AFRL-RI-RS-TR-2021-067

NANOSCALE ELECTRONICS DEVELOPMENT AND FABRICATION (RAPIDFAB)

SUNY POLYTECHNIC INSTITUTE COLLEGES OF NANOSCALE SCIENCE AND ENGINEERING

APRIL 2021

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-2021-067 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE CHIEF ENGINEER:

/ S /
JOSEPH E. VAN NOSTRAND
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

Form Approved
OMB No. 0704-0188

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
APRIL 2021	FINAL TECHNICAL REPORT	OCT 2017 – OCT 2020

4. TITLE AND SUBTITLE

NANOSCALE ELECTRONICS DEVELOPMENT AND FABRICATION (RAPIDFAB)

5a. CONTRACT NUMBER FA8750-18-2-0022

5b. GRANT NUMBER N/A

5c. PROGRAM ELEMENT NUMBER 61102F

6. AUTHOR(S)

Nathaniel C. Cady

5d. PROJECT NUMBER T2DY

5e. TASK NUMBER FA

5f. WORK UNIT NUMBER B2

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

SUNY Polytechnic Institute Colleges of Nanoscale Science and Engineering
257 Fuller Rd
Albany NY 12203-3613

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-2021-067

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 provide nanoscale electronic device design, layout, fabrication and testing services for AFRL projects and programs. This effort leveraged multiple ongoing efforts between SUNY Polytechnic Institute (SUNY Poly) and AFRL, including, but not limited to CMOS and resistive memory (ReRAM, memristor) platforms. The ultimate goal was delivery of cutting-edge nanoelectronic devices/systems and performance data to AFRL for support of R&D programs. The major outcomes of this effort include development of novel memristive materials/stacks, extending 300mm wafer based integration of CMOS/ReRAM to the far back end of the line (BEOL), characterization of CMOS/ReRAM device performance, and finally, evaluation of the impact of device variability and methods to mitigate the effects of such variability.

15. SUBJECT TERMS

Memristor, FPGA, CMOS, nanoelectronics

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
					JOSEPH E. VAN NOSTRAND
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER (*Include area code*)
U	U	U	UU	26	**N/A**

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

Table of Contents

1.0 SUMMARY **1**

2.0 INTRODUCTION **2**

3.0 METHODS, ASSUMPTIONS AND PROCEDURES **2**

3.1 FABRICATION 2
3.2 DEVICE TESTING 3

4.0 RESULTS AND DISCUSSION **6**

4.1 KEY ACCOMPLISHMENTS 6
4.1.1 Development of Novel Memristive Material Stacks 6
4.1.2 Integrated CMOS/ReRAM Development and Testing 8
4.1.3 Investigating Memristor Variability for Integrated CMOS/ReRAM 11
4.1.4 Assessing the Effects of Memristive Device Var. on Neural Networks 15

5.0 CONCLUSIONS **18**

6.0 PUBLICATIONS AND PATENT APPLICATIONS RESULTING FROM THIS PROJECT **19**

7.0 LIST OF ACRONYMS **20**

List of Figures / List of Tables

Figure 1. *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.* 2

Figure 2. *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

Figure 3. *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).* 4

Figure 4. *(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.* 5

Figure 5. *Memristor device stack consisting of a platinum bottom electrode, a two-layer switching dielectric (tantalum oxide and cerium oxide), a hafnium "oxygen exchange layer", and tungsten top electrode.* 6

Figure 6. *TaOx/CeOx multi-layer memristors showing multi-level switching behavior. These devices were 50x50 microns square and could be set at 2.5V and reset at -2.3V. To achieve multi-level switching, devices were initial set (LRS) and then reset pulses were applied with varying pulse-width. Pulse width ranged from 500ns to 1.5 microseconds, yielding the intermediate (multi-level) resistance states shown as different colrs in the figure. The top plot shows the HRS resistance level (achieved by varying pulse-width) while the bottom plot shows the corresponding LRS level for each cycle.* 7

Figure 7. *Endurance and retention data for the TaOx/CeOx memristors fabricated during this project.* 8

Figure 8. *a) Endurance measurement for 1T1R cell for up to 1 billion switching cycles, and b) HRS and LRS retention up to 10^4 seconds at 100°C.* 9

Figure 9. *Example of an 8x8 one-transistor/one-ReRAM (1T1R) array designed by the Cady group and fabricated in a hybrid 65 nm CMOS process at the SUNY Poly 300mm foundry.* 9

Figure 10. *BEOL integration on mrDANNA chips/wafers in the SUNY Poly 300mm fabrication facility. A) Demonstration of BEOL layers M1 through M3, and B) demonstration of additional metallization layers BA through BB, with the WA interconnect/via layer.* 10

Figure 11. *Image of wafer fabricated through LB level (aluminum pads in BEOL).* 11

Figure 12. *Hafnium oxide based 1T1R memory retention study. ReRAM were set to either LRS or HRS and then measured with a sub-threshold pulse (0.2 V) over a 10 day period.* 12

Figure 13. *Pulse-based switching of hafnium oxide 1T1R cells in 250 Ω increments using -1.1 V pulses with 2 ns rise/fall and 3 ns pulse width. Conductance value vs. the number of applied pulses is shown.* 13

Figure 14. *Analog switching performance for hafnium oxide 1T1R ReRAM as a function of maximum compliance current (left) and pulse width (right). As ReRAM cells are switched to higher resistance states (left), the magnitude of resistance is directly related to the peak compliance current applied, as controlled by Vg on the control transistor (red vs. blue curve). When the pulse width is increased from the 1.5 ns to 50 ns, incremental resistance changes are larger, as well as the variation in resistance state achieved.* 13

Figure 15. *Retention study for 1T1R ReRAM set to different resistance levels (via pulse-based analog switching).* 14

Figure 16. *Example of the graph of a spiking network evolved to solve a pole-balancing task. Each edge between neurons has a weight attached to it, which is being perturbed to determine the network's sensitivity to synaptic noise.* 16

Figure 17. *Results of perturbations made to noisy and noise-free networks. The noisy networks decay much less quickly than the noise-free networks, and the magnitude of perturbation required to reduce them to 50% performance is approximately 5 times greater.* 17

1.0 Summary

This technical report summarizes the R&D efforts for the AFRL project "Nanoscale Electronics Development and Fabrication. This research project was enabled to provide nanoscale electronic device design, layout, fabrication and testing services for AFRL projects and programs. This effort leveraged multiple ongoing efforts between SUNY Polytechnic Institute (SUNY Poly) and AFRL, including, but not limited to CMOS and resistive memory (ReRAM, memristor) platforms. The ultimate goal was delivery of cutting-edge nanoelectronic devices/systems and performance data to AFRL for support of R&D programs. The major outcomes of this effort include development of novel memristive materials/stacks, extending 300mm wafer based integration of CMOS/ReRAM to the far back end of the line (BEOL), characterization of CMOS/ReRAM device performance, and finally, evaluation of the impact of device variability and methods to mitigate the effects of such variability.

2.0 Introduction

The memristor is a two-terminal resistive switching memory (RRAM) nanoscale device, which is recognized as the fourth fundamental electrical element in addition to resistors, capacitors, and inductors. By utilizing different resistance values to store information, memristors can function as low-power, highly-scalable device elements – critically essential properties for memory and field programmable gate array applications. More importantly, memristors have exhibited unique self-learning properties and can execute complex logic functions, which enables innovative logic, data processing, and implementation of neuromorphic systems. And, in contrast with other emerging nanoelectronic devices, memristors can utilize CMOS-compatible oxide and electrode materials, which dramatically simplifies CMOS-memristor integration. These factors underscore the promise and feasibility of large-scale CMOS-memristor hybrid nanoelectronics and point to their significant impact on future IC applications.

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 1*** shows the cross-section of a circuit implemented in the CNSE CMOS/Memristor process under previous efforts.

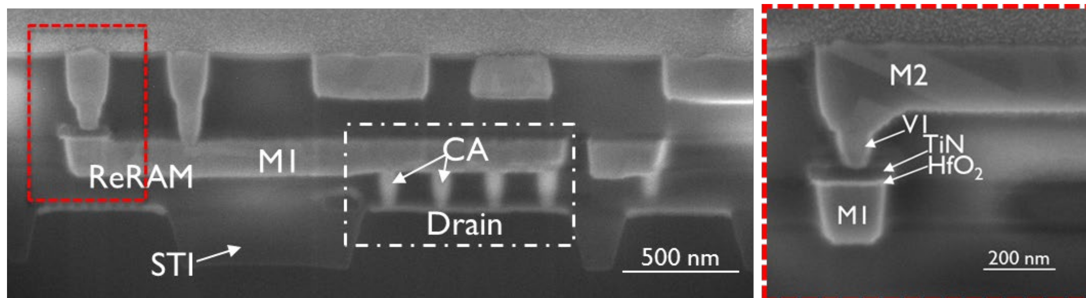


Figure 1. *Cross-section of a vertically integrated transistor / memristor (1T1R) device fabricated at CNSE. The bottom electrode of the memristor (ReRAM) device is connected to the drain contact stud (CA) via the M1 line. The figure at right is a cross-section of the memristor element alone, showing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

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. 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 1*** & ***Figure 2***.

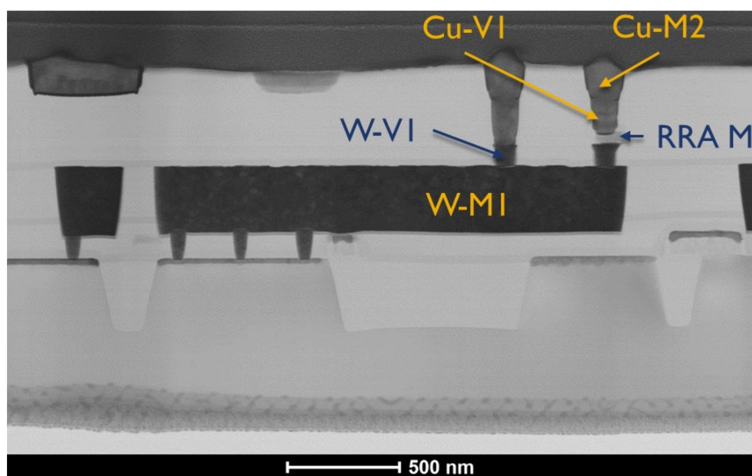


Figure 2. *Cross-sectional contrast image of our integrated CMOS/ReRAM circuit taken with a transmission electron microscope. This cross-section shows the W bottom electrode and ReRAM device, as fabricated in a FEOL-compatible process.*

Additional fabrication approaches were used to fabricate individual memristor / ReRAM devices using the 200mm fabrication facilities at SUNY Polytechnic Institute. The fabrication approach for these devices included fabrication of blanket bottom electrodes, switching layers and top electrode, followed by photolithographic patterning and reactive ion etching (and/or lift-off based patterning) to generate micrometer-scale memristive devices.

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 or other wafer sizes. 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 enabling 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.

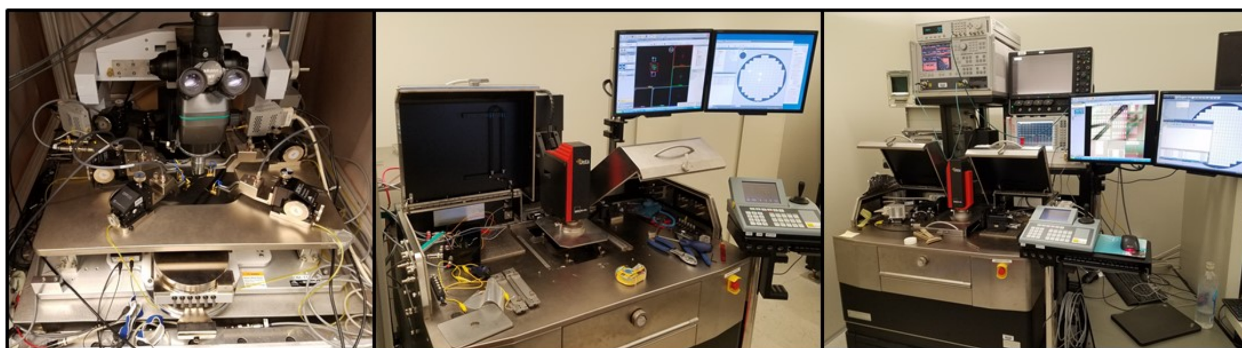


Figure 3. *Probe stations used for this effort. Manual probe station with B1500A analyzer (left), Suss Microtech semi-automatic probe station with Keysight E5270A, and Suss Microtech semi-automatic probe station with B1500A/B1530A (right).*

Electrical measurements primarily utilized an on-chip NMOS field-effect transistor (NFET) for current control during the forming and set operation, as seen in ***Figure 4*** (a). ***Figure 4*** (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.

Approved for Public Release; Distribution Unlimited.

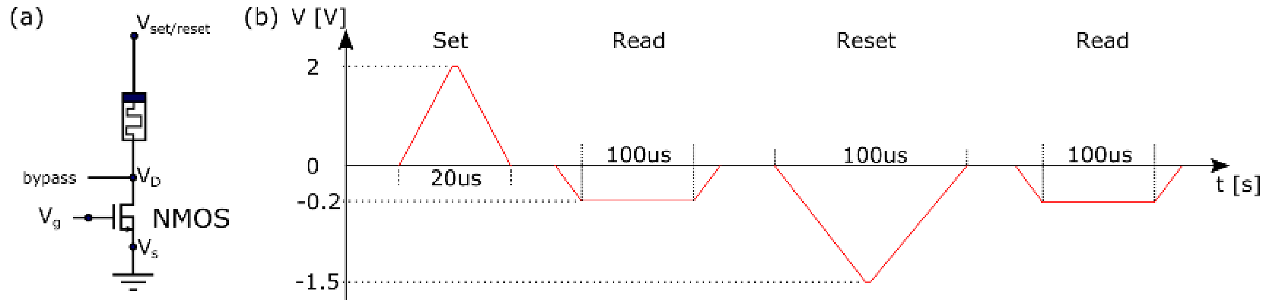


Figure 4. *(a) Schematic of a 1T1R structure with an NMOS that acts as the current limiting device during the forming and set operations. A parasitic base diode opens during the reset allowing for a higher current during the reset operation. The bypass connection enables the direct measurement of the 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.

Approved for Public Release; Distribution Unlimited.

4.0 Results and Discussion

4.1 Key Accomplishments

4.1.1 Development of Novel Memristive Material Stacks

One of the limiting factors for hafnium oxide memristors (in both our previous work and work by other groups) is that these devices can suffer from large variability when switching to incremental resistance states (vs. binary switching between a single low resistance state and a single high resistance state). This is a challenge for neuromorphic systems in which memristors are typically used to set the so-called "synaptic weight" for connections between neurons. Thus, we have been developing alternative memristor devices that exhibit intermediate switching behavior (>4 switching levels between the highest and lowest resistance states), with low switch to switch variability between those levels.

Figure 5 shows one of the memristor device stacks that was used during the performance period. Our previous work demonstrated that tantalum oxide memristors could achieve >10 resistance levels when modulating the peak voltage or pulse width during pulse-based switching[1]. One of the theories for why tantalum oxide devices can be switched to a larger number of intermediate states, with better switch-to-switch uniformity is that their fundamental switching properties are different than hafnium oxide devices, primarily in how the conductive filament is formed, and how oxygen vacancies are distributed (and re-distribute) within the oxide films.

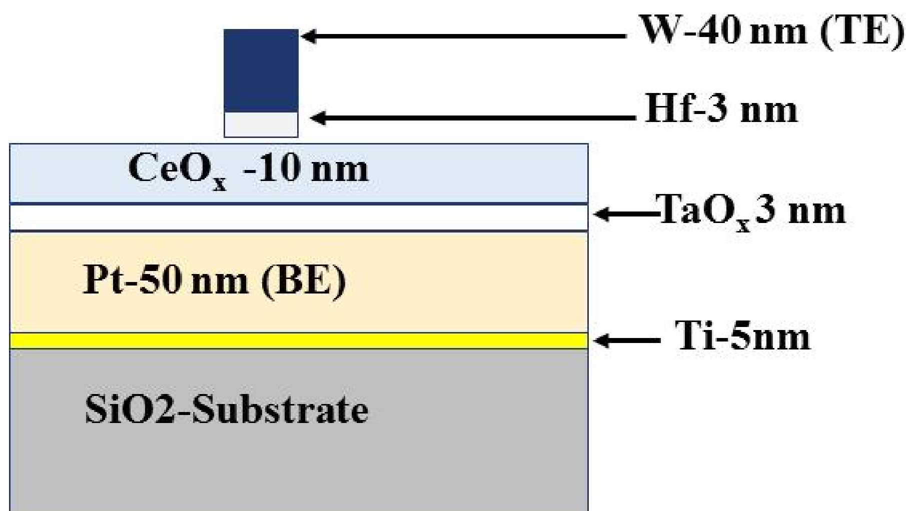


Figure 5. *Memristor device stack consisting of a platinum bottom electrode, a two-layer switching dielectric (tantalum oxide and cerium oxide), a hafnium "oxygen exchange layer", and tungsten top electrode.*

During this project, we fabricated and electrically tested devices with the cross-sectional architecture shown in ***Figure 5***. These devices showed excellent resistive switching behavior, as

---

[1] 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 http://dx.doi.org/10.1063/1.4993058

well as multi-level switching that was stable and reproducible for 5 different resistance states as shown in ***Figure 6***.

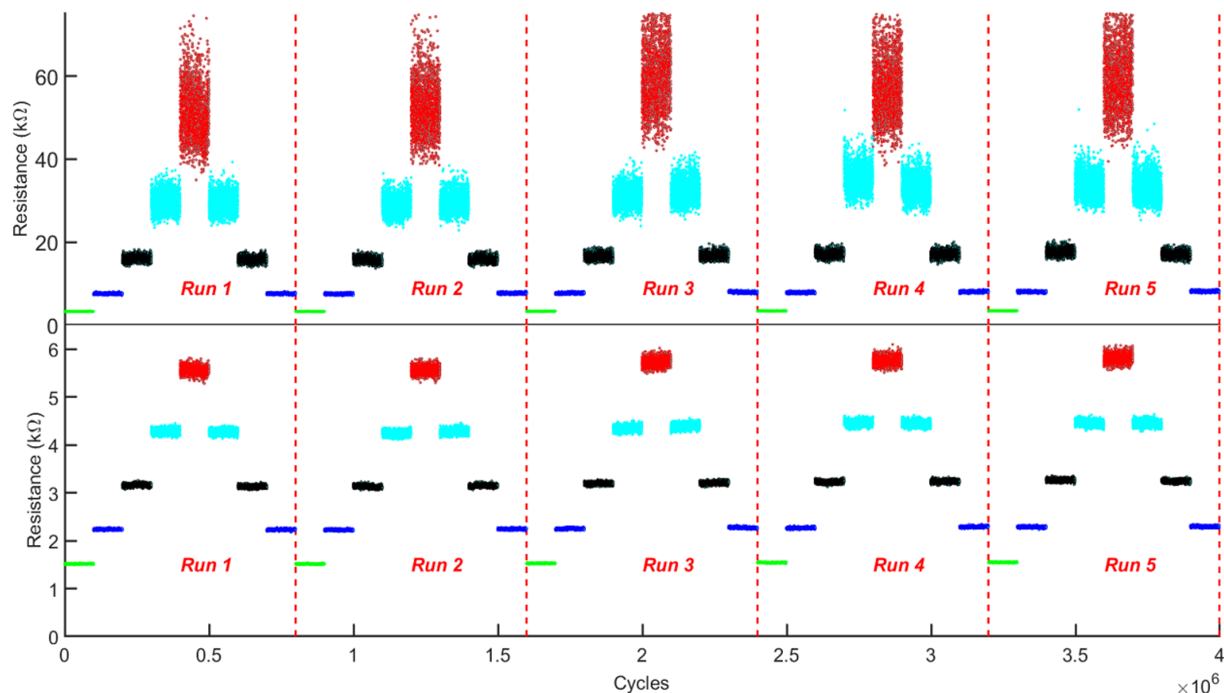


Figure 6. *TaOx/CeOx multi-layer memristors showing multi-level switching behavior. These devices were 50x50 microns square and could be set at 2.5V and reset at -2.3V. To achieve multi-level switching, devices were initial set (LRS) and then reset pulses were applied with varying pulse-width. Pulse width ranged from 500ns to 1.5 microseconds, yielding the intermediate (multi-level) resistance states shown as different colrs in the figure. The top plot shows the HRS resistance level (achieved by varying pulse-width) while the bottom plot shows the corresponding LRS level for each cycle.*

The TaOx/CeOx devices tested during this period also showed excellent retention properties, as shown in ***Figure 7***. These devices could be switched up to 10^11 times without breakdown. Some variation in the HRS was observed after 10^9 cycles, but was highly uniform for all prior cycles. These devices also showed retention of >10^4 seconds (7 days). Taken together, these data suggest that this multi-layer memristor has excellent reliability and would be a good candidate for further study, for integration into systems requiring non-volatile memory elements, or multi-level switching characteristics (such as neuromorphic systems).

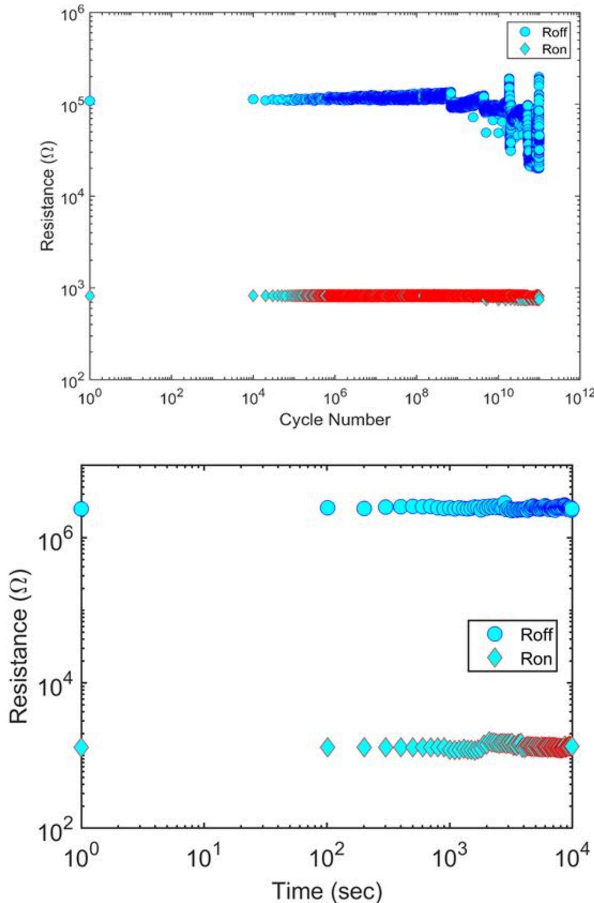


Figure 7. *Endurance and retention data for the TaOx/CeOx memristors fabricated during this project.*

4.1.2 Integrated CMOS/ReRAM Development and Testing

For our ongoing ReRAM fabrication efforts (and integration with CMOS), we have been able to achieve close to 100% yield 1T1R cells (across an entire 300mm wafer) with significantly lower cell-to-cell variability by optimizing deposition conditions for the HfO_2 switching layer of ReRAM cells. The key variable that we have investigated is the deposition method used to form the hafnium oxide switching layer. We utilized atomic layer deposition (ALD) methods with different precursors (ranging from metal-organic, to metal halide based percursors). We found that chlorine based hafnium oxide ALD precursors resulted in the best performance, with respect to yield, endurance, retention, and analog behavior. 1T1R cells fabricated with chlorine based precursors were investigated for high temperature retention up to 100 °C and long-term endurance, exhibiting excellent endurance of up to 1 billion switching cycles with an average R_{off}/R_{on} ratio of 10:1 (***Figure 8***). We have also performed full wafer measurements of ReRAM cells showing excellent consistency for threshold voltages and resistance values across full 300 mm wafer.

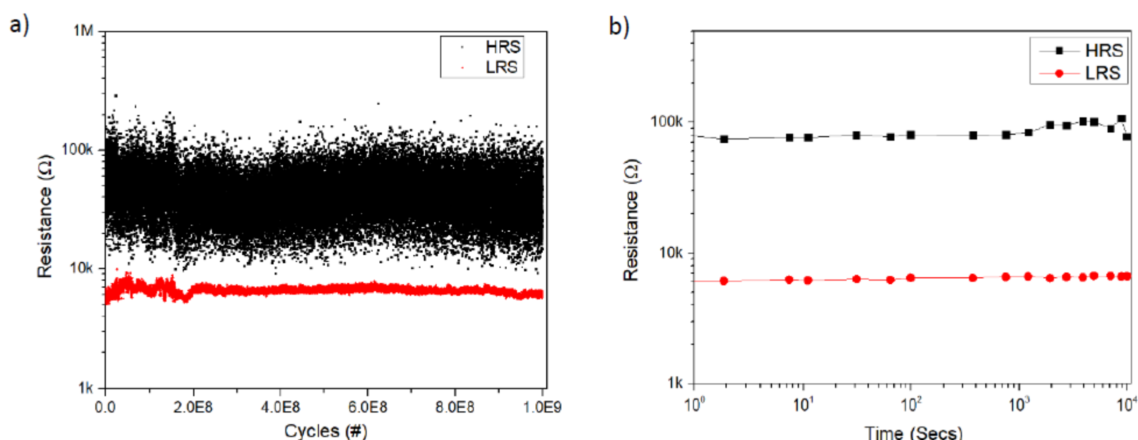


Figure 8. *a) Endurance measurement for 1T1R cell for up to 1 billion switching cycles, and b) HRS and LRS retention up to 10^4 seconds at 100°C.*

In addition to optimizing memristor performance, we also focused on utilizing the 8x8 1T1R memory arrays that we designed in our AFRL mrDANNA project. These arrays do not have a decoder circuit, and are addressed using a 12 x 2 probe card and switching matrix. Our results show that we can achieve 100% yield of 1T1R cells within these arrays, and that every cell in the array can be set in analog fashion (***Figure 9***). The array shown in ***Figure 9*** (left) was programmed (set) with two different current compliance levels to form the AFRL logo. The array shown on the right of ***Figure 9*** was then programmed (set) with a gradient of current compliance levels, according to the intensity of each pixel in a greyscale image. This demonstrates the multi-level / analog capabilities of our 1T1R cells. For this project, the ability to read/write data into 1T1R arrays is paramount, since we will eventually need to combine hybrid CMOS/memristor neuron circuits with memory arrays, to achieve fully-functional neuromorphic circuits (for the larger mrDANNA architecture, and for our follow-on efforts in our RAVENS project). Thus, demonstrating 100% yield and analog switching in these arrays is an important step.

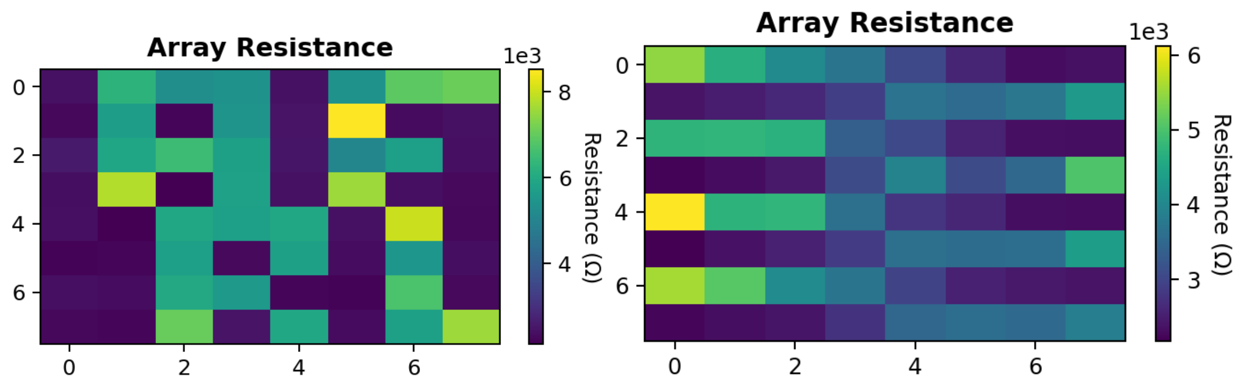


Figure 9. *Example of an 8x8 one-transistor/one-ReRAM (1T1R) array designed by the Cady group and fabricated in a hybrid 65 nm CMOS process at the SUNY Poly 300mm foundry.*

In addition to individual device and array testing, we have also focused on completing the back end of the line (BEOL) integration flow for the mrDANNA design / mask-set. To this end, we focused on finalizing the metallization modules for the following layers: metal 3 (M3), 1st level 2X metal line (BA), connecting via (WA) 2nd level 2X metal line (BB), and the via /aluminum wiring level (VV/LB). The final layer (LB) is an aluminum layer that will enable contact pads for packaging. In addition, these layers will enable full testing of the neuron circuits in the mrDANNA design, as well as reconfiguration of circuits (i.e., connecting multiple neurons and even 1T1R arrays on the mrDANNA chip, to demonstrate more complex functionality). We implemented and optimized each of these layers individually, and are now running wafers to yield the entire flow (CMOS through the full BEOL stack) ***Figure 10*** & ***Figure 11***.

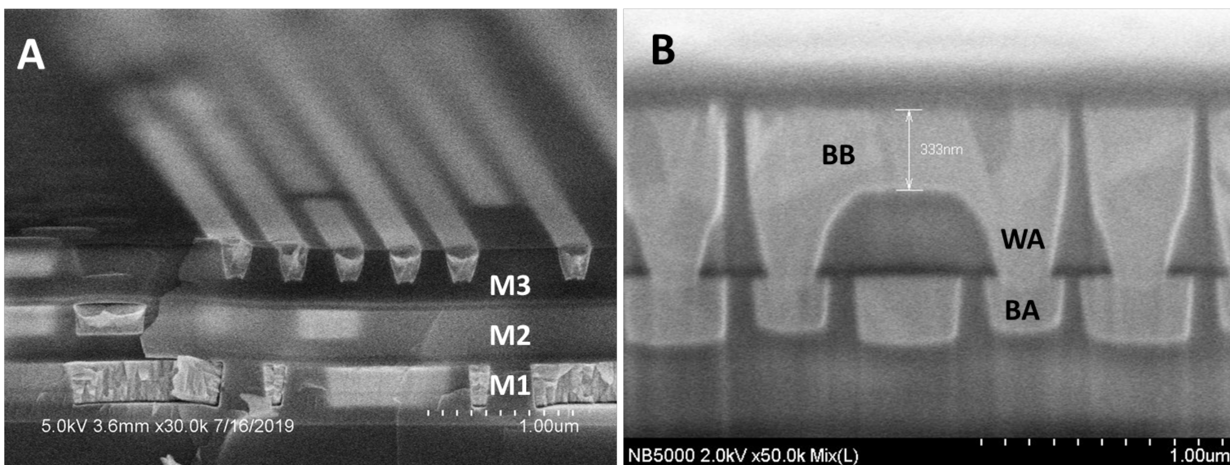


Figure 10. *BEOL integration on mrDANNA chips/wafers in the SUNY Poly 300mm fabrication facility. A) Demonstration of BEOL layers M1 through M3, and B) demonstration of additional metallization layers BA through BB, with the WA interconnect/via layer.*

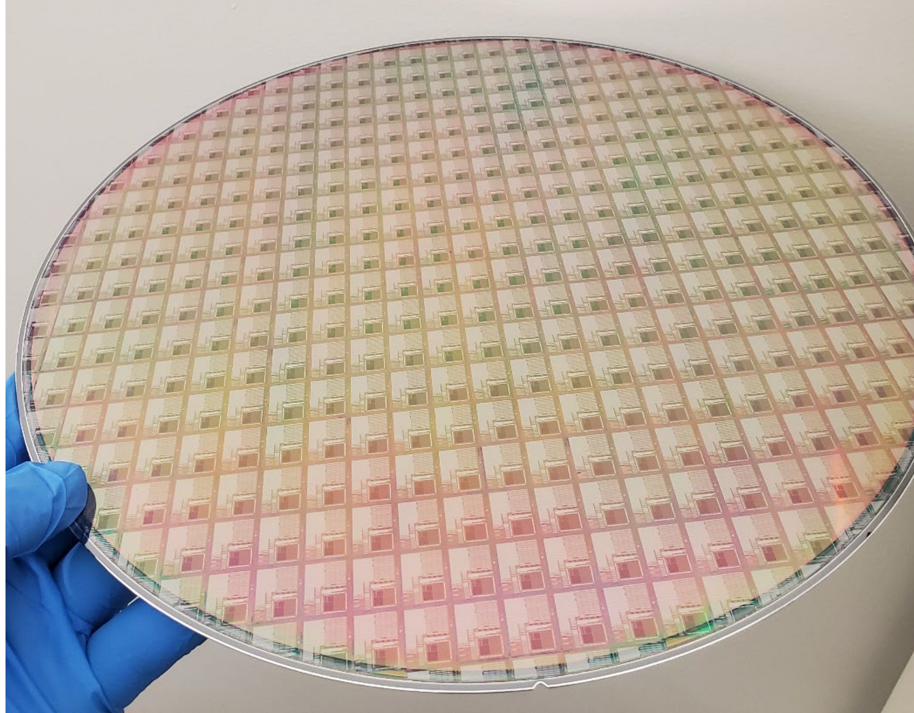


Figure 11. *Image of wafer fabricated through LB level (aluminum pads in BEOL).*

4.1.3 Investigating Memristor Variability for Integrated CMOS/ReRAM Hardware

In this portion of the research effort, we focused on investigating ReRAM device variability, in the context of memory retention and analog switching performance. Previously, we showed that when ReRAM are used to encode synaptic weights, the relative variation in their resistance state can affect the accuracy of a neural network. Further, we showed that by training a neural network with so-called "noisy integrate and fire" neurons, we can compensate for some of the variability in ReRAM resistance[2]. These results also prompted us to focus on ReRAM variability and reliability, specifically improvement of these factors.

ReRAM retention (stability of a set resistance value) in neuromorphic circuits (and in general memory applications) is important, since changes in resistance, including drift and bit switching, can cause loss of memory or significant changes to the synaptic values that the ReRAM elements are encoding. We performed a series of ReRAM retention studies to demonstrate the stability of hafnium oxide based 1T1R cells on our 300mm 65nm CMOS platform. ***Figure 12***, below, shows ReRAM retention for 8 devices, over a 10 day period, and for both the HRS and LRS. As we have seen previously, the HRS is much more variable (device-to-device) than LRS. The devices tested showed good retention of memory window for the entire 10 day test, and stability in the LRS state. Ongoing studies are being performed to include a larger number of devices and to accelerate testing by measuring retention at increased temperatures (above room temp).

[2] W. Olin-Ammentorp, N.C. Cady. Training Spiking Networks via Natural Evolutionary Strategies. (2019) *Proceedings of the International Conference on Neuromorphic Systems (ICONS '19). Association for Computing Machinery.* 18:1–6.

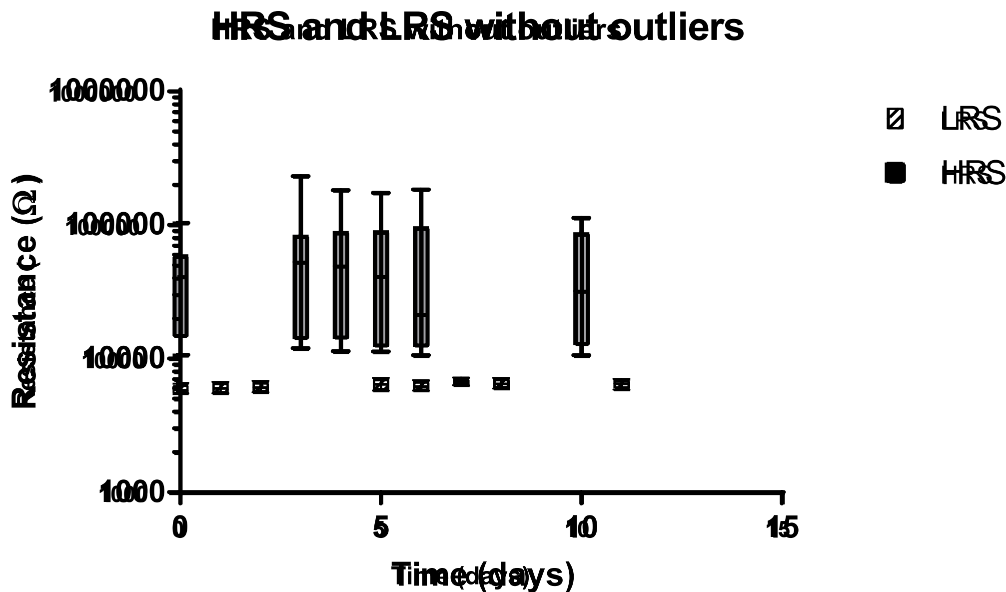


Figure 12. *Hafnium oxide based 1T1R memory retention study. ReRAM were set to either LRS or HRS and then measured with a sub-threshold pulse (0.2 V) over a 10 day period.*

In addition to long-term ReRAM retention studies, we also examined the analog performance of 1T1R cells when using short (< 5 ns) applied pulses (***Figure 13***) and how the peak compliance current (as set by the 1T1R control transistor) and the pulse length affect the resulting magnitude of resistance change, and the variability of the resistance value for repeated switching (***Figure 14***). Our results show that increasing the peak current during switching will increase the magnitude of resistance, but that it also affects device variability (***Figure 14***, left). Further, when the pulse length is increased from 1.5 ns to 50 ns, both the magnitude of resistance change and the variability of the resistance state increase dramatically. These data help us to understand how to control ReRAM variability for analog switching, especially as related to neuromorphic circuits / neural networks.

In our final study, we set devices to different resistance states (using pulsing, followed by a read-verify procedure). Retention of this resistance state was then measured repeatedly (at a sub-threshold voltage) for 20 min. As we have noticed previously, the achieved resistance values are more variable at high resistance levels. This can be seen in ***Figure 15***, where repeated studies of devices set to 8000 Ω (bottom panel) showed more variability than at 6000 Ω (top panel). What was notable, however, is that once devices were set to a particular resistance level, that level was highly stable over the entire 20 min retention test.

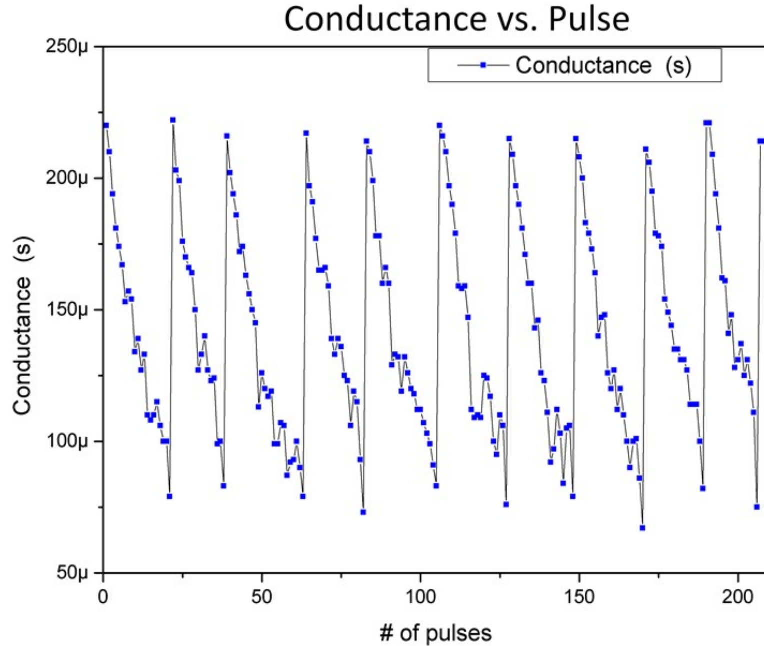


Figure 13. *Pulse-based switching of hafnium oxide 1T1R cells in 250 Ω increments using -1.1 V pulses with 2 ns rise/fall and 3 ns pulse width. Conductance value vs. the number of applied pulses is shown.*

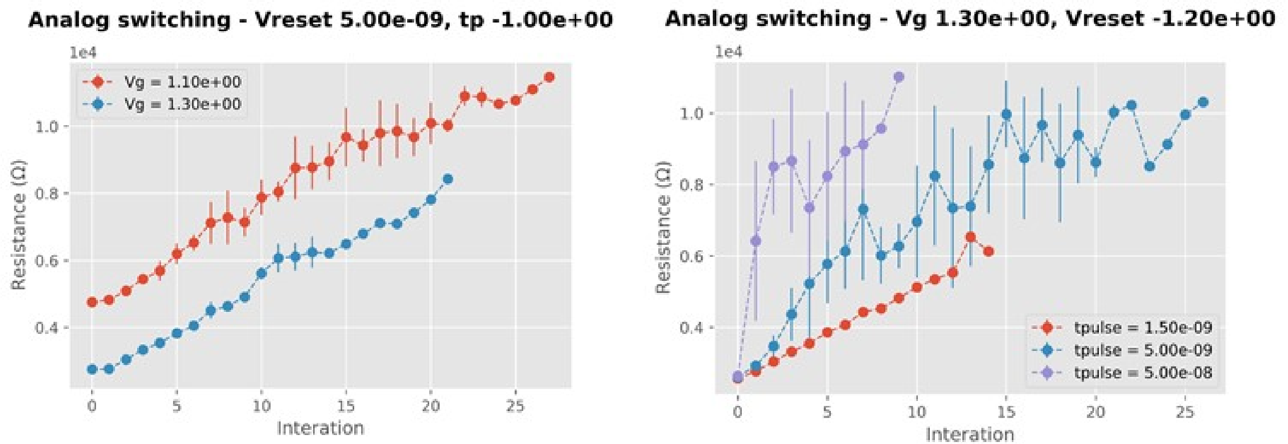
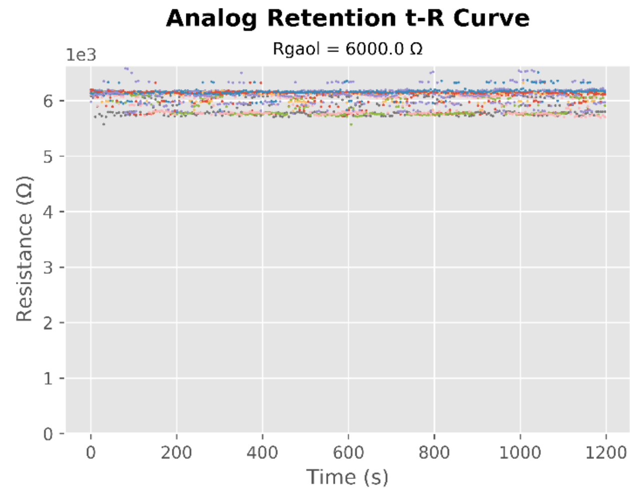
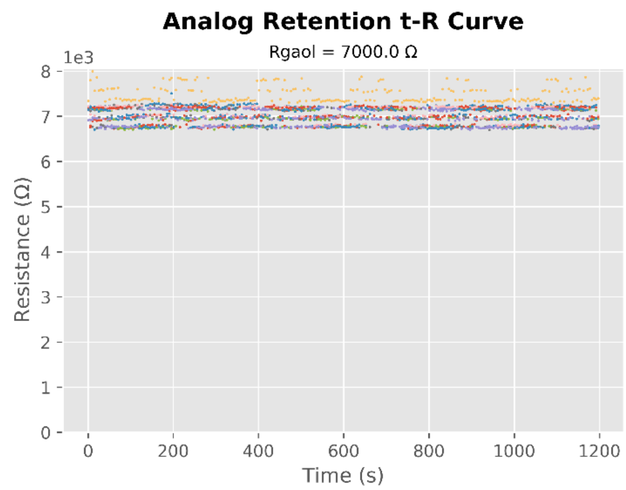


Figure 14. *Analog switching performance for hafnium oxide 1T1R ReRAM as a function of maximum compliance current (left) and pulse width (right). As ReRAM cells are switched to higher resistance states (left), the magnitude of resistance is directly related to the peak compliance current applied, as controlled by Vg on the control transistor (red vs. blue curve). When the pulse width is increased from the 1.5 ns to 50 ns, incremental resistance changes are larger, as well as the variation in resistance state achieved.*





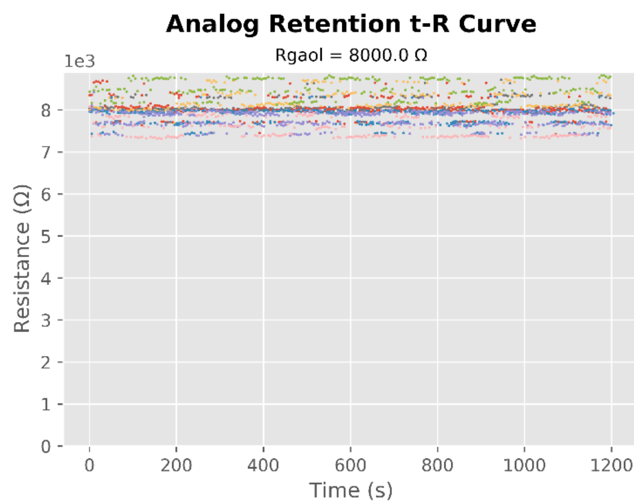


Figure 15. *Retention study for 1T1R ReRAM set to different resistance levels (via pulse-based analog switching).*

4.1.4 Assessing the Effects of Memristive Device Variability on Neural Networks

For a neural network, processing capability is dependent on the weight of these synapses, and many synapses may be incident on a single neuron. While traditional CMOS technologies can be used to implement synapses, emerging memories such as resistive random access memory (ReRAM) offer a unique and potentially revolutionary set of advantages over standard CMOS. ReRAM is a class of memories which operate by different mechanisms, but share the common architecture of a two-terminal device with a resistance that can be either read or modulated depending on specific electrical conditions.

In mrDANNA (the CMOS/RRAM neural network developed under our previous effort), a pair of hafnium oxide (HfOx) based vacancy-change ReRAM elements encode the synaptic weights[3]. These memristors are vertically-integrated between fabrication layers to create a hybrid CMOS-ReRAM process, in so-called 1 transistor 1 ReRAM (1T1R) configuration[4]. As a result, synapses in this architecture have features such as no static power dissipation, persistent memory, and very small footprint. Our 1T1R cells are capable of encoding multiple resistance states, as shown in Figure 6. A current challenge with ReRAM (and many other non-volatile memory cells) is the relative level of variability in the resistance values that it can encode. During binary memory encoding, a so-called memory window of ~10X difference between the high resistance state (HRS) and low resistance state (LRS) can be maintained.

During multi-level switching, however, the variability in the resistance state can make it difficult to differentiate between intermediate resistance levels. This variability can potentially limit the number of well-defined 'levels' of weights which can be used in a synapse, and as a result, impact the functionality of networks running on this architecture.

Biological neuronal networks can carry out their functions despite exhibiting variability[5], and recent research suggests this stochasticity can play an important role in neuronal networks. For instance, it can enable spiking networks to carry out operations such as sampling from a distribution[6]. In this light, we began to carry out investigations of the impacts and uses of variability within spiking architectures.

To begin, we modified an integrate-and-fire neuron and changed its strict integrate-and-fire behavior to one which produces an increasing possibility of firing the closer its charge is to its threshold[7]. As a result, the same set of impulses applied to a neuron with the same synaptic weights will not always produce the same behavior. This creates two distinct sources of variability within the system: variability from the memristive synapses, and variability within the neuron's potential.

While the introduction of additional variability might seem to only make a challenging situation worse, we find that noise on the neuron can counteract noisy synapses to make the system more robust. Even if weights programmed into the memristive synapses stray from their ideal values, the slightly-random firing of the neuron can compensate for this. In other words, when a

---

[3] K. Beckmann, J. Holt, H. Manem, J. Van Nostrand, and N. C. Cady, "Nanoscale hafnium oxide RRAM devices exhibit pulse dependent behavior and multi-level resistance capability," *MRS Adv.*, vol. 1, no. 49, pp. 3355–3360, 2016.
[4] K. Beckmann, J. Holt, W. Olin-Ammentorp, Z. Alamgir, J. Van Nostrand, and N. C. Cady, "The effect of reactive ion etch (RIE) process conditions on ReRAM device performance," *Semicond. Sci. Technol.*, vol. 32, no. 9, 2017.
[5] T. Branco and K. Staras, "The probability of neurotransmitter release: Variability and feedback control at single synapses," *Nat. Rev. Neurosci.*, vol. 10, no. 5, pp. 373–383, 2009.
[6] L. Buesing, J. Bill, B. Nessler, and W. Maass, "Neural dynamics as sampling: A model for stochastic computation in recurrent networks of spiking neurons," *PLoS Comput. Biol.*, vol. 7, no. 11, 2011.
[7] H. E. Plesser and W. Gerstner, "Noise in integrate-and-fire neurons: From stochastic input to escape rates," *Neural Comput.*, vol. 12, no. 2, pp. 367–384, 2000.

deterministic neuron produces the wrong behavior, it is always wrong – but when the neuron has a stochastic component, there is still a probability it can produce the correct behavior.

To demonstrate this, we created two sets of spiking networks for a pole-balancing control problem via evolutionary optimization[8]. One set of these networks used noise-free neurons, and the other used noisy neurons. The weights for all neurons in these networks were then collected and perturbed away from their ideal values as determined by the optimizer, and the fitness of the network – its ability to balance the pole - was re-measured. The magnitude of the perturbations was executed over a range of values to determine the sensitivity of the networks to lack of precision in the synapses.

We found that the set of networks utilizing noisy neurons displayed a significantly greater tolerance to weight perturbation. This tolerance was measured by estimating the magnitude of weight perturbation required to reduce a network's average fitness to 50% of its original value, and was approximately 5 times greater for noisy networks. Some exceptional networks, however, continue to display good performance even longer than this average threshold.

One disadvantage of including noise in the neuron is that the network's behavior becomes non-deterministic; its performance at the same task can vary from run-to-run. As a result, one important stage of optimizing these networks is verifying that they produce consistent behavior. To accomplish this, all solutions were run at least ten times to ensure that their average fitness was at least 95% of the desired goal.

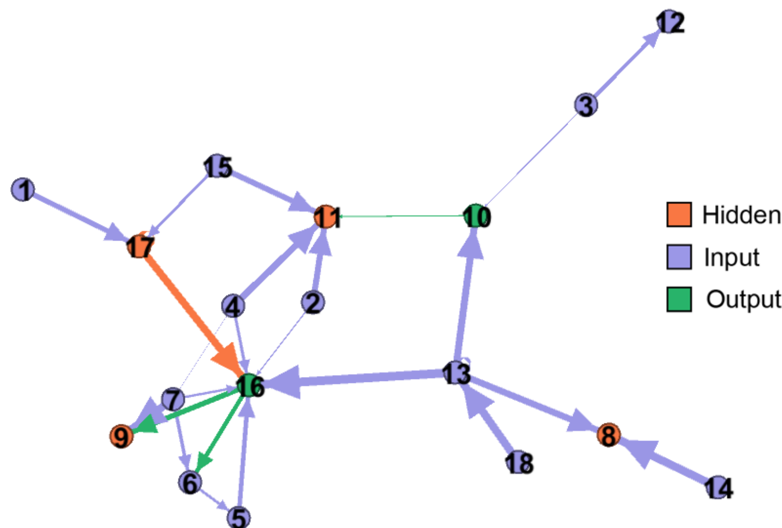


Figure 16. *Example of the graph of a spiking network evolved to solve a pole-balancing task. Each edge between neurons has a weight attached to it, which is being perturbed to determine the network's sensitivity to synaptic noise.*

[8] J. S. Plank *et al.*, "The TENNLab Exploratory Neuromorphic Computing Framework," 2018.

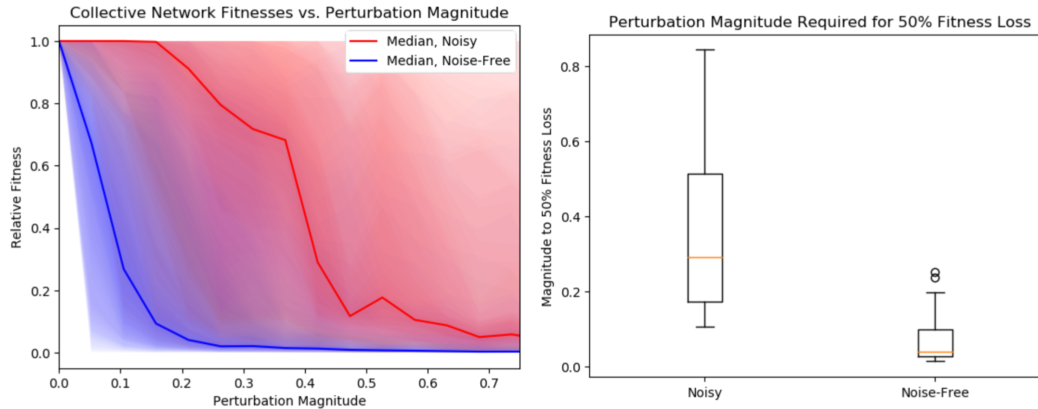


Figure 17. *Results of perturbations made to noisy and noise-free networks. The noisy networks decay much less quickly than the noise-free networks, and the magnitude of perturbation required to reduce them to 50% performance is approximately 5 times greater.*

Spiking networks operating on an architecture utilizing memristive synapses such as mrDANNA have the potential to create a highly flexible, powerful, low-power computing system. However, these networks must also be able to run under the conditions which physical hardware impose rather than in the perfect world of simulation. In this work, we investigated the effects which limited memristor precision have on these networks, and found that stochastic neurons can increase tolerance to imprecise synapses.

5.0 Conclusions

The following major accomplishments were achieved during this effort:

1. A novel memristive material stack (Hf/CeOx/TaOx/Pt) was developed that yielded outstanding endurance and retention behavior, as well as robust analog (multi-level) switching results.
2. Hafnium oxide ReRAM devices were integrated with CMOS (using our standard 1T1R approach) with optimization of the deposition and etching conditions of the ReRAM material stack, resulting in robust memory window and retention. This enabled interrogation of 8x8 1T1R arrays and demonstration of analog memory storage within these arrays with 100% device yield.
3. Back end of the line (BEOL) metallization was completed up to the aluminum pad level (LB) and wafers were delivered to collaborators at UT-Knoxville for testing.
4. Hafnium oxide 1T1R devices were characterized for variability (variation during retention experiments, and during analog switching). These results were used for subsequent studies on the effects of variability on neural network performance.
5. A so-called "noisy" neural network training approach was developed, in which the variability of synaptic devices was incorporated into the training model. The resulting "noisy" networks were more robust to perturbation and therefore are predicted to be more stable, especially when implemented with inherently stochastic devices (such as memristors).

The results of this work will be used to inform ongoing research efforts at SUNY Poly that are focused on memristive device fabrication, and development of neuromorphic hardware for AFRL applications. More broadly, our demonstration of a new resistive material stack and approaches towards training neural networks have application for a variety of machine learning, neuromorphic computing, and AI efforts.

6.0 Publications and Patent Applications Resulting from this Project

1. **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.* https://apps.dtic.mil/dtic/tr/fulltext/u2/1052240.pdf
2. J. Hazra, M. Liehr, K. Beckmann, C. Hobbs, M. Rodgers, **N.C. Cady.** Impact of Organic and Chlorine Based ALD Precursors on Hafnium Dioxide Based Nanoscale RRAM Devices. The 61st Electronic Materials Conference (EMC), Michigan, USA, 2019.
3. K. Beckmann, N. Suguitan, J. Van Nostrand, **N.C. Cady.** Interface Modification of HfO_2-based ReRAM via Low Temperature Anneal. (2019) *Semiconductor Science & Technology.* 34: 105021. https://doi.org/10.1088/1361-6641/ab362a
4. W. Olin-Ammentorp, **N.C. Cady.** Training Spiking Networks via Natural Evolutionary Strategies. (2019) *Proceedings of the International Conference on Neuromorphic Systems (ICONS '19). Association for Computing Machinery.* 18:1–6. https://doi.org/10.1145/3354265.3354283
5. W. Olin-Ammentorp, **N.C. Cady.** Biologically-inspired Neuromorphic Computing. 2019. *Science Progress.* doi.org/10.1177/0036850419850394
6. M. Liehr, J. Hazra, K. Beckmann, W. Olin-Ammentorp, **N. Cady**, R. Weiss, S. Sayyaparaju, G. Rose, J. Van Nostrand. Fabrication and Performance of Hybrid ReRAM-CMOS Circuit Elements for Dynamic Neural Networks. (2019) *Proceedings of the International Conference on Neuromorphic Systems (ICONS '19).* Association for Computing Machinery. 6:1–4. https://doi.org/10.1145/3354265.3354271
7. K. Beckmann, N. Suguitan, J. Van Nostrand, **N.C. Cady.** Interface Modification of HfO_2-based ReRAM via Low Temperature Anneal. 2019. *Semiconductor Science & Technology.* 34: 105021.
8. 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
9. W. Olin-Ammentorp, K. Beckmann, C.D. Schuman, J.S. Plank, **N.C. Cady.** Stochasticity and Robustness in Spiking Neural Networks. (2021) *Elsevier Journal of Neural Networks.* 419:1.

7.0 List of Acronyms

1T1R: 1 transistor 1 memristor (memory cell containing 1 transistor and 1 memristor)
AFRL: Air Force Research Laboratory
AI: Artificial Intelligence
BE: Bottom electrode
CMOS: Complimentary Metal Oxide Semiconductor
CNSE: College of Nanoscale Science and Engineering
DC: Direct Current
FWHM: Full Width Half Maximum
GUI: Graphical User Interface
HRS: High resistance state
I&F: Integrate and fire
IC: Integrated Circuit
LRS: Low resistance state
R&D: Research & Development
ReRAM: Resistive random access memory
RMD: Resistive memory device
RRAM: Resistive random access memory
SMU: Source/Measure Units
SUNY: State University of New York
TE: Top electrode
Vg: Gate voltage