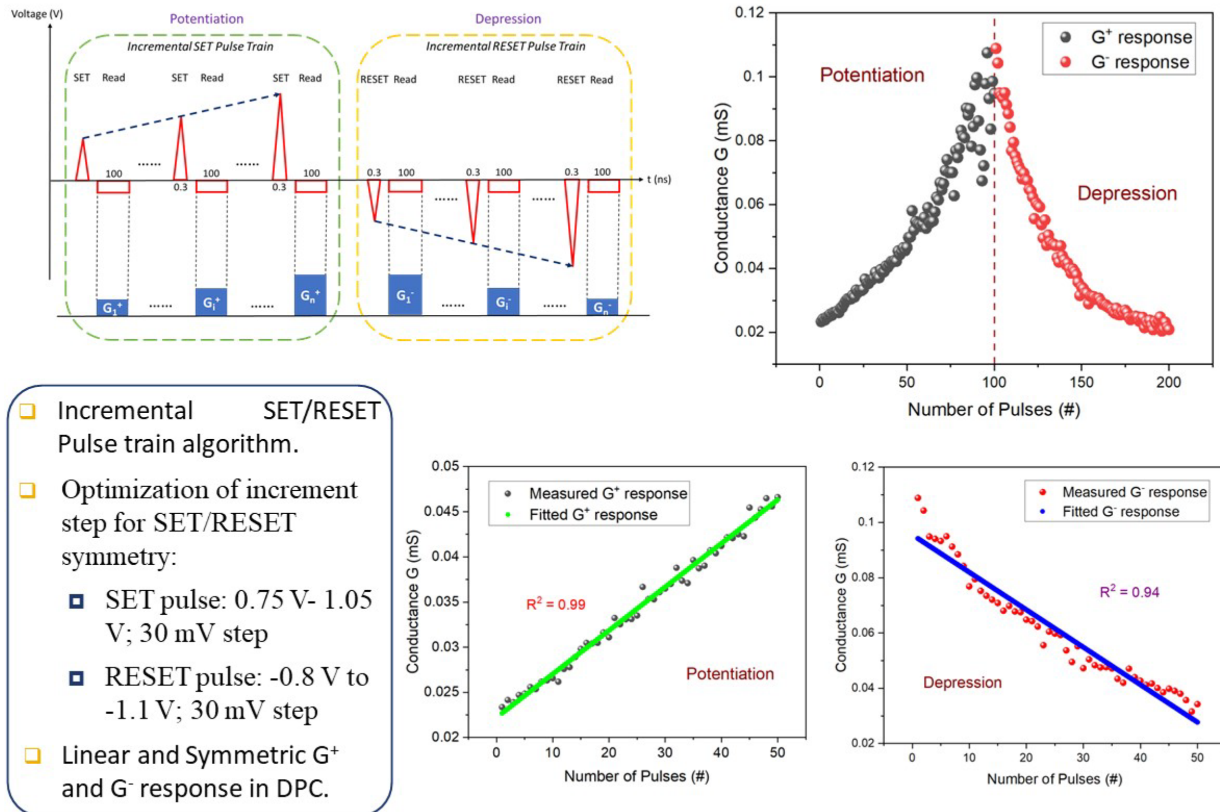


Figure 17. *Application of variable pulse height (set or reset voltage) results in improved linearity and symmetry of conductance change through the potentiation and depression cycle for hafnium oxide 1T1R cells.*

6.1.1.3 Delivery of early-stage hardware to AFRL collaborators (AFRL-RI and AFRL-RX)

As part of this effort we delivered multiple RRAM/CMOS test chips to our collaborators at AFRL-RI (Rome, NY) for evaluation and testing. This included mainly mrDANNA chips that contain individual RRAM (1R), individual 1 transistor 1 RRAM (1T1R), 1T1R arrays, and a number of other test circuits and neuromorphic computing elements. The goal of this interaction is to get AFRL-RI staff comfortable with handling and testing our chips, to ensure that the testing setup at AFRL-RI is capable of testing the RRAM/CMOS circuits, and that both binary and analog (multi-level) switching of RRAM devices can be performed. Further, this effort serves as a qualification of our chips by an outside laboratory, to confirm the results of our in-house studies.

To date the AFRL-RI team (Lombardi et al.) have focused on testing of 8x8 1T1R arrays have worked with Prof. Cady's group to optimize testing parameters to achieve successful device switching and measurement. This has included one visit by AFRL-RI staff to SUNY Polytechnic institute for training and information exchange. Currently the AFRL-RI team is capable of forming and writing data into the 1T1R arrays using a probe card interface (**Figure 18**) and have achieved data input / read-out for full arrays (**Figure 19**).

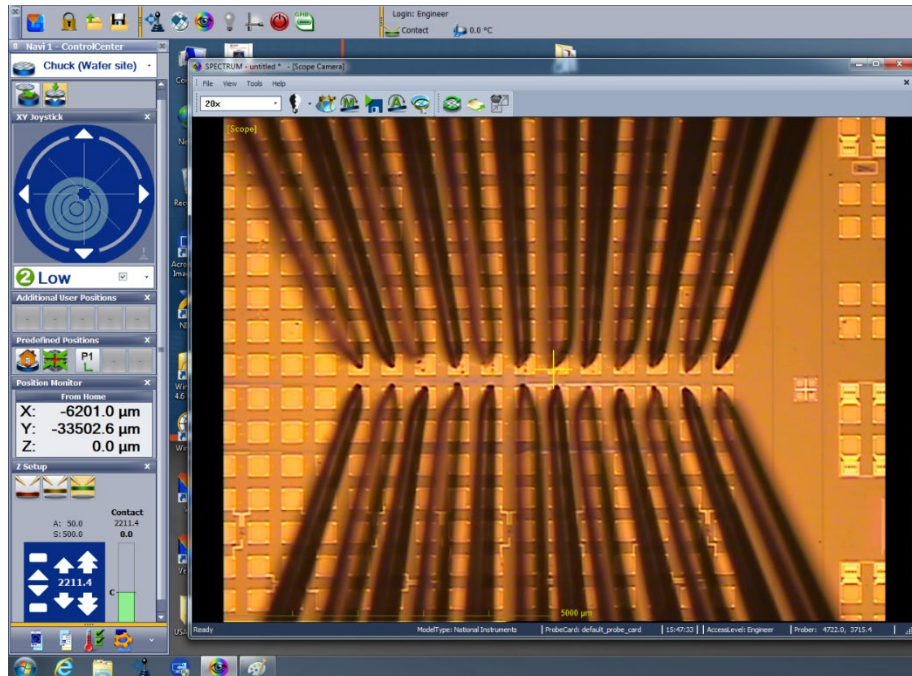


Figure 18. *Image of AFRL-RI probe card contact with SUNY Polytechnic 8x8 1T1R array pads (performed at AFRL-RI, Rome, NY).*

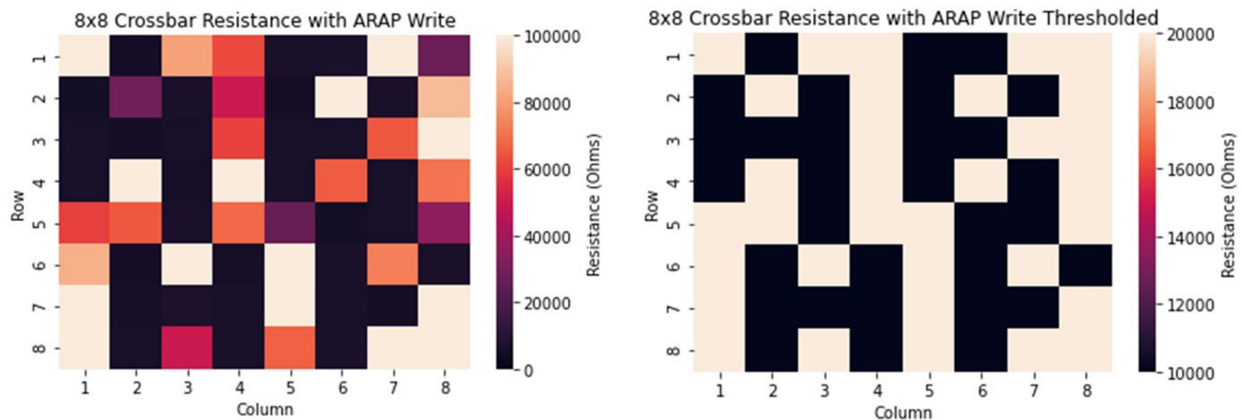
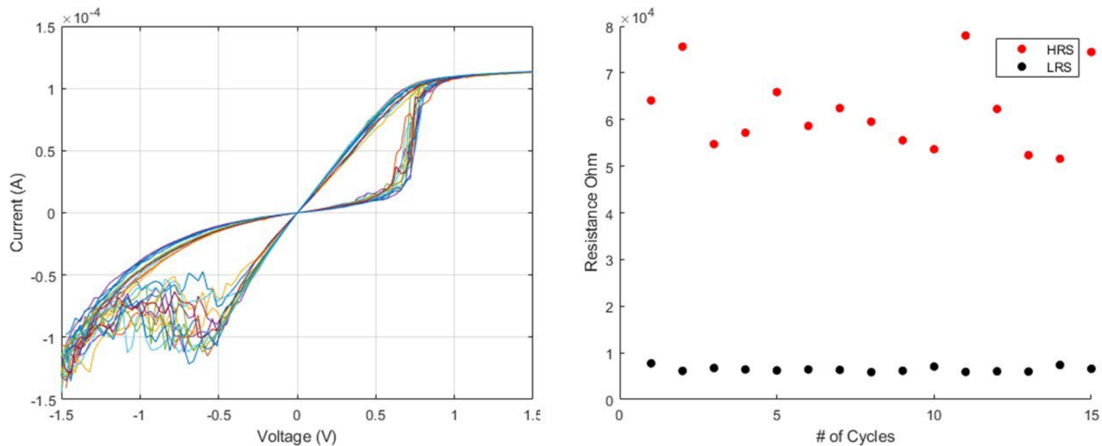


Figure 19. *Successful encoding of 1T1R array with high and low resistance states using a SUNY Polytechnic 8x8 1T1R array at AFRL-RI. Raw resistance data is shown at left, while a threshold-adjusted image of the array data is shown at right.*

RRAM arrays were also sent to AFRL-RX (Dr. Sabyasachi Ganguli) for evaluation. The AFRL-RX team was able to successfully measure 1T1R devices in a traditional probe station setup. Additionally, the SUNY Poly team prepared thin lamella of 1T1R devices on specialized "Protochip" sample holders for performing *in situ* TEM / electrical characterization at AFRL-RX. Experiments to test these devices and observe materials changes and filament formation during RRAM switching are still underway at the writing of this report.

A.



B.

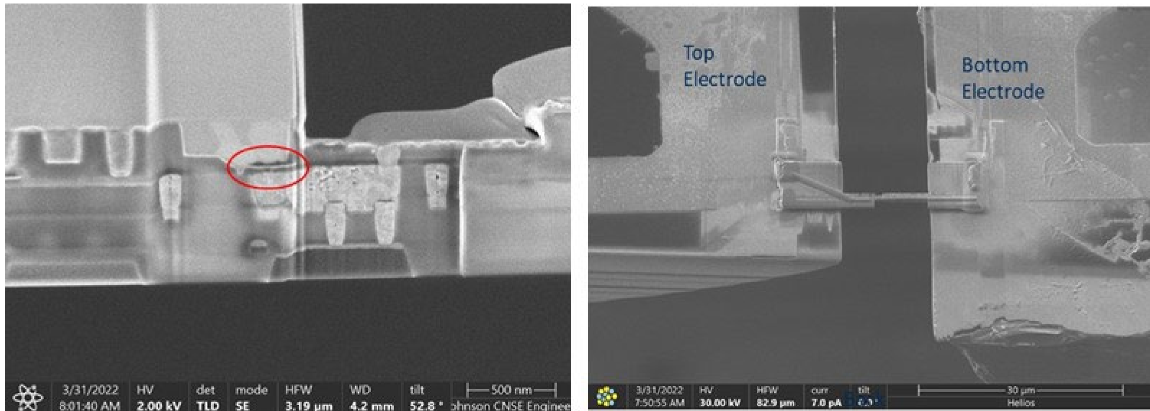


Figure 20. *(A) Successful linear sweep based switching of HfOx 1T1R devices (from the SUNY Polytechnic team) made by AFRL-RX (Dayton, OH). (B) Example of a 1T1R cell prepared for in situ electrical switching in transmission electron microscope (TEM) at AFRL-RX to observe conductive filament formation / rupture during the electroform/set/reset processes.*

6.1.2 Fabrication of fully reconfigurable neuromorphic hardware using hybrid CMOS/memristor architecture

6.1.2.1 Design Overview and Fabrication of RAVENS Wafers

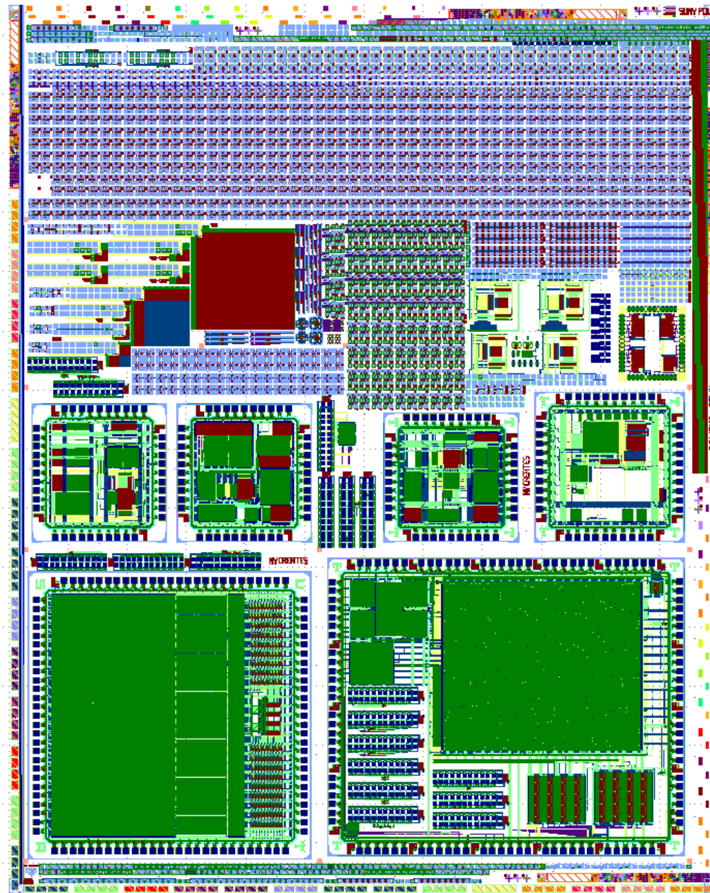
Over the duration of this project, the SUNY Poly team completed full tape-out of the RAVENS design and included designs from multiple collaborating groups. The chip design includes unique circuits from the SUNY Polytechnic team, including 1T1R arrays to be used for matrix vector operations and general RRAM testing, as well as neuron circuits for demonstrating learning behavior on-chip. We have also included the main RAVENS architecture that was designed by our collaborators on this effort (led by Prof. Garrett Rose, UT-Knoxville). We also included designs from AFRL-sponsored researcher, Prof. Dhireesha Kudithipudi (UT-San Antonio), which is a neuromorphic computing architecture design. We have also spearheaded a new collaboration with researchers from Arizona State University, including Profs. Yu Cao, Deliang Fan, and Jae-sun Seo. These investigators submitted small designs, including combined CMOS SRAM/RRAM circuits. Last, we started a new collaboration with Prof. Jaydeep Kulkani

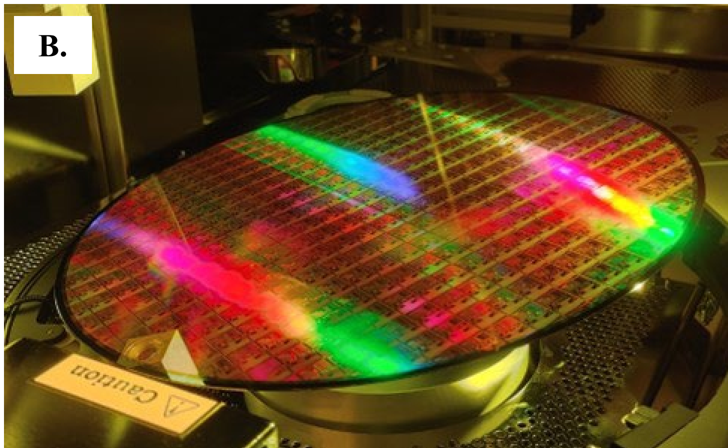
at UT-Austin, and have included small CMOS/RRAM circuits from his group. The UT-Knoxville designs were included as part of the AFRL RAVENS effort, while the ASU, UT-SA and UT-A designs were supported by these collaborators purchasing space on the reticle set (i.e. pay to add space on the multi-project wafer effort). Screen shots of the full reticle design, including sub-designs from the different groups, as well as a bulleted list of the different designs and collaborators:

- **SUNY Polytechnic Institute (N. Cady) - AFRL sponsored**
    - RRAM / CMOS test structures, memory arrays, in-memory computing demonstration circuits
- **UT-Knoxville (G. Rose) - AFRL sponsored**
    - RAVENS RISC-V CMOS/RRAM neuromorphic processor, CMOS-RRAM learning circuits / demonstration circuits
- **UT-San Antonio (D. Kudithipudi) - AFRL sponsored**
    - Genesis CMOS/RRAM neuromorphic computing demonstration
- **UT-Austin (J. Kulkarni) – non-AFRL**
    - Novel CMOS/RRAM circuits for memory, neuromorphic computing & machine learning applications
- **ASU (Y. Cao, J. Seo, D. Fan) – non-AFRL (SRC JUMP)**
    - Novel CMOS/RRAM/SRAM circuits for accurate inference and on-line adaptation

Approved for Public Release; Distribution Unlimited.

A.





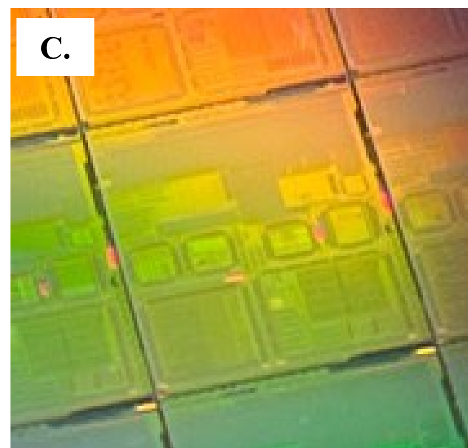


Figure 21. *(A) Full-field reticle view of the tape-out with the UTK "RAVENS" processor, as well as designs from other collaborating groups. (B) Image of full 300mm RAVENS wafer and (C) close-up image of an individual RAVENS die.*

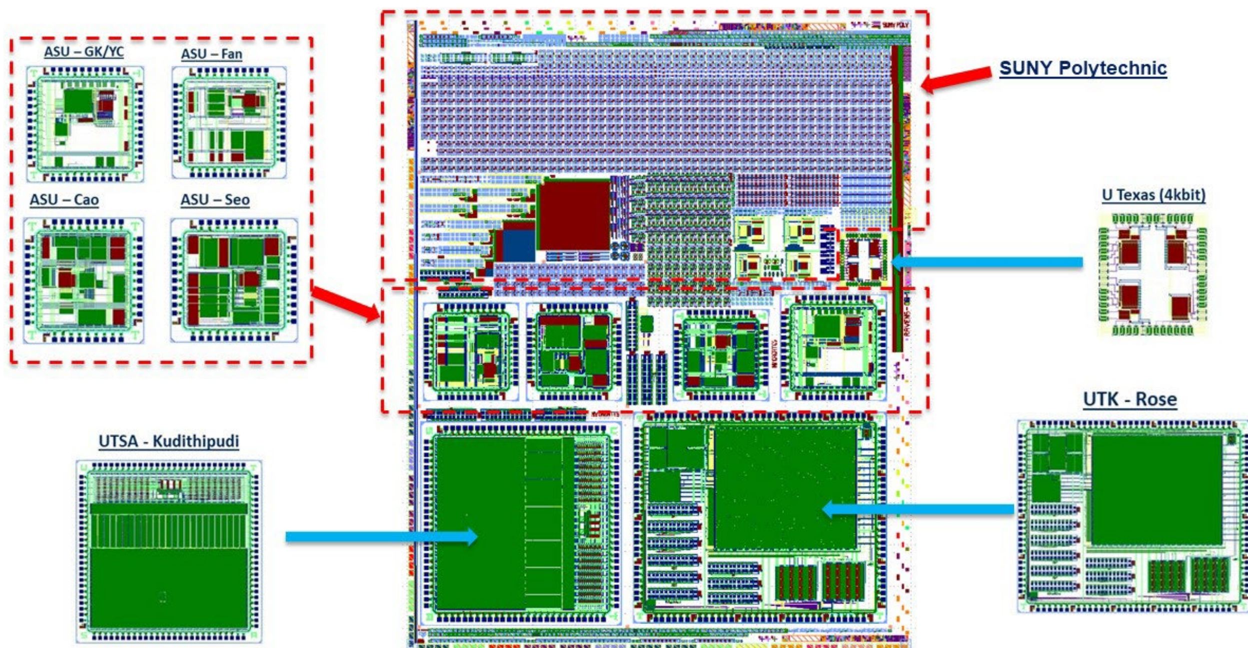


Figure 22. *Labeled sub-die map of the RAVENS design, showing the location of various designs on the reticle.*

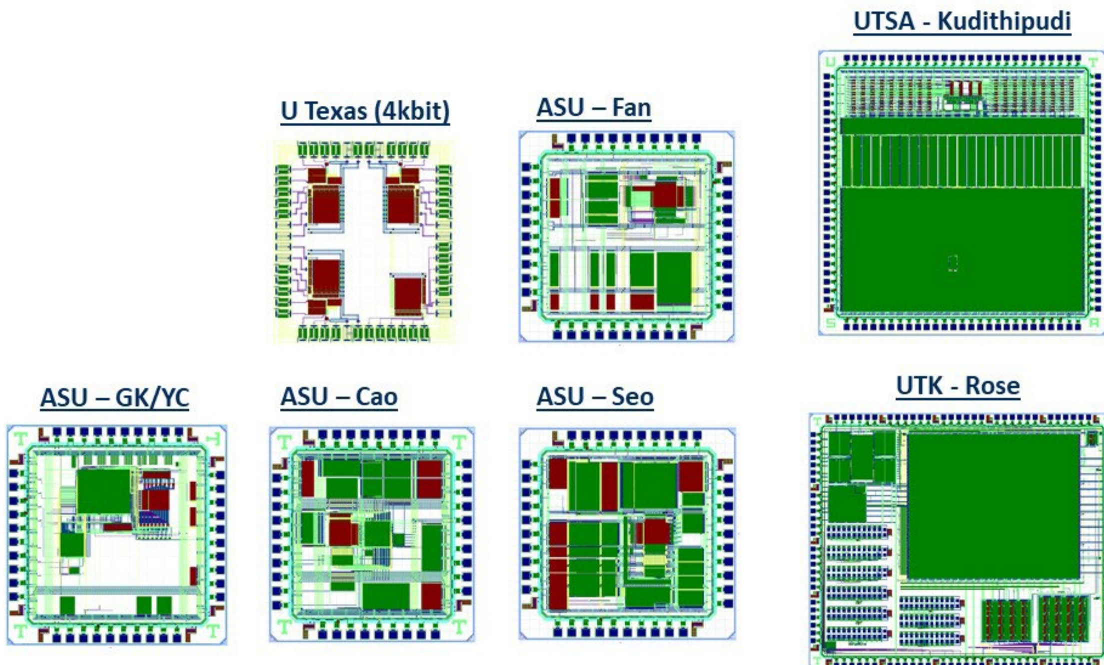


Figure 23. *Sub-designs from collaborating research groups.*

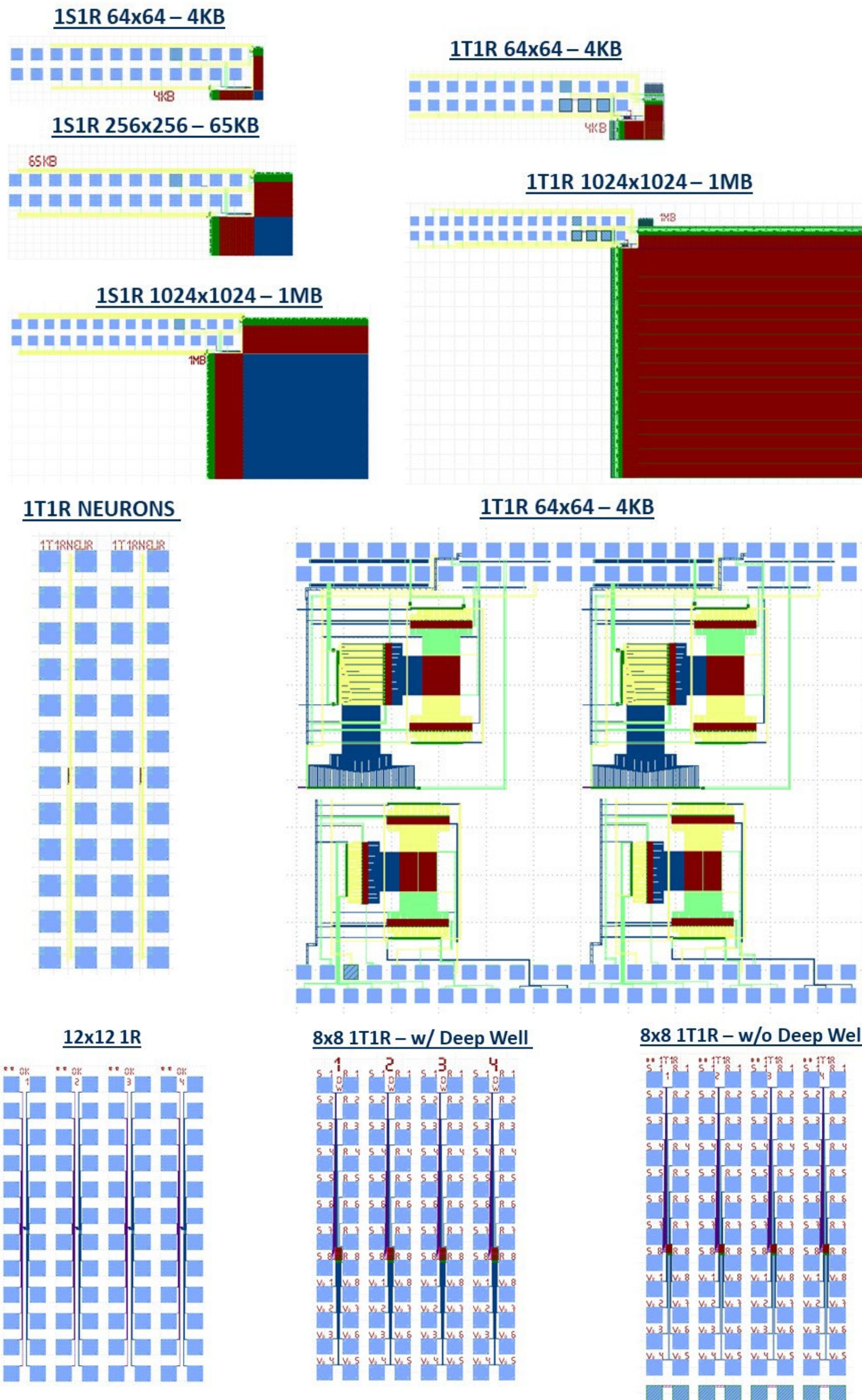


Figure 24. *SUNY Polytechnic designs from the RAVENS tape-out. Designs include memory arrays up to 1MB (with decoder circuit), neurons for demonstrating RRAM-based learning behavior, a 64x64 1T1R array for performing vector matrix operations, a 12x12 RRAM array (no transistors), and finally novel 8x8 1T1R arrays with deep well transistors for enabling RRAM switching with either polarity.*

For the SUNY Poly designs (**Figure 24**) the memory arrays (ranging from 4KB to 1MB will enable larger-scale demonstration of RRAM memory, and encoding of synaptic weight information for large data sets / large neuromorphic computing functions. It will also serve as an excellent testbed for determining yield over larger arrays (vs our smaller arrays in previous runs). The 1T1R neurons are also demonstration circuits to test learning functionality for some novel neuron designs. The individual 1T1R 64x64 array was specifically designed for performing vector matrix operations (summing currents from columns, when input voltages are applied via rows). This will enable operations such as vector matrix multiplication or other matrix-based functions. Last, we have implemented a standard 12x12 1R array (essentially a crossbar array) and some additional 8x8 1T1R arrays that are driven with deep well transistors. This will enable us to use negative and positive voltage when switching the RRAM and performing vector matrix multiplications making the programming via our current setup easier. In addition, it does not require a dedicated ground contact either via the chuck or via the $25/26^{th}$ contact.

6.1.2.2 Initial Qualification Testing of RAVENS Wafers

The First RAVENS wafer finished processing through the aluminum contact module and was pulled off the line after final electrical inline verification. The wafer will be ready for packaging after further offline testing of the ReRAM devices and CMOS circuits. Figure and Figure 26 show a Scanning Tunneling Electron Microscope (STEM) image of the CMOS build up to the BA level and ReRAM devices embedded between M1 and M2, respectively. The images confirm a successful integration of all critical layers starting with the transistors continuing through the minimum feature size metallization layers (M1-M4) and finishing with the 2x metallization layers BA and BB for **Figure 25** and the ReRAM module for **Figure 26**.

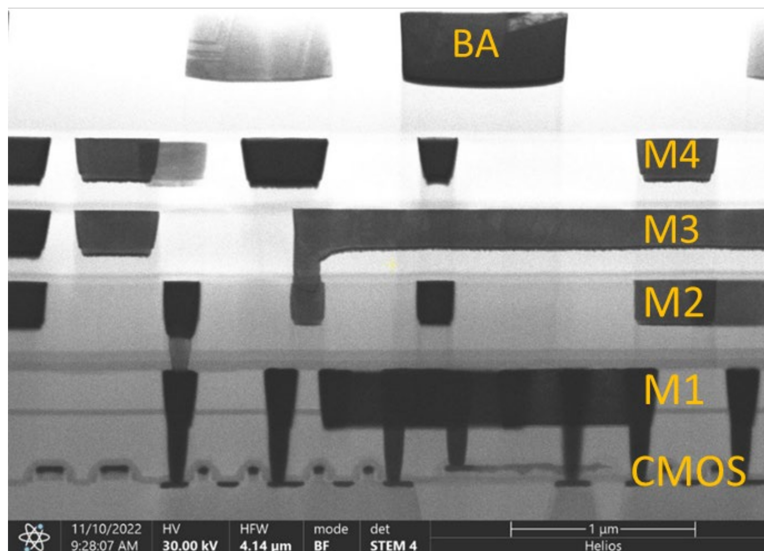


Figure 25. *30keV STEM image from the center of the first fully build RAVENS test chip. Visible are the transistors (bottom layer), four metallization layers (M1-M4) and BA. The two upper most layers (BB, LB) did not fit into the frame.*

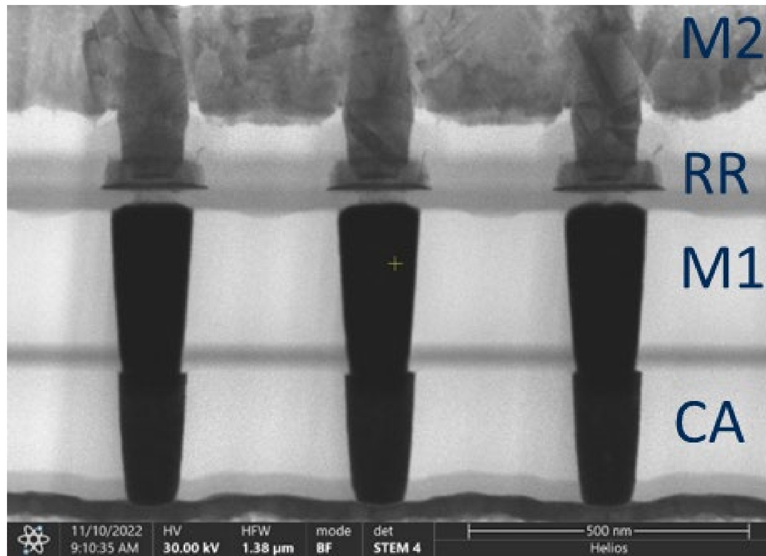


Figure 26. *30keV STEM image from the center of the first fully build RAVENS test chip. Visible are three ReRAM elements embedded between the tungsten M1 and copper M2. In addition, the M1 lines are connected to dummy poly silicon lines via CA.*

Inline Test (ILT) data confirms a successful fabrication of CMOS transistors low, regular, and high threshold voltage FETs and 1.8, 2.5 and 3.3V IO FETs. The later series of transistors is used for standard IO communication and is used to drive the ReRAM devices. In addition, the metallization layers do not show any shorts and opens across the usable part of the wafer and resistance values exceed expectations (tungsten M1 increases the line resistance by a factor of 6). An example of the transistor performance can be seen in **Figure 27** where the on-currents (I_{on}), off-currents (I_{off}) and the threshold voltages (V_{tlin}) are mapped. The three parameters as well as all monitored electrical metrics fall within the boundary conditions set by the PDK setting the ground for successful evaluation of the integrated circuits.

Initial characterization of the 1T1R test structures show excellent ReRAM performance with respect to set and reset voltages as well as on/off ratio (**Figure 28**). For this first pass, the forming voltage shows higher than expected values of 2.5 to 4 V which are beyond the highest IO voltage (V_{DD} = 3.3 V). However, temporary raising of the IO V_{DD} to 4 V during the forming event does not introduce any lasting damage to either the transistors or the ReRAM devices. Even with the higher forming voltage, the R_{off}/R_{on} ratio surpasses 10 for the 10000 measured cycles on the majority of dies.

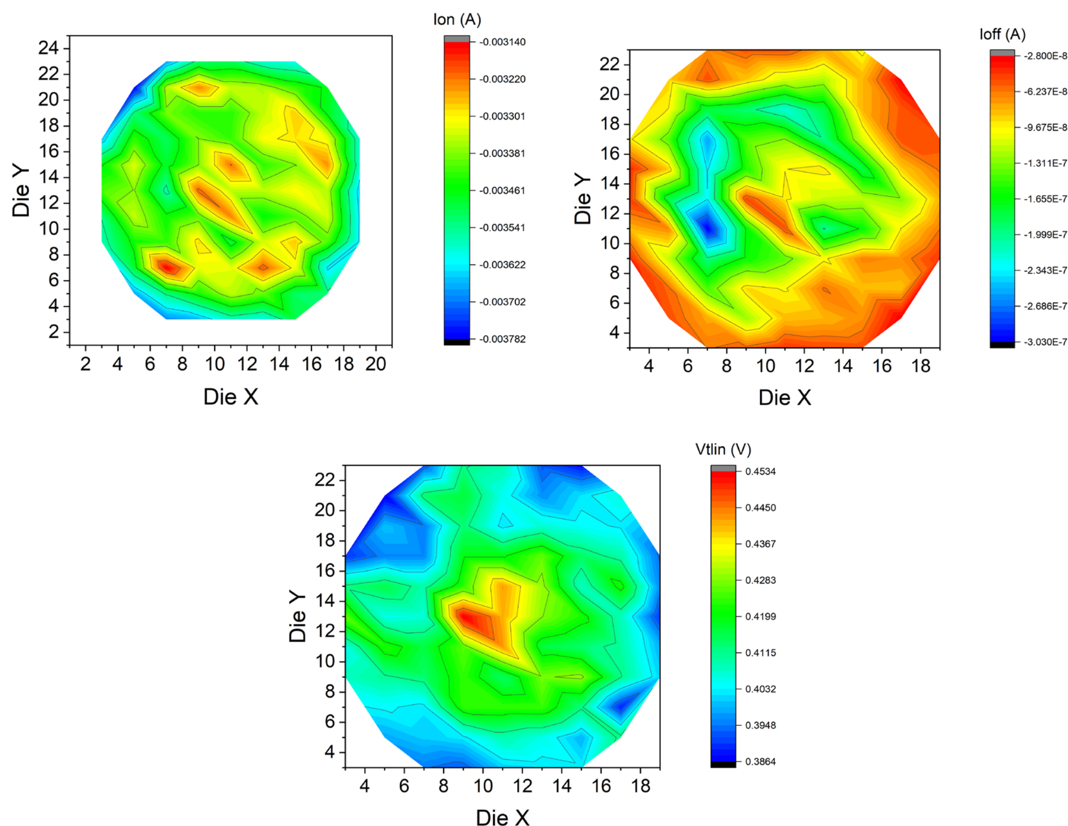


Figure 27. *Exemplary inline test data for a 5x minimum width low threshold voltage NFET. Top left, top right, and bottom show on current, off current and threshold voltage, respectively.*

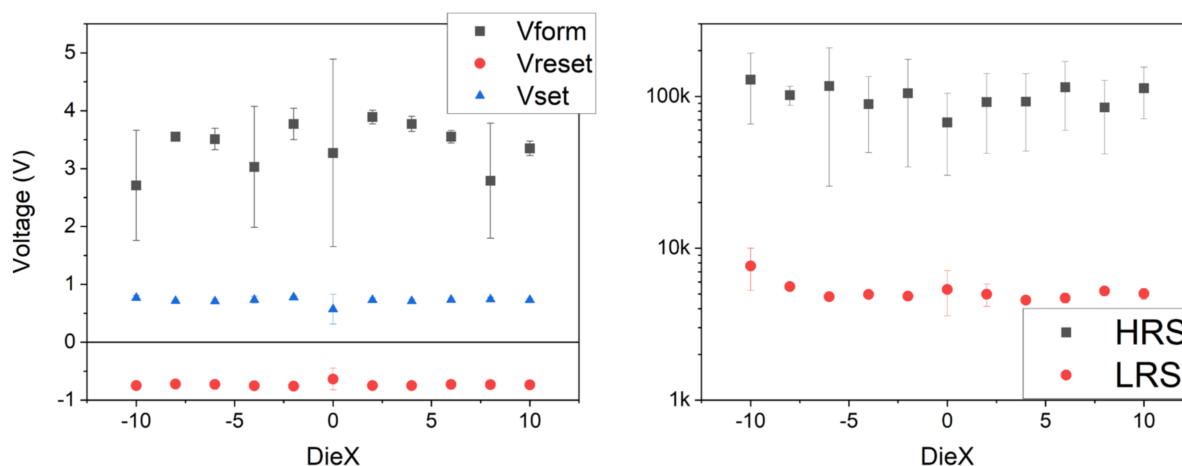


Figure 28. *Initial edge-center-edge pulse-based analysis of 1T1R test structures. 11 dies with each 5 devices for a total of 55 devices were measured. The devices were 10000 times cycled after an initial forming pulse. A rise and fall time of 10 µs was with +5, +2 and -1.5 V for the forming, set and reset voltages. The current during the forming and set event was limited to 100 µA.*

7.0 Conclusions

The goal of this effort was to fabricate hybrid CMOS/RRAM chips that could be used for a variety of Air Force specific applications, using a low-power, reconfigurable framework. Working with our collaborators at UT-Knoxville and a number of other institutions, the SUNY Poly team was able to compile a design containing an unique neuromorphic, reconfigurable test chip (UT-Knoxville design), a variety of hybrid CMOS/RRAM demonstration circuits (from UT-Austin, ASU), and a neuromorphic computing demonstrator chip (from UT-San Antonio). In addition, the SUNY Poly team designed multiple memory arrays for demonstration of compute in memory operations and analog information encoding. To this end, the SUNY Poly team also developed optimized RRAM testing parameters to 1) maintain high device yield and endurance, 2) enable a large device memory window, and 3) perform analog information encoding using sub-nanosecond pulses. Together, this work provides a framework for ongoing testing and evaluation of hybrid CMOS/RRAM circuits by the SUNY Poly team and collaborators. Full testing of the circuits and chips fabricated under this effort is underway at the partnering institutions and will continue under the ARAP NeuroPipes program.

Finally, through this AFRL program, SUNY Poly has also developed the following industry outreach efforts & partnerships:

Tokyo Electron (TEL)

- Hafnium oxide materials optimization for RRAM
- Hafnium zirconium oxide ferroelectric material & device integration for alternative memristors (e.g. ferroelectric tunnel junction / FTJ)
- Niobium oxide deposition and modification (in situ) procedures for “neuristor” fabrication efforts.

Xallent, LLC

- Utilize SUNY Poly baseline 1T1R process to develop novel RRAM technology and high frequency switching applications (through a DARPA-funded program)

IBM

- Utilize SUNY Poly 1T1R devices to perform benchmarking studies between hafnium oxide and tantalum oxide devices and determine readiness for neural network training and inference.

8.0 Publications and Patent Applications Resulting from this Project

1. M. Abedin, A. Roohi, M. Liehr, **N. Cady**, S. Angizi, MR-PIPA: An Integrated Multi-level RRAM (HfOx) based Processing-In-Pixel Accelerator. (2022) *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*. *Online Publication* https://doi.org/10.1109/JXCDC.2022.3210509
2. M. Liehr, K. Beckmann, **N. Cady**. Impact of Switching Variability, Memory Window, and Temperature on Vector Matrix Operations Using 65nm CMOS Integrated Hafnium Dioxide-based RRAM Devices. (2022) *IEEE 31st Microelectronics Design & Test Symposium (MDTS), 2022*. https://doi.org/10.1109/MDTS54894.2022.9826924
3. G. Krishnan, L. Yang, J. Sun, J. Hazra, X. Du, M. Liehr, Z. Li, K. Beckmann, R. Joshi, **N.C. Cady**, D. Fan, Y. Cao. Exploring Model Stability of Deep Neural Networks for Reliable RRAM-based In-Memory Acceleration. (2022) *IEEE Transactions on Computers*, https://doi.org/10.1109/TC.2022.3174585
4. M. Abedin, M. Liehr, K. Beckmann, J. Hazra, S. Rafiq, **N. C. Cady**. In-memory Computation of Error-Correcting Codes Using a Reconfigurable HfOx ReRAM 1T1R Array. (2021) *2021 IEEE International Midwest Symposium on Circuits and Systems (MWSCAS)*. p. 593-598, doi: https://doi.org/10.1109/MWSCAS47672.2021.9531717
5. G. Krishnan, J. Sun, J. Hazra, X. Du, M. Liehr, Z. Li, K. Beckmann, R. Joshi, **N. Cady**, Y. Cao. Robust RRAM-based In-Memory Computing in Light of Model Stability. (2021) 2021 IEEE International Reliability Physics Symposium (IRPS), 2021. p. 1-5, doi: https://doi.org/10.1109/IRPS46558.2021.9405092
6. G. Krishnan, J. Hazra, M. Liehr, X. Du, K. Beckmann, R. Joshi, **N. Cady**, Y. Cao. Design Limits of In-Memory Computing: Beyond the Crossbar. (2021) 2021 5th IEEE Electron Devices Technology & Manufacturing Conference (EDTM). pp. 1-3. doi: 10.1109/EDTM50988.2021.9421057
7. J. Hazra, M. Liehr, K. Beckmann, M. Abedin, S. Rafiq, **N.C. Cady.** Optimization of Switching Metrics for CMOS Integrated HfO_2 based Bipolar RRAM Devices on 300 mm Wafer Platform. (2021) *IEEE International Memory Workshop (IMW) 2021.* 1-4. DOI: 10.1109/IMW51353.2021.9439618
8. Rafiq, J. Hazra, M. Liehr, K. Beckmann, M. Abedin, J.S. Pannu, S.K. Jha, **N.C. Cady.** Investigation of ReRAM variability on flow-based edge detection computing using HfO_2-based ReRAM arrays. (2021) *IEEE Transactions on Circuits and Systems.* https://doi.org/10.1109/TCSI.2021.3072210
9. M. Liehr, J. Hazra, K. Beckmann, S. Rafiq and **N. Cady.** Impact of Switching Variability of 65nm CMOS Integrated Hafnium Dioxide-based ReRAM Devices on Distinct Level Operations (2020) *2020 IEEE International Integrated Reliability Workshop (IIRW)*, South Lake Tahoe, CA, pp. 1-4, https://doi.org/10.1109/IIRW49815.2020.9312855
10. J. Hazra, M. Liehr, K. Beckmann, S. Rafiq and N. Cady, Impact of Atomic Layer Deposition Co-Reactant Pulse Time on 65nm CMOS Integrated Hafnium Dioxide-based Nanoscale RRAM Devices. (2020) *2020 IEEE International Integrated Reliability Workshop (IIRW)*, South Lake Tahoe, CA, pp. 1-4, https://doi.org/10.1109/IIRW49815.2020.9312877

11. G. Charan, J. Hazra, K. Beckmann, X. Du, G. Krishnan, R.V. Joshi, **N.C. Cady**, Y. Cao. Accurate Inference with Inaccurate RRAM Devices: Statistical Data, Model Transfer, and On-line Adaptation. 2020 57th ACM/IEEE Design Automation Conference (DAC), San Francisco, CA, USA, 2020, pp. 1-6, https://doi.org/10.1109/DAC18072.2020.9218605

12. J. S. Pannu, S. Raj, S.L. Fernandes, D. Chakraborty, S. Rafiq, **N. Cady,** S.K. Jha. Design and Fabrication of Flow-based Edge Detection Memristor Crossbar Circuits. (2020) *IEEE Transactions on Circuits and Systems II: Express Briefs,* 67(5), 961-965. https://doi.org/10.1109/TCSII.2020.2984155

13. K. Beckmann, W. Olin-Ammentorp, C. Gangotree, S. Amer, G. Rose, J. Van Nostrand, **N.C. Cady.** Towards synaptic behavior of nanoscale ReRAM devices for neuromorphic computing applications. (2020) *ACM Journal on Emerging Technologies in Computing (JETC) Special Issue on New Trends in Nanoelectronic Device, Circuit and Architecture Design.* 16(3): 23. https://doi.org/10.1145/3381859

14. J. Hazra, M. Liehr, K. Beckmann, S. Rafiq, **N. Cady**. Improving the Memory Window/Resistance Variability Trade-Off for 65nm CMOS Integrated HfO2 Based Nanoscale RRAM Devices. (2020) *IEEE International Integrated Reliability Workshop (IIRW)*, South Lake Tahoe, CA, USA, p1-4. https://doi.org/10.1109/IIRW47491.2019.8989872

Patents / Technology Disclosures

1. Research Foundation for SUNY Ref No. 011-2138 (New Technology Disclosure), "Defect Density Manipulation in HfO2 for ReRAM Application via Angular Velocity Adjustments in a Semi-batch ALD Chamber", disclosed 10/14/2020, by Karsten Beckmann, Nathaniel Cady, and Jubin Hazra.

Approved for Public Release; Distribution Unlimited.

9.0 List of Acronyms

1T1R:	1 transistor 1 memristor (memory cell containing 1 transistor and 1 memristor)
AFRL:	Air Force Research Laboratory
AI:	Artificial Intelligence
ASU:	Arizona State University
BE:	Bottom electrode
CMOS:	Complementary Metal Oxide Semiconductor
DARPA:	Defense Advanced Research Projects Agency
FET:	Field Effect Transistor
FPGA:	Field Programmable Gate Array
GUI:	Graphical User Interface
HRS:	High resistance state
I&F:	Integrate and fire
I/O:	Input/Output
LRS:	Low resistance state
NMOS:	N-Type Metal Oxide Semiconductor
R&D:	Research & Development
ReRAM:	Resistive random access memory
RISC-V:	Reduced Instruction Set Computer - Five
RMD:	Resistive memory device
RRAM:	Resistive random access memory
SUNY:	State University of New York
TE:	Top electrode
UAV:	Unmanned Aerial Vehicle
UT-A:	University of Texas - Austin
UT-SA:	University of Texas – San Antonio
Vg:	Gate voltage